Karel Zvára

R & Regrese

Verze ze dne 19. prosince 2003

Obsah

1 Úvod **7**

2 Model **9**

2.1 Lineární model . . . 9
2.2 Odhad vektoru středních hodnot . . . 10
2.3 Rezidua . . . 11
2.4 Normální rovnice . . . 12
2.5 Odhadnutelné funkce . . . 13
2.6 Normální lineární model . . . 15
2.7 Normální model s plnou hodností . . . 17
2.8 Vážený lineární model . . . 18

3 Podmodel **21**

3.1 Podmodel . . . 21
3.2 Vypuštění sloupců . . . 23
3.3 Lineární omezení na parametry . . . 24
3.4 Předem daná hodnota regresního koeficientu . . . 25
3.5 Koeficient determinace . . . 26
3.6 Vážený lineární model . . . 29

4 Regresní funkce lineární v parametrech **31**

4.1 Jedna přímka . . . 31
4.2 Obecnější funkce . . . 33
4.3 Pás spolehlivosti pro regresní funkci . . . 34
4.4 Několik přímek . . . 35
4.5 Inverzní predikce . . . 39

5 Identifikace **43**

5.1 Nejkratší řešení . . . 43
5.2 Identifikační omezení . . . 44

6 Analýza rozptylu **47**

6.1 Jednoduché třídění . . . 47
6.2 Analýza rozptylu dvojného třídění . . . 57
6.3 Analýza kovariance . . . 62

7 Následky nesplnění předpokladů **65**

7.1 Prostor středních hodnot . . . 65
7.2 Případ s úplnou hodností . . . 68
7.3 Varianční matice . . . 70
7.4 Typ rozdělení . . . 74

8 Rezidua **79**
8.1 Vynechání jednoho pozorování 79
8.2 Studentizovaná rezidua 81
8.3 Vliv jednotlivých pozorování 84
8.4 Nabídka prostředí R 87
8.5 Nekorelovaná rezidua 89
8.6 Parciální rezidua 90
8.7 Grafy reziduí 91

9 Testy **93**
9.1 Tvar závislosti 93
9.2 Rozptyl 97
9.3 Normalita 105
9.4 Nezávislost 108

10 Multikolinearita **111**
10.1 Teorie 111
10.2 Regrese standardizovaných veličin 113

11 Hledání modelu **121**
11.1 Dvě kritéria 121
11.2 Porovnání modelu a podmodelu 123
11.3 Sekvenční postupy 127
11.4 Praxe hledání modelu 130
11.5 Transformace 133

12 Logistická regrese **137**
12.1 Tvar závislosti 137
12.2 Odhad parametrů 138
12.3 Interpretace parametrů 140
12.4 Testování podmodelu 145
12.5 Tři druhy studií 150
12.6 Diagnostika 153

13 Zobecněný lineární model **155**
13.1 Rozdělení exponenciálního typu 155
13.2 Zobecněný lineární model 158

14 Model nelineární regrese **169**
14.1 Předpoklady 169
14.2 Lineární aproximace 170
14.3 Testování jednoduché hypotézy o $\boldsymbol{\theta}$ 171
14.4 Testování složené hypotézy 173

15 Parametrizace v NLR **177**
15.1 Označení 177
15.2 Odhad vychýlení 179
15.3 Dvojí parametrizace 181
15.4 Míry křivosti 183

A Pomocná tvrzení, označení **185**
A.1 Tvrzení o maticích 185
A.2 Některé vlastnosti náhodných veličin 190
A.3 Metoda maximální věrohodnosti 190

B Prostředí R **193**
B.1 Procedura `lm()` 193
B.2 Vlastní procedury 199

C Data **201**

Text je určen především studentům matematické statistky či ekonometrie na matfyzu Univerzity Karlovy v Praze. Doufám však, že bude užitečný i dalším zájemcům, kteří budou ochotni překonat použitý formalismus. Nijak nechci skrývat dojem, který na mě udělalo programové prostředí R, proto jsem jej učinil svým hlavním fámulem. Doufám, že jsem tím neodradil příliš zájemců o regresní model, *erko opravdu stojí za to*!
KZv.

V Praze dne 19. prosince 2003.

1. Úvod

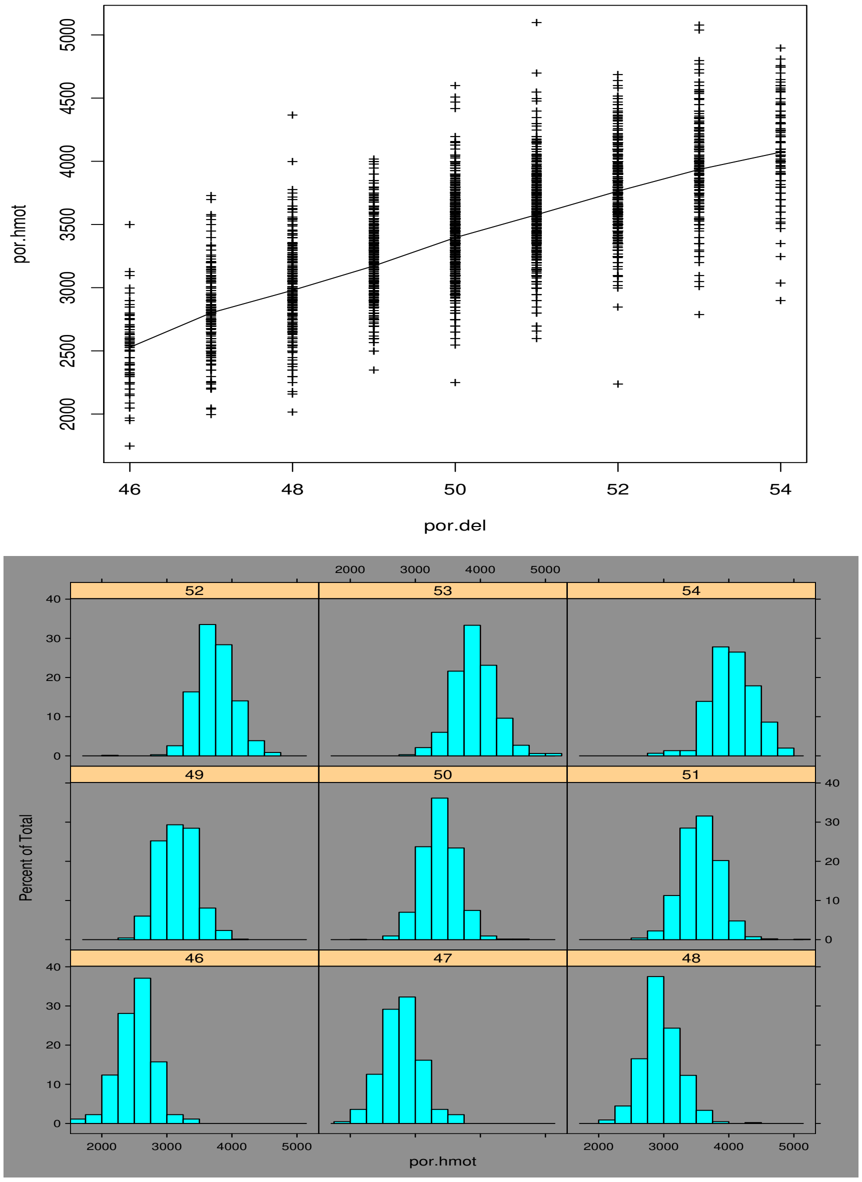
Začněme původem slova regrese. Ve stejném smyslu, jak jej používá tato kniha, použil pojem regrese jako první Francis Galton, když vyšetřoval závislost výšky postavy syna na výšce otce (Galton (1886a), Galton (1886b)).

Představme si dvě skupiny synů. První je charakterizována tím, že otcové mají výšku řekněme 170 cm, která je současně také průměrnou výškou v populaci otců. Druhá skupina synů je charakterizována tím, že výška jejich otců je rovna 180 cm, tedy o 10 cm více, než je průměrná výška všech otců. Ukazuje se, že průměrná výška synů z druhé skupiny je jen o 5 cm vyšší, než průměrná výška synů první skupiny. Odchylka výšky synů tedy sleduje odchylku výšky otců, ale nereprodukuje ji celou, redukuje ji na polovinu. Jde „o zpětný pochod, postup“ (Petráčková et al., 1995, heslo regrese) k průměru. Jak by se asi jmenovala tato kniha, kdyby se Galton zabýval jinou problematikou, např. exponenciálně rostoucími populacemi. Progrese?

Příklad 1.1 (hmotnost hochů) Jako první ukázku použijeme data, která obsahují porodní hmotnost a porodní délku celkem 4838 chlapců. V horní části obrázku 1.1 je patrné, že s rostoucí porodní délkou celkem pravidelně roste také *průměrná* porodní hmotnost. Na spodním obrázku jsou histogramy porodní hmotnosti podmíněné konkrétní hodnotou porodní délky. Je zde patrné, že pro každou porodní délku má porodní hmotnost přibližně stejné rozdělení, ovšem až na průměr, který vcelku pravidelně s hodnotou porodní délky roste.

```
> print(tapply(por.hmot,por.del,mean),digit=5)
    46     47     48     49     50     51     52     53     54
2528.1 2801.3 2979.1 3172.5 3396.1 3577.5 3763.9 3935.8 4072.5
> print(diff(tapply(por.hmot,por.del,mean)),digit=4)
   47    48    49    50    51    52    53    54
273.2 177.9 193.3 223.6 181.4 186.4 171.9 136.7
> print(mean(diff(tapply(por.hmot,por.del,mean))),digit=4)
[1] 193.1
```

S každým centimetrem porodní délky se tedy průměrná porodní hmotnost zvětšuje o necelých 200 gramů. ◯



Obrázek 1.1: Závislost porodní hmotnosti chlapců na jejich porodní délce (u histogramů uvedena v záhlaví)

2. Model

Co nového si o regresi (a lineárních modelech) můžeme říci, když je těmto tématům věnováno v každé praktičtější knížce o statistice tolik místa? Pokusíme se o jiný pohled. Uvidíme, že vlastní odhad parametrů v regresi je jen jednou dílčí úlohou, že v mnoha ohledech důležitější (a zajímavější) úlohou je odhad vektoru středních hodnot závisle proměnné. Na poslední úloze je založena například téměř celá diagnostika. Samotný výklad bude do značné míry založen na geometrickém pohledu.

2.1. Lineární model

Předpokládejme se, že střední hodnoty nekorelovaných náhodných veličin $Y_1, \ldots, Y_n$ lze popsat pomocí $k+1$ neznámých lineárních parametrů jako

$$\mathsf{E}\, Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \tag{2.1}$$

kde x_{ij} jsou známé konstanty. Až na výjimky budeme dál předpokládat $\mathsf{var}\, Y_i = \sigma^2$, kde $\sigma > 0$ je další zpravidla neznámý parametr. Známé konstanty x_{ij} uspořádáme do matice konstant o n řádcích a $k+1$ sloupcích

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} \tag{2.2}$$

takové, že $\mathsf{h}(\mathbf{X}) = r > 0$ a $n > r$. Náhodný vektor $\mathbf{Y}$ má pak střední hodnotu $\mathbf{X}\boldsymbol{\beta}$ a varianční matici $\sigma^2\mathbf{I}$. Požadavek na střední hodnotu je vlastně požadavkem $\mathsf{E}\,\mathbf{Y} \in \mathcal{M}(\mathbf{X})$, neboť lineární prostor $\mathcal{M}(\mathbf{X})$ je tvořen právě všemi lineárními kombinacemi sloupců matice $\mathbf{X}$ (viz Appendix). Předpokládaná varianční matice znamená stejný rozptyl a nekorelovanost jednotlivých složek náhodného vektoru $\mathbf{Y}$. Uvedené předpoklady budeme stručně zapisovat jako $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Ekvivalentně můžeme lineární model zapsat pomocí chybového členu $\mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I})$ jako $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$.

Zaveďme speciální označení. Nechť sloupce matice $\mathbf{Q}$ tvoří nějakou ortonormální bázi *regresního prostoru* $\mathcal{M}(\mathbf{X})$, nechť sloupce matice $\mathbf{N}$ doplní tuto bázi na ortonormální bázi prostoru $\mathbb{R}^n$. Dostaneme tak ortonormální matici $\mathbf{P} = (\mathbf{Q}, \mathbf{N})$ takovou, že $\mathcal{M}(\mathbf{X}) = \mathcal{M}(\mathbf{Q})$, $\mathbf{PP}' = \mathbf{I}_n$ a $\mathbf{P}'\mathbf{P} = \mathbf{I}_n$. Z toho, že sloupce matice $\mathbf{P}$ jsou ortonormální, plynou vztahy

$$\mathbf{QQ}' + \mathbf{NN}' = \mathbf{I}_n, \qquad \mathbf{Q}'\mathbf{Q} = \mathbf{I}_r, \qquad \mathbf{N}'\mathbf{N} = \mathbf{I}_{n-r}, \qquad \mathbf{Q}'\mathbf{N} = \mathbf{O}.$$

Označme $\mathbf{H} = \mathbf{QQ}'$ a $\mathbf{M} = \mathbf{NN}'$. Obě nově zavedené matice jsou symetrické a idempotentní. Protože platí $\mathbf{HM} = \mathbf{O}$, jsou sčítanci na pravé straně vztahu

$$\mathbf{y} = \mathbf{Hy} + \mathbf{My}$$

navzájem ortogonální, takže jde o průměty obecného vektoru $\mathbf{y} \in \mathbb{R}^n$ do regresního prostoru $\mathcal{M}(\mathbf{X})$ a *reziduálního prostoru* $\mathcal{M}(\mathbf{X})^\perp$. Ze známých vlastností projekce jsou tyto průměty a tedy také projekční matice $\mathbf{H}, \mathbf{M}$ dány jednoznačně. Navíc je vektor $\hat{\mathbf{y}} = \mathbf{Hy}$ nejbližším prvkem regresního prostoru $\mathcal{M}(\mathbf{X})$ k danému vektoru $\mathbf{y}$. V dalším bude užitečné znát explicitní vyjádření projekční matice $\mathbf{H}$ pomocí regresní matice $\mathbf{X}$, která tuto projekční matici generuje. Ze známého vztahu (např. (Anděl, 1978, věta IV.15 b))) $\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X} = \mathbf{X}$ plyne, že je $(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}')\mathbf{X} = \mathbf{O}$, takže jsou sloupce matice $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$ ortogonální na $\mathcal{M}(\mathbf{X})$ a

$$\mathbf{I} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}' + (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}')$$

je hledaný rozklad $\mathbf{I} = \mathbf{H} + \mathbf{M}$. Je tedy

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}', \tag{2.3}$$

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'. \tag{2.4}$$

2.2. Odhad vektoru středních hodnot

Nejprve se budeme zabývat odhadem vektoru $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. K náhodnému vektoru $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ najdeme v podprostoru $\mathcal{M}(\mathbf{X})$ nejbližší prvek, který opět označíme stříškou, tedy $\hat{\mathbf{Y}}$.

Při porovnávání nestranných odhadů vektorového parametru použijeme jejich varianční matice. Jsou-li $\hat{\mathbf{Y}}, \tilde{\mathbf{Y}}$ dva odhady vektoru $\boldsymbol{\mu}$, pak je odhad $\hat{\mathbf{Y}}$ lepší, když je matice $\operatorname{var}\tilde{\mathbf{Y}} - \operatorname{var}\hat{\mathbf{Y}}$ pozitivně semidefinitní. Znamená to, že také pro každý vektor $\mathbf{q} \in \mathbb{R}^n$ je $\operatorname{var}\mathbf{q}'\hat{\mathbf{Y}} \leq \operatorname{var}\mathbf{q}\tilde{\mathbf{Y}}$.

Věta 2.1. (Gaussova-Markovova) V modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ je $\hat{\mathbf{Y}}$ *nejlepším nestranným lineárním odhadem* (NNLO) vektoru $\mathbf{X}\boldsymbol{\beta}$, přičemž platí $\operatorname{var}\hat{\mathbf{Y}} = \sigma^2\mathbf{H}$.

D ů k a z: Nestrannost odhadu plyne ze známé vlastnosti projekce do podprostoru. Prvek podprostoru se promítne sám na sebe (je sám sobě nejbližším prvkem podprostoru), což má za následek mimo jiné, že platí nutně

$$\mathbf{H}\mathbf{X} = \mathbf{X}. \tag{2.5}$$

Proto pro každé $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ platí

$$\mathsf{E}\,\hat{\mathbf{Y}} = \mathsf{E}\,\mathbf{H}\mathbf{Y} = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}.$$

Vezměme nyní nějaký lineární odhad vektoru $\mathbf{X}\boldsymbol{\beta}$ tvaru $\tilde{\mathbf{Y}} = \mathbf{a}+\mathbf{B}\mathbf{Y}$. Požadavek nestrannosti vede k požadavku

$$\mathbf{a} + \mathbf{B}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} \qquad \text{pro všechna } \boldsymbol{\beta},$$

což je ekvivalentní s dvojicí identit $\mathbf{a} = \mathbf{0}$ a $\mathbf{B}\mathbf{X} = \mathbf{X}$. Z druhé identity postupným násobením zprava maticemi $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}', \mathbf{X}$ dostaneme

$$\mathbf{B}\mathbf{X} = \mathbf{X} \Rightarrow \mathbf{B}\mathbf{H} = \mathbf{H} \Rightarrow \mathbf{B}\mathbf{X} = \mathbf{X},$$

což znamená, že nestrannost dohadu $\tilde{\mathbf{Y}}$ je ekvivalentní s dvojicí identit $\mathbf{a} = \mathbf{0}$ a $\mathbf{B}\mathbf{H} = \mathbf{H}$.

Spočítejme varianční matici statistiky $\tilde{\mathbf{Y}}$. S ohledem na požadavek $\mathbf{B}\mathbf{H} = \mathbf{H}$ platí

$$\begin{aligned}\mathsf{var}\,\tilde{\mathbf{Y}} &= \mathbf{B}\sigma^2\mathbf{I}\mathbf{B}' = \sigma^2\,[\mathbf{H} + (\mathbf{B} - \mathbf{H})]\,[\mathbf{H} + (\mathbf{B} - \mathbf{H})]' \\ &= \sigma^2\mathbf{H}\mathbf{H}' + \sigma^2(\mathbf{B} - \mathbf{H})(\mathbf{B} - \mathbf{H})' \geq \sigma^2\mathbf{H}\mathbf{H}',\end{aligned}$$

neboť je opravdu $\mathsf{var}\,\hat{\mathbf{Y}} = \mathsf{var}\,\mathbf{H}\mathbf{Y} = \sigma^2\mathbf{H}\mathbf{H}' = \sigma^2\mathbf{H}$. □

2.3. Rezidua

Nyní se budeme zabývat průmětem vektoru $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ do prostoru reziduí $\mathcal{M}(\mathbf{X})^{\perp}$ a zavedeme nestranný odhad rozptylu σ^2. *Vektor reziduí* zavedený vztahem $\mathbf{u} = \mathbf{Y} - \hat{\mathbf{Y}}$ porovnává pozorované hodnoty vysvětlované proměnné s odhadem jejich středních hodnot. *Reziduální součet čtverců* definovaný jako $RSS = \|\mathbf{u}\|^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ porovnává tyto dva vektory, udává čtverec jejich vzdálenosti. *Reziduální rozptyl* zavedeme jako $S^2 = RSS/(n-r)$.

Věta 2.2. (O reziduích) V lineárním modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ platí

$$\mathbf{u} = \mathbf{M}\mathbf{Y} = \mathbf{M}\mathbf{e}, \tag{2.6}$$

$$\mathbf{u} \sim (\mathbf{0}, \sigma^2\mathbf{M}), \tag{2.7}$$

$$RSS = \mathbf{e}'\mathbf{M}\mathbf{e}, \tag{2.8}$$

$$\mathsf{E}\, RSS = (n-r)\sigma^2, \tag{2.9}$$

$$\mathsf{E}\, S^2 = \sigma^2, \tag{2.10}$$

$$\mathbf{X}'\mathbf{u} = \mathbf{0}. \tag{2.11}$$

D ů k a z: První a poslední tvrzení plyne z $\mathbf{M}\mathbf{X} = \mathbf{O}$, druhé je jednoduchým důsledkem prvního. Vztah (2.8) je prostým přepisem čtverce délky vektoru reziduí. Při důkazu tvrzení (2.9) lze použít tvrzení o stopě projekční matice (A.18)

$$\mathsf{E}\,\mathbf{e}'\mathbf{M}\mathbf{e} = \operatorname{tr}\mathsf{E}\,\mathbf{e}'\mathbf{M}\mathbf{e} = \operatorname{tr}\mathbf{M}\mathsf{E}\,\mathbf{e}\mathbf{e}' = \operatorname{tr}\mathbf{M}\sigma^2\mathbf{I} = \sigma^2\operatorname{tr}\mathbf{M} = \sigma^2(n - \mathsf{h}(\mathbf{X})).$$

Vztah (2.9) je triviálním důsledkem předchozího. □

Vektor reziduí $\mathbf{u}$ lze interpretovat jako jakýsi odhad rozdílu $\mathbf{e} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$. Reziduální rozptyl S^2 je podle (2.10) nestranným odhadem rozptylu σ^2.

2.4. Normální rovnice

Zatím jsme se nezabývali odhadem vektoru $\boldsymbol{\beta}$, který vyjadřuje střední hodnotu náhodného vektoru $\mathbf{Y}$ jako konkrétní lineární kombinaci sloupců matice $\mathbf{X}$. Pokud nemá matice $\mathbf{X}$ lineárně nezávislé sloupce, nebudou koeficienty této lineární kombinace dány jednoznačně, takže lineární odhad neexistuje. (Připomeňme si, že odhad či odhadová statistika má být *funkcí* náhodných veličin.)

Symbolem $\mathbf{b}$ označíme libovolné řešení soustavy $\mathbf{X}\mathbf{b} = \hat{\mathbf{Y}}$. Vektor $\mathbf{b}$ tedy tvoří právě hledané koeficienty lineární kombinace. Skutečnost, že $\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{u}$ je ortogonální rozklad, je ekvivalentní s požadavkem, aby vektor reziduí $\mathbf{u}$ byl ortogonální vůči regresnímu prostoru $\mathcal{M}(\mathbf{X})$, tedy s požadavkem

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{b}) = \mathbf{0},$$

což je opět ekvivalentní s *normální rovnicí* pro $\mathbf{b}$

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}. \tag{2.12}$$

Všimněte si, že tato soustava lineárních rovnic je vždy řešitelná, neboť na obou stranách je nějaká lineární kombinace řádků matice $\mathbf{X}$.

2.5. Odhadnutelné funkce

I v případě, že vektor $\boldsymbol{\beta}$ nelze odhadnout, protože rovnice (2.12) může mít nekonečně mnoho řešení, mohou být odhadnutelné některé jeho lineární funkce. Například k vektoru $\mathbf{X}\boldsymbol{\beta}$ známe dokonce nejlepší nestranný lineární odhad a každá složka tohoto vektoru je lineární funkcí vektoru $\boldsymbol{\beta}$.

Připomeňme si význam Gaussovy-Markovovy věty. Pro každé $\mathbf{q} \in \mathbb{R}^n$ je statistika $\mathbf{q}'\hat{\mathbf{Y}}$ nejlepším nestranným lineárním odhadem své střední hodnoty, tedy odhadem funkce

$$\mathsf{E}\,\mathbf{q}'\mathbf{Y} = \mathbf{q}'\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}'\mathbf{q})'\boldsymbol{\beta} = \mathbf{t}'\boldsymbol{\beta}.$$

Řekneme, že $\mathbf{t}'\boldsymbol{\beta}$ je *odhadnutelná funkce* v modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, když existuje její nestranný lineární odhad.

Věta 2.3. (Odhadnutelná funkce) Funkce $\mathbf{t}'\boldsymbol{\beta}$ je odhadnutelná právě, když platí některá z podmínek

a) $\mathbf{t} \in \mathcal{M}(\mathbf{X}') = \mathcal{M}(\mathbf{X}'\mathbf{X})$,

b) $\mathbf{t}'\mathbf{b}$ nezávisí na volbě řešení normální rovnice.

Jsou-li $\mathbf{t}'\boldsymbol{\beta}$, $\mathbf{t}_1'\boldsymbol{\beta}$ a $\mathbf{t}_2'\boldsymbol{\beta}$ odhadnutelné funkce, pak bez ohledu na volbu zobecněné inverzní matice platí

$$\begin{aligned} \mathsf{E}\,\mathbf{t}'\mathbf{b} &= \mathbf{t}'\boldsymbol{\beta}, \\ \mathsf{var}\,\mathbf{t}'\mathbf{b} &= \sigma^2\mathbf{t}'(\mathbf{X}'\mathbf{X})^-\mathbf{t}, \\ \mathsf{cov}(\mathbf{t}_1'\mathbf{b}, \mathbf{t}_2'\mathbf{b}) &= \sigma^2\mathbf{t}_1'(\mathbf{X}'\mathbf{X})^-\mathbf{t}_2. \end{aligned}$$

D ů k a z: Má-li být $\mathbf{t}'\boldsymbol{\beta}$ odhadnutelná funkce, musí existovat taková lineární funkce $\mathbf{q}'\mathbf{Y}$, že její střední hodnota $(\mathbf{X}'\mathbf{q})'\boldsymbol{\beta}$ je rovna $\mathbf{t}'\boldsymbol{\beta}$ pro všechna $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$. To je ekvivalentní s požadavkem $\mathbf{t} = \mathbf{X}'\mathbf{q}$, což je ekvivalentní s požadavkem $\mathbf{t} \in \mathcal{M}(\mathbf{X}')$. Vztah $\mathcal{M}(\mathbf{X}') = \mathcal{M}(\mathbf{X}'\mathbf{X})$ je také zřejmý, neboť oba prostory patří do $\mathbb{R}^{k+1}$ a jejich ortogonální doplňky jsou totožné. Obráceně, je-li $\mathbf{t} \in \mathcal{M}(\mathbf{X}')$, pak pro nějaké $\mathbf{q}$ lze $\mathbf{t}$ vyjádřit jako $\mathbf{X}'\mathbf{q}$. Lineární funkce $\mathbf{q}'\mathbf{Y}$ má střední hodnotu $\mathsf{E}\,\mathbf{q}'\mathbf{Y} = \mathbf{q}'\mathbf{X}\boldsymbol{\beta} = \mathbf{t}'\boldsymbol{\beta}$, takže je hledaným nestranným lineárním odhadem $\mathbf{t}'\boldsymbol{\beta}$.

Je-li $\mathbf{t} \in \mathcal{M}(\mathbf{X}')$, je $\mathbf{t} = \mathbf{X}'\mathbf{q}$ pro nějaké $\mathbf{q}$. Pak je ale výraz $\mathbf{t}'\mathbf{b} = \mathbf{q}'\mathbf{X}\mathbf{b} = \mathbf{q}'\hat{\mathbf{Y}}$ nezávislý na volbě řešení normální rovnice $\mathbf{b}$. Obráceně, každé řešení normální rovnice lze zapsat jako $\mathbf{b} = \mathbf{b}^+ + \mathbf{c}$, kde $\mathbf{c}$ je libovolný vektor, který splňuje $\mathbf{X}'\mathbf{X}\mathbf{c} = \mathbf{0}$ a tedy $\mathbf{X}\mathbf{c} = \mathbf{0}$, resp. $\mathbf{c} \in \mathcal{M}(\mathbf{X}')^\perp$. Výraz $\mathbf{t}'\mathbf{b} = \mathbf{t}'\mathbf{b}^+ + \mathbf{t}'\mathbf{c}$ tedy nezávisí na volbě řešení normální rovnice, když pro všechna $\mathbf{c} \in \mathcal{M}(\mathbf{X}')^\perp$ platí $\mathbf{t}'\mathbf{c} = 0$, tedy když je $\mathbf{t} \in (\mathcal{M}(\mathbf{X}')^\perp)^\perp = \mathcal{M}(\mathbf{X}')$.

Doplňme ještě, že v případě odhadnutelné funkce je $\mathbf{t}'\mathbf{b}$ hledaným nestranným odhadem $\mathbf{t}'\boldsymbol{\beta}$, neboť platí $\mathsf{E}\,\mathbf{t}'\mathbf{b} = \mathbf{q}'\mathsf{E}\,\hat{\mathbf{Y}} = \mathbf{q}'\mathbf{X}\boldsymbol{\beta} = \mathbf{t}'\boldsymbol{\beta}$. Jsou-li $\mathbf{t}_1'\boldsymbol{\beta}$ a $\mathbf{t}_2'\boldsymbol{\beta}$ dvě odhadnutelné funkce, můžeme pakbez ohledu na volbu pseudoinverze psát

$$\begin{aligned} \mathsf{cov}(\mathbf{t}_1'\mathbf{b}, \mathbf{t}_2'\mathbf{b}) = \mathsf{cov}(\mathbf{q}_1'\mathbf{X}\mathbf{b}, \mathbf{q}_2'\mathbf{X}\mathbf{b}) &= \mathsf{cov}(\mathbf{q}_1'\hat{\mathbf{Y}}, \mathbf{q}_2'\hat{\mathbf{Y}}) \\ &= \sigma^2\mathbf{q}_1'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{q}_2 = \sigma^2\mathbf{t}_1'(\mathbf{X}'\mathbf{X})^-\mathbf{t}_2. \qquad \square \end{aligned}$$

Jednoduchým důsledkem právě dokázané věty je následující tvrzení.

Věta 2.4. Vektor $\mathbf{T}'\boldsymbol{\beta}$ je vektorem odhadnutelných funkcí právě, když platí $\mathcal{M}(\mathbf{T}) \subset \mathcal{M}(\mathbf{X}')$. Potom pro každé řešení normální rovnice platí

$$\mathbf{T}'\mathbf{b} \sim (\mathbf{T}'\boldsymbol{\beta}, \sigma^2\mathbf{T}'(\mathbf{X}'\mathbf{X})^-\mathbf{T}),$$

přičemž nezávisí na volbě zobecněné inverzní matice.

Příklad 2.1 (jednoduché třídění) Úloha analýzy rozptylu jednoduchého třídění předpokládá, že pro nezávislé náhodné veličiny Y_{it}, kde je $1 \leq t \leq n_I$, $1 \leq i \leq I$, platí $Y_{it} \sim \mathsf{N}(\mu_i, \sigma^2)$. Takto máme vlastně I nezávislých náhodných výběrů z normálních rozdělení, které mají obecně nestejné střední hodnoty, ale stejné rozptyly. V praktických úlohách vlastně třídíme hodnoty spojité veličiny Y podle nějakého *faktoru*, tedy podle znaku (veličiny) měřeného v nominálním měřítku. Jednotlivé hodnoty faktoru se nazývají úrovně či *ošetření*.

Častěji se používá parametrické vyjádření středních hodnot ve tvaru

$$\mathsf{E}\,Y_{it} = \mu + \alpha_i, \tag{2.13}$$

kde α_i jsou *efekty* (také někdy hlavní efekty), odpovídající jednotlivým úrovním sledovaného faktoru (jednotlivým ošetřením). Model můžeme zapsat jako

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_I \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{1} \end{pmatrix} \begin{pmatrix} \mu \\ \boldsymbol{\alpha} \end{pmatrix} + \mathbf{e}, \tag{2.14}$$

kde $\mathbf{e} \sim \mathsf{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Snadno zjistíme, že matice modelu $\mathbf{X}$ má hodnost I, kdežto sloupců má $I+1$, takže celý vektor parametrů není odhadnutelný. Snadno se také zjistí, že každou lineární kombinaci řádků matice $\mathbf{X}$, tedy každý vektor $\mathbf{t}'$ určující odhadnutelnou lineární funkci $\mathbf{t}'\boldsymbol{\beta}$, lze zapsat jako

$$\mathbf{t}' = \left(\sum_{i=1}^{I} c_i, c_1, \dots, c_I \right), \tag{2.15}$$

kde c_i jsou libovolné konstanty. K odhadnutelným funkcím patří například střední hodnoty jednotlivých pozorování $\mathsf{E}\,Y_{it} = \mu + \alpha_i$ (volbou $\mathbf{t}' = (1, 0, \dots, 1, 0, \dots, 0)$). Volbou $\mathbf{t}' = (0, \dots, 1, 0, \dots, 0, -1, 0, \dots)$ můžeme pro $1 \leq i \neq j \leq I$ vyjádřit rozdíly hlavních efektů $\alpha_i - \alpha_{i'}$, které patří mezi *kontrasty*. ○

Příklad 2.2 (analýza kovariance) Uvažujme nyní poněkud složitější model, než v předchozím příkladě. Nechť platí

$$Y_{it} = \mu + \alpha_i + \beta x_{it} + e_{it}, \qquad 1 \leq t \leq n_i, 1 \leq i \leq I, \tag{2.16}$$

kde opět jsou $e_{11}, \ldots, e_{In_I}$ nezávislé náhodné veličiny s nulovou střední hodnotou a rozptylem σ^2 a $x_{11}, \ldots, x_{In_I}$ jsou známé konstanty. Zajímá nás, kdy je parametr β odhadnutelný. Tentokrát má regresní matice tvar

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{x}_1 \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} & \mathbf{x}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ \mathbf{1}_{n_I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_I} & \mathbf{x}_I \end{pmatrix}.$$

Abychom mohli vyjádřit vektor $\mathbf{t} = (0, 0, \ldots, 0, 1)'$ ve tvaru $\mathbf{q}'\mathbf{X}$, musí pro všechna i být $\mathbf{q}_i'\mathbf{1}_{n_i} = 0$. Odtud je ovšem zaručena také první nula vektoru $\mathbf{t}$. Abychom získali jedničku na posledním místě vektoru $\mathbf{t}$, nesmí pro všechna i být $\mathbf{q}_i'\mathbf{x}_i = 0$. Stačí tedy, aby aspoň pro nějaké i^* bylo $\mathbf{q}_{i^*}'\mathbf{x}_{i^*} \neq 0$. Vezmeme-li v úvahu, požadavek $\mathbf{q}_{i^*}'\mathbf{1}_{n_i} = 0$, je zřejmé, že stačí, aby vektor $\mathbf{x}_{i^*}$ měl aspoň dvě nestejné složky.

Prakticky použijeme popisovaný model, když potřebujeme nejprve hodnoty závisle proměnné Y_{it} *adjustovat* vůči nějaké doprovodné veličině x. Model předpokládá lineární závislost střední hodnoty Y na x, přičemž regresní přímky v jednotlivých skupinách jsou rovnoběžné (mají stejnou směrnici β). Potom se zpravidla zajímáme, zda jsou tyto přímky dokonce totožné ($\alpha_1 = \ldots = \alpha_I$). ○

2.6. Normální lineární model

Předpokládejme navíc, že náhodný vektor $\mathbf{Y}$ má normální rozdělení, tedy že platí $\mathbf{Y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. V takovém případě hovoříme o *normálním lineárním modelu.* Připomeňme si ortonormální bázi prostoru $\mathbb{R}^n$ určenou maticí $\mathbf{P} = (\mathbf{Q}, \mathbf{N})$ s předpokladem $\mathsf{h}(\mathbf{X}) = r > 0$ a upřesněme vlastnosti statistik $\hat{\mathbf{Y}}, \mathbf{u}, RSS, S^2$.

Pro $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ můžeme psát

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{He} + \mathbf{Me} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Q}(\mathbf{Q}'\mathbf{e}) + \mathbf{N}(\mathbf{N}'\mathbf{e}) \\ &= (\mathbf{X}\boldsymbol{\beta} + \mathbf{Q}\mathbf{V}) + \mathbf{N}\mathbf{U} \\ &= \hat{\mathbf{Y}} + \mathbf{u}, \end{aligned} \tag{2.17}$$

kde náhodný vektor

$$\frac{1}{\sigma}\mathbf{P}'\mathbf{e} = \frac{1}{\sigma}\begin{pmatrix} \mathbf{Q}' \\ \mathbf{N}' \end{pmatrix}\mathbf{e} = \frac{1}{\sigma}\begin{pmatrix} \mathbf{V} \\ \mathbf{U} \end{pmatrix} \tag{2.18}$$

vzniklý ortonormální lineární transformací z vektoru $(1/\sigma)\mathbf{e}$ s rozdělením $\mathsf{N}(\mathbf{0}, \mathbf{I})$ má zřejmě opět rozdělení $\mathsf{N}(\mathbf{0}, \mathbf{I})$. Tato vlastnost, spolu s rozkladem (2.17), umožní dokázat následující větu.

Věta 2.5. (Normální lineární model) V modelu $\mathbf{Y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ platí a)

$$\hat{\mathbf{Y}} \sim \mathsf{N}\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H}\big);$$

b)

$$\mathbf{u} \sim \mathsf{N}\big(\mathbf{0}, \sigma^2\mathbf{M}\big);$$

c) náhodné vektory $\hat{\mathbf{Y}}, \mathbf{u}$ jsou nezávislé;

d)

$$\frac{1}{\sigma^2}||\hat{\mathbf{Y}}||^2 \sim \chi^2\big(r; ||\mathbf{X}\boldsymbol{\beta}||^2/\sigma^2\big);$$

e)

$$\frac{1}{\sigma^2}RSS = \frac{1}{\sigma^2}\|\mathbf{u}\|^2 = \frac{1}{\sigma^2}\|\mathbf{N}\mathbf{U}\|^2 = \frac{1}{\sigma^2}\|\mathbf{U}\|^2 \sim \chi^2(n-r);$$

f) je-li $\mathbf{T}'\boldsymbol{\beta}$ vektor odhadnutelných parametrů, pak statistiky $\mathbf{T}'\mathbf{b}$ a S^2 nezávisí na volbě pseudoinverze a platí

$$\mathbf{T}'\mathbf{b} \sim \mathsf{N}\big(\mathbf{T}'\boldsymbol{\beta}, \sigma^2\mathbf{T}'(\mathbf{X}'\mathbf{X})^-\mathbf{T}\big).$$

g) je-li $\mathbf{t}'\boldsymbol{\beta}$ odhadnutelná funkce, pak má statistika

$$\frac{\mathbf{t}'\mathbf{b} - \mathbf{t}'\boldsymbol{\beta}}{S\sqrt{\mathsf{var}\,\mathbf{t}'\mathbf{b}}} = \frac{\mathbf{t}'\mathbf{b} - \mathbf{t}'\boldsymbol{\beta}}{S\sqrt{\mathbf{t}'(\mathbf{X}'\mathbf{X})^-\mathbf{t}}}$$

rozdělení $\mathsf{t}(n-r)$.

D ů k a z: První dvě tvrzení jsou triviální, třetí plyne z $\mathbf{H}\mathbf{M} = \mathbf{O}$, což znamená nulovou matici kovariancí vektorů $\hat{\mathbf{Y}}$ a $\mathbf{u}$. Tvrzení d) plyne z vyjádření $\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Q}\mathbf{V}$, což je součet vektoru konstant a náhodného vektoru, pro který platí $||\mathbf{Q}\mathbf{V}||^2/\sigma^2 \sim \chi^2(r)$. Výraz uvedený v d) má tedy necentrální rozdělení χ^2. Další vztah plyne ze souvislosti mnohorozměrného normálního a χ^2 rozdělení. Ttvrzení f) je jen upřesněním tvrzení věty 2.4 pro normální lineární model. Poslední tvrzení je přímým důsledkem tvrzení f), e) a definice t rozdělení. □

Poznámka Náhodný vektor $\mathbf{Y}$ má v normálním lineárním modelu hustotu

$$(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2\right),$$

takže je zřejmě odhad vektoru $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ metodou maximální věrohodnosti totožný s odhadem metodou nejmenších čtverců $\hat{\mathbf{Y}}$. Naproti tomu odhad rozptylu σ^2 metodou maximální věrohodnosti je dán vztahem

$$\widehat{\sigma^2} = \frac{RSS}{n} = \frac{n-1}{n}S^2,$$

je tedy vychýlený, byť toto vychýlení s rostoucím n konverguje k nule.

2.7. Normální model s plnou hodností

Když má matice $\mathbf{X}$ lineárně nezávislé sloupce (platí $r = k + 1$), pak má normální rovnice (2.12) jediné řešení.

Věta 2.6. (Klasický model regrese) Má-li matice $\mathbf{X}$ v normálním modelu $\mathbf{Y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ hodnost rovnou počtu jejích sloupců, potom

a) řešením normální rovnice je statistika

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}; \tag{2.19}$$

b) $\mathbf{b}$ je NNLO vektoru $\boldsymbol{\beta}$;

c) platí (označme $\mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1}$ s indexy $0 \leq i, j \leq k$)

$$\mathbf{b} \sim \mathsf{N}(\boldsymbol{\beta}, \sigma^2\mathbf{V});$$

d) náhodné vektory $\mathbf{b}$ a $\mathbf{u}$ jsou nezávislé;

e) statistiky $\mathbf{b}$ a S^2 jsou nezávislé;

f) pro $j = 0, 1, \ldots, k$ platí

$$T_j = \frac{b_j - \beta_j}{S\sqrt{v_{jj}}} \sim \mathsf{t}(n - k - 1); \tag{2.20}$$

g) množina

$$\mathcal{K}_2 = \left\{\boldsymbol{\beta} \in \mathbb{R}^{k+1} : (\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) \leq (k + 1)S^2 F_{k+1,n-k-1}(\alpha)\right\} \tag{2.21}$$

tvoří konfidenční množinu pro $\boldsymbol{\beta}$ se spolehlivostí $1 - \alpha$.

D ů k a z: První tvrzení plyne z regularity matice $\mathbf{X}'\mathbf{X}$. Odhad $\mathbf{b}$ lze napsat ve tvaru $\mathbf{b} = \mathbf{V}\mathbf{X}'\hat{\mathbf{Y}}$, odkud je zřejmé, že tento vektor je lineární funkcí $\hat{\mathbf{Y}}$. Proto podle Gaussovy-Markovovy věty je nejlepším nestranným lineárním odhadem své střední hodnoty, tedy vektoru $\boldsymbol{\beta}$. Z věty 2.5 plyne nezávislost uvedená v bodech d) a e). K důkazu vztahu f) je třeba si uvědomit nezávislost uvedenou v e). Upravíme-li statistiku T_j na tvar

$$T_j = \frac{\dfrac{b_j - \beta_j}{\sqrt{\sigma^2 v_{jj}}}}{\sqrt{\dfrac{(n - k - 1)S^2}{\sigma^2}\dfrac{1}{n - k - 1}}},$$

je patrné, že symbolicky jde o zlomek

$$\frac{\mathsf{N}(0, 1)}{\sqrt{\dfrac{\chi^2(n - k - 1)}{n - k - 1}}}.$$

To, spolu se zmíněnou nezávislostí, k důkazu rozdělení statistiky T_j stačí. Podobně, s využitím c), dostaneme také konfidenční množinu popsanou v g). □

2.8. Vážený lineární model

Někdy je vhodné umět řešit poněkud obecnější úlohu, než jsme dělali až doposud. Nechť platí lineární model s obecnější varianční maticí, tj.

$$\mathbf{Y} \sim \big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1}\big). \tag{2.22}$$

Také tentokrát jsou $\boldsymbol{\beta}$ a $\sigma > 0$ neznámé parametry a $\mathbf{W}$ je zpravidla známá pozitivně definitní matice. Příkladem takového modelu je situace, kdy i-tá složka vektoru $\mathbf{Y}$ je průměrem n_i nezávislých pozorování se stejnou střední hodnotou a stejným rozptylem σ^2. Potom je $\operatorname{var} Y_i = \sigma^2/n_i$ pro každé i a matice $\mathbf{W}$ je diagonální s četnostmi $n_1, \dots, n_n$ na diagonále.

Abychom našli v modelu (2.22) protějšky $\hat{\mathbf{Y}}_W$ a S_W^2 ke statistikám $\hat{\mathbf{Y}}$ a S^2 (případně $\mathbf{b}_W$ jako protějšek k $\mathbf{b}$), převedeme nejprve model s obecnější varianční maticí na standardní model.

Protože matice $\mathbf{W}$ je pozitivně definitní, existuje regulární matice $\mathbf{C}$, která splňuje požadavek $\mathbf{C}'\mathbf{C} = \mathbf{W}$ (tuto odmocninovou matici lze zkonstruovat například pomocí spektrálního rozkladu matice $\mathbf{W}$). Zřejmě platí $\mathbf{C}\mathbf{W}^{-1}\mathbf{C}' = \mathbf{I}$.

Zaveďme matici $\mathbf{X}^* = \mathbf{C}\mathbf{X}$ a uvažujme náhodný vektor $\mathbf{Y}^* = \mathbf{C}\mathbf{Y}$, který již vyhovuje běžnému lineárnímu modelu

$$\mathbf{Y}^* \sim \big(\mathbf{C}\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{C}\mathbf{W}^{-1}\mathbf{C}'\big) = \big(\mathbf{X}^*\boldsymbol{\beta}, \sigma^2\mathbf{I}\big).$$

Spočítejme v novém (hvězdičkovém) modelu běžný odhad vektoru středních hodnot

$$\begin{aligned}\hat{\mathbf{Y}}^* &= \mathbf{H}^*\mathbf{Y}^* \\ &= \mathbf{C}\mathbf{X}(\mathbf{X}'\mathbf{C}'\mathbf{C}\mathbf{X})^-\mathbf{X}'\mathbf{C}\mathbf{C}\mathbf{Y} \\ &= \mathbf{C}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^-\mathbf{X}'\mathbf{W}\mathbf{Y}.\end{aligned}$$

Protože střední hodnota $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} = \mathbf{C}^{-1}\mathsf{E}\,\mathbf{Y}^*$ je lineární funkcí střední hodnoty $\mathsf{E}\,\mathbf{Y}^*$, platí stejný vztah i pro odhady. Je tedy odhad vektoru $\mathsf{E}\,\mathbf{Y}$ v původním modelu roven

$$\hat{\mathbf{Y}}_W = \mathbf{C}^{-1}\hat{\mathbf{Y}}^* = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^-\mathbf{X}'\mathbf{W}\mathbf{Y}.$$

Reziduální součet čtverců v modelu s hvězdičkami (jen tam má smysl, sčítáme srovnatelné hodnoty a můžeme tak najít běžný odhad σ^2) je roven

$$\begin{aligned}RSS_W = RSS^* &= \|\mathbf{Y}^* - \hat{\mathbf{Y}}^*\|^2 = \|\mathbf{C}\mathbf{Y} - \mathbf{C}\hat{\mathbf{Y}}_W\|^2 \\ &= (\mathbf{Y} - \hat{\mathbf{Y}}_W)'\mathbf{W}(\mathbf{Y} - \hat{\mathbf{Y}}_W),\end{aligned}$$

což v nejčastějším případě diagonální matice $\mathbf{W}$ vede ke statistice

$$RSS_W = \sum_{i=1}^{n} w_{ii}\left(Y_i - \hat{Y}_{Wi}\right)^2. \tag{2.23}$$

Nyní odhadneme rozptyl σ^2. Statistika

$$S_W^2 = S^{*2} = \frac{RSS^*}{n-r}$$

je zřejmě nestranným odhadem parametru σ^2. V normálním lineárním modelu $\mathbf{Y} \sim \mathsf{N}\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1}\big)$ má S_W^2 stejné rozdělení, jako statistika S^2 v běžném lineárním modelu $\mathsf{N}\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\big)$.

Má-li matice $\mathbf{X}$ lineárně nezávislé sloupce, je celý vektor $\boldsymbol{\beta}$ odhadnutelný. Řešením normální rovnice je pak

$$\begin{aligned} \mathbf{b}_W = \mathbf{b}^* &= (\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{Y}^* = (\mathbf{X}'\mathbf{C}'\mathbf{C}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}'\mathbf{C}\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}. \end{aligned} \tag{2.24}$$

Odhad vektoru středních hodnot $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ můžeme zřejmě psát jako

$$\hat{\mathbf{Y}}_W = \mathbf{X}\mathbf{b}_W. \tag{2.25}$$

Snadno se spočítá, že v modelu s úplnou hodností je $\mathbf{b}_W \sim \big(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\big)$.

Vztah (2.23) ukazuje, jak je (aspoň v případě diagonální matice $\mathbf{W}$) zobecněna metoda nejmenších čtverců. S výhodou lze tento vztah použít v programu STATISTICA, modul Nonlinear Estimation, při hledání odhadu $\mathbf{b}_W$. V programu R má procedura `lm` parametr `weights`, kterým se volí diagonální matice $\mathbf{W}$.

Shrňme dosažená zjištění.

Věta 2.7. (Zobecněná regrese) Nechť platí $\mathbf{Y} \sim \big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1}\big)$, kde $\mathbf{W} > 0$ je daná matice. Potom je vektor

$$\hat{\mathbf{Y}}_W = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}\mathbf{Y} \sim \big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\big)$$

nejlepším nestranným lineárním odhadem vektoru $\mathsf{E}\,\mathbf{Y}$. Statistika S_W^2 je nestranným odhadem rozptylu σ^2. Má-li matice $\mathbf{X}$ lineárně nezávislé sloupce, potom je také

$$\mathbf{b}_W \sim \big(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\big)$$

nejlepším nestranným lineárním odhadem vektoru $\boldsymbol{\beta}$. Jestliže má $\mathbf{Y}$ mnohorozměrné normální rozdělení, pak také $\hat{\mathbf{Y}}_W$, případně $\mathbf{b}_W$, má mnohorozměrné normální rozdělení a platí $RSS_W/\sigma^2 \sim \chi^2(n-r)$. Statistika RSS_W je v takovém případě nezávislá s $\hat{\mathbf{Y}}_W$, případně s $\mathbf{b}_W$.

3. Podmodel

Regresní metody slouží k vyhledávání a prokazování způsobu závislosti střední hodnoty nějaké náhodné veličiny na jiných veličinách. Snažíme se přitom najít model co možná nejjednodušší. Zde je klíčovým pojem podmodelu. Tato kapitola je věnována podmodelu, který v porovnání s modelem zmenšuje prostor pro možné střední hodnoty náhodného vektoru $\mathbf{Y}$.

3.1. Podmodel

Řekneme, že platí podmodel modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, když pro nějaký vektor $\boldsymbol{\beta}_0$ platí $\mathsf{E}\,\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0$, kde $\mathbf{X}_0$ je nějaká matice splňující požadavky $\mathcal{M}(\mathbf{X}_0) \subset \mathcal{M}(\mathbf{X})$, $0 < \mathsf{h}(\mathbf{X}_0) = r_0 < r$. Uvedené požadavky zaručují, že i za platnosti podmodelu je prostor možných středních hodnot netriviální, že je vlastním podprostorem původního prostoru středních hodnot modelu. Je tedy jakýmsi jeho speciálním případem.

Navážeme na úvahy o ortonormálních bázích. Vytvořme matici $\mathbf{Q}$ ze dvou podmatic, které mají po řadě r_0 a $r - r_0$ sloupců tak, aby sloupce matic $\mathbf{Q}_0$ a $(\mathbf{Q}_0, \mathbf{Q}_1)$ generovaly prostory $\mathcal{M}(\mathbf{X}_0)$ a $\mathcal{M}(\mathbf{X})$. Ortonormální matici $\mathbf{P}$, která generuje $\mathbb{R}^n$, lze pak zapsat ve tvaru

$$\mathbf{P} = (\mathbf{Q}_0, \mathbf{Q}_1, \mathbf{N}). \tag{3.1}$$

Napozorovaný vektor $\mathbf{Y}$ můžeme tedy rozložit na součet tří navzájem ortogonálních vektorů

$$\begin{aligned}
\mathbf{Y} &= \mathbf{Q}_0\mathbf{Q}_0'\mathbf{Y} + \mathbf{Q}_1\mathbf{Q}_1'\mathbf{Y} + \mathbf{N}\mathbf{N}'\mathbf{Y} && (3.2)\\
&= (\mathbf{Q}_0\mathbf{Q}_0'\mathbf{Y} + \mathbf{Q}_1\mathbf{Q}_1'\mathbf{Y}) + \mathbf{N}\mathbf{N}'\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{u} && (3.3)\\
&= \mathbf{Q}_0\mathbf{Q}_0'\mathbf{Y} + (\mathbf{Q}_1\mathbf{Q}_1'\mathbf{Y} + \mathbf{N}\mathbf{N}'\mathbf{Y}) = \hat{\mathbf{Y}}_0 + \mathbf{u}_0, && (3.4)
\end{aligned}$$

kde $\hat{\mathbf{Y}}_0, \mathbf{u}_0$ jsou po řadě odhad $\mathsf{E}\,\mathbf{Y}$ a vektor reziduí spočítané v podmodelu. Dva odhady vektoru středních hodnot i dva vektory reziduí se liší o vektor

$$\mathbf{d} = \mathbf{Q}_1\mathbf{Q}_1'\mathbf{Y}. \tag{3.5}$$

Za platnosti podmodelu pak speciálně platí

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{Q}_0(\mathbf{Q}_0'\mathbf{e}) + \mathbf{Q}_1(\mathbf{Q}_1'\mathbf{e}) + \mathbf{N}(\mathbf{N}'\mathbf{e}) \\ &= (\mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{Q}_0\mathbf{V}_0) + (\mathbf{Q}_1\mathbf{V}_1 + \mathbf{N}\mathbf{U}) = \hat{\mathbf{Y}}_0 + \mathbf{u}_0 \qquad (3.6)\\ &= (\mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{Q}_0\mathbf{V}_0 + \mathbf{Q}_1\mathbf{V}_1) + (\mathbf{N}\mathbf{U}) = \hat{\mathbf{Y}} + \mathbf{u}. \qquad (3.7)\end{aligned}$$

Máme tedy dva rozklady, které se liší podle toho, kam umístíme vektor $\mathbf{d} = \mathbf{Q}_1\mathbf{V}_1$, získaný jako průmět do podprostoru $\mathcal{M}(\mathbf{Q}_1)$, o který jsme zmenšili původní množinu možných středních hodnot vektoru $\mathbf{Y}$. Všimněme si dále, jak se chovají průměty náhodného vektoru $\mathbf{e}$ (ať už platí model či podmodel):

$$\frac{1}{\sigma}\mathbf{P}'\mathbf{e} = \frac{1}{\sigma}\begin{pmatrix}\mathbf{Q}_0'\\ \mathbf{Q}_1'\\ \mathbf{N}'\end{pmatrix}\mathbf{e} = \frac{1}{\sigma}\begin{pmatrix}\mathbf{V}_0\\ \mathbf{V}_1\\ \mathbf{U}\end{pmatrix} \sim (\mathbf{0}, \mathbf{I}). \qquad (3.8)$$

Tento rozklad použijeme k důkazu následující věty, dříve však ještě označíme reziduální součet čtverců v podmodelu $RSS_0 = \|\mathbf{u}_0\|^2$ a reziduální rozptyl v podmodelu $S_0^2 = RSS_0/(n - r_0)$.

Věta 3.1. (O podmodelu) Platí-li v lineárním modelu podmodel, potom

a) $\hat{\mathbf{Y}}_0$ je NNLO vektoru $\mathbf{X}_0\boldsymbol{\beta}_0$;

b) statistika S_0^2 je nestranným odhadem rozptylu σ^2;

c) pro vektor $\mathbf{d} = \hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0 = \mathbf{u}_0 - \mathbf{u}$ platí

$$\|\mathbf{d}\|^2 = RSS_0 - RSS; \qquad (3.9)$$

d) má-li $\mathbf{Y}$ v modelu normální rozdělení a platí-li podmodel, je

$$F = \frac{(RSS_0 - RSS)/(r - r_0)}{RSS/(n - r)} \sim \mathsf{F}(r - r_0, n - r). \qquad (3.10)$$

D ů k a z: První dvě tvrzení jsou triviálním důsledkem vět 2.1 a 2.2. Vztah c) je důsledkem ortogonality sloupců matice $\mathbf{P} = (\mathbf{Q}_0, \mathbf{Q}_1, \mathbf{N})$ a toho, že je $\mathbf{u}_0 = \mathbf{u} + \mathbf{d}$. Protože v normálním modelu (za platnosti podmodelu) platí

$$\begin{aligned}\frac{1}{\sigma^2}RSS &= \|\mathbf{R}\mathbf{R}'\mathbf{e}\|^2 = \|\mathbf{R}'\mathbf{e}\|^2 = \|\mathbf{U}\|^2 \sim \chi^2(n - r),\\ \frac{1}{\sigma^2}\|\mathbf{d}\|^2 &= \|\mathbf{P}_1\mathbf{P}_1'\mathbf{e}\|^2 = \|\mathbf{P}_1'\mathbf{e}\|^2 = \|\mathbf{V}_1\|^2 \sim \chi^2(r - r_0),\end{aligned}$$

přičemž náhodné veličiny jsou nezávislé, plyne z rozkladu (3.8) také tvrzení d). □

Můžeme uvažovat posloupnost podmodelů, které jsou do sebe postupně vloženy, které ponechávají vektoru $\mathsf{E}\,\mathbf{Y}$ stále méně stupňů volnosti. Podstatné stačí ukázat u dvojice podmodelů. Mějme tedy n-řádkové matice $\mathbf{X}_{01}, \mathbf{X}_0, \mathbf{X}$ splňující $\mathcal{M}(\mathbf{X}_{01}) \subset \mathcal{M}(\mathbf{X}_0) \subset \mathcal{M}(\mathbf{X})$, pro které platí $0 < r_{01} = \mathsf{h}(\mathbf{X}_{01}) < r_0 = \mathsf{h}(\mathbf{X}_0) < r = \mathsf{h}(\mathbf{X}) < n$. Ortonormální matici $\mathbf{Q}$ pak můžeme vyjádřit jako $(\mathbf{Q}_{01}, \mathbf{Q}_{02}, \mathbf{Q}_1, \mathbf{N})$ s tím, že platí $\mathbf{Q}_0 = (\mathbf{Q}_{01}, \mathbf{Q}_{02})$. Označme ještě jako $\hat{\mathbf{Y}}_{01}$ odhad $\mathsf{E}\,\mathbf{Y}$ metodou nejmenších čtverců v podmodelu $\mathbf{Y} \sim \mathsf{N}\big(\mathbf{X}_{01}\boldsymbol{\beta}_{01}, \sigma^2\mathbf{I}\big)$ a jako RSS_{01} reziduální součet čtverců v tomto podmodelu. Podobně jako nahoře dojdeme k následujícím tvrzením.

Věta 3.2. (O podmodelech) Uvažujme model $\mathbf{Y} \sim \mathsf{N}\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\big)$. Platí-li podmodel $\mathbf{Y} \sim \mathsf{N}\big(\mathbf{X}_0\boldsymbol{\beta}_0, \sigma^2\mathbf{I}\big)$, pak platí

$$F_0 = \frac{(RSS_0 - RSS)/(r - r_0)}{S^2} \sim \mathsf{F}(r - r_0, n - r). \tag{3.11}$$

Platí-li podmodel $\mathbf{Y} \sim \mathsf{N}\big(\mathbf{X}_{01}\boldsymbol{\beta}_{01}, \sigma^2\mathbf{I}\big)$, pak platí

$$F_{01} = \frac{(RSS_{01} - RSS_0)/(r_0 - r_{01})}{S^2} \sim \mathsf{F}(r_0 - r_{01}, n - r). \tag{3.12}$$

D ů k a z: V důkazu se využije především skutečnost, že platí $RSS_{01} - RSS_0 = ||\hat{\mathbf{Y}}_0 - \hat{\mathbf{Y}}_{01}||^2$, přičemž je tento vektor ortonormální vůči vektoru $\mathbf{u}$. $\square$

Poznámka Nepřehlédněte, prosím, že v obou vzorcích uvedených ve větě je ve jmenovateli stejný odhad rozptylu σ^2. Ve vztahu (3.12) jsme ve jmenovateli mohli použít také odhad rozptylu S_0^2. Tím bychom jen přepsali tvrzení (3.11) s jiným označením.

K podmodelu můžeme dojít několika způsoby, zde uvedeme dva. Budeme se zajímat především o možnost výpočtu přímo vektoru $\mathbf{d}$ nebo čtverce jeho délky.

3.2. Vypuštění sloupců

Podmodel může být dán požadavkem vynechat z regresní matice $\mathbf{X}$ některé sloupce. Bez újmy na obecnosti předpokládejme, že matice, které určují model a podmodel, se liší právě posledními sloupci matice $\mathbf{X}$, totiž $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1)$. Aby šlo o podmodel, musí být $0 < \mathsf{h}(\mathbf{X}_0) = r_0 < \mathsf{h}(\mathbf{X}) = r$. Označíme-li $\mathbf{H}_0 = \mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^-\mathbf{X}_0'$ a $\mathbf{M}_0 = \mathbf{I} - \mathbf{H}_0$, bude zřejmě $\hat{\mathbf{Y}}_0 = \mathbf{H}_0\mathbf{Y}$ a $\mathbf{u}_0 = \mathbf{M}_0\mathbf{Y}$. Dále platí

$$\mathcal{M}(\mathbf{X}) = \mathcal{M}((\mathbf{X}_0, \mathbf{X}_1)) = \mathcal{M}((\mathbf{X}_0, \mathbf{M}_0\mathbf{X}_1)), \tag{3.13}$$

neboť oba poslední lineární obaly jsou totožné. Protože poslední matice $\mathbf{X}_0$ a $\mathbf{M}_0\mathbf{X}_1$ mají navzájem ortogonální sloupce, musí platit $\mathcal{M}(\mathbf{M}_0\mathbf{X}_1) = \mathcal{M}(\mathbf{Q}_1)$. Odtud s použitím (A.15) je projekční matice, která počítá vektor $\mathbf{d}$, dána vztahem (viz (2.3))

$$\mathbf{Q}_1\mathbf{Q}_1' = \mathbf{M}_0\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_0\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{M}_0,$$

takže vektor $\mathbf{d}$ dostaneme jako

$$\begin{aligned}\mathbf{d} &= \mathbf{Q}_1\mathbf{Q}_1'\mathbf{Y} = \mathbf{M}_0\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_0\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{M}_0\mathbf{Y} \\ &= \mathbf{M}_0\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_0\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{u}_0. \end{aligned} \tag{3.14}$$

Podobně vyjde

$$\|\mathbf{d}\|^2 = \mathbf{u}_0'\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_0\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{u}_0. \tag{3.15}$$

3.3. Lineární omezení na parametry

Tentokrát dovolíme pouze některé hodnoty vektoru parametrů $\boldsymbol{\beta}$, totiž takové, které vyhovují zvolenému lineárnímu omezení. Například, složky vektoru $\boldsymbol{\beta}$ mohou znamenat dělení celku do několika částí, takže součet složek musí být roven jedničce.

Nechť $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ je konzistentní soustava takových lineárních rovnic, že platí $\mathcal{M}(\mathbf{A}') \subset \mathcal{M}(\mathbf{X}')$ (každý řádek matice $\mathbf{A}$ je nějakou lineární kombinací řádků matice $\mathbf{X}$). V tomto případě je každá složka vektoru $\mathbf{A}\boldsymbol{\beta}$ odhadnutelná. Hledejme v $\mathcal{M}(\mathbf{X})$ bod $\hat{\mathbf{Y}}_0 = \mathbf{X}\mathbf{b}_0$, který je k danému $\mathbf{Y}$ nejbližší, ale navíc splňuje požadavek $\mathbf{A}\mathbf{b}_0 = \mathbf{c}$. Pomůžeme si známou metodou Lagrangeových multiplikátorů. Označme

$$\varphi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + 2\boldsymbol{\lambda}'(\mathbf{A}\boldsymbol{\beta} - \mathbf{c}).$$

Derivováním podle složek sloupcového vektoru $\boldsymbol{\beta}$ dojdeme k soustavě rovnic

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y} - \mathbf{A}'\boldsymbol{\lambda},$$

která je v důsledku předpokladu $\mathcal{M}(\mathbf{A}') \subset \mathcal{M}(\mathbf{X}')$ konzistentní. Odtud máme nějaké řešení soustavy rovnic (záleží na volbě pseudoinverze)

$$\mathbf{b}_0 = (\mathbf{X}'\mathbf{X})^-\mathbf{X}\mathbf{Y} - (\mathbf{X}'\mathbf{X})^-\mathbf{A}'\boldsymbol{\lambda} = \mathbf{b} - (\mathbf{X}'\mathbf{X})^-\mathbf{A}'\boldsymbol{\lambda}.$$

Vezmeme-li v úvahu omezení $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ (nebo derivaci funkce φ podle $\boldsymbol{\lambda}$), po dosazení za $\boldsymbol{\beta}$ dostaneme konzistentní soustavu pro $\boldsymbol{\lambda}$

$$\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}'\boldsymbol{\lambda} = \mathbf{A}\mathbf{b} - \mathbf{c}.$$

Vektor $\mathbf{b}_0$, který splňuje požadovaná lineární omezení a který určuje hledaný nejbližší bod v $\mathcal{M}(\mathbf{X})$, má po dosazení za $\boldsymbol{\lambda}$ tvar

$$\mathbf{b}_0 = \mathbf{b} - (\mathbf{X}'\mathbf{X})^-\mathbf{A}'\left(\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}'\right)^-(\mathbf{A}\mathbf{b} - \mathbf{c}).$$

Samotný nejbližší bod (a odhad vektoru $\mathsf{E}\,\mathbf{Y}$ za platnosti hypotézy $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$) je pak dán jednoznačně vztahem

$$\hat{\mathbf{Y}}_0 = \mathbf{X}\mathbf{b}_0.$$

Odtud je

$$\begin{aligned}\mathbf{d} &= \hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0 = \mathbf{X}(\mathbf{b} - \mathbf{b}_0)\\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}'\left(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}'\right)^{-}(\mathbf{A}\mathbf{b} - \mathbf{c}),\end{aligned}$$

takže pro testování nejzajímavější výsledek je

$$\|\mathbf{d}\|^2 = (\mathbf{A}\mathbf{b} - \mathbf{c})'\left(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}'\right)^{-}(\mathbf{A}\mathbf{b} - \mathbf{c}).$$

Rozdíl reziduálních součtů čtverců v modelu a za hypotézy tedy měří, nakolik klasické řešení normální rovnice (bez omezení) splňuje hypotézu.

Pokud speciálně má matice $\mathbf{X}$ lineárně nezávislé sloupce a matice $\mathbf{A}$ nemá žádné zbytečné řádky (které by byly lineární kombinací ostatních řádků), potom v posledních dvou vztazích můžeme pseudoinverzní matice nahradit klasickými inverzními maticemi:

$$\mathbf{b}_0 = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\left(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\right)^{-1}(\mathbf{A}\mathbf{b} - \mathbf{c}), \tag{3.16}$$

$$\mathbf{d} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\left(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\right)^{-1}(\mathbf{A}\mathbf{b} - \mathbf{c}), \tag{3.17}$$

$$\|\mathbf{d}\|^2 = (\mathbf{A}\mathbf{b} - \mathbf{c})'\left(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\right)^{-1}(\mathbf{A}\mathbf{b} - \mathbf{c}). \tag{3.18}$$

3.4. Předem daná hodnota regresního koeficientu

Jako ukázku lineárního omezení si popišme situaci, kdy požadujeme, aby platilo $\beta_k = \beta_k^0$, kde β_k^0 je zvolená konstanta. Zvolíme-li speciálně $\beta_k^0 = 0$, znamená to, že chceme vynechat z modelu poslední sloupec matice $\mathbf{X}$.

Pro jednoduchost předpokládejme lineární nezávislost sloupců matice $\mathbf{X}$. Příslušné omezení na $\boldsymbol{\beta}$ můžeme zapsat pomocí $\mathbf{A} = (0, \ldots, 0, 1) = \mathbf{j}_k'$ a $\mathbf{c} = \beta_k^0$. Použijeme-li dříve zavedené označení $\mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1}$, máme pak postupně (označení $\mathbf{v}_{\bullet k}$ pro k-tý sloupec matice $\mathbf{V}$ je zavedeno v Appendixu)

$$\begin{aligned}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' &= \mathbf{V}\mathbf{j}_k = \mathbf{v}_{\bullet k},\\ \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' &= \mathbf{j}_k'\mathbf{V}\mathbf{j}_k = v_{kk},\end{aligned}$$

$$\|\mathbf{d}\|^2 = \frac{(b_k - \beta_k^0)^2}{v_{kk}}, \tag{3.19}$$

$$\mathbf{b}_0 = \mathbf{b} - \frac{b_k - \beta_k^0}{v_{kk}}\mathbf{v}_{\bullet k}. \tag{3.20}$$

S uvážením, jaká je varianční matice odhadu $\mathbf{b}$, lze poslední vztah (po rozšíření konstantou σ^2) psát ve tvaru

$$\mathbf{b}_0 = \mathbf{b} - \frac{b_k - \beta_k^0}{\operatorname{var} b_k} \operatorname{cov}(\mathbf{b}, b_k).$$

Poslední vyjádření lze interpretovat tak, že pokud je některá složka odhadu $\mathbf{b}$ nekorelovaná s k-tou složkou tohoto odhadu b_k, pak se odhad této složky vektoru $\boldsymbol{\beta}$ po fixování regresního koeficientu (tedy také po vyloučení k-té nezávisle proměnné, tj. po vyloučení k-tého sloupce matice $\mathbf{X}$) nezmění.

Poznámka K obdobě vztahu (3.19) se dostaneme v kapitole o parametrizaci v nelineární regresi při zavádění pojmu profilový diagram. V lineárním regresním modelu je zvýšení reziduálního součtu čtverců způsobené požadavkem na konkrétní hodnotu β_k^0 parametru β_k úměrné čtverci rozdílu $b_k - \beta_k^0$.

3.5. Koeficient determinace

Důležitý speciální případ podmodelu dostaneme, když využijeme náš předpoklad, že první sloupec matice $\mathbf{X}$ je tvořen jedničkami (v modelu je absolutní člen, stačilo by předpokládat, že platí $\mathbf{1} \in \mathcal{M}(\mathbf{X})$). Potom požadavek $\mathsf{E}\,\mathbf{Y} = \mathbf{1}\beta_0$ určuje podmodel modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

Snadno spočítáme, že v tomto podmodelu je $b_0 = \bar{Y}$ a $\hat{\mathbf{Y}}_0 = \bar{Y}\mathbf{1}$. Odtud je $\mathbf{d} = \hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0 = \hat{\mathbf{Y}} - \bar{Y}\mathbf{1}$, takže podle (3.9) je

$$RSS_0 = RSS + ||\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}||^2.$$

Spočítejme *výběrový* korelační koeficient mezi $\mathbf{Y}$ a $\hat{\mathbf{Y}}$. Z předpokladu $\mathbf{1} \in \mathcal{M}(\mathbf{X})$ plyne, že platí

$$0 = \mathbf{1}'\mathbf{u} = \mathbf{1}'(\mathbf{Y} - \hat{\mathbf{Y}}),$$

takže výběrové průměry složek vektorů $\mathbf{Y}, \hat{\mathbf{Y}}$ jsou shodné. Proto lze psát

$$\begin{aligned} r^2_{Y,\hat{Y}} &= \frac{(\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}))^2}{\sum(Y_i - \bar{Y})^2 \sum(\hat{Y}_i - \bar{Y})^2} = \frac{((\mathbf{Y} - \bar{Y}\mathbf{1})'(\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}))^2}{||\mathbf{Y} - \bar{Y}\mathbf{1}||^2 ||\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}||^2} \\ &= \frac{((\mathbf{Y} - \hat{\mathbf{Y}}_0)'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0))^2}{||\mathbf{Y} - \hat{\mathbf{Y}}_0||^2 ||\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0||^2} = \frac{((\mathbf{d} + \mathbf{u})'\mathbf{d})^2}{||\mathbf{u}_0||^2 ||\mathbf{d}||^2} \\ &= \frac{||\mathbf{d}||^2}{||\mathbf{u}_0||^2} = \frac{RSS_0 - RSS}{RSS_0} \\ &= 1 - \frac{RSS}{\sum(Y_i - \bar{Y})^2} \;=\; R^2. \end{aligned} \tag{3.21}$$

Druhá identita v posledním řádku je nejčastější definicí *koeficientu determinace* R^2, který je v případě lineárního modelu shodný se čtvercem výběrového koeficientu mnohonásobné korelace spočítaného z vektoru $\mathbf{Y}$ a odpovídajících netriviálních (nekonstantních) sloupců matice $\mathbf{X}$.

Koeficient determinace ukazuje, jak velký díl výchozí variability hodnot závisle proměnné charakterizované výrazem

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2 = \|\mathbf{u}_0\|^2$$

se nám podařilo vysvětlit, když nevysvětlená variabilita je dána reziduálním součtem čtverců RSS, v této souvislosti označovaným také jako SSE. Variabilita vysvětlená modelem (uvažovanou závislostí) je tedy dána výrazem

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \|\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}\|^2 = \|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|^2 = \|\mathbf{d}\|^2.$$

V normálním modelu můžeme testovou statistiku F pro testování podmodelu určeného požadavkem $\mathsf{E}\,\mathbf{Y} = \mathbf{1}\beta_0$ vyjádřit pomocí koeficientu determinace R^2:

$$\begin{aligned} F &= \frac{SSR}{RSS}\frac{n-r}{r-1} = \frac{1 - RSS/RSS_0}{RSS/RSS_0}\frac{n-r}{r-1} \\ &= \frac{R^2}{1-R^2}\frac{n-r}{r-1}. \end{aligned}$$

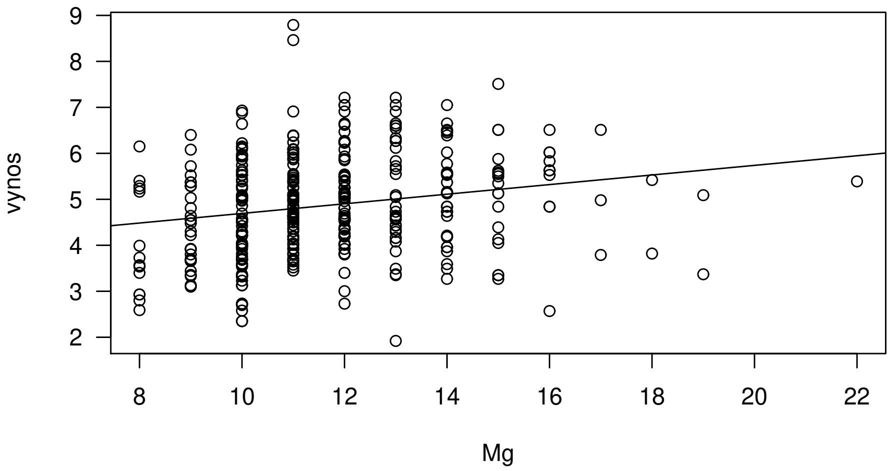
Na tomto místě je snad užitečné připomenout, že při testování nulové hypotézy o nezávislosti složek dvourozměrného normálního rozdělení se používá statistika

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

přičemž za platnosti nulové hypotézy platí $T^2 \sim \mathsf{F}(1, n-2)$.

Příklad 3.1 (DRIS) Na základě dat z velkého polního pokusu, který zkoumal předpovědi výnosu podle známého obsahu hořčíku v sušině rostliny během vegetace, vyšla ve zvolených jednotkách předpověď ve tvaru $\widehat{\mathsf{výnos}} = 3{,}648 + 0{,}104\ \mathsf{Mg}$, přičemž směrnice přímky byla odhadnuta se střední chybou 0,026. Odtud je hodnota t statistiky rovna $t = 3{,}955$ s dosaženou hladinou $p < 0{,}0001$. O tom, že střední hodnota výnosů závisí na obsahu hořčíku tedy není pochyb. Reziduální součet čtverců je roven $SSE = 422{,}43$, kdežto v podmodelu požadujícím, aby výnos byl konstantní, je reziduální součet čtverců roven $SST = 440{,}48$, tedy jen nepatrně víc. Odtud vyjde $R^2 = 0{,}041$. Tedy pouze 4,1 % variability výnosů lze vysvětlit závislostí na obsahu hořčíku. Tak slabou závislost asi prakticky nedokážeme využít, přestože je směrnice regresní přímky průkazně nenulová.

Následuje výpočet v prostředí R:



Obrázek 3.1: Závislost výnosů na koncentraci hořčíku v sušině

```
> summary(vynos.Mg<-lm(vynos~Mg,data=Dris))

Call:
lm(formula = vynos ~ Mg, data = Dris)

Residuals:
     Min       1Q   Median       3Q      Max
-3.08671 -0.75509 -0.08571  0.70128  3.99429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.64823    0.31260  11.670  < 2e-16
Mg           0.10450    0.02642   3.955  9.2e-05

Residual standard error: 1.074 on 366 degrees of freedom
Multiple R-Squared: 0.04098,    Adjusted R-squared: 0.03836
F-statistic: 15.64 on 1 and 366 degrees of freedom,     p-value: 9.206e-005

> anova(vynos.Mg)
Analysis of Variance Table

Response: vynos
           Df Sum Sq Mean Sq F value    Pr(>F)
Mg          1  18.05   18.05  15.639 9.206e-05
Residuals 366 422.43    1.15
```

```
> plot(Dris$Mg,Dris$vynos)
> abline(vynos.Mg)
```

Jistě nebude obtížné vysvětlit, proč jsou dosažené hladiny v řádku `Mg` v `summary()` a v `anova()` stejné, když testová statistika v `anova()` je druhou mocninou statistiky v `summary()`. ○

Příklad 3.2 (hmotnost hochů) Snadno se přesvědčíme, že naše zjištění o rychlosti, s jakou roste porodní hmotnost s porodní délkou. Lineární regresní model odhaduje, že s každým centimetrem porodní délky roste porodní hmotnost přibližně o 192 gramů. Tento regresní koefiient je průkazně nenulový. Nestejné porodní hmotnosti hochů vyvětlíme jejich porodními délkami téměř z 57 %:

```
> summary(lm(por.hmot~por.del,data=Hosi0))

Call:
lm(formula = por.hmot ~ por.del, data = Hosi0)

Residuals:
     Min       1Q   Median       3Q      Max
-1520.33  -188.20   -10.33   189.67  1531.80

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6230.146    121.095  -51.45   <2e-16
por.del       192.124      2.407   79.81   <2e-16

Residual standard error: 291.7 on 4836 degrees of freedom
Multiple R-Squared: 0.5685,     Adjusted R-squared: 0.5684
F-statistic:  6370 on 1 and 4836 DF,  p-value: < 2.2e-16
```

○

3.6. Vážený lineární model

Pokud pracujeme s modelem $\mathbf{Y} \sim \left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1}\right)$, pak koeficient determinace převezmeme z modelu pro transformovaný vektor $\mathbf{Y}^*$ s varianční maticí $\sigma^2\mathbf{I}$. Odhadem parametru β_0 z podmodelu $\mathsf{E}\,\mathbf{Y}^* = \mathbf{1}\beta_0$ je pak

$$\bar{Y}_W = \bar{Y}^* = (\mathbf{1}'\mathbf{W}\mathbf{1})^{-1}\mathbf{1}'\mathbf{W}\mathbf{Y},$$

takže v podmodelu je reziduální součet čtverců roven

$$RSS_{W0} = RSS_0^* = (\mathbf{Y} - \bar{Y}_W)'\mathbf{W}(\mathbf{Y} - \bar{Y}_W).$$

Koeficient determinace bude tedy

$$R_W^2 = 1 - \frac{RSS_W}{RSS_{W0}}.$$

V případě vážené regrese s diagonální maticí $\mathbf{W}$ dostaneme

$$R_W^2 = 1 - \frac{\sum_{i=1}^n w_{ii}(Y_i - \hat{Y}_{Wi})^2}{\sum_{i=1}^n w_{ii}(Y_i - \bar{Y}_W)^2},$$

kde jsme použili vážený průměr vysvětlované proměnné Y

$$\bar{Y}_W = \frac{\sum_{i=1}^n w_{ii} Y_i}{\sum_{i=1}^n w_{ii}}.$$

Testy hypotéz o nulovosti jednotlivých složek vektoru $\boldsymbol{\beta}$ ve váženém lineárním modelu jsou totožné s testy stejných hypotéz v modelu pro transformovaný vektor $\mathbf{Y}^*$.

4. Regresní funkce lineární v parametrech

Nejčastěji se v regresi vyšetřuje regresní přímka. V této kapitole se budeme zabývat zejména přímkou a porovnáváním přímek. Všimneme si také závislostí, které lze popsat pomocí funkce, která je v neznámých parametrech lineární, avšak na jediné nezávisle proměnné t může záviset i nelineárně. Příkladem může být polynom v t.

4.1. Jedna přímka

Tuto jednoduchou situaci pouze shrneme. Předpokládá se n nezávislých náhodných veličin $Y_i \sim \mathsf{N}\big(\beta_0 + \beta_1 x_i, \sigma^2\big)$, kde konstanty $x_1, \ldots, x_n$ nejsou všechny stejné, β_0, β_1 a $\sigma > 0$ jsou neznámé parametry.

Odhady regresních koeficientů jsou dány vztahy

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad b_0 = \bar{Y} - b_1 \bar{x}. \tag{4.1}$$

Reziduální součet čtverců lze vyjádřit jako

$$RSS = \sum_{i=1}^n (Y_i - \bar{Y})^2 - b_1 \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}),$$

nestranným odhadem rozptylu je zřejmě

$$S^2 = \frac{RSS}{n-2}.$$

Všimněme si dvou modifikací naší úlohy. Odhad b_1 z (4.1) můžeme přepsat na tvar

$$b_1 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\sum_{t=1}^n (x_t - \bar{x})^2} \frac{Y_i - \bar{Y}}{x_i - \bar{x}},$$

když v případě $x_i = \bar{x}$ nebereme nulový sčítanec v úvahu. Směrnice b_1 je tedy váženým průměrem směrnic $(Y_i - \bar{Y})/(x_i - \bar{x})$ přímek spojujících vždy bod $[x_i, Y_i]$ s těžištěm $[\bar{x}, \bar{Y}]$.

Zajímavou modifikaci dostaneme, když přímku zapíšeme ve tvaru $y = \beta_0^* + \beta_1^*(x - \bar{x})$, kde je samozřejmě $\beta_0^* = \beta_0 + \beta_1\bar{x}$ a $\beta_1^* = \beta_1$. Regresní matice $\mathbf{X}^*$ má v tomto případě tvar

$$\mathbf{X}^* = (\mathbf{1} \quad \mathbf{x} - \bar{x}\mathbf{1}),$$

takže vyjde

$$\mathbf{X}^{*'}\mathbf{X}^* = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2 \end{pmatrix}, \quad \mathbf{X}^{*'}\mathbf{Y} = \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}). \end{pmatrix}.$$

Odhady parametrů dostaneme snadno. Jako odhad směrnice dostaneme ihned vzorec identický s odhadem (4.1), pro absolutní člen vyjde $b_0^* = \bar{Y}$, takže ze vztahu mezi β_0^* a β_0, β_1 ihned vyjde také odhad b_0.

Lineární obaly sloupců matic $\mathbf{X}, \mathbf{X}^*$ jsou totožné, takže totožné odhady $\hat{Y}_i$ včetně jejich rozdělení. Rozptyl statistiky $\hat{Y}_i$ snáze spočítáme z hvězdičkového modelu. Když využijeme skutečnost, že matice $\mathbf{X}^{*'}\mathbf{X}^*$ je diagonální a tudíž odhady b_0^*, b_1 jsou nekorelované, dostaneme

$$\begin{aligned} \operatorname{var} \hat{Y}_i &= \operatorname{var}(b_0^* + b_1(x_i - \bar{x})) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{t=1}^n (x_t - \bar{x})^2} \right). \end{aligned} \tag{4.2}$$

Podobně vyjde

$$\begin{aligned} \operatorname{cov}(\hat{Y}_i, \hat{Y}_j) &= \operatorname{cov}(b_0^* + b_1(x_i - \bar{x}), b_0^* + b_1(x_j - \bar{x})) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2} \right), \end{aligned}$$

takže projekční matice $\mathbf{H}$ má prvky (s ohledem na $\operatorname{var} \hat{\mathbf{Y}} = \sigma^2\mathbf{H}$)

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2}. \tag{4.3}$$

Matice $\mathbf{M}$ má tedy prvky (δ_{ij} je Kroneckerovo delta)

$$m_{ij} = \delta_{ij} - \frac{1}{n} - \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2}.$$

Výsledek (matice $\mathbf{H}, \mathbf{M}$) se týká středních hodnot Y_i, nikoliv třeba regresních koeficientů. Nezávisí na zvoleném parametrickém vyjádření, platí pro obojí parametrické vyjádření.

4.2. Obecnější funkce

Uvažujme nyní závislost $y = \boldsymbol{\beta}'\mathbf{x}(t)$, kde $\mathbf{x}(t)$ je vektor známých spojitých funkcí. Předpokládejme dále, že parametr $\boldsymbol{\beta}$ odhadneme z n nezávislých pozorování $Y_i \sim \mathsf{N}\big(\boldsymbol{\beta}'\mathbf{x}(t_i), \sigma^2\big)$ s takovými hodnotami $t_1, \ldots, t_n$, že matice $\mathbf{X}$ s i-tým řádkem $\mathbf{x}(t_i)$ má lineárně nezávislé sloupce. Vektor $\boldsymbol{\beta}$ je pak odhadnutelný, odhad $\mathbf{b}$ má varianční matici $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Uvažujme nejprve jedinu pevnou hodnotu t. Větu 2.6 použijeme pro hledání *intervalu spolehlivosti* (*konfidenčního intervalu*) pro $\mathsf{E}\,Y(t) = \boldsymbol{\beta}'\mathbf{x}(t)$. Bodovým odhadem bude zřejmě statistika $\mathbf{b}'\mathbf{x}(t)$ s rozptylem $\sigma^2\mathbf{x}(t)'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}(t) = \sigma^2 d^2(t)$, když jsme zavedli nezápornou funkci $d(t)$. Můžeme vyjít z testování hypotézy, že $\mathsf{E}\,Y(t) = y_0$. Protože jde o odhad lineární parametrické funkce, má zřejmě podle tvrzení g) zmíněné věty statistika

$$\frac{\mathbf{b}'\mathbf{x}(t) - y_0}{S\ d(t)}$$

rozdělení $\mathsf{t}(n-k-1)$. Interval spolehlivosti pro $\mathsf{E}\,Y(t)$ dostaneme jako množinu všech y_0, pro která nulovou hypotézu nezamítneme, tedy

$$\big(\mathbf{b}'\mathbf{x}(t) - t_{n-k-1}(\alpha)S\ d(t); \mathbf{b}'\mathbf{x}(t) + t_{n-k-1}(\alpha)S\ d(t)\big)\,. \tag{4.4}$$

Hledejme nyní *predikční* interval s vlastností, že s předem danou pravděpodobností obsahuje nezávislé budoucí pozorování $Y(t)$ (opět pro pevně zvolené jediné t). Zajímáme se o $\boldsymbol{\beta}'\mathbf{x}(t) + e$, kde $e \sim \mathsf{N}\big(0, \sigma^2\big)$. Bodovým odhadem bude opět $\mathbf{b}'\mathbf{x}(t)$, ale rozdíl $Y(t) - \mathbf{b}'\mathbf{x}(t)$ bude mít tentokrát rozptyl $\sigma^2(1 + d^2(t))$. Příslušný interval tedy má tvar

$$\Big(\mathbf{b}'\mathbf{x}(t) - t_{n-k-1}(\alpha)S\sqrt{1+d^2(t)}; \mathbf{b}'\mathbf{x}(t) + t_{n-k-1}(\alpha)S\sqrt{1+d^2(t)}\Big)\,. \tag{4.5}$$

V obou případech se často graficky znázorňují intervaly (4.4), (4.5) pro všechna t z nějakého intervalu spolu s funkcí $\mathbf{b}'\mathbf{x}(t)$. Dostaneme tak pás spolehlivosti kolem regresní funkce.

Speciálně pro regresní přímku dostaneme

$$d^2(x) = \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

takže na místě (4.4) interval s krajními body (viz pás spolehlivosti kolem regresní přímky (Anděl, 1978, odst. VI. 3) nebo (Anděl, 1998, odst. 12. 2. B))

$$b_0 + b_1 x \pm S \cdot t_{n-2}(\alpha)\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{t=1}^n (x_t - \bar{x})^2}}. \tag{4.6}$$

Podobně jsou krajní body predikčního intervalu jsou v případě regresní přímky dány

$$b_0 + b_1 x \pm S \cdot t_{n-2}(\alpha)\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{t=1}^n (x_t - \bar{x})^2}}. \tag{4.7}$$

4.3. Pás spolehlivosti pro regresní funkci

Uvažujme opět stejnou situaci jako v předchozím oddíle. Místo jediného pevného t nás bude tentokrát zajímat průběh regresní funkce $\boldsymbol{\beta}'\mathbf{x}(t)$ pro $t \in T$, zpravidla pro $t \in \mathbb{R}$. Uvedeme konstrukci, která vede k *pásu spolehlivosti pro regresní funkci.*

Nechť $\mathcal{K}$ je konfidenční množina pro $\boldsymbol{\beta}$. Zvolme funkce

$$L(t) = \sup_{\boldsymbol{\beta} \in \mathcal{K}} \boldsymbol{\beta}'\mathbf{x}(t), \qquad U(t) = \sup_{\boldsymbol{\beta} \in \mathcal{K}} \boldsymbol{\beta}'\mathbf{x}(t). \tag{4.8}$$

Pás spolehlivosti pro regresní funkci sestrojíme jako

$$\mathcal{L} = \{(t, y)' : L(t) \leq y \leq U(t), t \in T\}.$$

Z toho, jak jsme množinu $\mathcal{L}$ zavedli, plyne, že pro každé $\boldsymbol{\beta} \in \mathcal{K}$ platí $L(t) \leq \boldsymbol{\beta}'\mathbf{x}(t) \leq U(t)$ pro všechna $t \in T$. Je-li spolehlivost $\mathcal{K}$ rovna $1 - \alpha$, pak pás $\mathcal{L}$ pokryje funkci $\boldsymbol{\beta}'\mathbf{x}(t)$ současně pro všechna $t \in T$ s pravděpodobností aspoň $1 - \alpha$.

Pokud je u pásu $\mathcal{L}$ zaručena rovnost, hovoří se o *přesném* pásu spolehlivosti. Přesnost pásu může být zaručena, pokud je s pravděpodobností 1 výchozí konfidenční množina konvexní a ohraničená (Zvára (1979)).

Hledejme pás spolehlivosti pro lineární regresní funkci $\boldsymbol{\beta}'\mathbf{x}(t)$. Jako výchozí konfidenční množinu použijeme elipsoid $\mathcal{K}_2$ z věty 2.6

$$\mathcal{K} = \left\{\boldsymbol{\beta} \in \mathbb{R}_{k+1} : (\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) \leq (k+1)S^2 F_{k+1,n-k-1}(\alpha)\right\}.$$

Vzhledem k tvaru konfidenční množiny nastanou extrémy definující funkce $L(t)$ a $U(t)$ v hraničních bodech $\mathcal{K}_2$, takže k jejich nalezení lze použít metodu Lagrangeových multiplikátorů. Hledáme extrém funkce

$$\varphi(\boldsymbol{\beta}, \lambda) = \boldsymbol{\beta}'\mathbf{x}(t) - \frac{\lambda}{2}\left((\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) - c\right),$$

kde jsme pro stručnost označili $c = (k+1)S^2 F_{k+1,n-k-1}(\alpha)$.

Derivace podle $\boldsymbol{\beta}$ jsou nulové pro $\mathbf{x}(t) = \lambda\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b})$. Odtud $\tilde{\boldsymbol{\beta}}$, v němž nastává extrém, splňuje

$$\tilde{\boldsymbol{\beta}} = \mathbf{b} + \frac{1}{\lambda}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}(t).$$

Po dosazení do podmínky dostaneme $1/\lambda = \pm\sqrt{c}/d(t)$, což vede k extrému v $\tilde{\boldsymbol{\beta}} = \mathbf{b} \pm \sqrt{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}(t)/d(t)$. Extrémní funkční hodnota je tedy

$$\tilde{\boldsymbol{\beta}}'\mathbf{x}(t) = \mathbf{b}'\mathbf{x}(t) \pm \frac{\sqrt{c}}{d(t)}\mathbf{x}(t)'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}(t) = \mathbf{b}'\mathbf{x}(t) \pm \sqrt{c}d(t).$$

Vzhledem k nezápornosti funkce $d(t)$ máme výsledné meze pásu spolehlivosti

$$L(t) = \mathbf{b}'\mathbf{x}(t) - S\ d(t)\sqrt{(k+1)\ F_{k+1,n-k-1}(\alpha)}, \tag{4.9}$$

$$U(t) = \mathbf{b}'\mathbf{x}(t) + S\ d(t)\sqrt{(k+1)\ F_{k+1,n-k-1}(\alpha)}. \tag{4.10}$$

Ve speciálním případě regresní přímky vyjde pás spolehlivosti (viz např. Anděl (1978, str. 149))

$$b_0 + b_1 x \pm S\sqrt{2F_{2,n-2}(\alpha)\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{t=1}^{n}(x_t-\bar{x})^2}\right)}. \tag{4.11}$$

4.4. Několik přímek

Vyšetřujme nyní k nezávisle odhadovaných regresních přímek. Máme k disposici nezávislé náhodné veličiny $Y_{ij} \sim \mathsf{N}\big(\beta_{0i} + \beta_{1i}x_{ij}, \sigma^2\big)$, přičemž u i-té přímky máme n_i pozorování. Máme celkem $n = \sum_{i=1}^{I} n_i$ pozorování. Parametry $\beta_{0i}, \beta_{1i}, \sigma > 0$ odhadujeme.

Všechna pozorování lze zapsat maticově

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{I1} \\ \vdots \\ Y_{In_I} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & 0 & 0 \\ 1 & x_{12} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_1} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & x_{I1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & x_{In_I} \end{pmatrix} + \begin{pmatrix} \beta_{01} \\ \beta_{11} \\ \vdots \\ \beta_{0I} \\ \beta_{1I} \end{pmatrix} + \mathbf{e}, \tag{4.12}$$

kde náhodný vektor $\mathbf{e}$ má rozdělení $\mathsf{N}\big(\mathbf{0}, \sigma^2\mathbf{I}\big)$.

Z blokově diagonální struktury regresní matice je zřejmé, že odhady přímek jsou nezávislé, že reziduální součet čtverců v modelu je součtem reziduálních součtů čtverců u jednotlivých přímek. Snadno lze z blokové struktury matice $\mathbf{X}'\mathbf{X}$ odvodit, že její determinant je roven součinu determinantů jednotlivých diagonálních bloků

$$\det(\mathbf{X}'\mathbf{X}) = \prod_{i=1}^{I} n_i \sum_{t=1}^{n_i}(x_{it} - \bar{x}_i)^2.$$

Odtud plyne, že matice modelu bude mít lineárně nezávislé sloupce, právě když pro každou přímku máme pozorování aspoň ve dvou různých bodech x_{ij}.

Testujme podmodel, který vyjadřuje předpoklad, že směrnice všech přímek jsou shodné, tedy přímky jsou rovnoběžné. Podmodel znamená, že platí

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{I1} \\ \vdots \\ Y_{In_I} \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 0 & x_{11} \\ 1 & \cdots & 0 & x_{12} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \cdots & 0 & x_{1n_1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & x_{I1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & x_{In_I} \end{pmatrix} + \begin{pmatrix} \beta_{01} \\ \vdots \\ \beta_{0I} \\ \beta_1 \end{pmatrix} + \mathbf{e}, \tag{4.13}$$

Že jde o podmodel je zřejmé z toho, že sloupce nové regresní matice lze snadno získat z původní: sloupce s jedničkami a nulami ponecháme, ostatní sloupce sečteme. Pokud výchozí matice měla úplnou hodnost, nová matice má stejnou vlastnost. Podrobněji je hodnost této regresní matice vyšetřena v příkladu 2.2.

Příklad 4.1 (listy) Součástí většího pokusu bylo také opakované měření délky prvních tří listů rostlinky pšenice. Na obrázku 4.1 jsou znázorněna data a příslušné regresní přímky. Odhady ve výchozím modelu jsou (`list` je faktor, nechali jsme standardní nastavení kontrastů v R na `contr.treatment` – viz str. 52)

```
> summary(a<-lm(delka~den*List,data=Listy))

Call:
lm(formula = delka ~ den * List, data = Listy)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91073 -0.17127 -0.05549  0.22735  0.92575

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.57660    0.79354  -3.247    0.007 **
den           1.20319    0.07261  16.570 1.24e-09 ***
List2       -36.20834    1.92114 -18.847 2.79e-10 ***
List3       -48.81182    2.30132 -21.210 7.02e-11 ***
den:List2     1.55845    0.12236  12.737 2.48e-08 ***
den:List3     1.45131    0.11210  12.947 2.07e-08 ***

Residual standard error: 0.5322 on 12 degrees of freedom
Multiple R-Squared: 0.9951,     Adjusted R-squared: 0.9931
F-statistic: 488.2 on 5 and 12 degrees of freedom,   p-value: 1.996e-013
```

Jednotlivé přímky mají rovnice

$$\begin{aligned} y &= -2{,}577 + 1{,}203x && \text{1. přímka} \\ y &= (-2{,}577 - 36{,}208) + (1{,}203 + 1{,}558)x && \text{2. přímka} \\ y &= (-2{,}577 - 48{,}812) + (1{,}203 + 1{,}451)x && \text{3. přímka} \end{aligned}$$

Zkusme vyšetřit podmodel, v němž jsou všechny tři přímky rovnoběžné:

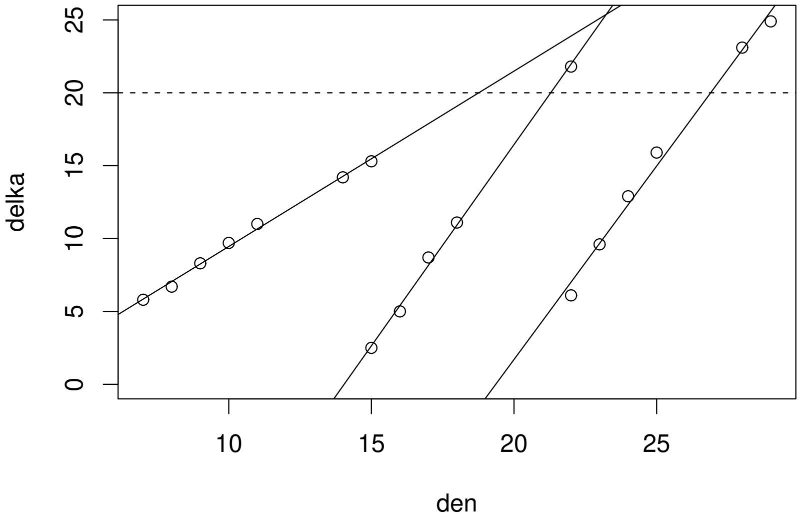
```
> summary(a.rovno<-lm(delka~den+List,data=Listy))

Call:
lm(formula = delka ~ den + List, data = Listy)

Residuals:
   Min     1Q Median     3Q    Max
-3.877 -1.516  0.284  1.588  3.004

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.4217     2.3175  -4.928 0.000222 ***
den           2.0399     0.2039  10.003 9.31e-08 ***
List2       -14.6604     1.9469  -7.530 2.75e-06 ***
List3       -24.4989     3.2289  -7.587 2.52e-06 ***

Residual standard error: 2.25 on 14 degrees of freedom
Multiple R-Squared: 0.898,      Adjusted R-squared: 0.8761
F-statistic: 41.08 on 3 and 14 DF,  p-value: 3.449e-07
```

O podmodelu rozhodneme pomocí F testu

```
> anova(a.rovno,a)
Analysis of Variance Table

Model 1: delka ~ den + List
Model 2: delka ~ den + List + den:List
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     14 70.883
2     12  3.399  2    67.484 119.14 1.215e-08 ***
```

Po shlédnutí obrázku 4.1 nepřekvapí, že jsme hypotézu o rovnoběžnosti zamítli. Jinak by to dopadlo s testem nulové hypotézy, podle které se neliší rychlosti růstu druhého listu a třetího listu. Tato hypotéza má svoje biologické vysvětlení, navíc souvisí s původní otázkou experimentátora, totiž, zda jsou konstantní časové odstupy mezi okamžiky, kdy jednotlivé listy dosahují předem zvolené pevné délky 20 mm.

```
> summary(a.rovno23<-lm(delka~den+List+(List!=1):den,data=Listy))
```



Obrázek 4.1: Závislost délky listu na době pro jednotlivé listy

```
Call:
lm(formula = delka ~ den + List + (List != 1):den, data = Listy)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86854 -0.26686  0.03317  0.23346  0.93341

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -2.57660    0.78357  -3.288  0.00588 **
den                 1.20319    0.07170  16.781 3.43e-10 ***
List2             -35.13203    1.38800 -25.311 1.91e-12 ***
List3             -49.96907    1.79745 -27.800 5.76e-13 ***
den:List != 1TRUE   1.49730    0.09592  15.610 8.43e-10 ***

Residual standard error: 0.5255 on 13 degrees of freedom
Multiple R-Squared: 0.9948,     Adjusted R-squared: 0.9932
F-statistic: 625.8 on 4 and 13 DF,  p-value: 1.021e-14

> anova(a,a.rovno23)
```

```
Analysis of Variance Table

Model 1: delka ~ den + List + den:List
Model 2: delka ~ den + List + den:List != 1
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     12  3.3986
2     13  3.5899 -1   -0.1913 0.6755 0.4272
```

Z výsledku je patrné, že se problémům způsobeným nerovnoběžností přímek nevyhneme. Druhou a třetí přímku lze považovat za rovnoběžné, první má však průkazně povlovnější stoupání. ○

4.5. Inverzní predikce

V praxi často narazíme na úlohu odhadnout ze známé hodnoty závisle proměnné odpovídající hodnotu nezávisle proměnné. Podrobně se této a podobným úlohám věnuje Jílkova (1988) kniha. Pokud hledáme postup, jak k nekonečně mnoha budoucím pozorováním závisle proměnné najít odpovídající hodnoty nezávisle proměnné, jedná se o úlohu *kalibrace*.

Zde uvedeme nejprve jednoduché přibližné řešení úlohy pro jedinou realizaci závisle proměnné (Netter, Wasserman, Kutner (1985), oddíl 5.8), které je použitelné v případě, kdy data jsou velmi dobře popsána regresní přímkou, což se projeví ve velké hodnotě koeficientu determinace.

Předpokládejme, že jsme již odhadli parametry regresní přímky. Získali jsme nové stochasticky nezávislé pozorování Y závisle proměnné, které se řídí stejným modelem, tj. $Y \sim \mathsf{N}(\beta_0 + \beta_1 x, \sigma^2)$. Problém je v tom, že neznáme hodnotu x, takže cílem je najít jednoduchý bodový a intervalový odhad pro x.

Vyjdeme z „naivního odhadu“ $\hat{x}$ ze vztahu $Y = \bar{Y} + b_1(\hat{x} - \bar{x})$. Po úpravě dostaneme

$$\hat{x} = \bar{x} + \frac{Y - \bar{Y}}{b_1}. \tag{4.14}$$

Rozptyl odhadu určíme pomocí tzv. δ-metody (viz např. Rao (1978, str. 431)) z lineární aproximace odhadové statistiky, která je funkcí tří nezávislých náhodných veličin: $Y, \bar{Y}, b_1$ (připomeňte si druhou parametrizaci přímky). Protože je

$$\frac{\partial \hat{x}}{\partial Y} = \frac{1}{b_1}, \quad \frac{\partial \hat{x}}{\partial \bar{Y}} = -\frac{1}{b_1}, \quad \frac{\partial \hat{x}}{\partial b_1} = -\frac{Y - \bar{Y}}{b_1^2},$$

aproximaci rozptylu statistiky $\hat{x}$

$$\begin{aligned}\mathsf{var}\,\hat{x} &\doteq \begin{pmatrix} \frac{1}{b_1} \\ -\frac{1}{b_1} \\ -\frac{Y-\bar{Y}}{b_1^2} \end{pmatrix}' \sigma^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{n} & 0 \\ 0 & 0 & \frac{1}{T_{xx}} \end{pmatrix} \begin{pmatrix} \frac{1}{b_1} \\ -\frac{1}{b_1} \\ -\frac{Y-\bar{Y}}{b_1^2} \end{pmatrix} \\ &= \frac{\sigma^2}{b_1^2}\left(1 + \frac{1}{n} + \frac{(Y-\bar{Y})^2}{b_1^2}\frac{1}{T_{xx}}\right),\end{aligned}$$

když jsme zavedli označení $T_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Použijeme-li vztah $Y - \bar{Y} = b_1(\hat{x} - \bar{x})$ a neznámý rozptyl σ^2 nahradíme jeho odhadem S^2, dostaneme nakonec přibližný odhad rozptylu $\hat{x}$

$$\widehat{\mathsf{var}}\,\hat{x} \doteq \frac{S^2}{b_1^2}\left(1 + \frac{1}{n} + \frac{(\hat{x}-\bar{x})^2}{T_{xx}}\right). \tag{4.15}$$

Přibližný interval spolehlivosti pro hledanou hodnotu x má tedy krajní body

$$\hat{x} \pm \frac{S}{|b_1|} t_{n-2}(\alpha)\sqrt{1 + \frac{1}{n} + \frac{(\hat{x}-\bar{x})^2}{T_{xx}}}. \tag{4.16}$$

Všimněte si nápadné podoby s predikčním intervalem (4.7). Interval (4.16) je totiž vzorem predikčního intervalu (4.7), když ke zobrazení použijeme odhad regresní funkce.

Věnujme se ještě malé modifikaci úlohy. Kdybychom hledali hodnotu nezávisle proměnné k *dané střední hodnotě* μ závisle proměnné, dostali bychom přibližný interval s krajními body (srovnej s (4.6))

$$\bar{x} + \frac{\mu - \bar{Y}}{b_1} \pm \frac{S}{|b_1|} t_{n-2}(\alpha)\sqrt{\frac{1}{n} + \frac{(\hat{x}-\bar{x})^2}{T_{xx}}}. \tag{4.17}$$

Příklad 4.2 (listy) Vraťme se k pokusu s růstem prvních tří listů a pro jednotlivé přímky odhadněme okamžik, kdy list dosáhl délky 20 mm. Bodový odhad je jednoduchý:

```
> d.0 <- 20
> summary(a.1<-lm(delka~den,subset=List==1))

Call:
lm(formula = delka ~ den, subset = List == 1)

Residuals:
       1        2        3        4        5        6        7 
-0.04574 -0.34894  0.04787  0.24468  0.34149 -0.06809 -0.17128 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.5766     0.3869   -6.66  0.00115 **
den           1.2032     0.0354   33.99 4.15e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2595 on 5 degrees of freedom
Multiple R-Squared: 0.9957,     Adjusted R-squared: 0.9948
F-statistic:  1155 on 1 and 5 DF,  p-value: 4.147e-07

> print(d.1<-(d.0-coef(a.1)[1])/coef(a.1)[2])
(Intercept)
   18.76393
```

Nyní si připravíme a uložíme mezivýsledky, které budeme dál potřebovat.

```
> print(x.bar<-mean(den[List==1]))
[1] 10.57143
> print(y.bar<-mean(delka[List==1]))
[1] 10.14286
> print(Txx<-sum((den[List==1]-x.bar)^2))
[1] 53.71429
> print(b1<-coef(a.1)[2])
     den
1.203191
```

Střední chybu odhadu pro den, kdy bylo dosaženo zvolené délky dostaneme ze střední chyby pro odhad $\mathsf{E}\,Y$ pro $x = \hat{x}$:

```
> SE.d.1<-predict(a.1,newdata=data.frame(den=d.1),se.fit=T)$se.fit/coef(a.1)[2]
> print(SE.d.1)
      den
0.2544561
```

Hledaný interval pak už najdeme snadno.

```
> print(t.1<-qt(.975,a.1$df.resid))
[1] 2.570582
> int.1<-c(d.1,SE.d.1)%*%matrix(c(1,0,1,-t.1,1,t.1),2,3)
> int.1
         [,1]     [,2]     [,3]
[1,] 18.76393 18.10983 19.41803
```

○

Naznačme ještě jednu metodu, tentokrát přesnou, nikoliv založenou na aproximaci. *Fiellerova metoda* spočítá v tom, že vyjdeme z testování nulové hypotézy, podle které je hledané x rovno danému x_0. Interval spolehlivosti bude pak tvořen množinou takových x_0, pro která nulovou hypotézu na zvolené hladině nezamítneme. Modifikací predikčního intervalu (4.7) jde o množinu danou nerovností

$$|Y - \bar{Y} - b_1(x_0 - \bar{x})| < S \cdot t_{n-2}(\alpha)\sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{T_{xx}}\right)}. \tag{4.18}$$

Této nerovnosti vyhoví všechna x_0 splňující

$$A(x_0 - \bar{x})^2 + B(x_0 - \bar{x}) + C < 0, \tag{4.19}$$

kde koeficient u druhé mocniny je roven

$$A = b_1^2 - \frac{S^2 t_{n-2}^2(\alpha)}{T_{xx}}.$$

Řešením (4.19) je interval jen když je A kladné, což je právě tehdy, když na hladině α je směrnice β_1 průkazně nenulová. Podobně se řeší úloha najít interval spolehlivosti pro x_0, v němž je regresní funkce rovna dané hodnotě μ, jen z výrazu pod odmocninou v (4.18) odpadne jednička. Jednoduchý program dal v příkladu listy následující 95% intervaly spolehlivosti:

```
        x.Hat   xHat.L   xHat.U
[1,] 18.76393 18.15339 19.46873
[2,] 21.28621 20.91857 21.71346
[3,] 26.89329 26.50909 27.34021
```

5. Identifikace

Tato kapitola se týká lineárního modelu, v němž regresní matice $\mathbf{X}$ nemá úplnou hodnost. Budeme se zabývat způsoby, jak z nekonečně mnoha možných řešení normální rovnice zvolit jediné řešení. Je sice pravda, že každý lineární model s neúplnou hodností lze reparametrizovat tak, aby regresní matice měla lineárně nezávislé sloupce (mohli bychom použít již několikrát zmíněnou ortonormální bázi $\mathbf{Q}$), ale mnohdy bychom si zkomplikovali samotný model a především interpretaci zjištěných závěrů. To platí zejména o modelech analýzy rozptylu.

5.1. Nejkratší řešení

Nejprve uvedeme řešení, které je spíše zajímavé, než aby bylo praktické.

Připomeňme, že Mooreova-Penroseho pseudoinverze $\mathbf{X}^+$ k matici $\mathbf{X}$ vyhovuje vztahům $\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}$, $\mathbf{X}^+\mathbf{X}\mathbf{X}^+ = \mathbf{X}^+$, přičemž matice $\mathbf{X}^+\mathbf{X}$ a $\mathbf{X}\mathbf{X}^+$ jsou symetrické (viz například (Rao, 1978, odst. 1b. 5 (VIII))) je dána jednoznačně.

Věta 5.1. Vektor $\mathbf{b}^+ = \mathbf{X}^+\mathbf{Y}$ je nejkratším řešením normální rovnice $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$.

D ů k a z: Nejprve dosadíme $\mathbf{b}^+$ do levé strany normální rovnice:

$$
\begin{aligned}
\mathbf{X}'\mathbf{X}\mathbf{b}^+ &= \mathbf{X}'\mathbf{X}\mathbf{X}^+\mathbf{Y} \\
&= \mathbf{X}'(\mathbf{X}\mathbf{X}^+)'\mathbf{Y} \qquad \text{(ze symetrie } \mathbf{X}\mathbf{X}^+\text{)} \\
&= (\mathbf{X}\mathbf{X}^+\mathbf{X})'\mathbf{Y} \qquad \text{(ale platí } \mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}\text{)} \\
&= \mathbf{X}'\mathbf{Y},
\end{aligned}
$$

což dokazuje, že $\mathbf{b}^+$ je řešením normální rovnice.

Z teorie lineárních rovnic je známo, že vektor $\mathbf{b}$ je řešením normální rovnice, právě když platí $\mathbf{b} = \mathbf{b}^+ + \mathbf{a}$, kde je $\mathbf{X}'\mathbf{X}\mathbf{a} = \mathbf{0}$, což je totéž, jako $\mathbf{X}\mathbf{a} = \mathbf{0}$. Proveďme pomocný výpočet

$$
\begin{aligned}
\mathbf{a}'\mathbf{b}^+ &= \mathbf{a}'\mathbf{X}^+\mathbf{Y} = \mathbf{a}'(\mathbf{X}^+\mathbf{X})\mathbf{X}^+\mathbf{Y} \\
&= \mathbf{a}'(\mathbf{X}^+\mathbf{X})'\mathbf{X}^+\mathbf{Y} = (\mathbf{a}'\mathbf{X}')\mathbf{X}^{+'}\mathbf{X}^+\mathbf{Y} = \mathbf{0}.
\end{aligned}
$$

Nyní můžeme zdola omezit čtverec délky vektoru $\mathbf{b}$:

$$\|\mathbf{b}\|^2 = \|\mathbf{b}^+ + \mathbf{a}\|^2 = \|\mathbf{b}^+\|^2 + 2\mathbf{a}'\mathbf{b}^+ + \|\mathbf{a}\|^2 \geq \|\mathbf{b}^+\|^2.$$

□

Poznámka Matici $\mathbf{X}^+$ lze zkonstruovat pomocí rozkladu podle singulárních hodnot (A.6) jako $\mathbf{X}^+ = \mathbf{V}^0\mathbf{D}^{-1}\mathbf{U}^{0'}$. Snadno se ověří, že jsou splněny všechny čtyři požadavky na Mooreovu-Penroseho matici. V prostředí R lze vektor $\mathbf{X}^+$ počítat pomocí následující procedury:

```
mp.inv <- function(X,eps=sqrt(.Machine$double.eps)){
 a <- svd(X)
 nn <- a$d>eps*a$d[1]
 if (any(nn)) a$v[,nn]%*%(t(a$u[,nn])/a$d[nn]) else X*0
}
```

K vysvětlení funkce `mp.inv()` je třeba poznamenat, že funkce `svd()` dá v prostředí R všechny tři matice z rozkladu podle singulárních hodnot (A.8), přičemž diagonála `a$d` matice $\mathbf{D}$ (tedy singulární hodnoty) tvoří nerostoucí posloupnost (a matice $\mathbf{U}, \mathbf{V}$ mají odpovídajícím způsobem uspořádané sloupce).

Příklad 5.1 (měď) Na pěti místech bylo nepřímo hodnoceno znečištění řeky tak, že vždy u sedmi vylovených ryb byl zjištěn logaritmus koncentrace mědi. Data jsou uvedena v knížce Zvára (1998). Jedná se o úlohu analýzy rozptylu jednoduchého třídění. Použijeme-li parametrizaci (2.13), nejsou hlavní efekty $\alpha_1, \ldots, \alpha_5$ odhadnutelné. K výpočtu nejkratšího řešení normální rovnice pro odhady parametrů $\mu, \alpha_1, \ldots, \alpha_5$ použijeme právě zavedenou funkci `mp.inv`.

```
> attach(Med)
> X <- 1; for (m in levels(Misto)) X <- cbind(X,Misto==m)
> print(b.plus <- as.vector(mp.inv(X)%*%lnCu))
[1]  0.30230952  0.26611905  0.18126190  0.19297619 -0.36502381  0.02697619
```

Snadno lze zjistit (např. pomocí `sqrt(crossprod())`), že je $\|\mathbf{b}^+\| = 0{,}605$, kdežto při standardní parametrizaci R vyjde $\|\mathbf{b}\| = 0{,}889$. ○

5.2. Identifikační omezení

Připomeňme, že pro $\mathcal{M}(\mathbf{A}') \subset \mathcal{M}(\mathbf{X}')$ jsou v modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ složky vektoru $\mathbf{A}\boldsymbol{\beta}$ odhadnutelné, takže požadavkem na splnění netriviální konzistentní soustavy lineárních rovnic $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ jsme v oddílu 3.3 určili podmodel. Lze očekávat, že k novému účelu (určení jediného řešení normální rovnice) musíme použít nějaká jiná lineární omezení. Jistě, podle věty 2.3 inkluse $\mathcal{M}(\mathbf{A}') \subset \mathcal{M}(\mathbf{X}')$ znamená, že vektor $\mathbf{Ab}$ je pro všechna řešení normální rovnice $\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{Y}$ stejný. K určení jediného řešení takovou matici použít nemůžeme.

Uvažujme jako určující (identifikační) omezení vektoru $\boldsymbol{\beta}$ soustavu lineárních rovnic. Řekneme, že omezení (tzv. *reparametrizační* rovnice)

$$\mathbf{A}\boldsymbol{\beta} = \mathbf{0} \tag{5.1}$$

identifikuje vektor $\boldsymbol{\beta}$ v modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, když ke každému $\boldsymbol{\mu} \in \mathcal{M}(\mathbf{X})$ existuje jediný vektor $\boldsymbol{\beta}$, který splňuje současně

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \quad \mathbf{A}\boldsymbol{\beta} = \mathbf{0}.$$

Věta 5.2. (Scheffého) Omezení (5.1) identifikuje vektor $\boldsymbol{\beta}$ právě, když platí

$$\mathcal{M}(\mathbf{A}') \cap \mathcal{M}(\mathbf{X}') = \{\mathbf{0}\}, \tag{5.2}$$
$$\mathsf{h}(\mathbf{X}) + \mathsf{h}(\mathbf{A}) = k + 1. \tag{5.3}$$

D ů k a z: První požadavek zajišťuje existenci $\boldsymbol{\beta}$, druhý jeho jednoznačnost. Začneme existencí (omezení na $\boldsymbol{\beta}$ nesmí být příliš silné). Pro každé $\boldsymbol{\mu} \in \mathcal{M}(\mathbf{X})$ musí mít rovnice v $\boldsymbol{\beta}$

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{A} \end{pmatrix} \boldsymbol{\beta} = \mathbf{D}\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{pmatrix}$$

nějaké řešení. Pro každé $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ tedy musí platit

$$\left\{ \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix} : \boldsymbol{\beta} \in \mathbb{R}^{k+1} \right\} \subset \mathcal{M}(\mathbf{D}),$$

což je postupně ekvivalentní se vztahy

$$\begin{aligned}
&\mathcal{M}(\mathbf{D})^{\perp} \subset \left\{ \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix} : \boldsymbol{\beta} \in \mathbb{R}^{k+1} \right\}^{\perp} && , \\
&(\mathbf{v}_1', \mathbf{v}_2')\,\mathbf{D} = \mathbf{0} \Rightarrow (\mathbf{v}_1', \mathbf{v}_2') \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix} = \mathbf{0} && \text{pro všechna } \boldsymbol{\beta}, \\
&\mathbf{v}_1'\mathbf{X} = -\mathbf{v}_2'\mathbf{A} \Rightarrow \mathbf{v}_1'\mathbf{X} = \mathbf{0}'.
\end{aligned}$$

Poslední implikaci lze interpretovat tak, že každý vektor, který je současně v $\mathcal{M}(\mathbf{X}')$ a $\mathcal{M}(\mathbf{A}')$, musí být nutně nulový, což je přesně požadavek (5.2).

Požadavek na jednoznačnost je požadavkem na hodnost matice $\mathbf{D}$. Protože řádky matice $\mathbf{X}$ hodnosti r jsou také řádky matice $\mathbf{D}$, musí platit $\mathsf{h}(\mathbf{A}) \geq k + 1 - r$. Protože ale lineární obaly řádků matic $\mathbf{X}', \mathbf{A}'$ mají společný pouze nulový vektor, musí nutně platit (5.3). □

Prakticky si můžeme představit hledání jediného řešení normální rovnice jako řešení soustavy rovnic

$$\begin{aligned}
\mathbf{X}'\mathbf{Xb} &= \mathbf{X}'\mathbf{Y} \\
\mathbf{A}'\mathbf{Ab} &= \mathbf{0},
\end{aligned}$$

neboť druhá rovnice je ekvivalentní se vztahem $\mathbf{Ab} = \mathbf{0}$. Řešení soustavy musí vyhovovat také rovnici $\mathbf{D}'\mathbf{D}\mathbf{b} = \mathbf{X}'\mathbf{Y}$, takže vyjde

$$\mathbf{b} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{X}'\mathbf{Y}. \tag{5.4}$$

Uvedený postup lze prakticky zařídit tak, že regresní matici $\mathbf{X}$ rozšíříme o řádky matice $\mathbf{A}$ na matici $\mathbf{D}$ a současně vektor $\mathbf{Y}$ rozšíříme o stejný počet nul.

Příklad 5.2 (jednoduché třídění) Model jednoduchého třídění jsme zavedli již v (2.13). Příslušnou matici plánu $\mathbf{X}$ jsme uvedli v (2.14). Jako reparametrizační podmínku (umožňující určení jediného řešení normální rovnice) lze použít každé omezení

$$a_0\mu + \sum_{i=1}^{I} a_i\alpha_i = 0,$$

jehož levá strana není odhadnutelná, tedy nemá tvar (2.15). Nesmí tedy být $\sum_{i=1}^{I} a_i = a_0$. Tomu odpovídají například následující matice a odpovídající podmínky:

$$\mathbf{A} = (0, 1, \ldots, 1) \longleftrightarrow \sum_{i=1}^{I} \alpha_i = 0, \tag{5.5}$$

$$\mathbf{A} = (0, n_1, \ldots, n_I) \longleftrightarrow \sum_{i=1}^{I} n_i\alpha_i = 0,$$

$$\mathbf{A} = \mathbf{j}_j' \longleftrightarrow \alpha_j = 0 \text{ pro zvolené } j. \tag{5.6}$$

Jak uvidíme v příští kapitole, omezení (5.5) a (5.6) lze v prostředí R uplatnit. ○

Příklad 5.3 (měď) Omezení (5.6) pro $j = 1$ dostaneme pomocí matice $\mathbf{A} = (0, 1, 0, 0, 0, 0)$. Navážeme na příklad 5.1.

```
> D <- rbind(X,c(0,1,0,0,0,0))
> print(as.vector(b.1 <- solve(t(D)%*%D)%*%t(X)%*%lnCu))
[1]  5.684286e-01  1.729283e-15 -8.485714e-02 -7.314286e-02 -6.311429e-01
[6] -2.391429e-01
> c(as.vector(crossprod(lnCu-X%*%b.1)),deviance(a<-lm(lnCu~Misto)))
  [1] 2.284876   2.284876
> summary(a)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.56843    0.10431   5.449 6.55e-06
MistoB      -0.08486    0.14752  -0.575 0.569416
MistoC      -0.07314    0.14752  -0.496 0.623625
MistoD      -0.63114    0.14752  -4.278 0.000177
MistoE      -0.23914    0.14752  -1.621 0.115452
F-statistic: 5.896 on 4 and 30 DF,  p-value: 0.001265
```

Snadno lze ověřit shodu jednotlivých složek vektoru $\mathbf{b}$ i reziduálního součtu čtverců 2,285. ○

6. Analýza rozptylu

6.1. Jednoduché třídění

Model analýzy rozptylu jednoduchého třídění jsme zavedli již v 2. kapitole. Předpokládáme, že máme nezávislé náhodné veličiny $Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{I1}, \ldots, Y_{In_I}$, pro které platí $Y_{ij} \sim \mathsf{N}\big(\mu_i, \sigma^2\big)$. Jde tedy o I nezávislých výběrů z normálního rozdělení, přičemž u každého výběru připouštíme obecně jinou střední hodnotu, rozptyl je ve všech výběrech stejný.

Úlohu zapíšeme jako normální lineární model $\mathbf{Y} \sim \mathsf{N}\big(\mathbf{X}_a\boldsymbol{\beta}, \sigma^2\mathbf{I}\big)$, zvolíme-li

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_I \end{pmatrix}, \quad \mathbf{X}_a = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{1}_{n_I} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{pmatrix}, \tag{6.1}$$

kde vektor $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ obsahuje pozorování z i-tého výběru. Zřejmě vyjde $b_i = \bar{Y}_{i\bullet}$ (průměr v i-tém výběru) a tedy reziduální součet čtverců je roven

$$SS_e = \sum_{i=1}^{I}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y}_{i\bullet})^2.$$

Běžně testovaná hypotéza $\mathsf{H}_0 : \mu_1 = \ldots = \mu_I$ vede k podmodelu, který je dán regresní maticí $\mathbf{X}_0 = \mathbf{1}_n$, kde $n = \sum_{i=1}^{I} n_i$. Tentokrát vyjde $b_0 = \bar{Y}$ (průměr ze všech n pozorování). Odtud je celkový součet čtverců roven

$$SS_T = \sum_{i=1}^{I}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y})^2.$$

Snadno lze spočítat také

$$\mathbf{d} = \hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0 = \begin{pmatrix} (\bar{Y}_{1\bullet} - \bar{Y})\mathbf{1}_{n_1} \\ \vdots \\ (\bar{Y}_{I\bullet} - \bar{Y})\mathbf{1}_{n_I} \end{pmatrix},$$

odkud snadno vyjde

$$\|\mathbf{d}\|^2 = SS_A = SS_T - SS_e = \sum_{i=1}^{I} n_i(\bar{Y}_{i\bullet} - \bar{Y})^2, \tag{6.2}$$

když jsme zavedli často používané označení SS_A pro součet čtverců vysvětlený (jediným) faktorem A.

Uveďme explicitně rozklad součtu čtverců v analýze rozptylu jednoduchého třídění (celková variabilita=variabilita uvnitř výběrů+variabilita mezi výběry), který vznikne úpravou (6.2)

$$\begin{aligned} \sum_{i=1}^{I}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y})^2 &= \sum_{i=1}^{I}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y}_{i\bullet})^2 + \sum_{i=1}^{I} n_i(\bar{Y}_{i\bullet} - \bar{Y})^2, \\ SS_T &= SS_e + SS_A. \end{aligned} \tag{6.3}$$

O nulové hypotéze rozhodujeme pomocí statistiky (3.10) z věty 7.6:

$$F = \frac{SS_A/(I-1)}{SS_e/(n-I)} = \frac{MS_A}{MS_e}.$$

Výpočet se často vyjadřuje pomocí *tabulky analýzy rozptylu*, jejíž schéma je uvedeno v tabulce 6.1.

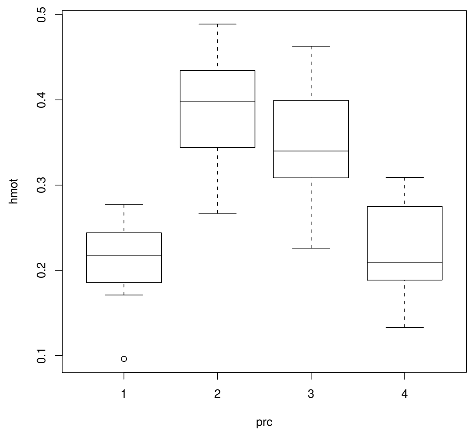
Tabulka 6.1: Tabulka analýzy rozptylu jednoduchého třídění

variabilita	stupně vol.	součet čtverců	průměrné čtverce	F	p
ošetření	$I-1$	SS_A	$MS_A = SS_A/(I-1)$	F	p
reziduální	$n-I$	SS_e	$MS_e = SS_e/(n-I)$	-	-
celková	$n-1$	SS_T	-	-	-

Příklad 6.1 (kořeny) Student zjišťoval hmotnost kořenového systému rostlin pěstovaných v živných roztocích s různými koncentracemi cukru (viz obrázek 6.1). Pomocí funkce `anova()` uplatněné na výsledek procedury `lm()` dostaneme tabulku analýzy rozptylu

```
> anova(lm(hmotnost~Procento,data=Koreny))
Analysis of Variance Table

Response: hmotnost
          Df   Sum Sq  Mean Sq F value    Pr(>F)
Procento   3 0.312687 0.104229  28.568 6.641e-11
Residuals 50 0.182422 0.003648
---
```

z níž je patrné, že rozdíl mezi roztoky je průkazný. Identický výsledek by dala procedura: `summary(aov(hmotnost~Procento,data=Koreny))`. ○



Obrázek 6.1: Závislost hmotnosti kořenové části na procentu cukru v živném roztoku

6.1.1. Kontrasty

Uvažujme nyní klasickou parametrizaci $\mathsf{E}\, Y_{it} = \mu + \alpha_i$ úlohy jednoduchého třídění. Vektor parametrů má tvar $\boldsymbol{\beta}' = (\mu, \boldsymbol{\alpha}') = (\mu, \alpha_1, \dots, \alpha_I)$, regresní matice pak $\mathbf{X} = (\mathbf{1}, \mathbf{X}_a)$, kde $\mathbf{X}_a$ jsme zavedli v (6.1). Připomeňme zjištění příkladu 2.1, podle kterého je v tomto modelu parametrická funkce $\mathbf{t}'\boldsymbol{\beta}$ odhadnutelná, když vektor $\mathbf{t}$ má tvar $\mathbf{t} = (\mathbf{1}'\mathbf{c}, \mathbf{c}')$. Speciální případ odhadnutelné funkce, kdy je $t_0 = \mathbf{1}'\mathbf{c} = \sum c_i = 0$, se nazývá *kontrast*. Je zřejmé, že kontrast závisí pouze na efektech α_i jednotlivých ošetření, nikoliv na μ.

Zabývejme se nyní odhadem kontrastu. Označme $\mathbf{D} = \mathbf{X}_a'\mathbf{X}_a = \operatorname{diag}\{n_1, \dots, n_I\}$ a $\mathbf{n} = (n_1, \dots, n_I)'$. Matice $\mathbf{X}'\mathbf{X}$ má tvar

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \mathbf{n}' \\ \mathbf{n} & \mathbf{D} \end{pmatrix}.$$

Není sice regulární, ale snadno se zjistí, k jejím pseudoinverzím patří také

$$(\mathbf{X}'\mathbf{X})^- = \begin{pmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{D}^{-1} \end{pmatrix}.$$

Označme $\mathbf{b}' = (m, \mathbf{a}')'$ jakékoliv řešení normální rovnice v modelu analýzy rozptylu jednoduchého třídění. Pro odhad $\mathbf{c}'\mathbf{a}$ kontrastu $(0, \mathbf{c}')'\boldsymbol{\beta} = \mathbf{c}'\boldsymbol{\alpha}$ tedy platí

$$\mathbf{c}'\mathbf{a} \sim \mathsf{N}\big(\mathbf{c}'\boldsymbol{\alpha}, \sigma^2\mathbf{c}\mathbf{D}^{-1}\mathbf{c}\big) = \mathsf{N}\left(\sum_{i=1}^{I} c_i\alpha_i, \sigma^2\sum_{i=1}^{I}\frac{c_i^2}{n_i}\right).$$

Kontrasty dané vektory $\mathbf{c}$ a $\mathbf{d}$ se nazývají *ortogonální kontrasty*, když jsou vektory $\mathbf{c}, \mathbf{d}$ ortogonální. Kovariance jejich odhadů je rovna

$$\sigma^2 \mathbf{c}' \mathbf{D}^{-1} \mathbf{d} = \sigma^2 \sum_{i=1}^{I} \frac{c_i d_i}{n_i}.$$

V případě, že model analýzy rozptylu je *vyvážený*, tj. platí $n_1 = \ldots = n_I = T$, budou pak odhady $\mathbf{c}'\mathbf{a}$ a $\mathbf{d}'\mathbf{a}$ nutně nezávislé (viz tvrzení f) věty 2.5).

6.1.2. Test lineární hypotézy pomocí kontrastů

Věnujme se nyní testování nulové hypotézy $\mathsf{H}_0 : \alpha_1 = \ldots = \alpha_I$. Pomocí $I - 1$ kontrastů

$$\alpha_1 - \alpha_I, \alpha_2 - \alpha_I, \ldots, \alpha_{I-1} - \alpha_I,$$

lze souhrnně zapsat nulovou hypotézu H_0 jako požadavek

$$\mathbf{C}'\boldsymbol{\alpha} = \mathbf{0}, \tag{6.4}$$

kde jsme použili označení

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \\ -1 & -1 & \ldots & -1 \end{pmatrix}. \tag{6.5}$$

V prostředí R je tato matice označována jako `contr.sum(I)`. Rozhodování o hypotéze H_0 (o nezávislosti Y na sledovaném faktoru) pomocí testování ověřitelné lineární hypotézy (6.4) s maticí (6.5) spočívá v porovnání jednotlivých efektů α_i s efektem I-tého ošetření α_I.

Jinou možností, jak vyjádřit H_0 ve tvaru lineárního omezení (6.4), je použít matici

$$\mathbf{C} = \begin{pmatrix} -1 & -1 & \ldots & -1 \\ 1 & -1 & \ldots & -1 \\ 0 & 2 & \ldots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & I-1 \end{pmatrix}. \tag{6.6}$$

Tato Helmertova matice (v prostředí R nazvaná `contr.helmert(I)`) odpovídá posloupnosti omezení

$$\begin{aligned} -\alpha_1 + \alpha_2 &= 0, \\ -\alpha_1 - \alpha_2 + 2\alpha_3 &= 0, \\ &\ldots \\ -\alpha_1 - \cdots - \alpha_{I-1} + (I-1)\alpha_I &= 0. \end{aligned}$$

Postupně srovnáváme druhý až I-tý efekt s aritmetickým průměrem efektů s nižšími indexy.

Je ihned zřejmé, že sloupce matice $\mathbf{C}$ z (6.5) tvoří kontrasty, sloupce matice $\mathbf{C}$ z (6.6) tvoří ortogonální kontrasty.

6.1.3. Reparametrizace pomocí kontrastů

Připomeňme zjištění z příkladu 5.2, že v modelu analýzy rozptylu jednoduchého třídění může mít identifikační omezení tvar $(0, \mathbf{c}')(\mu, \boldsymbol{\alpha}')' = 0$, kde ovšem součet $\mathbf{1}'\mathbf{c}$ složek vektoru $\mathbf{c}$ není nulový, nesmí tedy jít o kontrasty. Přesto však využijeme obě až dosud zavedené matice kontrastů. Přejdeme při tom k úloze s menším počtem parametrů. Později naznačíme, jak tento postup lze rozšířit i na složitější modely analýzy rozptylu.

Místo vektoru efektů $\boldsymbol{\alpha}$ zaveďme vektor $\boldsymbol{\alpha}^*$ o $I-1$ složkách předpisem

$$\boldsymbol{\alpha} = \mathbf{C}\boldsymbol{\alpha}^*, \tag{6.7}$$

kde $\mathbf{C}$ je libovolná matice rorměru $I \times (I-1)$ s lineárně nezávislými sloupci. Vzhledem k této poslední vlastnosti lze psát

$$\boldsymbol{\alpha}^* = (\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\boldsymbol{\alpha}. \tag{6.8}$$

Nyní vyjádříme vektor středních hodnot pomocí nových parametrů. Je zřejmé, že při popisu všech možných středních hodnot není třeba pracovat s celou maticí $\mathbf{X}$, že z identických řádků stačí zachovat pouze jediný. Budeme tak dál postupovat, aniž bychom to opakovaně připomínali. Skutečnou matici $Xmce$ bychom tedy z naší skromnější dostali n_1 násobným zopakováním prvního řádku, n_2 násobným zopakováním druhého řádku atd., zkrátka „svislým rozmazáním“ naší zhuštěné matice $\mathbf{X}$. Vektor středních hodnot v modelu analýzy rozptylu jednoduchého třídění pak lze postupně upravit na

$$\begin{aligned} \mathsf{E}\,\mathbf{Y} &= \mathbf{1}\mu + \mathbf{X}_a\boldsymbol{\alpha} \qquad (\text{připomeňme } \mathbf{X} = (\mathbf{1}, \mathbf{X}_a)) \\ &= \mathbf{1}\mu + \mathbf{X}_a\mathbf{C}\boldsymbol{\alpha}^* \\ &= \mathbf{X}_a(\mathbf{1}, \mathbf{C})\begin{pmatrix} \mu \\ \boldsymbol{\alpha}^* \end{pmatrix}, \end{aligned}$$

když jsme využili zřejmý vztah $\mathbf{X}_a\mathbf{1} = \mathbf{1}$. Abychom zachovali stejný prostor středních hodnot (tedy $\mathcal{M}(\mathbf{X}) = \mathcal{M}(\mathbf{X}_a)$), musí být matice $(\mathbf{1}, \mathbf{C})$ regulární s hodností I. Obě až dosud zavedené matice kontrastů tomuto požadavku vyhovují, navíc obě splňují $\mathbf{C}'\mathbf{1} = \mathbf{0}$, takže každý řádek matice $\mathbf{C}'$ určuje jeden kontrast. Přitom efekty $\boldsymbol{\alpha} = \mathbf{C}\boldsymbol{\alpha}^*$ vyhovují identifikačnímu omezení $\mathbf{1}'\boldsymbol{\alpha} = 0$, tedy (5.5).

Podobně je matice $(\mathbf{1}, \mathbf{C})$ regulární i pro matici

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & 1 \end{pmatrix}, \tag{6.9}$$

kterou prostředí R nabízí pod názvem `contr.treatment(I)`. Tentokrát nejsou součty jednotlivých sloupců nulové, takže složky matice $\mathbf{C}'\boldsymbol{\alpha}$ jsou sice odhadnutelné, ale nejsou to už kontrasty. Odpovídá to identifikačnímu omezeni $\alpha_j = 0$ (viz (5.6)) použitému na $\boldsymbol{\alpha} = \mathbf{C}\boldsymbol{\alpha}^*$ pro $j = 1$.

Všimněme si ještě varianční matice odhadu vektoru $(\mu, \boldsymbol{\alpha}^{*'})'$:

$$\begin{aligned} \operatorname{var}\begin{pmatrix} m \\ \mathbf{a}^* \end{pmatrix} &= \sigma^2 \left(\begin{pmatrix} \mathbf{1}' \\ \mathbf{C}' \end{pmatrix} \mathbf{X}_a' \mathbf{X}_a \begin{pmatrix} \mathbf{1} & \mathbf{C} \end{pmatrix} \right)^{-1} = \sigma^2 \left(\begin{pmatrix} \mathbf{1}' \\ \mathbf{C}' \end{pmatrix} \mathbf{D} \begin{pmatrix} \mathbf{1} & \mathbf{C} \end{pmatrix} \right)^{-1} \\ &= \sigma^2 \begin{pmatrix} n & \mathbf{1}'\mathbf{DC} \\ \mathbf{C}'\mathbf{D1} & \mathbf{C}'\mathbf{DC} \end{pmatrix}^{-1} = \sigma^2 \begin{pmatrix} n & \mathbf{n}'\mathbf{C} \\ \mathbf{C}'\mathbf{n} & \mathbf{C}'\mathbf{DC} \end{pmatrix}^{-1} \end{aligned}$$

Existuje situace, kdy je tato varianční matice diagonální, takže v normálním modelu jsou složky odhadu $\mathbf{a}^*$ vektoru $\boldsymbol{\alpha}^*$ nezávislé. Je to v případě, kdy jde opravdu o ortogonální kontrasty (platí $\mathbf{C}'\mathbf{1} = \mathbf{0}$ a matice $\mathbf{C}'\mathbf{C}$ je diagonální) a kdy je současně model *vyvážený*, $(n_1 = \dots = n_I (= T)$, tj. $\mathbf{n} = m\mathbf{1}$ a $\mathbf{D} = T\mathbf{I})$.

6.1.4. Interpretace kontrastů v R

V prostředí R se právě popsaná reparametrizace použije, kdykoliv pomocí funkce `lm()` hledáme závislost na nějakém faktoru.

Uvažujme stále model analýzy rozptylu jednoduchého třídění. Matice $\mathbf{X}_a$ je vlastně maticí umělých proměnných, každý sloupec odpovídá jedné z možných úrovní faktoru. Odhady složek vektoru $(\mu, \boldsymbol{\alpha}^{*'})'$ získáme v R, když na výsledek `lm()` použijeme `summary()`. Proberme nyní jednotlivé možné volby kontrastů, jak jsou dostupné v R.

`contr.treatment`

Jedna z úrovní faktoru se zvolí jako základní a ostatní se s touto úrovní porovnávaji. Identifikační omezení spočívá v tom, že složka vektoru $\boldsymbol{\alpha}$ odpovídající základní úrovni faktoru je nulová. Standardně je základní úrovní faktoru jeho první hodnota. Potom můžeme střední hodnoty v jednotlivých výběrech zapsat jako

$$\begin{aligned} \mathsf{E}\, Y_{1t} &= \mu, & 1 \le t \le n_1, \\ \mathsf{E}\, Y_{it} &= \mu + \alpha_i^*, & 1 \le t \le n_i, 2 \le i \le I. \end{aligned}$$

Snadno tedy můžeme porovnat vliv jednotlivých úrovní faktoru s vlivem jeho základní úrovně.

Příklad 6.2 (kořeny) Pokračujme v naší úloze jednoduchého třídění.

```
> a <- lm(hmotnost~Procento,data=Koreny,contr=list(Procento=contr.treatment))
> summary(a)

Call:
lm(formula = hmotnost ~ Procento, data = Koreny,
                                  contrasts = list(Procento = contr.treatment))

Residuals:
      Min        1Q    Median        3Q       Max
-0.123667 -0.037121 -0.002733  0.041271  0.114867

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.21180    0.01560  13.581  < 2e-16
Procento2    0.17887    0.02339   7.646 5.89e-10
Procento3    0.13633    0.02206   6.181 1.14e-07
Procento4    0.01428    0.02339   0.611    0.544

Residual standard error: 0.0604 on 50 degrees of freedom
Multiple R-Squared: 0.6316,     Adjusted R-squared: 0.6094
F-statistic: 28.57 on 3 and 50 DF,  p-value: 6.641e-11
```

Odhad uvedený v řádku `(Intercept)` je odhadem střední hodnoty v prvním výběru, součet zmíněného odhadu s odhadem `Procento2` dá odhad střední hodnoty ve druhém výběru atd. Snadno si to ověříme, když si tyto odhady (tj. výběrové průměry) necháme spočítat přímo:

```
> coef(a)[1]+c(0,coef(a)[-1])
          Procento2 Procento3 Procento4
0.2118000 0.3906667 0.3481333 0.2260833
> tapply(Koreny$hmotnost,Koreny$Procento,mean)
        1         2         3         4
0.2118000 0.3906667 0.3481333 0.2260833
```

Samozřejmě, odhady středních hodnot v jednotlivých výběrech můžeme s pomocí vztahu $\alpha = \mathbf{C}\alpha^*$ získat také jako

```
> coef(a)[1]+contr.treatment(4)%*%coef(a)[2:4]
```

○

Kontrast `contr.treatment` je standardním nastavením v R. Pokud jsme nastavení kontrastů nezměnili, nebylo třeba parametr `contrasts` uvádět.

`contr.helmert`

(Standardní nastavení v S+) Pro Helmertovu matici platí $\mathbf{C}'\mathbf{1} = \mathbf{0}$, takže jednotlivé složky vektoru $\mathbf{C}\boldsymbol{\alpha}$ jsou skutečně kontrasty. Dalším důsledkem tohoto vztahu je

$$\sum_i^I \alpha_i = \mathbf{1}'\boldsymbol{\alpha} = \mathbf{1}'\mathbf{C}\boldsymbol{\alpha}^* = \mathbf{0}'\boldsymbol{\alpha}^* = 0,$$

což je, jak víme z příkladu 5.2, identifikační omezení. Důsledkem je vztah $\mu = \sum_i \mu_i/I$, totéž platí pro odhady. Proto je odhadem μ nevážený průměr průměrů $\bar{Y}_{i\bullet}$ jednotlivých výběrů.

Matice $\mathbf{C}'\mathbf{C}$ pro $\mathbf{C}$ z (6.6) je zřejmě diagonální s prvky $i + i^2 = i(i+1)$ na diagonále. Proto lze snadno vyjádřit složky $\boldsymbol{\alpha}^*$ pomocí $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}^* = (\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\boldsymbol{\alpha},$$

odkud je (pro $i = 1, \ldots, I-1$)

$$\begin{aligned}\alpha_i^* &= \frac{1}{i(i+1)}\left(i\alpha_{i+1} - \sum_{t=1}^{i}\alpha_t\right) = \frac{1}{i+1}\left(\alpha_{i+1} - \frac{1}{i}\sum_{t=1}^{i}\alpha_t\right) \\ &= \frac{1}{i+1}\left(\mathsf{E}\,Y_{ij} - \frac{1}{i}\sum_{t=1}^{i}\mathsf{E}\,Y_{tj}\right).\end{aligned}$$

Porovnáváme tedy vždy další efekt s aritmetickým průměrem předchozích, resp. střední hodnotu v dalším výběru s průměrem středních hodnot výběrů s menšími indexy.

Příklad 6.3 (kořeny)

```
> summary(lm(hmotnost~Procento,contrast=list(Procento=contr.helmert),data=Koreny))

Call:
lm(formula = hmotnost ~ Procento, data = Koreny,
                                  contrasts = list(Procento = contr.helmert))

Residuals:
      Min        1Q    Median        3Q       Max
-0.123667 -0.037121 -0.002733  0.041271  0.114867

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.294171   0.008271  35.567  < 2e-16
Procento1    0.089433   0.011697   7.646 5.89e-10
Procento2    0.015633   0.006498   2.406   0.0199
Procento3   -0.022696   0.004949  -4.586 3.05e-05

Residual standard error: 0.0604 on 50 degrees of freedom
Multiple R-Squared: 0.6316,     Adjusted R-squared: 0.6094
F-statistic: 28.57 on 3 and 50 DF,  p-value: 6.641e-11
```

Například v řádku Procento2 je tedy uvedena třetina rozdílu průměrné hmotnosti ve třetí skupině a (neváženého!) průměru z hmotností v prvních dvou skupinách. ○

`contr.sum`

Také v tomto případě jsou složky vektoru $\mathbf{C}'\boldsymbol{\alpha}$ kontrasty, opět splňují identifikační podmínku $\sum \alpha_i = 0$, takže například odhad μ je identický s odhadem tohoto parametru pro `contr.helmert`. Vzhledem k tvaru matice $\mathbf{C}$ z (6.5) platí

$$\boldsymbol{\alpha} = \mathbf{C}\boldsymbol{\alpha}^* = \begin{pmatrix} \mathbf{I} \\ -\mathbf{1}' \end{pmatrix} \boldsymbol{\alpha}^* = \begin{pmatrix} \boldsymbol{\alpha}^* \\ -\mathbf{1}'\boldsymbol{\alpha}^*. \end{pmatrix}$$

Každá ze složek $\boldsymbol{\alpha}^*$ je tedy totožná odpovídající složce $\boldsymbol{\alpha}$ při identifikaci pomocí $\sum \alpha_i = 0$. Poslední složku α_I bychom dostali tak, že sečteme jejích prvních $I - 1$ složek a obrátíme znaménko.

Příklad 6.4 (kořeny)

```
> summary(lm(hmotnost~Procento,contrast=list(Procento=contr.sum),data=Koreny))

Call:
lm(formula = hmotnost ~ Procento, data = Koreny,
                                  contrasts = list(Procento = contr.sum))

Residuals:
      Min        1Q    Median        3Q       Max
-0.123667 -0.037121 -0.002733  0.041271  0.114867

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.294171   0.008271  35.567  < 2e-16
Procento1   -0.082371   0.013785  -5.975 2.39e-07
Procento2    0.096496   0.014847   6.499 3.64e-08
Procento3    0.053962   0.013785   3.915 0.000274

Residual standard error: 0.0604 on 50 degrees of freedom
Multiple R-Squared: 0.6316,     Adjusted R-squared: 0.6094
F-statistic: 28.57 on 3 and 50 DF,  p-value: 6.641e-11
```

○

6.1.5. Reparametrizace pro uspořádaný faktor

Hodnoty uspořádaného faktoru (`ordered`) jsou uspořádány. V proceduře `lm()` se uspořádanému faktoru standardně přiřazuje matice kontrastů `contr.poly(I)`, jejíž sloupce jsou dány ortogonálními polynomy. Například pro $I = 4$ je to matice

```
> contr.poly(4)
             .L   .Q         .C
[1,] -0.6708204  0.5 -0.2236068
[2,] -0.2236068 -0.5  0.6708204
[3,]  0.2236068 -0.5 -0.6708204
[4,]  0.6708204  0.5  0.2236068
```

Jak už označení sloupců naznačuje, souvisí jednotlivé sloupce této matice s lineárním, kvadratickým … trendem. Pokud je model vyvážený (četnosti n_i jsou shodné), jsou odhady složek α_i^* nezávislé.

Příklad 6.5 (kořeny) Teprve nyní bereme v úvahu, že úrovně použitého faktoru jsou uspořádány (jsou to procenta cukru v živném roztoku). Jednotlivé složky vektoru $\boldsymbol{\alpha}^*$ se tedy snaží zachytit lineární, kvadratický či kubický trend. Samozřejmě, za předpokladu, že hodnoty uspořádaného faktoru (ordinálního znaku) jsou od sebe ekvidistantně vzdálené (že jde vlastně o intervalové měřítko).

```
> summary(lm(hmotnost~Procento,contrast=list(Procento=contr.poly),data=Koreny))

Call:
lm(formula = hmotnost ~ Procento, data = Koreny,
                                  contrasts = list(Procento = contr.poly))

Residuals:
      Min        1Q    Median        3Q       Max
-0.123667 -0.037121 -0.002733  0.041271  0.114867

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.942e-01  8.271e-03  35.567  < 2e-16
Procento.L   7.081e-05  1.654e-02   0.004   0.9966
Procento.Q  -1.505e-01  1.654e-02  -9.096 3.53e-12
Procento.C   3.173e-02  1.654e-02   1.918   0.0608

Residual standard error: 0.0604 on 50 degrees of freedom
Multiple R-Squared: 0.6316,     Adjusted R-squared: 0.6094
F-statistic: 28.57 on 3 and 50 DF,  p-value: 6.641e-11
```

Tabulka analýzy rozptylu je samozřejmě totožná s výpočty při jiných volbách matice kontrastů. Ovšem z právě uvedených výsledků je zřejmé, co způsobilo zamítnutí nulové hypotézy o nezávislosti hmotnosti kořenových částí na procentu tuku v živném roztoku. Závislost bude zřejmě blízká kvadratické závislosti na koncentraci cukru v žívném roztoku. ◯

6.2. Analýza rozptylu dvojného třídění

Předpokládáme, že nezávislé náhodné veličiny Y_{ijt} mají normální rozdělení

$$\mathsf{N}\big(\mu + \alpha_i + \beta_j + \gamma_{ij}, \sigma^2\big),$$

přičemž je $1 \le t \le n_{ij}, 1 \le i \le I, 1 \le j \le J$. Vedle (hlavních) efektů se v našem modelu vyskytují také *interakce* γ_{ij}, které se někdy značí jako $(\alpha\beta)_{ij}$. Interakce ukazují, nakolik není vliv sledovaných dvou faktorů aditivní, nakolik není závislost střední hodnot závisle proměnné Y na faktoru A stejná pro různé úrovně faktoru B. Matice plánu je složena ze tří částí, které odpovídají po řadě koeficientům α, β, γ. Když podobně jako u jednoduchého třídění budeme psát z identických řádku vždy jediný, bude

$$\mathbf{X} = (\mathbf{1}, \mathbf{X}_a, \mathbf{X}_b, \mathbf{X}_{ab}),$$

kde je

$$\mathbf{1} = \mathbf{1}_I \otimes \mathbf{1}_J, \qquad \mathbf{X}_a = \mathbf{I}_I \otimes \mathbf{1}_J, \qquad \mathbf{X}_b = \mathbf{1}_I \otimes \mathbf{I}_J, \qquad \mathbf{X}_{ab} = \mathbf{I}_I \otimes \mathbf{I}_J. \tag{6.10}$$

K tomu, aby bylo možno s interakcemi pracovat, musíme mít více pozorování, než kolik činí hodnost regresní matice $\mathbf{X}$, tedy než $I \cdot J$. Celkový počet pozorování opět označíme $n = \sum n_{ij}$. Odhadem středních hodnot $\mathsf{E}\, Y_{ijt}$ jsou nepochybně průměry $\bar{Y}_{ij\bullet}$. Odtud je zřejmé, že reziduální součet čtverců je roven

$$SS_T = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{n_{ij}} \big(Y_{ijt} - \bar{Y}_{ij\bullet}\big)^2 .$$

K identifikaci se zpravidla používají vztahy

$$\sum_{i=1}^{I} \alpha_i = 0, \qquad \sum_{j=1}^{J} \beta_j = 0,$$
$$\sum_{i=1}^{I} \gamma_{ij} = 0 \qquad \text{pro všechna } j,$$
$$\sum_{j=1}^{J} \gamma_{ij} = 0 \qquad \text{pro všechna } i.$$

6.2.1. Reparametrizace pomocí kontrastů

K reparametrizaci lze znovu použít matic kontrastů

$$\mathsf{E}\, \mathbf{Y} = \mathbf{1}\mu + \mathbf{X}_a \mathbf{C}_a \alpha^* + \mathbf{X}_b \mathbf{C}_b \boldsymbol{\beta}^* + \mathbf{X}_{ab} \mathbf{C}_{ab} \boldsymbol{\gamma}^*.$$

Snadno se zjistí, že lze použít $\mathbf{C}_{ab} = \mathbf{C}_a \otimes \mathbf{C}_b$, přičemž dvě matice na pravé straně nemusí mít stejné vlastnosti, lze kombinovat například `contr.treatment` a `contr.sum`. K tomu, aby sloupce matice $\mathbf{C}_{ab}$ tvořily skutečné kontrasty stačí, aby aspoň jedna ze zúčastněných matic měla tuto vlastnost. Pak totiž platí

$$\mathbf{1}'\mathbf{C}_{ab} = \left(\mathbf{1}' \otimes \mathbf{1}'\right)(\mathbf{C}_a \otimes \mathbf{C}_b) = \left(\mathbf{1}'\mathbf{C}_a\right) \otimes \left(\mathbf{1}'\mathbf{C}_b\right) = \mathbf{0}'.$$

Pokud je pro každou kombinaci úrovní obou faktorů stejný počet pozorování, tj. pokud je $n_{ij} = T$ pro všechna i a j a pokud obsahují matice $\mathbf{C}_a$ a $\mathbf{C}_b$ ortogonální kontrasty, zjistíme stejně jako u jednoduchého třídění, že varianční matice odhadů parametrů $\mu, \alpha^*, \beta^*, \gamma^*$ je diagonální.

6.2.2. Tabulka analýzy rozptylu

V tabulce analýzy rozptylu jednoduchého třídění tab. 6.1 je uveden rozklad celkového součtu čtverců SS_T na dva sčítance, z nichž první udává variabilitu vysvětlenou uvažovanou závislostí a druhý udává variabilitu nevysvětlenou. Když však vysvětlujeme variabilitu několika faktory, lze tabulku zobecnit více způsoby. Při výkladu budeme pod členem rozumět buď faktor nebo interakci složenou z jednoho či (v obecnějším modelu) několika faktorů. Každému členu odpovídá v tabulce analýzy rozptylu jeden řádek. Pod řádem členu budeme rozumět řád interakce, pokud je člen interakcí, nebo nulu, pokud jde o samostatný faktor (tzv. *hlavní efekt*).

Procedura `anova()` v R s jediným argumentem třídy `lm` generuje *rozklad typu I*, přičemž jednotlivé řádky postupně udávají, o kolik se přidáním daného členu zmenší reziduální součet čtverců. Obecně tedy závisí na pořadí, v jakém se jednotlivé členy objevují. Ve sloupci *Součet čtverců* je uvedeno, nakolik daný člen (faktor, interakce) přispěl k vysvětlení variability vysvětlované proměnné nad to, co už vysvětlily členy výše uvedené. Testová statistika F pak vzniká jako podíl příslušného průměrného čtverce v daném řádku a odhadu rozptylu S^2 (průměrného čtverce v řádku *reziduální*). V každém řádku tedy statistika F (prostřednictvím příslušné dosažené hladiny testu p) vypovídá o tom, zda vysvětlovaná proměnná *po adjustaci* vůči všem výše uvedeným členům závisí na daném členu (faktoru, interakci). Vypovídá o významnosti té části variability závisle proměnné, kterou nelze vysvětlit pomocí všech výše uvedených členů. Program R má tu nevýhodu, že o pořadí jednotlivých členů můžeme rozhodnout jen do jisté míry, jen v rámci dané úrovně interakcí. Vy výstupu se vždy objeví nejprve základní efekty (interakce nultého řádu), pak interakce prvního řádu (dvojic faktorů) atd.

Rozklad typu II hodnotí přínos daného členu (faktoru, interakce) po adjustaci vůči ostatním členům, které jsou stejného či nižšího řádu (neobsahují více základních faktorů). Vyjdeme z modelu, který obsahuje testovaný člen a všechny členy modelu, které jsou stejného nebo nižšího řádu. Zjistíme, o kolik se zvětší reziduální součet čtverců, když testovaný člen odstraníme. Jako odhad rozptylu použijeme výchozí model, který jsme uvedli při volání funkce `lm()`. Test tedy vypovídá o možnosti vypustit daný člen z modelu když jej přidáváme jako poslední člen daného řádu. Výsledné statistiky nezáleží na tom, v jakém pořadí jsme uvedli faktory v definici modelu.

Rozklad typu III podobně hodnotí přínos daného členu po adjustaci vůči všem ostatním členům bez ohledu na jejich řád. Pro tento rozklad je obtížné hledat interpretaci, protože hodnotí vzrůst reziduálního součtu čtverců způsobený vyloučením daného členu, když v modelu zůstanou (je provedena adjustace vůči nim) všechny ostatní členy včetně případných interakcí, v nichž je člen obsažen. Výsledné statistiky opět nezáleží na tom, v jakém pořadí jsme uvedli faktory v definici modelu.

Rozklad typu II je určitým kompromisem mezi rozklady typu I a III.

V R lze počítat tabulky analýzy rozptylu typu II a III v knihovně `car` procedurou `Anova()`. S určitou námahou lze ke stejným statistikám dojít i pomocí klasické procedury `anova()`, když porovnáme vždy příslušný model a podmodel.

Příklad 6.6 (ICHS) Dlouhodobě byla sledována řada mužů středního věku, u nichž byl zjištěn právě jeden rizikový faktor ischemické chroby srdeční (silné kouření, vysoký krevní tlak, obezita, rodinná disposice). Zajímáme se o možnou závislost indexu obezity BMI (body mass index) na dosaženém vzdělání a na kouření. Použijeme jen údaje o silých kuřácích a o nekuřácích. Když připustíme interakce mezi oběma faktory, dostaneme tabulku analýzy typu I, která postupně vysvětluje celkovou variabilitu indexu obezity.

```
> anova(lm(bmi~Vzdel*Kurak,data=IchsN))
Analysis of Variance Table
Response: bmi
             Df  Sum Sq Mean Sq F value  Pr(>F)
Vzdel         2   14.90    7.45  0.9204 0.40044
Kurak         1   61.82   61.82  7.6356 0.00639
Vzdel:Kurak   2   12.17    6.09  0.7516 0.47324
Residuals   161 1303.44    8.10
```

Všimněme si, že změna pořadí faktorů vede k jiné tabulce:

```
> anova(lm(bmi~Kurak*Vzdel,data=IchsN))
Analysis of Variance Table
Response: bmi
             Df  Sum Sq Mean Sq F value  Pr(>F)
Kurak         1   41.01   41.01  5.0651 0.02577
Vzdel         2   35.71   17.86  2.2057 0.11349
Kurak:Vzdel   2   12.17    6.09  0.7516 0.47324
Residuals   161 1303.44    8.10
```

Rozklady typu II a III už nezávisí na pořadí faktorů:

```
> Anova(lm(bmi~Vzdel*Kurak,data=IchsN))
Anova Table (Type II tests)
Response: bmi
             Sum Sq  Df F value  Pr(>F)
Vzdel         35.71   2  2.2057 0.11349
```

```
Kurak         61.82   1  7.6356 0.00639
Vzdel:Kurak   12.17   2  0.7516 0.47324
Residuals   1303.44 161
```

V rozkladu typu II jsou součet čtverců, F statistika i dosažená hladina u proměnné Kurak totožné s odpovídajícími statistikami v té tabulce typu I, v níž je tento člen uveden jako poslední z členů daného řádu. Podobně lze shodu ověřit u členu Vzdel.

```
> Anova(lm(bmi~Vzdel*Kurak,data=IchsN),type="III")
Anova Table (Type III tests)
Response: bmi
            Sum Sq  Df   F value  Pr(>F)
(Intercept)  39879   1 4925.8226 < 2e-16
Vzdel           28   2    1.7436 0.17818
Kurak           54   1    6.6237 0.01096
Vzdel:Kurak     12   2    0.7516 0.47324
Residuals     1303 161
```

Máme-li ověřit, odkud pochází součet čtverců pro člen Kurak, musíme si trochu pomoci. Vytvoříme matici $\mathbf{X}$ našeho modelu a necháme spočítat reziduální součet čtverců. Potom vyloučíme sloupec odpovídající faktoru Kurak (faktoru Vzdel) a zhodnotíme příslušný vzrůst reziduálního součtu čtverců. Při přípravě matice $\mathbf{X}$ jsme přihlédli k tomu, že absolutní člen je do modelu vkládán automaticky, proto jsme jej do této matice nevložili. Funkce deviance() podá v případě lineárního modelu reziduální součet čtverců.

```
> attach(IchsN)
> X0 <- model.matrix(~Vzdel*Kurak-1)
> X0[1:5,]
  Vzdel1 Vzdel2 Vzdel3 KurakTRUE Vzdel.L:KurakTRUE Vzdel.Q:KurakTRUE
1      1      0      0         0      0.000000e+00         0.0000000
2      1      0      0         1     -7.071068e-01         0.4082483
3      0      0      1         0      0.000000e+00         0.0000000
4      1      0      0         1     -7.071068e-01         0.4082483
5      0      1      0         1      4.350720e-18        -0.8164966
> print(RSS0 <- deviance(lm(bmi~X0)))
1303.442
> print(RSS1 <- deviance(lm(bmi~X0[,-4])))
1357.067
> print(f0<-anova(lm(bmi~X0))[2,1])
161
> print(F <- (RSS1-RSS0)/(RSS0/f0))
6.623721

> print(RSS1 <- deviance(lm(bmi~X0[,-(1:3)])))
1331.674
> print(F <- (RSS1-RSS0)/2/(RSS0/f0))
1.743569
```

○

Příklad 6.7 (Howells) W. W. Howels dal na internetu k disposici rozsáhlý soubor dat zjištěných na exhumovaných lebkách z různých populací Země (Howells (1996)). Nás zajímá jen část údajů, která se týká tří míst (rakouský Berg, Austrálie a Burjati na Sibiři). Experti určili u každé exhumované nejen pohlaví, ale také řadu rozměrů, z nichž si vybereme dva: GOL (Glabell-Occipital Length, tj. největší délka mozkovny) a ACA (Occipital Angle, tj. týlní úhel) (viz též Zvára (1998)). Snadno se přesvědčíme, že tabulka analýzy roztylu (typu I) dá pro obě možná pořadí faktorů identické výsledky. Není to náhoda?

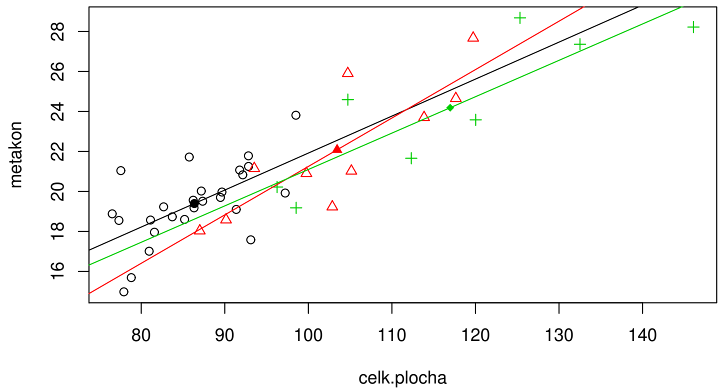
```
> anova(lm(oca~Sex*Popul,data=Howells))
Analysis of Variance Table

Response: oca
           Df Sum Sq Mean Sq F value  Pr(>F)
Sex         1   91.3    91.3  3.6888 0.05599 .
Popul       2  150.9    75.5  3.0497 0.04926 *
Sex:Popul   2  191.6    95.8  3.8722 0.02216 *
Residuals 234 5789.6    24.7
> anova(lm(oca~Popul*Sex,data=Howells))
Analysis of Variance Table

Response: oca
           Df Sum Sq Mean Sq F value  Pr(>F)
Popul       2  150.9    75.5  3.0497 0.04926 *
Sex         1   91.3    91.3  3.6888 0.05599 .
Popul:Sex   2  191.6    95.8  3.8722 0.02216 *
Residuals 234 5789.6    24.7
> anova(lm(gol~Sex*Popul,data=Howells))
Analysis of Variance Table

Response: gol
           Df Sum Sq Mean Sq  F value Pr(>F)
Sex         1 5170.8  5170.8 128.5753 <2e-16 ***
Popul       2 5242.1  2621.1  65.1743 <2e-16 ***
Sex:Popul   2    9.6     4.8   0.1198 0.8872
Residuals 234 9410.6    40.2
> anova(lm(gol~Popul*Sex,data=Howells))
Analysis of Variance Table

Response: gol
           Df Sum Sq Mean Sq  F value Pr(>F)
Popul       2 5242.1  2621.1  65.1743 <2e-16 ***
Sex         1 5170.8  5170.8 128.5753 <2e-16 ***
Popul:Sex   2    9.6     4.8   0.1198 0.8872
Residuals 234 9410.6    40.2
```

Tento příklad nám umožní ukázat význam a interpretaci interakcí.

○

6.3. Analýza kovariance

Příklad 6.8 (Protoconid) Budeme porovnávat velikost plochy jistého hrbolku dolní sedmičky (Protoconid) ve třech skupinách archeologických nálezů, které odpovídají různým vývojovým stádiím, zde značeným symboly *r, m, s*. Použitá data jsou uvedena v C.1. Kdybychom k porovnání použili analýzu rozptylu jednoduchého třídění, dostali bychom statistiku $F = 11{,}6935$, která je při 2 a 41 stupních volnosti významná ($p < 0{,}0001$). Lze ale namítat, že rozdíly velikosti sledovaného hrbolku mezi skupinami mohou být způsobeny přímo rozdíly velikosti zubů.



Obrázek 6.2: Velikosti ploch dolních stoliček s nezávisle odhadovanými přímkami

Proveďme tedy nejprve adjustaci na velikost zubu a teprve potom analýzu rozptylu. V modelu analýzy kovariance (2.16)

$$Y_{it} \sim \mathsf{N}\big(\mu + \alpha_i + \beta x_{it} + e_{it}, \sigma^2\big)$$

dostaneme odhady (v závorkách uvedeme střední chyby) $b = 0{,}1988$ $(0{,}0257)$, při reparametrizaci podle (5.6) s $j = 1$ dále $m = 2{,}2243$ $(2{,}2413)$, $a_1 = 0$, $a_2 =$

$-0{,}7089$ (0,7612), $a_3 = -1{,}2965$ (1,0368) s reziduálním součtem čtverců $RSS =$ 111,7786 při 40 stupních volnosti. Daty jsme proložili rovnoběžné přímky se směrnicí 0,1988, pro skupinu m (resp. s) jsou proti skupině r posunuty o $-0{,}7089$ (resp. $-1{,}2965$) jednotek.

Chceme-li rozhodnout, zda se skupiny po provedené adjustaci na velikost zubu ještě liší, testujeme hypotézu, že přímky jsou totožné. V tomto podmodelu daném požadavky $\alpha_1 = \alpha_2 = \alpha_3 = 0$ dostaneme odhady $b = 0{,}1733$ (0,0161) a $m =$ 4,2714 (1,5596) a reziduální součet čtverců $RSS_0 = 116{,}3392$. Test podmodelu vede podle (3.10) k testové statistice

$$F = \frac{116{,}3392 - 111{,}7786}{111{,}7786}\frac{40}{2} = 0{,}8160,$$

což při 2 a 40 stupních volnosti dá $p = 0{,}4494$. Po adjustaci na velikost zubu nelze mezi skupinami prokázat rozdíl.

Na obrázku 6.2 je patrné, proč po adjustaci vůči celkové ploše může vyjít rozdíl mezi skupinami nevýznamný. Stačí si představit, že všechna tři těžiště (tučné symboly) posuneme ve směru společné regresní přímky tak, aby měla stejnou x-ovou souřadnici.

```
> anova(a<-lm(metakon~Skupina))
Analysis of Variance Table

Response: metakon
          Df  Sum Sq Mean Sq F value    Pr(>F)
Skupina    2 159.347  79.673  11.694 9.588e-05
Residuals 41 279.353   6.813

> summary(aa<-lm(metakon~celk.plocha+Skupina))

Call:
lm(formula = metakon ~ celk.plocha + Skupina)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1574 -1.2329 -0.1113  0.6853  3.5631

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.22432    2.24130   0.992    0.327
celk.plocha  0.19883    0.02568   7.744 1.76e-09
Skupina2    -0.70892    0.76124  -0.931    0.357
Skupina3    -1.29649    1.03684  -1.250    0.218

Residual standard error: 1.672 on 40 degrees of freedom
Multiple R-Squared: 0.7452,     Adjusted R-squared: 0.7261
F-statistic:    39 on 3 and 40 DF,  p-value: 5.949e-012
```

```
> anova(aa)
Analysis of Variance Table

Response: metakon
            Df Sum Sq Mean Sq F value    Pr(>F)
celk.plocha  1 322.36  322.36 115.357 2.366e-13
Skupina      2   4.56    2.28   0.816    0.4494
Residuals   40 111.78    2.79

> anova(aaa<-lm(metakon~celk.plocha))
Analysis of Variance Table

Response: metakon
            Df Sum Sq Mean Sq F value    Pr(>F)
celk.plocha  1 322.36  322.36  116.38 1.111e-13
Residuals   42 116.34    2.77

> anova(aaa,aa)
Analysis of Variance Table

Model 1: metakon ~ celk.plocha
Model 2: metakon ~ celk.plocha + Skupina
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1     42 116.339
2     40 111.779  2     4.561 0.816 0.4494
```

Uveďme ještě postup, jak jsme nakreslili obrázek 6.2.

```
> plot(metakon~celk.plocha,pch=skupina,col=Skupina)
> points(tapply(celk.plocha,Skupina,mean),tapply(metakon,Skupina,mean),
+        col=1:3,pch=15+1:3)
> abline(lm(metakon~celk.plocha,subset=Skupina==1),col=1)
> abline(lm(metakon~celk.plocha,subset=Skupina==2),col=2)
> abline(lm(metakon~celk.plocha,subset=Skupina==3),col=3)
```

○

Poznámka Poslední příklad poskytuje pěknou ukázku toho, že statistiky uvedené v tabulce analýzy rozptylu, které poskytuje R, závisí na pořadí faktorů. Doporučuji porovnat součet čtverců vysvětlený příslušností ke třem skupinám se stejnou statistikou uvedenou v obou tabulkách analýzy rozptylu v posledním příkladu (modely `a` a `aa`).

```
> anova(lm(metakon~Skupina+celk.plocha,data=Metakon))
Analysis of Variance Table

Response: metakon
            Df  Sum Sq Mean Sq F value    Pr(>F)
Skupina      2 159.347  79.673  28.511 2.013e-08
celk.plocha  1 167.575 167.575  59.967 1.764e-09
Residuals   40 111.779   2.794
```

7. Následky nesplnění předpokladů

V lineárním modelu jsme předpokládali, že známe prostor možných středních hodnot, že všechna pozorování mají stejný rozptyl, že jsou nekorelovaná (resp. nezávislá) a že mají normální rozdělení. Nyní se pokusíme popsat následky, které má nesplnění některého z uvedených předpokladů.

7.1. Prostor středních hodnot

Předpokládejme, že platí

$$\mathbf{Y} = \mathsf{X}\boldsymbol{\beta} + \mathsf{Z}\gamma + \mathsf{e}, \qquad \mathsf{e} \sim (\mathbf{0}, \sigma^2\mathsf{I}), \tag{7.1}$$

přestože my nadále pracujeme s modelem $\mathbf{Y} \sim (\mathsf{X}\boldsymbol{\beta}, \sigma^2\mathsf{I})$.

Označme $\mathsf{G} = (\mathsf{X}, \mathsf{Z})$ a $\boldsymbol{\delta} = (\boldsymbol{\beta}', \gamma')'$ a veškeré statistiky vztažené k modelu $\mathbf{Y} \sim (\mathsf{G}\boldsymbol{\delta}, \sigma^2\mathsf{I})$ označíme dolním indexem g. Běžný odhad vektoru $\mathsf{E}\,\mathbf{Y}$ je tedy

$$\hat{\mathbf{Y}}_g = \mathsf{G}(\mathsf{G}'\mathsf{G})^-\mathsf{G}'\mathbf{Y}, \tag{7.2}$$

což je, jak víme např. z (3.13), průmět $\mathbf{Y}$ do $\mathcal{M}(\mathsf{X}, \mathsf{Z}) = \mathcal{M}(\mathsf{X}, \mathsf{MZ})$. S použitím druhého vyjádření pak dostaneme

$$\begin{aligned}
\hat{\mathbf{Y}}_g &= (\mathsf{X}, \mathsf{MZ}) \begin{pmatrix} \mathsf{X}'\mathsf{X} & \mathsf{O} \\ \mathsf{O} & \mathsf{Z}'\mathsf{MZ} \end{pmatrix}^- \begin{pmatrix} \mathsf{X}' \\ \mathsf{Z}'\mathsf{M} \end{pmatrix} \mathbf{Y} \\
&= \mathsf{X}(\mathsf{X}'\mathsf{X})^-\mathsf{X}'\mathbf{Y} + \mathsf{MZ}(\mathsf{Z}'\mathsf{MZ})^-\mathsf{Z}'\mathsf{M}\mathbf{Y} \\
&= \hat{\mathbf{Y}} + \mathsf{MZ}(\mathsf{Z}'\mathsf{MZ})^-\mathsf{Z}'\mathsf{u} \qquad (7.3) \\
&= \mathsf{X}\mathsf{b}_g + \mathsf{Z}\mathsf{c}_g, \qquad (7.4)
\end{aligned}$$

kde b_g a c_g jsou obecně nějaká řešení příslušné normální rovnice.

Když přepíšeme (7.4) tak, aby bylo patrné jakou lineární kombinací sloupců matic $\mathbf{X}, \mathbf{Z}$ je vektor $\hat{\mathbf{Y}}_g$ (co mohou být vektory $\mathbf{b}_g, \mathbf{c}_g$), dostaneme po úpravě (vyjádříme $\mathbf{M}$ pomocí $\mathbf{X}$)

$$\hat{\mathbf{Y}}_g = \mathbf{X}(\mathbf{b} - (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z}\mathbf{c}_g) + \mathbf{Z}\mathbf{c}_g, \tag{7.5}$$

když jsme označili

$$\mathbf{c}_g = (\mathbf{Z}'\mathbf{M}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{u}. \tag{7.6}$$

Můžeme tedy psát

$$\mathbf{b}_g = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z}\mathbf{c}_g, \tag{7.7}$$

odkud je zřetelný zejména vztah mezi $\mathbf{b}$ a $\mathbf{b}_g$.

Z (7.3) plyne, že rozdíl reziduálních součtů čtverců mezi uvažovaným modelem $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ a skutečně platným modelem $\mathbf{Y} \sim (\mathbf{G}\boldsymbol{\delta}, \sigma^2\mathbf{I})$ je

$$\begin{aligned} RSS - RSS_g &= ||\mathbf{M}\mathbf{Z}(\mathbf{Z}'\mathbf{M}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{u}||^2 \\ &= ||\mathbf{M}\mathbf{Z}\mathbf{c}_g||^2. \end{aligned} \tag{7.8}$$

Porovnejme ještě střední hodnoty obou reziduálních součtů čtverců. Protože platí model (7.1), je zřejmě $\mathsf{E}\,RSS_g = (n - \mathsf{h}(\mathbf{X}, \mathbf{Z}))\sigma^2$. Jinak to dopadne u reziduálního součtu čtverců RSS z (nesprávně) předpokládaného modelu. Postupnými úpravami dostaneme

$$\mathsf{E}\,RSS = \mathsf{E}\,||\mathbf{M}\mathbf{Y}||^2 = \mathsf{E}\,||\mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e})||^2 = \mathsf{E}\,||\mathbf{M}\mathbf{Z}\boldsymbol{\gamma} + \mathbf{M}\mathbf{e}||^2,$$

tedy

$$\begin{aligned} \mathsf{E}\,RSS &= ||\mathbf{M}\mathbf{Z}\boldsymbol{\gamma}||^2 + \mathsf{E}\,||\mathbf{M}\mathbf{e}||^2 \\ &= ||\mathbf{M}\mathbf{Z}\boldsymbol{\gamma}||^2 + (n - \mathsf{h}(\mathbf{X}))\sigma^2. \end{aligned} \tag{7.9}$$

Vraťme se k odhadu $\hat{\mathbf{Y}}$. Jeho střední hodnota je rovna

$$\mathsf{E}\,\hat{\mathbf{Y}} = \mathbf{H}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{H}\mathbf{Z}\boldsymbol{\gamma}.$$

Obecně tedy není nestranným odhadem pro $\mathsf{E}\,\mathbf{Y}$, má vychýlení

$$\mathsf{bias}\,\hat{\mathbf{Y}} = \mathsf{E}\,\hat{\mathbf{Y}} - \mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{H}\mathbf{Z}\boldsymbol{\gamma} - (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}) \;=\; -\mathbf{M}\mathbf{Z}\boldsymbol{\gamma}. \tag{7.10}$$

Střední čtvercovou chybu $\hat{\mathbf{Y}}$ jako odhadu pro $\mathsf{E}\,\mathbf{Y}$ lze psát

$$\mathsf{MSE}\,\hat{\mathbf{Y}} = \mathsf{var}\,\hat{\mathbf{Y}} + (\mathsf{bias}\,\hat{\mathbf{Y}})(\mathsf{bias}\,\hat{\mathbf{Y}})' = \sigma^2\mathbf{H} + \mathbf{M}\mathbf{Z}\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}.$$

Protože $\hat{\mathbf{Y}}_g$ je nestranným odhadem $\mathsf{E}\,\mathbf{Y}$, platí $\mathsf{MSE}\,\hat{\mathbf{Y}}_g = \mathsf{var}\,\hat{\mathbf{Y}}_g$, což lze upravit podobně jako při výpočtu $\hat{\mathbf{Y}}_g$ na

$$\begin{aligned}\mathsf{var}\,\hat{\mathbf{Y}}_g &= \sigma^2(\mathbf{X},\mathbf{MZ})\begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{O}\\ \mathbf{O} & \mathbf{Z}'\mathbf{MZ}\end{pmatrix}^{-}\begin{pmatrix}\mathbf{X}'\\ \mathbf{Z}'\mathbf{M}\end{pmatrix}\\ &= \sigma^2\left(\mathbf{H}+\mathbf{MZ}(\mathbf{Z}'\mathbf{MZ})^{-}\mathbf{Z}'\mathbf{M}\right). \qquad (7.11)\end{aligned}$$

Shrňme vlastnosti odhadů klasického modelu.

Věta 7.1. (Vlastnosti odhadů, platí-li širší model) Nechť platí $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\gamma},\sigma^2\mathbf{I})$. Pro statistiky odvozené z modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I})$ platí

$$\mathsf{bias}\,\hat{\mathbf{Y}} = -\mathbf{MZ}\boldsymbol{\gamma}, \qquad (7.12)$$

$$\mathsf{bias}\,S^2 = \frac{||\mathbf{MZ}\boldsymbol{\gamma}||^2}{n-\mathsf{h}(\mathbf{X})},. \qquad (7.13)$$

Porovnejme $\hat{\mathbf{Y}}$ a $\mathbf{Y}$ jako odhady pro $\mathsf{E}\,\mathbf{Y}$:

$$\mathsf{MSE}\,\hat{\mathbf{Y}}_g - \mathsf{MSE}\,\hat{\mathbf{Y}} = \sigma^2\left(\mathbf{MZ}(\mathbf{Z}'\mathbf{MZ})^{-}\mathbf{Z}'\mathbf{M} - \mathbf{MZ}\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}/\sigma^2\right).$$

Nyní stačí použít tvrzení věty A.7 pro $\mathbf{A} = \mathbf{MZ}$ a $\mathbf{c} = \boldsymbol{\gamma}/\sigma$, abychom zjistili, že rozdíl středních čtvercových chyb dá pozitivně semidefinitní matici, právě když je $||\mathbf{Ac}||^2 = ||\mathbf{MZ}\boldsymbol{\gamma}/\sigma||^2 \leq 1$. Došli jsme tak k tvrzení následující věty.

Věta 7.2. (Kdy je vychýlení malé) Nechť platí $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\gamma},\sigma^2\mathbf{I})$. Pro $\hat{\mathbf{Y}}_g$ z tohoto modelu a pro $\hat{\mathbf{Y}}$ z modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I})$ platí ekvivalence

$$\mathsf{MSE}\,\hat{\mathbf{Y}}_g \geq \mathsf{MSE}\,\hat{\mathbf{Y}} \iff ||\mathsf{bias}\,\hat{\mathbf{Y}}||^2 \leq \sigma^2. \qquad (7.14)$$

Při předpovědi budoucího pozorování tedy je výhodnější použít menší model, když je vychýlení způsobené touto volbou dostatečně malé.

Věta 7.3. (Důsledek) Nechť $\mathbf{b}$ je libovolné řešení normální rovnice $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$, nechť $\theta = \mathbf{p}'\boldsymbol{\beta}+\mathbf{s}'\boldsymbol{\gamma}$ je odhadnutelný parametr v modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\gamma},\sigma^2\mathbf{I})$. Potom je parametr $\tau = \mathbf{p}'\boldsymbol{\beta}$ odhadnutelný také v modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I})$ a platí

$$\mathsf{MSE}\,\hat{\theta} \geq \mathsf{MSE}\,\hat{\tau} \iff ||\mathbf{MZ}\boldsymbol{\gamma}||^2 \leq \sigma^2.$$

D ů k a z: Především je třeba dokázat, že τ je odhadnutelný parametr. Odhadnutelnost θ je podle věty 2.3 ekvivalentní s existencí vektoru $\mathbf{q} \in \mathbb{R}^n$, pro který platí $\mathbf{q}'(\mathbf{X},\mathbf{Z}) = (\mathbf{p}',\mathbf{s}')$. Speciálně to tedy znamená existenci $\mathbf{q}$, pro který platí $\mathbf{q}'\mathbf{X} = \mathbf{p}'$, tedy podle téže věty odhadnutelnost parametru τ v „menším“ modelu. Porovnání středních čtvercových chyb plyne z použití tvrzení věty 7.2, když se vezme ohled na $\mathsf{MSE}\,\hat{\tau} = \mathbf{q}'(\mathsf{MSE}\,\hat{\mathbf{Y}})\mathbf{q}$ a $\mathsf{MSE}\,\hat{\theta} = \mathbf{q}'(\mathsf{MSE}\,\hat{\mathbf{Y}}_g)\mathbf{q}$. □

Poznámka Totéž dostaneme, pokud v modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\gamma},\sigma^2\mathbf{I})$ je odhadnutelný parametr $\theta^* = \mathbf{p}'\boldsymbol{\beta}+\mathbf{0}'\boldsymbol{\gamma} = \mathbf{p}'\boldsymbol{\beta}$. Něco jiného vyjde, když platí „menší“ model, a my použijeme model větší, i když jen k odhadu odhadnutelné funkce $\mathbf{p}'\boldsymbol{\beta}$. Pak jsou oba odhady $\hat{\tau}^*_g = \mathbf{q}'\hat{\mathbf{Y}}_g$ a $\hat{\tau}^* = \mathbf{q}'\hat{\mathbf{Y}}$ nestranné. O vztahu obou středních čtvercových chyb pak rozhoduje porovnání rozptylů. Z Gaussovy-Markovovy věty plyne, že odhad $\hat{\tau}^*$ je nejlepší, takže $\hat{\tau}^*_g$ nemůže mít rozptyl menší. Použijeme vyjádření (7.11) pro rozptyl odhadu $\hat{\tau}^*_g$

$$\begin{aligned}\operatorname{var}\hat{\tau}^*_g &= \mathbf{q}'(\operatorname{var}\hat{\mathbf{Y}}_g)\mathbf{q}\\ &= \operatorname{var}\hat{\tau}^* + \mathbf{q}'\mathbf{MZ}(\mathbf{Z}'\mathbf{MZ})^-\mathbf{Z}'\mathbf{Mq}\end{aligned}$$

což ukazuje, nakolik je odhad ve zbytečně bohatém modelu méně přesný.

7.2. Případ s úplnou hodností

Předpokládejme nyní, že matice $\mathbf{G}$ má lineárně nezávislé sloupce. Odtud plyne, že také matice $\mathbf{X}$ a $\mathbf{Z}$ mají lineárně nezávislé sloupce, takže $\mathbf{X}'\mathbf{X}$ a $\mathbf{Z}'\mathbf{Z}$ jsou regulární. Regulární musí být také matice $\mathbf{Z}'\mathbf{MZ}$, neboť prostor $\mathcal{M}(\mathbf{MZ})$ musí mít stejnou dimenzi jako prostor $\mathcal{M}(\mathbf{Z})$. Můžeme tedy v tomto případě psát (viz (7.7), (7.6))

$$\mathbf{b}_g = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{c}_g, \tag{7.15}$$

$$\mathbf{c}_g = (\mathbf{Z}'\mathbf{MZ})^{-1}\mathbf{Z}'\mathbf{u}. \tag{7.16}$$

Ze vztahu (7.15) můžeme vyjádřit střední hodnotu odhadu $\mathbf{b}$:

$$\mathsf{E}\,\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\gamma}. \tag{7.17}$$

Invertováním matice rozdělené na pole (viz například (Anděl, 1978, kap. IV, věta 9)) dostaneme

$$\begin{aligned}\operatorname{var}\begin{pmatrix}\mathbf{b}_g\\ \mathbf{c}_g\end{pmatrix} &= \sigma^2\begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\\ \mathbf{X}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z}\end{pmatrix}^{-1}\\ &= \begin{pmatrix}\sigma^2(\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1} & *\\ * & \sigma^2(\mathbf{Z}'\mathbf{MZ})^{-1}\end{pmatrix},\end{aligned} \tag{7.18}$$

když jsme hvězdičkou označili matice kovariancí, jejichž explicitní vyjádření nyní nepotřebujeme.

Závěr Pro model $\mathbf{Y} \sim (\mathbf{G}\boldsymbol{\delta}, \sigma^2\mathbf{I})$ s úplnou hodností platí:

a) Je-li $\mathbf{X}'\mathbf{Z} = \mathbf{O}$, pak platí $\mathbf{b}_g = \mathbf{b}$ (se všemi důsledky).

b) Je-li $\mathbf{X}'\mathbf{Z} \neq \mathbf{O}$, pak je odhad $\mathbf{b}$ vychýleným odhadem $\boldsymbol{\beta}$, platí však
$$\operatorname{var} \mathbf{b}_g > \operatorname{var} \mathbf{b}. \tag{7.19}$$
Tvrzení o variančních maticích plyne z toho, že je
$$\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} < \mathbf{X}'\mathbf{X},$$
pak stačí použít větu A.5 z appendixu o srovnání kvadratických forem.

Příklad 7.1 (dva regresory) Nechť platí regresní model se dvěma nezávisle proměnnými
$$\begin{aligned} y &= \beta_0 + \beta x + \gamma z \\ &= \beta_0^* + \beta(x - \bar{x}) + \gamma(z - \bar{z}) \end{aligned}$$
kdežto my uvažujeme pouze závislost na nezávisle proměnné x. V takovém případě používáme odhad parametru β_1 tvaru
$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{T_{yx}}{T_{xx}}$$
s rozptylem
$$\operatorname{var} b = \sigma^2 / T_{xx}.$$
Odhadem parametru β_0^* je $\bar{Y}$ s rozptylem σ^2/n.

Ve skutečnosti jsme měli použít odhad založený na
$$\begin{pmatrix} b_g \\ c_g \end{pmatrix} = \begin{pmatrix} T_{xx} & T_{xz} \\ T_{zx} & T_{zz} \end{pmatrix}^{-1} \begin{pmatrix} T_{xy} \\ T_{zy} \end{pmatrix},$$
což po úpravě vede k odhadu
$$\begin{aligned} b_g &= \frac{T_{zz}T_{xy} - T_{xz}T_{zy}}{T_{xx}T_{zz} - T_{xz}^2} \\ &= \frac{b - (T_{xz}/T_{xx})(T_{zy}/T_{zz})}{1 - r_{xz}^2}, \end{aligned}$$
kde r_{xz}^2 je výběrový korelační koeficient mezi veličinami x, z. Rozptyl odhadové statistiky b_g můžeme zapsat jako
$$\begin{aligned} \operatorname{var} b_g &= \sigma^2 \frac{T_{zz}}{T_{xx}T_{zz} - T_{xz}^2} \\ &= \frac{\sigma^2}{T_{xx}} \frac{1}{1 - r_{xz}^2} = \frac{1}{1 - r_{xz}^2} \operatorname{var} b. \end{aligned}$$
Odtud je vidět zřetelně, že rozptyl b_g nemůže být nikdy menší, než rozptyl b. Naopak, při podobně se chovajících veličinách x a z bude rozptyl b_g mnohem větší.

Všimněme si také, že i ve správném modelu je odhad absolutního členu stejný, je to $\bar{Y}$.

Ze vztahu (7.17) o střední hodnotě $\mathbf{b}$ zde speciálně dostaneme vychýlení odhadu b
$$\operatorname{bias} b = \frac{T_{xz}}{T_{xx}} \gamma = \sqrt{\frac{T_{zz}}{T_{xx}}} r_{xz} \gamma.$$
○

7.3. Varianční matice

Předpokládejme, že ve skutečnosti platí

$$\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1}), \tag{7.20}$$

kde $\mathbf{W} > 0$ je známá pozitivně definitní matice. Vhodné statistiky jsme popsali v oddílu 2.8. Zde se pokusíme zjistit následky toho, že vycházíme z předpokladu

$$\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}). \tag{7.21}$$

Naším hlavním cílem je zjistit, kdy je takto získaný běžný odhad $\hat{\mathbf{Y}}$ totožný s optimálním odhadem $\hat{\mathbf{Y}}_W$.

Odhad $\hat{\mathbf{Y}}$ je i za platnosti modelu (7.20) nestranným odhadem $\mathsf{E}\,\mathbf{Y}$:

$$\mathsf{E}\,\hat{\mathbf{Y}} = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}.$$

Varianční matici odhadu $\hat{\mathbf{Y}}$ dostaneme také snadno:

$$\operatorname{var}\hat{\mathbf{Y}} = \operatorname{var}\mathbf{H}\mathbf{Y} = \mathbf{H}\sigma^2\mathbf{W}^{-1}\mathbf{H} = \sigma^2\mathbf{H}\mathbf{W}^{-1}\mathbf{H}.$$

Zaveďme pracovní označení. Nechť $\mathbf{P} = (\mathbf{Q}, \mathbf{N})$ je taková ortonormální matice, pro kterou platí $\mathcal{M}(\mathbf{X}) = \mathcal{M}(\mathbf{Q})$. Označíme-li

$$\mathbf{T}_{QQ} = \mathbf{Q}'\mathbf{W}\mathbf{Q}, \tag{7.22}$$
$$\mathbf{T}_{QN} = \mathbf{Q}'\mathbf{W}\mathbf{N}, \tag{7.23}$$
$$\mathbf{T}_{NN} = \mathbf{N}'\mathbf{W}\mathbf{N}, \tag{7.24}$$

můžeme matici $\mathbf{W}$ zapsat jako $\mathbf{W} = \mathbf{P}\mathbf{P}'\mathbf{W}\mathbf{P}\mathbf{P}'$, tedy

$$\mathbf{W} = (\mathbf{Q}, \mathbf{N}) \begin{pmatrix} \mathbf{T}_{QQ} & \mathbf{T}_{QN} \\ \mathbf{T}'_{QN} & \mathbf{T}_{NN} \end{pmatrix} \begin{pmatrix} \mathbf{Q}' \\ \mathbf{N}' \end{pmatrix} \tag{7.25}$$
$$= \mathbf{Q}\mathbf{T}_{QQ}\mathbf{Q}' + \mathbf{Q}\mathbf{T}_{QN}\mathbf{N}' + \mathbf{N}\mathbf{T}'_{QN}\mathbf{Q}' + \mathbf{N}\mathbf{T}_{NN}\mathbf{N}'. \tag{7.26}$$

Podobně lze vyjádřit matici $\mathbf{W}^{-1}$ jako

$$\mathbf{W}^{-1} = \mathbf{Q}\mathbf{T}^{QQ}\mathbf{Q}' + \mathbf{Q}\mathbf{T}^{QN}\mathbf{N}' + \mathbf{N}\mathbf{T}'^{QN}\mathbf{Q}' + \mathbf{N}\mathbf{T}^{NN}\mathbf{N}'.$$

7.3.1. Totožné odhady

Zajímá nás, kdy jsou odhady $\hat{\mathbf{Y}}_W$ a $\hat{\mathbf{Y}}$ totožné. Je to právě tehdy, když jsou obě projekční matice totožné, tedy když platí (viz též větu 2.7)

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}. \tag{7.27}$$

K maticím $\mathbf{X}$ a $\mathbf{Q}$ existuje matice $\mathbf{C}$ typu $r \times k$ taková, že je $\mathbf{X} = \mathbf{QC}$ (jsou to souřadnice jednotlivých sloupců matice $\mathbf{X}$ v bázi $\mathbf{Q}$). Protože řádky matice $\mathbf{C}$ musí být lineárně nezávislé, existuje její pravá inverzní matice $\mathbf{C}^-$. Když použijeme vyjádření $\mathbf{X} = \mathbf{QC}$, dostaneme s použitím (7.26) a vlastností matice $\mathbf{P}$

$$\mathbf{X}'\mathbf{WX} = \mathbf{C}'\mathbf{Q}'\mathbf{WQC} = \mathbf{C}'\mathbf{T}_{QQ}\mathbf{C}.$$

Odtud je snadno matice $\mathbf{C}^-\mathbf{T}_{QQ}^{-1}\mathbf{C}^{-'}$ nějakou pseudoinverzní maticí matice $\mathbf{X}'\mathbf{WX}$. Dosadíme-li do (7.27), dostaneme s využitím (7.25)

$$\begin{aligned}
\mathbf{QQ}' &= \mathbf{QC}(\mathbf{C}^-\mathbf{T}_{QQ}^{'-1}\mathbf{C}^{-'})\mathbf{C}'\mathbf{Q}'\mathbf{W} \\
&= \mathbf{QT}_{QQ}^{-1}(\mathbf{T}_{QQ}\mathbf{Q}' + \mathbf{T}_{QN}\mathbf{N}') \\
&= \mathbf{QQ}' + \mathbf{QT}_{QQ}^{-1}\mathbf{T}_{QN}\mathbf{N}'.
\end{aligned}$$

Uvážíme-li že matice $\mathbf{Q}$ a $\mathbf{N}$ mají lineárně nezávislé sloupce, došli jsme tedy k tvrzení následující věty:

Věta 7.4. Odhady $\hat{\mathbf{Y}}_W$ a $\hat{\mathbf{Y}}$ jsou totožné, právě když platí

$$\mathbf{O} = \mathbf{T}_{QN} = \mathbf{Q}'\mathbf{WN}, \tag{7.28}$$

což je ekvivalentní s podmínkou

$$\mathbf{O} = \mathbf{T}^{QN} = \mathbf{Q}'\mathbf{W}^{-1}\mathbf{N}. \tag{7.29}$$

D ů k a z: K dokončení důkazu stačí ukázat ekvivalenci obou podmínek. Stačí si však uvědomit, že inverzní matice k blokově diagonální matici je opět blokově diagonální. □

Totožnost obou odhadů je tedy zajištěna, když ortogonální skupiny sloupců matic $\mathbf{Q}, \mathbf{N}$ jsou vůči sobě ortogonální také v prostoru deformovaném maticí $\mathbf{W}$.

7.3.2. Odhad rozptylu

Jsou-li splněny klasické předpoklady, je S^2 nestranným odhadem rozptylu σ^2. Důkaz byl založen na tom, že v klasickém lineárním modelu platí $\mathsf{E}\,RSS = (n-r)\sigma^2$.

Zachováme-li označení z 2. kapitoly, můžeme psát

$$RSS = \|\mathbf{u}\|^2 = \|\mathbf{NN}'\mathbf{Y}\|^2 = \|\mathbf{N}'\mathbf{Y}\|^2,$$

když jsme použili ortonormalitu sloupců matice $\mathbf{N}$. Má-li náhodný vektor $\mathbf{Y}$ varianční matici $\sigma^2\mathbf{W}^{-1}$, má náhodný vektor $\mathbf{N}'\mathbf{Y}$ nulovou střední hodnotu a varianční matici

$$\begin{aligned}
\operatorname{var}\mathbf{N}'\mathbf{Y} &= \sigma^2\mathbf{N}'\mathbf{W}^{-1}\mathbf{N} \\
&= \sigma^2\mathbf{T}^{NN}
\end{aligned}$$

Došli jsme tedy k tvrzení

Věta 7.5. V modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1})$ je statistika S^2 nestranným odhadem rozptylu σ^2 právě, když platí $\operatorname{tr}\mathbf{N}'\mathbf{W}^{-1}\mathbf{N} = n - r$.

Žádáme tedy, aby varianční matice vektoru $\mathbf{N}'\mathbf{Y}$ měla stejnou stopu, ať už platí model $\mathbf{Y} \sim \big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1}\big)$ nebo model $\mathbf{Y} \sim \big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\big)$.

7.3.3. Test podmodelu

Tentokrát musíme předpokládat normální rozdělení $\mathbf{Y} \sim \mathsf{N}\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1}\big)$. Požadavek $\mathsf{E}\,\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0$ určí podmodel uvažovaného modelu, když platí $\mathcal{M}(\mathbf{X}_0) \subset \mathcal{M}(\mathbf{X})$ a současně $0 < \mathsf{h}(\mathbf{X}_0) = r_0 < \mathsf{h}(\mathbf{X}) = r$.

O platnosti podmodelu se rozhoduje pomocí F statistiky z věty 7.6, tvrzení d). V porovnání se zmiňovanou větou tentokrát má náhodný vektor $\mathbf{Y}$ jinou varianční matici. Tvrzení však zůstane v platnosti, pokud náhodný vektor

$$\begin{pmatrix} \mathbf{Q}_1' \\ \mathbf{N}' \end{pmatrix} \mathbf{Y}$$

má rozdělení $\mathsf{N}\big(\mathbf{0}, \sigma^2\mathbf{I}\big)$. Je tedy nutné a stačí, aby bylo současně

$$\mathbf{Q}_1'\mathbf{W}^{-1}\mathbf{Q}_1 = \mathbf{I} \tag{7.30}$$

$$\mathbf{Q}_1'\mathbf{W}^{-1}\mathbf{N} = \mathbf{O} \tag{7.31}$$

$$\mathbf{N}'\mathbf{W}^{-1}\mathbf{N} = \mathbf{I}. \tag{7.32}$$

Věta 7.6. Když existuje matice $\mathbf{D}$ tak, že platí

$$\mathbf{W}^{-1} = \mathbf{I} + \mathbf{X}_0\mathbf{D}' + \mathbf{D}\mathbf{X}_0', \tag{7.33}$$

a platí podmodel, pak statistika F z (3.10) má rozdělení $\mathsf{F}(r - r_0, n - r)$.

D ů k a z: Je třeba dokázat, že platí vztahy (7.30)–(7.32). Toho se snadno dosáhne, když se využije vztahů $\mathbf{X}_0'\mathbf{N} = \mathbf{O}$ a $\mathbf{Q}_1'\mathbf{X}_0 = \mathbf{O}$. □

7.3.4. Příklady

Zde uvedeme dva modely, které vedou k speciálním maticím $\mathbf{W}$.

Příklad 7.2 (náhodné bloky) Rozšiřme úlohu, která vedla na jednoduché třídění. Opět chceme porovnat I nějakých ošetření. Abychom co možná nejvíce zmenšili vliv variability pokusných objektů (zvířat, osob, políček), sestavíme nejprve J pokud možno homogenních skupin – *bloků* po I prvcích (myši z jednoho hnízda, sourozenci, velké pole, v němž vydělujeme políčka). V daném bloku pak náhodně přidělíme každému prvku jedno ošetření. Výsledný model by měl splňovat ($1 \leq i \leq I, 1 \leq j \leq J$)

$$Y_{ij} = \mu + \alpha_i + B_j + e_{ij}, \tag{7.34}$$

kde $e_{ij} \sim \mathsf{N}\big(0, \sigma^2\big)$, $B_j \sim \mathsf{N}\big(0, \sigma_B^2\big)$ je celkem $IJ + J$ nezávislých náhodných veličin. Neznámé konstanty (parametry) α_i se nazývají *pevné efekty*, kdežto B_j jsou *náhodné efekty* jednotlivých bloků.

Snadno zjistíme, že platí

$$\mathsf{cov}(Y_{ij}, Y_{pq}) = \mathsf{cov}(B_j + e_{ij}, B_q + e_{pq}) = \delta_{ip}\delta_{jq}\sigma^2 + \delta_{jq}\sigma_B^2,$$

což lze pomocí Kroneckerova součinu (viz (A.21)) zapsat jako

$$\begin{aligned} \mathsf{var}\,\mathbf{Y} &= \sigma^2(\mathbf{I}_I \otimes \mathbf{I}_J) + \sigma_B^2(\mathbf{11}' \otimes \mathbf{I}_J) \\ &= \sigma^2\left((\mathbf{I}_I \otimes \mathbf{I}_J) + \frac{\sigma_B^2}{\sigma^2}(\mathbf{11}' \otimes \mathbf{I}_J)\right) \end{aligned} \tag{7.35}$$

Protože v našem modelu jsou stejné parametry, jako v modelu analýzy rozptylu jednoduchého třídění, je stejná i matice $\mathbf{X}$. Matici $\mathbf{P} = (\mathbf{Q}, \mathbf{N})$ s ortonormální bází $\mathbb{R}^n$ snadno vyjádříme pomocí matice $\mathbf{N}_0$ typu $J \times (J-1)$, pro kterou je $(\sqrt{(1/J)}\mathbf{1}, \mathbf{N}_0)$ ortonormální. Snadno je

$$\mathbf{Q} = (\mathbf{I}_I \otimes \sqrt{(1/J)}\mathbf{1}), \tag{7.36}$$

$$\mathbf{N} = \mathbf{I}_I \otimes \mathbf{N}_0. \tag{7.37}$$

Ověříme, že jsou oba odhady $\hat{\mathbf{Y}}_W = \hat{\mathbf{Y}}$ v modelu náhodných bloků totožné. Podle věty 7.4 stačí ověřit podmínku (7.29):

$$\mathbf{Q}'\mathbf{W}^{-1}\mathbf{N} = \sqrt{\frac{1}{J}}(\mathbf{I}_I \otimes \mathbf{1}')\left((\mathbf{I}_I \otimes \mathbf{I}_J) + \frac{\sigma_B^2}{\sigma^2}(\mathbf{11}' \otimes \mathbf{I}_J)\right)(\mathbf{I}_I \otimes \mathbf{N}_0), \tag{7.38}$$

$$= \sqrt{\frac{1}{J}}\left((\mathbf{I}_I \otimes \mathbf{1}'\mathbf{N}_0) + \frac{\sigma_B^2}{\sigma^2}(\mathbf{11}' \otimes \mathbf{1}'\mathbf{N}_0)\right), \tag{7.39}$$

$$= \mathbf{O}, \tag{7.40}$$

neboť je $\mathbf{1}'\mathbf{N}_0 = \mathbf{0}'$.

Dál se snadno zjistí, že je $\mathsf{tr}\,(\mathbf{I}_I \otimes \mathbf{N}_0'\mathbf{N}_0 = I(J-1) = n - I$, takže odhad S^2 je nutně kladně vychýleným odhadem σ^2. Podobně se dá ukázat, že matici $\mathbf{W}^{-1}$ nelze pro test hypotézy, že pevné efekty jsou totožné, zapsat ve tvaru (7.33) z věty 7.6.

○

Příklad 7.3 (adjustace) Měřicí přístroj je třeba před použitím adjustovat, nastavit na něm nulu. K tomuto účelu se provádí n_0 měření Y_{0i}^* známého etalonu s hodnotou μ_0, takže k nastavení stupnice použijeme zjištěný průměr $\bar{Y}_0^* \sim \mathsf{N}\big(\mu_0, \sigma^2/n_0\big)$. Vlastní měření (vyjádřené na stupnici před nastavením nuly) vyhovuje modelu

$Y_i^* \sim \mathsf{N}\big(\beta_0^* + \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2\big)$ pro $i = 1, \ldots, n$. Ve skutečnosti však porovnáváme zjištěnou úroveň měřené veličiny s průměrnou hodnotou $\bar{Y}_0^*$ u etalonu, takže dál budeme zpracovávat náhodné veličiny Y_i vyhovující modelu

$$\begin{aligned} Y_i &= Y_i^* - \bar{Y}_0^* \\ &= (\beta_0^* - \mu_0) + \mathbf{x}_i'\boldsymbol{\beta} + (e_i^* - \bar{e}_0^*) \\ &= \beta_0 + \mathbf{x}_i'\boldsymbol{\beta} + e_i, \end{aligned}$$

kde $\bar{e}^*, e_1^*, \ldots, e_n^*$ jsou nezávislé náhodné veličiny. Protože platí

$$\begin{aligned} \mathsf{cov}(Y_i, Y_j) &= \mathsf{cov}(e_i^* - \bar{e}_0^*, e_j^* - \bar{e}_0^*) \\ &= \delta_{ij}\sigma^2 + \sigma^2/n_0, \end{aligned}$$

můžeme varianční matici psát ve tvaru

$$\mathsf{var}\,\mathbf{Y} = \sigma^2\left(\mathbf{I} + (1/n_0)\mathbf{1}\mathbf{1}'\right) \tag{7.41}$$

Každá složka vektoru $\mathbf{Y}$ má rozptyl $((n_0+1)/n_0)\sigma^2$ a každé dvě různé složky stejnou kovarianci $(1/n_0)\sigma^2$.

Lze snadno ukázat, že v popsaném modelu jsou odhady $\hat{\mathbf{Y}}$ a $\hat{\mathbf{Y}}_W$ totožné. Je-li podmodelem $\mathsf{E}\,\mathbf{Y} \sim (\mathbf{1}\gamma, \sigma^2\mathbf{W}^{-1})$, je také splněn předpoklad věty 7.6.

K popsané úloze se dojde například při měření fluorescence, které je vlastně měřením relativním. Neznáme totiž multiplikativní konstantu, která udává poměr mezi naměřeným elektrickým signálem a skutečně vyzářenou energií. K aditivnímu modelu, jako v našem příkladu, dojdeme po logaritmování. ○

7.4. Typ rozdělení

Nakonec pojednáme o vlivu nesplnění předpokladu normálního rozdělení. Budeme předpokládat model $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, přičemž náhodné veličiny jsou $Y_1, \ldots, Y_n$ nezávislé, mají stejné rozdělení s šikmostí γ_1 a špičatostí γ_2 (pro určitost: $\gamma_2 = \mathsf{E}\,(e_i/\sigma)^4 - 3$).

7.4.1. Optimalita odhadu rozptylu

Zavedli jsme odhad S^2 rozptylu σ^2, zjistili jsme, že je nestranný. Nezabývali jsme se otázkou, zda je tento odhad nejlepší. Pro jednoduchost budeme odhadovat jeho násobek, parametr $\theta = (n-r)\sigma^2$, pro který je nestranným odhadem statistika RSS.

V dalším budeme zjišťovat, za jakých předpokladů je ve zvolené třídě odhadů odhad RSS nejlepším odhadem θ.

Nechť A je libovolná pozitivně semidefinitní matice typu $n \times n$. Vyšetřujme vlastnosti statistiky $T = \mathbf{Y}'\mathsf{A}\mathbf{Y}$, která je vzhledem k předpokladu $\mathsf{A} \geq 0$ nezáporná. Má-li být tato statistika nestranným odhadem parametru θ, musí pro všechna $\boldsymbol{\beta}$ a $\sigma^2 > 0$ platit:

$$\begin{aligned} \mathsf{E}\,T = \mathsf{E}\,\mathbf{Y}'\mathsf{A}\mathbf{Y} \;&=\; \operatorname{tr}\mathsf{A}\mathsf{E}\,\mathbf{Y}\mathbf{Y}' \;=\; \operatorname{tr}\mathsf{A}((\mathsf{E}\,\mathbf{Y})(\mathsf{E}\,\mathbf{Y})' + \operatorname{var}\mathbf{Y}) \\ &= \operatorname{tr}\mathsf{A}\left(\mathsf{X}\boldsymbol{\beta}\boldsymbol{\beta}'\mathsf{X}' + \sigma^2\mathsf{I}\right) \;=\; \boldsymbol{\beta}'\mathsf{X}'\mathsf{A}\mathsf{X}\boldsymbol{\beta} + \sigma^2\operatorname{tr}\mathsf{A} = (n-r)\sigma^2. \end{aligned}$$

Vzhledem k požadované pozitivní semidefinitnosti matice A je nestrannost T ekvivalentní s dvojicí požadavků

$$\mathsf{A}\mathsf{X} = \mathsf{O}, \tag{7.42}$$

$$\operatorname{tr}\mathsf{A} = n - r. \tag{7.43}$$

Požadavek (7.42) umožňuje místo $\mathbf{Y}'\mathsf{A}\mathbf{Y}$ psát $\mathbf{e}'\mathsf{A}\mathbf{e}$. Podle věty A.11 dostaneme

$$\operatorname{var}\mathbf{Y}'\mathsf{A}\mathbf{Y} = \sigma^4\left(\gamma_2\sum a_{ii}^2 + 2\operatorname{tr}\mathsf{A}^2\right).$$

Protože je naším cílem konfrontovat odhad $T = \mathbf{Y}'\mathsf{A}\mathbf{Y}$ s odhadem $RSS = \mathbf{Y}'\mathsf{M}\mathbf{Y}$, zavedeme matici $\mathsf{D} = \mathsf{A} - \mathsf{M}$. Požadavek (7.43) přejde v požadavek

$$\operatorname{tr}\mathsf{D} = 0, \tag{7.44}$$

podobně požadavek (7.42) znamená $\mathsf{O} = (\mathsf{M} + \mathsf{D})\mathsf{X} = \mathsf{D}\mathsf{X}$. Je tedy nutně (nezapomeňme, že matice D je symetrická) $\mathcal{M}(\mathsf{D}) \subset \mathcal{M}(\mathsf{M})$, tedy

$$\mathsf{M}\mathsf{D} = \mathsf{D}. \tag{7.45}$$

Nyní budeme minimalizovat rozptyl kvadratické formy s maticí $\mathsf{A} = \mathsf{M} + \mathsf{D}$. K tomu budeme potřebovat druhou mocninu matice A. S využitím (7.45) a (7.44) dostaneme

$$\begin{aligned} \mathsf{A}^2 &= (\mathsf{M} + \mathsf{D})(\mathsf{M} + \mathsf{D}) \\ &= \mathsf{M} + 2\mathsf{D} + \mathsf{D}^2, \\ \operatorname{tr}\mathsf{A}^2 &= (n - r) + \operatorname{tr}\mathsf{D}^2. \end{aligned}$$

Proto nakonec vychází

$$\begin{aligned} \operatorname{var}\mathbf{Y}'\mathsf{A}\mathbf{Y} &= \sigma^4\left(\gamma_2\left(\sum m_{ii}^2 + 2\sum m_{ii}d_{ii} + \sum d_{ii}^2\right) + 2(n-r) + 2\operatorname{tr}\mathsf{D}^2\right) \\ &= \sigma^4\left(\gamma_2\sum m_{ii}^2 + 2(n-r)\right) \\ &\quad + 2\sigma^4\left(\gamma_2(\sum d_{ii}^2/2 + \sum m_{ii}d_{ii}) + \operatorname{tr}\mathsf{D}^2\right) \\ &= \operatorname{var}\mathbf{Y}'\mathsf{M}\mathbf{Y} + 2\sigma^4 g(\mathsf{D}), \end{aligned}$$

kde jsme zavedli

$$g(\mathsf{D}) = \gamma_2\left(\sum d_{ii}^2/2 + \sum m_{ii}d_{ii}\right) + \operatorname{tr}\mathsf{D}^2.$$

Popíšeme dvě situace, v nichž funkce $g(\mathsf{D})$ minimální právě pro $\mathsf{D} = \mathsf{O}$.

Případ $\gamma_2 = 0$. Tento předpoklad splňuje zejména normální rozdělení. Funkce $g(\mathbf{D}) = \operatorname{tr}\mathbf{D}^2$ je nezáporná, minimální je právě pro $\mathbf{D} = \mathbf{O}$.

Případ $m_{ii} = m$. Pokud jsou všechny diagonální prvky matice $\mathbf{M}$ stejné, musí být rovny hodnotě $(n-r)/n$, neboť stopa matice $\mathbf{M}$ je rovna $n-r$. Proto lze funkci $g(\mathbf{D})$ postupně (použij (7.44)) upravit na výraz

$$\begin{aligned} g(\mathbf{D}) &= \gamma_2 \sum d_{ii}^2/2 + \sum\sum d_{ij}^2 \\ &= (\gamma_2/2+1)\sum d_{ii}^2 + 2\sum_{i<j}\sum d_{ij}^2. \end{aligned}$$

Výraz je minimální opět pro $\mathbf{D} = \mathbf{O}$, neboť obecně platí $\gamma_2 \geq -2$.

Shrneme-li svá zjištění, dostaneme následující tvrzení.

Věta 7.7. Jestliže platí některá z podmínek

$$\gamma_2 = 0, \tag{7.46}$$

$$h_{ii} = h, \qquad 1 \leq i \leq n, \tag{7.47}$$

potom je odhad S^2 nejlepším kvadratickým nezáporným nestranným odhadem rozptylu σ^2. Je-li splněna podmínka (7.47), potom platí

$$\operatorname{var} S^2 = \frac{2\sigma^4}{n-r}\left(1 + \frac{\gamma_2}{2}\frac{n-r}{n}\right).$$

D ů k a z: K důkazu stačí si uvědomit, že platí $h_{ii} = 1 - m_{ii}$, zbytek důkazu plyne z úvah uvedených před zněním tvrzení. □

Splňuje-li lineární model podmínku (7.47), říkáme, že je to **kvadraticky vyvážený** model. Mezi kvadraticky vyvážené patří zejména mnohé modely analýzy rozptylu.

7.4.2. Test podmodelu

Snadno se lze přesvědčit, že v normálním lineárním modelu lze statistiku F (3.10) pro testování podmodelu $\mathsf{E}\,\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0$ vyjádřit jako podíl dvou nezávislých nestranných odhadů rozptylu (pro zjednodušení označíme $\mathbf{Q}_2 = \mathbf{N}$, příslušné stupně volnosti jsou $f_1 = r - r_0$ a $f_2 = n - r$)

$$F = \frac{\mathbf{Y}'\mathbf{Q}_1\mathbf{Q}_1'\mathbf{Y}/f_1}{\mathbf{Y}'\mathbf{Q}_2\mathbf{Q}_2'\mathbf{Y}/f_2},$$

přičemž pozitivně semidefinitní idempotentní (projekční) matice $\mathbf{Q}_j\mathbf{Q}_j'$ mají hodnosti $\mathsf{h}(\mathbf{Q}_j\mathbf{Q}_j') = \mathsf{h}(\mathbf{Q}_j) = f_j$ a platí $\mathbf{Q}_1'\mathbf{Q}_2 = \mathbf{O}$. V dalším budeme aproximovat první dva momenty logaritmu statistiky F a pokusíme se vymezit, kdy budou tyto aproximace stejné, jako v případě normálního lineárního modelu.

Označme vektor diagonálních prvků matice $\mathbf{Q}_j\mathbf{Q}_j'$ symbolem $\mathbf{q}_j$. Potom pro j-tý odhad rozptylu

$$S_j^2 = \mathbf{Y}'\mathbf{Q}_j\mathbf{Q}_j'\mathbf{Y}/f_j \tag{7.48}$$

s použitím věty A.11 platí $\mathsf{E}\, S_j^2 = \sigma^2$ a také

$$\begin{aligned} \mathsf{var}\, S_j^2 &= \frac{\sigma^4}{f_j^2}(\gamma_2\mathbf{q}_j'\mathbf{q}_j + 2f_j), \qquad j = 1, 2, \\ \mathsf{cov}(S_1^2, S_2^2) &= \frac{\sigma^4}{f_1 f_2}\gamma_2\mathbf{q}_1'\mathbf{q}_2. \end{aligned}$$

K nekorelovanosti obou odhadů rozptylu zdánlivě není nutné $\gamma_2 = 0$ (např. normální rozdělení), stačila by „ortogonalita“ diagonálních prvků matic $\mathbf{Q}_1\mathbf{Q}_1'$ a $\mathbf{Q}_2\mathbf{Q}_2'$. Tyto matice jsou však pozitivně semidefinitní, takže vektory $\mathbf{q}_1, \mathbf{q}_2$ mají nezáporné prvky. K ortogonalitě by se tedy musel sejít každý nenulový prvek jednoho vektoru s nulovým prvkem druhého vektoru. Přitom přinejmenším u diagonálních prvků matice $\mathbf{Q}_2\mathbf{Q}_2' = \mathbf{M}$ v rozumných případech nulové prvky vyloučíme (viz větu 8.1).

Místo F budeme dál vyšetřovat rozdělení $Z = (1/2)\log F$, neboť i v normálním modelu je rozdělení statistiky Z mnohem více symetrické, lépe aproximovatelné normálním rozdělením. Pomocí Taylorova rozvoje

$$\log S_j^2 \doteq \log\sigma^2 + \frac{S_j^2 - \sigma^2}{1!}\frac{1}{\sigma^2} + \frac{(S_j^2 - \sigma^2)^2}{2!}\left(-\frac{1}{\sigma^4}\right)$$

dostaneme

$$\begin{aligned} \mathsf{E}\,\log S_j^2 &\doteq \log\sigma^2 - \frac{\mathsf{var}\, S_j^2}{2\sigma^4} && (7.49) \\ &= \log\sigma^2 - \frac{1}{f_j} - \frac{\gamma_2}{2f_j^2}\mathbf{q}_j'\mathbf{q}_j, && (7.50) \end{aligned}$$

takže pro $\mathsf{E}\, Z$ dostaneme aproximaci

$$\begin{aligned} \mathsf{E}\, Z &\doteq \frac{1}{2}(\mathsf{E}\,\log S_1^2 - \mathsf{E}\,\log S_2^2) \\ &= \frac{1}{2}\left(\frac{1}{f_2} - \frac{1}{f_1} + \frac{\gamma_2}{2}\left(\frac{1}{f_2^2}\mathbf{q}_2'\mathbf{q}_2 - \frac{1}{f_1^2}\mathbf{q}_1'\mathbf{q}_1\right)\right) \\ &= \frac{1}{2}\left(\frac{1}{f_2} - \frac{1}{f_1} + \frac{\gamma_2}{2f_1^2f_2^2}(f_1\mathbf{q}_2 - f_2\mathbf{q}_1)'(f_1\mathbf{q}_2 + f_2\mathbf{q}_1)\right). \end{aligned}$$

Podobně pomocí aproximace $\log S_j^2 \doteq \log\sigma^2 + (S^2 - \sigma^2)/\sigma^2$ dostaneme

$$\mathsf{var}\, Z \doteq \frac{1}{2}\left(\frac{1}{f_1} + \frac{1}{f_2}\right)\left(1 + \frac{\gamma_2}{2f_1f_2(f_1 + f_2)}(f_1\mathbf{q}_2 - f_2\mathbf{q}_1)(f_1\mathbf{q}_2 - f_2\mathbf{q}_1)\right).$$

Závěr je nasnadě. Aproximované první dva momenty statistiky Z nezávisí na hodnotě γ_2, když platí

$$f_1\mathbf{q}_2 = f_2\mathbf{q}_1. \tag{7.51}$$

Jednou ze situací, kdy je tato podmínka splněna, je případ kdy model i podmodel jsou kvadraticky vyvážené. Pak je totiž $\mathbf{q}_j = (f_j/n)\mathbf{1}$ a podmínka (7.51) je bezpečně splněna.

Poznámka. V článku Box, Watson (1962) je vyšetřován podmodel $\mathsf{E}\,\mathbf{Y} = \mathbf{1}\beta_0$. Technikou permutačních momentů je ukázáno, že rozptyl testové statistiky nezávisí na γ_2 v případě, že se řádky matice $\mathbf{X}$ (nebereme v úvahu sloupec $\mathbf{1}$, jehož přítomnost v $\mathbf{X}$ se předpokládá) chovají jako náhodný výběr z mnohonásobného normálního rozdělení.

7.4.3. Příklady

Ukažme si příklad kvadraticky vyváženého modelu.

Příklad 7.4 (dvojné třídění) V oddílu 6.2 jsme zavedli model pro

$$Y_{ijt} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijt}, \quad 1 \le t \le n_{ij}, 1 \le i \le I, 1 \le j \le J,$$

přičemž náhodné veličiny $e_{ijt} \sim \mathsf{N}(0, \sigma^2)$ jsou nezávislé. Vysvětlili jsme, že je

$$\hat{Y}_{ijt} = \bar{Y}_{ij\bullet} = \frac{1}{n_{ij}} \sum_{t=1}^{n_{ij}} Y_{ijt}.$$

Je tedy $h_{ijt,ijt} = 1/n_{ij}$, takže o kvadraticky vyvážený model půjde v případě, že počty opakování n_{ij} budou shodné, tj. když bude $n_{ij} = T$ pro všechna i, j.

Když testujeme nulovou hypotézu, podle které je vliv faktorů A, B aditivní, ověřujeme vlastně podmodel daný omezeními $\gamma_{ij} = 0$ pro všechna i, j, tedy platí

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ijt}, \quad 1 \le t \le n_{ij}, 1 \le i \le I, 1 \le j \le J.$$

V případě $n_{ij} = T$ pro všechna i, j bude v podmodelu odhadem střední hodnoty $\mathsf{E}\,Y_{ijt}$ výraz

$$\begin{aligned}\hat{Y}^0_{ijt} &= \bar{Y}_{i\bullet\bullet} + \bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet} \\ &= \frac{1}{JT}\sum_{j't'} Y_{ij't'} + \frac{1}{IT}\sum_{i't'} Y_{i'jt'} - \frac{1}{IJT}\sum_{i'j't'} Y_{i'j't'},\end{aligned}$$

takže tentokrát je

$$h^0_{ijt,ijt} = \frac{1}{JT} + \frac{1}{IT} - \frac{1}{IJT}.$$

Vektor $\mathbf{q}_1$ z odstavce 7.4.2 (diagonála matice $\mathbf{Q}_1\mathbf{Q}_1'$) má tedy každém místě prvek

$$h_{ijt,ijt} - h^0_{ijt,ijt} = \frac{1}{T} - \left(\frac{1}{JT} + \frac{1}{IT} - \frac{1}{IJ}\right) = \frac{(I-1)(J-1)}{IJT}.$$

○

8. Rezidua

V této kapitole se budeme věnovat podrobně složkám u_i vektoru $\mathbf{u}$ a jednotlivým jejich „vylepšením“. Zavedeme dvojí upravená rezidua, vhodná zejména pro testování odlehlosti jednotlivých pozorování. Proto bude užitečné vyšetřit vlastnosti odhadů po vynechání jednoho pozorování.

8.1. Vynechání jednoho pozorování

Zvolíme pevně index t a budeme se snažit vyšetřit model bez tohoto pozorování (nazveme jej *model vynechaného pozorování*). Použijeme při tom označení zavedené na začátku appendixu:

$$\mathbf{Y}_{[t]} \sim (\mathbf{X}_{[t\bullet]}\boldsymbol{\beta}, \sigma^2\mathbf{I}). \tag{8.1}$$

Odhady v modelu (8.1) budeme porovnávat s jiným modelem, kde naopak přidáme jednu nezávisle proměnnou, specifickou pro jediné, t-té pozorování (nazveme *model odlehlého pozorování*).

$$\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta} + \mathbf{j}_t\gamma, \sigma^2\mathbf{I}). \tag{8.2}$$

V tomto druhém případě jde o speciální případ rozšířeného modelu (7.1), proto statistiky vztažené k tomuto modelu označíme dolním indexem g. Nejprve se budeme zajímat o předpoklady, které zajistí odhadnutelnost parametru γ.

Věta 8.1. Následující tři tvrzení jsou ekvivalentní:

$$\mathsf{h}(\mathbf{X}) = \mathsf{h}(\mathbf{X}_{[t\bullet]}), \tag{8.3}$$

$$m_{tt} > 0, \tag{8.4}$$

$$\gamma \text{ je v modelu (8.2) odhadnutelné.} \tag{8.5}$$

D ů k a z: Platí ekvivalence

$$m_{tt} = \mathbf{j}_t'\mathbf{M}\mathbf{j}_t = 0 \Leftrightarrow \mathbf{M}\mathbf{j}_t = \mathbf{0} \Leftrightarrow \mathbf{j}_t \in \mathcal{M}(\mathbf{X}).$$

To znamená, že $m_{tt} = 0$ právě tehdy, když existuje $\mathbf{a} \in \mathbb{R}^k$ tak, že je $\mathbf{X}\mathbf{a} = \mathbf{j}_t$. Jinými slovy právě tehdy, když existuje vektor $\mathbf{a}$, který je kolmý na všechny řádky matice

$\mathbf{X}$ s výjimkou t-tého. Poslední tvrzení však lze psát také tak, že $\mathcal{M}(\mathbf{X}')^\perp$ je vlastní podmnožinou $\mathcal{M}((\mathbf{X}_{[t\bullet]})')^\perp$, což je opět ekvivalentní s tvrzením, že $\mathcal{M}((\mathbf{X}_{[t\bullet]})')$ je vlastní podmnožinou $\mathcal{M}(\mathbf{X}')$, což je už naposled ekvivalentní s tvrzením $\mathsf{h}(\mathbf{X}_{[t\bullet]}) < \mathsf{h}(\mathbf{X})$. Protože nutně platí $\mathsf{h}(\mathbf{X}_{[t\bullet]}) \leq \mathsf{h}(\mathbf{X})$, dokázali jsme tak ekvivalenci (8.3) a (8.4).

Věnujme se nyní odhadnutelnosti parametru γ v modelu (8.2). Ta je ekvivalentní s existencí vektoru $\mathbf{q}$ splňujícího $(\mathbf{0}', 1) = \mathbf{q}'(\mathbf{X}, \mathbf{j}_t)$, tedy $1 = \mathbf{q}'\mathbf{j}_t = q_t$ a současně $\mathbf{q}'\mathbf{X} = \mathbf{0}'$. Druhý vztah je ekvivalentní s tvrzením $(\mathbf{x}_{t\bullet})' = (-\mathbf{q}_{[t]})'\mathbf{X}_{[t\bullet]}$. Je tedy $\mathbf{x}_{t\bullet} \in \mathcal{M}((\mathbf{X}_{[t\bullet]})')$, což je konečně ekvivalentní s (8.3). $\square$

Nyní vyjádříme v našem speciálním případě řešení c_g normální rovnice modelu (8.2) podle (7.6)

$$c_g = (\mathbf{j}_t'\mathbf{M}\mathbf{j}_t)^-\mathbf{j}_t'\mathbf{u}.$$

Je-li $m_{tt} > 0$, je parametr γ odhadnutelný a vyjde

$$c_g = \frac{u_t}{m_{tt}}. \tag{8.6}$$

Podobně podle (7.7) vyjde v tomto případě

$$\mathbf{b}_g = \mathbf{b} - \frac{u_t}{m_{tt}}(\mathbf{X}'\mathbf{X})^-\mathbf{x}_{t\bullet} \tag{8.7}$$

a také

$$\begin{aligned}\hat{\mathbf{Y}}_g &= \mathbf{X}\mathbf{b}_g + \mathbf{j}_t c_g \;=\; \mathbf{X}\left(\mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{j}_t c_g\right) + \mathbf{j}_t c_g \\ &= \hat{\mathbf{Y}} + \frac{u_t}{m_{tt}}\left(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\right)\mathbf{j}_t \;=\; \hat{\mathbf{Y}} + \frac{u_t}{m_{tt}}\mathbf{m}_{t\bullet}.\end{aligned}$$

Protože je $\mathbf{d} = \hat{\mathbf{Y}}_g - \hat{\mathbf{Y}}$, dostaneme ještě

$$RSS - RSS_g = ||\mathbf{d}||^2 = \frac{u_t^2}{m_{tt}^2}(\mathbf{m}_{t\bullet})'\mathbf{m}_{t\bullet} = \frac{u_t^2}{m_{tt}}. \tag{8.8}$$

Vraťme se ke vztahu modelů (8.1) a (8.2). Odhady v modelu (8.1) označíme dolním indexem $(t\bullet)$.

Věta 8.2. (Ekvivalence dvou modelů) Vektor $\mathbf{b}_g$ je řešením normální rovnice modelu (8.1) právě, když je spolu s $c_g = Y_t - (\mathbf{x}_{t\bullet})'\mathbf{b}_g$ řešením modelu (8.2). Reziduální součty čtverců jsou v obou modelech stejné. Je-li $m_{tt} > 0$, pak platí

$$\mathbf{b}_{(t\bullet)} = \mathbf{b} - \frac{u_t}{m_{tt}}(\mathbf{X}'\mathbf{X})^-\mathbf{x}_{t\bullet}, \tag{8.9}$$

$$RSS_{(t\bullet)} = RSS - \frac{u_t^2}{m_{tt}}, \tag{8.10}$$

$$\frac{S^2_{(t\bullet)}}{S^2} = \frac{n - r - v_t^2}{n - r - 1}, \tag{8.11}$$

kde jsme označili

$$v_t = \frac{u_t}{S\sqrt{m_{tt}}}. \tag{8.12}$$

D ů k a z: Důkaz plyne ze vztahu

$$||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{j}_t\gamma||^2 = ||\mathbf{Y}_{[t]} - \mathbf{X}_{[t\bullet]}\boldsymbol{\beta}||^2 + (Y_t - (\mathbf{x}_{t\bullet})'\boldsymbol{\beta} - \gamma)^2. \tag{8.13}$$

Je zřejmé, že pro každé $\boldsymbol{\beta}$ lze zvolit γ tak, aby se poslední člen na pravé straně anuloval. Vztahy (8.9) a (8.10) plynou pak bezprostředně z (8.7) a (8.8). Vztah (8.11) dostaneme postupnou úpravou založenou na $S^2_{(t\bullet)} = RSS_{(t\bullet)}/(n-1-r)$. □

Statistika v_t se nazývá *normované reziduum* (někdy také studentizované, ale toto označení použijeme později pro poněkud jinak definovanou statistiku). V prostředí R lze spočítat tato rezidua pomocí funkce `rstandard(a)`, kde `a` je výsledek použití funkce `lm()`. Jednoduchým důsledkem vztahu (8.11) je ekvivalence

$$S^2_{(t\bullet)} < S^2 \Leftrightarrow |v_t| > 1. \tag{8.14}$$

Věta 8.3. (Vlastnosti normovaného rezidua) V normálním lineárním modelu splňujícím $m_{tt} > 0$ platí $\mathsf{E}\, v_t = 0$ a $\mathsf{var}\, v_t = 1$.

D ů k a z: Statistiku v_t lze psát jako

$$v_t = \frac{(\mathbf{j}_t'\mathbf{N})(\mathbf{N}'\mathbf{Y})}{||\mathbf{N}'\mathbf{Y}||}\sqrt{\frac{n-r}{m_{tt}}} = \frac{\mathbf{j}_t'\mathbf{N}\mathbf{U}}{||\mathbf{U}||}\sqrt{\frac{n-r}{m_{tt}}},$$

kde je $\mathbf{U} = \mathbf{N}'\mathbf{Y} \sim \mathsf{N}(\mathbf{0}, \sigma^2\mathbf{I})$ (viz (2.18)). Protože se zřejmě v_t nezmění, když místo $\mathbf{U}$ pro $c > 0$ použijeme $c\mathbf{U}$, podle věty A.12 jsou náhodné veličiny S a v_t jsou nezávislé. Odtud plyne

$$0 = \mathsf{E}\, u_t = \mathsf{E}\,(v_t S\sqrt{m_{tt}}) = (\mathsf{E}\, v_t)(\mathsf{E}\, S)\sqrt{m_{tt}} \Rightarrow \mathsf{E}\, v_t = 0$$

a podobně

$$m_{tt}\sigma^2 = \mathsf{E}\, u_t^2 = (\mathsf{E}\, v_t^2)(\mathsf{E}\, S^2)m_{tt} = m_{tt}\sigma^2\mathsf{E}\, v_t^2 \Rightarrow \mathsf{E}\, v_t^2 = 1.$$

□

8.2. Studentizovaná rezidua

Jak jsme zjistili, pokud platí $m_{tt} > 0$, je parametr γ v modelu (8.2) odhadnutelný. Požadavek $\gamma = 0$ určuje podmodel, v němž platí $\mathbf{Y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Proto lze podmodel testovat pomocí F statistiky

$$F = \frac{RSS - RSS_g}{RSS_g} \frac{n-r-1}{1} \tag{8.15}$$

$$= \frac{u_t^2/m_{tt}}{S^2_{(t\bullet)}} = \frac{u_t^2/(S^2 m_{tt})}{S^2_{(t\bullet)}/S^2}$$

$$= \frac{v_t^2}{\dfrac{n-r-v_t^2}{n-r-1}}$$

$$= v_t^2 \frac{n-r-1}{n-r-v_t^2} \sim \mathsf{F}(1, n-r-1) \tag{8.16}$$

Zkusme použít model (8.1) k tomu, abychom odhadli neznámé parametry a pak ověřili, zda i t-té pozorování klasického modelu $\mathbf{Y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ vyhovuje stejnému modelu.

Odhadněme nejprve střední hodnotu $\mathsf{E}\, Y_t = (\mathbf{x}_{t\bullet})'\boldsymbol{\beta}$ pomocí modelu (8.1), který náhodnou veličinu Y_t neobsahuje. Parametrická funkce $(\mathbf{x}_{t\bullet})'\boldsymbol{\beta}$ je nutně odhadnutelná, neboť předpoklad $m_{tt} > 0$ je podle věty 8.1 ekvivalentní s tím, že matice $\mathbf{X}$ a $\mathbf{X}_{[t\bullet]}$ mají stejnou hodnost. Rozdíl mezi skutečným pozorováním a odhadem jeho střední hodnoty

$$Y_t - (\mathbf{x}_{t\bullet})'\mathbf{b}_{(t\bullet)} = c_g = \frac{u_t}{m_{tt}} \tag{8.17}$$

se v počítačových výstupech často nazývá *deleted residual*. Jeho rozptyl je zřejmě roven

$$\mathsf{var}\,(Y_t - (\mathbf{x}_{t\bullet})'\mathbf{b}_{(t\bullet)}) = \mathsf{var}\,\frac{u_t}{m_{tt}} = \frac{\sigma^2}{m_{tt}}.$$

K testování hypotézy, že střední hodnota t-tého pozorování je rovna $(\mathbf{x}_{t\bullet})'\boldsymbol{\beta}$, tedy použijeme statistiku, která ve jmenovateli používá odhad rozptylu nepochybně nezávislý s čitatelem

$$T = \frac{Y_t - (\mathbf{x}_{t\bullet})'\mathbf{b}_{(t\bullet)}}{\sqrt{\widehat{\mathsf{var}}\,(Y_t - (\mathbf{x}_{t\bullet})'\mathbf{b}_{(t\bullet)})}} \sim \mathsf{t}(n-r-1)$$

$$= \frac{u_t/m_{tt}}{\sqrt{S^2_{(t\bullet)}/m_{tt}}}$$

$$= \frac{u_t}{S_{(t\bullet)}\sqrt{m_{tt}}} = v_t^*. \tag{8.18}$$

Statistika v_t^* se nazývá *studentizované reziduum*, někdy též jackkniffe reziduum. V R se počítá pomocí `rstudent(a)`, kde `a` je výsledek použití `lm()`. Z odvození je zřejmé i jeho rozdělení. Všimněte si, že F statistika z (8.15) je rovna právě v_t^{*2}. Ze vztahu (8.16) je zřejmý také vztah obojích reziduí:

$$v_t^* = \sqrt{\frac{n-r-1}{n-r-v_t^2}}\, v_t.$$

Věta 8.4. (Vlastnosti studentizovaných reziduí) Nechť pro dané t, $1 \leq t \leq n$, v normálním lineárním modelu $\mathbf{Y} \sim \mathsf{N}\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\big)$ platí $m_{tt} > 0$. Potom má studentizované reziduum v_t^* Studentovo t rozdělení s $n - r - 1$ stupni volnosti a platí

$$\text{je-li } n - r > 2, \text{ pak } \mathsf{E}\, v_t^* = 0, \tag{8.19}$$

$$\text{je-li } n - r > 3, \text{ pak } \mathsf{var}\, v_t^* = \frac{n - r - 1}{n - r - 3}. \tag{8.20}$$

D ů k a z: K dokončení důkazu stačí připomenout vlastnosti Studentova rozdělení, viz například (Anděl, 1998, odst. 4.5). □

Poslední úvahou jsme vlastně sledovali smysl modelu odlehlého pozorování (8.2). Parametr γ slouží k tomu, aby střední hodnota t-tého pozorování mohla být zcela individuální, nezávislá na středních hodnotách ostatních pozorování. Pouze v případě $\gamma = 0$ je použitý model pro všechna pozorování stejný. Odtud dostáváme nejčastější použití studentizovaných reziduí, kdy pomocí v_t^* testujeme, zda t-té pozorování je odlehlé, tj. má střední hodnotu jinou, než určuje model.

Uvedený postup je adekvátní v případě, že index t (které pozorování má být odlehlé) známe předem, nezávisle na náhodném vektoru $\mathbf{Y}$. Na hladině α označíme t-té pozorování (t předem dáno) za odlehlé, když platí $|v_t^*| \geq t_{n-r-1}(\alpha)$.

V praxi je mnohem častější jiná situace, kdy nevíme předem, které pozorování by mohlo být odlehlé. Nejčastěji podezříváme z odlehlosti takové pozorování, které má v absolutní hodnotě největší reziduum, případně v absolutní hodnotě největší studentizované reziduum (nebo normované reziduum, což je totéž). Řešená úloha patří k mnohonásobným srovnáním.

Pro nějaké $\delta \in (0,1)$ a pro $i = 1, \ldots, n$ zaveďme náhodné jevy $W_i(\delta) = \{|v_i^*| \geq t_{n-r-1}(\delta)\}$. Některé z n pozorování bychom měli na hladině nejvýše α označit za odlehlé, pokud platí $\mathsf{P}(\cup_{i=1}^n W_i(\delta)) \leq \alpha$. Problém jak zvolit δ pomůže vyřešit Bonferroniho nerovnost (viz též A.13 z appendixu pro $A_i = W_i(\delta)$). Zvolíme-li $\delta = \alpha/n$, bude zajištěno

$$\mathsf{P}(\cup_{i=1}^n W_i(\alpha/n)) \leq \sum_{i=1}^n \mathsf{P}(W_i(\alpha/n)) = \alpha.$$

Prakticky to znamená použít kritickou hodnotu $t_{n-r-1}(\alpha/n)$. Soudobé programové vybavení je schopno udat ke každému studentizovanému reziduu v_i^* hodnotu $p_i = \mathsf{P}(|T_{n-r-1}| \geq v_i^*)$, kde T_{n-r-1} je náhodná veličina s rozdělením $\mathsf{t}(n - r - 1)$. Za odlehlé pak označíme každé pozorování, pro které vyjde $p_i \leq \alpha/n$, což je totéž, jako $|v_i^*| \geq t_{n-r-1}(\alpha/n)$.

Poněkud jemnější Holmovu metodu mnohonásobných srovnání lze nalézt u Havránka (1993) od str. 174.

8.3. Vliv jednotlivých pozorování

Připomeňme význam dolního indexu $(t\bullet)$, který jsme zavedli na str. 185, který označuje odhad získaný z modelu (8.1), ať už jej použijeme k jakémukoliv účelu. Symbolem $\hat{\mathbf{Y}}_{(t\bullet)}$ tedy označíme odhad celého n-členného vektoru $\mathsf{E}\,\mathbf{Y}$.

O vlivu jednotlivých pozorování vypovídají rezidua. Další pohled dostaneme, když porovnáme odhady pro $\mathsf{E}\,Y_t$ a pro $\boldsymbol{\beta}$ založené na všech pozorováních s odhady získanými po vyloučení jediného pozorování. Zpravidla se při tom předpokládá, že vyloučení jednoho pozorování nesníží hodnost regresní matice $\mathbf{X}$, tedy že pro všechna t platí $m_{tt} > 0$.

Nejprve se budeme zabývat citlivostí odhadů na případné vyloučení t-tého pozorování.

8.3.1. Diagonála H

Především připomeňme, v tomto textu uvažujeme model s absolutním členem, takový, že první sloupec matice $\mathbf{X}$ je tvořen jedničkami. Použijme označení

$$\mathbf{X} = \left(\mathbf{1}, \mathbf{x}_{1\bullet}, \cdots, \mathbf{x}_{k\bullet}\right)$$

Symbolem $\mathbf{x}_{\bullet j}$ tedy značíme hodnoty j-tého regresoru a symbolem $\bar{x}_j$ označíme průměr tohoto regresoru. Symbolem $\tilde{\mathbf{X}}$ označíme matici s k sloupci

$$\tilde{\mathbf{X}} = \left(\mathbf{x}_{\bullet 1} - \bar{x}_1\mathbf{1}, \mathbf{x}_{\bullet 2} - \bar{x}_2\mathbf{1}, \cdots, \mathbf{x}_{\bullet k} - \bar{x}_k\mathbf{1}\right).$$

Platí zřejmě $\mathcal{M}(\mathbf{X}) = \mathcal{M}((\mathbf{1}, \tilde{\mathbf{X}}))$, takže projekční matici $\mathbf{H}$ lze zapsat také ve tvaru

$$\mathbf{H} = (\mathbf{1}, \tilde{\mathbf{X}}) \begin{pmatrix} n & \mathbf{0}' \\ \mathbf{0} & \tilde{\mathbf{X}}'\tilde{\mathbf{X}} \end{pmatrix}^{-} (\mathbf{1}, \tilde{\mathbf{X}})' = \frac{1}{n}\mathbf{1}\mathbf{1}' + \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-}\tilde{\mathbf{X}}'.$$

Je tedy

$$h_{tt} = \frac{1}{n} + (x_{t1} - \bar{x}_1, \ldots, x_{tk} - \bar{x}_k)(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-}(x_{t1} - \bar{x}_1, \ldots, x_{tk} - \bar{x}_k)',$$

takže t-tý diagonální prvek matice $\mathbf{H}$ můžeme interpretovat jako o číslo $1/n$ zvětšenou zobecněnou vzdálenost t-tého řádku matice $\mathbf{X}$ od těžiště všech jejích řádků. Samotná hodnota h_{tt} je v počítačových výstupech uváděna pod označením *leverage.* Pozorování s velkou hodnotou h_{tt} mohou značně ovlivnit odhad parametru $\boldsymbol{\beta}$, zpravidla se za mezní hodnotu považuje hodnota $2r/n$. (Je hodnota h_{tt} dána jednoznačně?)

Pro regresní přímku (viz (4.3)) platí

$$h_{tt} = \frac{1}{n} + \frac{(x_t - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Nejvíce tedy ovlivňují odhad parametrů regresní přímky ta pozorování, jejichž nezávisle proměnná je nejdále od průměru této proměnné.

8.3.2. DFBETAS

Abychom mohli porovnávat dva odhady vektoru $\boldsymbol{\beta}$, musíme zajistit jeho odhadnutelnost. Proto zde předpokládáme úplnou hodnost matice $\mathbf{X}$. Podle (8.9) z věty 8.2 platí (použijeme opět označení $\mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1}$)

$$\mathbf{b} - \mathbf{b}_{(t\bullet)} = \frac{u_t}{m_{tt}} \mathbf{V} \mathbf{x}_{t\bullet}. \tag{8.21}$$

Tyto rozdíly ukazují změny v odhadech jednotlivých regresních koeficientů způsobené vynecháním t-tého pozorování. Častěji se uvedené rozdíly škálují tak, že jsou děleny odhadem střední chyby příslušné složky vektoru $\mathbf{b}$, takže j-tá složka škálovaného rozdílu je rovna

$$\triangle_t(\beta_j) = \frac{b_j - b_{j(t\bullet)}}{S_{(t\bullet)}\sqrt{v_{jj}}}. \tag{8.22}$$

Uvedené rozdíly bývají označovány jako $DFBETAS$. Neškálovanou verzi rozdílu uvedenou v (8.21) bychom pak označili jako $DFBETA$.

8.3.3. DFFITS

Podobně se můžeme zajímat o odhad parametrické funkce $\mu_t = (\mathbf{x}_{t\bullet})'\boldsymbol{\beta}$, která je vždy odhadnutelná. Předpoklad $m_{tt} > 0$ zajistí, že je odhadnutelná i po vynechání t-tého pozorování. Proto bez ohledu na hodnost matice $\mathbf{X}$ platí

$$\begin{aligned} \hat{Y}_{t(t\bullet)} &= (\mathbf{x}_{t\bullet})'\mathbf{b}_{(t\bullet)} = \hat{Y} - (\mathbf{x}_{t\bullet})'(\mathbf{X}'\mathbf{X})^{-}\mathbf{x}_{t\bullet}\frac{u_t}{m_{tt}} \\ &= \hat{Y}_t - \frac{h_{tt}}{m_{tt}} u_t \end{aligned}$$

Rozdíl odhadů střední hodnoty $\mathsf{E}\, Y_i$ lze tedy vyjádřit jako

$$\hat{Y}_t - \hat{Y}_{t(t\bullet)} = \frac{h_{tt}}{m_{tt}} u_t. \tag{8.23}$$

Uvedený rozdíl bývá někdy označen jako $DFFIT$. Podobně jako u rozdílu odhadů regresních koeficientů provedeme škálování, přičemž použijeme $\mathsf{var}\, \hat{Y}_t = \sigma^2 m_{tt}$. Postupnými úpravami dojdeme k vyjádření pomocí studentizovaného rezidua

$$\begin{aligned} \triangle_t(\mathsf{E}\, Y_t) &= \frac{\hat{Y}_t - \hat{Y}_{t(t\bullet)}}{\sqrt{\mathsf{var}\, \hat{Y}_t}} = \frac{h_{tt}}{m_{tt}} \frac{u_t}{S_{(t\bullet)}\sqrt{h_{tt}}} = \sqrt{\frac{h_{tt}}{m_{tt}}} \frac{u_t}{S_{(t\bullet)}\sqrt{m_{tt}}} \\ &= \sqrt{\frac{h_{tt}}{m_{tt}}} v_t^* \end{aligned} \tag{8.24}$$

Pro tuto statistiku se používá označení $DFFITS$.

8.3.4. Cookova vzdálenost

Pokusme se vyjádřit vliv t-tého pozorování na odhad celé střední hodnoty $\mathsf{E}\,\mathbf{Y}$ pomocí jediného čísla tak, že zjistíme čtverec délky rozdílu obou odhadů:

$$\begin{aligned}
||\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(t\bullet)}||^2 &= ||\mathbf{Xb} - \mathbf{Xb}_{(t\bullet)}||^2 \; = \; ||\mathbf{X}(\mathbf{b} - \mathbf{b}_{(t\bullet)})||^2 \\
&= (\mathbf{b} - \mathbf{b}_{(t\bullet)})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{b}_{(t\bullet)}) \\
&= \left(\frac{u_t}{m_{tt}}(\mathbf{X}'\mathbf{X})^{-}\mathbf{x}'_{t\bullet}\right)' \mathbf{X}'\mathbf{X} \left(\frac{u_t}{m_{tt}}(\mathbf{X}'\mathbf{X})^{-}\mathbf{x}'_{t\bullet}\right) \\
&= \frac{u_t^2}{m_{tt}^2} h_{tt}.
\end{aligned}$$

Drobnou modifikací (např. abychom dostali bezrozměrnou charakteristiku) dostaneme odtud *Cookovu vzdálenost*

$$D_t = \frac{1}{rS^2}||\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(t\bullet)}||^2 = v_t^2 \frac{h_{tt}}{m_{tt}} \frac{1}{r}. \tag{8.25}$$

Cookova vzdálenost je tedy součinem tří členů. První z nich ukazuje nakolik se střední hodnota závisle proměnné Y_t odlišuje od střední hodnoty dané modelem. Druhý člen je monotonní funkcí h_{tt}, kterážto hodnota ukazuje, jak daleko je řádek $\mathbf{x}_{t\bullet}$ od těžiště všech řádků matice $\mathbf{X}$, jak bude vidět z následující úvahy. Tato charakteristika je podobná (až na dělení hodností matice $\mathbf{X}$) čtverci statistiky $\triangle_t(\mathsf{E}\,Y_t)$, jen je použito normovaneé reziduum v_t na místo rezidua studentizovaného v_t^*.

8.3.5. COVRATIO

Nyní budeme hodnotit vliv vynechání t-tého pozorování na přesnost odhadů regresních koeficientů. Budeme tedy opět předpokládat model s úplnou hodností. Abychom místo odhadu varianční matice dostali jednorozměrnou charakteristiku, použijeme determinant tohoto odhadu. Statistika $COVRATIO$ je dána podílem těchto determinantů, přičemž v čitateli se determinant odkazuje na odhady s vynecháním t-tého pozorování.

Dříve než uvedeme vzorec, pomocí často používané identity pro determinanty (viz např. (Anděl, 1978, Věta IV. 4)) najdeme vztah mezi determinanty dvou souvisejících matic:

$$\begin{aligned}
\begin{vmatrix} \mathbf{X}'\mathbf{X} & \mathbf{x}_{t\bullet} \\ (\mathbf{x}_{t\bullet})' & 1 \end{vmatrix} &= \left|\mathbf{X}'\mathbf{X}\right| \left(1 - (\mathbf{x}_{t\bullet})'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{t\bullet}\right) = |\mathbf{X}'\mathbf{X}| m_{tt} \\
&= 1 \cdot \left|\mathbf{X}'\mathbf{X} - \mathbf{x}_{t\bullet}(\mathbf{x}_{t\bullet})'\right| = |(\mathbf{X}_{[t\bullet]})'\mathbf{X}_{[t\bullet]}|.
\end{aligned}$$

Hledaný podíl je tedy

$$
\begin{aligned}
\frac{|\widehat{\operatorname{var}}\,\mathbf{b}_{(t\bullet)}|}{|\widehat{\operatorname{var}}\,\mathbf{b}|} &= \left(\frac{S_{(i\bullet)}}{S}\right)^{2(k+1)} \frac{|\mathbf{X}'\mathbf{X}|}{|(\mathbf{X}_{[t\bullet]})'\mathbf{X}_{[t\bullet]}|} \\
&= \left(\frac{S_{(t\bullet)}}{S}\right)^{2(k+1)} \frac{1}{m_{tt}}, \\
&= \frac{1}{m_{tt}} \left(\frac{n-k-1-v_t^2}{n-k-2}\right)^{k+1}.
\end{aligned}
\tag{8.26}
$$

Přesnost odhadu regresních koeficientů se tedy zlepší například tehdy, když je příslušné studentizované reziduum příliš velké (daleko od nuly) a zvláště tehdy, když jde hodnoty regresorů blízké hodnotám průměrným. Pak je totiž h_{tt} malé, takže $1/m_{tt} = 1/(1-h_{tt})$ také malé.

8.4. Nabídka prostředí R

V prostředí R je k disposici zejména funkce `influence.measures()`, kterou lze použít na objekt třídy `lm`. Výsledkem je objekt třídy `infl`, který je složen ze tří prvků: `infmat`, `is.inf` a `call`.

V matici nazvané `infmat` jsou soustředěny hlavní diagnostické statistiky. Každý řádek odpovídá jednomu pozorování, tedy jednomu řádku matice $(\mathbf{Y}, \mathbf{X})$.

Prvních $k+1$ sloupců tvoří matici statistik $DFBETAS$, jejíž (i,j)-tý prvek je dán vztahem (8.22). Tyto sloupce jsou nazvány `dfb.`, kde za tečkou následuje (někdy přiměřeně zkrácený) název příslušného regresoru. Následuje sloupec statistik $DFFITS$ označený `dffit`. Další sloupce, nazvané `cov.r`, `cook.d`, `hat` obsahují odpovídající statistiky $COVRATIO$, D_t a h_{tt}.

Matice `is.inf` má stejný rozměr jako `infmat`. Jednotlivé prvky odpovídají prvkům matice `infmat`, jsou `TRUE`, pokud příslušný prvek ukazuje na problém, tj. pokud překračuje (mnohdy velmi arbitrárně) zvolenou mez. Je to tehdy, když

$$|\triangle_t(\beta_j)| > 1, \tag{8.27}$$

$$|\triangle_t(\mathsf{E}\,Y_t)| > 3\sqrt{\frac{k+1}{n-k-1}}, \tag{8.28}$$

$$|1 - COVRATIO| > 3\frac{k+1}{n-k-1}, \tag{8.29}$$

$$F_{k+1,n-k-1}(D_t) > 0{,}5, \qquad (F \text{ je distr. funkce } F \text{ rozdělení}) \tag{8.30}$$

$$h_{tt} > 3\frac{k+1}{n}. \tag{8.31}$$

V případě statistik, které lze spočítat, i když nemá regresní matice lineárně nezávislé sloupce ($DFFITS$, h_{tt}) je hodnota $k+1$ nahrazena skutečnou hodností regresní matice.

Pokud tiskneme matici `infmat` funkcí `print()`, nejprve se připomene tvar vyšetřované závislosti uložený v `call`. Pak se tiskne matice `infmat`, přičemž na konec každého řádku je doplněna buď hvězdička nebo mezera podle toho, zda je v daném řádku matice `is.inf` aspoň jednou `TRUE` či nikoliv. Výstup pomocí `summary` obsahuje pouze ty řádky, které v bohatším výstupu pomocí `print` obsahují hvězdičku. Hvězdičky jsou tentokrát umístěny u příslušné statistiky.

Normovaná rezidua lze v R spočítat, když se na objekt třídy `lm` použije funkce `rstandard`. Podobně lze spočítat vektor studentizovaných reziduí pomocí funkce `rstudent`, a další statistiky pomocí funkcí `dffits`, `dfbetas`, `covratio`, `cooks.distance`, které se všechny používají na objekt třídy `lm`. Podobně lze spočítat diagonální prvky regresní matice pomocí funkce `hat`, jejímž argumentem je regresní matice. Tu můžeme získat funkcí `model.matrix` uplatněnou na objekt třídy `lm`.

Příklad 8.1 (procento tuku) Vyšetřuje se závislost procenta tuku u mladých mužů v závislosti na jejich výšce a hmotnosti.

```
> summary(f.hw<-lm(fat~height+weight))

Call:
lm(formula = fat ~ height + weight)

Residuals:
     Min       1Q   Median       3Q      Max
-6.40111 -2.94819 -0.02106  2.30723  7.29683

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.55309   15.24621   1.086   0.2831
height      -0.24362    0.09728  -2.504   0.0158 *
weight       0.50418    0.05095   9.896 4.49e-13 ***
---
Residual standard error: 3.731 on 47 degrees of freedom
Multiple R-Squared: 0.714,      Adjusted R-squared: 0.7018
F-statistic: 58.66 on 2 and 47 degrees of freedom,   p-value: 1.681e-013

> anova(lm(fat~height+weight))
Analysis of Variance Table

Response: fat
          Df  Sum Sq Mean Sq F value    Pr(>F)
height     1  270.06  270.06  19.398 6.096e-05 ***
weight     1 1363.26 1363.26  97.922 4.490e-13 ***
Residuals 47  654.33   13.92
---
```

```
> summary(f.hw.infl<-influence.measures(f.hw))
Potentially influential observations of
         lm(formula = fat ~ height + weight) :

  dfb.1_ dfb.hght dfb.wght dffit   cov.r   cook.d hat
2 -0.43   0.60    -0.98    -1.02_*  1.30_*  0.34   0.30_*
4  0.01  -0.01     0.01    -0.01    1.22_*  0.00   0.12
6 -0.60   0.52     0.10     0.79_*  0.98    0.20   0.14
```

○

8.5. Nekorelovaná rezidua

Dvě až dosud uvedené modifikace reziduí odstraňují jeden z problémů klasických reziduí, totiž jejich nestejné rozptyly. Nemohou však odstranit další nedostatek reziduí v porovnání s chybovým členem $\mathbf{e}$, totiž jejich vzájemnou závislost. Vektor reziduí $\mathbf{u}$ leží v podprostoru $\mathcal{M}(\mathbf{X})^{\perp}$, jehož dimenze je nutně menší, než počet jeho složek n. Budeme-li tedy hledat skutečně nekorelovaná (v normálním modelu nezávislá) rezidua, musíme zmenšit jejich počet.

Klasická rezidua můžeme pomocí jakékoliv matice $\mathbf{N}$, jejíž sloupce tvoří ortonormální bázi prostoru $\mathcal{M}(\mathbf{X})^{\perp}$ (tj. která splňuje $\mathbf{N}'\mathbf{N} = \mathbf{I}, \mathbf{N}\mathbf{N}' = \mathbf{M}$), psát v tvaru

$$\mathbf{u} = \mathbf{N}(\mathbf{N}'\mathbf{Y}) = \mathbf{N}\mathbf{n}.$$

Složky vektoru $\mathbf{n}$ nazveme *nekorelovaná rezidua.* Jsou to tedy koeficienty jednoznačně určeného vektoru $\mathbf{u}$ vyjádřeného v některé z nekonečně mnoha ortonormálních bází prostoru $\mathcal{M}(\mathbf{X})^{\perp}$. Snadno zjistíme, že $\mathbf{u}$ má mnohorozměrné normální rozdělení:

$$\mathbf{n} \sim \mathsf{N}\big(\mathbf{N}'\mathbf{X}\beta, \sigma^2\mathbf{N}'\mathbf{N}\big) = \mathsf{N}\big(\mathbf{0}, \sigma^2\mathbf{I}_{n-r}\big).$$

V normální lineárním modelu jsou tedy složky vektoru $\mathbf{n}$ nezávislé, mají nulové střední hodnoty a stejné rozptyly σ^2.

Volbou různých bází prostoru $\mathcal{M}(\mathbf{X})^{\perp}$ dostaneme různá nekorelovaná rezidua. Zajímavou interpretaci mají *rekurzivní rezidua.* Tato rezidua závisí na pořadí řádků matice $\mathbf{X}$, tedy zpravidla na pořadí, v jakém data získáváme.

Vyjdeme z prvního řádku matice $\mathbf{X}$ a postupně budeme přidávat jednotlivé řádky. V každém kroku, kdy se *nezvýší* hodnost postupně rozšiřované matice, spočítáme rozdíl mezi nově přidanou hodnotou Y_t a predikcí její střední hodnoty spočítanou pomocí všech již dřív zavedených pozorování (s menšími indexy). Tento rozdíl ještě normujeme tak, aby vzniklá statistika měla rozptyl rovný σ^2. Předpokládejme, že jsme takto do modelu zavedli prvních t řádků matice $(\mathbf{Y}, \mathbf{X})$, označme je jako

$(\mathbf{Y}_t, \mathbf{X}_t)$ a že při zavedení dalšího pozorování $(Y_{t+1}, (\mathbf{x}_{t+1,\bullet})')$ se hodnost matice regresorů nezvýší. Tuto hodnost označíme jako r_t (tj. platí $\mathsf{h}(\mathbf{X}_t) = \mathsf{h}(\mathbf{X}_{t+1}) = r_t$). Řešení normální rovnice, která používá prvních t pozorování označme jako $\mathbf{b}_t$. Potom bude

$$n_{t-r_t+1} = \frac{Y_{t+1} - (\mathbf{x}_{t+1,\bullet})'\mathbf{b}_t}{\sqrt{1 + (\mathbf{x}_{t+1,\bullet})'(\mathbf{X}_t'\mathbf{X}_t)^-\mathbf{x}_{t+1,\bullet}}}. \tag{8.32}$$

Střední hodnota $\mathsf{E}\, Y_{t+1} = (\mathbf{x}_{t+1,\bullet})'\mathbf{b}_t$ je odhadnutelným parametrem podle věty 2.3, neboť jsme předpokládali, že přidáním $(t+1)$. řádku hodnost matice regresorů nevzrostla. Výraz v čitateli i ve jmenovateli (8.32) je proto jednoznačný pro každé řešení normální rovnice.

Podle (8.32) dostaneme postupně statistiky $n_1, \dots, n_{n-r}$, které mají důležitou vlastnost. Každá z nich je nekorelovaná se všemi statistikami s nižším indexem. Pro $j = 1, \dots, t$ totiž platí

$$\begin{aligned}
&\mathsf{cov}(Y_{t+1} - (\mathbf{x}_{t+1,\bullet})'\mathbf{b}_t, Y_{t+1-j} - (\mathbf{x}_{t+1-j,\bullet})'\mathbf{b}_{t-j})\\
&= \mathsf{cov}(Y_{t+1} - (\mathbf{x}_{t+1,\bullet})'(\mathbf{X}_t'\mathbf{X}_t)^-\mathbf{X}_t'\mathbf{Y}_t,\\
&\qquad Y_{t+1-j} - (\mathbf{x}_{t+1-j,\bullet})'(\mathbf{X}_{t-j}'\mathbf{X}_{t-j})^-\mathbf{X}_{t-j}'\mathbf{Y}_{t-j})\\
&= \sigma^2\Big(0 - 0 - (\mathbf{x}_{t+1,\bullet})'(\mathbf{X}_t'\mathbf{X}_t)^-\mathbf{X}_t'\mathbf{j}_{t+1-j}\\
&\qquad + (\mathbf{x}_{t+1,\bullet})'(\mathbf{X}_t'\mathbf{X}_t)^-\mathbf{X}_t'\begin{pmatrix}\mathbf{I}_{t-j}\\ \mathbf{O}_{t\times j}\end{pmatrix}\mathbf{X}_{t-j}(\mathbf{X}_{t-j}'\mathbf{X}_{t-j})^-\mathbf{x}_{t+1-j,\bullet}\Big)\\
&= \sigma^2\Big(-(\mathbf{x}_{t+1,\bullet})'(\mathbf{X}_t'\mathbf{X}_t)^-\mathbf{x}_{t+1-j,\bullet}\\
&\qquad + (\mathbf{x}_{t+1,\bullet})'(\mathbf{X}_t'\mathbf{X}_t)^-\mathbf{X}_{t-j}'\mathbf{X}_{t-j}(\mathbf{X}_{t-j}'\mathbf{X}_{t-j})^-\mathbf{x}_{t+1-j,\bullet}\Big)\\
&= \sigma^2\Big(-(\mathbf{x}_{t+1,\bullet})'(\mathbf{X}_t'\mathbf{X}_t)^-\mathbf{x}_{t+1-j,\bullet} + (\mathbf{x}_{t+1,\bullet})'(\mathbf{X}_t'\mathbf{X}_t)^-\mathbf{x}_{t+1-j,\bullet}\Big)\\
&= 0.
\end{aligned}$$

Rekurzivní rezidua mají interpretaci, pokud má smysl uspořádání řádků matice $(\mathbf{Y}, \mathbf{X})$. Ukazují, nakolik další pozorování odpovídá modelu obsahujícímu všechna předchozí pozorování. Proto se používají tam, kde se zajímáme o stabilitu závislosti.

8.6. Parciální rezidua

Také parciální rezidua budeme používat tam, kde se budeme zajímat o správnost zvoleného modelu. Tentokrát půjde o vhodnost zařazení toho kterého regresoru.

Zvolme pevně index j sloupce matice $\mathbf{X}$ takový, že platí $\mathsf{h}(\mathbf{X}_{[\bullet j]}) = r - 1$. V takovém případě je parametr β_j odhadnutelný, neboť pseudoinvertovanou maticí v (7.6) je zřejmě nenulové číslo (použili jsme $\mathbf{X}_{[\bullet j]}$ místo $\mathbf{X}$ a $\mathbf{x}_{\bullet j}$ místo $\mathbf{Z}$). Zaveďme vektor *parciálních reziduí* $\mathbf{u}^{(\bullet j)}$ se složkami

$$u_i^{(\bullet j)} = u_i + x_{ij} b_j. \tag{8.33}$$

Protože lze psát

$$u_i^{(\bullet j)} = Y_i - \sum_{\nu \neq j} x_{i\nu} b_\nu,$$

lze vektor $\mathbf{u}^{(\bullet j)}$ interpretovat jako tu složku vektoru hodnot závisle proměnné, kterou se nepodařilo vysvětlit pomocí ostatních regresorů, tedy jako tu složku, jejíž vysvětlení zbylo na j-tý regresor $\mathbf{x}_{\bullet j}$.

Parciální rezidua jsou užitečná především při grafickém vyjádření, v němž se znázorňují body o souřadnicích $[x_{ij}, u_i^{(\bullet j)}]$. Těmito body se prokládá běžná regresní přímka. Užitečné je zjištění, že směrnice této přímky je rovna právě odhadu b_j parametru β_j. Platí totiž

$$\begin{aligned} ||\mathbf{u}^{(\bullet j)} - \mathbf{x}_{\bullet j}\beta||^2 &= ||(\mathbf{Y} - \mathbf{X}_{[\bullet j]}\mathbf{b}_{[j]}) - \mathbf{x}_{\bullet j}\beta||^2 \\ &\geq ||\mathbf{Y} - \mathbf{X}\mathbf{b}||^2. \end{aligned}$$

Jen je třeba opatrně interpretovat těsnost rozmístění bodů kolem přímky, neboť grafické znázornění odpovídá formálně modelu $\mathbf{u}^{(\bullet j)} \sim (\mathbf{x}_{\bullet j}\beta, \sigma^2\mathbf{I})$, v němž má odhad pro β obecně menší rozptyl, než je skutečný rozptyl odhadu b_j v původním modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

Některé programy při grafickém znázornění používají vektor

$$\mathbf{u}^{(\bullet j)} + (\bar{Y} - b_j \bar{x}_j)\mathbf{1} \tag{8.34}$$

místo $\mathbf{u}^{(\bullet j)}$, což má smysl, jen když je $\mathbf{1} \in \mathcal{M}(\mathbf{X})$. Graf potom opravdu připomíná „očištěnou závislost“ Y na j-tém regresoru, neboť průměr souřadnic na svislé ose je roven $\bar{Y}$.

V prostředí R existuje `predict.lm()` použitelná na objekt třídy `lm`. Je-li `a` onen objekt třídy `lm` v modelu s absolutním členem, pak funkce

```
mean(fitted(a))+resid(a)+predict(a,type="terms")
```

dá matici, jejíž sloupce tvoří jednotlivá parciální rezidua počítaná podle (8.34).

8.7. Grafy reziduí

Rezidua poskytují řadu možností, jak diagnostikovat porušení toho kterého z předpokladů, na nichž je lineární model založen.

Při diagnostice *nesprávného tvaru závislosti* jsou užitečné diagramy znázorňující body $[\hat{Y}_i, Y_i]$, $[\hat{Y}_i, u_i]$, $[x_{ij}, u_i]$ pro nezávisle proměnné, které jsou v matici $\mathbf{X}$ nebo body $[z_{ij}, u_i]$ pro potenciální nezávisle proměnné, které v matici $\mathbf{X}$ zahrnuty nejsou. Velmi používaná jsou také parciální rezidua $\mathbf{u}^{(\bullet j)}$ pro jednotlivé nezávisle proměnné z matice $\mathbf{X}$.

Při diagnostice *nekonstantního rozptylu* jsou užitečné diagramy pro $[\hat{Y}_i, u_i]$, $[\hat{Y}_i, u_i^2]$ nebo pro $[x_{ij}, u_i]$ resp. $[x_{ij}, u_i^2]$ pro v regresní matici $\mathbf{X}$ uplatněné či $[z_{ij}, u_i]$ resp. $[z_{ij}, u_i^2]$ pro neuplatněné nezávisle proměnné.

Při diagnostice *nenormálního rozdělení* chybového členu se používá zejména *normální diagram*, který znázorňuje $[g_i, u_{(i)}]$, případně $[u_{(i)}, g_i]$. Při tom je $g_i = \mathsf{E}\, Z_{(i)}$, kde $Z_1, \ldots, Z_n$ je náhodný výběr z rozdělení $\mathsf{N}(0, 1)$. Závorky u indexů tentokrát klasicky odkazují na to, že rezidua jsou uspořádaná.

Hodnocení je založeno na představě, že kdyby byl $U_1, \ldots, U_n$ náhodný výběr z rozdělení $\mathsf{N}\big(\mu, \sigma^2\big)$, platilo by $\mathsf{E}\, U_{(i)} = \mu + \sigma g_i$. To znamená, že body $[g_i, U_{(i)}]$ by měly náhodně kolísat kolem přímky $y = \mu + \sigma x$. Pokud body $[g_i, U_{(i)}]$ naznačují konkávní závislost, je to známka záporné šikmosti rozdělení náhodné veličiny U (tedy její nenormality). Konvexní průběh je známkou kladné šikmosti. Naproti tomu esovitý průběh naznačuje špičatost jinou, než předpokládáme u normálního rozdělení. Menší, než průměrný růst v okrajových částech naznačuje špičatost spíš menší, kdežto větší růst v okrajových částech naznačuje spíš větší špičatost.

Uvedený postup se používá pro rezidua $u_1, \ldots, u_n$ přesto, že ta nejsou nezávislá a obecně nemají stejný rozptyl. Upozorňuji na to, že některé programy (například STATISTICA) zaměňují pořadí obou os. Potom musíme odpovídajícím způsobem upravit také interpretaci normálního diagramu.

9. Testy

Na rozdíl od poslední části předchozí kapitoly se budeme zabývat možnostmi ověřovat splnění předpokladů lineární regrese statistickými testy, nikoliv jen možností jejich nesplnění dodatečně diagnostikovat.

9.1. Tvar závislosti

9.1.1. Opakovaná pozorování

Podstatným (a často nesplnitelným) požadavkem pro řadu testů je to, že pro stejnou hodnotu všech nezávisle proměnných máme několik pozorování. Tomu také přizpůsobíme označení. Mějme tedy n *nezávislých* náhodných veličin, které splňují

$$Y_{ij} = \mu_i + e_{ij}, \qquad 1 \leq j \leq n_i, \quad 1 \leq i \leq I, \tag{9.1}$$

kde e_{ij} jsou nezávislé náhodné veličiny s rozdělením $\mathsf{N}(0, \sigma^2)$. Jde vlastně o model analýzy rozptylu jednoduchého třídění. Jak víme, reziduální součet čtverců je v tomto modelu roven

$$RSS = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2 \tag{9.2}$$

a má celkem $f = n - I$ stupňů volnosti.

Pro testování zvoleného tvaru závislosti uvedeme zobecnění postupu, který je uveden v IX. kapitole knihy prof. Anděla (1978). Předpokládaný tvar závislosti udává podmodel

$$Y_{ij} = \sum_{\ell=1}^{L} g_\ell(\mathbf{t}_i)\gamma_\ell + e_{ij}, \qquad 1 \leq j \leq n_i, \quad 1 \leq i \leq I. \tag{9.3}$$

Přitom $g_\ell(\mathbf{t})$ jsou pro $\ell = 1, \dots L$, $L < I$, známé funkce, jejichž argumentem je vektor nezávisle proměnných. Funkční hodnoty lze nazývat pro odlišení jako *regresory.*

Několik regresorů (např. mocnin) lze získat z jediné nezávisle proměnné. Maticový zápis podmodelu má tvar

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_I \end{pmatrix} = \begin{pmatrix} \mathbf{1}(\mathbf{g}(\mathbf{t}_1))' \\ \vdots \\ \mathbf{1}(\mathbf{g}(\mathbf{t}_I))' \end{pmatrix} \gamma + \mathbf{e}.$$

Je zřejmé, že sloupce regresní matice podmodelu jsou lineární kombinací sloupců matice modelu, koeficienty příslušných lineárních kombinací tvoří hodnoty $g_\ell(\mathbf{t}_i)$. Předpokládejme, že matice

$$\begin{pmatrix} \mathbf{g}(\mathbf{t}_1)' \\ \vdots \\ \mathbf{g}(\mathbf{t}_I)' \end{pmatrix}$$

má lineárně nezávislé sloupce, tedy hodnost L. Stejnou hodnost má také regresní matice podmodelu. Test podmodelu je podle (3.10) založen na statistice

$$F = \frac{(RSS_0 - RSS)/(I - L)}{S^2}, \tag{9.4}$$

kde RSS_0 je reziduální součet čtverců v podmodelu.

Uvedený postup je velmi účinný, ale hrozí nebezpečí nesprávného použití v případě, že pozorování pro pevné $\mathbf{t}_i$ (pevné i) nejsou nezávislá. Potom snadno dá použitý model velmi podhodnocený odhad rozptylu σ^2 a tudíž nadhodnocenou hodnotu statistiky F.

Příklad 9.1 (brzdná dráha) Zajímáme se o brzdnou dráhu 63 automobilů v závislosti na výchozí rychlosti. K disposici je celkem $n = 63$ měření, přičemž pro většinu z $I = 29$ různých výchozích rychlostí máme k disposici více než jedno pozorování.

Pro model lineární závislosti veličiny `draha/rychlost` na veličině `rychlost` provedeme test dobré shody podle (9.4):

```
> anova(a.ANOVA1<-lm(draha/rychlost~factor(rychlost)))
Analysis of Variance Table

Response: draha/rychlost
                 Df  Sum Sq Mean Sq F value    Pr(>F)
factor(rychlost) 28 25.7720  0.9204  4.0678 7.096e-05 ***
Residuals        34  7.6932  0.2263
---
> anova(a.kvadrat<-lm(draha/rychlost~rychlost))
Analysis of Variance Table

Response: draha/rychlost
          Df  Sum Sq Mean Sq F value    Pr(>F)
rychlost   1 21.1640 21.1640  104.95 6.994e-15 ***
```

```
Residuals 61 12.3012  0.2017
---
> anova(a.kvadrat,a.ANOVA1)
Analysis of Variance Table

Model 1: draha/rychlost ~ rychlost
Model 2: draha/rychlost ~ factor(rychlost)
  Res.Df Res.Sum Sq Df  Sum Sq F value Pr(>F)
1     61    12.3012
2     34     7.6932 27  4.6080  0.7543 0.7728
>
```

Výsledná testová statistika $F = 0{,}7543$ s dosaženou hladinou $p = 0{,}7728$ nikterak nesvědčí proti předpokládané závislosti. ○

9.1.2. Testy o parametru

Typickou situací je model

$$Y_i = (\mathbf{x}_{i\bullet})'\boldsymbol{\beta} + \gamma g(\mathbf{x}_{i\bullet}) + e_i, \tag{9.5}$$

kde $g(\mathbf{x})$ je nějaká známá funkce. Testujeme pak nulovou hypotézu $\gamma = 0$. Nejčastěji je $g(\mathbf{x})$ funkcí jediné složky vektoru $\mathbf{x}$. Pokud funkci $g(\mathbf{x})$ neznáme, volíme nějakou aproximaci, například polynom. Tento postup je účinný zvláště tehdy, když je skutečná funkce $g(\mathbf{x})$ konvexní nebo konkávní funkcí pouze skalárního x.

Příklad 9.2 (kořeny) Vraťme se k příkladu o závislosti hmotnosti kořenové části rostliny na obsahu cukru v živném roztoku. Tentokrát se zajímáme o závislost na podílu cukru v živném roztoku (vyjádřeném v procentech). Porovnáme závislost kvadratickou a lineární.

```
> summary(a<-lm(hmotnost~procento+I(procento**2)))

Call:
lm(formula = hmotnost ~ procento + I(procento^2))

Residuals:
       Min         1Q     Median         3Q        Max
-0.1410511 -0.0352009 -0.0006059  0.0508703  0.1219806

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.218106   0.015640  13.945  < 2e-16 ***
procento       0.111677   0.012900   8.657 1.38e-11 ***
I(procento^2) -0.018610   0.002119  -8.784 8.85e-12 ***
---
```

```
Residual standard error: 0.06197 on 51 degrees of freedom
Multiple R-Squared: 0.6044,     Adjusted R-squared: 0.5889
F-statistic: 38.97 on 2 and 51 degrees of freedom,   p-value: 5.355e-011
```

Závěr je nepochybný, bez kvadratického členu (nebo jiného konkávního) se neobejdeme. ○

9.1.3. Použití rekurzivních reziduí

Harvey a Collier (1977) navrhli použít rekurzivní rezidua k ověřování linearity závislosti na zvolené nezávisle proměnné proti alternativě, že je tato závislost konvexní či konkávní, tento test nazvali *ψ-test*.

Předem je třeba pozorování uspořádat tak, zmíněná nezávisle proměnná, řekněme j-tá, splňovala požadavek $x_{1j} \leq x_{2j} \leq \ldots \leq x_{nj}$. Pokud je skutečná závislost na j-té nezávisle proměnné například konvexní, pak lze očekávat, že rekurzivní rezidua budou spíše kladná. Testová statistika tedy spočívá v testování nulové hypotézy, že střední hodnota rekurzivních reziduí je nulová.

V knihovně `lmtest` prostředí R je tento test uveden jako funkce `harvtest()`.

9.1.4. Durbinův-Watsonův test

Durbinův-Watsonův (viz oddíl 9.4) test je původně určen k testování hypotézy o nezávislosti jednotlivých pozorování. Testová statistika je citlivá při testování nulové hypotézy $\mathsf{H}_0 : \gamma = 0$ v modelu (9.5), když je funkce $g(\mathbf{x})$ konvexní nebo konkávní funkcí některé složky $\mathbf{x}$. K smysluplnému použití je však třeba, aby funkční hodnoty x_i byly monotonní vůči pořadí pozorování i.

V knihovně `lmtest` prostředí R je tento test uveden jako funkce `dwtest()`.

9.1.5. Chowův test

Následující postup (viz například (Anděl, 1998, kap. 12.5)) lze použít v mnoha variantách, vždy jde o efektivní použití umělých proměnných.

Základní myšlenkou testu je ověřit stabilitu parametru $\boldsymbol{\beta}$, jeho případnou závislost na nějaké doprovodné veličině. Data rozdělíme na dvě až tři disjunktní podmnožiny dat. Dělení provedeme tak, aby ve skupině I byly velké hodnoty této doprovodné proměnné, ve skupině II naopak její malé hodnoty. Zbývající skupina III obsahuje pozorování s „prostředními“ hodnotami doprovodné veličiny, může být i prázdná. Odhadneme stejnou regresní závislost ve skupinách I a II. Statistiky vztažené k jednotlivým skupinám označíme příslušným indexem. Pro jednoduchost předpokládejme, že ve skupinách I a II má regresní matice úplnou hodnost rovnou $k + 1$.

Dál pracujeme se skupinami I a II buď jednotlivě (model) nebo spojenými (podmodel). Reziduální součet čtverců v modelu bude $RSS = RSS_I + RSS_{II}$. Použijeme-li data z obou skupin dohromady a odhadneme parametry, dostaneme výsledný reziduální součet čtverců v podmodelu RSS_0. Testujeme tak nulovou hypotézu, že parametry v obou částech dat jsou totožné.

Rozhodujeme pomocí statistiky

$$F = \frac{RSS_0 - (RSS_I + RSS_{II})}{RSS_I + RSS_{II}} \frac{n_I + n_{II} - 2k - 2}{k+1},$$

která má na platnosti nulové hypotézy rozdělení $\mathsf{F}(k+1, n_I + n_{II} - 2k - 2)$.

9.2. Rozptyl

V tomto oddílu se budeme zabývat ověřováním předpokladu *homoskedasticity*, tedy předpokladu konstantního rozptylu závisle proměnné. Když uvedený předpoklad není splněn, nastává *heteroskedasticita*.

9.2.1. Opakovaná pozorování

Předpokládejme opět, že platí model (9.1), tentokrát je však $e_{ij} \sim \mathsf{N}(0, \sigma_i^2)$. Znamená to tedy, že připouštíme jakoukoliv regresní funkci s libovolnými parametry. Je třeba rozhodnout o shodě všech rozptylů σ_i^2, tedy o nulové hypotéze $\mathsf{H}_0 : \sigma_1^2 = \ldots = \sigma_k^2 (= \sigma^2)$.

Řada použitelných testů je pomocí simulací porovnána v článku Conover et al. (1981). Uveďme nejprve klasický *Bartlettův test*, který je modifikací testu poměrem věrohodnosti. Označme odhady rozptylu pro jednotlivé střední hodnoty závisle proměnné symbolem

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2.$$

Odhadem společné hodnoty rozptylů σ^2 je reziduální rozptyl v modelu

$$S^2 = \frac{1}{n-I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2 = \sum_{i=1}^{I} \frac{n_i - 1}{n - I} S_i^2,$$

což je nepochybně vážený průměr odhadů jednotlivých odhadů s vahami

$(n_i - 1)/(n - I)$. Testová statistika Bartlettova testu má tvar

$$B = \frac{1}{C}\left((n-I)\log S^2 - \sum_{i=1}^{I}(n_i-1)\log S_i^2\right) \tag{9.6}$$
$$= \frac{n-I}{C}\left(\log S^2 - \sum_{i=1}^{I}\frac{n_i-1}{n-I}\log S_i^2\right).$$

Je zřejmé, že test je založen na porovnání logaritmu váženého průměru odhadů rozptylu pro jednotlivá i s váženým průměrem logaritmů těchto odhadů. Konstanta C je dána vztahem

$$C = 1 + \frac{1}{3(I-1)}\left(\sum_{i=1}^{I}\frac{1}{n_i-1} - \frac{1}{n-I}\right),$$

je zpravidla jen nepatrně větší než 1.

Rozdělení statistiky B lze za platnosti nulové hypotézy při dostatečně velkých četnostech aproximovat rozdělením $\chi^2(I-1)$. Udává se, že tuto vlastnost lze použít, platí-li pro všechna i nerovnost $n_i \geq 7$. Nulovou hypotézu pak zamítáme, je-li $B \geq \chi^2_{I-1}(\alpha)$.

Vážnou nevýhodnou Bartlettova testu je jeho velká citlivost na případné porušení předpokladu o normálním rozdělení. V knihovně `ctest` je prostředí R vedle Bartlettova testu (`bartlett.test`) implementován také test Flignerův-Killeenův (`fligner.test`)), který je robustní vůči porušení předpokladu normality. Postup vychází z uspořádaných hodnot $|Y_{it} - \tilde{Y}_{i\bullet}|$, kde $\tilde{Y}_{i\bullet}$ je medián $Y_{i1}, \ldots, Y_{in_i}$. Takto získáme celkem n veličin, které uspořádáme. Nechť R_{it} je pořadí $|Y_{it} - \tilde{Y}_{i\bullet}|$. Veličiny

$$a_{it} = \Phi^{-1}(1/2 + (R_{it}/2(n+1)))$$

se zpracují podobně, jako samotná pořadí v Kruskalově-Wallisově testu. Použije se tedy statistika

$$Q = \frac{\sum_{i=1}^{I}\left(\sum_{t=1}^{n_i} a_{it}\right)^2/n_i - n\left(\bar{a}\right)^2}{v_a},$$

kde v_a je výběrový rozptyl hodnot a_{it}. Za platnosti nulové hypotézy (rozptyly jsou shodné) má statistika Q asymptoticky rozdělení $\chi^2(I-1)$.

Příklad 9.3 (kořeny)

```
> bartlett.test(hmotnost,procentoF)

         Bartlett test for homogeneity of variances

data:  hmotnost and procentoF
Bartlett's K-square = 2.872, df = 3, p-value = 0.4118
```

```
> fligner.test(hmotnost,procentoF)

        Fligner-Killeen test for homogeneity of variances

data:  hmotnost and procentoF
Fligner-Killeen:med chi-square = 2.6522, df = 3, p-value = 0.4484
```

Je patrné, že homoskedasticitu můžeme předpokládat. ○

9.2.2. Leveneův test

V poslední době je Bartlettův test nahrazován postupem, který navrhl Levene.

Základní myšlenkou je vlastnost normálního rozdělení, kterou pro naše nezávislé náhodné veličiny Y_{ij} s rozdělením $\mathsf{N}(\mu_i, \sigma_i^2)$ můžeme zapsat jako

$$\mathsf{E}\,|Y_{ij} - \mu_i| = \sqrt{\frac{2}{\pi}}\sigma_i.$$

Spočítají se pomocné veličiny $Y_{ij}^* = |Y_{ij} - \bar{Y}_{i\bullet}|$ a potom se s nimi provede běžná analýza rozptylu jednoduchého třídění. Nulovou hypotézu, podle které jsou rozptyly σ_i^2 stejné, tedy zamítneme, když klasická F statistika vyjde významná.

Někdy se používá (například NCSS) modifikace, kterou navrhli Brown a Forsythe, místo s Y_{ij}^* s veličinami $Y_{ij}^{**} = |Y_{ij} - \tilde{Y}_{i\bullet}|$, kde $\tilde{Y}_{i\bullet}$ je opět medián veličin $Y_{i1}, \ldots, Y_{in_i}$.

Příklad 9.4 (kořeny) Veličiny `hmotnost.1` a `hmotnost.2` obsahují hodnoty závisle proměnné zmenšené o průměr (medián) zjištěný v dané skupině.

```
> hmotnost.mean <- unlist(tapply(hmotnost,Procento,function(u) u-mean(u)))
> anova(lm(abs(hmotnost.mean)~Procento))
Analysis of Variance Table

Response: abs(hmotnost.mean)
          Df   Sum Sq  Mean Sq F value Pr(>F)
Procento   3 0.003552 0.001184  0.9306 0.4329
Residuals 50 0.063613 0.001272
>
> hmotnost.med <- unlist(tapply(hmotnost,Procento,function(u) u-median(u)))
> anova(lm(abs(hmotnost.med)~Procento))
Analysis of Variance Table

Response: abs(hmotnost.med)
          Df   Sum Sq  Mean Sq F value Pr(>F)
Procento   3 0.003652 0.001217  0.8302 0.4836
Residuals 50 0.073319 0.001466
```

Je zřejmé, že žádná z variant Leveneova testu neukazuje na heteroskedasticitu. ○

9.2.3. Goldfeldův-Quandtův test

Tento postup je v mnohém podobný Chowovu testu.

Testujeme nulovou hypotézu, podle které je rozptyl Y_{ij} konstantní proti alternativní hypotéze, že rozptyl je monotonní funkcí pořadového indexu. Má-li být monotonní funkcí nějaké doprovodné veličiny, musíme nejprve data příslušným způsobem uspořádat.

Postup je založen na porovnání dvou nezávislých odhadů rozptylu. Nejprve vydělíme asi třetinu pozorování s malými indexy a zde provedeme odhad parametrů stejného lineárního modelu, jako jsme použili pro všechna data. Zejména spočítáme odhad rozptylu S_I^2. Podobně odhadneme rozptyl z poslední třetiny dat, takto získáme odhad S_{II}^2. Za platnosti nulové hypotézy má statistika $F = S_I^2/S_{II}^2$ rozdělení $\mathsf{F}(n_I - r_I, n_{II} - r_{II})$.

Goldfeldův-Quandtův test lze považovat za zobecnění klasického F testu shody rozptylů, jen poněkud jinak získáme dva nezávislé odhady rozptylu.

9.2.4. Obecný model

Nejprve popíšeme poměrně obecný model pro nekonstantní rozptyl, v dalších oddílech jej konkretizujeme na důležité speciální případy. Postup je založen na metodě maximální věrohodnosti a to na použití skórů.

Uvažujme model (speciální případ modelu z oddílu 2.8)

$$\mathbf{Y} \sim \mathsf{N}\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{W}^{-1}\big), \tag{9.7}$$

kde $\mathbf{W}$ je diagonální matice s diagonálními prvky w_i, přičemž

$$w_i^{-1} = \omega_i = \omega_i(\boldsymbol{\beta}, \boldsymbol{\lambda}). \tag{9.8}$$

Připouštíme tedy, že prostřednictvím známých funkcí ω_i může rozptyl záviset na neznámém parametru $\boldsymbol{\beta}$ (který slouží k popisu středních hodnot) a na nějakém dalším parametru $\boldsymbol{\lambda}$. Pro stručnost zápisu budeme v dalším někdy argumenty funkcí ω_i vynechávat. Věrohodnostní funkci modelu (9.7) lze zapsat jako

$$\ell(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\log\omega_i - \frac{1}{2}\sum_{i=1}^{n}\frac{(Y_i - (\mathbf{x}_{i\bullet})'\boldsymbol{\beta})^2}{\sigma^2\omega_i}.$$

Odtud plyne (po úpravě a s označením $e_i = Y_i - (\mathbf{x}_{i\bullet})'\boldsymbol{\beta}$)

$$\begin{aligned}
\frac{\partial\ell}{\partial\boldsymbol{\beta}} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{e_i}{\omega_i}\mathbf{x}_{i\bullet} + \frac{1}{2}\sum_{i=1}^{n}\left(\left(\frac{e_i}{\sigma\sqrt{\omega_i}}\right)^2 - 1\right)\frac{\partial\log\omega_i}{\partial\boldsymbol{\beta}},\\
\frac{\partial\ell}{\partial\sigma^2} &= \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\left(\frac{e_i}{\sigma\sqrt{\omega_i}}\right)^2 - 1\right),\\
\frac{\partial\ell}{\partial\boldsymbol{\lambda}} &= \frac{1}{2}\sum_{i=1}^{n}\left(\left(\frac{e_i}{\sigma\sqrt{\omega_i}}\right)^2 - 1\right)\frac{\partial\log\omega_i}{\partial\boldsymbol{\lambda}}.
\end{aligned}$$

Označíme-li symbolem $\mathbf{D}_\beta$ matici typu $n \times (k+1)$ parciálních derivací $\partial \log \omega_i / \partial \beta_j$ a podobně symbolem $\mathbf{D}_\lambda$ matici parciálních derivací $\partial \log \omega_i / \partial \lambda_j$ a uvážíme-li, že platí $(1 \le i, j \le n)$

$$\mathsf{E}\, e_i e_j = \delta_{ij}\sigma^2\omega_i$$

$$\mathsf{E}\, e_i \left(\left(\frac{e_j}{\sigma\sqrt{\omega_j}} \right)^2 - 1 \right) = 0$$

$$\mathsf{E} \left(\left(\frac{e_i}{\sigma\sqrt{\omega_i}} \right)^2 - 1 \right) \left(\left(\frac{e_j}{\sigma\sqrt{\omega_j}} \right)^2 - 1 \right) = 2\delta_{ij},$$

bude výsledná Fisherova informační matice rovna

$$\mathbf{J}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda}) = \mathsf{E} \begin{pmatrix} \dfrac{\partial \ell}{\partial \boldsymbol{\beta}} \dfrac{\partial \ell}{\partial \boldsymbol{\beta}'} & \dfrac{\partial \ell}{\partial \boldsymbol{\beta}} \dfrac{\partial \ell}{\partial \sigma^2} & \dfrac{\partial \ell}{\partial \boldsymbol{\beta}} \dfrac{\partial \ell}{\partial \boldsymbol{\lambda}'} \\ \dfrac{\partial \ell}{\partial \sigma^2} \dfrac{\partial \ell}{\partial \boldsymbol{\beta}'} & \dfrac{\partial \ell}{\partial \sigma^2} \dfrac{\partial \ell}{\partial \sigma^2} & \dfrac{\partial \ell}{\partial \sigma^2} \dfrac{\partial \ell}{\partial \boldsymbol{\lambda}'} \\ \dfrac{\partial \ell}{\partial \boldsymbol{\lambda}} \dfrac{\partial \ell}{\partial \boldsymbol{\beta}'} & \dfrac{\partial \ell}{\partial \boldsymbol{\lambda}} \dfrac{\partial \ell}{\partial \sigma^2} & \dfrac{\partial \ell}{\partial \boldsymbol{\lambda}} \dfrac{\partial \ell}{\partial \boldsymbol{\lambda}'} \end{pmatrix}$$

$$= \begin{pmatrix} \dfrac{1}{\sigma^2}\mathbf{X}'\mathbf{W}\mathbf{X} + \dfrac{1}{2}\mathbf{D}_\beta'\mathbf{D}_\beta & \dfrac{1}{2\sigma^2}\mathbf{D}_\beta'\mathbf{1} & \frac{1}{2}\mathbf{D}_\beta'\mathbf{D}_\lambda \\ \dfrac{1}{2\sigma^2}\mathbf{1}'\mathbf{D}_\beta & \dfrac{n}{2\sigma^4} & \dfrac{1}{2\sigma^2}\mathbf{1}'\mathbf{D}_\lambda \\ \dfrac{1}{2}\mathbf{D}_\lambda'\mathbf{D}_\beta & \dfrac{1}{2\sigma^2}\mathbf{D}_\lambda'\mathbf{1} & \dfrac{1}{2}\mathbf{D}_\lambda'\mathbf{D}_\lambda. \end{pmatrix} \tag{9.9}$$

Testová statistika je podle (A.34) rovna kvadratické formě

$$\left(\frac{\partial \ell}{\partial \boldsymbol{\beta}} \quad \frac{\partial \ell}{\partial \sigma^2} \quad \frac{\partial \ell}{\partial \lambda} \right)_{\tilde\beta, \tilde{\sigma^2}, \tilde\lambda} \left(\mathbf{J}(\tilde\beta, \tilde{\sigma^2}, \tilde\lambda) \right)^{-1} \left(\frac{\partial \ell}{\partial \boldsymbol{\beta}} \quad \frac{\partial \ell}{\partial \sigma^2} \quad \frac{\partial \ell}{\partial \lambda} \right)'_{\tilde\beta, \tilde{\sigma^2}, \tilde\lambda}.$$

9.2.5. Závislost na střední hodnotě

Velmi častým případem porušení předpokladu o konstantním rozptylu (tedy případem *heteroskedasticity*) je monotonní závislost rozptylu na střední hodnotě Y. Odvodíme testovou statistiku, která je založena na metodě skórů (viz Appendix A.3).

Předpokládejme, že je $\omega_i = \exp(\lambda(\mathbf{x}_{i\bullet})'\boldsymbol{\beta})$. Potom je $\mathbf{D}_\beta = \lambda\mathbf{X}$ a $\mathbf{D}_\lambda = \mathbf{X}\boldsymbol{\beta}$. Konstantní rozptyly (homoskedasticitu) zaručí nulová hypotéza $\mathsf{H}_0 : \lambda = 0$. Za platnosti H_0 je tedy $\mathbf{D}_\beta = \mathbf{O}$ a $\mathbf{D}_\lambda = \mathbf{X}\boldsymbol{\beta}$. Odtud je informační matice rovna

$$\mathbf{J}(\boldsymbol{\beta}, \sigma^2, 0) = \begin{pmatrix} \dfrac{1}{\sigma^2}\mathbf{X}'\mathbf{X} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & \dfrac{n}{2\sigma^4} & \dfrac{1}{2\sigma^2}\mathbf{1}'\mathbf{X}\boldsymbol{\beta} \\ \mathbf{0}' & \dfrac{1}{2\sigma^2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{1} & \dfrac{1}{2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{pmatrix}.$$

Když počítáme odhady metodou maximální věrohodnosti za nulové hypotézy, dostaneme $\tilde{\boldsymbol{\beta}} = \mathbf{b}$, $\tilde{\sigma^2} = RSS/n$ a samozřejmě $\tilde{\lambda} = 0$. Odtud vyjde

$$\begin{pmatrix} \frac{\partial \ell}{\partial \boldsymbol{\beta}} \\ \frac{\partial \ell}{\partial \sigma^2} \\ \frac{\partial \ell}{\partial \lambda} \end{pmatrix}_{\tilde{\beta},\tilde{\sigma^2},\tilde{\lambda}} = \begin{pmatrix} \mathbf{0} \\ 0 \\ \frac{1}{2\tilde{\sigma^2}} \sum_{i=1}^{n} (u_i^2 - \tilde{\sigma^2}) \hat{Y}_i \end{pmatrix}.$$

Když ještě vezmeme v úvahu, že odhad $\tilde{\sigma^2}$ je průměrem hodnot u_i^2 a když označíme průměrnou hodnotu z $\hat{Y}_i$ symbolem $\bar{\hat{Y}}$, můžeme jediný obecně nenulový prvek vektoru parciálních derivací logaritmické věrohodnostní funkce zapsat také jako

$$\frac{1}{2\tilde{\sigma^2}} \sum_{i=1}^{n} u_i^2 (\hat{Y}_i - \bar{\hat{Y}}).$$

Když také do Fisherovy informační matice dosadíme odhady za nulové hypotézy a výsledek dosadíme do (A.34), po úpravě (nezapomeňte invertovat matici $\mathbf{J}(\mathbf{b}, \tilde{\sigma^2}, 0)$) dostaneme statistiku

$$S_f = \frac{\left(\sum_{i=1}^{n} u_i^2 (\hat{Y}_i - \bar{\hat{Y}})\right)^2}{2\left(\tilde{\sigma^2}\right)^2 \sum_{i=1}^{n} (\hat{Y}_i - \bar{\hat{Y}})^2}. \tag{9.10}$$

Podle obecné teorie by za platnosti nulové hypotézy měla mít statistika S_f asymptoticky rozdělení $\chi^2(1)$.

Pokusme se nalezenou statistiku nějak názorně interpretovat. Až na dvojnásobek čtverce odhadu rozptylu $2\left(\tilde{\sigma^2}\right)^2$ je statistika S_f formálně rovna regresnímu součtu čtverců u lineární závislosti u_i^2 na $\hat{Y}_i$. Nebo jinak, je to polovina regresního součtu čtverců závislosti $u_i^2/\hat{\sigma^2}$ na $\hat{Y}_i$. Uvážíme-li, že v této pomocné úvaze statistika u_i^2 nahrazuje veličinu e_i^2, která má rozptyl $2\sigma^4$, můžeme považovat výraz $2(\tilde{\sigma^2})^2$ za odhad tohoto rozptylu. Statistika S_f tedy vypovídá o nulovosti směrnice regresní přímky závislosti u_i^2 na $\hat{Y}_i$.

Na místě je tedy zjednodušená varianta statistiky S_f, totiž čtverec testové t statistiky k testu hypotézy o nulové směrnici v uvažované pomocné regresní úloze.

Program R obsahuje popsaný test v knihovně `car` jako funkci `ncv.test()`.

Příklad 9.5 (brzdná dráha)

```
> summary(a<-lm(draha~rychlost+I(rychlost^2),data=Draha))

Call:
lm(formula = draha ~ rychlost + I(rychlost^2), data = Draha)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-22.499  -5.468  -0.425   3.932  28.106

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.83871    5.06085   0.363    0.718
rychlost       0.36935    0.54943   0.672    0.504
I(rychlost^2)  0.06664    0.01287   5.177 2.76e-06

Residual standard error: 9.891 on 60 degrees of freedom
Multiple R-Squared: 0.9137,     Adjusted R-squared: 0.9108
F-statistic: 317.7 on 2 and 60 DF,  p-value: < 2.2e-16

> ncv.test(a)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 23.08760    Df = 1     p = 1.547860e-06
```

Výsledek bylo lze očekávat, když si prohlédneme závislost reziduí na vyrovnaných hodnotách znázorněnou na obrázku 9.1. Ještě nahoře zmíněná přibližná varianta testu:

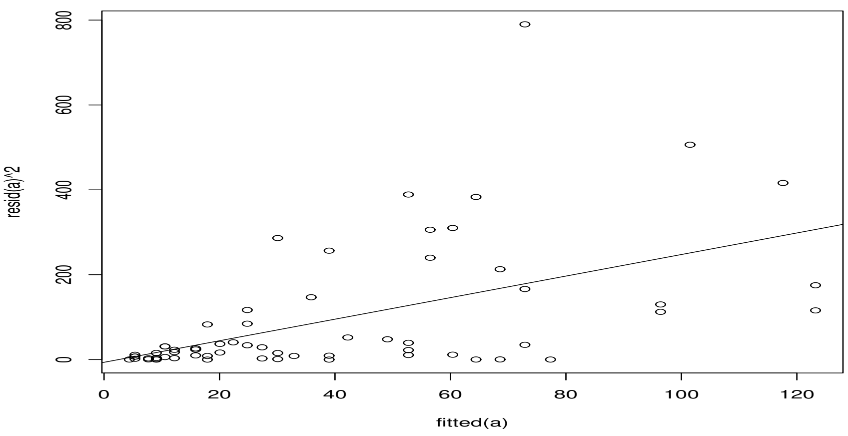
```
> anova(lm(resid(aa)^2~fitted(aa)))
Analysis of Variance Table

Response: resid(aa)^2
          Df  Sum Sq Mean Sq F value    Pr(>F)
fitted(aa)  1  400923  400923  24.133 7.077e-06 ***
Residuals 61 1013399   16613
```

○

9.2.6. Závislost na doprovodných veličinách

Předpokládejme nyní, že heteroskedasticita je způsobena monotonní závislostí rozptylu na lineární kombinaci nějakých doprovodných veličin, mezi něž mohou patřit i některé použité regresory.

Předpokládejme, že je $\omega_i = \exp(\boldsymbol{\lambda}'\mathbf{z}_{i\bullet})$, kde $\mathbf{z}_{i\bullet}$ je i-tý řádek matice známých konstant s lineárně nezávislými sloupci $\mathbf{Z}$. Pro matice derivací evidentně platí $\mathbf{D}_\beta = \mathbf{O}$ a $\mathbf{D}_\lambda = \mathbf{Z}$, a to ať už nulová hypotéza $\mathsf{H}_0 : \boldsymbol{\lambda} = \mathbf{0}$ platí nebo neplatí. Vektor parciálních derivací věrohodnostní funkce má za platnosti nulové hypotézy (po dosazení odhadů za nulové hypotézy) opět první dva bloky nulové. Nenulová je pouze derivace $\partial\ell/\partial\boldsymbol{\lambda}$. Po dosazení zmíněných odhadů dostaneme podobně jako



Obrázek 9.1: Závislost reziduí na vyhlazených hodnotách v modelu kvadratické závislosti brzdné dráhy na rychlosti

v předchozí kapitolce výraz

$$\frac{\partial \ell}{\partial \boldsymbol{\lambda}} = \frac{1}{2\tilde{\sigma^2}} \sum_{i=1}^{n} u_i^2 (\mathbf{z}_{i\bullet} - \bar{\mathbf{z}}).$$

Odpovídající prvek inverzní matice k Fisherově informační matici je inverzní matice k matici

$$\frac{1}{2}(\mathbf{Z} - \mathbf{1}\bar{\mathbf{z}}')'(\mathbf{Z} - \mathbf{1}\bar{\mathbf{z}}'),$$

takže výsledná statistika metody skórů typu (A.34) je

$$S_z = \frac{1}{2\left(\tilde{\sigma^2}\right)^2} \left(\sum_{i=1}^{n} u_i^2 (\mathbf{z}_{i\bullet} - \bar{\mathbf{z}}) \right)' \left((\mathbf{Z} - \mathbf{1}\bar{\mathbf{z}}')'(\mathbf{Z} - \mathbf{1}\bar{\mathbf{z}}') \right)^{-1} \left(\sum_{i=1}^{n} u_i^2 (\mathbf{z}_{i\bullet} - \bar{\mathbf{z}}) \right).$$

Platí-li nulová hypotéza (homoskedasticita), má statistika S_z asymptoticky rozdělení $\chi^2(q)$, kde q je počet složek vektoru λ.

Interpretace statistiky S_z je podobná, jako u S_f. Lze ji chápat jako míru těsnosti závislosti čtverců reziduí u_i^2 na nezávisle proměnných obsažených v matici $\mathbf{Z}$ (v modelu, který kromě nich obsahuje také absolutní člen). I zde si lze představit zjednodušenou variantu a k rozhodování použít tabulku analýzy rozptylu mnohonásobné regrese (s absolutním členem) čtverců reziduí na regresorech z matice $\mathbf{Z}$.

Samozřejmě, na místě doprovodných proměnných lze použít také některé nebo všechny nezávisle proměnné z matice modelu. Speciálně, když u regresní přímky budeme vyšetřovat závislost rozptylu na (jediné) nezávisle proměnné, musí vyjít přesně stejná testová statistika jako při testování závislosti na střední hodnotě, tedy $S_z = S_f$.

Také tato varianta testu homockedasticity je implemetována v R v knihovnách `car` (funkce `ncv.test()` s parametrem `var.formula`) a `lmtest` (funkce `bptest()` s parametrem `varformula`).

Příklad 9.6 (brzdná dráha)

```
> ncv.test(a,var.formula=~rychlost)
Non-constant Variance Score Test
Variance formula: ~ rychlost
Chisquare = 23.44439    Df = 1     p = 1.285769e-06
```

I tento výsledek bylo lze očekávat, když si prohlédneme závislost reziduí na vyrovnaných hodnotách znázorněnou na obrázku 9.1. ○

9.3. Normalita

V případě testování normality v lineárním modelu nastává zajímavá situace. Existují sice testové statistiky, jejichž rozdělení za platnosti nulové hypotézy (normálního rozdělení) bezpečně známe, ale takové testy mají slabou sílu. Mnohem užitečnější je aplikovat některé přibližné postupy, které použijí klasická rezidua u_i. Použití normovaných nebo studentizovaných reziduí vede ke snížení síly testu (viz např. diplomku Mgr. Štefka (1994)).

Často se používají *šikmost* a *špičatost*, vždy počítané z běžných reziduí. Velmi užitečné jsou transformace, které navrhl D'Agostino a které jsou použitelné pro poměrně malé počty pozorování. Transformovanou šikmost Z_3 lze použít již pro $n \geq 9$, transformovanou špičatost Z_4 již pro $n \geq 20$. Podrobně jsou transformace popsány například v Andělově (1998) knížce.

V kapitolce 8.7 jsme se již seznámili s *diagramem normality*, který znázorňuje body o souřadnicích $[g_i, u_{(i)}]$, kde g_i je střední hodnota i-té pořádkové statistiky prostého náhodného výběru z rozdělení $\mathsf{N}(0,1)$. Když předpokládáme běžný lineární model s absolutním členem, potom je součet reziduí nutně nulový, takže pak lze čtverec výběrového korelačního koeficientu psát jako

$$W' = \frac{\left(\sum_{i=1}^n g_i u_{(i)}\right)^2}{\sum_{i=1}^n g_i^2 \sum_{i=1}^n u_{(i)}^2}. \tag{9.11}$$

Gardiner (1997) uvádí přibližné kritické hodnoty pro výběrový korelační koeficient

$\sqrt{W'}$:

$$1{,}0063 - \frac{0{,}1288}{\sqrt{n}} - \frac{0{,}6118}{n} + \frac{1{,}3505}{n^2} \qquad \text{pro } \alpha = 5~\%,$$
$$1{,}0071 - \frac{0{,}1371}{\sqrt{n}} - \frac{0{,}3682}{n} + \frac{0{,}7780}{n^2} \qquad \text{pro } \alpha = 10~\%.$$

Postup založený na korelačním koeficientu $\sqrt{W'}$ bývá uváděn jako *Ryanův-Joinerův test*. Statistika W' je zjednodušenou alternativou k původní *statistice Shapira a Wilka*, která má tvar

$$W = \frac{1}{S^2}\left(\sum_{i=1}^{[n/2]} a_{i,n}(u_{(n-i+1)} - u_{(i)})\right)^2. \tag{9.12}$$

Koeficienty $a_{i,n}$ jsou odvozeny ze středních hodnot a varianční matice pořádkových statistik prostého náhodného výběru z $\mathsf{N}(0,1)$ rozsahu n. Spolu s kritickými hodnotami jsou tabelovány např. v knize Hahn, Shapiro (1967).

Uvedený test je v R součástí standardní knihovny `ctest` jako `shapiro.test`.

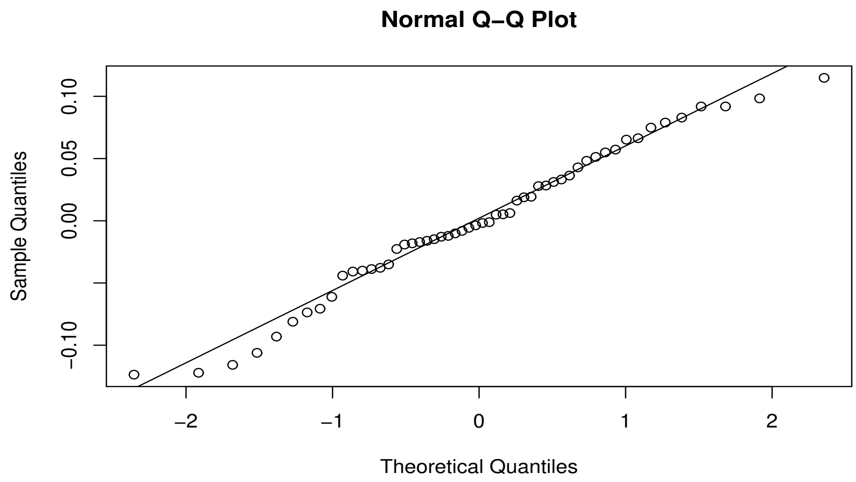
Příklad 9.7 (brzdná dráha)

```
> shapiro.test(resid(a))

        Shapiro-Wilk normality test

data:  resid(a)
W = 0.9744, p-value = 0.2126
```

○

Často se používá *test Kolmogorovův-Smirnovův*, který porovnává empirickou a teoretickou distribuční funkci. Protože jde o testování složené hypotézy (nulová hypotéza určuje pouze tvar rozdělení, nikoliv jeho parametry), je třeba pracovat s modifikací Kolmogorovova-Smirnovova testu, která známa jako *test Lillieforsův*. Rozdíl je pouze v použitých kritických hodnotách.

Pozor, dostupné programové vybavení je třeba používat opatrně. Jinak zajímavý program NCSS používá zmíněnou Lillieforsovu modifikaci automaticky a bez upozornění, kdežto Statistica udává dvojí hodnocení zjištěné statistiky Kolmogorova-Smirnova. Není mi známo, že by tuto modifikaci nabízel program R. Procedure `ks.test()` předpokládá u jednovýběrového testu nulovou hypotézou jednoznačně určenou distribuční funkci.



Obrázek 9.2: Normální diagram reziduí

Příklad 9.8 (kořeny) Opět se budeme věnovat známému příkladu. Začněme normálním diagramem reziduí (obrázek 9.2).

```
> u˜<- resid(lm(hmotnost˜procentoF,data=koren))
> shapiro.test(u)

         Shapiro-Wilk normality test

data:
u˜W = 0.9794, p-value = 0.476

> skewness.test(u)

         D'Agostino skewness normality test

data:
u˜Z3 = -0.7078, p-value = 0.4791

> kurtosis.test(u)

         D'Agostino kurtosis normality test

data:
u˜Z4 = -0.5144, p-value = 0.607

> omnibus.test(u)
```

```
        D'Agostino omnibus normality test

data:
u~Chi2 = 0.7656, df = 2, p-value = 0.682
```

Všechny použité testy naznačují totéž, co normální diagram. Není důvod nepředpokládat v modelu analýzy rozptylu normální rozdělení. Pilnému čtenáři doporučuji vyzkoušet si testy normality na stejných datech, ovšem v modelech lineární a kvadratické závislosti na obsahu cukru. ○

9.4. Nezávislost

Problém se stochastickou závislostí pozorování se vyskytuje zejména tehdy, když data získáváme postupně, takže hodnoty závisle proměnné vlastně tvoří časovou řadu. Každopádně musí mít pořadí pozorování nějaký význam, aby mělo smysl formálně se zabývat ověřováním předpokladu nezávislosti jednotlivých pozorování.

Mějme opět náhodné veličiny $Y_i = (\mathbf{x}_{i\bullet})'\boldsymbol{\beta} + e_i$, kde $e_i \sim \mathsf{N}\big(0, \sigma^2\big)$. Tentokrát připouštíme, že náhodné veličiny $e_1, \dots, e_n$ jsou závislé, speciálně, že tvoří autoregresní proces prvního řádu $e_i = \rho e_{i-1} + \epsilon_i$, v němž ϵ_i jsou již nezávislé. Pro $\rho = 0$ dostaneme klasický normální lineární model.

Statistika Durbina a Watsona má tvar

$$d = \frac{\sum_{i=1}^{n-1}(u_{i+1} - u_i)^2}{\sum_{i=1}^{n} u_i^2} = \frac{\mathbf{u}'\mathbf{A}\mathbf{u}}{\mathbf{u}'\mathbf{u}}, \tag{9.13}$$

kde matice

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}$$

je zřejmě symetrická a pozitivně semidefinitní (vyjadřuje nezápornou kvadratickou funkci z čitatele, součet řádků dá nulový vektor).

Zajímá nás rozdělení statistiky d za platnosti nulové hypotézy $\mathsf{H}_0 : \rho = 0$. Připomeňme, že je $\mathbf{u} = \mathbf{M}\mathbf{e}$. Přitom matici $\mathbf{M}$ lze vyjádřit pomocí mnohokrát použité ortonormální báze jako $\mathbf{M} = \mathbf{N}\mathbf{N}'$. Když zavedeme náhodný vektor

$$\mathbf{t} = \frac{1}{\sigma}\mathbf{N}'\mathbf{e} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{n-r}),$$

můžeme statistiku d přepsat jako

$$d = \frac{\mathbf{t}'\mathbf{N}'\mathbf{A}\mathbf{N}\mathbf{t}}{\mathbf{t}'\mathbf{t}}.$$

Nyní najdeme k pozitivně semidefinitní matici $\mathbf{N}'\mathbf{A}\mathbf{N}$ její spektrální rozklad $\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}'$, kde $\mathbf{Q}$ je nějaká ortonormální matice řádu $n-r$ a $\boldsymbol{\Lambda}$ je diagonální matice s diagonálními prvky $\lambda_1 \geq \ldots \geq \lambda_{n-r}$. Zaveďme nyní náhodný vektor $\mathbf{Z} = \mathbf{Q}'\mathbf{t}$. Snadno zjistíme, že je $\mathbf{Z} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{n-r})$, takže statistika

$$d = \frac{\mathbf{Z}'\boldsymbol{\Lambda}\mathbf{Z}}{\mathbf{Z}'\mathbf{Z}} = \frac{\sum_{i=1}^{n-r} \lambda_i Z_i^2}{\sum_{i=1}^{n-r} Z_i^2}$$

je podílem lineární kombinace náhodných veličin s rozdělením $\chi^2(1)$ a součtu těchto náhodných veličin.

Problémem je, že koeficienty lineární kombinace (konstanty λ_i) závisí na výchozí regresní matici $\mathbf{X}$. Naštěstí lze podle Poincarého věty (viz větu A.10 v Dodatku) tato vlastní čísla omezit pomocí vlastních čísel matice $\mathbf{A}$. Předpokládejme, že platí $\mathbf{1} \in \mathcal{M}(\mathbf{X})$ (například v modelu existuje absolutní člen). Potom platí $\mathbf{N1} = \mathbf{0}$ a protože je $\mathbf{1}$ vlastním vektorem matice $\mathbf{A}$ odpovídajícím jejímu nejmenšímu vlastnímu číslu, můžeme použít nerovnosti (A.22) a (A.24). Odtud bude (máme $q = n - r$)

$$d = \frac{\sum_{i=1}^{n-r} \lambda_i Z_i^2}{\sum_{i=1}^{n-r} Z_i^2} \leq \frac{\sum_{i=1}^{n-r} \alpha_i Z_i^2}{\sum_{i=1}^{n-r} Z_i^2} = d_U$$

a podobně (zvolme v (A.24) $j = n - r - i + 1$)

$$d = \frac{\sum_{j=1}^{n-r} \lambda_i Z_i^2}{\sum_{j=1}^{n-r} Z_i^2} \geq \frac{\sum_{j=1}^{n-r} \alpha_{j+r-1} Z_j^2}{\sum_{i=1}^{n-r} Z_j^2} = d_L.$$

Rozdělení náhodných veličin d_L, d_U závisí již pouze na n a r. Existují tabulky kritických hodnot pro náhodné veličiny d_L, d_U, např. Likeš, Laga (1978).

Při testování nulové hypotézy $\mathsf{H}_0 : \rho = 0$ proti alternativní hypotéze $\mathsf{H}_1 : \rho > 0$ pak ve prospěch alternativní hypotézy budou svědčit spíše malé hodnoty statistiky d (sousední rezidua jsou spíš podobná). Nulovou hypotézu zamítneme, když bude platit $d \leq d_L(\alpha)$, nezamítneme ji v případě, že vyjde $d > d_U(\alpha)$.

Ve zbývajících případech ($d_L(\alpha) < d \leq d_U(\alpha)$) rozhodnout takto snadno nelze. Pak je možno skutečné rozdělení statistiky $d/4$ aproximovat pomocí beta rozdělení s takovými parametry, aby se shodovaly první dva momenty. O možnostech aproximací rozdělení d pojednává podrobně přehledný článek autorů metody Durbin, Watson (1971). V poslední době se stále častěji používají k hodnocení statistiky d simulace. Výsledkem je pak přibližná dosažená hladina testu (p hodnota).

Snadno se zjistí, že statistika d těsně souvisí s odhadem koeficientu ρ: $d \doteq 2(1 - \hat{\rho})$. Při kladném parametru ρ mají body tendenci sdružovat se podle přímky $y = x$, při záporném ρ pak podle přímky $y = -x$.

K diagnostice problémů s nenulovým autokorelačním koeficientem ρ se používá diagram, který znázorňuje $n-1$ bodů $[u_{i-1}, u_i]$.

Předpokládejme, že data jsou uspořádána tak, že hodnoty nezávisle proměnné rostou s pořadovým indexem pozorování. Když se vyšetřuje kvadratická závislost na nezávisle proměnné a použije se pouze závislost lineární, výsledná sousední rezidua mají tendenci být si blízká, což je podobná situace, jako při kladném autokorelačním koeficientu ρ. Proto lze Durbinův-Watsonův test použít někdy také k diagnostice nesprávného tvaru regresní funkce.

10. Multikolinearita

Ve vlastní regresi se zpravidla předpokládá, že regresní matice $\mathbf{X}$ má lineárně nezávislé sloupce. Teoreticky matice má nebo nemá lineárně závislé sloupce. Ovšem u reálných matic je někdy obtížné rozhodnout, která z obou možností opravdu nastala.

O *multikolinearitě* tedy hovoříme tehdy, kdy matice $\mathbf{X}$ má sice lineárně nezávislé sloupce, ale v nějakém smyslu jsou tyto sloupce téměř lineárně závislé. O způsobem, jak multikolinearitu odhalit, pojednáme postupně.

10.1. Teorie

Nejprve uvedeme dvě důležité vlastnosti odhadů v lineárním modelu.

Věta 10.1. V modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ platí

$$\mathsf{E}\,||\hat{\mathbf{Y}}||^2 = ||\mathbf{X}\boldsymbol{\beta}||^2 + \sigma^2 \mathsf{h}(\mathbf{X}). \tag{10.1}$$

Má-li matice $\mathbf{X}$ lineárně nezávislé sloupce, pak platí

$$\mathsf{E}\,||\mathbf{b}||^2 = ||\boldsymbol{\beta}||^2 + \sigma^2\,\mathsf{tr}\,(\mathbf{X}'\mathbf{X})^{-1}. \tag{10.2}$$

D ů k a z: Výraz $\mathsf{E}\,||\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}||^2$ můžeme upravit dvěma způsoby. Jednak je to

$$\begin{aligned} \mathsf{E}\,(\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta})'(\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}) &= \mathsf{tr}\,\mathsf{E}\,(\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta})'(\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathsf{tr}\,\mathsf{var}\,\hat{\mathbf{Y}} \;=\; \sigma^2\,\mathsf{tr}\,\mathbf{H} \;=\; \sigma^2\mathsf{h}(\mathbf{X}), \end{aligned}$$

a také

$$\begin{aligned} \mathsf{E}\,||\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}||^2 &= \mathsf{E}\,||\hat{\mathbf{Y}}||^2 - 2\boldsymbol{\beta}'\mathbf{X}'\mathsf{E}\,\hat{\mathbf{Y}} + ||\mathbf{X}\boldsymbol{\beta}||^2 \\ &= \mathsf{E}\,||\hat{\mathbf{Y}}||^2 - ||\mathbf{X}\boldsymbol{\beta}||^2. \end{aligned}$$

Tvrzení (10.1) dostaneme porovnáním obou vyjádření. Druhé tvrzení věty dostaneme podobně, když dvěma způsoby vyjádříme výraz $\mathsf{E}\,||\mathbf{b}-\boldsymbol{\beta}||^2$:

$$\begin{aligned}\mathsf{E}\,||\mathbf{b}-\boldsymbol{\beta}||^2 &= \operatorname{tr}\operatorname{var}\mathbf{b} \;=\; \sigma^2\operatorname{tr}(\mathbf{X}'\mathbf{X})^{-1}\\ &= \mathsf{E}\,||\mathbf{b}||^2 - ||\boldsymbol{\beta}||^2. \qquad \square\end{aligned}$$

Ze vztahu (10.1) je zřejmé, že střední hodnota čtverce délky odhadu vektoru $\mathsf{E}\,\mathbf{Y}$ závisí pouze na skutečné hodnosti matice $\mathbf{X}$, nikoliv na tom, jak „dobře“ jsou její sloupce lineárně nezávislé. Multikolinearita tu tedy nehraje žádnou roli. Totéž však neplatí pro odhad vektoru regresních koeficientů $\boldsymbol{\beta}$. Při tom právě tento vektor udává, která lineární kombinace sloupců matice $\mathbf{X}$ tvoří jednoznačně určený vektor $\hat{\mathbf{Y}}$.

Dál budeme v této kapitole předpokládat, že platí $\mathsf{h}(\mathbf{X}) = k+1$. Nechť $\mathbf{X}'\mathbf{X}$ má spektrální rozklad podle (A.5) (s vlastními čísly $\lambda_1,\dots,\lambda_{k+1}$) tvaru:

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{k+1}\lambda_i\mathbf{q}_i\mathbf{q}_i'. \tag{10.3}$$

Potom platí

$$\mathsf{E}\,||\mathbf{b}||^2 = ||\boldsymbol{\beta}||^2 + \sigma^2\sum_{i=1}^{k+1}\frac{1}{\lambda_i}.$$

Malá vlastní čísla se tedy projeví velikou neshodou mezi $\mathsf{E}\,||\mathbf{b}||^2$ a $||\boldsymbol{\beta}||^2$.

Předpokládejme, že vlastní čísla jsou označena indexy tak, aby platilo

$$\lambda_1 \ge \ldots \ge \lambda_{k+1} > 0,$$

když poslední nerovnost plyne z našeho předpokladu o hodnosti matice $\mathbf{X}$. O nebezpečí multikolinearity do značné míry vypovídá *číslo podmíněnosti* matice $\mathbf{X}'\mathbf{X}$, které je definováno jako λ_1/λ_{k+1}. Podobně číslo podmíněnosti matice $\mathbf{X}$ je rovno $\sqrt{\lambda_1/\lambda_{k+1}}$. Podrobnější informaci dají *indexy podmíněnosti* matice $\mathbf{X}'\mathbf{X}$

$$\eta_j = \frac{\lambda_1}{\lambda_j}, \qquad 1 \le j \le k+1.$$

Číslo podmíněnosti matice $\mathbf{X}'\mathbf{X}$ je rovno η_{k+1} a číslo podmíněnosti matice $\mathbf{X}$ je rovno $\sqrt{\eta_{k+1}}$.

Je třeba upozornit na jednu velmi nepříjemnou vlastnost vlastních čísel, totiž jejich závislost na zvoleném měřítku. Porovnejme dvě matice:

$$A = \begin{pmatrix}30 & 2 & 1\\ 2 & 30 & 5\\ 1 & 5 & 10\end{pmatrix}, \qquad B = \begin{pmatrix}30 & 0{,}02 & 1000\\ 0{,}02 & 0{,}0030 & 50\\ 1000 & 50 & 10000000\end{pmatrix}.$$

Může jít o dvě matice typu $\mathbf{X}'\mathbf{X}$, které se liší pouze měřítkem, v jakém jsou vyjádřena data. Matice $\mathbf{X}$ má tři sloupce, z nichž první obsahuje jedničky (pro absolutní člen). Druhý sloupec obsahuje délkové údaje vyjádřené v centimetrech (matice $\mathbf{A}$) nebo v metrech (matice $\mathbf{B}$), třetí sloupec obsahuje údaje o hmotnosti vyjádření v kilogramech nebo v gramech. Jedná se tedy vlastně o stejnou úlohu, ovšem čísla podmíněnosti matice $\mathbf{X}'\mathbf{X}$ jsou velmi různá: $\eta_{k+1}(A) = 3{,}730$ je poměrně malé, kdežto $\eta_{k+1}(B) = 3{,}646 \cdot 10^9$.

Někdy se tedy, dříve než se spočítají vlastní čísla, matice $\mathbf{X}$ normuje tak, aby všechny její sloupce měly stejnou délku (viz program NCSS). Má to význam zejména tehdy, když máme interpretaci pro absolutní člen modelu.

Druhým používaným normováním je přechod ke korelačním koeficientům, jak to provedeme v následující kapitolce. Tento postup však nelze použít tehdy, když má ve vyšetřovaném modelu absolutní člen vlastní věcnou interpretaci.

10.2. Regrese standardizovaných veličin

Mnohé programy nabízejí diagnostické prostředky, které jsou založeny na standardizovaných veličinách a jejich kovariancích, tedy na korelačních koeficientech.

Uvažujme model tvaru

$$Y_i = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j + e_i, \qquad 1 \le i \le n, \tag{10.4}$$

kde nezávislé náhodné veličiny $e_1, \ldots, e_n$ mají rozdělení $\mathsf{N}(0, \sigma^2)$. Předpokládáme opět, že matice

$$\mathbf{X} = (\mathbf{1}, \mathbf{x}_{\bullet 1}, \cdots, \mathbf{x}_{\bullet k})$$

má lineárně nezávislé sloupce, tedy hodnost $k+1$. Označme

$$T_j^2 = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2, \qquad T_0^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

a zaveďme standardizované veličiny

$$Y_i^* = \frac{Y_i - \bar{Y}}{T_0}, \qquad x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{T_j},$$

pro které platí

$$\sum_{i=1}^{n} Y_i^* = 0, \qquad \sum_{i=1}^{n} Y_i^{*2} = 1, \qquad \sum_{i=1}^{n} x_{ij}^* = 0, \qquad \sum_{i=1}^{n} x_{ij}^{*2} = 1.$$

Označme dále

$$r_{jt} = \sum_{i=1}^{n} x_{ij}^* x_{it}^*, \qquad r_{j0} = \sum_{i=1}^{n} x_{ij}^* Y_i^*.$$

Snadno nahlédneme, že r_{jt}, r_{j0} jsou výběrové korelační koeficienty. Nyní vyjádříme původní pozorování pomocí odhadů

$$\begin{aligned} Y_i = \hat{Y}_i + u_i \ &= \ b_0 + \sum_{j=1}^{k} x_{ij} b_j + u_i \\ &= \left(b_0 + \sum_{j=1}^{k} \bar{x}_j b_j \right) + \sum_{j=1}^{k} (x_{ij} - \bar{x}_j) b_j + u_i \\ &= \bar{Y} + \sum_{j=1}^{k} (x_{ij} - \bar{x}_j) b_j + u_i, \end{aligned}$$

když jsme využili skutečnosti, že v modelu s absolutním členem prochází odhadnutá závislost těžištěm, takže platí $\bar{Y} = b_0 + \sum_{j=1}^{k} \bar{x}_j b_j$.

Poslední vztah vyjádříme pomocí standardizovaných veličin označených hvězdičkou, dostaneme tak *standardizovaný model*

$$\begin{aligned} Y_i^* = \frac{Y_i - \bar{Y}}{T_0} \ &= \ \sum_{j=1}^{k} \frac{x_{ij} - \bar{x}_j}{T_j} \frac{T_{jj}}{T_0} b_j + \frac{u_i}{T_{00}} \\ &= \sum_{j=1}^{k} x_{ij}^* b_j^* + u_i^*, \end{aligned}$$

když jsme zavedli *standardizované koeficienty* $b_j^* = (T_{jj}/T_{00}) b_j$ a rezidua standardizovaného modelu $u_i^* = u_i / T_{00}$. Reziduální součet čtverců standardizovaného modelu RSS^* zřejmě těsně souvisí s koeficientem determinace

$$RSS^* = \sum_{i=1}^{n} u_i^* = \sum_{i=1}^{n} \left(\frac{u_i}{T_{00}} \right)^2 = \frac{RSS}{T_{00}^2} = 1 - \left(1 - \frac{RSS}{T_{00}^2} \right) = 1 - R^2. \qquad (10.5)$$

Pokusme se vyjádřit hledání odhadů regresních koeficientů. Když shromáždíme standardizované veličiny x_{ij}^* a Y_i^* do matice $\mathbf{X}^*$ a vektoru $\mathbf{Y}^*$, bude vektor $\mathbf{b}^* = (b_1^*, \dots, b_k^*)'$ řešením normální rovnice (standardizovaný model má absolutní člen identicky nulový)

$$(\mathbf{X}^{*'} \mathbf{X}^*) \mathbf{b}^* = \mathbf{X}^{*'} \mathbf{Y}^*.$$

Označíme-li matici korelačních koeficientů r_{jt} jako $\mathbf{R}_{xx}$ a podobně vektor korelačních koeficientů r_{j0} symbolem $\mathbf{r}_{xy}$, můžeme poslední vztah vyjádřit také jako

$$\mathbf{R}_{xx} \mathbf{b}^* = \mathbf{r}_{xy}.$$

Vyjádříme ještě odhad varianční matice statistiky $\mathbf{b}^*$:

$$\widehat{\operatorname{var}\mathbf{b}^*} = S^{*2}\mathbf{R}_{xx}^{-1} = \frac{RSS^*}{n-k-1}\mathbf{R}_{xx}^{-1} = \frac{1-R^2}{n-k-1}\mathbf{R}_{xx}^{-1}.$$

Použijeme-li běžné označení prvků inverzní matice pomocí horních indexů, dostaneme vyjádření

$$\widehat{\operatorname{var}b_j^*} = \frac{1-R^2}{n-k-1}r^{jj}.$$

V dalším bude užitečné další vyjádření koeficientu determinace. Postupně upravíme inverzní matici k výběrové korelační matici veličin $Y^*, x_1^*, \ldots, x_k^*$ (která je totožná s korelační maticí veličin $Y, x_1, \ldots, x_k$):

$$\begin{aligned}
\begin{pmatrix} 1 & \mathbf{r}'_{xy} \\ \mathbf{r}_{xy} & \mathbf{R}_{xx} \end{pmatrix}^{-1} &= \begin{pmatrix} (1-\mathbf{r}'_{xy}\mathbf{R}_{xx}^{-1}\mathbf{r}_{xy})^{-1} & * \\ * & * \end{pmatrix} \\
&= \begin{pmatrix} (\mathbf{Y}^{*'}\mathbf{Y}^* - \mathbf{Y}^{*'}\mathbf{X}^*(\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{Y}^*)^{-1} & * \\ * & * \end{pmatrix} \\
&= \begin{pmatrix} RSS^{*-1} & * \\ * & * \end{pmatrix} = \begin{pmatrix} (1-R^2)^{-1} & * \\ * & * \end{pmatrix}
\end{aligned}$$

Nyní vyjádříme jemněji j-tý diagonální prvek matice $\mathbf{R}_{xx}^{-1}$. Představme si nyní, že na místě veličiny Y je jedna z veličin x_j. Označme symbolem R_j^2 koeficient determinace závislosti $\mathbf{x}_{\bullet j}$ na ostatních veličinách, tedy na veličinách $\mathbf{x}_{\bullet 1}, \ldots, \mathbf{x}_{\bullet(j-1)}$, $\mathbf{x}_{\bullet(j+1)}, \ldots, \mathbf{x}_{\bullet k}$. Z úvahy o inverzní matici ke korelační matici zřejmě plyne, že platí

$$r^{jj} = \frac{1}{1-R_j^2}$$

Můžeme tedy vyjádřit odhad rozptylu odhadu b_j^* ve tvaru

$$\widehat{\operatorname{var}b_j^*} = \frac{1-R^2}{n-k-1}\frac{1}{1-R_j^2}. \tag{10.6}$$

Nejmenší možný rozptyl dostaneme, když je $R_j^2 = 0$, s rostoucí hodnotou R_j^2 se rozptyl odhadu b_j^* zvětšuje. Charakteristika $1-R_j^2$ se zpravidla nazývá *tolerance*, její převrácená hodnota se označuje *VIF* (Variance Inflation Factor) a ukazuje, kolikrát se zhorší rozptyl odhadu b_j^* v důsledku korelovanosti j-tého regresoru s ostatními regresory.

Ukažme ještě souvislost s původními parametry. Protože je $b_j = (T_0/T_j)b_j^*$, platí

$$\widehat{\operatorname{var}b_j} = \frac{1-R^2}{n-k-1}\frac{1}{1-R_j^2}\left(\frac{T_0}{T_j}\right)^2.$$

Poslední poznámka patří testování nulovosti regresních koeficientů β_j. Testovou statistiku lze vyjádřit následovně:

$$\frac{b_j}{\sqrt{\widehat{\operatorname{var} b_j}}} = \frac{(T_0/T_j)b_j^*}{\sqrt{\widehat{\operatorname{var}((T_0/T_j)b_j^*)}}} = \frac{b_j^*}{\sqrt{\widehat{\operatorname{var} b_j^*}}}$$
$$= b_j^*\sqrt{\frac{n-k-1}{1-R^2}}\sqrt{1-R_j^2}.$$

Rozhodovat lze tedy buď v původní nebo v upravené (hvězdičkové) parametrizaci. Dále je zřejmé, jak závisí na vnitřní závislosti mezi regresory. Malá tolerance (velký inflační faktor VIF) vyžaduje větší hodnotu $|b_j^*|$ k tomu, abychom mohli prokázat nenulovost parametru β_j.

Ve výstupu programu NCSS lze koeficienty b_j^* nalézt v oddílu nazvaném Regression Coefficient Section pod názvem Standardized Coefficient. Program STATISTICA uvádí tyto odhady ve sloupci nadepsaném BETA. V R si můžeme pomoci procedurou `scale()`, která provádí normování (přechod od x_{ij} k x_{ij}^*).

Příklad 10.1 (měření IQ) Použijme data, zjištěná na velké škole při pedagogickém výzkumu. Pro každého ze 111 žáků známe jeho pohlaví, průměrný prospěch v pololetí sedmé a osmé třídy a hodnotu IQ. Naším cílem je ověřit možnost odhadovat IQ nepřímo, ze známých průměrných známek, případně s přihlédnutím k pohlaví, kdy dívky jsou kódovány jedničkou a hoši nulou. Výběrové korelační koeficienty zjistíme snadno:

```
> cor(cbind(iq,divka,zn7,zn8))
                 iq      divka        zn7        zn8
iq        1.0000000  0.1217568 -0.6887396 -0.6571046
pohlavi   0.1217568  1.0000000 -0.3666488 -0.3802419
zn7      -0.6887396 -0.3666488  1.0000000  0.9545902
zn8      -0.6571046 -0.3802419  0.9545902  1.0000000
```

Při výpočtu odhadů standardizovaného modelu b_j^* ponecháme přednastavené parametry funkce `scale` (odečte průměr, vydělí směrodatnou odchylkou). I když je ve standardizovaném modelu absolutní člen identicky nulový, my jej v popisu závislosti ponecháme, abychom zachovali správný počet stupňů volnosti (absolutní člen je v upraveném modelu pouze skryt).

```
> summary(lm(scale(iq)~scale(pohlavi)+scale(zn7)+scale(zn8),data=Iq))

Call:
lm(formula = scale(iq) ~ scale(pohlavi) + scale(zn7) + scale(zn8))

Residuals:
     Min       1Q   Median       3Q      Max
-1.47790 -0.50164 -0.02892  0.47855  1.76069
```

```
Coefficients:
                Estimate Std. Error   t value Pr(>|t|)
(Intercept)    -1.455e-16  6.844e-02 -2.13e-15  1.00000
scale(pohlavi) -1.528e-01  7.434e-02    -2.055  0.04232 *
scale(zn7)     -6.989e-01  2.308e-01    -3.029  0.00308 **
scale(zn8)     -4.800e-02  2.321e-01    -0.207  0.83658

Residual standard error: 0.721 on 107 degrees of freedom
Multiple R-Squared: 0.4943,     Adjusted R-squared: 0.4801
F-statistic: 34.87 on 3 and 107 degrees of freedom,  p-value: 8.882e-016
```

Pro srovnání uveďme také klasické odhady b_j:

```
> summary(lm(IQ~pohlavi+zn7+zn8,data=Iq))

Call:
lm(formula = IQ ~ pohlavi + zn7 + zn8)

Residuals:
     Min       1Q   Median       3Q      Max
-22.1677  -7.5243  -0.4338   7.1780  26.4095

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  142.785      3.869  36.909  < 2e-16 ***
pohlavi       -4.563      2.221  -2.055  0.04232 *
zn7          -16.767      5.536  -3.029  0.00308 **
zn8           -1.149      5.557  -0.207  0.83658

Residual standard error: 10.81 on 107 degrees of freedom
Multiple R-Squared: 0.4943,     Adjusted R-squared: 0.4801
F-statistic: 34.87 on 3 and 107 degrees of freedom,  p-value: 8.882e-016
```

Všimněme si především stejných hodnot jednotlivých t-statistik a odpovídajích dosažených hladin testu v běžném a standardizovaném modelu. Totéž platí pro koeficient determinace i pro adjustovaný koeficient determinace.

Ponechme zatím stranou velkou dosaženou hladinu u průměru z 8. třídy, která svědčí o tom, že tento regresor bychom mohli vynechat. O multikolinearitě svědčí velký korelační koeficient mezi oběma průměrnými známkami: Absolutní člen tentokrát nemá v modelu vlastní význam, proto při hodnocení multikolinearity vyjdeme z korelační matice. Indexy podmíněnosti a další charakteristiky odvozené z korelační matice spočítáme jednoduchou procedurou

```
VIF <- function(lmobj)
#  počítá diagnostické statistiky související s multikolinearitou
#  založené na korelační matici
{
if (!is.null(weights(lmobj)))
    stop("requires unweighted model")
```

```
if (!(any(names(coefficients(lmobj))=="(Intercept)")))
    stop("requires model with intercept")
X0 <- scale(model.matrix(lmobj))[,-1]   # standardizace regresorů
nam <- names(coefficients(lmobj))[-1]
y0 <- scale(lmobj$model[,1])            # standardizace regresandu
lmobj0 <- lm(y0~X0)                     # standardizovaná regrese
VIF <- diag(solve(cor(X0)))
tol <- 1/VIF; R2 <- 1-tol
b.star <- coef(lmobj0)[-1]
out <- cbind(b.star,VIF,R2,tol)
rownames(out) <- names(coefficients(lmobj))[-1]
return(out)
}
```

Vyšetřovaný model dal tyto výsledky:

```
VIF(lm(iq~divka+zn7+zn8,data=Iq))
          b.star       VIF        R2        tol
divka -0.15275544  1.169230 0.1447359 0.85526408
zn7   -0.69892795 11.268657 0.9112583 0.08874172
zn8   -0.04799886 11.402400 0.9122992 0.08770084
```

Sloupec nazvaný `beta0` obsahuje odhady b_j^*. Sloupec nazvaný `R2` obsahuje koeficienty determinace R_j^2 v regresních modelech, kdy se snažíme vysvětlit regresor x_j jako lineární funkci všech ostatních regresorů.

Ukazuje se, že vzájemná závislost některých regresorů zvětšila rozptyl odhadů koeficientů u standardizovaných průměrů více než desetkrát (*VIF*). Velikost vzájemné závislosti charakterizují velké koeficienty determinace. Například průměr v 8. třídě lze vysvětlit více než z 90 % pomocí ostatních regresorů. Diagonální prvky `vjj` matice $(\mathbf{X}'\mathbf{X})^{-1}$ udávají (až na S^2) rozptyl odhadů b_j.

Pro zajímavost, když odstraníme z modelu průměr známek z 8. třídy, jsou obě inflační čísla *VIF* rovna 1,155310 (Pročpak jsou *obě* inflační čísla stejná?):

```
> VIF(lm(iq~divka+zn7,data=Iq))
          beta0      VIF        R2       tol
divka -0.1510784 1.155310 0.1344313 0.8655687
zn7   -0.7441323 1.155310 0.1344313 0.8655687
```

Všimněme si také odhadů regresních koeficientů.

```
> summary(lm(iq~divka+zn7,data=Iq))

Call:
lm(formula = iq ~ divka + zn7, data = Iq)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9606  -7.4290  -0.1927   7.0047  26.5244
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  142.607      3.755  37.982   <2e-16 ***
divka         -4.513      2.198  -2.054   0.0424 *
zn7          -17.852      1.765 -10.116   <2e-16 ***

Residual standard error: 10.77 on 108 degrees of freedom
Multiple R-Squared: 0.4941,     Adjusted R-squared: 0.4848
F-statistic: 52.74 on 2 and 108 DF,  p-value: < 2.2e-16
```

Je jistě patrné, jak byl krátkozraká případná iterpretace velké dosažené hladiny u proměnné `zn8`. Vůbec neznamenala, že by hodnota IQ nesouvisela se známkovým průměrem. Pouze tento průměr neuměl říci nic podstatně nového o IQ, co bychom nevěděli z proměnných `divka, zn7`.

Ještě k charakteristikám podmíněnosti. Největší index podmíněnosti 48,330 z modelu s obojími průměrnými známkami založený na zhodnocení korelační matice (absolutní člen nás nezajímá) se zmenší na 2,158 u zjednodušeného modelu:

```
> ind.podm <- function(A) {e <- eigen(A); e$val[1]/e$val}
> ind.podm(cor(cbind(pohlavi,zn7,zn8)))
[1]  1.000000  2.859583 48.330483
> ind.podm(cor(cbind(pohlavi,zn7)))
[1] 1.000000 2.157806
```

○

11. Hledání modelu

V následující kapitole uvedeme některé charakteristiky a postupy, které lze použít v souvislosti s hledáním modelu. Nepochybně není na škodu připomenout, že nejlepší je situace, kdy model je odvozen z představy o fungování vyšetřovaných dějů. Je-li to možné, takovému postupu je třeba vždy dát přednost. To se týká také plánování pokusu (pro jaké hodnoty nezávisle proměnné zjišťovat hodnotu závisle proměnné).

11.1. Dvě kritéria

Nejprve provedeme dvě obecné úvahy o praktických možnostech srovnání modelu a podmodelu.

11.1.1. Silné kritérium

Připomeňme si větu 7.2. Tehdy jsme při porovnávání standardního modelu s nějakých obsáhlejším modelem zjistili, že menší klasický model nedá horší střední čtvercové chyby, pokud je čtverec délky vychýlení nejvýše roven rozptylu (tj. $||\mathsf{bias}\,\hat{\mathbf{Y}}||^2 \leq \sigma^2$). Předpokládejme nyní, že vektory parametrů $\boldsymbol{\beta}, \boldsymbol{\gamma}$ jsou oba odhadnutelné, což je zaručeno například tím, že matice $\mathbf{X}$ a $\mathbf{MZ}$ mají lineárně nezávislé sloupce, tj. platí $\mathsf{h}(\mathbf{X}) = k + 1$ a $\mathsf{h}(\mathbf{MZ}) = m$. Pod m si můžeme představovat počet nových regresorů v matici $\mathbf{Z}$.

Podle (7.10) vyjádříme vychýlení odhadu $\hat{\mathbf{Y}}$ jako $-\mathbf{MZ}\boldsymbol{\gamma}$ a do tohoto výrazu za $\boldsymbol{\gamma}$ i za σ^2 dosadíme běžné odhady, dostaneme *silné kritérium*

$$||\mathbf{MZc}_g||^2 \leq S_g^2. \tag{11.1}$$

Nyní tuto nerovnost vyjádříme praktičtějším způsobem. Protože podle (8.8) platí $RSS - RSS_g = ||\mathbf{MZc}_g||^2$, má testová statistika podmodelu (zde je jím klasický model) tvar

$$F = \frac{||\mathbf{MZc}_g||^2/m}{S_g^2}. \tag{11.2}$$

Silné kritérium je tedy ekvivalentní s požadavkem

$$F \le \frac{1}{m}. \tag{11.3}$$

V běžném regresním výstupu máme vedle odhadů jednotlivých regresních koeficientů uvedeny t statistiky. Můžeme je nějak v souvislosti s ověřováním (11.3) použít?

Připomeňme, že platí (7.18), takže varianční matici odhadu $\mathbf{c}_g$ můžeme odhadnout pomocí $\widehat{\mathsf{var}}\,\mathbf{c}_g = S_g^2(\mathbf{Z}'\mathbf{MZ})^{-1}$. Proto platí

$$(\mathbf{c}_g)'\left(\widehat{\mathsf{var}}\,\mathbf{c}_g\right)^{-1}\mathbf{c}_g = \frac{1}{S_g^2}(\mathbf{c}_g)'\left(\mathbf{Z}'\mathbf{MZ}\right)\mathbf{c}_g = \frac{||\mathbf{MZc}_g||^2}{S_g^2} = mF.$$

Se silným kritériem je ekvivalentní nerovnost $\mathbf{c}_g'\left(\widehat{\mathsf{var}}\,\mathbf{c}_g\right)^{-1}\mathbf{c}_g \le 1$. Podle věty A.8 je tato nerovnost ekvivalentní s tím, že matice $\widehat{\mathsf{var}}\,\mathbf{c}_g - \mathbf{c}_g\mathbf{c}_g'$ je pozitivně semidefinitní. K tomu je ale nutné (ale nemusí stačit), aby všechny diagonální prvky této matice byly nezáporné, tedy aby pro všechny t statistiky pro testy hypotéz, že je $\gamma_j = 0$, platilo

$$T_{\gamma_j} = \frac{|c_{gj}|}{\sqrt{\left(\widehat{\mathsf{var}}\,\mathbf{c}_g\right)_{jj}}} = \frac{|c_{gj}|}{\mathsf{S.E.}(c_{gj})} \le 1. \tag{11.4}$$

Odtud plyne užitečný závěr: *mezi kandidáty na „zbytečné“ regresory ve smyslu silného kritéria mohou patřit jen takové, u nichž je t statistika nejvýše rovna jedničce.*

11.1.2. Slabé kritérium

Když se nebudeme zajímat o všechny lineární funkce parametrů $\boldsymbol{\beta}, \boldsymbol{\gamma}$ (s tím je ekvivalentní vyšetřování $\hat{\mathbf{Y}}$), ale jen o kombinace „vyzkoušené“ v datech, můžeme porovnat střední čtvercové chyby odhadů $(\mathbf{x}_{i\bullet})'\mathbf{b}$ a $(\mathbf{x}_{i\bullet})'\mathbf{b}_g + (\mathbf{z}_{i\bullet})'\mathbf{c}_g$ pro lineární funkce parametrů $(\mathbf{x}_{i\bullet})'\boldsymbol{\beta} + (\mathbf{z}_{i\bullet})'\boldsymbol{\gamma}$, kde $i = 1, \dots, n$.

Zajímá nás tedy, kdy bude splněn požadavek (*slabé kritérium*)

$$\sum_{i=1}^{n} \mathsf{MSE}\,(\hat{Y}_i) \le \sum_{i=1}^{n} \mathsf{MSE}\,(\hat{Y}_{gi}). \tag{11.5}$$

Po dosazení postupně tento požadavek upravíme na

$$\begin{aligned}
\sum_{i=1}^{n}\left(\mathsf{var}\,\hat{Y}_i + (\mathsf{bias}\,\hat{Y}_i)^2\right) &\le \sum_{i=1}^{n}\mathsf{var}\,\hat{Y}_{gi},\\
\sum_{i=1}^{n}\left(\sigma^2 h_{ii} + ((-\mathbf{m}_{i\bullet})'\mathbf{Z}\boldsymbol{\gamma})^2\right) &\le \sum_{i=1}^{n}\sigma^2 h_{gii},\\
\sigma^2(k+1) + ||\mathbf{MZ}\boldsymbol{\gamma}||^2 &\le \sigma^2(k+1+m).
\end{aligned} \tag{11.6}$$

Výsledkem je tedy požadavek

$$||\mathbf{MZ}\gamma||^2 \leq m\sigma^2, \tag{11.7}$$

který nahradil podobný požadavek (11.1) silného kritéria. Protože se obě nerovnosti liší pouze koeficientem m na pravé straně (11.7), je zřejmé, že nerovnost (11.3) můžeme v případě slabého kritéria nahradit požadavkem $F \leq 1$ a nutnou podmínku (11.4) slabším požadavkem $|T_{\gamma_j}| \leq \sqrt{m}$.

Mezi kandidáty na „zbytečné“ regresory ve smyslu slabého kritéria mohou patřit jen takové, u nichž je t statistika nejvýše rovna $\sqrt{m}$.

11.2. Porovnání modelu a podmodelu

Zde shrneme zpravidla již známá tvrzení o možnostech porovnání kvality modelu a podmodelu.

11.2.1. Reziduální součet čtverců RSS

Podle (8.8) víme, že platí

$$RSS_g = RSS - ||\mathbf{MZc}_g||^2 \leq RSS,$$

takže reziduální součet čtverců v podmodelu je zdola omezen reziduálním součtem čtverců v modelu. Přejdeme-li k podmodelu, nemůže reziduální součet čtverců klesnout.

11.2.2. Koeficient determinace R^2

Vzhledem ke vztahu mezi RSS_g a RSS platí

$$R_g^2 = 1 - \frac{RSS_g}{||\mathbf{Y} - \bar{Y}\mathbf{1}||^2} \geq 1 - \frac{RSS}{||\mathbf{Y} - \bar{Y}\mathbf{1}||^2} = R^2.$$

Při zjednodušení modelu na podmodel nemůže koeficient determinace vzrůst. Uspořádání posloupnosti do sebe vřazených podmodelů podle klesajícího koeficientu determinace je stejné, jako uspořádání týchž podmodelů podle rostoucího reziduálního součtu čtverců.

11.2.3. Reziduální rozptyl S^2

Nejprve vyjádříme požadavky silného a slabého kritéria pomocí nestranných odhadů rozptylu v modelu a podmodelu. Pomocí obou reziduálních součtů čtverců můžeme statistiku F ze vztahu (11.2) upravit postupně jako

$$\begin{aligned} F &= \frac{RSS - RSS_g}{RSS_g}\frac{n-k-1-m}{m} \\ &= \frac{(n-k-1)S^2 - (n-k-1-m)S_g^2}{mS_g^2} \\ &= \frac{n-k-1}{mS_g^2}\left(S^2 - S_g^2\right) + 1, \end{aligned} \tag{11.8}$$

takže požadavek slabého kritéria lze zapsat jako $S^2 \le S_g^2$.

Podobně požadavek silného kritéria $F \le 1/m$ vede k nerovnosti

$$(n-k-1)S^2 - (n-k-1-m)S_g^2 \le S_g^2,$$

která je ekvivalentní s nerovností

$$S^2 \le \frac{n-k-m}{n-k-1}S_g^2. \tag{11.9}$$

O možnostech splnění poslední nerovnosti vypoví následující úvaha. Nerovnost $RSS_g \le RSS$ je ekvivalentní s nerovností $(n-k-1-m)S_g^2 \le (n-k-1)S^2$, která dá omezení zdola pro odhad rozptylu S^2, které je téměř totožné s omezením shora uvedeným v (11.9). Platí-li silné kritérium, musí být současně splněny nerovnosti

$$\frac{n-k-1-m}{n-k-1}S_g^2 \le S^2 \le \frac{n-k-m}{n-k-1}S_g^2.$$

Je vidět, že silné kritérium dává jen velmi málo „svobody“ pro možné hodnoty reziduálního rozptylu S^2.

11.2.4. Adjustovaný koeficient determinace R^2_{adj}

Klasický koeficient determinace R^2 lze vyjádřit pomocí odhadů rozptylu metodou maximální věrohodnosti v modelu a ve speciálním podmodelu, který má pouze absolutní člen, totiž $\mathsf{E}\,\mathbf{Y} = \mathbf{1}\gamma$, jako

$$R^2 = 1 - \frac{RSS/n}{\sum(Y_i - \bar{Y})^2/n} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}.$$

Když nyní nahradíme odhady metodou maximální věrohodností příslušnými nestrannými odhady, dostaneme *adjustovaný (upravený) koeficient determinace*

$$R^2_{adj} = 1 - \frac{RSS/(n-k-1)}{\sum (Y_i - \bar{Y})^2/(n-1)} = 1 - \frac{n-1}{n-k-1}\left(1 - R^2\right).$$

Protože lze tento koeficient vyjádřit jako monotonní funkci výběrového rozptylu S^2 (S_0^2 je odhad rozptylu v podmodelu)

$$R^2_{adj} = 1 - \frac{S^2}{S_0^2},$$

je uspořádání posloupnosti do sebe vnořených podmodelů podle klesajícího upraveného koeficientu determinace stejné, jako podle rostoucího výběrového rozptylu.

11.2.5. Mallowsovo C_p

Myšlenka statistiky C_p je založena na porovnání odhadu celkové střední čtvercové chyby z (11.5) s „bezpečným“ odhadem rozptylu.

Nechť platí „bezpečný“ model $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \sigma^2\mathbf{I})$. Použijeme-li střední hodnotu $\mathsf{E}\,RSS$ ze vztahu (7.9), dostaneme v předpokládaném modelu s úplnou hodností vztah

$$\mathsf{E}\,RSS = (n-k-1)\sigma^2 + ||\mathbf{MZ}\boldsymbol{\gamma}||^2.$$

Když vyjádříme celkovou střední chybu podle (11.6), dostaneme

$$\sum_{i=1}^{n} \mathsf{MSE}\,(\hat{Y}_i) = (k+1)\sigma^2 + ||\mathbf{MZ}\boldsymbol{\gamma}||^2.$$

Když ze dvou posledních rovnic vyloučíme neznámý čtverec délky vychýlení $||\mathbf{MZ}\boldsymbol{\gamma}||^2$ a celkovou střední čtvercovou chyby podělíme rozptylem, dostaneme

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} \mathsf{MSE}\,(\hat{Y}_i) = \frac{(k+1)\sigma^2 + \mathsf{E}\,RSS - (n-k-1)\sigma^2}{\sigma^2} = 2(k+1) - n + \frac{\mathsf{E}\,RSS}{\sigma^2}.$$

Nahradíme-li nyní neznámý rozptyl σ^2 jeho nestranným odhadem S_g^2 a střední hodnotu statistiky RSS její skutečnou hodnotou, dostaneme *Mallowsovo C_p*

$$C_p = 2(k+1) - n + \frac{RSS}{S_g^2}. \tag{11.10}$$

Zbývá ukázat souvislost s nahoře uvedeným slabým kritériem. Použijme vyjádření F statistiky podle (11.8). Snadnou úpravou dostaneme

$$m(F-1) = \frac{n-k-1}{S_g^2}\left(S^2 - S_g^2\right) = \frac{RSS}{S_g^2} - (n-k-1) = C_p - k - 1.$$

Slabé kritérium $F \le 1$ je tedy ekvivalentní s nerovností $C_p \le k+1$. Protože je dále

$$m(F - \frac{1}{m}) = C_p - k - 2 + m,$$

je silné kritérium $F < 1/m$ ekvivalentní s požadavkem $C_p \le k + 2 - m$.

11.2.6. Průměrný rozptyl předpovědi

Následující úvaha již není založena na porovnání modelu a podmodelu, už se nesnažíme model zjednodušit vylučováním některých regresorů. Tentokrát se budeme zamýšlet na přesností předpovědi budoucích pozorování,

Pro každý řádek matice $\mathbf{X}$ máme předpovídat nové pozorování $Y(\mathbf{x}_{i\bullet})$, nezávislé na těch, s jejichž pomocí jsme odhadli všechny parametry. Bodovým odhadem bude samozřejmě $\hat{Y}_i$. Ovšem rozptyl chyby předpovědi $\hat{Y}_i - Y(\mathbf{x}_{i\bullet})$ bude $\sigma^2 h_{ii} + \sigma^2$. Součet těchto rozptylů (celkový rozptyl) je tedy roven výrazu

$$\frac{1}{n}\sum_{i=1}^{n} \sigma^2 (1 + h_{ii}) = \sigma^2 \left(1 + \frac{k+1}{n}\right).$$

Když ještě neznámý parametr σ^2 nahradíme jeho nestranným odhadem S^2, dostaneme statistiku

$$J_k = S^2 \left(1 + \frac{k+1}{n}\right), \tag{11.11}$$

která na rozdíl od samotného rozptylu penalizuje počet parametrů použitých v modelu.

11.2.7. Akaikeho informační kritérium

V poslední době se k porovnání různých modelů často používá funkce založená na logaritmu odhadu rozptylu zvětšeném o penalizaci počtu odhadovaných parametrů (viz Anděl (1998, str. 187)). Akaikeho informační kritérium bylo navrženo jako

$$AIC = -2\ell(\hat{\boldsymbol{\theta}}) + 2q,$$

kde ℓ je logaritmická věrohodnostní funkce a q je počet složek maximálně věrohodného odhadu $\hat{\boldsymbol{\theta}}$. V případě lineárního normálního modelu se *známým* rozptylem σ^2 po dosazení do logaritmické věrohodnostní funkce dostaneme

$$AIC = n \log 2\pi\sigma^2 + \frac{RSS}{\sigma^2} + 2r,$$

což se až na konstantu velice podobá Malowsovu C_p.

Pokud odhadujeme také rozptyl σ^2, dostaneme (funkce `AIC()` v R)

$$\begin{aligned} AIC &= n\,(1 + \log(2\pi) + \log(RSS) - \log(n)) + 2(r+1) \\ &= n\left(1 + \log(2\pi\widehat{\sigma^2})\right) + 2(r+1), \end{aligned} \tag{11.12}$$

kde $\widehat{\sigma^2}$ je odhad σ^2 metodou maximální věrohodnosti a r je hodnost matice $\mathbf{X}$. V případě modelu s úplnou hodností a s absolutním členem je tedy $r = k + 2$ (nezapomeňme na to, že i σ^2 je pak odhadovaným parametrem).

11.2.8. Odhad stupně polynomu

Nechť je závislost $\mathsf{E}\,Y$ na nezávisle proměnné x popsána polynomem $\beta_0 + \beta_1 x + \ldots + \beta_k x^k$, přičemž platí $\beta_k \neq 0$. Máme k disposici $n > k+1$ nezávislých pozorování

$$Y_i = \sum_{j=0}^{k} \beta_j x^j + e_i,$$

kde $e_i \sim \mathsf{N}\big(0, \sigma^2\big)$. Předpokládáme, že stupeň k polynomu neznáme, že je dalším neznámým parametrem. V parametru k je úloha nelineární. V tomto odstavci popíšeme některé metody, které vedou ke konzistentnímu odhadu tohoto parametru.

Připomeňme vztah (7.13) z věty 7.1, podle kterého reziduální rozptyl nadhodnocuje skutečný rozptyl v případě, že použitý model opomíjí některé regresory, které skutečně ovlivňují střední hodnotu závisle proměnné. Na druhé straně, když použijeme některé regresory zbytečně, odhad rozptylu zůstane nestranným.

Zdálo by se tedy, že stačí odhadovat regresní modely postupně s rostoucím stupněm a skončit tehdy, když reziduální rozptyly (označíme je S_k^2) přestanou klesat, kdy začnou kolísat kolem nějaké konstanty. Tento postup ale nevede ke konzistentnímu odhadu stupně polynomu. Je třeba nějak penalizovat počet parametrů.

Kupodivu, i když statistika J_k z (11.11) se o takovou penalizaci snaží, nestačí to, minimalizace J_k přes stupeň polynomu nevede ke konzistentnímu odhadu. Podobně nemusí vést ke správné hodnotě ani Akaikeho kritérium z (11.12).

Ke konzistentním odhadům vede minimalizace řady funkcí, například

$$A(k) = S_k^2 \left(1 + c(k+1)n^{-\alpha}\right), \qquad \alpha \in (0, 0{,}5), c > 0, \tag{11.13}$$

$$SR(k) = \log S_k^2 + (k+1)\frac{\log n}{n}, \tag{11.14}$$

$$HQ(k) = \log S_k^2 + 2c(k+1)\frac{\log\log n}{n}, \quad c > 0. \tag{11.15}$$

11.3. Sekvenční postupy

Běžně používané programové vybavení nabízí zpravidla automatizovaný výběr regresorů z množiny možných regresorů, které zvolí uživatel. K tomu se používají v zásadě dva postupy a zejména jejich kombinace.

11.3.1. Sestupný výběr

Nejprve se spočítá nejbohatší model, pak se jednotlivé regresory postupně z modelu vylučují. V každém kroku se vylučuje takový regresor, který v daném modelu nejméně přispívá k vysvětlení. Označme symbolem t_j hodnotu t statistiky pro test hypotézy, že v daném modelu je koeficient u j-tého regresoru nulový. Zpravidla k rozhodování se používá čtverec této statistiky $F_j = t_j^2$. Končí se tehdy, když všechny tyto F statistiky pro vyloučení jsou větší, než nějaké předem zvolené kritické číslo F^{**}. Někdy se nevolí přímo toto číslo, ale spíš číslo α^{**}, z něhož se kritické číslo odvodí jako kritická hodnota $F^{**} = F_{1,n-k-1}(\alpha)$.

11.3.2. Vzestupný výběr

Jde o pravý opak předchozího postupu. Vyjde se z „prázdné“ množiny regresorů, do níž se pak v každém kroku přidá vždy ten z ještě nezařazených regresorů, který v daném kroku co možná nejlépe zlepší vysvětlení závisle proměnné. Představme si, že bychom zkusili jeden regresor vložit a jako F_j označíme čtverec t statistiky pro jeho vyloučení. V daném kroku vložíme takový regresor z dostupných kandidátů, u něhož je hodnota F největší. Skončíme, když toto F není dost velké, když je menší, než předem zvolené F^*. Také zde lze postup někdy řídit volbou α^*, z něhož se vlastní kritické číslo odvozuje.

11.3.3. Kroková regrese

Kroková (stepwise) regrese kombinuje oba právě popsané postupy. Vzestupný výběr je v každém kroku kombinován pokusem o zjednodušení pomocí sestupného výběru. Kdyby ovšem bylo $F^* \leq F^{**}$, mohlo by se stát, že dojde k zacyklení algoritmu, kdy bude právě vložený regresor okamžitě vyloučen, poté znovu vložen, vyloučen atd. Musí tedy být $F^* > F^{**}$, což je ekvivalentní s požadavkem $\alpha^* < \alpha^{**}$.

Každá z popsaných metod může dát jiný výsledný model, kromě jiného závisí také na volbě kritických čísel F^*, F^{**} resp. α^*, α^{**}. Výsledný model lze považovat nejvýše za doporučení, nikoliv za nějaký důkaz. Zejména u krokové regrese se doporučuje najít několik téměř optimálních modelů a pokusit se najít mezi nimi ten, který má nejlepší interpretaci.

11.3.4. Kroková volba modelu v R

V programu R je k dispozici procedura `step()`, která hledá model s nejmenší hodnotou AIC. Ve výstupu je však uváděna hodnota AIC z (11.12) zmenšená o konstantu $n + n\log(2\pi) + 2$. Jako ukázku hledejme v příkladu procento tuku nejlepší vysvětlení procenta tuku pomocí dostupných veličin:

```
> a<-step(lm(fat~1),
     scope=list(lower=~1,upper=~react+height+weight+pulse+diast))

Start:  AIC= 193.16
```

```
 fat ~ 1

         Df Sum of Sq     RSS     AIC
+ weight  1   1546.01  741.65  138.84
+ height  1    270.06 2017.60  188.88
+ react   1    129.92 2157.74  192.24
<none>                2287.66  193.16
+ pulse   1     21.06 2266.59  194.70
+ diast   1      0.57 2287.09  195.15

Step:  AIC= 138.84
 fat ~ weight

         Df Sum of Sq     RSS     AIC
+ pulse   1    111.52  630.14  132.70
+ height  1     87.32  654.33  134.58
<none>                 741.65  138.84
+ diast   1      2.92  738.73  140.65
+ react   1      2.87  738.79  140.65
- weight  1   1546.01 2287.66  193.16

Step:  AIC= 132.7
 fat ~ weight + pulse

         Df Sum of Sq     RSS     AIC
+ height  1    101.53  528.61  125.91
<none>                 630.14  132.70
+ diast   1      7.52  622.62  134.10
+ react   1      0.55  629.59  134.65
- pulse   1    111.52  741.65  138.84
- weight  1   1636.46 2266.59  194.70

Step:  AIC= 125.91
 fat ~ weight + pulse + height

         Df Sum of Sq     RSS     AIC
<none>                 528.61  125.91
+ react   1      0.94  527.66  127.82
+ diast   1      0.78  527.82  127.84
- height  1    101.53  630.14  132.70
- pulse   1    125.73  654.33  134.58
- weight  1   1485.84 2014.44  190.80
> summary(a)

Call: lm(formula = fat ~ weight + pulse + height)

Residuals:
     Min       1Q   Median       3Q      Max
-5.17474 -2.89827  0.09504  1.47482  7.63024
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.66934   14.17048   0.471  0.64011
weight       0.55847    0.04911  11.371 5.85e-15
pulse        0.12020    0.03635   3.307  0.00184
height      -0.26330    0.08858  -2.973  0.00469

Residual standard error: 3.39 on 46 degrees of
freedom Multiple R-Squared: 0.7689,     Adjusted
R-squared: 0.7539 F-statistic: 51.03 on 3 and 46
DF,  p-value: 1.126e-14
```

Z výpisu je patrné, jak se algoritmus v každém kroku pokusil přidat postupně každou proměnnou mimo stávající model a také ubrat každou proměnnou ze stávajícího modelu. Skončil tehdy, když žádná taková *jednokroková* změna nevede ke zmenšení AIC. Standardně má totiž parametr `direction` hodnotu `"both"`. Lze však nastavit vzestupný (`"forward"`) i setupný (`"backward"`) výběr.

Je třeba upozornit, že dosažené hodnoty u jednotlivých proměnných v modelu získané pomocí `summary(a)` je třeba interpretovat velice opatrně. Kdybychom dokázali vzít v úvahu cestu, jakou jsme došli v výslednému modelu, byly by tyto hodnoty nepochybně větší.

11.4. Praxe hledání modelu

Pokud hledáme pouze možnost predikce hodnot závisle proměnné, zpravidla nám dobře poslouží ten nejbohatší model. Zde je vhodné připomenout tvrzení věty 10.1, podle které je velký rozdíl v přesnosti odhadů $\hat{\mathbf{Y}}$ a $\mathbf{b}$.

Velmi často nás však zajímá vliv zvoleného regresoru nebo chceme modelovat vzájemné vztahy veličin. Potom je naším cílem odhadnout některý regresní koeficient či některé regresní koeficienty.

11.4.1. Interakce a confounding

Velmi často je při vyšetřování závislosti nějaké veličiny y na regresoru x třeba vzít v úvahu také další veličiny, které budeme v tomto odstavci značit symbolem z. Je při tom třeba rozlišovat dvě různé situace.

Interakce (effect modification) je taková situace, kdy skutečná hodnota veličiny z ovlivňuje závislost y na x. Interakce v tom nejjednodušším případě vyjadřují pomocí součinu $x \cdot z$. Příkladem by mohlo být například vyšetřování závislosti platu na délce praxe, když se zjistí, že směrnice příslušné přímky je jiná u mužů a jiná u žen. Kdyby byly přímky rovnoběžné, byl by vliv veličin *délka praxe* a *pohlaví* aditivní. Každý rok praxe by v průměru přidal stejnou částku k platu mužům i ženám. Vliv délky praxe by naopak byl modifikován proměnnou pohlaví, kdyby tyto průměrné přírůstky byly u mužů a u žen různé.

Jiná situace se popisuje anglickým slovem *confounding*. K matení dochází tehdy, když vedle nezávisle proměnné x a závisle proměnné y existuje jiná (matoucí) veličina z, která ovlivňuje y nezávisle na hodnotě x, přičemž sama z také souvisí s x. Neexistuje však příčinný řetězec $x \to z \to y$. Příkladem může být výskyt rakoviny jícnu y (měřený například počtem onemocnění na 100 000 obyvatel), který je ovlivňován podílem x kuřáků v populaci a současně spotřebou alkoholu z. Tyto dvě doprovodné veličiny spolu nepochybně také souvisí.

Jiným příkladem je tolikrát zmiňovaná závislost procenta tuku o mužů y v závislosti na výšce x a hmotnosti z. Dá se očekávat, že pro každou zvolenou hmotnost z bude s rostoucí výškou procento tuku klesat, takže jistě nejde o interakci. Ovšem, když vyšetřujeme závislost procenta tuku na výšce bez ohledu na hmotnost, skutečná závislost procenta tuku na výšce bude „překryta“ závislostí procenta na hmotnosti, protože hmotnost s výškou souvisí také.

Skutečnost, že se přihlédlo k závislosti na další veličině či veličinách se vyjadřuje slovy, že závislost byla adjustována vůči něčemu (adjusted for), že bylo přihlédnuto k závislosti ...

O nějaké veličině začneme uvažovat jako o matoucí teprve tehdy, když jsme vyloučili možnost interakcí.

11.4.2. Hierrarchicky dobře formulované modely (HWD)

S každou mocninou veličiny musí být v modelu všechny mocniny nižšího stupně, se součinem veličin musí být v modelu také všechny složky tohoto součinu.

Důvod k tomuto požadavku na hierarchicky dobře formulované hypotézy (Hierarchically Well-Formulated) je prostý. Zajistíme tak nezávislost na parametrizaci úlohy. Ukažme to na jednoduchém příkladu. Model kvadratické závislosti

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

vyjádříme pomocí nové nezávisle proměnné t zavedené vztahem $x = \delta(t - \varphi)$. Po dosazení postupně dostaneme

$$\begin{aligned} y &= \beta_0 + \beta_1 \delta(t - \varphi) + \beta_2 \left(\delta(t - \varphi)\right)^2 \\ &= (\beta_0 - \beta_1 \delta\varphi + \beta_2 \delta^2 \varphi^2) + (\beta_1 \delta - 2\beta_2 \delta^2 \varphi) t + \beta_2 \delta^2 t^2 \\ &= \gamma_0 + \gamma_1 t + \gamma_2 t^2. \end{aligned}$$

Kdybychom připustili model pouze s kvadratickým členem, bez členu lineárního, tj. s $\beta_1 = 0$, potom by se po netriviální lineární transformaci nezávisle proměnné tento člen v modelu znovu objevil. Podobnou úvahu bychom mohli udělat pro součin nezávisle proměnných.

11.4.3. Vyjádření nominální veličiny s více než dvěma hodnotami

Pokud střední hodnota závisle proměnné může být závislá na hodnotě nějakého nominálního znaku (faktoru), zpravidla v regresním modelu používáme umělé proměnné. U dvouhodnotového faktoru vystačíme s jedinou nula-jedničkovou veličinou, u faktoru s q různými hodnotami použijeme $q-1$ umělých proměnných, z nichž j-tá je rovna jedničce právě, když faktor nabyl své $(j+1)$. hodnoty. Koeficient u j-té umělé proměnné interpretujeme jako opravu absolutního členu, který popisuje závislost pro základní hodnotu faktoru (nepřísluší mu žádná umělá proměnná) na absolutní člen pro závislost při j-té hodnotě faktoru.

Při hledání modelu je třeba dodržovat pravidlo, že v modelu jsou a nebo nejsou současně zařazeny buď všechny umělé proměnné k jednomu faktoru nebo žádná z nich.

Čtenář si jistě uvědomil, že jsme právě použili reparametrizaci založenou na `contr.treatment`, která je u běžných faktorů v prostředí R nastavena standardně. Analogicky bychom mohli použít i jinou z nabízených reparametrizací.

11.4.4. Tři fáze (Kleinabaumův postup)

Podle Davida G. Kleinbauma (1994) se při hledání vhodného modelu použijí postupně tři fáze: najde se dobrý výchozí model, vyloučí se některé interakce, při vylučování dalších nezávisle proměnných se identifikují matoucí proměnné. Při zjednodušování modelu se dodržují obě dosud zmíněná pravidla: pravidlo hierarchicky dobře definovaného modelu a pravidlo o umělých proměnných.

Před provedením prvního kroku se samozřejmě seznámíme se všemi dostupnými modely, které se pokusily osvětlit vyšetřovanou závislost.

V prvním kroku zařadíme do modelu všechny dostupně proměnné, které by *mohly přispět* k vysvětlení variability závisle proměnné. Vedle proměnné x, jejíž vliv na střední hodnotu závisle proměnné nás zajímá, do modelu zařadíme také její druhou mocninu, pokud připouštíme možnost nelineární závislosti na x, dále všechny další doprovodné veličiny z, případně také součiny typu $x \cdot z$, které modelují možné interakce. Výjimečně se uvažují také mocniny veličin z, případně součiny typu $x \cdot z^2$. Při tom všem je třeba dbát na to, aby výsledek příliš neovlivnila multikolinearita. Další možností, jak sestavit vhodný výchozí model, je použít vhodně transformace závisle proměnné y a zejména x a z.

Ve druhém kroku se snažíme eliminovat interakční členy, tedy ty členy, které obsahují x a některá z. Při tom používáme standardní statistické testování. Doporučuje se nejprve se pokusit vyloučit naráz všechny takové členy.

Po ukončení druhého kroku si poznamenáme odhady regresních koeficientů u x a interakčních členů $x \cdot z$ a jejich střední chyby. Cílem třetího kroku je dál co nejvíc zjednodušit model, zmenšit střední chyby odhadů koeficientů u x a $x \cdot z$, ale jen tak, aby se odhad koeficientu u x číselně příliš nezměnil.

Pokud ve druhém kroku v modelu zůstal interakční člen, je situace složitější, protože příliš závisí na hodnotách doprovodné proměnné z z interakčního členu. Abychom se dostali k minimalizaci jedné střední chyby, zvolíme „typickou" hodnotu veličin x a z z interakčního členu a zajímáme se o odhad střední hodnoty y pro tuto hodnotu.

Za přijatelnou změnu se považuje změna do pěti až deseti procent výsledného odhadu z druhého kroku. Při vlastním zjednodušování modelu ve třetím kroku se vůbec nezajímáme o statistickou významnost vylučovaných členů, zejména necháme v modelu ty „nevýznamné" členy, po jejichž vyloučení by došlo k velké změně odhadů.

11.5. Transformace

Při práci s reálnými daty se mnohdy musíme uchýlit k transformacím. Pokud učiníme bohatším množinu možných středních hodnot tak, že jako regresor použijeme funkci některé nezávisle proměnné, nejde o nový problém. Ostatně polynomy patří mezi takové funkce také. Kvalitativně velmi odlišná situace nastane, když transformujeme závisle proměnnou.

11.5.1. Boxova-Coxova transformace

Boxova-Coxova transformace je pro kladné y zavedena předpisem

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \lambda \neq 0, \\ \log y & \lambda = 0. \end{cases} \tag{11.16}$$

Snadno se ověří, že funkce $y^{(\lambda)}$ je spojitou funkcí proměnné λ i v bodě 0.

Vektor se složkami $y_i^{(\lambda)}$ označíme symbolem $\mathbf{y}^{(\lambda)}$. Běžný lineární model modifikujeme tak, že předpokládáme (aspoň přibližnou) platnost

$$\mathbf{Y}^{(\lambda)} \sim \mathsf{N}\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\big). \tag{11.17}$$

Všechny parametry modelu (vedle $\boldsymbol{\beta}$ a σ^2 také λ) odhadneme metodou maximální věrohodnosti. Uvážíme-li, že platí

$$\frac{\mathrm{d}}{\mathrm{d}y} y^{(\lambda)} = y^{\lambda-1},$$

je logaritmická věrohodnostní funkce netransformovaného náhodného vektoru $\mathbf{Y}$ rovna

$$\ell(\boldsymbol{\beta}, \sigma^2, \lambda) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(y_i^{(\lambda)} - (\mathbf{x}_{i\bullet})'\boldsymbol{\beta}\Big)^2 + n(\lambda-1)\log\dot{Y},$$

kde $\dot{Y}$ je geometrický průměr hodnot $Y_1, \ldots, Y_n$. Pro pevné λ minimalizuje tuto funkci odhad metodou nejmenších čtverců $\mathbf{b}^{(\lambda)}$ v modelu (11.17).

Pokusme se však o poněkud jiné vyjádření, kde by v logaritmické věrohodnostní zmizel (nestandardní) poslední člen. Abychom jej zařadili do prvního členu se σ^2, musíme tento rozptyl nahradit výrazem

$$\left(\frac{\sigma}{\dot{Y}^{\lambda-1}}\right)^2.$$

Tomu ovšem odpovídá úprava součtu čtverců pomocí veličin $Z_i^{(\lambda)} = Y_i^{(\lambda)}/\dot{Y}^{\lambda-1}$ a nového vektoru parametrů $\boldsymbol{\gamma}^{(\lambda)} = (1/\dot{Y}^{\lambda-1})\boldsymbol{\beta}^{(\lambda)}$. Přejdeme tedy pro dané λ formálně k modelu

$$\mathbf{Z}^{(\lambda)} \sim \mathsf{N}\left(\mathbf{X}\boldsymbol{\gamma}^{(\lambda)}, \left(\frac{\sigma}{\dot{Y}^{\lambda-1}}\right)^2 \mathbf{I}\right)$$

a provedeme pouze jednorozměrnou minimalizaci reziduálního součtu čtverců $RSS_Z(\lambda)$ v posledním modelu. Reziduální součet čtverců původního modelu je dán jednoduchým vztahem

$$RSS_Y(\lambda) = \dot{Y}^{2(\lambda-1)} RSS_Z(\lambda),$$

který vyplývá například ze zvolené transformace z $Y^{(\lambda)}$ na $Z^{(\lambda)}$. Když použijeme asymptotickou vlastnost odhadu $\hat{\lambda}$ metodou maximální věrohodnosti a vyjádříme-li hodnotu věrohodnostní funkce pomocí reziduálního součtu čtverců (viz (A.28)), můžeme hledat řešením nerovnosti

$$RSS_Z(\lambda) \leq RSS_Z(\hat{\lambda}) \exp(\chi_1^2(\alpha)/n),$$

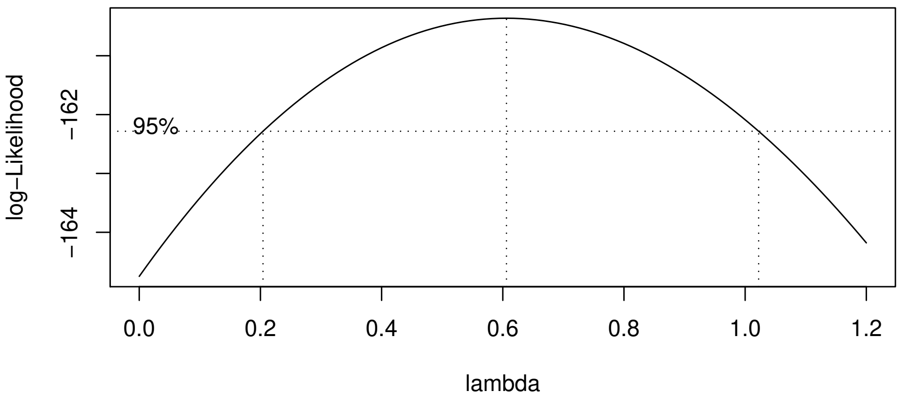
kde $\chi_1^2(\alpha)$ je kritická hodnota rozdělení $\chi^2(1)$, přibližný interval spolehlivosti pro λ.

Příklad 11.1 (procento tuku) V příkladu 8.1 jsme se zabývali závislostí procenta tuku v těle mladých mužů na jejich výšce a hmotnosti. při podrobnější analýze řešení narazíme na možné problémy, kdy je téměř průkazná závislost rozptylu na výšce. Zkusme tedy hledat vhodnou mocninu procenta tuku, kterou bychom vysvětlovali. Použití funkce `boxcox(lm(fat~height+weight,data=Police),plotit=T)` z knihovny `MASS` dá obrázek 11.1, z něhož usuzujeme, že vhodnou volbou bude $\lambda = 0{,}6$, i když hodnota $\lambda = 1$, která znamená identickou transformaci, patří také do 95% intervalu spolehlivosti. Zmíněná knihovna `MASS` doprovází velmi pěknou knihu Venables, Ripley (1997). ○

11.5.2. Žebřík transformací

Při hledání vhodné transformace pro závislost závisle proměnné s kladnými hodnotami na jediné nezávisle proměnné s kladnými hodnotami je užitečnou pomůckou posloupnost mocninných transformací

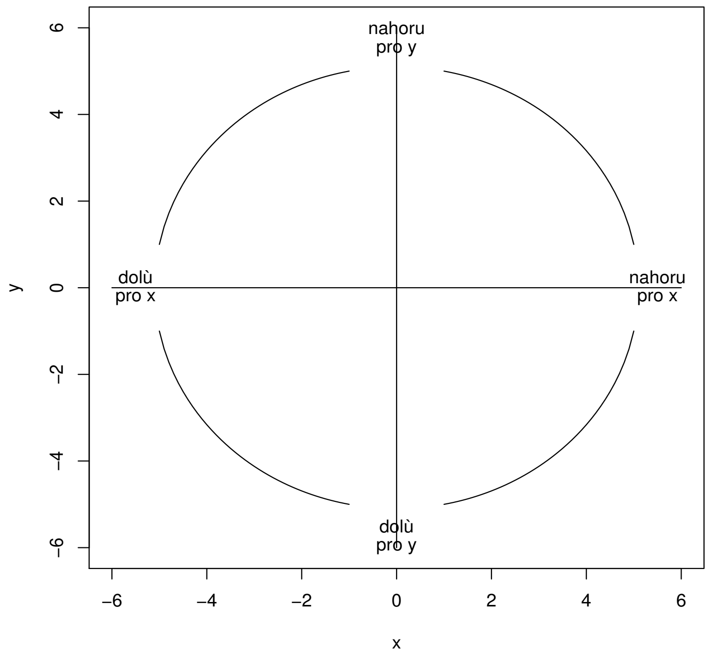
$$\ldots, -1/x^2, -1/x, -1/\sqrt{x}, \log x, \sqrt{x}, x, x^2, \ldots.$$



Obrázek 11.1: Maximálně věrohodný odhad parametru λ Boxovy-Coxovy transformace s vyznačeným 95% intervalem spolehlivosti

Po tomto žebříku transformací se můžeme pohybovat buď nahoru (k vyšším mocninám) nebo dolů. Cílem je především linearizace závislosti. Když dosáhneme pohybem po zvoleném žebříku (na ose x nebo ose y) přibližně lineární závislosti, potom současným pohybem po obou žebřících se pokusíme také o stabilizaci rozptylu.

Při volbě směru pohybu, který má vést k lineárnímu průběhu, je užitečný obrázek 11.2. Například když je závislost konvexní a rostoucí, k linearizaci vede zvyšování mocnin proměnné x nebo snižování mocnin proměnné y.



Obrázek 11.2: Linearizující transformace

12. Logistická regrese

Než přistoupíme k zobecněnému lineárnímu modelu, popíšeme jeho speciální případ, logistickou regresi. U lineárního modelu, kterým jsme se doposud zabývali, byla vysvětlovaná proměnná spojitá. Nyní se pokusíme vysvětlit chování nula-jedničkové veličiny, která modeluje nevýskyt či výskyt sledovaného jevu. Stejně jako u lineárního modelu budeme vyjadřovat střední hodnotu vysvětlované závisle proměnné jako funkci nezávisle proměnných. Tentokrát bude ovšem střední hodnota rovna pravděpodobnosti jedničky, tedy pravděpodobnosti výskytu sledovaného jevu.

12.1. Tvar závislosti

Uvažujme nezávislé náhodné veličiny $Y_1, \dots, Y_n$ s alternativními rozděleními s parametry μ_i. Střední hodnoty jsou totožné s pravděpodobnostmi μ_i, ty mohou záviset na nějakých nenáhodných doprovodných veličinách $\mathbf{x}_{i\bullet}$. Je zřejmé, že platí $\operatorname{\mathsf{var}} Y_i = \mu_i(1-\mu_i)$. To je první podstatný rozdíl v porovnání s normálním lineárním modelem, kde byl rozptyl konstatní.

Pokud bychom předpokládali, jako v lineárním modelu, závislost tvaru

$$\mu_i = \mathsf{E}\, Y_i = \boldsymbol{\beta}'\mathbf{x}_{i\bullet},$$

bude problém s interpretací, protože nelze zaručit, že pro libovolné $\mathbf{x}_{i\bullet}$ bude μ_i ležet v intervalu $(0,1)$. Hledejme tedy jiný interpretovatelný tvar závislosti a motivaci hledejme v odhadech metodou maximální věrohodnosti.

Pravděpodobnosti dvou možných hodnot $Y_i = 1$ a $Y_i = 0$ lze psát souhrnně jako

$$\mathsf{P}(Y_i = j) = \mu_i^j\,(1-\mu_i)^{1-j}\,, \qquad j = 0, 1.$$

Logaritmickou věrohodnostní funkci lze tedy zapsat

$$\begin{aligned}\ell(\boldsymbol{\mu}) &= \log \prod_{i=1}^{n} \mu_i^{Y_i} (1-\mu_i)^{1-Y_i} \\ &= \sum_{i=1}^{n} (Y_i \log \mu_i + (1-Y_i)\log(1-\mu_i)) \qquad (12.1) \\ &= \sum_{i=1}^{n} Y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) + \sum_{i=1}^{n} \log(1-\mu_i).\end{aligned}$$

Jak je vidět, pozorované náhodné veličiny se v logaritmické věrohodnostní funkci projevují pouze v součinech s výrazy $\log(\mu_i/(1-\mu_i))$. Podíl

$$\omega(\mathbf{x}_{i\bullet}) = \frac{\mu_i}{1-\mu_i} = \frac{\mathsf{P}_{\mathbf{x}_{i\bullet}}(Y_i = 1)}{\mathsf{P}_{\mathbf{x}_{i\bullet}}(Y_i = 0)} \qquad (12.2)$$

má bezprostřední interpretaci. Porovnává pravděpodobnost jedničky (výskyt sledovaného jevu) a nuly (nevýskyt jevu). Pro tento podíl se v angličtině používá označení *odds*. Tomu odpovídá české *šance*. Samotné funkci $\eta(\mu) = \log(\mu/(1-\mu))$ se říká *logit*. V kontextu *zobecněných lineárních modelů* (generalized linear model – GLM) je logit speciálním případem *spojovací* funkce (*link function*).

Předpokládejme speciálně, že logit pravděpodobnosti je lineární funkcí neznámých parametrů

$$\eta_i(\boldsymbol{\beta}) = \boldsymbol{\beta}'\mathbf{x}_{i\bullet}.$$

Někdy se i v tomto obecném zápisu explicitně uvádí absolutní člen, protože, jak uvidíme, ne vždy jej budeme schopni odhadnout. Pak se místo regresní matice $\mathbf{X}$ uvádí matice $(\mathbf{1}, \mathbf{X})$.

Střední hodnotu pak v našem modelu můžeme vyjádřit ve tvaru

$$\begin{aligned}\mu_i(\boldsymbol{\beta}) &= \frac{\exp(\eta_i(\boldsymbol{\beta}))}{1+\exp(\eta_i(\boldsymbol{\beta}))} = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{i\bullet})}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_{i\bullet})} \\ &= \frac{1}{1+\exp(-(\boldsymbol{\beta}'\mathbf{x}_{i\bullet}))}, \qquad (12.3)\end{aligned}$$

což zaručí, že platí $0 < \mu_i < 1$ a odstraní jeden z naznačených problémů.

12.2. Odhad parametrů

Naznačme ještě odhad parametrů metodou maximální věrohodnosti. Protože platí

$$\frac{\partial}{\partial \eta_i} \log(1-\mu_i) = -\frac{\partial}{\partial \eta_i} \log(1+\mathrm{e}^{\eta_i}) = -\frac{\mathrm{e}^{\eta_i}}{1+\mathrm{e}^{\eta_i}} = -\mu_i$$

a logaritmickou věrohodnostní funkci jsme upravili na tvar

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \eta_i(\boldsymbol{\beta}) + \sum_{i=1}^{n} \log(1 - \mu_i(\boldsymbol{\beta})),$$

jsou parciální derivace logaritmické věrohodnostní funkce rovny

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} = \sum_{i=1}^{n} (Y_i - \mu_i(\boldsymbol{\beta}))\, \mathbf{x}_{i\bullet}. \tag{12.4}$$

Po malé úpravě zjistíme, že soustavu normálních rovnic (nelineární v $\boldsymbol{\beta}$) lze psát

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{X}' (\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \mathbf{0}, \tag{12.5}$$

Snadno zjistíme, že platí

$$\frac{\partial \mu}{\partial \eta} = \frac{\mathrm{e}^{\eta}}{(1 + \mathrm{e}^{\eta})^2} = \mu(1 - \mu),$$

odkud dostaneme

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \mu_i (1 - \mu_i) \mathbf{x}_{i\bullet}.$$

Když zavedeme diagonální matici rozptylů jednotlivých pozorování

$$\mathbf{D}(\boldsymbol{\beta}) = \mathrm{diag}\{\mu_1(1 - \mu_1), \ldots, \mu_n(1 - \mu_n)\},$$

můžeme Fisherovu informační matici podle (A.30) (vzhledem k (12.4)) zapsat jako

$$\begin{aligned} \mathbf{J}(\boldsymbol{\beta}) &= \mathbf{X}'\mathbf{D}(\boldsymbol{\beta})\mathbf{X} \qquad (12.6) \\ &= \sum_{i=1}^{n} \mu_i(\boldsymbol{\beta})\, (1 - \mu_i(\boldsymbol{\beta}))\, \mathbf{x}_{i\bullet} \mathbf{x}_{i\bullet}'. \end{aligned}$$

Vzhledem k tomu, že matice $\mathbf{D}$ je pozitivně definitní, je Fisherova informační matice přinejmenším pozitivně semidefinitní a v případě úplné sloupcové hodnosti matice $\mathbf{X}$ dokonce pozitivně definitní. Tato skutečnost usnadňuje iterační řešení soustavy normálních rovnic.

Označme řešení normální rovnice (12.5) jako $\mathbf{b}$. Asymptotickou varianční maticí je inverzní matice k Fisherově informační matici. V praxi při jejím výpočtu za neznámé parametry do $\mathbf{J}(\boldsymbol{\beta})$ dosadíme odhady metodou maximální věrohodnosti, které jsou konzistentní, takže také $\mathbf{J}(\mathbf{b})$ je konzistentním odhadem $\mathbf{J}(\boldsymbol{\beta})$. Všimněme si, že, na rozdíl od lineárního modelu, v asymptotické varianční matici nevystupuje parametr měřítka (rozptyl σ^2). Na druhé straně, jak jsme už upozornili, závisí rozptyl odhadů na odhadovaných parametrech $\boldsymbol{\beta}$.

12.3. Interpretace parametrů

Věnujme se interpretaci parametrů β_0, β_1 v modelu $\eta_i = \beta_0 + \beta_1 x_i$.

12.3.1. Binární nezávisle proměnná

Předpokládejme, že jednorozměrná veličina x nabývá právě dvou hodnot. Bez újmy na obecnosti to jsou hodnoty 0 a 1, takže x je umělá proměnná k dvouhodnotovému faktoru a vyjadřuje nepřítomnost či přítomnost nějakého jevu.

Pro $x = 0$ jsou šance rovny

$$\omega(0) = \frac{\mathsf{P}(Y=1)}{\mathsf{P}(Y=0)} = \frac{\dfrac{\exp(\beta_0)}{1+\exp(\beta_0)}}{\dfrac{1}{1+\exp(\beta_0)}} = \mathrm{e}^{\beta_0}. \tag{12.7}$$

Parametr β_0 je tedy roven logitu pravděpodobnosti výskytu sledovaného jevu pro $x = 0$

$$\beta_0 = \log \frac{\mathsf{P}(Y=1)}{\mathsf{P}(Y=0)}. \tag{12.8}$$

Pro $x = 1$ je odpovídající šance rovna

$$\omega(1) = \frac{\dfrac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}}{\dfrac{1}{1+\exp(\beta_0+\beta_1)}} = \mathrm{e}^{\beta_0+\beta_1}. \tag{12.9}$$

Poměr šancí (*odds ratio*) pro dvě hodnoty x je pak roven

$$\frac{\omega(1)}{\omega(0)} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \mathrm{e}^{\beta_1},$$

takže parametr β_1 je roven logaritmu poměru šancí. Pokud pravděpodobnost sledovaného jevu na hodnotě x nezávisí, je poměr šancí roven jedné, tedy $\beta_1 = 0$.

Když známe odhad b_1 parametru β_1 i jeho (asymptotický) rozptyl v_{11} (označme jsme $\mathbf{V} = (\mathbf{J}(b_0, b_1))^{-1}$ s řádky a sloupci číslovanými od nuly), můžeme testovat nulovou hypotézu $\mathsf{H}_0 : \beta_1 = 0$ pomocí statistiky

$$Z = \frac{b_1}{\sqrt{v_{11}}}, \tag{12.10}$$

která má za platnosti nulové hypotézy asymptoticky rozdělení $\mathsf{N}(0,1)$. Některé programy zde předpokládají rozdělení $\mathsf{t}(n-k-1)$ (například STATISTICA), jiné

(NCSS) vycházejí z toho, že za platnosti hypotézy je $Z^2 \sim \chi^2(1)$. V tomto druhém případě se jedná o bezprostřední aplikaci *Waldova testu* (viz (A.32)). Protože při oboustranné alternativě je rozhodování pomocí Z nebo pomocí Z^2 ekvivalentní, o Waldově testu se hovoří také při použití samotného Z.

V případě binárního x nalezneme odhady b_0, b_1 snadno, přímo z odhadů šancí $\omega(0), \omega(1)$. Pro $x = i$ a $Y = j$ označme zjištěnou četnost jako N_{ij}. Celkem tedy máme $n_{i\bullet} = N_{i0} + N_{i1}$ pozorování s hodnotou $x = i$. Hledané odhady jsou

$$\hat{\omega}(0) = \frac{N_{01}/n_{0\bullet}}{N_{00}/n_{0\bullet}} = \frac{N_{01}}{N_{00}},$$
$$\hat{\omega}(1) = \frac{N_{11}/n_{1\bullet}}{N_{10}/n_{1\bullet}} = \frac{N_{11}}{N_{10}}.$$

Odtud snadno dostaneme

$$b_0 = \log\frac{N_{01}}{N_{00}}, \qquad b_1 = \log\frac{N_{00}N_{11}}{N_{01}N_{10}}. \tag{12.11}$$

Pokusme se explicitně vyjádřit rozptyl v_{11}. Diagonální matice $\mathbf{D}(b_0, b_1)$ má pouze dvojí diagonální prvky, $n_{0\bullet}$ prvků s odhadem rozptylu pro $x = 0$ a $n_{1\bullet}$ prvků s odhadem rozptylu pro $x = 1$. Zmíněné odhady rozptylu závisle proměnné jsou rovny $N_{x0}N_{x1}/n_{x\bullet}^2$. Odhad Fisherovy informační matice má tedy tvar

$$\mathbf{J}(b_0, b_1) = \begin{pmatrix} \dfrac{N_{00}N_{01}}{n_{0\bullet}} + \dfrac{N_{10}N_{11}}{n_{1\bullet}} & \dfrac{N_{00}N_{01}}{n_{0\bullet}} \\ \dfrac{N_{00}N_{01}}{n_{0\bullet}} & \dfrac{N_{00}N_{01}}{n_{0\bullet}} \end{pmatrix}.$$

Protože determinant této matice je roven $N_{00}N_{01}N_{10}N_{11}/(n_{0\bullet}n_{1\bullet})$, dostaneme příslušný prvek (vpravo dole) matice $\mathbf{J}(b_0, b_1)^{-1}$ jako

$$v_{11} = \frac{1}{N_{00}} + \frac{1}{N_{01}} + \frac{1}{N_{10}} + \frac{1}{N_{11}}.$$

Doporučuji porovnat výslednou statistiku Z z (12.10) s testováním logaritmických interakcí v (Anděl, 1998, kap. 11.4).

Příklad 12.1 (kojení) Data, upravená podle diplomové práce Hajná (1995), podávají mimo jiné informace o tom, zda matka kojila své dítě ještě ve 24. týdnu. Zabývejme se nejprve tím, zda tato skutečnost závisí na tom, jak matka odpověděla na otázku, zda těhotenství bylo plánováno. Příslušné četnosti dostaneme snadno

```
> print(kt<-xtabs(~Plan+Koj24,data=Kojeni))
    Koj24
Plan  0  1
   0 35  6
   1 36 22
```

Je zřejmé, že u plánovaných těhotenství kojilo ve 24. týdnu života dítěte relativně více matek, než u těhotenství neplánovaných. Proveďme explicitní výpočty, které jsou ekvivalentní hodnocení logaritmických interakcí:

```
> print(b0<-log(kt[1,2]/kt[1,1]))
[1] -1.763589
> print(b1<-log(kt[1,1]*kt[2,2]/(kt[1,2]*kt[2,1])))
[1] 1.271112
> print(v11<-sum(1/kt))
[1] 0.2684704
> print(z<-b1/sqrt(v11));print(2*(1-pnorm(abs(z))))
[1] 2.453215
[1] 0.01415856
```

Velmi podobný výsledek dá program R, který však počítá iteračně:

```
> summary(Koj24.Plan <- glm(Koj24~Plan,family="binomial",data=Kojeni))

Call:
glm(formula = Koj24 ~ Plan, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9767  -0.9767  -0.5625   1.3924   1.9605

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7636     0.4418  -3.992 6.55e-05 ***
Plan1         1.2711     0.5180   2.454   0.0141 *
---

Number of Fisher Scoring iterations: 3
```

Pokud bychom použili proceduru `glm()` s menším parametrem ε, který určuje konec iteračního výpočtu, volbou `control=glm.control(epsilon=1e-11)`, dostali bychom odhady prakticky identické s odhady přímými.

Nyní je ideální příležitost porovnat všechny tři asymptoticky ekvivalentní postupy testování hypotézy popsané v Appedixu A.3. Právě jsme provedli dvěma způsoby *Waldův test*, když jsme porovnali odhad parametru β_1 metodou maximální věrohodnosti s jeho hypotetickou hodnotou. Test poměrem věrohodnosti dostaneme v obdobě tabulky analýzy rozptylu:

```
> anova(Koj24.Plan,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Koj24
```

```
Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                   98    117.930
Plan  1    6.800       97    111.130     0.009
```

Zbývá *metoda skórů (Lagrangeova multiplikátoru)*. Když dosadíme do (12.5) s maticí $\mathbf{X}' = \begin{pmatrix} \mathbf{1}', \mathbf{1}' \\ \mathbf{1}', \mathbf{0}' \end{pmatrix}$, odhad vektoru $\boldsymbol{\beta}$ za nulové hypotézy, totiž

$$\tilde{\mathbf{b}} = \begin{pmatrix} N_{\bullet 1}/n \\ 0 \end{pmatrix},$$

dostaneme vektor skórů (A.33)

$$\mathbf{U}(\hat{\mathbf{b}}) = \frac{\partial \ell(\tilde{\mathbf{b}})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} 0 \\ N_{11} - n_{1\bullet} N_{\bullet 1}/n \end{pmatrix}.$$

Matice $\mathbf{D}(\tilde{\boldsymbol{\beta}})$ z (12.6) je rovna $N_{\bullet 1} N_{\bullet 0}/n^2$ násobku jednotkové matice, takže Fisherova informační matice vyjde

$$\mathbf{J}(\tilde{\mathbf{b}}) = \frac{N_{\bullet 1} N_{\bullet 0}}{n^2} \begin{pmatrix} n & n_{1\bullet} \\ n_{1\bullet} & n_{1\bullet}. \end{pmatrix}$$

Odtud je kvadratická forma statistiky skórového testu (A.34) rovna

$$LM = \frac{n(N_{00}N_{11} - N_{01}N_{10})^2}{n_{0\bullet} n_{1\bullet} N_{\bullet 1} N_{\bullet 0}},$$

což je známá χ^2-statistika k hodnocení čtyřpolní tabulky. V našem příkladě tak dostaneme (bez Yatesovy korekce)

```
> chisq.test(kt,correct=F)

        Pearson's Chi-squared test

data:  kt
X-squared = 6.4273, df = 1, p-value = 0.01124
```

Pro zajímavost, Fisherův exaktní (faktoriálový) test dá v našem případě hodnotu $p = 0{,}01310$. ○

12.3.2. Spojitá nezávisle proměnná

Také tentokrát má zajímavou interpretaci opět především parametr β_1. Uvažujme jedinou nezávisle proměnnou, zobecnění pro více proměnných se provede obdobně jako v případě lineární regrese. Především, samotná šance je při zvolené hodnotě x rovna

$$\omega(x) = \frac{\dfrac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}}{\dfrac{1}{1 + \exp(\beta_0 + \beta_1 x)}} = \mathrm{e}^{\beta_0 + \beta_1 x}.$$

Porovnejme šance sledovaného výsledku pokusu pro dvě hodnoty nezávisle proměnné, které se liší o jednotku:

$$\frac{\omega(x+1)}{\omega(x)} = \frac{\mathrm{e}^{\beta_0 + \beta_1 (x+1)}}{\mathrm{e}^{\beta_0 + \beta_1 x}} = \mathrm{e}^{\beta_1}.$$

Opět tedy β_1 vypovídá o změně vztažené k jednotkovému přírůstku nezávisle proměnné x, tentokrát je to změna logaritmu poměru šancí. Stejná je také nulová hypotéza, kterou nejčastěji testujeme. Hypotéza $\mathsf{H}_0 : \beta_1 = 0$ odpovídá stejným šancím při obou hodnotách nezávisle proměnné, tedy nezávislosti šancí (a tudíž pravděpodobnosti $\mathsf{P}(Y = 1)$) na nezávisle proměnné x. Opět lze použít Waldův test založený na statistice (12.10), která porovnává odhad β_1 s jeho asymptotickou střední chybou.

Příklad 12.2 (kojení) Podívejme se nyní na možnou závislost skutečnosti, že matka ve 24. týdnu věku dítěte ještě kojí, na jejím věku.

```
> summary(Koj24.vek.m<-glm(Koj24~vek.m,family="binomial",data=Kojeni))

Call:
glm(formula = Koj24 ~ vek.m, family = "binomial", data = Kojeni)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2733  -0.8104  -0.6838   1.2210   1.9431

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.31866    1.48629  -2.906  0.00366 **
vek.m        0.12975    0.05551   2.337  0.01942 *
---
Number of Fisher Scoring iterations: 3
```

Závislost je tedy na 5% hladině průkazná. S každým rokem věku matky roste šance, že matka dlouho kojí, násobí se koeficientem exp(0,12975) = 1,138548. S každým rokem věku tedy šance v průměru roste téměř o 13 %. ○

12.4. Testování podmodelu

K testování podmodelu lze použít test poměrem věrohodnosti (A.31) popsaný v apendixu založený na odhadech $\mathbf{b}$ v modelu a $\tilde{\mathbf{b}}$ v podmodelu. Test se však zpravidla provádí prostřednitvím tzv. deviancí, které nyní zavedeme.

Uvažujme nejprve nejbohatší možný model, který má právě tolik parametrů, kolik je pozorování, tedy n. Přiléhavější model (s větší hodnotou věrohodnostní funkce) neexistuje. Tento nejbohatší model se nazývá *saturovaný*. Označme maximální hodnotu věrohodnostní funkce v saturovaném modelu symbolem $\ell_{\max}$. Každý jiný představitelný model je podmodelem saturovaného modelu. Přiléhavost běžného modelu můžeme posoudit pomocí *deviance*

$$D(\mathbf{b}) = 2\left(\ell_{\max} - \ell(\mathbf{b})\right). \tag{12.12}$$

Čím je náš model méně přiléhavý, tím je hodnota deviance D větší, podobně, jako je větší reziduální součet čtverců v lineárním modelu v méně výstižném modelu.

Pojem deviance jsme zde zavedli zejména proto, že se používá i v souvislosti s logistickou regresí, byť je zde hodnota $\ell_{\max}$ triviální. Ve jiných zobecněných lineárních modelech (viz kapitolu 13) slouží použití $\ell_{\max}$ k odstranění nadbytečných konstant. Saturovaný model má n parametrů $\mu_1, \ldots, \mu_n$. Odhadem střední hodnoty μ_i je v případě logistické regrese přímo Y_i, takže podle (12.1) je

$$\ell_{\max} = \sum_{i=1}^{n} \left(Y_i \log Y_i + (1 - Y_i)\log(1 - Y_i)\right) = 0. \tag{12.13}$$

Označíme-li odhady pravděpodobnosti jedničky v běžném modelu jako $\hat{\mu}_i = \mu(\mathbf{x}_{i\bullet})$, devianci v modelu logistické regrese vyjádříme s pomocí (12.13) jako

$$\begin{aligned} D(\mathbf{b}) &= 2\sum_{i=1}^{n}\left(Y_i \log\frac{Y_i}{\hat{\mu}_i} + (1 - Y_i)\log\left(\frac{1 - Y_i}{1 - \hat{\mu}_i}\right)\right) \\ &= -2\sum_{i=1}^{n}\left(Y_i \log\hat{\mu}_i + (1 - Y_i)\log(1 - \hat{\mu}_i)\right). \end{aligned} \tag{12.14}$$

Vraťme se k obecné situaci. V našem běžném modelu teď uvažujme nějaký podmodel například po vyloučení části regresorů. Testovou statistiku testu poměrem věrohodnosti (viz (A.31)) lze vyjádřit pomocí deviance modelu a podmodelu (s odhady parametrů $\tilde{b}_0, \tilde{\mathbf{b}}$)

$$\begin{aligned} 2\Big(\ell(\mathbf{b}) - \ell(\tilde{\mathbf{b}})\Big) &= \Big(2\big(\ell_{\max} - \ell(\tilde{\mathbf{b}})\big)\Big) - \Big(2\Big(\ell_{\max} - \ell(\mathbf{b})\Big)\Big) \\ &= D(\tilde{\mathbf{b}}) - D(\mathbf{b}). \end{aligned}$$

Tato testová statistika (rozdíl deviancí) má (za platnosti testovaného podmodelu) asymptoticky rozdělení $\chi^2(f)$, kde f je rovno rozdílu počtů nezávislých parametrů v porovnávaných modelech.

Je tedy zřejmé, že hypotézu $\mathsf{H}_0 : \beta_1 = 0$ (a podobné hypotézy o nulovosti jediné složky vektoru $\boldsymbol{\beta}$) lze testovat nejen pomocí zmíněného Waldova testu, ale také testem poměrem věrohodnosti. V literatuře (viz Hauck, Donner (1977)) lze nalézt zjištění, podle kterých může být v tomto případě test poměrem věrohodnosti (využívající deviance) silnější, zvláště když je skutečnost daleko od nulové hypotézy.

Podobnost deviance k reziduálnímu součtu čtverců vedla ke snaze rozšířit pojem koeficientu determinace také na logistickou regresi. K tomu účelu nejprve zavedeme pojem *nulového modelu*. Jde o model, kde jsou všechny střední hodnoty $\mu_i = \mathsf{E}\,Y_i$ shodné. Hodnotu věrohodnostní funkce či deviance označíme ℓ_0 resp. D_0. V appendixu jsme v (A.28) vyjádřili hodnotu logaritmické věrohodnostní funkce normálního lineárního modelu jako

$$\ell(\mathbf{b}) = -\frac{n}{2}\left(1 + \log(2\pi) - \log n\right) - \frac{n}{2}\log RSS.$$

Koeficient determinace v lineárním modelu lze vyjádřit jako

$$\begin{aligned} R^2 &= 1 - \frac{RSS}{RSS_0} \\ &= 1 - \exp\left(-\frac{2}{n}(\ell(\mathbf{b}) - \ell_0)\right) \\ &= 1 - \exp\left(\frac{1}{n}(D(\mathbf{b}) - D_0)\right) \end{aligned} \tag{12.15}$$

V (12.15) je uveden návod k výpočtu i pro případ logistické regrese. Přiléhavější model, než je saturovaný, nalézt nelze. Deviance saturovaného modelu je, jak víme, rovna nule, takže koeficient z (12.15) nemůže překročit hodnotu

$$R^2_{\max} = 1 - \exp\left(-\frac{1}{n}D_0\right).$$

Po dosazení do (12.14) dostaneme

$$D_0 = -2\left(\sum_{i=1}^n Y_i \log\left(\sum_{i=1}^n Y_i\right) + \left(n - \sum_{i=1}^n Y_i\right)\log\left(n - \sum_{i=1}^n Y_i\right) - n\log n\right),$$

neboť pro všechna i je odhadem střední hodnoty relativní četnost jedniček, totiž $\sum_{i=1}^n Y_i/n$.

Nagelkerke (1991) proto navrhl upravit definici zobecněného koeficientu determinace na

$$R^2_{\mathrm{N}} = \frac{R2}{R^2_{\max}} \tag{12.16}$$

$$= \frac{1 - \exp\left((D(\mathbf{b}) - D_0)/n\right)}{1 - \exp\left(-D_0/n\right)}. \tag{12.17}$$

Příklad 12.3 (kojení) Data, upravená podle diplomové práce (Hajná (1995)), podávají mimo jiné informace o tom, zda matka kojila své dítě ještě ve 24. týdnu. Zabývejme se nejprve tím, zda závisí na nejvyšším dosaženém vzdělání matky.

```
> anova(a.Vzdelani<-glm(Koj24~Vzdelani,family="binomial"),test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Koj24

Terms added sequentially (first to last)

         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                        98     117.93
Vzdelani  2     3.32        96     114.61      0.19
```

Příkazem `summary` místo `anova` bychom dostali bodové odhady složek $\boldsymbol{\beta}$, které nás však v tomto „zobecněném“ modelu analýzy rozptylu nezajímají (závisejí, kromě jiného, na zvolených kontrastech). Všimněme si poklesu deviance z hodnoty 114,61 v modelu, kdy rozlišujeme tři úrovně vzdělání matky na 117,93 v modelu, kdy je pravděpodobnost toho, že matka kojí ještě ve 24. týdnu nezávislá na jejím vzdělání. Rozdíl deviancí vede k testové statistice s hodnotou 3,32, což při 2 stupních volnosti dá dosaženou hladinu $p = 0{,}19$. Neprokázali jsme tedy závislost na vzdělání matky.

Zjištěné deviance lze zhodnotit také pomocí Nagelkerkeho zobecněného koeficientu determinace (procedura je uvedena v oddílu B.2.1).

```
> nagel.R2(a.Vzdelani)
[1] 0.04736801
> nagel.R2(a.Vzdelani,Nagelkerke=F)
[1] 0.03297507
```

Porovnáním Nagelkerkeho R^2_{N} s běžným R^2 zjistíme, že je

$$R^2_{\max} = 0{,}032975/0{,}047368 = 0{,}6961465,$$

což je možno zjistit i přímým výpočtem, přičemž si jistě uvědomíme, že vystačíme s hodnotami závisle proměnné:

```
> print(n0<-summary(Koj24)[1])
 0
71
> print(n1<-summary(Koj24)[2])
 1
28
> print(n<-n0+n1)
 0
99
```

```
> print(D0<- -2*(n1*log(n1/n)+n0*log(n0/n)))
       1
117.9297
> print(R2.max<-1-exp(-D0/n))
        1
0.6961465
```

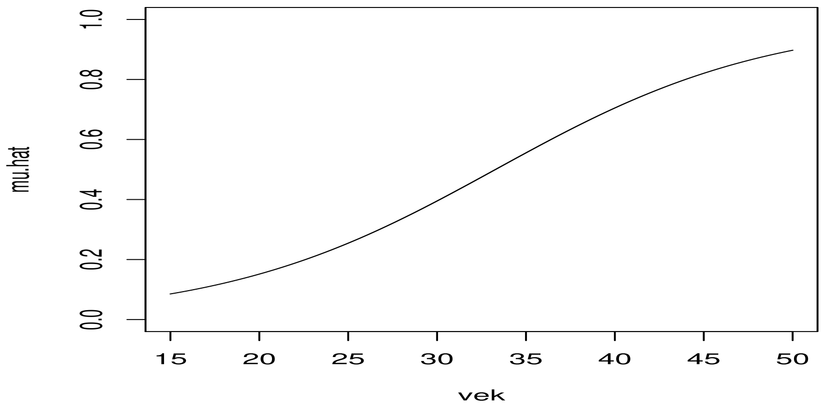
Na okraj poznamenejme, že rozhodnout jsme mohli také pomocí kontingenční tabulky (s použitím knihovny `ctest`):

```
> chisq.test(table(Vzdelani,Koj24))

         Pearson's Chi-square test

data:  table(Vzdelani, Koj24)
X-squared = 3.2001, df = 2, p-value = 0.2019
```

Oba testy jsou asymptoticky ekvivalentní. ◯

Příklad 12.4 (kojení) Použijme znovu stejná data a rozhodněme, zda stejná závisle proměnná jako v předchozím příkladu závisí na věku matky.

```
> summary(a.vek.m<-glm(Koj24~vek.m,family="binomial"))

Call:
glm(formula = Koj24 ~ vek.m, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2733  -0.8104  -0.6838   1.2210   1.9431

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.31866    1.48629  -2.906  0.00366 **
vek.m        0.12975    0.05551   2.337  0.01942 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 117.93  on 98  degrees of freedom
Residual deviance: 112.20  on 97  degrees of freedom
AIC: 116.20

Number of Fisher Scoring iterations: 3

> nagel.R2(a.vek.m)
[1] 0.08083891

> anova(a.vek.m,test="Chisq")
Analysis of Deviance Table
```



Obrázek 12.1: Odhadnutá závislost na věku pro pravděpodobnost, že matka ve 24. týdnu věku dítěte ještě kojí

```
Model: binomial, link: logit

Response: Koj24

Terms added sequentially (first to last)

      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     98    117.930
vek.m  1    5.734        97    112.196     0.017
```

Tentokrát jsme použili nejprve podrobnější výstup (`summary`), protože nás zajímá také odhad koeficientu β_1. Zjistili jsme, že na každý rok věku matky se šance ke kojení ve 24. týdnu zvětšuje (násobí) $\mathrm{e}^{0{,}12975} \doteq 1{,}14$, tedy asi o 14 %. Pozor však na interpretaci, je nutno rozlišovat mezi šancí a pravděpodobností. Když porovnáváme matku třicetiletou s matkou dvacetiletou, tak šance stoupnou přibližně 3,66 krát (o 266 %), neboť je $\mathrm{e}^{10\cdot 0{,}12975} = \mathrm{e}^{1{,}2975} \doteq 3{,}66$. Porovnání pravděpodobností závisí také na odhadu parametru β_0.

Když považujeme skupinu 99 matek za reprezentativní vzorek všech našich matek, pak můžeme počítat i odhad pravděpodobnosti sledovaného jevu pro daný věk t. Vyjde

$$\hat{\mathsf{P}}_t(Y = 1) = \hat{\mu}(t) = \frac{\mathrm{e}^{-4{,}31866+0{,}12975t}}{1 + \mathrm{e}^{-4{,}31866+0{,}12975t}}.$$

O průběhu této funkce si lze učinit představu z obrázku 12.1, v němž jsme pro názornost vědomě zvolili větší rozsah hodnot nezávisle proměnné, než jaké máme v datech (tam se věk matky pohybuje od 18 do 38 roků). Pro dvacetiletou matku dostaneme odhad $\hat{\mu}(20) = 0{,}151$, kdežto pro třicetiletou matku $\hat{\mu}(30) = 0{,}395$. Podíl pravděpodobností je pouze asi 2,61.

O průkaznosti závislosti na věku můžeme rozhodnout dvěma způsoby, oba dají podobné výsledky s dosaženými hladinami $p = 0{,}017$ (test poměrem věrohodnosti, vypočteno příkazem `anova()`) a $p = 0{,}019$ (Waldův test, v řádku s odhadem příkazem `glm()`). Závislost je tedy na 5% hladině průkazná. $\bigcirc$

12.5. Tři druhy studií

V této části se budeme věnovat zvláštnosti modelu logistické regrese, která v určité situaci umožní odhadnout vektor $\boldsymbol{\beta}$ aniž by byl odhadnutelný také parametr β_0. Půjde o tzv. *studie případů a kontrol* (case-control).

Zhruba můžeme tři různé v úvahu připadající situace ukázat na jednoduchém příkladu jediné dvouhodnotové nezávisle proměnné. Budeme se zabývat možnostmi získat data a s tím související možností odhadovat parametr β_0. Jedná se o situaci, kdy bychom mohli použít čtyřpolní tabulku. Pro naši diskusi bude rozhodující, zda jsou marginální četnosti v této tabulce pevné nebo náhodné. Kvalitu marginálních četností vyjadřujeme v následujících tabulkách velikostí písmen (pevné četnosti jako n, kdežto četnosti náhodné jako N. Terminologie pochází z epidemiologie, avšak podobné úlohy se vyskytují i jinde.

	Y		
X	0	1	
0	N_{00}	N_{01}	$n_{0\bullet}$
1	N_{10}	N_{11}	$n_{1\bullet}$
	$N_{\bullet 0}$	$N_{\bullet 1}$	n

prospektivní studie

	Y		
X	0	1	
0	N_{00}	N_{01}	$N_{0\bullet}$
1	N_{10}	N_{11}	$N_{1\bullet}$
	$N_{\bullet 0}$	$N_{\bullet 1}$	n

průřezová studie

	Y		
X	0	1	
0	N_{00}	N_{01}	$N_{0\bullet}$
1	N_{10}	N_{11}	$N_{1\bullet}$
	$n_{\bullet 0}$	$n_{\bullet 1}$	n

retrospektivní studie

12.5.1. Prospektivní studie (cohort)

Vybíráme ze dvou populací například podle toho, zda sledované osoby byly či nebyly zatíženy působením rizikového faktoru. Protože jsme sami rozhodli o velikosti obou výběrů, jsou řádkové marginální četnosti pevné. Jde o podobnou situaci, jaká vede například k dvouvýběrovému t testu. Obojí šance mají jednoduché vyjádření

$$\omega(0) = \frac{\mathsf{P}_0(Y = 1)}{\mathsf{P}_0(Y = 0)}, \qquad \omega(1) = \frac{\mathsf{P}_1(Y = 1)}{\mathsf{P}_1(Y = 0)},$$

přičemž dolním indexem vždy vyjadřujeme příslušnost k té které populaci. Odhady šancí i obou parametrů jsme již uvedli v (12.11). Je zřejmé, že odhady šancí jsou identické s odhady v průřezové studii uvedenými v (12.18). Proto jsou identické i odhady parametrů β_0, β_1.

Je třeba poznamenat, že prospektivní studie bývají velmi náročné, zejména na čas (a tudíž i peníze). V prospektivní studii se pořídí dva výběry (kohorty), s rizikovým faktorem a bez něho, a pak se zvolenou dobu prvky obou výběrů sledují. Zvláště tehdy, když je pravděpodobnost události malá (doufejme, že to platí například pro AIDS), je třeba mít výběry značně velké. Navíc je zpravidla nutno vypořádat se s problémem chybějících pozorování.

Ze statistického pohledu jsme data sbírali klasickým způsobem. Hodnotu nejdůležitější nezávisle proměnné v modelu (12.3) považujeme za pevnou, konkrétními hodnotami ostatních regresorů další postup podmíníme.

12.5.2. Průřezová studie (cross-sectional)

Náhodné jsou obojí marginální četnosti, což je případ studie, kdy z populace náhodně vybíráme objekty, na nich zjišťujeme hodnoty obou veličin X, Y. Přesto, že jsou tyto veličiny náhodné, zajímáme se o regresní model, tedy o závislost střední hodnoty Y podmíněnou danou hodnotou $X = x$. Postupně vyjádříme šance $\omega(0)$ a $\omega(1)$ z (12.7) a (12.9).

$$\omega(0) = \frac{\mathsf{P}(Y=1|X=0)}{\mathsf{P}(Y=0|X=0)} = \frac{\mathsf{P}(Y=1,X=0)/\mathsf{P}(X=0)}{\mathsf{P}(Y=0,X=0)/\mathsf{P}(X=0)} = \frac{\mathsf{P}(Y=1,X=0)}{\mathsf{P}(Y=0,X=0)},$$
$$\omega(1) = \frac{\mathsf{P}(Y=1|X=1)}{\mathsf{P}(Y=0|X=1)} = \frac{\mathsf{P}(Y=1,X=1)/\mathsf{P}(X=1)}{\mathsf{P}(Y=0,X=1)/\mathsf{P}(X=1)} = \frac{\mathsf{P}(Y=1,X=1)}{\mathsf{P}(Y=0,X=1)}.$$

Je zřejmé, že všechny čtyři pravděpodobnosti, které se vyskytují v těchto výrazech, lze snadno odhadnout z relativních četností, takže obě šance snadno odhadneme

$$\hat{\omega}(0) = \frac{N_{01}}{N_{00}}, \qquad \hat{\omega}(1) = \frac{N_{11}}{N_{10}}. \tag{12.18}$$

Odtud jsou odhady parametrů saturovaného modelu (je jasné, proč je saturovaný?) jsou dány

$$b_0 = \log\frac{N_{01}}{N_{00}}, \qquad b_1 = \log\frac{N_{11}N_{00}}{N_{01}N_{10}}.$$

Můžeme tedy ke zvolené hodnotě x odhadnout odpovídající pravděpodobnost jevu $Y = 1$.

V obecném modelu (12.3) platí poznámka o podmiňování pro všechny zúčastněné regresory.

12.5.3. Retrospektivní studie (case-control)

Tentokrát jsou dvě porovnávané populace dány hodnotou závisle proměnné. Pořídíme z nich dva výběry a *retrospektivně* sledujeme, zda (někdy v minulosti) nastalo $X = 1$ nebo $X = 0$. Pevné jsou tedy sloupcové marginální četnosti. Výklad založený na Bayesově vzorci je poněkud komplikovaný, provedeme jej obecněji než jen pro dvouhodnotový regresor.

Máme k disposici $n_{\bullet 1}$ pozorování (objektů, pacientů) s hodnotou $Y_i = 1$ a $n_{\bullet 0} = n - n_{\bullet 1}$ pozorování s hodnotou $Y_i = 0$. Skutečnost, že nějaký objekt se dostal do našeho výběru, modelujeme tak, že jemu odpovídající hodnota indikátorové náhodné veličiny V je rovna jedničce, jinak je nulová. Nyní učiníme důležitý předpoklad, že pravděpodobnost zahrnutí do výběru nezávisí na hodnotě nezávisle proměnné $\mathbf{X}$ ($\mathbf{X}, V$ jsou nezávislé), tedy

$$\begin{aligned}
\mathsf{P}(V = 1|Y = 1, \mathbf{X} = \mathbf{x}) &= \mathsf{P}(V = 1|Y = 1) = \vartheta_1,\\
\mathsf{P}(V = 1|Y = 0, \mathbf{X} = \mathbf{x}) &= \mathsf{P}(V = 1|Y = 0) = \vartheta_0.
\end{aligned}$$

Pravděpodobnosti zahrnutí nejsou pochopitelně stejné. Zvláště u řídce se vyskytujících nemocí musíme zvolit parametr ϑ_1 veliký, kdežto při výběru kontrol je populace odkud vybíráme veliká, takže pravděpodobnost zahrnutí ϑ_0 lze zvolit poměrně malou.

Předpokládejme ještě, že opravdu platí logistický model, tedy podmíněná pravděpodobnost jevu $Y = 1$ při daném $\mathbf{x}$ (osoba s hodnotou nezávisle proměnných $\mathbf{x}$ onemocní) je dána vztahem (tentokrát je třeba explicitně rozlišit absolutní člen od ostatních složek vektoru β)

$$\mathsf{P}(Y = 1|\mathbf{X} = \mathbf{x}) = \mu(\mathbf{x}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}'\mathbf{x})}{1 + \exp(\beta_0 + \boldsymbol{\beta}'\mathbf{x})}.$$

Použití Bayesovy věty dá pro vybrané objekty

$$\mathsf{P}(Y = 1|\mathbf{X} = \mathbf{x}, V = 1) = \frac{\mathsf{P}(V = 1|Y = 1, \mathbf{X} = \mathbf{x})\mu(\mathbf{x})}{\mathsf{P}(V = 1|\mathbf{X} = \mathbf{x})},$$

kde je

$$\begin{aligned}
\mathsf{P}(V = 1|\mathbf{X} = \mathbf{x}) = \mathsf{P}(V = 1|Y = 1, \mathbf{X} = \mathbf{x})\mu(\mathbf{x})&\\
+ \mathsf{P}(V = 1|Y = 0, \mathbf{X} = \mathbf{x})(1 - \mu(\mathbf{x})).&
\end{aligned}$$

Po dosazení pravděpodobností zahrnutí dostaneme

$$
\begin{aligned}
\mathsf{P}(Y=1|\mathbf{X}=\mathbf{x},V=1) &= \frac{\vartheta_1\mu(\mathbf{x})}{\vartheta_1\mu(\mathbf{x})+\vartheta_0\,(1-\mu(\mathbf{x}))}\\
&= \frac{\dfrac{\vartheta_1}{\vartheta_0}\dfrac{\mu(\mathbf{x})}{1-\mu(\mathbf{x})}}{1+\dfrac{\vartheta_1}{\vartheta_0}\dfrac{\mu(\mathbf{x})}{1-\mu(\mathbf{x})}}\\
&= \frac{\exp\left(\beta_0+\log(\vartheta_1/\vartheta_0)+\boldsymbol{\beta}'\mathbf{x}\right)}{1+\exp\left(\beta_0+\log(\vartheta_1/\vartheta_0)+\boldsymbol{\beta}'\mathbf{x}\right)}\\
&= \frac{\exp\left(\beta_0^*+\boldsymbol{\beta}'\mathbf{x}\right)}{1+\exp\left(\beta_0^*+\boldsymbol{\beta}'\mathbf{x}\right)}.
\end{aligned}
$$

Výsledkem je tedy opět logistický model pro pravděpodobnost náhodného jevu $Y=1$, dokonce se stejným vektorem parametrů u vektoru $\mathbf{x}$, ale s modifikovaným absolutním členem. Z dat zřejmě odhadneme parametry β_0^* a $\boldsymbol{\beta}$. Abychom odhadli také původní parametr β_0, musíme znát poměr pravděpodobností zahrnutí ϑ_1/ϑ_0. Zpravidla bude platit $\vartheta_1 >> \vartheta_0$, takže je absolutní člen β_0^* větší než β_0. Zdánlivě je tedy pro dané $\mathbf{X}=\mathbf{x}$ pravděpodobnost jevu $Y=1$ mnohem větší, než je ve skutečnosti.

12.6. Diagnostika

Také v logistické regresi jsou rezidua účinným diagnostickým nástrojem. Rezidua lze zde zavést několika způsoby, které dají v případě normálního lineárního modelu identické výsledky.

Označme jako $\hat{\mu}_i=\hat{\mu}(\mathbf{x}_{i\bullet})$ odhad pravděpodobnosti $Y=1$ pro dané $\mathbf{x}=\mathbf{x}_{i\bullet}$. Rezidua známá z lineárního modelu, tedy rozdíly $u_i=Y_i-\hat{\mu}_i$, se v logistické regresi anglicky nazývají *response residuals*. Za standardní jsou považována spíše *deviančnί rezidua* definovaná jako

$$
{}_{\text{dev}}u_i=\sqrt{-2\left(Y_i\log\hat{\mu}_i)+(1-Y_i)\log(1-\hat{\mu}_i)\right)}\ \operatorname{sign}(Y_i-\hat{\mu}_i).
$$

Součet čtverců těchto reziduí je roven devianci modelu, jak se snadno přesvědčíme porovnáním se vztahem (12.14).

Někdy se používají *Pearsonova rezidua*, která dostaneme, když obyčejné reziduum vydělíme odhadem směrodatné odchylky Y_i

$$
{}_{\text{Pearson}}u_i=\frac{Y_i-\hat{\mu}_i}{\sqrt{\hat{\mu}_i(1-\hat{\mu}_i)}}.
$$

Nevýhodou těchto reziduí je skutečnost, že v logistické regresi mají více šikmé rozdělení, než rezidua deviančí. Součet čtverců těchto reziduí dá běžnou statistiku χ^2. U té však v obecné logistické regresi se spojitou nezávisle proměnnou s mnoha hodnotami nelze očekávat rozdělení blízké rozdělení χ^2. Podobně jako devianci lze chápat hodnotu statistiky χ^2 pouze jako míru přiléhavosti modelu.

13. Zobecněný lineární model

Tato kapitola do značné míry vychází z poznámek k přednášce dr. Kulicha (Kulich (2000)) a z přehledné 2. kapitoly monografie McCullagh, Nelder (1989).

13.1. Rozdělení exponenciálního typu

Podobnou úvahu, jako jsme udělali v logistické regresi pro alternativní rozdělení, lze provést i pro mnohá další rozdělení.

Uvažujme rozdělení exponenciálního typu s rozptylovým parametrem φ, které má hustotu (u diskrétního rozdělení pravděpodobnostní funkci) tvaru

$$f(y, \theta, \varphi) = \exp\left(\frac{y \cdot \theta - b(\theta)}{a(\varphi)} + c(y, \varphi)\right), \tag{13.1}$$

kde $c(x, \varphi)$ nezávisí na přirozeném parametru θ, $a(.)$ je kladná funkce. U funkce $b(\theta)$ předpokládáme spojitost druhé derivace. Abychom určili střední hodnotu a rozptyl náhodné veličiny Y s hustotou (13.1), najdeme nejprve momentovou vytvořující funkci. Integrační obor A je dán jen těmi y, pro která je hustota kladná.

$$\begin{aligned}
M(t) = \mathsf{E}\, \mathrm{e}^{tY} &= \int_A \exp(ty) \exp\left(\frac{y \cdot \theta - b(\theta)}{a(\varphi)} + c(y, \varphi)\right) \mathrm{d}y \\
&= \int_A \exp\left(\frac{y(\theta + a(\varphi)t) - b(\theta)}{a(\varphi)} + c(y, \varphi)\right) \mathrm{d}y \\
&= \exp\left(\frac{b(\theta + a(\varphi)t) - b(\theta)}{a(\varphi)}\right) W(\theta, \varphi) \\
&= \exp\left(\frac{b(\theta + a(\varphi)t) - b(\theta)}{a(\varphi)}\right),
\end{aligned}$$

neboť platí

$$W(\theta, \varphi) = \int_A \exp\left(\frac{y(\theta + a(\varphi)t) - b(\theta + a(\varphi)t)}{a(\varphi)} + c(y, \varphi)\right) \mathrm{d}y = 1$$

protože jde o integrál z hustoty.

K určení obou momentů potřebujeme spočítat první dvě derivace momentové vytvořující funkce:

$$
\begin{aligned}
M'(t) &= \exp\left(\frac{b(\theta + a(\varphi)t) - b(\theta)}{a(\varphi)}\right) b'(\theta + a(\varphi)t),\\
M''(t) &= \exp\left(\frac{b(\theta + a(\varphi)t) - b(\theta)}{a(\varphi)}\right) (b'(\theta + a(\varphi)t))^2 + \\
&\quad \exp\left(\frac{b(\theta + a(\varphi)t) - b(\theta)}{a(\varphi)}\right) b''(\theta + a(\varphi)t)a(\varphi),
\end{aligned}
$$

takže je

$$
\begin{aligned}
\mathsf{E}\, X &= b'(\theta), && (13.2)\\
\mathsf{var}\, X &= a(\varphi)b''(\theta)\\
&= a(\varphi)V(\theta). && (13.3)
\end{aligned}
$$

Jak vidíme, funkce $b(.)$ umožňuje popsat momenty resp. kumulanty rozdělení Y, proto se někdy nazývá *kumulantová* funkce. Funkce $V(.)$ ukazuje nakolik souvisí rozptyl vysvětlované veličiny Y s její střední hodnotou. Parametr φ zpravidla považujeme za *rušivý*.

Všimněme si některých důležitých rozdělení exponenciálního typu.

13.1.1. Normální rozdělení

V regresi se zabýváme vysvětlením chování střední hodnoty, takže nás zajímá parametr μ, kdežto σ^2 chápeme jako rušivý parametr. Hustotu rozdělení $\mathsf{N}(\mu, \sigma^2)$ můžeme postupně upravit

$$
\begin{aligned}
f(y;\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)\\
&= \exp\left(-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right)\\
&= \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\log(2\pi\sigma^2) + \frac{y^2}{\sigma^2}\right)\right).
\end{aligned}
$$

Zvolíme $\theta = \mu, b(\theta) = \mu^2/2, a(\varphi) = \varphi$ a $\varphi = \sigma^2$, dostaneme $V(\theta) = b''(\theta) = 1$.

13.1.2. Binomické rozdělení

Pravděpodobnostní funkci (hustotu vzhledem k čítací míře) binomického rozdělení $\mathsf{bi}(n, \pi)$ lze vyjádřit jako

$$\begin{aligned} f(y; \pi) &= \binom{m}{y} \pi^y (1-\pi)^{m-\pi} \\ &= \exp\left(y \log\left(\frac{\mu}{1-\mu}\right) + n \log(1-\pi) + \log\binom{m}{y}\right). \end{aligned}$$

Zvolíme

$$\theta = \frac{\pi}{1-\pi}, \qquad b(\theta) = m \log(1 + \mathrm{e}^\theta), \qquad a(\varphi) = \varphi, \qquad \varphi = 1,$$

takže dostaneme

$$\begin{aligned} \mathsf{E}\, Y = \mu = b'(\theta) &= m \frac{\mathrm{e}^\theta}{1+\mathrm{e}^\theta} = m\pi, \\ V(\theta) = b''(\theta) &= m \frac{\mathrm{e}^\theta}{(1+\mathrm{e}^\theta)^2} = m\pi(1-\pi). \end{aligned}$$

Přirozeným parametrem θ je zřejmě logit, který jsme zavedli v logistické regresi. Nemůže to překvapit, neboť tam uvažovaná závisle proměnná Y má alternativní rozdělení, které je speciálním případem rozdělení binomického.

13.1.3. Poissonovo rozdělení

Pravděpodobnostní funkci (hustotu vzhledem k čítací míře) Poissonova rozdělení $\mathsf{Po}(\lambda)$ lze vyjádřit jako

$$\begin{aligned} f(y; \lambda) &= \frac{\lambda^y}{y!} \mathrm{e}^{-\lambda} \\ &= \exp\left(y \log \lambda - \lambda - \log y!\right). \end{aligned}$$

Opět jde o kanonický tvar hustoty, tentokrát s přirozeným parametrem $\log \lambda$. Volíme tedy

$$\theta = \log \lambda, \qquad b(\theta) = \mathrm{e}^\theta, \qquad a(\varphi) = \varphi, \qquad \varphi = 1.$$

Odtud je

$$V(\theta) = b''(\theta) = \mathrm{e}^\theta.$$

13.2. Zobecněný lineární model

Zaveďme nyní pojem *zobecněného lineárního modelu* (GLM – generalized linear model). Předpokládejme, že $Y_1, \ldots, Y_n$ jsou nezávislé náhodné veličiny, jejichž rozdělení závisí na pevných vektorech $\mathbf{x}_i \in \mathbb{R}_k$ prostřednictvím neznámého parametru $\boldsymbol{\beta} \in \mathbb{R}_k$. Matice $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ má hodnost k. Nechť platí:

1. Y_i má rozdělení s hustotou

$$\exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi)\right),$$

   kde *kumulantová* funkce $b(\theta)$ má spojitou druhou derivaci.

2. *Přirozený* parametr θ_i závisí na $\mathbf{x}_i$ a $\boldsymbol{\beta}$ prostřednictvím $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$.

3. Existuje známá ryze monotonní *spojovací funkce* (linková funkce) se spojitou druhou derivací taková, že je $\eta_i = g(\mu_i)$, kde je $\mu_i = \mathsf{E}\, Y_i$.

První požadavek znamená, že Y_i má exponenciální rozdělení s dispersním parametrem, avšak v poněkud zjednodušené formě. Věrohodnostní funkci můžeme zapsat jako

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{Y_i\theta_i - b(\theta_i)}{a(\varphi)} + c(Y_i, \varphi)\right). \tag{13.4}$$

Poznámka V dalším budeme střídavě podle potřeby používat několik sad parametrů. Mezi vektory parametrů $\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\mu}$ jsou vzájemně jednoznačné vztahy dané funkcemi

$$\begin{aligned} &\mu_i = \mathsf{E}\, Y_i = b'(\theta_i) \qquad (b \text{ má kladnou druhou derivaci!}) \\ &\eta_i = g(\mu_i). \end{aligned}$$

Zmíněné n-rozměrné parametry jsou ovšem funkcí k-rozměrného parametru $\boldsymbol{\beta}$.

Kanonický link

Pokud platí $g(\mu_i) = \theta_i$, funkce $g()$ se nazývá *kanonická* spojovací funkce (kanonický link). Pak je přímo $g(\mu_i) = \theta_i$, parametr η_i jakoby nepotřebujeme. V případě kanonického linku je spojovací funkce funkcí inverzní k $b'(\theta)$.

13.2.1. Extrémní případy

Saturovaný model

Uvažujme nejbohatší model, v němž má každé pozorování vlastní parametr. Vektor $\boldsymbol{\beta}$ lze ztotožnit s vektorem $\boldsymbol{\mu}$. Maximálně věrohodný odhad pak dostaneme z požadavku

$$\frac{\partial \ell}{\partial \theta_i} = \frac{Y_i - b'(\theta_i)}{a(\varphi)} = 0,$$

což vede k odhadům

$$\hat{\mu}_i = b'(\hat{\theta}_i) = Y_i, \qquad \hat{\theta}_i = (b')^{-1}(Y_i).$$

Vyhlazená hodnota $\hat{\mu}_i$ je tedy rovna přímo pozorování Y_i. Počet parametrů tu roste s počtem pozorování, takže nelze použít tvrzení o asymptotickém chování maximálně věrohodných odhadů.

Nulový model

Opačným extrémem je případ, kdy jsou všechny střední hodnoty (tedy i jiné n-členné parametry) stejné, tj. kdy platí

$$\theta_1 = \ldots = \theta_n (= \theta).$$

K maximálně věrohodnému odhadu dojdeme řešením rovnice

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} \frac{Y_i - b'(\theta)}{a(\varphi)} = 0.$$

V nulovém modelu je tedy

$$\hat{\mu} = b'(\hat{\theta}) = \bar{Y}, \qquad a(\varphi), \qquad \hat{\theta} = (b')^{-1}(\bar{Y}).$$

13.2.2. Odhad $\boldsymbol{\beta}$

Běžný model bude někde mezi právě popsanými extrémy. Při výpočtu derivací logaritmické věrohodnostní funkce vyjdeme z vyjádření

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}}.$$

Když použijeme souvislost derivace inverzní funkce s derivací funkce původní, stačí připravit následující výrazy:

$$
\begin{aligned}
\frac{\partial \ell}{\partial \theta_i} &= \frac{1}{a(\varphi)}(Y_i - \mu_i),, \\
\frac{\partial \mu_i}{\partial \theta_i} &= b''(\theta_i) = V(\theta_i) \Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{V(\mu_i)} \\
\frac{\partial \eta_i}{\partial \mu_i} &= g(\mu_i) \Rightarrow \frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}, \\
\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} &= \mathbf{x}_i.
\end{aligned}
$$

Hledaný vektor skórů je tedy roven

$$
\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{a(\varphi)} \sum_{i=1}^{n} (Y_i - \mu_i) \frac{1}{V(\mu_i) g'(\mu_i)} \mathbf{x}_i.
$$

Nezapomínejme při tom, že parametry μ_i jsou závislé na $\boldsymbol{\beta}$, kdežto φ je na $\boldsymbol{\beta}$ nezávislé. Odhad metodou maximální věrohodnosti tedy budeme hledat mezi řešeními soustavy nelineárních rovnic

$$
\sum_{i=1}^{n} \frac{1}{V(\mu_i) g'(\mu_i)} (Y_i - \mu_i) \mathbf{x}_i = \mathbf{0}, \tag{13.5}
$$

která už nijak na φ nezávisí. V případě kanonického linku je spojovací funkce inverzní funkcí k funkci b'), takže součin jejich derivací (jmenovatel v (13.5) je roven jedné. To znamená, že se normální rovnice v tom případě zjednoduší na

$$
\sum_{i=1}^{n} (Y_i - \mu_i) \mathbf{x}_i = \mathbf{0}, \tag{13.6}
$$

Běžné iterační metody řešení soustavy nelineárních rovnic používají na matici derivací těchto funkcí, kterou nazýváme *informační matice*. V našem případě určíme pouze střední hodnotu informační matice, tedy *Fisherovu informační matici*. Funkce $\mathbf{U}(\boldsymbol{\beta})$ je součtem součinů tří funkcí $\boldsymbol{\beta}$, z nichž jedna má nulovou střední hodnotu, proto zůstanou ve výsledné střední hodnotě pouze derivace členů $Y_i - \mu_i$.

$$
\begin{aligned}
-\mathsf{E}\, \frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} &= \frac{1}{a(\varphi)} \sum_{i=1}^{n} \frac{1}{V(\mu_i) g'(\mu_i)} \mathbf{x}_i \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}'} \\
&= \frac{1}{a(\varphi)} \sum_{i=1}^{n} \frac{1}{V(\mu_i) (g'(\mu_i))^2} \mathbf{x}_i \mathbf{x}_i' \\
&= \frac{1}{a(\varphi)} \sum_{i=1}^{n} w(\mu_i) \mathbf{x}_i \mathbf{x}_i',
\end{aligned}
$$

kde jsme označili
$$w(\mu_i) = \frac{1}{V(\mu_i)(g'(\mu_i))^2}.$$
V případě kanonického linku by vyšlo
$$w(\mu_i) = \frac{1}{g'(\mu_i)} = V(\mu_i).$$
Zavedeme-li diagonální matici jakýchsi obecných kladných vah
$$\mathbf{W}(\boldsymbol{\mu}) = \operatorname{diag}\{w(\mu_1), \ldots, w(\mu_n)\},$$
můžeme hledanou matici druhých derivací psát jako pozitivně definitní matice, když předpokládáme, že samotná matice $\mathbf{X}$ má lineárně nezávislé sloupce.
$$-\mathsf{E}\,\frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = \mathbf{X}'\mathbf{W}\mathbf{X}.$$

Lze ukázat, že v případě kanonického linku je informační matice rovna přímo svojí střední hodnotě.

Vraťme se k rovnici (13.5), která připomíná vážený lineární model s diagonální maticí $\mathbf{W}$ popsaný v kapitole 2.8. V zobecněném lineárním modelu má být lineární funkcí regresorů parametr $\eta_i = g(\mu_i)$. Všimněme si, že lze psát přibližnou rovnost
$$\begin{aligned} \mathsf{var}\, g(Y_i) &\doteq \mathsf{var}\,\left(g(\mu_i) + g'(\mu_i)(Y_i - \mu_i)\right) \\ &= \left(g'(\mu_i)\right)^2 a(\varphi) \cdot V(\mu_i) = \frac{a(\varphi)}{w(\mu_i)}, \end{aligned}$$
což dává interpretaci funkci $w(\mu)$. Doporučuji připomenout si na tomto místě vážený lineární model z oddílu 2.8, zvláště jeho speciální případ s rozptylem $\mathsf{var}\, Y_i = \sigma^2/w_i$.

Zaveďme nyní upravenou závisle proměnnou (adjusted dependent variable)
$$Z_i = g(\hat{\mu}_i) + g'(\hat{\mu}_i)(Y_i - \hat{\mu}_i) = \hat{\eta}_i + g'(\hat{\mu}_i)(Y_i - \hat{\mu}_i). \tag{13.7}$$
Jde vlastně o lineární část Taylorova rozvoje výrazu $g(Y_i)$ v bodě $\hat{\mu}_i$. Když tuto upravenou proměnnou použijeme ve vzorci (2.25) pro výpočet odhadu $\mathbf{b}_W$ ve váženém lineárním modelu, dostaneme návod pro výpočet i v zobecněném lineárním modelu. Ukažme, že hledaný maximálně věrohodný odhad, tedy řešení $\hat{\boldsymbol{\beta}}$ soustavy (13.5) je řešením soustavy rovnic $\mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}\mathbf{Z}$. Když použijeme skutečnost, že $\hat{\mu}_i$ je řešením soustavy (13.5), dostaneme totiž
$$\begin{aligned} \mathbf{X}'\mathbf{W}\mathbf{Z} &= \sum_{i=1}^{n} \mathbf{x}_i w(\hat{\mu}_i)\left(\hat{\eta}_i + g'(\hat{\mu}_i)(Y_i - \hat{\mu}_i)\right) \\ &= \sum_{i=1}^{n} \mathbf{x}_i w(\hat{\mu}_i)\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}} + g'(\hat{\mu}_i)(Y_i - \hat{\mu}_i)\right) \\ &= \sum_{i=1}^{n} \mathbf{x}_i w(\hat{\mu}_i)\mathbf{x}_i'\hat{\boldsymbol{\beta}} + \sum_{i=1}^{n} \mathbf{x}_i w(\hat{\mu}_i) g'(\hat{\mu}_i)(Y_i - \hat{\mu}_i) \\ &= \mathbf{X}'\mathbf{W}\mathbf{X}. \end{aligned}$$

Prakticky se nelineární soustava rovnic $\mathbf{XW}(\hat{\mu})\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{XW}(\hat{\boldsymbol{\mu}})\mathbf{Z}(\hat{\boldsymbol{\mu}})$ řeší iteračně. Jako výchozí aproximace se zpravidla volí $\hat{\mu} = \mathbf{Y}$. Ke každé aproximaci pro $\hat{\mu}$ se spočítá řešením soustavy lineárních rovnic nová aproximace pro $\hat{\boldsymbol{\beta}}$ (jako ve váženém lineárním modelu), k ní se spočítá další aproximace pro $\hat{\mu}$ atd. Iterace se opakují, dokud se v nastavené toleranci odhady ještě mění. Uvedený postup se nazývá *iterační vážená metoda nejmenších čtverců* (IWLS, IRLS).

13.2.3. Míry dobré shody

Vraťme se k saturovanému modelu, který maximalizuje logaritmickou věrohodnostní funkci pro daný vektor závisle proměnné $\mathbf{Y}$. Označme všechny odhady v saturovaném modelu hvězdičkou $\hat{\mu}_i^* = Y_i, \hat{\theta}_i^* = (b')^{-1}(Y_i)$. Maximální dosažitelná hodnota logaritmické věrohodnostní funkce je tedy vzhledem k (13.4)

$$\ell^* = \frac{1}{a(\varphi)} \sum_{i=1}^{n} (Y_i \hat{\theta}_i^* - b(\hat{\mu}_i)) + \sum_{i=1}^{n} c(Y_i, \varphi).$$

K porovnání běžného modelu s modelem saturovaným slouží *škálovaná deviance* (scaled deviance)

$$\begin{aligned} D_{\text{scale}}(\hat{\boldsymbol{\beta}}) &= 2\left(\ell^* - \ell(\hat{\boldsymbol{\beta}})\right) \\ &= \frac{2}{a(\varphi)}\left(\sum_{i=1}^{n}(Y_i\hat{\theta}_i^* - b(\hat{\mu}_i^*)) - \sum_{i=1}^{n}(Y_i\hat{\theta}_i - b(\hat{\mu}_i))\right) \\ &= \frac{2}{a(\varphi)}\sum_{i=1}^{n}\left(Y_i(\hat{\theta}_i^* - \hat{\theta}_i) - (b(\hat{\theta}_i^*) - b(\hat{\theta}_i))\right), \end{aligned}$$

když se členy s funkcí $c(.,.)$ vyrušily, neboť jsou v saturovaném modelu a v podmodelu totožné. Samotný dvojnásobek součtu na pravé straně (bez vydělení dispersním parametrem) nazveme *deviance*

$$D(\hat{\boldsymbol{\beta}}) = 2\sum_{i=1}^{n}\left(Y_i(\hat{\theta}_i^* - \hat{\theta}_i) - (b(\hat{\theta}_i^*) - b(\hat{\theta}_i))\right).$$

Viděli jsme, že v případě binomického či Poissonova rozdělení máme $a(\varphi) = 1$, takže obě definice deviance splývají.

Specielně pro normální rozdělení, kde je střední hodnota μ totožná s kanonickým parametrem θ, dostaneme v saturovaném modelu $\hat{\mu}_i^* = \hat{\theta}_i^* = Y_i$, takže dostaneme

$$\begin{aligned} D(\hat{\boldsymbol{\beta}}) &= 2\sum_{i=1}^{n}\left(Y_i(Y_i - \hat{\mu}_i) - (Y_i^2/2 - \hat{\mu}_i^2/2)\right) \\ &= \sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2, \end{aligned}$$

což není nic jiného, než reziduální součet čtverců známý z lineárního modelu. Můžeme tedy devianci chápat jako zobecnění reziduálního součtu čtverců.

Pomocí škálovaných deviancí lze zapsat také testovou statistiku testu poměrem věrohodností. Je-li $\tilde{\boldsymbol{\beta}}$ odhad parametru $\boldsymbol{\beta}$ za hypotézy (v podmodelu), pak lze statistiku $LR = \ell(\hat{\boldsymbol{\beta}}) - \ell(\tilde{\boldsymbol{\beta}})$ z (A.31) psát

$$LR = D^*(\tilde{\boldsymbol{\beta}}) - D^*(\hat{\boldsymbol{\beta}}).$$

Připomeňme, že za platnosti testované hypotézy má tento rozdíl asymptoticky rozdělení $\chi^2(q)$, kde q je počet omezení kladených hypotézou na parametry.

Další mírou kvality vyhlazení je *Pearsonovo* X^2 definované jako

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

Snadno zjistíme, že v případě normálního rozdělení dostaneme opět reziduální součet čtverců, kdežto například v případě binomického rozdělení s odhadnutými pravděpodobnostmi $\hat{\pi}_i = \hat{\mu}_i / m_i$ dostaneme

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1 - \hat{\pi}_i)},$$

což je běžná statistika χ^2-testu dobré shody.

Deviance porovnává skutečnou hodnotu logaritmické věrohodnostní funkce s její maximální možnou hodnotou, kdežto Pearsonovo χ^2 porovnává skutečně napozorované hodnoty s odhadem jejich střední hodnoty v daném modelu.

13.2.4. Rezidua

V normálním lineárním modelu je reziduální součet čtverců dán součtem čtverců reziduí definovaných jako rozdíly $u_i = Y_i - \hat{\mu}_i$. V zobecněných lineárních modelech se prakticky nepoužívají, program R je počítá pod názvem response residuals. Jak jsme viděli, reziduální součet čtverců lze zapsat dvěma dalšími způsoby, jimž odpovídají dvě různé definice reziduí. *Pearsonova rezidua* zavedeme jako

$$_{\text{Pearson}}u_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}, \tag{13.8}$$

deviančni rezidua jako

$$_{\text{dev}}u_i = \sqrt{2\left(Y_i(\hat{\theta}_i^* - \hat{\theta}_i) - (b(\hat{\theta}_i^*) - b(\hat{\theta}_i))\right)}\,\text{sign}(Y_i - \hat{\mu}_i). \tag{13.9}$$

Program R dá k disposici také rezidua

$$_{\text{work}}u_i = g'(\hat{\mu}_i)(Y_i - \hat{\mu}_i), \tag{13.10}$$

která lze chápat jako rezidua v linearizovaném modelu (13.7). Odtud lze odvodit také parciální rezidua pro GLM.

13.2.5. Odhad disperzního parametru

Připomeňme, že v zobecněném lineárním modelu máme $\mathsf{var}\, Y_i = \varphi V(\mu_i)$. Připomeňme také, že ve speciálním případě normálního lineárního modelu je Pearsonovo χ^2 rovno reziduálnímu součtu čtverců a statistika

$$\hat{\varphi} = \frac{X^2}{n-k}$$

je nestranným odhadem rozptylu $\varphi = \sigma^2$. Lze dokázat (např. (McCullagh, Nelder, 1989, str. 298) nebo text přednášky dr. Kulicha STP126 *Zobecněné lineární modely*), že uvedená statistika je konzistentním odhadem v obecném lineárním modelu, nejen v modelu s normálním rozdělením.

13.2.6. Standardizovaná rezidua

V kapitole 8 jsme zavedli normovaná a studentizovaná rezidua s cílem zajistit stejný rozptyl upravených reziduí. Navíc jsme tam odstraňovali (dvěma různými způsoby) závislost rozptylu upravených reziduí na parametru σ^2. Také v GLM se zavádějí standardizovaná rezidua, přičemž na místě matice $\mathbf{H}$ z lineárního modelu slouží matice

$$\mathbf{H}_W = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$$

s diagonálními prvky h_{ii}. V lineárním modelu spočívala standardizace především v použití hodnot $m_{ii} = 1 - h_{ii}$. Standardizovaná rezidua v GLM spočítáme tak, že rezidua zavedená v (13.8) a (13.9) vydělíme buď $\sqrt{1-h_{ii}}$ nebo $\sqrt{\hat{\varphi}(1-h_{ii})}$.

13.2.7. Binomické rozdělení

Závisle proměnná Y_i nabývá hodnot $0, \ldots, m_i$, kde m_i je pro každé i známá konstanta. Originálním parametrem je π, přirozeným je $\theta = \log \frac{\pi}{1-\pi}$, rozptylový parametr φ je identicky roven jedné. Kumulantová funkce je dána vztahem $b(\theta) = m \log(1 + \mathrm{e}^\theta) = -m \log(1-\pi)$. Saturovaný model vede k odhadům

$$\hat{\mu}_i^* = Y_i, \quad \hat{\pi}_i^* = \bar{Y}_i, \quad \hat{\theta}_i^* = \log \frac{\bar{Y}_i}{1-\bar{Y}_i}, \quad b(\hat{\theta}_i^*) = -\log(1-\bar{Y}_i),$$

takže deviance vyjde

$$\begin{aligned} D(\hat{\beta}) &= 2\sum_{i=1}^n \left(Y_i\left(\log\frac{\bar{Y}_i}{1-\bar{Y}_i} - \log\frac{\hat{\mu}_i}{1-\hat{\mu}_i}\right) + m_i \log\frac{1-\bar{Y}_i}{1-\hat{\pi}_i}\right) \\ &= 2\sum_{i=1}^n m_i\left(\bar{Y}_i \log\frac{\bar{Y}_i}{\hat{\pi}_i} + (1-\bar{Y}_i)\log\frac{1-\bar{Y}_i}{1-\hat{\pi}_i}\right). \end{aligned}$$

Devianční reziduum tedy má tvar

$$_{\mathrm{dev}}u_i = \sqrt{2m_i\left(\bar{Y}_i \log\frac{\bar{Y}_i}{\hat{\pi}_i} + (1-\bar{Y}_i)\log\frac{1-\bar{Y}_i}{1-\hat{\pi}_i}\right)}\,\mathrm{sign}(\bar{Y}_i - \hat{\pi}_i).$$

Pearsonovo X^2 lze vyjádřit ve tvaru

$$X^2 = \sum_{i=1}^{n} m_i \frac{(\bar{Y}_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)}.$$

Explicitní vyjádření Pearsonova rezidua je tedy

$$_{\mathrm{Pearson}}u_i = \sqrt{m_i}\frac{\bar{Y}_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} = \frac{Y_i - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

Protože je

$$\mu = b'(\theta) = m\frac{\exp(\theta)}{1+\exp(\theta)} = m\pi,$$

kanonická spojovací funkce, která musí být inverzní k $b'(.)$, je tedy rovna

$$\theta = g(\mu) = \log\frac{\mu}{m-\mu} = \log\frac{\pi}{1-\pi}.$$

Úloha logistické regrese je speciálním případem, když se použije logistická funkce jako funkce spojovací a platí $m_i = 1$ pro všechna i. Jde o alternativní (Bernoulliovo) rozdělení.

13.2.8. Poissonovo rozdělení

V případě Poissonova rozdělení máme

$$\varphi = 1 \quad \theta = \log\lambda, \quad \lambda = \mathrm{e}^{\theta}, \quad \mu = \lambda = \mathrm{e}^{\theta}, \quad b(\theta) = \mathrm{e}^{\theta}.$$

Pro saturovaný model platí

$$\hat{\mu}_i^* = Y_i, \quad \hat{\theta}_i^* = \log Y_i,$$

takže deviance je rovna

$$\begin{aligned} D(\hat{\boldsymbol{\beta}}) &= 2\sum_{i=1}^{n}\left(Y_i(\log Y_i - \log\hat{\mu}_i) - (Y_i - \hat{\mu}_i)\right) \\ &= 2\sum_{i=1}^{n}\left(Y_i\log\frac{Y_i}{\hat{\mu}_i} - (Y_i - \hat{\mu}_i)\right). \end{aligned}$$

Pearsonovo X^2 je rovno

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

Kanonická spojovací funkce je $g(\mu) = \log \mu$.

Příklad 13.5 (zachovalost kostí) Vedlejším produktem antropologického zkoumání archeologických nálezů bylo vyšetřování závislosti počtu patologických odchylek (PPO) na stupni zachovalosti kostí (IKZ) (podle zprávy o nalezišti v Afalou-bou-Rhummel v západním Alžírsku, Černý (1994)). U počtu patologických odchylek lze předpokládat Poissonovo rozdělení. Data jsou uvedena v tabulce 13.1

Tabulka 13.1: Počty patologických odchylek PPO a stupeň zachovalosti kostí IKZ podle obrázku 2 z článku Černý (1994)

PPO	2	2	2	5	5	6	4	9
IKZ	27	32	41	42	51	58	61	74

```
> summary(a<-glm(ppo~ikz,poisson,data=Afalou))

Call:
glm(formula = ppo ~ ikz, family = poisson, data = Afalou)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.80173  -0.37957   0.03465   0.30076   0.88929

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.09534    0.65552  -0.145  0.88436
ikz          0.03045    0.01153   2.641  0.00826 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 9.3346  on 7  degrees of freedom
Residual deviance: 2.2135  on 6  degrees of freedom
AIC: 31.993

Number of Fisher Scoring iterations: 3
> anova(a,test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: ppo

Terms added sequentially (first to last)
```

```
     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    7     9.3346
ikz   1   7.1211        6     2.2135    0.0076
```

Z výstupu je zřejmé, že závislost je bezpečně prokázaná. Navíc je znovu vidět rozdíl mezi dvěma možnými testy, kdy Waldův test dal dosaženou hladinu $p =$ 0,0083, kdežto test poměrem věrohodnosti 0,0076. ◯

14. Model nelineární regrese

Až doposud jsme se s výjimkou logistické regrese resp. zobecněného lineárního modelu zabývali lineárním modelem, tedy takovým případem, kdy je množina všech možných středních hodnot vektoru $\mathbf{Y}$ lineární. Předpokládali jsme dokonce, že je $\mathsf{E}\,\mathbf{Y} \in \mathcal{M}(\mathbf{X})$, i když v zásadě jsme mohli předpokládat, že platí $\mathsf{E}\,\mathbf{Y} - \boldsymbol{\mu} \in \mathcal{M}(\mathbf{X})$ pro nějaké pevné známé $\boldsymbol{\mu}$.

14.1. Předpoklady

V dalším budeme předpokládat, že platí:

a) $\mathbf{Y} = \mathbf{f}(\boldsymbol{\theta}) + \mathbf{e}$, kde $\mathbf{e} \sim \mathsf{N}(\mathbf{0}, \sigma^2\mathbf{I})$ a $\mathbf{f}(\boldsymbol{\theta}) = (f(x_1, \boldsymbol{\theta}), \dots, f(x_n, \boldsymbol{\theta}))'$, přičemž $f(x, \boldsymbol{\theta})$ je známá *regresní* funkce,

b) $\boldsymbol{\theta} \in \Omega$, kde parametrický prostor $\Omega \in \mathbb{R}^k$ je otevřená konvexní množina,

c) funkce $f_j(x, \boldsymbol{\theta}) = \frac{\partial}{\partial\theta_j} f(x, \boldsymbol{\theta})$ a $f_{jt}(x, \boldsymbol{\theta}) = \frac{\partial^2}{\partial\theta_j \partial\theta_t} f(x, \boldsymbol{\theta})$ jsou pro všechna $x \in \mathcal{X}$ spojitou funkcí $\boldsymbol{\theta}$,

d) matice prvních derivací regresní funkce typu $n \times k$ daná vztahem $F(\boldsymbol{\theta}) = (f_j(x_i, \boldsymbol{\theta}))$ má přinejmenším v okolí správné hodnoty parametru $\boldsymbol{\theta}$ hodnost k,

Zaveďme funkci

$$\mathcal{S}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (Y_i - f(x_i, \boldsymbol{\theta}))^2 .$$

Odhad metodou nejmenších čtverců $\mathbf{t}$ je takový prvek Ω, který minimalizuje $S(\boldsymbol{\theta})$. Jako odhad rozptylu použijeme (podobně jako u lineárního modelu)

$$S^2 = \frac{\mathcal{S}(\mathbf{t})}{n - k}.$$

Protože jsme předpokládali normální rozdělení, je $\mathbf{t}$ odhadem metodou nejmenších čtverců a S^2 je asymptoticky ekvivalentní s odhadem rozptylu metodou maximální věrohodnosti daným $\mathcal{S}(\mathbf{t})/n$.

V bodě $\mathbf{t}$, který minimalizuje na otevřené množině Ω funkci $\mathcal{S}(\boldsymbol{\theta})$, by měl být vektor parciálních derivací nulový, což vede k *normální rovnici*

$$\mathbf{F}(\boldsymbol{\theta})'(\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta})) = \mathbf{0}. \tag{14.1}$$

Je dobré porovnat tuto rovnici s normální rovnicí (12.5) pro logistickou regresi, kde je rozdíl $\mathbf{Y} - \mathsf{E}\,\mathbf{Y}$ je násoben maticí konstant $\mathbf{X}'$, kdežto zde matice $\mathbf{F}(\theta)'$ je funkcí odhadovaného parametru. Stejně jednoduchou rovnici jako v případě logistické regrese dostaneme v každém zobecněném lineárním modelu s kanonickou spojovací funkcí (viz (13.6)).

14.2. Lineární aproximace

Pro $\boldsymbol{\theta}$, které je dostatečně blízko správné hodnoty $\boldsymbol{\theta}^*$, můžeme použít aproximace

$$\mathbf{f}(\boldsymbol{\theta}) \doteq \mathbf{f}^* + \mathbf{F}^*(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \tag{14.2}$$

$$\mathbf{F}(\boldsymbol{\theta}) \doteq \mathbf{F}^*, \tag{14.3}$$

když jsme zavedli stručný zápis

$$\mathbf{f}^* = \mathbf{f}(\boldsymbol{\theta}^*), \qquad \mathbf{F}^* = \mathbf{F}(\boldsymbol{\theta}^*).$$

Dosaďme uvedené aproximace do normální rovnice

$$\begin{aligned}\mathbf{0} &\doteq \mathbf{F}^{*'}(\mathbf{Y} - \mathbf{f}^* - \mathbf{F}^*(\boldsymbol{\theta} - \boldsymbol{\theta}^*))\\ &\doteq \mathbf{F}^{*'}(\mathbf{e} - \mathbf{F}^*(\boldsymbol{\theta} - \boldsymbol{\theta}^*)).\end{aligned}$$

Odtud je

$$\mathbf{t} \doteq \boldsymbol{\theta}^* + (\mathbf{F}^{*'}\mathbf{F}^*)^{-1}\mathbf{F}^{*'}(\mathbf{Y} - \mathbf{f}^*),$$

odkud máme aproximaci pro rozdělení odhadu $\mathbf{t}$

$$\mathbf{t} \sim \mathsf{N}\Big(\boldsymbol{\theta}^*, \sigma^2(\mathbf{F}^{*'}\mathbf{F}^*)^{-1}\Big). \tag{14.4}$$

Pro reziduální součet čtverců $\mathcal{S}(\mathbf{t})$ dostaneme podobně

$$\begin{aligned}\mathcal{S}(\mathbf{t}) &= ||\mathbf{Y} - \mathbf{f}^* - \mathbf{F}^*(\mathbf{t} - \boldsymbol{\theta}^*)||^2\\ &= ||(\mathbf{I} - \mathbf{F}^*(\mathbf{F}^{*'}\mathbf{F}^*)^{-1}\mathbf{F}^{*'})\mathbf{e}||^2 \mathrel{\dot\sim} \sigma^2\chi^2(n-k).\end{aligned}$$

Protože jsou $\mathbf{t}$ a $\mathcal{S}(\mathbf{t})$ asymptoticky nezávislé a protože je $\mathbf{t}$ konzistentním odhadem $\boldsymbol{\theta}^*$, aproximuje se pro každé $j = 1, \ldots, k$ rozdělení výrazu

$$\frac{t_j - \theta_j^*}{S\sqrt{v_{jj}}}, \tag{14.5}$$

rozdělením $\mathsf{t}(n-k)$. Při tom jsme použili označení $\mathbf{V} = (\mathbf{F}(\mathbf{t})'\mathbf{F}(\mathbf{t}))^{-1}$.

14.3. Testování jednoduché hypotézy o θ

Věnujme se nyní testování hypotézy $\boldsymbol{\theta} = \boldsymbol{\theta}^0$, která úplně určuje vektor regresních koeficientů. V souvislosti s tím nalezneme konfidenční množiny pro tento vektor. Použití aproximací způsobí, že testy i konfidenční množiny budou pouze přibližné.

Pokud je regresní funkce $f(x, \boldsymbol{\theta})$ lineární v $\boldsymbol{\theta}$, jsou dále uvedené konfidenční množiny $\mathcal{K}_W, \mathcal{K}_{LR}$ totožné s konfidenční množinou (2.21).

Waldův test

Waldův test je založen na hodnocení toho, nakolik odhad $\mathbf{t}$ metodou maximální věrohodnosti vyhovuje omezení $\boldsymbol{\theta} = \boldsymbol{\theta}^0$, které klade testovaná hypotéza.

Z předchozího výkladu (zejména z (14.4)) plyne, že za platnosti nulové hypotézy má statistika

$$\frac{(\mathbf{t} - \boldsymbol{\theta}^0)' \mathsf{F}(\boldsymbol{\theta}^0)' \mathsf{F}(\boldsymbol{\theta}^0)(\mathbf{t} - \boldsymbol{\theta}^0)}{ks^2},$$

přibližně rozdělení $\mathsf{F}(k, n-k)$. Proto je přibližný kritický obor dán nerovností

$$(\mathbf{t} - \boldsymbol{\theta}^0)' \mathsf{F}(\boldsymbol{\theta}^0)' \mathsf{F}(\boldsymbol{\theta}^0)(\mathbf{t} - \boldsymbol{\theta}^0) \geq k\ S^2 F_{k,n-k}(\alpha).$$

Chceme-li hledat interval spolehlivosti pro $\boldsymbol{\theta}$, nejjednodušší řešení dostaneme, když v matici $\mathsf{F}(\boldsymbol{\theta})' \mathsf{F}(\boldsymbol{\theta})$ použijeme konzistentní odhad $\mathbf{t}$ parametru $\boldsymbol{\theta}$. Odpovídající přibližná konfidenční množina má tedy tvar

$$\mathcal{K}_W = \left\{ \boldsymbol{\theta} \in \Omega : (\boldsymbol{\theta} - \mathbf{t})' \mathsf{F}(\mathbf{t})' \mathsf{F}(\mathbf{t})(\boldsymbol{\theta} - \mathbf{t}) < k\ S^2 F_{k,n-k}(\alpha) \right\}. \tag{14.6}$$

Pro každé $\mathbf{t}$ jde o elipsoid se středem v bodě $\mathbf{t}$.

Waldův test lze takto použít, jen když je nelinearita úlohy dostatečně zanedbatelná.

Test poměrem věrohodnosti

Test poměrem věrohodnosti porovnává hodnotu věrohodnostní funkce pro $\mathbf{t}$ a $\boldsymbol{\theta}^0$. Logaritmická věrohodnostní funkce je při předpokládaném normálním rozdělení rovna

$$\ell(\boldsymbol{\theta}, \sigma^2) = c - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \mathcal{S}(\boldsymbol{\theta}).$$

K testování hypotézy použijeme vlastnost testu poměrem věrohodnosti, podle které (při známém rozptylu σ^2) má rozdíl $2(\ell(\mathbf{t}, \sigma^2) - \ell(\boldsymbol{\theta}^0, \sigma^2))$ asymptoticky rozdělení $\chi^2(k)$. Nyní použijeme místo neznámého σ^2 jeho odhad S^2, takže za platnosti testované hypotézy přibližně platí

$$\frac{\mathcal{S}(\boldsymbol{\theta}^0) - \mathcal{S}(\mathbf{t})}{kS^2} \sim \mathsf{F}(k, n-k).$$

Proto je přibližný kritický obor dán nerovností

$$S(\boldsymbol{\theta}^0) \geq \mathcal{S}(\mathbf{t}) + kS^2 F_{k,n-k}(\alpha).$$

Když navíc vyjádříme odhad S^2 pomocí $\mathcal{S}(\mathbf{t})$, dostaneme přibližnou konfidenční množinu ve tvaru

$$\mathcal{K}_{LR} = \left\{ \boldsymbol{\theta} \in \Omega : \mathcal{S}(\boldsymbol{\theta}) < \mathcal{S}(\mathbf{t}) \left(1 + \frac{k}{n-k} F_{k,n-k}(\alpha) \right) \right\}. \tag{14.7}$$

Tato konfidenční množina má obecně složitější tvar. Obsahuje takové hodnoty $\boldsymbol{\theta}$, pro něž funkční hodnota $\mathcal{S}(\boldsymbol{\theta})$ nepřekračuje příliš minimální možnou hodnotu $\mathcal{S}(\mathbf{t})$. Dovolené překročení je určeno výrazem v kulaté závorce v (14.7).

Přesný test

V tomto oddílu naznačíme, jak by bylo možno sestavit kritický obor přesného testu. Jak ale uvidíme, metoda má jen velmi omezené použití.

Nechť platí nulová hypotéza $\boldsymbol{\theta} = \boldsymbol{\theta}^0$, nechť $\mathbf{H}$ je nějaká pevná idempotentní matice typu $n \times n$ hodnosti k. Potom má výraz

$$F_H = \frac{(\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}^0))'\mathbf{H}(\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}^0))}{(\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}^0))'(\mathbf{I} - \mathbf{H})(\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}^0))} \frac{n-k}{k}$$

rozdělení $\mathsf{F}(k, n-k)$. Snadno tedy sestrojíme kritický obor testu, který má *přesně* zvolenou hladinu α. Je však třeba, aby matice $\mathbf{H}$ byla zvolena tak, aby test měl také co největší sílu.

Jednou z možností je *nezávisle na* $\mathbf{Y}$ zvolit vektory $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^k$ tak, aby matice

$$\mathbf{X} = \Big(\mathbf{f}(\boldsymbol{\theta}^1) - \mathbf{f}(\boldsymbol{\theta}^0), \ldots, \mathbf{f}(\boldsymbol{\theta}^k) - \mathbf{f}(\boldsymbol{\theta}^0) \Big)$$

měla hodnost k. Potom má matice

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

požadované vlastnosti. Lze ukázat, že test založený na F_H je citlivý vůči alternativám $\boldsymbol{\theta}^* = \boldsymbol{\theta}^j$, $j = 1, \ldots, k$.

Jinou možností je použít projekční matici

$$\mathbf{H}^0 = \mathbf{F}(\boldsymbol{\theta}^0)(\mathbf{F}(\boldsymbol{\theta}^0)'\mathbf{F}(\boldsymbol{\theta}^0))^{-1}\mathbf{F}(\boldsymbol{\theta}^0).$$

14.4. Testování složené hypotézy

Rozdělme nyní parametr $\boldsymbol{\theta}$ na dvě složky jako $\boldsymbol{\theta} = (\boldsymbol{\gamma}', \boldsymbol{\delta}')'$. Testujeme nulovou hypotézu $\boldsymbol{\delta} = \boldsymbol{\delta}^0$, kde $\boldsymbol{\delta}^0 \in \mathbb{R}^q$ je pevný vektor.

První řešení založíme na Waldově postupu. Podobně jako $\boldsymbol{\theta}$ rozdělme odhad metodou nejmenších čtverců $\mathbf{t} = (\mathbf{c}', \mathbf{d}')'$ a také odhad varianční matice

$$\sigma^2 \mathbf{V} = \sigma^2 \left((\mathbf{F}(\mathbf{t}))'\mathbf{F}(\mathbf{t})\right)^{-1} = \sigma^2 \begin{pmatrix} \mathbf{V}_{\gamma\gamma} & \mathbf{V}_{\gamma\delta} \\ \mathbf{V}_{\delta\gamma} & \mathbf{V}_{\delta\delta} \end{pmatrix}.$$

Speciálním případem přibližného rozdělení $\mathbf{t}$ z (14.4) je $\mathbf{d} \sim \mathsf{N}\big(\boldsymbol{\delta}, \sigma^2 \mathbf{V}_{\delta\delta}\big)$ a zejména přibližný interval spolehlivosti pro $\boldsymbol{\delta}$ (protějšek eliptické konfidenční množiny podle (14.6))

$$\left\{\boldsymbol{\delta} : (\mathbf{d} - \boldsymbol{\delta})'\mathbf{V}_{\delta\delta}^{-1}(\mathbf{d} - \boldsymbol{\delta}) < qS^2 F_{q,n-k}(\alpha)\right\}.$$

Speciálním případem pro $q = 1$ jsou přibližné intervaly spolehlivosti

$$\left(t_j - S\sqrt{v_{jj}}\, t_{n-k}(\alpha), t_j + S\sqrt{v_{jj}}\, t_{n-k}(\alpha)\right)$$

založené na přímém použití (14.5).

Další možné řešení, které vychází z testu poměrem věrohodnosti, je výpočetně náročnější. Nechť $\tilde{\mathbf{c}}(\boldsymbol{\delta})$ je odhad vektoru $\boldsymbol{\gamma}$ pro dané $\boldsymbol{\delta}$. Označme $\tilde{\mathbf{t}} = \tilde{\mathbf{t}}(\boldsymbol{\delta}) = (\tilde{\mathbf{c}}(\boldsymbol{\delta})', \boldsymbol{\delta}')'$. Platí-li nulová hypotéza $\boldsymbol{\delta} = \boldsymbol{\delta}^0$, pak má statistika

$$2\left(\ell(\mathbf{t}) - \ell(\tilde{\mathbf{t}}(\boldsymbol{\delta}^0))\right) = \frac{1}{\sigma^2}\left(\mathcal{S}(\tilde{\mathbf{t}}(\boldsymbol{\delta}^0)) - \mathcal{S}(\mathbf{t})\right)$$

asymptoticky rozdělení $\chi^2(q)$. Použijeme-li opět konzistentní odhad S^2 parametru σ^2, dostaneme přibližný kritický obor

$$\mathcal{S}\left(\tilde{\mathbf{t}}(\boldsymbol{\delta}^0)\right) \geq \mathcal{S}(\mathbf{t}) + qS^2 F_{q,n-k}(\alpha)$$

tj.

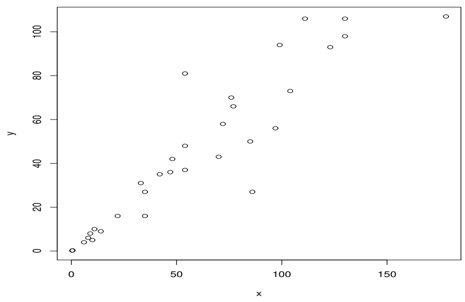
$$\mathcal{S}\left(\tilde{\mathbf{t}}(\boldsymbol{\delta}^0)\right) \geq \mathcal{S}(\mathbf{t}) \left(1 + \frac{q}{n-k} F_{q,n-k}(\alpha)\right).$$

Interval spolehlivosti by tedy byl

$$\left\{\boldsymbol{\delta} : \mathcal{S}\left(\tilde{\mathbf{t}}(\boldsymbol{\delta})\right) < \mathcal{S}(\mathbf{t}) \left(1 + \frac{q}{n-k} F_{q,n-k}(\alpha)\right)\right\}.$$

Speciálně pro $q = 1$ označme $\tilde{\mathbf{t}}_j(\theta)$ vektor parametrů, který minimalizuje $\mathcal{S}(\boldsymbol{\theta})$ za podmínky, že $\theta_j = \theta$. Potom má výraz

$$\tau(\theta) = \frac{\sqrt{\mathcal{S}(\tilde{\mathbf{t}}_j(\theta)) - \mathcal{S}(\mathbf{t})}}{S} \operatorname{sign}(\theta - t_j)$$



Obrázek 14.1: Farmakologická závislost

přibližně rozdělení $\mathtt{t}(n-k)$. V normálním lineárním modelu s úplnou hodností to platí přesně, jak plyne z (3.19).

Odtud lze opět nalézt přibližný interval spolehlivosti pro θ_j. Míra nelinearity je patrná z *profilového diagramu*, který znázorňuje body $[\theta, \tau(\theta)]$ (případně $[\theta, |\tau(\theta)|]$) v okolí bodového odhadu t_j parametru θ_j.

Příklad 14.1 Farmakolog vyšetřuje u dat znázorněných na obrázku 14.1 závislost tvaru

$$f(x;\beta,\gamma) = \frac{1}{\gamma}\left(x + (625-x)\left(1-\exp(\beta x/(625-x))\right)\right).$$

Výpočet pomocí knihovny `nls` programu R dal

```
> a.Kan<-nls(y~(x+(625-x)*(1-exp(-b*x/(625-x))))/c,
                                  start=list(b=5,c=10),data=In.Kan)
> summary(a.Kan)

Formula: y ~ (x + (625 - x) * (1 - exp(-b * x/(625 - x))))/c

Parameters:
  Estimate Std. Error t value Pr(>|t|)
b    2.417      1.317   1.836  0.07629 .
c    3.881      1.081   3.591  0.00116 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.34 on 30 degrees of freedom

Correlation of Parameter Estimates:
       b
c 0.9883
```

```
> plot(profile(a.Kan,1))
> plot(profile(a.Kan))
```

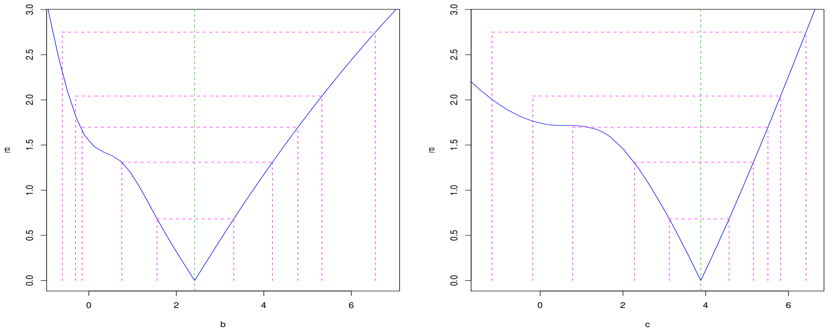
Z výstupu je vidět, že je-li platná použitá lineární aproximace, parametr β není průkazně nenulový. Za hypotézy $\beta = 0$ bychom dostali přímku. O případné silné nelinearitě se můžeme přesvědčit na profilových diagramech (obr. 14.2), které jsme připravili posledními dvěma příkazy. Z grafů je patrné, že v úloze se silně projevuje nelinearita. Například intervaly spolehlivosti pro γ budou velmi nesymetrické vzhledem k bodovému odhadu. (Na obrázku jsou znázorněny intervaly spolehlivosti se spolehlivostí po řadě 99 %, 95 %, 90 %, 80 % a 50 %).

O hypotéze, že $\beta = 0$ můžeme rozhodovat také pomocí přibližného F-testu, který porovná reziduální součty čtverců.

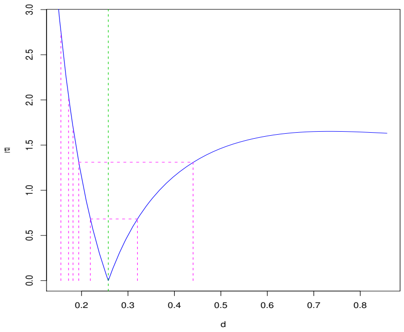
```
> ap.Kan<-nls(y~x/c,start=list(c=1),data=In.Kan)
> summary(ap.Kan)

Formula: y ~ x/c

Parameters:
  Estimate Std. Error t value Pr(>|t|)
c  1.34890    0.05897   22.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.71 on 31 degrees of freedom

> anova(ap.Kan,a.Kan)
Analysis of Variance Table

Model 1: y ~ x/c
Model 2: y ~ (x + (625 - x) * (1 - exp(-b * x/(625 - x))))/c
  Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)
1     31     5829.6
2     30     5341.0  1  488.6  2.7447  0.108
```

Jak je patrno, přímka je možným modelem pro naše data.

Původně byla úloha parametrizována jinak, místo γ byl v definici regresní funkce parametr $\delta = 1/\gamma$, takže regresní funkce byla v δ lineární. Přesto bylo chování odhadů δ mnohem méně lineární, jak naznačuje obrázek 14.3. ○



Obrázek 14.2: Profilové diagramy pro parametry β (vlevo) a γ (vpravo)



Obrázek 14.3: Profilový diagram pro parametry $\delta = 1/\gamma$

15. Parametrizace v NLR

V nelineární regresi se setkáme s novým jevem v porovnání s regresí lineární. Když použijeme aproximace jemnější, než lineární, zjistíme, že odhad vektoru parametrů $\boldsymbol{\theta}$ obecně není nestranný a že jeho vychýlení závisí na tom, jak jsme regresní funkci vyjádřili pomocí parametrů.

15.1. Označení

Rozšiřme označení zavedené v odstavci 14.2. Symbolem $\ddot{\mathbf{F}}(\boldsymbol{\theta})$ označíme trojrozměrnou matici typu $n \times k \times k$ danou vztahem

$$\ddot{\mathbf{F}}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mathbf{f}(\boldsymbol{\theta}) \tag{15.1}$$

$$= \left(\ddot{\mathbf{F}}_{i\bullet\bullet}\right)_{i=1,\dots,n} = \left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} f(x_i,\boldsymbol{\theta})\right)_{i=1,\dots,n} \tag{15.2}$$

$$= \left(\ddot{\mathbf{f}}_{\bullet jr}\right)_{j,r=1,\dots,k} = \left(\frac{\partial^2}{\partial\theta_j\partial\theta_r}\mathbf{f}(\boldsymbol{\theta})\right)_{j,r=1,\dots,k}. \tag{15.3}$$

Matici $\ddot{\mathbf{F}}(\boldsymbol{\theta})$ si můžeme představit jako trojrozměrný objekt se čtvercovou základnou a výškou n, jehož i-tá vrstva je tvořena maticí $\ddot{\mathbf{F}}_{i\bullet\bullet}$ a jr-tý sloupec vektorem $\ddot{\mathbf{f}}_{\bullet jr}$. Podobně jako dříve označíme $\ddot{\mathbf{F}}(\boldsymbol{\theta}^*)$ symbolem $\ddot{\mathbf{F}}^*$.

Jako kvadratickou aproximaci vektoru $\mathbf{f}(\boldsymbol{\theta})$ použijeme

$$\mathbf{f}(\boldsymbol{\theta}) \doteq \mathbf{f}^* + \mathbf{F}^*(\boldsymbol{\theta}-\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)'\,\ddot{\mathbf{F}}^*(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\,. \tag{15.4}$$

Jde o rozšíření lineární aproximace (14.3) o kvadratický člen, v němž se při násobení trojrozměrné matice $\ddot{\mathbf{F}}(\boldsymbol{\theta})$ provádí součin přes druhý a třetí rozměr.

Ukažme si budoucí problémy na dvou velmi jednoduchých příkladech.

Příklad 15.1 Mějme regresní funkci $f(x,\theta) = \mathrm{e}^{x\theta}$. Zvolíme-li $\beta = \mathrm{e}^{\theta}$, můžeme stejnou funkci zapsat jako $\tilde{f}(x,\beta) = \beta^x$. Je tedy

$$\begin{aligned}
\frac{\partial}{\partial\theta} f(x,\theta) &= x\mathrm{e}^{x\theta} &&= x f(x,\theta),\\
\frac{\partial^2}{\partial\theta^2} f(x,\theta) &= x^2\mathrm{e}^{x\theta} &&= x^2 f(x,\theta),\\
\frac{\partial}{\partial\beta} \tilde{f}(x,\beta) &= x\beta^{x-1} &&= \frac{x}{\beta}\tilde{f}(x,\beta),\\
\frac{\partial^2}{\partial\beta^2} f(x,\beta) &= x(x-1)\beta^{x-2} &&= \frac{x(x-1)}{\beta^2}\tilde{f}(x,\beta).
\end{aligned}$$

Zvolíme-li $x_1 = 0$, $x_2 = 1$, bude

$$\mathbf{f}(\theta) = \begin{pmatrix}1\\ \mathrm{e}^{\theta}\end{pmatrix}, \qquad \mathbf{F}(\theta) = \begin{pmatrix}0\\ \mathrm{e}^{\theta}\end{pmatrix}, \qquad \ddot{\mathbf{F}}(\theta) = \begin{pmatrix}0\\ \mathrm{e}^{\theta}\end{pmatrix},$$

$$\tilde{\mathbf{f}}(\beta) = \begin{pmatrix}1\\ \beta\end{pmatrix}, \qquad \tilde{\mathbf{F}}(\beta) = \begin{pmatrix}0\\ 1\end{pmatrix}, \qquad \ddot{\tilde{\mathbf{F}}}(\beta) = \begin{pmatrix}0\\ 0\end{pmatrix}.$$

Zvolíme $\theta^\star = 0$, čemuž odpovídá $\beta^\star = 1$. Výsledné *lineární* aproximace jsou

$$\begin{aligned}
\mathbf{f}(\theta) &= \begin{pmatrix}1\\ \mathrm{e}^{\theta}\end{pmatrix} \doteq \begin{pmatrix}1\\ 1\end{pmatrix} + \begin{pmatrix}0\\ 1\end{pmatrix}\theta = \begin{pmatrix}1\\ 1+\theta\end{pmatrix},\\
\tilde{\mathbf{f}}(\beta) &= \begin{pmatrix}1\\ \beta\end{pmatrix} = \begin{pmatrix}1\\ 1\end{pmatrix} + \begin{pmatrix}0\\ 1\end{pmatrix}(\beta-1) = \begin{pmatrix}1\\ \beta\end{pmatrix}.
\end{aligned}$$

Zatímco v prvním případě jde skutečně o aproximaci, ve druhém případě máme místo aproximace identitu. Zvolíme-li $\theta \neq 0$ a odpovídající $\beta = \mathrm{e}^{\theta}$, budou vektory $\mathbf{f}(\theta)$ a $\tilde{\mathbf{f}}(\beta)$ neshodné. Dále stojí za povšimnutí, že množina aproximujících vektorů je v obou případech stejná. $\bigcirc$

Příklad 15.2 Zvolme nyní pro stejné funkce $f, \tilde{f}$ jako v příkladu 15.1, ale $x_1 = 1$ a $x_2 = 3$, dostaneme poněkud jiné matice

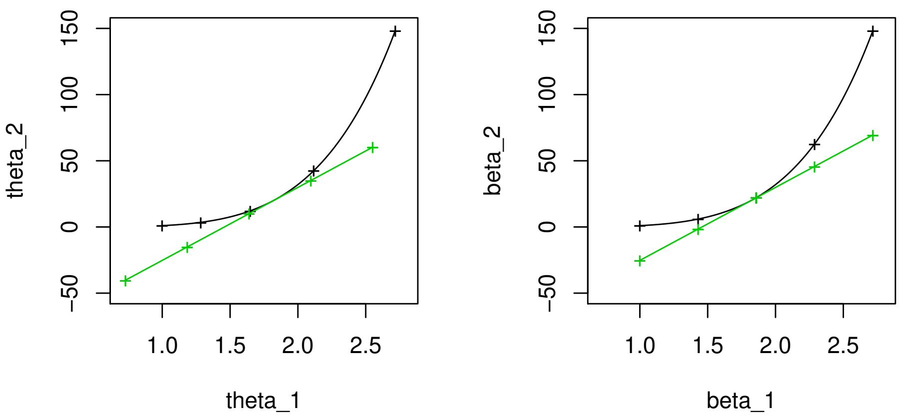
$$\mathbf{f}(\theta) = \begin{pmatrix}\mathrm{e}^{\theta}\\ \mathrm{e}^{3\theta}\end{pmatrix}, \qquad \mathbf{F}(\theta) = \begin{pmatrix}e^{\theta}\\ 3\mathrm{e}^{3\theta}\end{pmatrix}, \qquad \ddot{\mathbf{F}}(\theta) = \begin{pmatrix}e^{\theta}\\ 9\mathrm{e}^{3\theta}\end{pmatrix},$$

$$\tilde{\mathbf{f}}(\beta) = \begin{pmatrix}\beta\\ \beta^3\end{pmatrix}, \qquad \tilde{\mathbf{F}}(\beta) = \begin{pmatrix}1\\ 3\beta^2\end{pmatrix}, \qquad \ddot{\tilde{\mathbf{F}}}(\beta) = \begin{pmatrix}0\\ 6\beta\end{pmatrix}$$

a také jiné aproximace

$$\begin{aligned}
\mathbf{f}(\theta) &= \begin{pmatrix}\mathrm{e}^{\theta}\\ \mathrm{e}^{3\theta}\end{pmatrix} \doteq \begin{pmatrix}1\\ 1\end{pmatrix} + \begin{pmatrix}1\\ 3\end{pmatrix}\theta = \begin{pmatrix}1+\theta\\ 1+3\theta\end{pmatrix},\\
\tilde{\mathbf{f}}(\beta) &= \begin{pmatrix}\beta\\ \beta^3\end{pmatrix} \doteq \begin{pmatrix}1\\ 1\end{pmatrix} + \begin{pmatrix}1\\ 3\end{pmatrix}(\beta-1) = \begin{pmatrix}\beta\\ 3\beta-2\end{pmatrix}.
\end{aligned}$$

Na obrázku 15.1 jso znázorněny části množin možných středních hodnot. Zvlášť jsou vyznačeny střední hodnoty pro lineárně se měnící parametry θ (vlevo) a β (vpravo). Všimněte si, že odstupy těchto bodů nejsou stejné, na pravém obrázku se tolik nemění. Necháme-li β probíhat všechna celá čísla a θ všechna čísla větší, než -1, jsou množiny aproximujících vektorů totožné. Dále je zajímavé porovnat, jak si navzájem odpovídají dvojice bodů na křivce (množina možných středních hodnot) a aproximujíxí přímce. Parametrizace pomocí β vypadá lépe, vzdálenosti mezi sobě odpovídajícími body nejsou tak veliké. ○



Obrázek 15.1: Množiny možných středních hodnot z příkladu 15.2

15.2. Odhad vychýlení

Nyní se pokusíme vyjádřit vychýlení odhadu $\mathbf{t}$. Učiníme to nepřímo tak, že porovnáme lineární a kvadratickou aproximaci.

Další postup založíme na následujícím předpokladu: *Střední hodnota průmětu vektoru* $\mathbf{f}(\mathbf{t})$ *do tečné nadroviny k množině možných středních hodnot v bodě* $\mathbf{f}(\boldsymbol{\theta}^*)$ *je rovna* $\mathbf{f}(\boldsymbol{\theta}^*)$.

Použijme nyní kvadratickou aproximaci (15.4) na odhad $\mathbf{t}$ a vynásobme rozdíl

$\mathbf{f}(\mathbf{t}) - \mathbf{f}(\boldsymbol{\theta}^*)$ maticí $\mathbf{H}^* = \mathbf{F}^* \left(\mathbf{F}^{*\prime}\mathbf{F}^*\right)^{-1} \mathbf{F}^{*\prime}$. Aplikujeme-li na tento součin střední hodnotu, dostaneme

$$\mathbf{0} \doteq \mathbf{F}^* \mathsf{bias}\,(\mathbf{t}) + \frac{1}{2}\mathbf{H}^* \mathsf{E}\,(\mathbf{t} - \boldsymbol{\theta}^*)' \ddot{\mathbf{F}}^* (\mathbf{t} - \boldsymbol{\theta}^*).$$

Spočítejme nyní i-tou složku vektoru $\mathsf{E}\,(\mathbf{t} - \boldsymbol{\theta}^*)' \ddot{\mathbf{F}}^* (\mathbf{t} - \boldsymbol{\theta}^*)$.

$$\begin{aligned}
\mathsf{E}\,(\mathbf{t} - \boldsymbol{\theta}^*)' \ddot{\mathbf{F}}^*_{i\bullet\bullet} (\mathbf{t} - \boldsymbol{\theta}^*) &= \mathsf{E}\,(\mathbf{t} - \boldsymbol{\theta}^*)' \ddot{\mathbf{F}}^*_{i\bullet\bullet} (\mathbf{t} - \boldsymbol{\theta}^*) \\
&= \operatorname{tr} \ddot{\mathbf{F}}^*_{i\bullet\bullet} \mathsf{E}\,(\mathbf{t} - \boldsymbol{\theta}^*)(\mathbf{t} - \boldsymbol{\theta}^*)' \\
&\doteq \sigma^2 \operatorname{tr} \ddot{\mathbf{F}}^*_{i\bullet\bullet} \sigma^2 \left(\mathbf{F}^{*\prime}\mathbf{F}^*\right)^{-1} \\
&= \sigma^2 \operatorname{tr} \ddot{\mathbf{F}}^*_{i\bullet\bullet} \left(\mathbf{F}^{*\prime}\mathbf{F}^*\right)^{-1} \\
&= \sigma^2 m_i.
\end{aligned}$$

Výslednou aproximaci pro vychýlení můžeme pomocí právě zavedeného vektoru $\mathbf{m}$ zapsat jako

$$\mathsf{bias}\,\boldsymbol{\theta} = -\frac{\sigma^2}{2} \left(\mathbf{F}^{*\prime}\mathbf{F}^*\right)^{-1} \mathbf{F}^{*\prime}\mathbf{m}. \tag{15.5}$$

Příklad 15.1 (pokračování) Pokračujme v příkladu 15.1. Postupně spočítáme

$$\mathbf{F}^{*\prime}\mathbf{F}^* = \mathrm{e}^{2\theta^*}, \qquad \mathbf{m}(\theta^*) = \mathrm{e}^{-2\theta^*} \begin{pmatrix} 0 \\ \mathrm{e}^{\theta^*} \end{pmatrix},$$

takže vychýlení odhadu t je dáno vztahem

$$\mathsf{bias}\,t = -\frac{\sigma^2}{2}\mathrm{e}^{-2\theta^*}.$$

Pro naše $\theta^* = 0$ vyjde $\mathsf{bias}\,t = -\sigma^2/2$. Parametrizace pomocí β vede k nestrannému odhadu parametru β bez ohledu na jeho hodnotu, neboť je nutně $\mathbf{m} = \mathbf{0}$. ○

Příklad 15.2 (pokračování) Pokračujme v příkladu 15.2. Snadno zjistíme, že je

$$\mathbf{F}^{*\prime}\mathbf{F}^* = \mathrm{e}^{2\theta^*} + 9\mathrm{e}^{6\theta^*}$$

a také

$$\mathbf{m}(\theta^*) = \frac{1}{\mathrm{e}^{2\theta^*} + 9\mathrm{e}^{6\theta^*}} \begin{pmatrix} \mathrm{e}^{\theta^*} \\ 9\mathrm{e}^{3\theta^*} \end{pmatrix},$$

takže nakonec aproximace pro vychýlení je dána vztahem

$$\mathsf{bias}\,t = -\frac{\sigma^2}{2} \frac{\mathrm{e}^{2\theta^*} + 27\mathrm{e}^{6\theta^*}}{\left(\mathrm{e}^{2\theta^*} + 9\mathrm{e}^{6\theta^*}\right)^2}.$$

Speciálně pro $\theta^* = 0$ vyjde
$$\mathsf{bias}\, t = -\frac{\sigma^2}{2}\frac{28}{100}.$$
Analogické výpočty pro regresní funkci parametrizovanou pomocí β vede k aproximaci vychýlení
$$\mathsf{bias}\, b = -\frac{\sigma^2}{2}\frac{18\beta^{*3}}{(1+9\beta^{*4})^2},$$
což pro $\beta^* = 1$ vede k aproximaci vychýlení
$$\mathsf{bias}\, b = -\frac{\sigma^2}{2}\frac{18}{100}.$$
○

15.3. Dvojí parametrizace

V obou příkladech jsme porovnávali dvě parametrická vyjádření téže regresní funkce. Proveďme tuto úvahu obecněji.

Nechť $\boldsymbol{\beta} = \mathbf{g}(\boldsymbol{\theta})$ je regulární a prosté zobrazení jednoho parametrického prostoru na druhý. To znamená, že existuje také inverzní zobrazení $\boldsymbol{\theta} = \mathbf{g}^{-1}(\boldsymbol{\beta})$ a že čtvercová matice řádu k
$$\mathbf{G}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}'}\mathbf{g}(\boldsymbol{\theta})$$
je regulární. Souvislost mezi dvěma parametrickými vyjádřeními lze popsat jako
$$f(x,\boldsymbol{\theta}) = f(x,\mathbf{g}^{-1}(\boldsymbol{\beta}) \equiv \tilde{f}(x,\boldsymbol{\beta}) = \tilde{f}(x,\mathbf{g}(\boldsymbol{\theta})).$$

Souvislost derivací podle parametrů v obou parametrických vyjádřeních je dána vztahem
$$\begin{aligned}\frac{\partial}{\partial\theta_j}f(x,\boldsymbol{\theta}) &= \frac{\partial}{\partial\theta_j}\tilde{f}(x,\mathbf{g}(\boldsymbol{\theta}))\\ &= \sum_{r=1}^{k}\frac{\partial}{\partial\beta_r}\tilde{f}(x,\boldsymbol{\beta})\frac{\partial}{\partial\theta_j}g_r(\boldsymbol{\theta}).\end{aligned}$$

Pro matice prvních parciálních derivací pro $\boldsymbol{\beta} = \mathbf{g}(\boldsymbol{\theta})$ odtud dostáváme
$$\mathbf{F}(\boldsymbol{\theta}) = \tilde{\mathbf{F}}(\boldsymbol{\beta})\mathbf{G}(\boldsymbol{\theta}). \tag{15.6}$$

Lineární obaly matic $\mathbf{F}(\boldsymbol{\theta})$ a $\tilde{\mathbf{F}}(\boldsymbol{\beta})$ jsou tedy totožné.

Příklad 15.3 Tentokrát budeme vyšetřovat úlohu, klasicky řešenou dvouvýběrovým t testem. Mějme regresní funkci

$$f(x,\boldsymbol{\theta}) = \theta_1 x + \theta_1\theta_2(1-x),$$

přičemž $\theta_2 \neq 1$ a

$$x_i = \begin{cases} 1, & 1 \leq i \leq m, \\ 0, & m+1 \leq i \leq n. \end{cases}$$

Máme vlastně dva nezávislé výběry z normálního rozdělení se středními hodnotami $\beta_1 = \theta_1$ a $\beta_2 = \theta_1\theta_2$. Vektory možných středních hodnot mají tvar

$$\mathbf{f}(\boldsymbol{\theta}) = \begin{pmatrix} \theta_1\mathbf{1} \\ \theta_1\theta_2\mathbf{1} \end{pmatrix}, \qquad \tilde{\mathbf{f}}(\boldsymbol{\beta}) = \begin{pmatrix} \beta_1\mathbf{1} \\ \beta_2\mathbf{1} \end{pmatrix}.$$

Protože vektory prvních parciálních derivací jsou

$$\frac{\partial f(x,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \begin{pmatrix} x+\theta_2(1-x) \\ \theta_1(1-x) \end{pmatrix}, \qquad \frac{\partial\tilde{f}(x,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \begin{pmatrix} x \\ 1-x \end{pmatrix},$$

můžeme matice prvních parciálních derivací zapsat jako

$$\mathbf{F}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \theta_2\mathbf{1} & \theta_1\mathbf{1} \end{pmatrix}, \qquad \tilde{\mathbf{F}}(\boldsymbol{\beta}) = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}.$$

Snadno se ověří, že náš předpoklad $\theta_1 \neq 0$ stačí k tomu, aby obě poslední matice generovaly týž prostor. ◯

Odhad vektoru $\boldsymbol{\beta}$ metodou nejmenších čtverců je zřejmě roven $\mathbf{b} = \mathbf{g}(\mathbf{t})$. Podobně jako v (15.4) použijeme kvadratickou aproximaci a dostaneme

$$\mathbf{b} = \mathbf{g}(\mathbf{t}) \doteq \mathbf{g}(\boldsymbol{\theta}^*) + \frac{\partial\mathbf{g}(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}'}(\mathbf{t}-\boldsymbol{\theta}^*) + \frac{1}{2}(\mathbf{t}-\boldsymbol{\theta}^*)'\frac{\partial^2\mathbf{g}(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\mathbf{t}-\boldsymbol{\theta}^*).$$

Když aplikujeme na obě strany operátor střední hodnoty a použijeme vlastnost stopy matice, dostaneme po úpravách

$$\mathsf{bias}\,\mathbf{b} \doteq \mathbf{G}(\boldsymbol{\theta}^*)\,\mathsf{bias}\,\mathbf{t} + \frac{1}{2}\begin{pmatrix} \mathsf{tr}\,\dfrac{\partial^2 g_1(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mathsf{var}\,\mathbf{t} \\ \vdots \\ \mathsf{tr}\,\dfrac{\partial^2 g_k(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mathsf{var}\,\mathbf{t} \end{pmatrix}.$$

15.4. Míry křivosti

Křivost (nelinearitu) je třeba měřit. Uvedeme tedy míry křivosti a popíšeme jejich těsné spojení se skutečnou spolehlivostí konfidenčních množin (14.6) a (14.7).

Vlastní měření křivosti spočívá v porovnání lineární a kvadratické aproximace. Pro τ z otevřeného okolí nuly uvažujme křivku $\mathbf{f}(\boldsymbol{\theta}+\tau\mathbf{h})$, která leží na ploše všech možných středních hodnot. Porovnáme tuto křivku s lineární a kvadratickou aproximací:

$$\begin{aligned}\mathbf{f}(\boldsymbol{\theta}+\tau\mathbf{h}) &= \mathbf{f}(\boldsymbol{\theta}) + \mathbf{F}(\boldsymbol{\theta})\tau\mathbf{h} + \frac{\tau^2}{2}\mathbf{h}'\ddot{\mathbf{F}}(\boldsymbol{\theta})\mathbf{h} \\ &= \mathbf{f}(\boldsymbol{\theta}) + \tau\dot{\mathbf{f}}_h + \frac{\tau^2}{2}\ddot{\mathbf{f}}_h,\end{aligned}$$

kde jsme zavedli označení pro vektor oprav lineární a kvadratické aproximaci. Oba vektory závisí na volbě nenulového vektoru $\mathbf{h}$.

Parametr τ zvolíme tak, aby oprava lineární aproximace měla zvolenou velikost.

$$||\tau\dot{\mathbf{f}}_h||^2 = \sigma\sqrt{k} \qquad \Rightarrow \qquad \tau^2 = \frac{\sigma\sqrt{k}}{||\dot{\mathbf{f}}_h||^2}$$

Dvojnásobek opravy kvadratické aproximace vůči lineární aproximaci rozložíme do dvou složek, z nichž jedna je ortogonální vůči tečné nadrovině (označená horním indexem N) a rovnoběžné s touto nadrovinou (označená horním indexem T)

$$\frac{\sigma\sqrt{k}}{||\dot{\mathbf{f}}_h||^2}\ddot{\mathbf{f}}_h^N + \frac{\sigma\sqrt{k}}{||\dot{\mathbf{f}}_h||^2}\ddot{\mathbf{f}}_h^T.$$

Velikosti těchto složek nazveme po řadě jako *vnitřní křivost* ve směru $\mathbf{h}$

$$K_h^N = \frac{\sigma\sqrt{k}}{||\dot{\mathbf{f}}_h||^2}||\ddot{\mathbf{f}}_h^N||$$

a jako *parametrická křivost* ve směru $\mathbf{h}$

$$K_h^T = \frac{\sigma\sqrt{k}}{||\dot{\mathbf{f}}_h||^2}||\ddot{\mathbf{f}}_h^T||.$$

Pracuje se zpravidla s maximálními hodnotami těchto křivostí

$$K^N = \max K_h^N, \qquad K^T = \max K_h^T,$$

kde se hledá maximum přes všechny nenulové vektory $\mathbf{h}$.

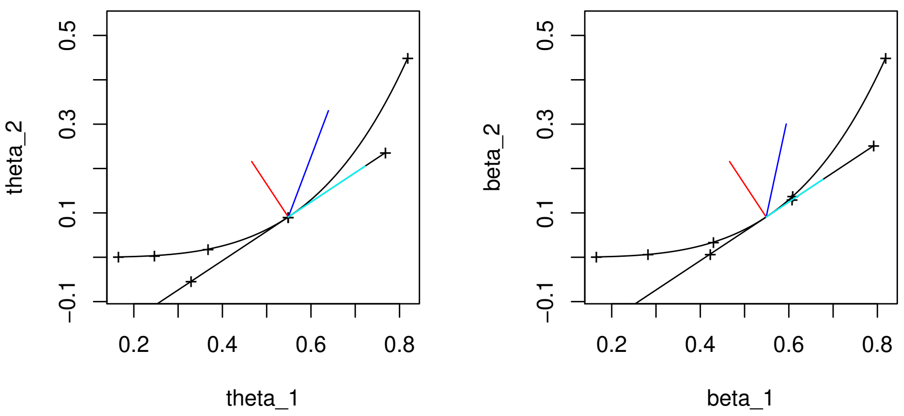
Když si uvědomíme, že $\dot{\mathbf{f}}_h$ je lineární funkcí vektoru $\mathbf{h}$, kdežto $\ddot{\mathbf{f}}_h$ je kvadratickou funkcí tohoto vektoru, je zřejmé, že stačí hledat maximum přes všechny vektory splňující $||\mathbf{h}|| = 1$.

Velikost křivosti se někdy hodnotí porovnáním s hodnotou $(F_{k,n-k}(\alpha))^{-1/2}$. V rozsáhlém simulačním experimentu založeném na datech z reálných úloh zjistili Donaldsonová a Schnabel (viz Donaldson, Schnabel (1987)), že skutečná spolehlivost eliptické konfidenční množiny (14.6) těsně souvisí s hodnotou

$$\log(K^T \sqrt{F_{k,n-k}(\alpha)}).$$

Pro parametrickou křivost K^T větší, než uvedená mezní hodnota, skutečná spolehlivost této eliptické konfidenční množiny velmi rychle klesá s rostoucí hodnotou parametrické křivosti.

Příklad 15.4 Vraťme se k příkladu 15.1, ale zvolíme $x_1 = 2, x_2 = 8$. Zvolíme-li dále $\sigma = 1/\sqrt{(2)}$, dostaneme v bodě $\theta = -0{,}3$ vnitřní křivost 2,1 a parametrickou křivost 2,9. Přejdeme-li k druhé parametrizaci, vyjde parametrická křivost 2,1, vnitřní křivost zůstane stejná. Na obrázku 15.2 je znázorněn rozklad vektoru druhých derivací. Je patrné, že ve druhém parametrickém vyjádření je průmět tohoto vektoru do tečné nadroviny kratší. ○



Obrázek 15.2: Množiny možných středních hodnot a rozklad vektoru druhých derivací z příkladu 15.4

A. Pomocná tvrzení, označení

Zde jsou uvedena některá tvrzení (například o maticích), užitečná v ostatních kapitolách.

A.1. Tvrzení o maticích

Chceme-li označit j-tý sloupec (i-tý řádek) matice $\mathbf{A}$, použijeme symbol $\mathbf{a}_{\bullet j}$ ($\mathbf{a}'_{i\bullet}$) Chceme-li vyjádřit, že matice vznikla z $\mathbf{A}$ vynecháním jejího j-tého sloupce, napíšeme $\mathbf{A}_{[\bullet j]}$, když vznikla vynecháním i-tého řádku, pak píšeme $\mathbf{A}_{[i\bullet]}$. Je tedy například

$$\mathbf{A} = \left(\mathbf{a}_{\bullet 1}, \mathbf{A}_{[\bullet 1]}\right) = \begin{pmatrix} \mathbf{a}'_{1\bullet} \\ \mathbf{A}_{[1\bullet]} \end{pmatrix} \tag{A.1}$$

Speciálně r-tý sloupec jednotkové matice $\mathbf{I}$ označíme symbolem $\mathbf{j}_r$, vektor ze samých jedniček symbolem $\mathbf{1}$, případně $\mathbf{1}_n$, pokud chceme explicitně vyjádřit počet složek.

Nechť $\mathbf{X}_{n\times k}$ je pevná matice. Symbolem $\mathcal{M}(\mathbf{X})$ označíme podprostor $\mathbb{R}^n$ tvořený všemi lineárními kombinacemi sloupců matice $\mathbf{X}$. Tento prostor, nazývaný **lineární obal sloupců matice X**, vlastně splňuje

$$\mathcal{M}(\mathbf{X}) = \{\mathbf{X}\mathbf{t} : \mathbf{t} \in \mathbb{R}^k\}.$$

Je-li matice $\mathbf{X}$ nějaká matice typu $n \times k$, pak **pseudoinverzní matice** k matici $\mathbf{X}$ je libovolná matice $\mathbf{X}^-$ typu $k \times n$, která vyhovuje vztahu $\mathbf{X}\mathbf{X}^-\mathbf{X} = \mathbf{X}$. Pseudoinverzní matice obecně není dána jednoznačně.

Jednoznačně je však dána **Mooreova-Penroseho pseudoinverzní matice**, která musí vyhovovat požadavkům:

$$\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}, \qquad \mathbf{X}^+\mathbf{X}\mathbf{X}^+ = \mathbf{X}^+, \tag{A.2}$$

$$(\mathbf{X}\mathbf{X}^+)' = \mathbf{X}\mathbf{X}^+, \qquad (\mathbf{X}^+\mathbf{X})' = \mathbf{X}^+\mathbf{X}. \tag{A.3}$$

Věta A.1. (Spektrální rozklad) Nechť $\mathbf{A}$ je symetrická matice řádu n. Potom existují ortonormální matice $\mathbf{Q}$ a diagonální matice $\mathbf{\Lambda}$ s diagonálními prvky $\lambda_1 \geq \ldots \geq \lambda_n$ tak, že platí

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'. \tag{A.4}$$

Je zřejmé, že λ_i jsou vlastní čísla matice $\mathbf{A}$ a že sloupce $\mathbf{q}_{\bullet i}$ matice $\mathbf{Q}$ jsou odpovídající ortonormální vlastní vektory s jednotkovou délkou. Matici $\mathbf{A}$ lze vyjádřit ve tvaru

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{q}_{\bullet i}\mathbf{q}'_{\bullet i}. \tag{A.5}$$

Věta A.2. (SVD – rozklad podle singulárních hodnot) Nechť $\mathbf{X}_{n\times k}$, kde je $n \geq k$ je matice s kladnou hodností r. Potom existují matice s ortonormálními sloupci $\mathbf{U}^0_{n\times r}, \mathbf{V}^0_{k\times r}$ a diagonální matice $\mathbf{D}^0_{r\times r}$ s reálnými čísly $d_1 \geq \ldots \geq d_r > 0$ na diagonále tak, že platí

$$\mathbf{X} = \mathbf{U}^0\mathbf{D}^0\mathbf{V}^{0\prime} \tag{A.6}$$

D ů k a z: Uvažujme zřejmě pozitivně semidefinitní matici $\mathbf{X}'\mathbf{X}$ s vlastními čísly $d_1^2 \geq \ldots \geq d_r^2 > d_{r+1}^2 = \ldots = d_k^2 = 0$ a jim odpovídajícími ortonormálními vlastními vektory $\mathbf{v}_1, \ldots, \mathbf{v}_k$. Pro $1 \leq i \leq r$ zaveďme vektory

$$\mathbf{u}_i = \frac{1}{d_i}\mathbf{X}\mathbf{v}_i. \tag{A.7}$$

Snadno zjistíme, že tyto vektory jsou ortonormální:

$$\mathbf{u}_i'\mathbf{u}_j = \frac{1}{d_i d_j}\mathbf{v}_i'\mathbf{X}'\mathbf{X}\mathbf{v}_j = \frac{d_j^2}{d_i d_j}\mathbf{v}_i'\mathbf{v}_j = \begin{cases} 0 & \text{pro } i \neq j, \\ 1 & \text{pro } i = j. \end{cases}$$

Vztah z (A.7) lze přepsat jako

$$\mathbf{u}_i d_i = \mathbf{X}\mathbf{v}_i,$$

a to dokonce pro všechna $1 \leq i \leq k$, když libovolně přidáme vektory $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_k$ tak, aby sloupce matice $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_k)$ měla ortonormální sloupce. Zavedeme-li ještě čtvercovou matici $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_k)$ a diagonální matici $\mathbf{D}$ s diagonálními prvky $d_1, \ldots, d_k$, můžeme všech k vztahů souhrnně zapsat jako $\mathbf{UD} = \mathbf{XV}$. Odtud přímo plyne vztah

$$\mathbf{X} = \mathbf{UDV}' = \sum_{i=1}^{k} d_i \mathbf{u}_i\mathbf{v}_i' = \sum_{i=1}^{r} d_i \mathbf{u}_i\mathbf{v}_i'. \tag{A.8}$$

Přitom je vidět, že vystačíme s prvními r sloupci matic $\mathbf{U}, \mathbf{D}, \mathbf{V}$. Označíme-li horním indexem 0 odpovídající podmatice, dostaneme vztah (A.6). □

Věta A.3. (QR rozklad) Nechť $\mathbf{X}_{n\times k}$ je matice konstant. Potom existují matice $\mathbf{Q}_{n\times k}$ s ortonormálními sloupci a horní trojúhelníková čtvercová matice $\mathbf{R}$ řádu k tak, že platí

$$\mathbf{X} = \mathbf{QR}. \tag{A.9}$$

Je-li hodnost r matice $\mathbf{X}$ kladná, existují matice $\mathbf{Q}^0_{n\times r}$ s ortonormálními sloupci a matice $\mathbf{R}^0$ s r řádky a k sloupci taková, že je $r^0_{ij} = 0$ pro $i > j$ a že platí

$$\mathbf{X} = \mathbf{Q}^0\mathbf{R}^0. \tag{A.10}$$

Je-li hodnost matice $\mathbf{X}$ rovna počtu jejích sloupců, pak existuje jediná matice $\mathbf{R}$ splňující (A.9), která má kladné diagonální prvky, nazývá se *Choleského faktor*.

Existence rozkladu (A.9) je dokázána v oddílu 1b.2 (VII) knihy Rao (1978). V jednotlivých sloupcích matice $\mathbf{R}$ jsou souřadnice odpovídajících sloupců matice $\mathbf{X}$ v ortonormální bázi tvořené sloupci matice $\mathbf{Q}$. Pokud nemá matice $\mathbf{X}$ lineárně nezávislé sloupce, pak se v součinu (A.9) nesmí projevit některé sloupce matice $\mathbf{Q}$. To je zajištěno, když jsou odpovídající řádeky $\mathbf{R}$ nulové. Jednoznačnost $\mathbf{R}$ v případě matice $\mathbf{X}$ s lineárně nezávislými sloupci lze dokázat indukcí ((Zvára, 1989, věta 12.1)). Z jednoznačnosti $\mathbf{R}$ plyne v tomto případě také jednoznačnost matice $\mathbf{Q}$.

Věta A.4. (Odmocninová matice) Nechť $\mathbf{A}$ je pozitivně semidefinitní matice. Pak existuje pozitivně semidefinitní matice $\mathbf{C}$ taková, že platí

$$\mathbf{A} = \mathbf{CC}.$$

D ů k a z: Nechť $\mathbf{A} = \mathbf{Q\Lambda Q}'$ je spektrální rozklad matice $\mathbf{A}$. Pozitivní semidefinitnost $\mathbf{A}$ je ekvivalentní se stejnou vlastností $\mathbf{\Lambda}$. Označme jako $\mathbf{\Lambda}^{1/2}$ diagonální matici, která má na diagonále odmocniny ze stejných prvků matice Λ. Snadno se ověří, že matice $\mathbf{C} = \mathbf{Q\Lambda}^{1/2}\mathbf{Q}'$ má požadované vlastnosti. □

Pozitivně semidefinitní matice budeme značit $\mathbf{A} \geq 0$, podobně zápis $\mathbf{A} \geq \mathbf{B}$ znamená, že matice $\mathbf{A} - \mathbf{B}$ je pozitivně semidefinitní. Analogicky použijeme symbol $>$ k vyjádření pozitivní definitnosti.

Věta A.5. (Porovnání kvadratických forem) Nechť $\mathbf{A}, \mathbf{B}$ jsou dvě pozitivně definitní matice. Potom platí

$$\mathbf{A} \geq \mathbf{B} \Leftrightarrow \mathbf{B}^{-1} \geq \mathbf{A}^{-1}, \tag{A.11}$$

$$\mathbf{A} > \mathbf{B} \Leftrightarrow \mathbf{B}^{-1} > \mathbf{A}^{-1}. \tag{A.12}$$

Věta A.6. (Projekce do podprostoru) Nechť $\mathbf{X}_{n\times k}$ je matice, jejíž hodnost r je kladná. Potom

a) rozklad $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2$, kde $\mathbf{y}_1 \in \mathcal{M}(\mathbf{X})$ a $\mathbf{y}_2 \perp \mathcal{M}(\mathbf{X})$, je dán jednoznačně;

b) nechť $\mathbf{P} = (\mathbf{Q}, \mathbf{N})$ je ortonormální matice taková, že je $\mathcal{M}(\mathbf{X}) = \mathcal{M}(\mathbf{Q})$. Projekční matice $\mathbf{H}_X$ a $\mathbf{M}_X$, které zajišťují průměty $\mathbf{y}_1, \mathbf{y}_2$, jsou dány jednoznačně a platí

$$\mathbf{H}_X = \mathbf{QQ}', \tag{A.13}$$

$$\mathbf{M}_X = \mathbf{NN}'. \tag{A.14}$$

c) Platí

$$\mathbf{H}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}', \tag{A.15}$$
$$\mathbf{M}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'; \tag{A.16}$$

d) matice $\mathbf{H}_X, \mathbf{M}_X$ jsou symetrické a idempotentní.

e) Platí

$$\operatorname{tr}(\mathbf{H}_X) = r, \tag{A.17}$$
$$\operatorname{tr}(\mathbf{M}_X) = n - r. \tag{A.18}$$

Věta A.7. (Porovnání délky vektoru s jedničkou) Pro matici $\mathbf{A}_{m\times n}$ a vektor $\mathbf{c} \in \mathbb{R}^n$ platí nerovnost $||\mathbf{Ac}||^2 \le 1$ právě tehdy, když je matice

$$\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}' - \mathbf{Acc}'\mathbf{A}' \tag{A.19}$$

pozitivně semidefinitní.

D ů k a z: Pro $\mathbf{Ac} = \mathbf{0}$ je tvrzení triviální. Nechť je tedy $\mathbf{Ac} \neq \mathbf{0}$. Potom platí $\mathcal{M}(\mathbf{Ac}) \subset \mathcal{M}(\mathbf{A})$, takže rozdíl projekčních matic na $\mathcal{M}(\mathbf{A})$ a na $\mathcal{M}(\mathbf{Ac})$ je projekční maticí na ortogonální doplněk $\mathcal{M}(\mathbf{Ac})$ prostoru $\mathcal{M}(\mathbf{A})$. Pozitivně semidefinitní je tedy

$$0 \le \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}' - \mathbf{Ac}(\mathbf{c}'\mathbf{A}'\mathbf{Ac})^{-1}\mathbf{c}'\mathbf{A}'. \tag{A.20}$$

Předpoklad $||\mathbf{Ac}||^2 \le 1$ je však ekvivalentní s $-(\mathbf{c}'\mathbf{A}'\mathbf{Ac})^{-1} \le -1$, takže pravou stranu nerovnosti (A.20) můžeme shora omezit maticí $\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}' - \mathbf{Acc}'\mathbf{A}'$, která je tedy nutně pozitivně semidefinitní a je dokázána implikace jedním směrem.

Obráceně, nechť je matice (A.19) pozitivně semidefinitní. Když ji vynásobíme zprava vektorem $\mathbf{Ac}$ a zleva transpozicí tohoto vektoru, dostaneme po malé úpravě (použitím definice pseudoinverzní matice)

$$0 \le ||\mathbf{Ac}||^2 - ||\mathbf{Ac}||^4 = ||\mathbf{Ac}||^2(1 - ||\mathbf{Ac}||^2),$$

což je ekvivalentní s dokazovanou nerovností $||\mathbf{Ac}||^2 \le 1$. □

Věta A.8. (Porovnání délky vektoru s jedničkou*) Nechť $\mathbf{V}$ je pozitivně definitní matice řádu k, nechť $\mathbf{b} \in \mathbb{R}^k$ je libovolný vektor. Potom platí nerovnost $\mathbf{b}'\mathbf{V}^{-1}\mathbf{b} \le 1$ právě tehdy, když je matice $\mathbf{V} - \mathbf{bb}'$ pozitivně semidefinitní.

D ů k a z: Pozitivně definitní matici $\mathbf{V}^{-1}$ lze zapsat pomocí symetrické a regulární odmocninové matice (viz větu A.4) jako $\mathbf{V}^{-1} = \mathbf{AA}$. Kvadratickou formu $\mathbf{b}'\mathbf{V}^{-1}\mathbf{b}$ lze tedy přepsat jako

$$\mathbf{b}'\mathbf{AAb} = ||\mathbf{Ab}||^2.$$

Podle věty A.7 je tedy nerovnost $\mathbf{b}'\mathbf{V}^{-1}\mathbf{b} \leq 1$ ekvivalentní s tím, že je pozitivně semidefinitní matice

$$\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A} - \mathbf{A}\mathbf{b}\mathbf{b}'\mathbf{A} = \mathbf{A}\left(\mathbf{V} - \mathbf{b}\mathbf{b}'\right)\mathbf{A}.$$

Protože je matice $\mathbf{A}$ regulární, je ona nerovnost ekvivalentní s pozitivní semidefinitností matice $\mathbf{V} - \mathbf{b}\mathbf{b}'$, což bylo dokázat. □

Když pracujeme s vektory označenými dvojitými indexy (například v modelech analýzy rozptylu dvojného třídění), je užitečný pojem **Kroneckerova součinu**. Jsou-li $\mathbf{A}$ typu $m \times n$ a $\mathbf{B}$ typu $p \times q$, pak označíme jako $\mathbf{A} \otimes \mathbf{B}$ matici typu $mp \times nq$, jejíž blok (i,j) je roven $a_{ij}\mathbf{B}$, tedy

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix}. \tag{A.21}$$

Následující vlastnosti lze snadno dokázat.

Věta A.9. (Vlastnosti Kroneckerova součinu) Pro Kroneckerův součin platí

$$\begin{aligned} \mathbf{O} \otimes \mathbf{A} = \mathbf{A} \otimes \mathbf{O} &= \mathbf{O}, \\ (\mathbf{A}_1 + \mathbf{A}_2) \otimes \mathbf{B} &= (\mathbf{A}_1 \otimes \mathbf{B}) + (\mathbf{A}_2 \otimes \mathbf{B}), \\ \mathbf{A} \otimes (\mathbf{B}_1 + \mathbf{B}_2) &= (\mathbf{A} \otimes \mathbf{B}_1) + (\mathbf{A} \otimes \mathbf{B}_2). \\ c\mathbf{A} \otimes d\mathbf{B} &= cd(\mathbf{A} \otimes \mathbf{B}), \\ \mathbf{A}_1\mathbf{A}_2 \otimes \mathbf{B}_1\mathbf{B}_2 &= (\mathbf{A}_1 \otimes \mathbf{B}_1)(\mathbf{A}_2 \otimes \mathbf{B}_2), \\ (\mathbf{A} \otimes \mathbf{B})^{-1} &= \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}, \quad \text{pokud inverze existují}, \\ (\mathbf{A} \otimes \mathbf{B})^{-} &= \mathbf{A}^{-} \otimes \mathbf{B}^{-}, \quad \text{pro libovolné pseudoinverze}, \\ (\mathbf{A} \otimes \mathbf{B})' &= \mathbf{A}' \otimes \mathbf{B}'. \end{aligned}$$

Věta A.10. (Poincaréova věta o separaci) Nechť $\mathbf{R}$ je matice typu $n \times q$ s ortonormálními sloupci, nechť $\alpha_1 \geq \ldots \geq \alpha_n$ jsou vlastní čísla nějaké symetrické matice $\mathbf{A}$, nechť $\lambda_1 \geq \ldots \geq \lambda_q$ jsou vlastní čísla matice $\mathbf{R}'\mathbf{A}\mathbf{R}$. Potom platí

$$\lambda_i \leq \alpha_i, \qquad 1 \leq i \leq q, \tag{A.22}$$

$$\lambda_{q-i+1} \geq \alpha_{n-i+1}, \qquad 1 \leq i \leq q. \tag{A.23}$$

Platí-li navíc pro vlastní vektor $\mathbf{q}_n$ matice $\mathbf{A}$ odpovídající jejímu vlastnímu číslu α_n vztah $\mathbf{R}'\mathbf{q}_n = \mathbf{0}$, lze nerovnost (A.23) upravit na

$$\lambda_{q-i+1} \geq \alpha_{n-i}, \qquad 1 \leq i \leq q. \tag{A.24}$$

Tvrzení lze nalézt v 1. kapitole knihy Rao (1978) resp. ve cvičeních 1.4, 1.5 2. kapitoly knihy Zvára (1989).

A.2. Některé vlastnosti náhodných veličin

Věta A.11. (Vlastnosti kvadratické formy) Nechť $e_1, \ldots, e_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, nechť $\mathsf{E}\, e_i = 0, \mathsf{E}\, e_i^2 = \sigma^2, \mathsf{E}\, e_i^4 = \sigma^4(\gamma_2+3)$. Nechť $\mathbf{A}$ je symetrická matice. Potom platí

$$\mathsf{E}\, \mathbf{e}'\mathbf{A}\mathbf{e} = \sigma^2 \operatorname{tr} \mathbf{A}, \tag{A.25}$$

$$\operatorname{var} \mathbf{e}'\mathbf{A}\mathbf{e} = \sigma^4 \left(\gamma_2 \sum a_{ii}^2 + 2 \operatorname{tr} \mathbf{A}^2\right). \tag{A.26}$$

Věta A.12. (Vlastnost normálního rozděleni) Nechť měřitelná funkce $T(\mathbf{x})$ splňuje $T(c\mathbf{x}) = T(\mathbf{x})$ pro každé $c > 0$ a pro každé $\mathbf{x} \in \mathbb{R}^n$. Má-li náhodný vektor $\mathbf{X}$ rozdělení $\mathsf{N}_n\big(\mathbf{0}, \sigma^2\mathbf{I}\big)$, pak jsou náhodné veličiny $T(\mathbf{X})$ a $||\mathbf{X}||$ nezávislé.

D ů k a z: Stačí přejít k polárním souřadnicím. Potom vzdálenost náhodného bodu od počátku a jeho směr od počátku jsou nezávislé. Ovšem vzdálenost od počátku je rovna $||\mathbf{X}||$ a funkční hodnota $T(\mathbf{X})$ je vzhledem k požadované vlastnosti závisí pouze na směru od počátku. □

Věta A.13. (Bonferroniho nerovnost) Pro náhodné jevy $A_1, \ldots, A_n$ platí

$$\mathsf{P}\left(\cup_{i=1}^n A_i\right) \le \sum_{i=1}^n \mathsf{P}\left(A_i\right),$$

$$\mathsf{P}\left(\cap_{i=1}^n A_i\right) \ge 1 - \sum_{i=1}^n \left(1 - \mathsf{P}(A_i)\right).$$

A.3. Metoda maximální věrohodnosti

Nechť má náhodný vektor $\mathbf{X}$ hustotu $f_\theta(\mathbf{x})$, která závisí na parametru $\boldsymbol{\theta} \in \Omega$, přičemž Ω je parametrický prostor. V případě diskrétního rozdělení míníme pod hustotou pravděpodobnostní funkci (hustotu vůči čítací míře). Jako logaritmickou věrohodnostní funkci označíme funkci

$$\ell(\boldsymbol{\theta}) = \log(f_\theta(\mathbf{X})), \tag{A.27}$$

je tedy pro každé $\boldsymbol{\theta}$ náhodnou veličinou.

Odhad $\hat{\boldsymbol{\theta}}$ metodou maximální věrohodnosti je takový prvek parametrického prostoru, v němž je logaritmická věrohodnostní funkce maximální. Například v lineárním modelu $\mathbf{Y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ dá metoda maximální věrohodnosti odhady

$$\hat{\boldsymbol{\beta}} = \mathbf{b}, \qquad \hat{\sigma^2} = \frac{RSS}{n}.$$

Logaritmická věrohodnostní funkce je rovna

$$\ell(\hat{\boldsymbol{\beta}}, \hat{\sigma^2}) = -\frac{n}{2} - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(RSS/n). \tag{A.28}$$

Pokud bychom považovali rozptyl σ^2 za známý (neodhadovaný), vyšla by logaritmická věrohodnostní funkce

$$\ell(\hat{\boldsymbol{\beta}}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}RSS \tag{A.29}$$

Jsou-li splněny podmínky regularity, potom lze dokázat mnohé užitečné vlastnosti odhadu $\hat{\boldsymbol{\theta}}$. Asymptoticky má rozdělení $\mathsf{N}(\boldsymbol{\beta}, \mathbf{J}^{-1})$, kde $\mathbf{J}$ je *Fisherova informační matice* s prvky

$$J_{jt}(\boldsymbol{\theta}) = \mathsf{E}\,\frac{\partial\ell(\boldsymbol{\theta})}{\partial\theta_j}\frac{\partial\ell(\boldsymbol{\theta})}{\partial\theta_t} = -\mathsf{E}\,\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_t}. \tag{A.30}$$

Ke zmíněným podmínkám regularity patří požadavek, aby množina $\{\mathbf{x} : f_\theta(\mathbf{x}) > 0\}$ nezávisela na parametru $\boldsymbol{\theta}$ nebo požadavek, aby parametrický prostor byl otevřená množina.

Podmodel je určen vlastním podprostorem $\omega \subset \Omega$. Odhad $\tilde{\boldsymbol{\theta}}$ v podmodelu je takovým prvkem ω, který maximalizuje logaritmickou věrohodnostní ℓ na ω. Testování podmodelu lze založit na některé ze tří statistik, které mají všechny stejné asymptotické rozdělení. Je jím rozdělení $\chi^2(q)$, kde q je rozdíl dimenze prostorů Ω a ω, resp. počet nezávislých omezení, jejichž aplikace vede k náhradě parametrického prostoru Ω parametrickým prostorem ω.

Test poměrem věrohodnosti (Wilksův test) porovnává hodnoty logaritmické věrohodnostní funkce pro $\hat{\boldsymbol{\theta}}$ a $\tilde{\boldsymbol{\theta}}$ pomocí statistiky

$$LR = 2\left(\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})\right). \tag{A.31}$$

Platí-li podmodel, pak za předpokladu splnění podmínek regularity má statistika LR asymptoticky rozdělení $\chi^2(q)$.

Waldův test předpokládá, že se od Ω dostaneme k ω tak, že požadujeme, aby parametr $\boldsymbol{\theta}$ vyhovoval omezením $g_j(\boldsymbol{\theta}) = 0, j = 1, \ldots, q$. Tato omezení lze psát vektorově jako $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{0}$. Myšlenka je založena na zjištění, nakolik odhad $\hat{\boldsymbol{\theta}}$ vyhovuje uvedeným omezením.

Označme jako $\mathbf{A}(\boldsymbol{\theta})$ matici parciálních derivací $\partial \mathbf{g}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}'$. Asymptotická varianční matice vektoru $\mathbf{g}(\hat{\boldsymbol{\theta}})$ je rovna výrazu $\mathbf{A}(\boldsymbol{\theta})\mathbf{J}(\boldsymbol{\theta})^{-1}\mathbf{A}(\boldsymbol{\theta})'$. Prakticky sem musíme za neznámý parametr dosadit jeho odhad. Asymptoticky má výraz

$$W = \mathbf{g}(\hat{\boldsymbol{\theta}})' \left(\mathbf{A}(\hat{\boldsymbol{\theta}})\mathbf{J}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{A}(\hat{\boldsymbol{\theta}})'\right)^{-1} \mathbf{g}(\hat{\boldsymbol{\theta}}) \tag{A.32}$$

rozdělení $\chi^2(q)$.

Metoda skórů (Lagrangeova multiplikátoru) využívá na rozdíl od Waldova testu pouze odhad v podmodelu. Maximálně věrohodný odhad, protože maximalizuje logaritmickou věrohodnostní funkci, musí anulovat vektor parciálních derivací $\partial \ell / \partial \boldsymbol{\theta}$. Vyzkoušíme tedy, nakolik také odhad v podmodelu $\tilde{\boldsymbol{\theta}}$ anuluje tento vektor.

Zaveďme náhodný vektor (*vektor skórů*)

$$\mathbf{U}(\tilde{\boldsymbol{\theta}}) = \frac{\partial \ell(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}. \tag{A.33}$$

Platí-li podmodel, má tento vektor nutně nulovou střední hodnotu, takže jeho varianční matice je právě rovna Fisherově informační matici, jak je zřejmé z definice (A.30) prvků této matice. Proto má, platí-li podmodel, statistika

$$LM = \frac{\partial \ell(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'} \left(\mathbf{J}(\tilde{\boldsymbol{\theta}})\right)^{-1} \frac{\partial \ell(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \tag{A.34}$$

asymptoticky rozdělení $\chi^2(q)$.

B. Prostředí R

V této části shrneme některé informace spíše technického charakteru o programovém prostředí R. Pro podrobnou informaci a získání jeho volně šiřitelného kódu doporučuji především internetovou adresu `http://www.r-project.org/`, kde lze nalézt program, jednotlivé knihovny i manuály. Pro začátek je velmi instruktivní projít si ukázkovou lekci.

B.1. Procedura `lm()`

V prostředí R metodě nejmenších čtverců odpovídá procedura `lm`, věnujme se jí podrobněji. Viděli jsme, že metodu nejmenších čtverců můžeme do značné míry vyjádřit pomocí ortogonálního rozkladu regresní matice. Základem procedury `lm()` je rozklad matice $\mathbf{X}$ na součin matice $\mathbf{Q}$ s ortonormálními sloupci a horní trojúhelníkové matice $\mathbf{R}$, která obsahuje „souřadnice“ jednotlivých sloupců matice $\mathbf{X}$, vyjádřených pomocí sloupců matice $\mathbf{Q}$:

$$\mathbf{X} = \mathbf{QR}. \tag{B.1}$$

Existence tohoto *QR rozkladu* je dokázána například v oddílu 1b.2 (VII) knihy Rao (1978). Samotný výpočet je založen na Householderových transformacích, kdy matice $\mathbf{P} = (\mathbf{Q}, \mathbf{N})$ vzniká jako součin ortonormálních matic tvaru $\mathbf{I} - 2\mathbf{qq}'$, kde $\mathbf{q}$ je vhodný vektor jednotkové délky. Zajímavý výklad poskytne oddíl 2.7 knihy Antoch, Vorlíčková (1992).

V případě, že matice $\mathbf{X}$ nemá lineárně nezávislé sloupce, není matice $\mathbf{Q}$ z QR rozkladu totožná s maticí $\mathbf{Q}$ z úvodu této kapitoly, jejíž sloupce tvoří ortonormální bázi prostoru $\mathcal{M}(\mathbf{X})$, nýbrž generuje větší lineární prostor. Abychom dostali z QR rozkladu skutečnou bázi $\mathcal{M}(\mathbf{X})$, musíme z matice $\mathbf{Q}$ použít jen ty sloupce, jimž odpovídající řádky matice $\mathbf{R}$ jsou nenulové. To znamená použít rozklad (A.10). Algoritmus QR rozkladu v R je modifikací procedury DQRDC souboru programů LINPACK.

Možno říci, že matice $\mathbf{Q}$ (přesněji by to byla matice $\mathbf{Q}^0$ z (A.10)) vypovídá o lineárním prostoru $\mathcal{M}(\mathbf{X})$, kde se hledá odhad $\hat{\mathbf{Y}}$. Tato matice rozhoduje o varianční matici zmíněného odhadu. Na druhé straně matice $\mathbf{R}$ (přesněji $\mathbf{R}^0$ z (A.10)) zachycuje vztahy mezi sloupci matice $\mathbf{X}$, rozhoduje tedy o rozptylu každé odhadnutelné funkce β, v případě úplné hodnosti o varianční matici $\mathbf{b}$.

Ukažme si funkci `lm()` na primitivním příkladu s následujícími daty:

$$\mathbf{X} = \begin{pmatrix} 1 & -3 & 9 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{X}_a \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} -9 \\ -11 \\ 1 \\ 19 \end{pmatrix}, \qquad \mathbf{w} = \begin{pmatrix} 1 \\ 1 \\ 4 \\ 1 \end{pmatrix}, \tag{B.2}$$

přičemž diagonální matice $\mathbf{W}$ má na diagonále prvky vektoru $\mathbf{w}$. Začneme však bez vážení, tedy bez $\mathbf{W}$ resp. $\mathbf{w}$.

B.1.1. Úloha bez vah

Provedeme-li standardní Gramovu-Schmidtovu ortogonalizaci sloupců matice $\mathbf{X}$ a přidáme zbývající vektor, dostaneme ortonormální matici, jejíž sloupce tvoří bázi $\mathbb{R}^4$. Je třeba mít na paměti, že tato matice není dána jednoznačně, že když například vynásobíme některé (nebo všechny) sloupce konstantou -1, dostaneme matici se stejnými vlastnostmi. Následující vyjádření má znaménka zvolena tak, aby bylo konzistentní s výsledkem programu R.

$$\mathbf{P} = (\mathbf{Q}, \mathbf{N}) = \left(\begin{pmatrix} -1/2 & 3/\sqrt{20} & 1/2 \\ -1/2 & 1/\sqrt{20} & -1/2 \\ -1/2 & -1/\sqrt{20} & -1/2 \\ -1/2 & -3/\sqrt{20} & 1/2 \end{pmatrix}, \begin{pmatrix} 1/\sqrt{20} \\ -3/\sqrt{20} \\ 3/\sqrt{20} \\ -1/\sqrt{20} \end{pmatrix} \right).$$

Souřadnice jednotlivých sloupců matice $\mathbf{X}$ obsahuje matice $\mathbf{R}$

$$\mathbf{R} = \mathbf{Q}'\mathbf{X} = \begin{pmatrix} -2 & 0 & -10 \\ 0 & -\sqrt{20} & 0 \\ 0 & 0 & 8 \end{pmatrix}. \tag{B.3}$$

Souřadnice vektoru $\mathbf{y}$ v bázi tvořené sloupci matice $\mathbf{P}$ jsou dány vztahem

$$\mathbf{P}'\mathbf{y} = \begin{pmatrix} 0 \\ -96/\sqrt{20} \\ 10 \\ 8/\sqrt{20} \end{pmatrix} = \begin{pmatrix} 0 \\ -21{,}466253 \\ 10 \\ 1{,}788854 \end{pmatrix}.$$

Odtud je pomocí prvních tří složek vektoru $\mathbf{P}'\mathbf{y}$

$$\hat{\mathbf{y}} = 0 \begin{pmatrix} -1/2 \\ -1/2 \\ -1/2 \\ -1/2 \end{pmatrix} - \frac{96}{\sqrt{20}} \begin{pmatrix} 3/\sqrt{20} \\ 1/\sqrt{20} \\ -1/\sqrt{20} \\ -3/\sqrt{20} \end{pmatrix} + 10 \begin{pmatrix} 1/2 \\ -1/2 \\ -1/2 \\ +1/2 \end{pmatrix} = \begin{pmatrix} -9{,}4 \\ -9{,}8 \\ -0{,}2 \\ 19{,}4 \end{pmatrix} \tag{B.4}$$

a podobně s použitím poslední složky $\mathbf{P}'\mathbf{y}$

$$\mathbf{u} = \frac{8}{\sqrt{20}} \begin{pmatrix} 1/\sqrt{20} \\ -3/\sqrt{20} \\ 3/\sqrt{20} \\ -1/\sqrt{20} \end{pmatrix} = \begin{pmatrix} 0{,}4 \\ -1{,}2 \\ 1{,}2 \\ -0{,}4 \end{pmatrix}. \tag{B.5}$$

Protože sloupce matice $\mathbf{P}$ mají jednotkovou délku a v našem případě je vektor reziduí $\mathbf{u}$ násobkem jediného (posledního) sloupce matice $\mathbf{P}$, je koeficient $8/\sqrt{20}$ nutně roven odmocnině S reziduálního rozptylu S^2.

Snadno ověříme, že vektor $\hat{\mathbf{y}}$ můžeme vyjádřit jako

$$\hat{\mathbf{y}} = \begin{pmatrix} -9{,}4 \\ -9{,}8 \\ -0{,}2 \\ 19{,}4 \end{pmatrix} = \begin{pmatrix} 1 & -3 & 9 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \end{pmatrix} \begin{pmatrix} -6{,}25 \\ 4{,}80 \\ 1{,}25 \end{pmatrix},$$

takže je $\mathbf{b} = (-6{,}25, 4{,}8, 1{,}25)'$.

Místo matice $\mathbf{X}$ při vyvolání funkce `a <- lm(y~Xa)` použijeme pouze $\mathbf{X}_a$, protože absolutní člen je do modelu vkládán standardně. Kdybychom chtěli použít celou matici $\mathbf{X}$, zvolili bychom příkaz `a <- lm(y~X-1)`, abychom zabránili standardnímu přidávání absolutního členu. (Pozor, objekt `X` resp. `Xa` musí být matice!)

Výsledkem je objekt `a`, který je složen z řady položek. Jejich názvy lze získat příkazem `names(a)`:

```
> names(a)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

V položce `a$qr` je uložen zašifrovaný QR rozklad matice $\mathbf{X}$, souřadnice $\mathbf{P}'\mathbf{y}$ vektoru $\mathbf{y}$ v ortonormální bázi obsahuje `a$effects`. Vektor reziduí $\mathbf{u}$ je uložen v `a$residuals`, vektor $\hat{\mathbf{y}}$ vyrovnaných hodnot je v `a$fitted.values`. Koeficienty vyjádření $\hat{\mathbf{y}}$ pomocí sloupců matice $\mathbf{X}$ jsou v `a$coefficients`. Pokud by matice $\mathbf{X}$ neměla sloupce lineárně nezávislé (platí `a$rank`<`ncol(X)`), nebudou některé souřadnice tohoto vektoru definovány – stačí tam doplnit nuly. Matice vstupních dat $(\mathbf{y}, \mathbf{X}_a)$ je součástí objektu `a` jako `a$model`. Některé z uvedených statistik lze z objektu `a` získat použitím funkcí `coefficients(a)`, `effects(a)`, `residuals(a)` a `fitted.values(a)`. Existují i zkrácená volání, jako např. `coef()`, `resid()` nebo `fitted()`.

Použijeme-li příkaz `print(a)`, dostaneme text:

```
Call:
lm(formula = y ~ Xa)

Coefficients:
(Intercept)         Z2          Z3
     -0.625      0.500       0.125
```

V řádku `coefficients` jsou uvedeny složky vektoru **b**. Příkaz `summary(a)` vytiskne podrobnější informaci o lineárním modelu:

```
Call:
lm(formula = y ~ Xa)

Residuals:
   1    2    3    4
 0.4 -1.2  1.2 -0.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.2500     1.4318  -4.365   0.1434
Xa1           4.8000     0.4000  12.000   0.0529 .
Xa2           1.2500     0.2236   5.590   0.1127
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

Residual standard error: 1.789 on 1 degrees of freedom
Multiple R-Squared: 0.9943,     Adjusted R-squared: 0.983
F-statistic: 87.2 on 2 and 1 degrees of freedom,       p-value: 0.07532
```

V odstavci `Coefficients` je vždy vedle bodového odhadu b_j uvedena střední chyba tohoto odhadu $S\sqrt{v_{jj}}$, testová statistika T_j podle (2.20) pro test nulové hypotézy $\mathsf{H}_0 : \beta_j = 0$ a odpovídající dosažená hladina testu při oboustranné alternativě. Případná významnost testových statistik je označena běžným způsobem pomocí hvězdiček. Pod označením `Residual standard error` je statistika S, dále následují koeficient determinace R^2 a upravený koeficient determinace R^2_{adj}, o kterých bude řeč později. Později podrobněji uvedeme testy podmodelu, k nimž se vztahuje také F statistika a dosažená hladina testu.

Abychom vypsali rozklad matice **X** na součin **QR**, použijeme příkaz `a$qr`:

```
> a$qr
$qr
  X.1       X.2           X.3
1 -2.0  0.0000000 -1.000000e+01
2  0.5 -4.4721360 -8.881784e-16
3  0.5  0.4472136  8.000000e+00
4  0.5  0.8944272 -9.296181e-01
$qrattr(,"assign")
[1] 1 1 1

$qraux
[1] 1.500000 1.000000 1.368524

$pivot
[1] 1 2 3

$tol
[1] 1e-07
```

```
$rank
[1] 3
```

Zcela stejný výsledek bychom dostali pomocí funkce qr(cbind(1,Xa)) nebo qr(X). Pod označením $qr jsme dostali matici stejného rozměru jako **X**, jejíž horní trojúhelníková část obsahuje horní trojúhelník matice **R**. Zbytek matice spolu s vektorem $qraux obsahuje informaci potřebnou k rekonstrukci matice **Q**. Zjištěná hodnost matice **X** uvedena jako $rank. Tato hodnota do jisté míry (v případě špatné podmíněnosti matice **X**) závisí na volbě tolerance $tol.

Matice **Q** a **R** získáme, když na kompaktní zápis použijeme funkce qr.Q() a qr.R():

```
> qr.Q(a$qr)
     [,1]       [,2] [,3]
[1,] -0.5  0.6708204  0.5
[2,] -0.5  0.2236068 -0.5
[3,] -0.5 -0.2236068 -0.5
[4,] -0.5 -0.6708204  0.5
> qr.R(a$qr)
  X.1       X.2           X.3
1  -2  0.000000 -1.000000e+01
2   0 -4.472136 -8.881784e-16
3   0  0.000000  8.000000e+00
```

Lze si nechat spočítat celou čtvercovou ortonormální matici **P**. Stačí ve funkci qr.Q() nastavit volitelný parametr complete=T:

```
> qr.Q(qr(X),complete=T)
     [,1]       [,2] [,3]       [,4]
[1,] -0.5  0.6708204  0.5  0.2236068
[2,] -0.5  0.2236068 -0.5 -0.6708204
[3,] -0.5 -0.2236068 -0.5  0.6708204
[4,] -0.5 -0.6708204  0.5 -0.2236068
```

Vraťme se ještě k příkazu summary.lm(). Výsledkem je objekt, složený z dalších zajímavých informací:

```
> names(s<-summary(a))
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "sigma"         "df"            "r.squared"     "adj.r.squared"
 [9] "fstatistic"    "cov.unscaled"
```

Upozorňuji zejména na informace o odhadech regresních koeficientů

```
> s$coefficients
            Estimate Std. Error   t value   Pr(>|t|)
(Intercept)    -6.25  1.4317821 -4.365189 0.14336634
Xa1             4.80  0.4000000 12.000000 0.05292935
Xa2             1.25  0.2236068  5.590170 0.11269007
```

a na (odhadnutou) varianční matici těchto koeficientů:

```
> s$cov.unscaled
              (Intercept)           Xa1           Xa2
(Intercept)  6.406250e-01  1.551584e-17 -7.812500e-02
Xa1          1.551584e-17  5.000000e-02 -3.103168e-18
Xa2         -7.812500e-02 -3.103168e-18  1.562500e-02
```

B.1.2. Úloha s vahami

V oddílu 2.8 jsme ukázali, jak převedeme lineární model $\mathbf{Y} \sim \left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1}\right)$ s obecnější varianční maticí na model s varianční maticí $\sigma^2\mathbf{I}$. Procedura `lm` s parametrem `weights=w` použije QR rozklad matice $\mathbf{X}^*$. Proto dostaneme poněkud jiné bodové odhady, než v modelu bez vah

```
> summary(a.w <- lm(y~Xa,weight=w))

Call:
lm(formula = y ~ Xa, weights = w)

Residuals:
      1       2       3       4
 0.6038 -1.8113  0.9057 -0.6038

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.4858     1.1680  -4.697   0.1335
        Xa1   4.8679     0.4773  10.198   0.0622 .
        Xa2   1.1651     0.2326   5.009   0.1255
---
Signif. codes:  0 ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 2.198 on 1 degrees of freedom
Multiple R-Squared: 0.9915,     Adjusted R-squared: 0.9744
F-statistic:  58.06 on 2 and 1 degrees of freedom,       p-value: 0.0924
```

Samozřejmě, dostaneme poněkud jiný QR rozklad:

```
> qr.Q(a.w$qr)
           [,1]       [,2]       [,3]
[1,] -0.3779645 -0.7357672  0.4902222
[2,] -0.3779645 -0.3065697 -0.2896767
[3,] -0.7559289  0.2452557 -0.4456565
[4,] -0.3779645  0.5518254  0.6907676
> qr.R(a.w$qr)
       XX1       XX2       XX3
1 -2.645751 -1.133893 -8.693183
2  0.000000  4.659859 -1.471534
3  0.000000  0.000000  9.447918
```

Protože máme

$$\mathbf{X}^* = \mathbf{W}^{1/2}\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \sqrt{4} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -3 & 9 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \end{pmatrix} = \begin{pmatrix} 1 & -3 & 9 \\ 1 & -1 & 1 \\ 2 & 2 & 2 \\ 1 & 3 & 9 \end{pmatrix},$$

vyjde skutečně například normováním prvního sloupce matice $\mathbf{X}^*$ první sloupce matice $\mathbf{Q}$ jako

$$\pm\frac{1}{\sqrt{7}}\begin{pmatrix} 1 \\ 1 \\ 2 \\ 1 \end{pmatrix} = \pm\begin{pmatrix} 0{,}377964 \\ 0{,}377964 \\ 0{,}755929 \\ 0{,}377964 \end{pmatrix}.$$

Porovnáme-li nyní vektory `fitted(aw)` a `X%*%coefficients(aw))`, zjistíme, že jsou totožné:

```
> cbind(fitted(a.w),X%*%coefficients(a.w),y-residuals(a.w))
        [,1]       [,2]       [,3]
1 -9.6037736 -9.6037736 -9.6037736
2 -9.1886792 -9.1886792 -9.1886792
3  0.5471698  0.5471698  0.5471698
4 19.6037736 19.6037736 19.6037736
```

Je tedy zřejmé, že vyrovnané hodnoty odpovídají modelu s vahami, jsou vyjádřené v původním modelu, nikoliv v modelu s hvězdičkami.

B.2. Vlastní procedury

Zde uvedeme souhrnně drobné vlastní procedury, které si autor připravil sám.

B.2.1. Nagelkerkeho koeficient determinace

```
nagel.R2 <- function(a,Nagelkerke=T){
#  výpočet Nagelkerkeho zobecněného koeficientu
#  determinace
#  a   výsledek požití glm
D1 <- deviance(a)
D0 <- a$null.deviance
n <- length(residuals(a))
R2 <- 1-exp((D1-D0)/n)
```

```
R2.max <- 1-exp(-D0/n)
R2.Nagel <- R2/R2.max
R2.out <- ifelse(Nagelkerke,R2.Nagel,R2)
return(R2.out)
}
```

C. Data

Příklad C.1 (Protoconid) Sleduje se velikost plochy jistého hrbolku dolní sedmičky (Protoconid) a její celková plocha ve třech skupinách archeologických nálezů, které odpovídají různým vývojovým stádiím, zde značeným symboly r, m, s. Použitá data jsou uvedena v tabulce C.1.

○

skup.	celková plocha	plocha Metaconid	skup.	celková plocha	plocha Metaconid
r	89,66	19,97	r	82,68	19,23
r	81,11	18,57	r	86,32	19,18
r	85,19	18,60	r	83,72	18,73
r	78,81	15,69	r	91,37	19,10
r	97,23	19,92	r	85,75	21,72
r	87,19	20,02	r	92,84	21,25
r	76,53	18,88	r	81,58	17,96
r	98,51	23,81	r	80,95	17,01
r	87,35	19,51	r	77,56	21,04
r	92,83	21,78	r	89,48	19,70
r	77,33	18,55	r	93,11	17,58
r	92,15	20,83	r	91,78	21,07
r	77,92	14,98	r	86,22	19,56
m	102,87	19,23	m	87,01	18,03
m	113,85	23,70	m	119,73	27,67
m	105,15	21,02	m	117,65	24,65
m	99,77	20,90	m	104,72	25,90
m	93,53	21,15	m	90,15	18,58
s	146,09	28,22	s	125,33	28,68
s	112,32	21,66	s	98,54	19,98
s	96,26	20,22	s	120,03	23,58
s	132,51	27,36	s	104,73	24,59

Tabulka C.1: Velikosti ploch dolních sedmiček

Literatura

J. Anděl (1978). *Matematická statistika.* SNTL, Praha.

J. Anděl (1998). *Statistické metody.* MATFYZPRESS, Praha.

J. Antoch, D. Vorlíčková (1992). *Vybrané metody statistické analýzy dat.* Academia, Praha.

G. E. Box, G. S. Watson (1962). Robustness to non-normality of regression tests. *Biometrika*, 62, 93–106.

W. J. Conover, M. E. Johnson, M. M. Johnson (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23, 351–361.

J. R. Donaldson, R. B. Schnabel (1987). Computational experience with confidence regions and confidence intervals for nonlinear least squares. *Technometrics*, 29, 67–82.

J. Durbin, G. S. Watson (1971). Testing for serial correlation and least squares regression. *Biometrika*, 58, 1–19.

V. Černý (1994). Výpovědní hodnota epipaleotických souborů Taforalt a Afalou-bou-Rhumel z hlediska paleoepidemiologie. Sborník *Soudobá česká antropologie*, str. 97–102. Masarykova univerzita Brno.

F. Galton (1886a). Family likeness in stature. *Proc. Roy. Soc.*, XL, 42.

F. Galton (1886b). Regression towards mediocrity in hereditary stature. *Journ. Anthrop. Inst.*, XV, 246.

W. P. Gardiner (1997). *Statistics for Biosciences.* Prentice Hall.

G. J. Hahn, S. S. Shapiro (1967). *Statistical Models in Engineering.* Wiley, New York. Existuje ruský překlad.

P. Hajná (1995). Vliv biosociálních faktorů na délku kojení a závislost vybraných antropometrických charakteristik na způsob výživy dítěte v prvních šesti měsících života. Diplomová práce, PřF UK, Praha.

A. C. Harvey, P. Collier (1977). Testing for functional misspecification in regression analysis. *Journal of the Econometrics*, 6, 103–119.

W. W. Hauck, Allan Donner (1977). Wald's test as applied to hypotheses in logit regression. *Journal of the American Statistical Society*, 72, 851–853.

T. Havránek (1993). *Statistika pro biologické a lékařské vědy.* Academia, Praha.

W. W. Howells (1996). Howells' craniometric data on the internet. *American Journal of Physical Anthropology*, str. 441–442.

M. Jílek (1988). *Toleranční meze.* SNTL, Praha.

D. G. Kleinbaum (1994). *Logistic regression: a self-learning text.* Springer, New York.

M. Kulich (2000). Zobecněné lineární modely. Poznámky k semestrální přednášce 1999/2000.

J. Likeš, J. Laga (1978). *Základní statistcké tabulky.* SNTL, Praha.

P. McCullagh, J. A Nelder (1989). *Generalized Linear Models.* Chapman Hall.

N. J. D. Nagelkerke (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691–692.

J. Netter, W. Wasserman, M. H. Kutner (1985). *Applied linear statistical models.* Irwin, Homewood, Illinois.

V. Petráčková, J. Kraus, kol. (1995). *Akademický slovník cizích slov II.* Academia, Praha.

C. R. Rao (1978). *Lineární metody statistické indukce a jejich aplikace.* Academia, Praha.

M. Štefek (1994). Porušení předpokladu o normálním rozdělení v lineárním modelu. Diplomová práce, MFF UK, Praha.

W. N. Venables, B. D. Ripley (1997). *Modern applied statistics with S-PLUS.* Springer, New York, second edition.

K. Zvára (1979). On exact confidence regions for linear regression functions. *Math. Operationsforsch. Statist., Ser. Statistics*, 10, 55–62.

K. Zvára (1989). *Regresní analýza.* Academia, Praha.

K. Zvára (1998). *Biostatistika.* Karolinum, Praha.

Rejstřík

$COVRATIO$, 86
$DFBETA$, 85
$DFBETAS$, 85
$DFFITS$, 85
$DFIT$, 85
SSA, 48
SSE, 27, 47
SSR, 27
SST, 27, 47
VIF, 115
`contr.helmert`, 50
`contr.poly`, 55
`contr.sum`, 50
`contr.treatment`, 51
`ordered`, 55

bloky
  náhodné, 73

case-control, 152
Choleského faktor, 185
cohort, 150
confounding, 131

deviance, 145
diagram
  profilový, 174

efekt, 14
  hlavní, 57
  náhodný, 73
  pevný, 73

faktor, 14
  uspořádaný, 55
faktor Choleského, 185
funkce
  linková, 158
  odhadnutelná, 13
  regresní, 169
  spojovací, 138, 158
    kanonická, 158

GLM, 138

heteroskedasticita, 97
homoskedasticita, 97

identifikace, 45
index
  podmíněnosti, 112
interakce, 56, 130
interval
  konfidenční, 33
  predikční, 33
  spolehlivosti, 33

kalibrace, 39
koeficient
  determinace, 27
    adjustovaný, 124
  korelační
    výběrový, 26
  regresní
    standardizovaný, 114
kontrast, 14, 49
  ortogonální, 50
kritérim
  silné, 121
kritérium
  slabé, 122
Kroneckerův součin, 187

leverage, 84
link function, 138
logit, 138

Malowsovo C_p, 125
matice
    Helmertova, 50
    informační, 160
        Fisherova, 160, 189
    odmocninová, 185
    pseudoinverzní, 183
        Mooreova-Penroseho, 183
metoda
    Fiellerova, 41
    Lagrangeova multiplikátoru, 143, 190
    maximální věrohodnosti, 188
    skórů, 143, 190
model
    kvadraticky vyvážený, 76
    lineární
        zobecněný, 138, 158
    nulový, 146
    odlehlého pozorování, 79
    saturovaný, 145, 162
    standardizovaný, 114
    vynechaného pozorování, 79
    vyvážený, 50, 52
multikolinearita, 111

nerovnost
    Bonferroniho, 83, 188

odds, 138
odds ratio, 140
ošetření, 14

Pearsonovo X^2, 163
podmodel, 21
poměr šancí, 140
pozorování
    odlehlé, 83
proměnná
    nezávisle, 94
prostor
    regresní, 10
    reziduální, 10
pás spolehivosti
    kolem regresní funkce, 33
pás spolehlivosti
    kolem regresní přímky, 33
    pro regresní funkci, 34
    přesný, 34
příklad
    adjustace, 73
    analýza kovariance, 14
    brzdná dráha, 94, 102, 105, 106
    DRIS, 27
    dva regresory, 69
    dvojné třídění, 78
    hmotnost hochů, 7, 29
    ICHS, 58
    jednoduché třídění, 14, 46
    kojení, 141, 144, 147, 148
    kořeny, 48, 52, 54, 55, 95, 98, 99, 107
    listy, 36, 40
    měď, 44, 46
    měření IQ, 116
    náhodné bloky, 73
    procento tuku, 88, 134
    Protoconid, 60, 199
    zachovalost kostí, 166
    procento tuku, 128

QR rozklad, 184, 191

regresor, 93
rezidua, 11
    deviančni, 153, 163
    jackkniffe, 82
    nekorelovaná, 89
    normovaná, 81, 88
    Pearsonova, 153, 163
    rekurzivní, 89
    studentizovaná, 82, 88
reziduální rozptyl, 11
reziduální součet čtverců, 11
rovnice
    reparmetrizační, 45
rozdělení
    alternativní, 165

Bernoulliovo, 165
rozklad
podle singulárních hodnot, 44, 184
QR, 184, 191
spektrální, 184
typu I, 57
typu II, 58
typu III, 58
rozptyl
reziduální, 11

skóry, 190
součet čtverců
reziduální, 11
součin
Kroneckerův, 187
součin Kroneckerův, 187
srovnání
mnohonásobná, 83
studie
prospektivní, 150
průřezová, 151
případů a kontrol, 150
retrospektivní, 152
study
cross-sectional, 151

tabulka
analýzy rozptylu, 48
test
Bartlettův, 97
Durbinův-Watsonův, 96, 108
Flignerův-Killeenův, 98
Goldfeldův-Quandtův, 100
Kolmogorovův-Smirnovův, 106
Leveneův, 99
Lillieforsův, 106
poměrem věrohodnosti, 189
Ryanův-Joinerův, 106
Waldův, 141, 142, 189
Wilksův, 189
tolerance, 115
transformace
Boxova-Coxova, 133

vektor
skórů, 190
vektor reziduí, 11
vzdálenost
Cookova, 86, 88

šance, 138

číslo
podmíněnosti, 112