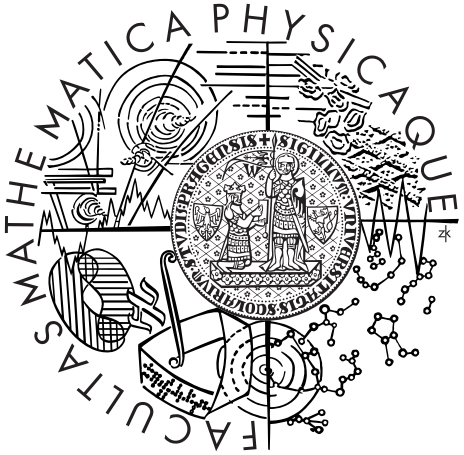
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Jan Vávra

Testy homoskedasticity v lineárním modelu

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	doc. RNDr. Arnošt Komárek, Ph.D.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2016

Poděkování

Rád bych na tomto místě věnoval poděkování vedoucímu bakalářské práce doc. RNDr. Arnoštu Komárkovi, Ph.D. za to, že si vždy našel chvíli, aby mi poskytnul nesčetné množství odborných rad a připomínek, které dopomohly k finální podobě této práce.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Testy homoskedasticity v lineárním modelu

Autor: Jan Vávra

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Arnošt Komárek, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá testováním předpokladu homoskedasticity v lineárním modelu, neboli předpokladu o konstantním rozptylu chyb tohoto modelu. Takových testů existuje celá řada, ale ne všechny se dají aplikovat na konkrétním modelu a ne všechny dosahují uspokojivých výsledků za různých okolností. Práce se zaměří na testy, které lze odvodit na základě asymptotické teorie maximální věrohodnosti, zvláště pak teorie testů s rušivými parametry. Odvozeny jsou dva základní testy, první v situaci modelu analýzy rozptylu jednoduchého třídění a druhý v situaci, kdy je připuštěna závislost rozptylu na doprovodných veličinách. V následných numerických studiích jsou prověřeny vlastnosti odvozených testových statistik.

Klíčová slova: homoskedasticita, heteroskedastický lineární model

Title: Homoscedasticity Tests in a Linear Model

Author: Jan Vávra

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Arnošt Komárek, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This thesis deals with testing the assumption of homoscedasticity in linear model, that is the assumption of constant variance of this model. There is plenty of such tests, but not all of them can be applied to specific model and not all of them reach satisfactory results under various circumstances. Thesis focuses on tests which can be derived on the basis of the asymptotic theory for maximum likelihood estimation, particularly the test theory with nuisance parameters. There are derived two basic tests, the first one in the situation of analysis of variance model and the second one in the situation when we allow the dependence of variance to concomitant quantities. In subsequent numerical studies there are examined characteristics of derived test statistics.

Keywords: homoscedasticity, heteroscedastic linear model

Obsah

Používané značení

V této práci se budeme držet následujících konvencí:

- ▷ Malými římskými a řeckými písmeny budeme označovat jednorozměrné konstanty, parametry či funkce. Písmena i, j, k, l, m, n budou označovat přirozená čísla.
- ▷ Velkými římskými písmeny budeme (až na výjimky - ε) označovat náhodné veličiny (X, Y, Z), jejich realizace potom příslušnými malými písmeny (x, y, z). Dále také v několika případech použijeme velkých římských a řeckých písmen pro označení množin (M, Ω, Θ, ...).
- ▷ Tučnými písmeny budeme značit vektory. Každý vektor bude chápán jako sloupcový. Pro náhodné vektory budeme (až na výjimky - $\boldsymbol{u}$) užívat velká římská písmena ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$) a jejich realizace příslušnými malými písmeny ($\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$).
- ▷ Pro matice budeme používat zdvojených velkých římských písmen ($\mathbb{A}$, $\mathbb{X}$) nebo $\boldsymbol{\Sigma}$. V textu budeme používat značení $\mathbb{X}$ jak pro náhodnou matici, tak i její realizaci, je tedy třeba význam symbolu určit z kontextu. Symboly $\mathbb{N}$ a $\mathbb{R}$ máme vyhrazeny pro označení množiny přirozených a reálných čísel.
- ▷ Pro míry máme vyhrazený speciální font řeckých písmen (μ, ν, λ).
- ▷ Bezpatkové písmo je používáno pro značení pravděpodobnostních rozdělení (N, Mult, LM, ...) nebo určitých funkcionálů či operátorů (E, var, Tr, ...).
- ▷ Strojové písmo budeme používat pro pojmy spojené s výpočetním softwarem (`bartlett.test`, `bptest`, ...).
- ▷ Použití stříšky či vlnky nad písmenem značí, že se jedná o odhad tohoto parametru, často je následován indexem n označující rozsah náhodného výběru. Například $\widetilde{\boldsymbol{\beta}}_n$, $\widehat{\sigma}^2_n$, ...

Speciální symboly:

$(\Omega, \mathcal{A}, \mathsf{P})$	pravděpodobnostní prostor s množinou jevů Ω, σ-algebrou $\mathcal{A}$ a pravděpodobnostní mírou P
s.j.	skoro jistě vzhledem k pravděpodobnostní míře P
$[\mu]$ s.v.	skoro všude/všechna vzhledem k míře μ
λ, λ^n	Lebesgueova (n-rozměrná) míra
$\mathcal{B}_0$	borelovská σ-algebra na $\mathbb{R}$

$Y\|\boldsymbol{X}$	podmíněné rozdělení náhodné veličiny Y při $\boldsymbol{X}$
$\mathsf{E}[Y\|\boldsymbol{X}]$	podmíněná střední hodnota náhodné veličiny Y při $\boldsymbol{X}$
$\mathsf{var}[Y\|\boldsymbol{X}]$	podmíněný rozptyl náhodné veličiny Y při $\boldsymbol{X}$
$\mathsf{diag}(\ldots)$	diagonální matice s danými prvky na diagonále
$\mathsf{Tr}(\mathbb{A})$	stopa matice $\mathbb{A}$
$\mathsf{det}(\mathbb{A})$	determinant matice $\mathbb{A}$
$\boldsymbol{x}^\top, \mathbb{X}^\top$	operátor transpozice vektoru $\boldsymbol{x}$ a matice $\mathbb{X}$
$\mathbb{A}^{-1}$	inverzní matice k matici $\mathbb{A}$
$\|\boldsymbol{x}\|$	eukleidovská norma vektoru $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$, $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}$
$\mathbb{1}_n$	n-složkový sloupcový vektor, jehož prvky jsou jen 1
$\mathbb{I}_n$	jednotková matice řádu n, neboli $\mathsf{diag}(1, \ldots, 1)$
$\mathbf{0}^n$, $\mathbf{0}^{n\times 1}$	n-složkový sloupcový nulový vektor
$\mathbf{0}^{1\times n}$	n-složkový řádkový nulový vektor
$\mathbb{O}^{n\times m}$	nulová matice řádu $n \times m$
δ_{ij}	Kroneckerovo delta, $\delta_{ij} = 1$ jedině tehdy, když $i = j$, jinak $\delta_{ij} = 0$
ℓ	logaritmická věrohodnost
$\mathcal{F}_m$	systém hustot generovaný parametrickým prostorem o dimenzi m
H_0, H_1	nulová hypotéza a alternativa
$\xrightarrow{\mathsf{D}}$	konvergence v distribuci
χ^2_q	chí kvadrát rozdělení o $q \in \mathbb{N}$ stupních volnosti
$\chi^2_q(\alpha)$	α-kvantil rozdělení chí kvadrát o q stupních volnosti
R	symbol pro označení statistického výpočetního softwaru **R**

Úvod

Úkolem této práce bude nejdříve představit klasický lineární model, kde se předpokládá konstantní rozptyl chyb, a následně ho zobecnit na heteroskedastický lineární model, který už bude obecně předpokládat proměnlivost tohoto rozptylu. Jednodušší definice lineárního modelu předpokládá, že uvažované regresory jsou předem známé konstanty. V tomto textu ovšem budeme předpokládat, že regresory budou obecně náhodné vektory. Toto zobecnění nás tedy nutí pracovat s podmíněným rozdělením, podmíněnou střední hodnotou a podmíněným rozptylem. Dále jelikož teorie maximální věrohodnosti a z ní vycházející teorie testů s rušivými parametry tvoří základy, ze kterých budeme při odvozování testů vycházet, shrneme ve druhé kapitole stěžejní definice a poznatky těchto přístupů, na které se budeme odvolávat. Ve třetí kapitole si představíme model analýzy rozptylu jednoduchého třídění, kde budeme obecně předpokládat nejen rozdílné střední hodnoty, ale i rozptyly v jednotlivých skupinách. Tento model si zobecníme do pojmů definovaných v první kapitole a následně podrobně odvodíme test homoskedasticity poměrem věrohodnosti. Seznámíme se také s velmi podobnou statistikou, kterou již v roce 1937 navrhl anglický statistik Maurice Stevenson Bartlett a dáme ji do souvislosti s odvozeným testem poměrem věrohodností. Ve čtvrté kapitole se podíváme na heteroskedastický lineární model, kde připouštíme závislost rozptylu na doprovodných regresorech. Důkladně v tomto modelu odvodíme takzvaný skórový test, který navrhli v roce 1979 australští statistici Trevor Breusch a Adrian Pagan. V závěrečné kapitole za pomoci výpočetního prostředí **R** (R Core Team, 2016)prostudujeme vlastnosti odvozených testových statistik. Zejména nás bude zajímat, zda dodržují předepsanou hladinu a jak přesná je použitá asymptotika při malém rozsahu výběru. Dále se zaměříme na studium síly těchto testů a chování příslušných statistik při nesplnění předpokladu normality.

Kapitola 1

Lineární model

V této kapitole se seznámíme s klasickým a heteroskedastickým lineárním modelem a základními pojmy, se kterými budeme dále pracovat. Uvedeme si také základní tvrzení a věty, které z těchto definic vyplývají.

1.1 Klasický lineární model

Nejprve si zavedeme klasický lineární model. Uvažme reálný náhodný vektor $(Y, \boldsymbol{X})$ na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathsf{P})$ se sdruženou hustotou $f(y, \boldsymbol{x})$ vzhledem k σ-konečné míře $\nu = \nu_y \times \nu_{\boldsymbol{x}}$, kde Y je náhodná veličina a $\boldsymbol{X} = (X_0, X_1, \ldots, X_k)^\top$ je náhodný vektor dimenze $k+1$ pro $k \in \mathbb{N}_0$. Dále budeme předpokládat, že ν_y je Lebesgueova míra λ na $(\mathbb{R}, \mathcal{B}_0)$.

Definice 1. *Řekneme, že $(Y, \boldsymbol{X})$ se řídí lineárním modelem, pokud*

$$\begin{aligned} \mathsf{E}[Y|\boldsymbol{X}] &= \boldsymbol{X}^\top \boldsymbol{\beta} \quad \text{s.j.}, \\ \mathsf{var}[Y|\boldsymbol{X}] &= \sigma^2 \qquad \text{s.j.}, \end{aligned} \tag{1.1}$$

kde $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ a $\sigma^2 > 0$ jsou neznámé parametry, a pokud rozdělení náhodného vektoru $\boldsymbol{X}$ na těchto parametrech nezávisí.
Nadále budeme psát $(Y, \boldsymbol{X}) \sim \mathsf{LM}\left(\boldsymbol{X}^\top \boldsymbol{\beta}, \sigma^2\right)$.

Poznámka. Několik doplňujících poznámek k definici 1.

a) Náhodné veličině Y se říká *vysvětlovaná proměnná* (nebo také závisle proměnná) a náhodnému vektoru $\boldsymbol{X}$ říkáme *regresory* (nebo také nezávisle proměnné). Dále říkáme, že se jedná o *lineární* model, protože podmíněná střední hodnota Y při $\boldsymbol{X}$ závisí na $\boldsymbol{\beta}$ lineárně.

b) Častým předpokladem v lineárních modelech bývá, že náhodná veličina X_0 má degenerované (nenáhodné) rozdělení $X_0 = 1$ skoro jistě. Náhodný vektor $\boldsymbol{X}$ jsme tedy zavedli jako $(k+1)$-složkový, abychom tento případ mohli odlišit ve zvláštní (nulté) složce.

c) V textu budeme dále pracovat s podmíněným rozdělením $Y|\boldsymbol{X}$, jehož hustotu budeme značit $f_{Y|\boldsymbol{X}}$. Tato hustota závisí na parametrech $\boldsymbol{\beta}$ a σ^2. Označme $f_{\boldsymbol{X}}$ hustotu náhodného vektoru $\boldsymbol{X}$, tato už na $\boldsymbol{\beta}$ a σ^2 nezávisí. Ze známých vlastností hustot z teorie pravděpodobnosti platí

$$f(y, \boldsymbol{x}; \boldsymbol{\beta}, \sigma^2) = f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}; \boldsymbol{\beta}, \sigma^2) f_{\boldsymbol{X}}(\boldsymbol{x}) \quad [\nu] \text{ s.v.} \tag{1.2}$$

Definice 2. *V lineárním modelu* $(Y, \boldsymbol{X}) \sim \mathsf{LM}\left(\boldsymbol{X}^\top\boldsymbol{\beta}, \sigma^2\right)$ *zavedeme náhodnou veličinu* ε *jako*

$$\varepsilon := Y - \boldsymbol{X}^\top\boldsymbol{\beta}$$

a budeme ji nazývat chybový člen modelu.

Poznámka. ε z definice 2 odpovídá náhodné veličině popisující odchylku (či chybu - „error“) vysvětlované proměnné od její podmíněné střední hodnoty. Jakožto náhodnou veličinu bychom ji tedy měli dle naší konvence značit velkým římským písmenem. Ovšem použití písmena E by mohlo snadno vést k záměně s operátorem střední hodnoty E, proto si zde dovolíme od naší konvence upustit a značit tuto veličinu právě písmenem ε.

Lemma 1. *Nechť* $(Y, \boldsymbol{X}) \sim \mathsf{LM}\left(\boldsymbol{X}^\top\boldsymbol{\beta}, \sigma^2\right)$*, potom pro* ε *z definice 2 platí*

$$\mathsf{E}\,\varepsilon = 0, \qquad \mathsf{var}\,\varepsilon = \sigma^2.$$

Důkaz. Oba vzorce dostaneme použitím základních vlastností podmíněné střední hodnoty a vzorce (1.1) z definice 1 lineárního modelu.

Pro střední hodnotu náhodné veličiny ε platí

$$\mathsf{E}\,\varepsilon = \mathsf{E}\left(\mathsf{E}\left[\varepsilon|\boldsymbol{X}\right]\right) = \mathsf{E}\left(\mathsf{E}\left[Y|\boldsymbol{X}\right] - \mathsf{E}\left[\boldsymbol{X}^\top\boldsymbol{\beta}\middle|\,\boldsymbol{X}\right]\right) = \mathsf{E}\left(\boldsymbol{X}^\top\boldsymbol{\beta} - \boldsymbol{X}^\top\boldsymbol{\beta}\right) = 0.$$

Pro rozptyl ε si nejprve připomeneme vzorec (1.3) pro $\boldsymbol{U}$,$\boldsymbol{V}$ obecné náhodné vektory. Platí, že

$$\mathsf{var}\,\boldsymbol{U} = \mathsf{E}\left(\mathsf{var}\left[\boldsymbol{U}|\boldsymbol{V}\right]\right) + \mathsf{var}\left(\mathsf{E}\left[\boldsymbol{U}|\boldsymbol{V}\right]\right). \tag{1.3}$$

Tento vzorec aplikujeme na $U = \varepsilon$ a $\boldsymbol{V} = \boldsymbol{X}$. Už v minulém výpočtu jsme došli ke zjištění, že

$$\mathsf{E}\left[\varepsilon|\boldsymbol{X}\right] = 0 \quad \text{s.j.},$$

proto druhý člen ve vzorci (1.3) bude také nulový. Zaměřme se proto na výpočet $\mathsf{var}\left[\varepsilon|\boldsymbol{X}\right]$, pro který platí

$$\begin{aligned}
\mathsf{var}\left[\varepsilon|\boldsymbol{X}\right] &= \mathsf{var}\left[Y - \boldsymbol{X}^\top\boldsymbol{\beta}\middle|\,\boldsymbol{X}\right] \\
&= \mathsf{E}\left[\left(Y - \boldsymbol{X}^\top\boldsymbol{\beta} - \mathsf{E}\left[Y - \boldsymbol{X}^\top\boldsymbol{\beta}\middle|\,\boldsymbol{X}\right]\right)^2\middle|\,\boldsymbol{X}\right] \\
&= \mathsf{E}\left[\left(Y - \mathsf{E}\left[Y|\boldsymbol{X}\right]\right)^2\middle|\,\boldsymbol{X}\right] \\
&= \mathsf{var}\left[Y|\boldsymbol{X}\right] = \sigma^2 \quad \text{s.j.}
\end{aligned}$$

Odtud již plyne, že $\mathsf{var}\,\varepsilon = \mathsf{E}\left(\mathsf{var}\left[\varepsilon|\boldsymbol{X}\right]\right) = \mathsf{E}\,\sigma^2 = \sigma^2$.

□

Standardní situací bývá, že parametry $\boldsymbol{\beta}$ a σ^2 jsou neznámé a naším cílem je odhadnout tyto parametry na základě naměřených dat, které reprezentujeme náhodným výběrem z příslušného rozdělení. Dále si definujeme řadu užitečných charakteristik, které nám přiblíží chování našeho modelu.

Mějme náhodný výběr $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ z rozdělení

$$(Y, \boldsymbol{X}) \sim \mathsf{LM}\left(\boldsymbol{X}^\top\boldsymbol{\beta}, \sigma^2\right).$$

Zaveďme si náhodný vektor $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ a náhodnou matici $\mathbb{X}$ s řádky $\boldsymbol{X}_i^\top, i \in \{1, \ldots, n\}$. Pak z nezávislosti a definice 1 plyne, že

$$\begin{aligned} \mathsf{E}\left[\boldsymbol{Y}|\mathbb{X}\right] &= \mathbb{X}\boldsymbol{\beta} \quad \text{s.j.}, \\ \mathsf{var}\left[\boldsymbol{Y}|\mathbb{X}\right] &= \sigma^2\mathbb{I}_n \quad \text{s.j.}, \end{aligned}$$

kde $\mathbb{I}_n$ značí jednotkovou matici řádu n.

Definice 3. *Budeme říkat, že* $(\boldsymbol{Y}, \mathbb{X})$ *se řídí lineárním modelem, a zapisovat to budeme také jako*

$$(\boldsymbol{Y}, \mathbb{X}) \sim \mathsf{LM}\left(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbb{I}_n\right).$$

Danou realizaci náhodného výběru budeme značit $(y_i, \boldsymbol{x}_i)$ pro $i \in \{1, \ldots, n\}$ a dále $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$. Symbolem $\mathbb{X}$ budeme rozumět jak onu realizaci

$$\mathbb{X} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix},$$

tak i náhodnou matici $\mathbb{X}$. V celém textu budeme pro jednoduchost předpokládat, že $n > k + 1$ a že tato matice má skoro jistě plnou sloupcovou hodnost $k + 1$, kterou budeme značit r, tj. $r = k + 1$.

Náhodný výběr z definice 3 pak má dle (1.2) sdruženou hustotu vzhledem k součinové míře $\nu^n = \nu \times \cdots \times \nu$ tvaru

$$\begin{aligned} f_n(\boldsymbol{y}, \mathbb{X}; \boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} f(y_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f_{Y|\boldsymbol{X}}(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2) f_{\boldsymbol{X}}(\boldsymbol{x}_i) \\ &= f_{\boldsymbol{Y}|\mathbb{X}}(\boldsymbol{y}|\mathbb{X}; \boldsymbol{\beta}, \sigma^2) f_{\mathbb{X}}(\mathbb{X}) \quad [\nu^n] \text{ s.v.}, \end{aligned} \tag{1.4}$$

kde $f_{\boldsymbol{Y}|\mathbb{X}}$ značí hustotu sdruženého podmíněného rozdělení $\boldsymbol{Y}|\mathbb{X}$ a $f_{\mathbb{X}}$ značí sdruženou hustotu náhodné matice $\mathbb{X}$.

Obdobně si také zavedeme vektor chybových členů modelu $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, kde $\varepsilon_i = Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}$ pro každé $i \in \{1, \ldots, n\}$, tedy $\boldsymbol{\varepsilon} = (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})$. Potom pro něj dle důkazu lemmatu 1 platí

$$\mathsf{E}\left[\boldsymbol{\varepsilon}|\mathbb{X}\right] = 0 \quad \text{s.j.}, \quad \mathsf{var}\left[\boldsymbol{\varepsilon}|\mathbb{X}\right] = \sigma^2\mathbb{I}_n \quad \text{s.j.}$$

1.2 Metoda nejmenších čtverců

Nyní můžeme přistoupit k odhadu vektorového parametru $\boldsymbol{\beta}$. Standardní metodou pro nalezení tohoto odhadu je *metoda nejmenších čtverců*. Součtem čtverců rozumíme funkci $SS(\boldsymbol{\beta})$ při daných $(\boldsymbol{Y}, \mathbb{X})$ definovanou jako

$$SS(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}\right)^2.$$

Odhadem $\boldsymbol{b}$ vektorového parametru $\boldsymbol{\beta}$ metodou nejmenších čtverců potom rozumíme takovou hodnotu $\boldsymbol{\beta}$, která minimalizuje funkci $SS(\boldsymbol{\beta})$, to jest

$$\boldsymbol{b} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{k+1}}{\arg\min}\, SS(\boldsymbol{\beta}).$$

Za našeho předpokladu o plné sloupcové hodnosti matice $\mathbb{X}$ lze toto $\boldsymbol{b}$ jednoznačně určit jako řešení tzv. *soustavy normálních rovnic*

$$\mathbb{X}^\top \mathbb{X}\boldsymbol{b} = \mathbb{X}^\top \boldsymbol{Y} \quad \text{ve tvaru} \quad \boldsymbol{b} = \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \boldsymbol{Y}.$$

Definice 4. *V modelu* $(\boldsymbol{Y}, \mathbb{X}) \sim \mathsf{LM}\left(\mathbb{X}\boldsymbol{\beta}, \sigma^2 \mathbb{I}_n\right)$ *definujeme*

- *vektor vyrovnaných hodnot* $\hat{\boldsymbol{Y}} := \mathbb{X}\boldsymbol{b} = \mathbb{X}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \boldsymbol{Y}$,
- *vektor reziduí* $\boldsymbol{u} := \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \left(\mathbb{I}_n - \mathbb{X}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top\right) \boldsymbol{Y}$,
- *reziduální součet čtverců* $RSS := SS(\boldsymbol{b}) = \|\boldsymbol{u}\|^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$,
- *reziduální rozptyl* $S^2 := \dfrac{RSS}{n-r}$, *kde* $r = k + 1$ *je hodnost matice* $\mathbb{X}$.

Poznámka. $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$ je opět náhodný vektor, tedy bychom měli užívat pro označení reziduí velkého písmene $\boldsymbol{U}$. Zde se držíme značení zavedeného v učebnici Zvára (2008), ze které primárně vycházíme.

Poznámka. Vektor $\hat{\boldsymbol{Y}}$ je nestranným odhadem vektoru $\mathbb{X}\boldsymbol{\beta}$, neboť

$$\mathsf{E}\left[\hat{\boldsymbol{Y}} \middle| \mathbb{X}\right] = \mathbb{X}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \mathsf{E}\left[\boldsymbol{Y} \middle| \mathbb{X}\right] = \mathbb{X}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \mathbb{X}\boldsymbol{\beta} = \mathbb{X}\boldsymbol{\beta} \quad \text{s.j.}$$

Pro podmíněný rozptyl tohoto odhadu platí

$$\mathsf{var}\left[\hat{\boldsymbol{Y}} \middle| \mathbb{X}\right] = \mathbb{X}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \left(\sigma^2 \mathbb{I}_n\right) \mathbb{X}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top = \sigma^2 \mathbb{X}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \quad \text{s.j.}$$

Reziduální součet čtverců odpovídá dle této definice součtu čtverců vzdáleností odhadů $\hat{Y}_i$ od Y_i, tedy nám vypovídá o tom, jak moc dobře nám $\hat{\boldsymbol{Y}}$ aproximuje $\boldsymbol{Y}$. Jeho vztah k podmíněném rozptylu σ^2 je popsán v následujícím lemmatu.

Lemma 2. *V modelu* $\boldsymbol{Y} \sim \mathsf{LM}\left(\mathbb{X}\boldsymbol{\beta}, \sigma^2 \mathbb{I}_n\right)$, *kde daná matice* $\mathbb{X}$ *má hodnost* r, *platí*

$$\mathsf{E}\left[RSS \middle| \mathbb{X}\right] = (n-r)\sigma^2 \quad \textit{s.j.}, \qquad \mathsf{E}\left[S^2 \middle| \mathbb{X}\right] = \sigma^2 \quad \textit{s.j.}$$

Důkaz. Důkaz je analogický důkazu věty 2.2. v práci Zvára (2008), uvedeme si však několik úprav, které je v důkazu nutné provést.

Označíme-li si matice $\mathbb{H} = \mathbb{X}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top$ a $\mathbb{M} = \mathbb{I}_n - \mathbb{H}$, pak tyto matice jsou symetrické a idempotentní se stopami po řadě r a $(n-r)$ a platí

$$\hat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y}, \quad \boldsymbol{u} = \mathbb{M}\boldsymbol{Y} = \mathbb{M}\boldsymbol{\varepsilon}.$$

Mimochodem se potom dá z těchto vyjádření získat

$$\mathsf{E}\left[\boldsymbol{u} \middle| \mathbb{X}\right] = \boldsymbol{0} \quad \text{s.j.}, \quad \mathsf{var}\left[\boldsymbol{u} \middle| \mathbb{X}\right] = \sigma^2 \mathbb{M} \quad \text{s.j.}$$

Chceme získat podmíněnou střední hodnotu RSS při $\mathbb{X}$, k tomu využijeme ekvivalentních zápisů

$$RSS = \|\boldsymbol{u}\|^2 = \boldsymbol{u}^\top \boldsymbol{u} = \boldsymbol{\varepsilon}^\top \mathbb{M}^\top \mathbb{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top \mathbb{M}\boldsymbol{\varepsilon}.$$

Nyní už se jen využije vlastností funkcionálu stopy matice Tr, abychom mohli provést následující úpravy

$$\begin{aligned}\mathsf{E}\left[RSS|\mathbb{X}\right] &= \mathsf{E}\left[\boldsymbol{\varepsilon}^\top\mathbb{M}\boldsymbol{\varepsilon}|\mathbb{X}\right] = \mathsf{E}\left[\mathsf{Tr}\left(\boldsymbol{\varepsilon}^\top\mathbb{M}\boldsymbol{\varepsilon}\right)|\mathbb{X}\right] = \mathsf{Tr}\left(\mathbb{M}\,\mathsf{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top|\mathbb{X}\right]\right)\\ &= \mathsf{Tr}\left(\mathbb{M}\,\mathsf{var}\left[\boldsymbol{\varepsilon}|\mathbb{X}\right]\right) = \mathsf{Tr}\left(\mathbb{M}\sigma^2\mathbb{I}_n\right) = \sigma^2\,\mathsf{Tr}\left(\mathbb{M}\right) = (n-r)\sigma^2 \quad \text{s.j.}\end{aligned}$$

Odtud již snadno plyne i druhé tvrzení. □

Při práci s lineárními modely se často předpokládá normalita podmíněného rozdělení $Y|\boldsymbol{X}$.

Definice 5. *Řekneme, že $(Y, \boldsymbol{X})$ se řídí normálním lineárním modelem, jestliže $Y|\boldsymbol{X}$ má normální rozdělení s parametrem střední hodnoty $\boldsymbol{X}^\top\boldsymbol{\beta}$ a rozptylem σ^2. Píšeme $(Y, \boldsymbol{X}) \sim \mathsf{NLM}\left(\boldsymbol{X}^\top\boldsymbol{\beta}, \sigma^2\right)$.*

Máme-li náhodný výběr $(Y_1, \boldsymbol{X}_1)$, …, $(Y_n, \boldsymbol{X}_n)$ z takovéhoto rozdělení $(Y, \boldsymbol{X})$, pak říkáme, že se $(\boldsymbol{Y}, \mathbb{X})$ řídí normálním lineárním modelem a píšeme $(\boldsymbol{Y}, \mathbb{X}) \sim \mathsf{NLM}\left(\mathbb{X}^\top\boldsymbol{\beta}, \sigma^2\mathbb{I}_n\right)$.

Podmíněná hustota $f_{Y|\boldsymbol{X}}$ vzhledem k Lebesgueově míře λ je tedy tvaru

$$f_{Y|\boldsymbol{X}}(y|\boldsymbol{x};\boldsymbol{\beta},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\left(y-\boldsymbol{x}^\top\boldsymbol{\beta}\right)^2}{2\sigma^2}\right\}.$$

Podmíněná hustota $f_{\boldsymbol{Y}|\mathbb{X}}$ rozdělení $(\boldsymbol{Y}|\mathbb{X})$ vzhledem k Lebesgueově n-rozměrné míře λ^n je potom rovna

$$f_{\boldsymbol{Y}|\mathbb{X}}(\boldsymbol{y}|\mathbb{X};\boldsymbol{\beta},\sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2\right\}.$$

Věta 3. *V modelu $(\boldsymbol{Y}, \mathbb{X}) \sim \mathsf{NLM}\left(\mathbb{X}^\top\boldsymbol{\beta}, \sigma^2\mathbb{I}_n\right)$ platí*

$$\left(\frac{1}{\sigma^2}RSS\middle|\,\mathbb{X}\right) = \left(\frac{1}{\sigma^2}\|\boldsymbol{u}\|^2\middle|\,\mathbb{X}\right) \sim \chi^2_{n-r},$$

kde r je hodnost matice $\mathbb{X}$.

Důkaz. Lze provést analogicky jako v práci Zvára (2008, Věta 2.6.). □

1.3 Heteroskedastické modely

Doposud jsme pracovali s klasickým tzv. *homoskedastickým* modelem, kde $\mathsf{var}\left[Y|\boldsymbol{X}\right]$ je rovno skoro jistě konstantě. Tento předpoklad nám značně usnadňuje práci. Obecně je však nutno uvážit situaci, kdy toto není nutně pravda a rozptyl závisí na regresorech. Naším cílem je tedy nalézt nějaké testy, které by dokázaly určit, zda naše data jsou v souladu s předpokladem *homoskedasticity* nebo naopak tento předpoklad silně vyvracejí.

Existuje řada obecných způsobů, jak v lineárním modelu zavést heteroskedasticitu. V této práci budeme říkat, že náhodný vektor $(Y, \boldsymbol{X})$ se řídí heteroskedastickým lineárním modelem, pokud platí

$$\begin{aligned}\mathsf{E}\left[Y|\boldsymbol{X}\right] &= \boldsymbol{X}^\top\boldsymbol{\beta} \quad &\text{s.j.},\\ \mathsf{var}\left[Y|\boldsymbol{X}\right] &= \sigma^2 h(\boldsymbol{X};\boldsymbol{\tau}) \quad &\text{s.j.},\end{aligned} \tag{1.5}$$

kde $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$, $\sigma^2 > 0$ a $\boldsymbol{\tau} \in T \subset \mathbb{R}^q, q \in \mathbb{N}$ jsou neznámé parametry, h je kladná měřitelná funkce, a pokud rozdělení náhodného vektoru $\boldsymbol{X}$ na parametrech $\boldsymbol{\beta}$, σ^2 a $\boldsymbol{\tau}$ nezávisí. Obecně by se dalo předpokládat, že funkce h může navíc záviset na $\boldsymbol{\beta}$, tuto situaci však zkoumat nebudeme.

Nastane-li tato situace, tak budeme analogicky jako v klasickém lineárním modelu psát

$$(Y, \boldsymbol{X}) \sim \mathsf{LM}\left(\boldsymbol{X}^\top\boldsymbol{\beta}, \sigma^2 h(\boldsymbol{X};\boldsymbol{\tau})\right) \quad \text{anebo} \quad (Y, \boldsymbol{X}) \sim \mathsf{NLM}\left(\boldsymbol{X}^\top\boldsymbol{\beta}, \sigma^2 h(\boldsymbol{X};\boldsymbol{\tau})\right)$$

v případě, že podmíněné rozdělení $(Y|\boldsymbol{X})$ je normální. Pro hustotu vzhledem k míře ν tohoto rozdělení platí

$$f(y, \boldsymbol{x}; \boldsymbol{\tau}, \boldsymbol{\beta}, \sigma^2) = f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}; \boldsymbol{\tau}, \boldsymbol{\beta}, \sigma^2) \cdot f_{\boldsymbol{X}}(\boldsymbol{x}) \quad [\nu] \text{ s.v.}$$

Když budeme opět uvažovat náhodný výběr $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ z tohoto rozdělení, tak ponecháváme značení $\boldsymbol{Y}$ a $\mathbb{X}$ a budeme říkat, že $(\boldsymbol{Y}, \mathbb{X})$ se řídí heteroskedasticitním modelem s podmíněnou rozptylovou maticí

$$\boldsymbol{\Sigma}(\mathbb{X}) := \mathsf{var}\left[\boldsymbol{Y}|\mathbb{X}\right] = \sigma^2 \,\mathsf{diag}\left(h(\boldsymbol{X}_1;\boldsymbol{\tau}), \ldots, h(\boldsymbol{X}_n;\boldsymbol{\tau})\right) \quad \text{s.j.}$$

a psát budeme $(\boldsymbol{Y}, \mathbb{X}) \sim \mathsf{LM}\left(\mathbb{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\mathbb{X})\right)$. V případě, že podmíněné rozdělení $(\boldsymbol{Y}|\mathbb{X})$ je normální, budeme analogicky psát $(\boldsymbol{Y}, \mathbb{X}) \sim \mathsf{NLM}\left(\mathbb{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\mathbb{X})\right)$.

Vzhledem k tomu, že funkce $h(\boldsymbol{X}_i;\boldsymbol{\tau}), i \in \{1, \ldots, n\}$ jsou kladné, tak existují matice

$$\begin{aligned}
\boldsymbol{\Sigma}^{-1}(\mathbb{X}) &= \frac{1}{\sigma^2}\,\mathsf{diag}\left(\frac{1}{h(\boldsymbol{X}_1;\boldsymbol{\tau})}, \ldots, \frac{1}{h(\boldsymbol{X}_n;\boldsymbol{\tau})}\right), \\
\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbb{X}) &= \frac{1}{\sigma}\,\mathsf{diag}\left(\frac{1}{\sqrt{h(\boldsymbol{X}_1;\boldsymbol{\tau})}}, \ldots, \frac{1}{\sqrt{h(\boldsymbol{X}_n;\boldsymbol{\tau})}}\right), \\
\boldsymbol{\Sigma}^{\frac{1}{2}}(\mathbb{X}) &= \sigma\,\mathsf{diag}\left(\sqrt{h(\boldsymbol{X}_1;\boldsymbol{\tau})}, \ldots, \sqrt{h(\boldsymbol{X}_n;\boldsymbol{\tau})}\right).
\end{aligned}$$

Lemma 4. *Nechť* $(\boldsymbol{Y}, \mathbb{X}) \sim \mathsf{LM}\left(\mathbb{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\mathbb{X})\right)$. *Potom*

$$\boldsymbol{Y}^* = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbb{X})\boldsymbol{Y} \sim \mathsf{LM}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbb{X})\mathbb{X}\boldsymbol{\beta}, \mathbb{I}_n\right).$$

Důkaz. Jelikož je $\boldsymbol{Y}^* = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbb{X})\boldsymbol{Y} = \left(\frac{Y_1}{\sigma\sqrt{h(\boldsymbol{X}_1;\boldsymbol{\tau})}}, \ldots, \frac{Y_n}{\sigma\sqrt{h(\boldsymbol{X}_n;\boldsymbol{\tau})}}\right)^\top$ a pro každé $i \in \{1, \ldots, n\}$ platí

$$\begin{aligned}
\mathsf{E}\left[\left.\frac{Y_i}{\sigma\sqrt{h(\boldsymbol{X}_i;\boldsymbol{\tau})}}\right| \boldsymbol{X}_i\right] &= \frac{\mathsf{E}\left[Y_i|\,\boldsymbol{X}_i\right]}{\sigma\sqrt{h(\boldsymbol{X}_i;\boldsymbol{\tau})}} = \frac{\boldsymbol{X}_i^\top\boldsymbol{\beta}}{\sigma\sqrt{h(\boldsymbol{X}_i;\boldsymbol{\tau})}} \quad \text{s.j.}, \\
\mathsf{var}\left[\left.\frac{Y_i}{\sigma\sqrt{h(\boldsymbol{X}_i;\boldsymbol{\tau})}}\right| \boldsymbol{X}_i\right] &= \frac{\mathsf{var}\left[Y_i|\,\boldsymbol{X}_i\right]}{\sigma^2 h(\boldsymbol{X}_i;\boldsymbol{\tau})} = \frac{\sigma^2 h(\boldsymbol{X}_i;\boldsymbol{\tau})}{\sigma^2 h(\boldsymbol{X}_i;\boldsymbol{\tau})} = 1 \quad \text{s.j.},
\end{aligned}$$

tak platí tvrzení lemmatu. □

Na vektor $\boldsymbol{Y}^*$ tedy můžeme aplikovat poznatky z podkapitol 1.1 a 1.2 a pomocí tohoto lemmatu je převést do heteroskedastického modelu.

Kapitola 2

Maximální věrohodnost

Důležitou roli při odvozování celé řady testů homoskedasticity hraje teorie maximální věrohodnosti. V této kapitole si představíme základní pojmy z této teorie a následně si ukážeme, jak se aplikuje v heteroskedastickém lineárním modelu.

2.1 Metoda maximální věrohodnosti

Definice 6. *Mějme* $\mathbb{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)^\top$ *náhodnou matici, kde* $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$ *jsou nezávislé stejně rozdělené* k*-složkové náhodné vektory s hustotou* $f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta})$ *vůči* σ*-konečné míře* μ *a* $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ *je parametr. Pro pevnou realizaci* $\mathbb{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)^\top$ *nazveme funkci* $L(\boldsymbol{\theta}) := \prod_{i=1}^{n} f_{\boldsymbol{V}}(\boldsymbol{v}_i; \boldsymbol{\theta})$ *věrohodnostní funkcí.*

Definice 7. *Hodnota* $\widehat{\boldsymbol{\theta}}_n \in \Theta$*, která maximalizuje věrohodnostní funkci* $L(\boldsymbol{\theta})$ *pro danou realizaci náhodné matice* $\mathbb{V}$*, se nazývá maximálně věrohodný odhad.*

Při hledání maximálně věrohodného odhadu se využívá toho, že libovolná ryze rostoucí transformace věrohodnostní funkce je maximalizována ve stejném bodě. Nejčastěji se využívá logaritmická transformace. Potom při dané realizaci $\mathbb{V}$ nazveme *logaritmickou věrohodností* funkci

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}).$$

Tato funkce se dá nadále chápat jako náhodná veličina, uvažujeme-li ji také jako funkci obecné náhodné matice $\mathbb{V}$.

Existují-li derivace věrohodnostní funkce podle všech složek parametru $\boldsymbol{\theta}$ pro každé $\boldsymbol{\theta} \in \Theta$, tak se za dalších předpokladů (které nám zajistí existenci a maximalitu řešení) dá $\widehat{\boldsymbol{\theta}}_n$ nalézt jako řešení takzvaného *systému věrohodnostních rovnic*:

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j} = 0 \qquad \forall j \in \{1, \ldots, m\}.$$

Definice 8. *Nechť* $\boldsymbol{V}$ *je* k*-složkový náhodný vektor s hustotou* $f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta})$ *vzhledem k* σ*-konečné míře* μ *a* $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$*. A nechť dále platí podmínky regularity:*

(R1) Θ *je neprázdná otevřená množina.*

(R2) $M = \{\boldsymbol{v} \in \mathbb{R}^k : f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta}) > 0\}$ *nezávisí na parametru* $\boldsymbol{\theta}$*.*

(R3) Pro $[\mu]$ *skoro všechna* $\boldsymbol{v} \in M$ *a pro každé* $j \in \{1, \ldots, m\}$ *existují*

$$f'_j(\boldsymbol{v}; \boldsymbol{\theta}) := \frac{\partial f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta})}{\partial \theta_j}.$$

(R4) Pro všechna $\boldsymbol{\theta} \in \Theta$ *a všechna* $j \in \{1, \ldots, m\}$ *platí* $\int\limits_M f'_j(\boldsymbol{v}; \boldsymbol{\theta}) \mathrm{d}\mu(\boldsymbol{v}) = 0.$

(R5) Pro všechny dvojice $i, j \in \{1, \ldots, m\}$ *existuje konečný integrál*

$$J_{ij}(\boldsymbol{\theta}) := \int\limits_M \frac{f'_i(\boldsymbol{v}; \boldsymbol{\theta})}{f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta})} \frac{f'_j(\boldsymbol{v}; \boldsymbol{\theta})}{f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta})} f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta}) \mathrm{d}\mu(\boldsymbol{v}).$$

(R6) Matice $\mathbb{J}(\boldsymbol{\theta}) = (J_{ij}(\boldsymbol{\theta}))_{i,j=1}^m$ *je pozitivně definitní pro každé* $\boldsymbol{\theta} \in \Theta$.

Potom se systém hustot $\mathcal{F}_m = \{f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ *nazývá regulární.*
Matice $\mathbb{J}(\boldsymbol{\theta})$ *se nazývá Fisherova informační matice pro náhodný vektor* $\boldsymbol{V}$.

Věta 5. *Nechť systém hustot* $\mathcal{F}_m = \{f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ *je regulární. Předpokládejme, že pro* $[\mu]$ *skoro všechna* $\boldsymbol{v} \in M$ *existují derivace*

$$f''_{ij}(\boldsymbol{v}; \boldsymbol{\theta}) = \frac{\partial^2 f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}, \qquad i, j \in \{1, \ldots, m\},$$

a že pro všechna $\boldsymbol{\theta} \in \Theta$ *platí*

$$\int\limits_M f''_{ij}(\boldsymbol{v}; \boldsymbol{\theta}) \mathrm{d}\mu(\boldsymbol{v}) = 0, \qquad i, j \in \{1, \ldots, m\}.$$

Pak jednotlivé prvky Fisherovy informační matice lze počítat jako

$$J_{ij}(\boldsymbol{\theta}) = -\int\limits_M \frac{\partial^2 \log f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta}) \mathrm{d}\mu(\boldsymbol{v}),$$

kde $i, j \in \{1, \ldots, m\}$.

Důkaz. Důkaz je analogický důkazu verze věty s jednorozměrným parametrem (viz Anděl, 2007, Věta 7.19). □

Máme-li náhodný výběr $\mathbb{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)^\top$, pak Fisherova informační matice náhodné matice $\mathbb{V}$, kterou označíme $\mathbb{J}^{(n)}(\boldsymbol{\theta})$, se dá zřejmě počítat jako

$$\mathbb{J}^{(n)}(\boldsymbol{\theta}) = n\mathbb{J}(\boldsymbol{\theta}).$$

Podle věty 5 lze potom její jednotlivé prvky počítat jako

$$J_{ij}^{(n)}(\boldsymbol{\theta}) = \mathsf{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right], \quad i, j \in \{1, \ldots, m\}.$$

Příklad (Maximální věrohodnost v lineárním modelu). Nyní si ukážeme, jak aplikovat maximální věrohodnost v situaci heteroskedastického lineárního modelu. Nechť tedy $(Y, \boldsymbol{X}) \sim \mathsf{LM}\left(\boldsymbol{X}^\top\boldsymbol{\beta}, \sigma^2 h(\boldsymbol{X};\boldsymbol{\tau})\right)$, viz (1.5).

Označme tedy $\boldsymbol{\theta} = (\tau_1, \ldots, \tau_q, \beta_0, \ldots, \beta_k, \sigma^2)^\top$ a $\Theta = T \times \mathbb{R}^{k+1} \times (0, \infty)$. Hustota náhodného vektoru vzhledem k σ-konečné míře $\boldsymbol{\nu}$ má tvar

$$f(y, \boldsymbol{x}; \boldsymbol{\theta}) = f_{Y|\boldsymbol{X}}(y|\boldsymbol{x};\boldsymbol{\theta}) \cdot f_{\boldsymbol{X}}(\boldsymbol{x}) \quad [\boldsymbol{\nu}] \text{ s.v.}$$

Pak pokud systém hustot

$$\mathcal{F}_{q+k+2} = \{f(y, \boldsymbol{x}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$$

je regulární, tak funkce

$$\log f(y, \boldsymbol{x}; \boldsymbol{\theta}) = \log f_{Y|\boldsymbol{X}}(y|\boldsymbol{x};\boldsymbol{\theta}) + \log f_{\boldsymbol{X}}(\boldsymbol{x}) \quad [\boldsymbol{\nu}] \text{ s.v.},$$

a její derivace podle libovolného $\theta_j, j \in \{1, \ldots, q+k+2\}$ je rovna

$$\frac{\partial \log f(y, \boldsymbol{x}, \boldsymbol{\theta})}{\partial \theta_j} = \frac{\partial \log f_{Y|\boldsymbol{X}}(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_j} \quad [\boldsymbol{\nu}] \text{ s.v.},$$

neboť funkce $\log f_{\boldsymbol{X}}(\boldsymbol{x})$ na parametru $\boldsymbol{\theta}$ vůbec nezávisí.

Mějme náhodný výběr $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ z rozdělení

$$(Y, \boldsymbol{X}) \sim \mathsf{LM}\left(\boldsymbol{X}^\top\boldsymbol{\beta}, \sigma^2 h(\boldsymbol{X};\boldsymbol{\tau})\right).$$

Potom logaritmickou věrohodností je dle (1.4) funkce

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \log f_n(\boldsymbol{Y}, \mathbb{X}; \boldsymbol{\theta}) = \log f_{\boldsymbol{Y}|\mathbb{X}}(\boldsymbol{Y}|\mathbb{X};\boldsymbol{\theta}) + \log f_{\mathbb{X}}(\mathbb{X}) \\ &= \sum_{i=1}^n \log f_{Y|\boldsymbol{X}}(Y_i|\boldsymbol{X}_i;\boldsymbol{\theta}) + \sum_{i=1}^n \log f_{\boldsymbol{X}}(\boldsymbol{X}_i) \quad [\boldsymbol{\nu}^n] \text{ s.v.}\end{aligned}$$

Pro její derivaci dle proměnné $\theta_j, j \in \{1, \ldots, q+k+2\}$ platí

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j} = \sum_{i=1}^n \frac{\partial \log f_{Y|\boldsymbol{X}}(Y_i|\boldsymbol{X}_i;\boldsymbol{\theta})}{\partial \theta_j} \quad [\boldsymbol{\nu}^n] \text{ s.v.}$$

Tedy maximálně věrohodný odhad $\widehat{\boldsymbol{\theta}}_n$ parametru $\boldsymbol{\theta}$ lze najít jen ze znalosti podmíněného rozdělení $Y|\boldsymbol{X}$ jako řešení systému věrohodnostních rovnic

$$\sum_{i=1}^n \frac{\partial \log f_{Y|\boldsymbol{X}}(Y_i|\boldsymbol{X}_i;\boldsymbol{\theta})}{\partial \theta_j} = 0, \quad j \in \{1, \ldots, q+k+2\}.$$

Dále pokud jsou navíc splněny předpoklady věty 5, tak k výpočtu Fisherovy matice stačí pracovat jen s hustotou podmíněného rozdělení. Potom Fisherova informační matice $\mathbb{J}^{(n)}(\boldsymbol{\theta})$ pro náhodnou matici $(\boldsymbol{Y}, \mathbb{X})$ se skládá z prvků

$$J_{ij}^{(n)}(\boldsymbol{\theta}) = -\sum_{l=1}^n \mathsf{E}\left(\frac{\partial^2 \log f_{Y|\boldsymbol{X}}(Y_l|\boldsymbol{X}_l;\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right), \quad i, j \in \{1, \ldots, q+k+2\}.$$

V této práci budeme dále předpokládat, že ono podmíněné rozdělení je normální, které zřejmě předpoklady regularity (R1)-(R6) a věty 5 splňuje. Jediné, co budeme při naší parametrizaci muset dodržet navíc, jsou předpoklady o parametrickém prostoru Θ.

2.2 Testy s rušivými parametry

V modelu $(Y, \boldsymbol{X}) \sim \mathsf{LM}\left(\boldsymbol{X}^\top\boldsymbol{\beta}, \sigma^2 h(\boldsymbol{X};\boldsymbol{\tau})\right)$ budeme vzhledem k naší parametrizaci (1.5) chtít testovat, zda je funkce h identicky rovna jedné, tj. $h \equiv 1$, neboť potom bude platit homoskedastický model $(Y, \boldsymbol{X}) \sim \mathsf{LM}\left(\boldsymbol{X}^\top\boldsymbol{\beta}, \sigma^2\right)$. Hypotéza tvaru $h \equiv 1$ potom bude odpovídat specifické hodnotě parametru $\boldsymbol{\tau}$, popř. $\boldsymbol{\beta}$, pokud bychom uvažovali závislost na střední hodnotě.

Vyvstává nám zde ovšem problém. Chceme testovat hypotézu o parametru $\boldsymbol{\tau}$, ale v naší parametrizaci ještě vystupují další neznámé parametry $\boldsymbol{\beta}$ a σ^2. Návod, jak testovat hypotézy jen o části neznámých parametrů, poskytuje právě teorie *testů s rušivými parametry*.

Uvažme opět obecně náhodný vektor $\boldsymbol{V}$ s hustotou $f_{\boldsymbol{V}}(\boldsymbol{v};\boldsymbol{\theta})$ vůči σ-konečné míře μ a $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top$ je m-rozměrný parametr, kde tentokrát $m \geq 2$. Tento parametr si rozdělme na dvě části. Buď $1 \leq q < m$ a označme

$$\boldsymbol{\tau} = (\theta_1, \ldots, \theta_q)^\top, \qquad \boldsymbol{\psi} = (\theta_{q+1}, \ldots, \theta_m)^\top.$$

Parametr $\boldsymbol{\tau}$ bude náš *cílový parametr*, o kterém budeme chtít provádět test hypotézy

$$H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0 \quad \text{proti alternativě } H_1 : \boldsymbol{\tau} \neq \boldsymbol{\tau}_0.$$

Parametr $\boldsymbol{\psi}$ je potřebný jen k plnému popisu modelu, ale o něm žádnou hypotézu testovat nehodláme, nazveme jej tedy *rušivým parametrem*.

Naše testy budou založené na teorii maximální věrohodnosti, proto předpokládejme, že systém hustot $\mathcal{F}_m$ je regulární (viz definice 8) a jsou splněny další předpoklady věty 5. Mějme dále náhodný výběr $\mathbb{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)^\top$ z rozdělení s hustotou $f_{\boldsymbol{V}}(\boldsymbol{v};\boldsymbol{\theta})$. Označme

$$\boldsymbol{U}_1(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\tau}} = \begin{pmatrix} \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_q} \end{pmatrix}, \quad \boldsymbol{U}_2(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\psi}} = \begin{pmatrix} \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_{q+1}} \\ \vdots \\ \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_m} \end{pmatrix}$$

a dále si Fisherovu informační matici pro celý náhodný výběr rozdělme na bloky

$$\mathbb{J}^{(n)}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbb{J}_{11}^{(n)}(\boldsymbol{\theta}) & \mathbb{J}_{12}^{(n)}(\boldsymbol{\theta}) \\ \mathbb{J}_{21}^{(n)}(\boldsymbol{\theta}) & \mathbb{J}_{22}^{(n)}(\boldsymbol{\theta}) \end{pmatrix},$$

kde $\mathbb{J}_{11}^{(n)}$ je matice typu $q \times q$ a ostatní matice mají takový rozměr, aby $\mathbb{J}^{(n)}$ byla matice typu $m \times m$.

Dále bude nutné zavést značení, které nám umožní zapsat inverzi blokové diagonální matice. Vše nám shrnuje následující lemma.

Lemma 6. *Nechť*

$$\mathbb{J} = \begin{pmatrix} \mathbb{J}_{11} & \mathbb{J}_{12} \\ \mathbb{J}_{21} & \mathbb{J}_{22} \end{pmatrix}$$

je regulární matice, přičemž bloky $\mathbb{J}_{11}$ a $\mathbb{J}_{22}$ jsou čtvercové a regulární. Položme

$$\begin{aligned} \mathbb{J}_{11.2} &= \mathbb{J}_{11} - \mathbb{J}_{12}\mathbb{J}_{22}^{-1}\mathbb{J}_{21}, & \mathbb{J}^{11} &= \mathbb{J}_{11.2}^{-1}, & \mathbb{J}^{12} &= -\mathbb{J}_{11.2}^{-1}\mathbb{J}_{12}\mathbb{J}_{22}^{-1}, \\ \mathbb{J}_{22.1} &= \mathbb{J}_{22} - \mathbb{J}_{21}\mathbb{J}_{11}^{-1}\mathbb{J}_{12}, & \mathbb{J}^{22} &= \mathbb{J}_{22.1}^{-1}, & \mathbb{J}^{21} &= -\mathbb{J}_{22.1}^{-1}\mathbb{J}_{21}\mathbb{J}_{11}^{-1}. \end{aligned}$$

Pak

$$\mathbb{J}^{-1} = \begin{pmatrix} \mathbb{J}^{11} & \mathbb{J}^{12} \\ \mathbb{J}^{21} & \mathbb{J}^{22} \end{pmatrix}.$$

Důkaz. Snadnými algebraickými úpravami se ukáže, že součin matice $\mathbb{J}$ a $\mathbb{J}^{-1}$ dává jednotkovou matici. □

Pro zavedení testových statistik potřebujeme maximálně věrohodné odhady parametru $\boldsymbol{\theta}$. Musíme ovšem rozlišit dva odhady. První z nich, který budeme značit $\widehat{\boldsymbol{\theta}}_n$, bude maximálně věrohodný odhad parametru $\boldsymbol{\theta}$, který není svazován žádnými dalšími podmínkami. Tento odhad rozdělíme podle parametrů $\boldsymbol{\tau}$ a $\boldsymbol{\psi}$ na

$$\widehat{\boldsymbol{\theta}}_n = \begin{pmatrix} \widehat{\boldsymbol{\tau}}_n \\ \widehat{\boldsymbol{\psi}}_n \end{pmatrix}.$$

Druhý z nich, který budeme značit $\widetilde{\boldsymbol{\theta}}_n$, bude maximálně věrohodný odhad za platnosti nulové hypotézy $\boldsymbol{\tau} = \boldsymbol{\tau}_0$. Zde se tedy maximalizace týká jen parametru $\boldsymbol{\psi}$, jehož maximálně věrohodný odhad za platnosti hypotézy budeme značit $\widetilde{\boldsymbol{\psi}}_n$. Potom $\widetilde{\boldsymbol{\theta}}_n$ lze zapsat jako

$$\widetilde{\boldsymbol{\theta}}_n = \begin{pmatrix} \boldsymbol{\tau}_0 \\ \widetilde{\boldsymbol{\psi}}_n \end{pmatrix} = \underset{\boldsymbol{\theta}\in\Theta:\boldsymbol{\tau}=\boldsymbol{\tau}_0}{\arg\max}\, \ell(\boldsymbol{\theta}).$$

Nyní si zavedeme následující statistiky

$$LM = \left[\boldsymbol{U}_1\left(\widetilde{\boldsymbol{\theta}}_n\right)\right]^\top \left[\mathbb{J}^{(n)}_{11.2}\left(\widetilde{\boldsymbol{\theta}}_n\right)\right]^{-1} \boldsymbol{U}_1\left(\widetilde{\boldsymbol{\theta}}_n\right), \tag{2.1}$$

$$W = \left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right)^\top \mathbb{J}^{(n)}_{11.2}\left(\widehat{\boldsymbol{\theta}}_n\right)\left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right), \tag{2.2}$$

$$LR = 2\left[\ell\left(\widehat{\boldsymbol{\theta}}_n\right) - \ell\left(\widetilde{\boldsymbol{\theta}}_n\right)\right]. \tag{2.3}$$

U těchto statistik jsme schopni nalézt jejich asymptotické rozdělení za platnosti $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$, ovšem jen za platnosti dalších nutných předpokladů. Potom tedy využijeme této znalosti asymptotického rozdělení, abychom sestavili test na hladině $\alpha \in (0, 1)$ o hypotéze $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ proti alternativě, že tomu tak není.

V následující větě se neobejdeme bez těchto dvou obecných předpokladů:

(O1) $\Theta \subset \mathbb{R}^m$ je parametrický prostor, který obsahuje takové neprázdné otevřené okolí O, že skutečná hodnota parametru $\boldsymbol{\theta}$ náleží do tohoto okolí O.

(O2) Nechť $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$. Pak $f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta}_1) = f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta}_2)$ [μ] s.v. platí právě tehdy, je-li $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

Abychom mohli odvodit asymptotické rozdělení testových statistik LM, W a LR za platnosti $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$, budeme potřebovat předpoklady (A1)-(A3):

(A1) Pro [μ] skoro všechna $\boldsymbol{v}$, pro všechna $\boldsymbol{\theta} \in O$ a pro všechna $i, j, l \in \{1, \ldots, m\}$ existuje derivace

$$\frac{\partial^3 f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta})}{\partial\theta_i \partial\theta_j \partial\theta_l}.$$

(A2) Pro všechna $\boldsymbol{\theta} \in O$ a pro každé $i, j \in \{1, \ldots, m\}$ platí

$$\int_M f''_{ij}(\boldsymbol{v}; \boldsymbol{\theta}) \mathrm{d}\mu(\boldsymbol{v}) = 0.$$

(A3) Pro všechna $i, j, l \in \{1, \ldots, m\}$ existují funkce $M_{ijl}(\boldsymbol{v}) \geq 0$ takové, že

$$\mathsf{E}_{\boldsymbol{\theta}}\, M_{ijl}(\boldsymbol{V}) < \infty \quad \text{a} \quad \left| \frac{\partial^3 \log f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_l} \right| \leq M_{ijl}(\boldsymbol{v})$$

pro všechna $\boldsymbol{\theta} \in O$ a $[\mu]$ skoro všechna $\boldsymbol{v} \in M$.

Věta 7. *Mějme náhodný výběr* $\mathbb{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)^\top$ *z rozdělení s hustotou* $f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta})$ *vzhledem k* σ*-konečné míře* μ*. Nechť systém* $\mathcal{F}_m = \{f_{\boldsymbol{V}}(\boldsymbol{v}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ *je regulární (viz definice 8).*

Nechť jsou navíc splněny předpoklady (O1), (O2), (A1), (A2), (A3) a navíc je Fisherova informační matice $\mathbb{J}^{(n)}(\boldsymbol{\theta})$ *spojitá ve skutečné hodnotě parametru* $\boldsymbol{\theta}$*. Jestliže* $n \to \infty$*, pak za platnosti hypotézy* $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ *platí*

$$LM \xrightarrow{D} \chi^2_q, \qquad W \xrightarrow{D} \chi^2_q, \qquad LR \xrightarrow{D} \chi^2_q.$$

Důkaz. Důkaz je proveden v Anděl (2007). □

Známe tedy asymptotické rozdělení našich statistik za platnosti nulové hypotézy $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$. Můžeme tedy sestavit testy na asymptotické hladině významnosti $\alpha \in (0, 1)$. Tyto testy dosáhnou největší síly, pokud zvolíme kritický obor ve tvaru $\big(\chi^2_q(1-\alpha), \infty\big)$, tedy když

$$H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0 \text{ zamítáme } \iff LM, W \text{ nebo } LR \geq \chi^2_q(1-\alpha).$$

Poznámka. Několik poznámek k těmto testům.

a) Test založený na statistice LM (2.1) se nazývá *skórový test*, ale dříve se také používal název *test založený na Lagrangeových multiplikátorech*. Výhodou tohoto testu je, že stačí znát maximálně věrohodný odhad za platnosti nulové hypotézy. Výpočet inverzní matice nebývá náročný, neboť řád této matice odpovídá dimenzi q parametru $\boldsymbol{\tau}$ a není obvykle příliš velký.

b) Test založený na statistice W (2.2) se nazývá *Waldův test*. Narozdíl od skórového testu nevyžaduje invertování informační matice, ale na druhou stranu vyžaduje maximálně věrohodný odhad bez předpokladu platnosti nulové hypotézy, což může být výpočetně náročné.

c) Test založený na statistice LR (2.3) se nazývá *test založený na věrohodnostním poměru (likelihood ratio test)*. Tento test narozdíl od předchozích dvou nevyžaduje znalost Fisherovy informační matice.

Kapitola 3

Bartlettův test

V Andělově učebnici Anděl (2007, str. 210) je zaveden model analýzy rozptylu jednoduchého třídění. Základní myšlenkou je, že máme k dispozici I nezávislých náhodných výběrů z normálního rozdělení s obecně různými středními hodnotami a shodným rozptylem. Na základě analýzy rozptylu se potom odvozuje test o shodnosti středních hodnot. Ovšem při tomto testování je zásadní předpoklad shodnosti rozptylů ve všech skupinách, který ovšem nemusí být vždy splněn. Naším úkolem je tedy sestavit test, který by na hladině $\alpha \in (0,1)$ zamítl (nebo nevyvrátil) tento předpoklad homoskedasticity.

Uvažme tedy obecně následující model:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathsf{N}\left(0, \sigma_i^2\right), \quad i \in \{1, \ldots, I\}, n_i \geq 2, j \in \{1, \ldots, n_i\}, \quad (3.1)$$

kde $\mu_i \in \mathbb{R}, \sigma_i^2 > 0, i \in \{1, \ldots, I\}$ jsou neznámé parametry, I nám označuje počet skupin a $n := n_1 + \cdots + n_I$. Hypotéza homoskedasticity tedy odpovídá tomu, že $\sigma_1^2 = \ldots = \sigma_I^2 (= \sigma^2)$. Později si ukážeme, že se tento model dá chápat jako heteroskedastický lineární model tak, jak jsme ho zavedli v podkapitole 1.3.

Nejprve se seznámíme s postupem, který navrhl Bartlett ve svém článku Bartlett (1937). Označme výběrové rozptyly v jednotlivých skupinách symboly

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2, \quad \text{kde } \overline{Y}_{i\bullet} = \frac{1}{n} \sum_{j=1}^{n_i} Y_{ij}.$$

Za odhad společné hodnoty rozptylů σ^2 se dá považovat statistika

$$S^2 = \frac{1}{n - I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 = \sum_{i=1}^{I} \frac{n_i - 1}{n - I} S_i^2.$$

Jedná se o vážený průměr jednotlivých odhadů s vahami $(n_i - 1)/(n - I)$. Bartlettova testová statistika má potom tvar

$$\begin{aligned} B &= \frac{1}{c} \left((n - I) \log S^2 - \sum_{i=1}^{I} (n_i - 1) \log S_i^2 \right) \\ &= \frac{n - I}{c} \left(\log S^2 - \sum_{i=1}^{I} \frac{n_i - 1}{n - I} \log S_i^2 \right), \end{aligned} \quad (3.2)$$

kde konstanta c je dána vztahem

$$c = 1 + \frac{1}{3(I-1)} \left(\sum_{i=1}^{I} \frac{1}{n_i - 1} - \frac{1}{n - I} \right).$$

Bartlett (1937) odvodil, že statistika B má za platnosti nulové hypotézy asymptoticky χ^2_{I-1} rozdělení.

Vidíme, že statistika B je založena na porovnávání logaritmu váženého průměru odhadů rozptylu pro jednotlivá i s váženým průměrem logaritmů těchto odhadů. Tato statistika je tedy založena na principech věrohodnostního poměru.

V následující části odvodíme na základě podkapitoly 2.2 test založený na testové statistice LR (viz (2.3)). Nicméně se nebude jednat přímo o Bartlettovu statistiku (3.2). Jejich podobnost a asymptotické vlastnosti budeme diskutovat na závěr této kapitoly (podkapitola 3.2).

3.1 Test založený na věrohodnostním poměru

Náš model pojmeme obecněji, než bylo uvedeno na začátku této kapitoly. Nebudeme totiž předem předpokládat, do které skupiny je jaké pozorování zařazeno. Tento proces bude řízen náhodným vektorem $\boldsymbol{X}$.

Nechť $(Y, \boldsymbol{X}) \sim \mathsf{NLM}\left(\boldsymbol{X}^\top \boldsymbol{\mu}, \sigma^2 \boldsymbol{X}^\top \boldsymbol{w}\right)$. Buď $I \geq 2$. I-složkový náhodný vektor

$$\boldsymbol{X} \sim \mathsf{Mult}_I\left(1, \boldsymbol{p}\right), \text{ kde } \boldsymbol{p} = (p_1, \ldots, p_I)^\top \in (0,1)^I,$$

je neznámý parametr, který splňuje $\sum_{i=1}^{I} p_i = 1$. Dále $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_I)^\top \in \mathbb{R}^I$, $\sigma^2 > 0$ a $\boldsymbol{w} = (w_1, \ldots, w_I)^\top \in (0, \infty)^I$ jsou neznámé parametry až na w_1, o kterém předpokládejme, že platí $w_1 = 1$ za účelem identifikace, viz předpoklad (O2). Funkcí h ze zavedení heteroskedastického modelu je zjevně $h\left(\boldsymbol{X}; \boldsymbol{w}\right) = \boldsymbol{X}^\top \boldsymbol{w}$.

Uvědomme si, že náhodný vektor $\boldsymbol{X}$ může nabývat jen kanonických vektorů

$$\boldsymbol{e}_i = \left(\delta_{ij}\right)_{j=1}^{I}, \quad i \in \{1, \ldots, I\},$$

a to s pravděpodobností

$$p(\boldsymbol{e}_i; \boldsymbol{p}) = \mathsf{P}\left(\boldsymbol{X} = \boldsymbol{e}_i\right) = p_i.$$

Potom $\boldsymbol{X}^\top \boldsymbol{\mu} = \mu_i$ a $\boldsymbol{X}^\top \boldsymbol{w} = w_i$ pro nějaké $i \in \{1, \ldots, I\}$. Tedy

$$(Y | \boldsymbol{X} = \boldsymbol{e}_i) \sim \mathsf{N}\left(\mu_i, \sigma^2 w_i\right).$$

Tedy náhodný vektor $\boldsymbol{X}$ nám určuje, do které z I skupin zařadit pozorování Y. Dále předpokládáme, že náhodná veličina Y se v i-té skupině řídí normálním rozdělením se střední hodnotou μ_i a rozptylem $\sigma^2 w_i$, což mimochodem znamená, že v první skupině je rozptyl právě σ^2. Tedy připouštíme, že v každé skupině může být díky váhovému parametru $\boldsymbol{w}$ různý rozptyl.

Rozptyl ve všech skupinách bude stejný, pokud $w_1 = 1 = w_2 = \ldots = w_I$. Chceme tedy odvodit test založený na věrohodnostním poměru o hypotéze

$$H_0 : \boldsymbol{w} = \mathbb{1}_I \quad \text{proti alternativě } H_1 : \boldsymbol{w} \neq \mathbb{1}_I. \tag{3.3}$$

Postupujme tedy dle podkapitoly 2.2, nejprve se ovšem zaměřme na parametrický prostor tohoto modelu.

Parametr $\boldsymbol{w}$ bude cílovým parametrem (až na $w_1 = 1$), držíme-li se tedy značení zavedeného v sekci 2.2, tak $\boldsymbol{\tau} = (w_2, \ldots, w_I)$ a $\boldsymbol{\tau}_0 = \mathbb{1}_{I-1}$. Rušivými parametry budou $\boldsymbol{\mu}, \sigma^2, \boldsymbol{p}$. Parametr $\boldsymbol{p}$ ovšem nepochází přímo z otevřené množiny, neboť je svazován podmínkou $p_1 + \cdots + p_I = 1$. Definujme si tedy

$$P = \left\{(p_1, \ldots, p_{I-1})^\top \in (0,1)^{I-1} : p_1 + \cdots + p_{I-1} < 1\right\},$$

tato množina již otevřená je. Parametr p_I potom lze na základě $p_1, \ldots, p_{I-1}$ určit jako $p_I = 1 - (p_1 + \cdots + p_{I-1})$. Označme $\boldsymbol{\psi} = (\mu_1, \ldots, \mu_I, \sigma^2, p_1, \ldots, p_{I-1})^\top$ a $\boldsymbol{\theta} = (\boldsymbol{\tau}^\top, \boldsymbol{\psi}^\top)^\top$.

Parametrický prostor tedy bude

$$\Theta = (0, \infty)^{I-1} \times \mathbb{R}^I \times (0, \infty) \times P,$$

splňuje tak náš požadavek na otevřenost (R1) a navíc skutečná hodnota parametru $\boldsymbol{\theta}$ musí ležet v nějakém otevřeném okolí (O1).

Dále ať $f(y, \boldsymbol{x}; \boldsymbol{\theta})$ značí hustotu náhodného vektoru $(Y, \boldsymbol{X})$ vůči součinové míře $\lambda \times \nu$, $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\mu}, \sigma^2)$ značí hustotu podmíněného rozdělení $Y|\boldsymbol{X}$ vůči Lebesgueově míře λ a $p(\boldsymbol{x}; \boldsymbol{p})$ značí hustotu náhodného vektoru $\boldsymbol{X}$ vzhledem ke sčítací míře ν.

Nechť $y \in \mathbb{R}$ a $\boldsymbol{x}$ je nějaký kanonický vektor, potom pro každé $\boldsymbol{\theta} \in \Theta$ platí:

$$f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\mu}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 \boldsymbol{x}^\top \boldsymbol{w}}} \exp\left\{-\frac{\left(y - \boldsymbol{x}^\top \boldsymbol{\mu}\right)^2}{2\sigma^2 \boldsymbol{x}^\top \boldsymbol{w}}\right\}, \tag{3.4}$$

$$p(\boldsymbol{x}; \boldsymbol{p}) = \prod_{i=1}^{I} p_i^{x_i} = p_1^{x_1} \cdot \ldots \cdot p_I^{x_I}, \tag{3.5}$$

$$f(y, \boldsymbol{x}; \boldsymbol{\theta}) = f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\mu}, \sigma^2) \cdot p(\boldsymbol{x}; \boldsymbol{p}) \quad [\lambda \times \nu] \text{ s.v.} \tag{3.6}$$

Dále systém hustot

$$\mathcal{F}_{I-1+I+1+I-1} = \mathcal{F}_{3I-1} = \{f(y, \boldsymbol{x}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$$

je regulární (definice 8), což plyne z vlastností normálního a multinomického rozdělení.

Uvažme náhodný výběr $(Y_1, \boldsymbol{X}_1)$, …, $(Y_n, \boldsymbol{X}_n)$ z rozdělení $(Y, \boldsymbol{X})$. Pak

$$\boldsymbol{N} = (N_1, \ldots, N_I)^\top = \sum_{l=1}^{n} \boldsymbol{X}_l \sim \mathsf{Mult}_I(n, \boldsymbol{p})$$

a sdružená hustota $p_{\mathbb{X}}(\mathbb{X}; \boldsymbol{p})$ vzhledem k součinové míře $\nu^n = \nu \times \cdots \times \nu$ náhodné matice $\mathbb{X}$, která představuje náhodný výběr $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, je potom součinem hustot z rovnice (3.5). Platí tedy

$$p_{\mathbb{X}}(\mathbb{X}; \boldsymbol{p}) = \prod_{l=1}^{n} p_1^{x_{1l}} \cdot \ldots \cdot p_I^{x_{Il}} = \prod_{i=1}^{I} p_i^{n_i}, \tag{3.7}$$

kde $(n_1, \ldots, n_I)^\top$ jsou realizace náhodného vektoru $\boldsymbol{N}$, pro které $n_1 + \cdots + n_I = n$. Předpokládejme, že $n_i \geq 1, i \in \{1, \ldots, I\}$. Tedy n_i udává kolik pozorování $\boldsymbol{X}_l$

z náhodného výběru nabylo hodnoty $\boldsymbol{e}_i$, neboli kolika veličinám Y_l určilo $\boldsymbol{X}_l$ jejich příslušnost do skupiny $i \in \{1, \dots, I\}$.

Pro každé $l \in \{1, \dots, n\}$ určitě Y_l spadá do některé skupiny $i \in \{1, \dots, I\}$ reprezentované hodnotou náhodného vektoru $\boldsymbol{X}_l$. Přeznačme tedy veličiny Y_l na veličiny Y_{ij}, kde $i \in \{1, \dots, I\}$ označuje příslušnou skupinu a $j \in \{1, \dots, n_i\}$ bude index, pomocí kterého budu procházet veličiny ve skupinách. Používáme tedy značení shodné s tím, které jsme uvedli na začátku této kapitoly, viz (3.1).

Zaveďme si pro $i \in \{1, \dots, I\}$

a) průměr hodnot Y_{ij} v i-té skupině $\overline{Y}_{i\bullet} = \dfrac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ a

b) reziduální součet čtverců v i-té skupině $RSS_i = \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$.

Tvrzení 8. *V modelu uvedeném výše za platnosti $H_0 : \boldsymbol{w} = \mathbb{1}_I$ statistika*

$$LR = n \log\left(\frac{1}{n} \sum_{i=1}^{I} RSS_i\right) - \sum_{i=1}^{I} n_i \log\left(\frac{RSS_i}{n_i}\right) \xrightarrow{D} \chi^2_{I-1}, \qquad n \to \infty.$$

Důkaz. Ukážeme si, že statistika LR je statistikou pro test poměrem věrohodnosti, viz (2.3). K tomu potřebujeme nalézt maximálně věrohodné odhady parametrů $\boldsymbol{\mu}$, σ^2, $\boldsymbol{w}$ a $\boldsymbol{p}$, a to jak za platnosti nulové hypotézy H_0, tak bez platnosti tohoto předpokladu. Nakonec tyto odhady dosadíme do logaritmické věrohodnosti ℓ a odečteme vzniklé výrazy. Začněme tedy s určením funkce ℓ.

Sdruženou hustotu podmíněného rozdělení $f_{\boldsymbol{Y}|\mathbb{X}}(\boldsymbol{y}|\mathbb{X}; \boldsymbol{w}, \boldsymbol{\mu}, \sigma^2)$ lze odvodit pomocí vzorce (3.4) a dostaneme

$$f_{\boldsymbol{Y}|\mathbb{X}}(\boldsymbol{y}|\mathbb{X}; \boldsymbol{w}, \boldsymbol{\mu}, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \cdot \prod_{i=1}^{I} w_i^{\frac{n_i}{2}} \cdot \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{I} \frac{1}{w_i} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2\right\}. \tag{3.8}$$

Označme nyní

$$\ell_1(\boldsymbol{w}, \boldsymbol{\mu}, \sigma^2) = \log f_{\boldsymbol{Y}|\mathbb{X}}(\boldsymbol{Y}|\mathbb{X}; \boldsymbol{w}, \boldsymbol{\mu}, \sigma^2) \text{ a } \ell_2(\boldsymbol{p}) = \log p_{\mathbb{X}}(\mathbb{X}; \boldsymbol{p}).$$

Pak logaritmováním výrazů (3.8) a (3.7) dostaneme

$$\ell_1(\boldsymbol{w}, \boldsymbol{\mu}, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \\ - \sum_{i=1}^{I} \frac{n_i}{2} \log w_i - \frac{1}{2\sigma^2} \sum_{i=1}^{I} \frac{1}{w_i} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2, \tag{3.9}$$

$$\ell_2(\boldsymbol{p}) = \sum_{i=1}^{I} n_i \log p_i. \tag{3.10}$$

Směřujeme nyní k logaritmické věrohodnostní funkci $\ell(\boldsymbol{\theta}) = \log f(\boldsymbol{Y}, \mathbb{X}; \boldsymbol{\theta})$. Rozšířením rovnosti (3.6) na sdružené hustoty dostáváme spolu s rovnostmi (3.9)

a (3.10) následující

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \log f(\boldsymbol{Y}, \mathbb{X}; \boldsymbol{\theta}) = \ell_1(\boldsymbol{w}, \boldsymbol{\mu}, \sigma^2) + \ell_2(\boldsymbol{p}), \quad [\lambda^n \times \nu^n] \text{ s.v.}, \\
\ell(\boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \sum_{i=1}^{I}\frac{n_i}{2}\log w_i - \\
&\quad - \frac{1}{2\sigma^2}\sum_{i=1}^{I}\frac{1}{w_i}\sum_{j=1}^{n_i}(Y_{ij}-\mu_i)^2 + \sum_{i=1}^{I} n_i \log p_i.
\end{aligned} \tag{3.11}$$

Druhá rovnost také platí jen $[\lambda^n \times \nu^n]$ skoro všude, ovšem od této chvíle si dovolíme tento dovětek nadále nepsat.

Nyní můžeme přistoupit k hledání maximálně věrohodných odhadů. Začněme s hledáním odhadů $\widehat{\boldsymbol{p}}_n$ a $\widetilde{\boldsymbol{p}}_n$. Všimneme si, že parametr $\boldsymbol{p}$ není nijak svázán s ostatními parametry, tedy platnost nulové hypotézy se na našich odhadech nijak neprojeví. Ukáži konkrétně na výpočtech.

Jedná se nám v podstatě o maximalizaci funkce ℓ_2 za dodatečné podmínky $p_1 + \cdots + p_I = 1$, což vyřešíme metodou Lagrangeových multiplikátorů. Definujme si pomocnou funkci $g(\boldsymbol{p}, \lambda)$ jako

$$g(\boldsymbol{p}, \lambda) = \sum_{i=1}^{I} n_i \log p_i - \lambda\left(\sum_{i=1}^{I} p_i - 1\right).$$

Potom spočtěme derivace podle proměnných λ a $p_i, i \in \{1, \ldots, I\}$.

$$\begin{aligned}
\frac{\partial g(\boldsymbol{p}, \lambda)}{\partial \lambda} &= 1 - \sum_{i=1}^{I} p_i, \\
\frac{\partial g(\boldsymbol{p}, \lambda)}{\partial p_i} &= \frac{n_i}{p_i} - \lambda, \quad i \in \{1, \ldots, I\}.
\end{aligned}$$

Položíme-li tyto derivace rovny nule, tak pro každé $i \in \{1, \ldots, I\}$ dostáváme, že musí platit $\lambda = \frac{n_i}{\widehat{p}_i}$, tedy také $\widehat{p}_i = \frac{n_i}{\lambda}$. Potom musí také dle první rovnosti platit

$$1 = \sum_{i=1}^{I} \frac{n_i}{\lambda} = \frac{n_1 + \cdots + n_I}{\lambda} = \frac{n}{\lambda}.$$

Tím tedy dostáváme $\lambda = n$ a také odhady

$$\widehat{\boldsymbol{p}}_n = \widetilde{\boldsymbol{p}}_n = \left(\frac{n_1}{n}, \ldots, \frac{n_I}{n}\right)^\top. \tag{3.12}$$

U ostatních parametrů budeme opravdu muset rozlišit případy, kdy nulová hypotéza (3.3) platí a kdy ne. Začněme s jednodušším případem, tedy s tím, když tato hypotéza platí.

Pak tedy víme, že $w_1 = 1 = w_2 = \ldots = w_I$ a proto má funkce ℓ z (3.11) tvar

$$\ell(\boldsymbol{\tau}_0, \boldsymbol{\psi}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\mu_i)^2 + \sum_{i=1}^{I} n_i \log p_i.$$

Spočtěme tedy derivace podle proměnných $\boldsymbol{\mu}$ a σ^2 (podle $\boldsymbol{p}$ už nemusíme):

$$\frac{\partial \ell(\boldsymbol{\tau}_0, \boldsymbol{\psi})}{\partial \mu_i} = \frac{1}{\sigma^2} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i), \quad i \in \{1, \ldots, I\}, \tag{3.13}$$

$$\frac{\partial \ell(\boldsymbol{\tau}_0, \boldsymbol{\psi})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2. \tag{3.14}$$

Pokud položíme výraz (3.13) roven nule, tak pro každé $i \in \{1, \ldots, I\}$ dostaneme odhady

$$\widetilde{\boldsymbol{\mu}}_n = \left(\frac{1}{n_1} \sum_{j=1}^{n_1} Y_{1j}, \ldots, \frac{1}{n_I} \sum_{j=1}^{n_I} Y_{Ij} \right) = \left(\overline{Y}_{1\bullet}, \ldots, \overline{Y}_{I\bullet} \right). \tag{3.15}$$

Dále pokud položíme výraz (3.14) roven nule a dosadíme odhad $\widetilde{\boldsymbol{\mu}}_n$, obdržíme

$$\widetilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet} \right)^2 = \frac{1}{n} \sum_{i=1}^{I} RSS_i. \tag{3.16}$$

Získali jsme tedy maximálně věrohodné odhady za platnosti nulové hypotézy. Dále pokračujme nalezením takových odhadů bez tohoto předpokladu. Postupujme zcela obdobně. Funkce ℓ má tedy tvar

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \sum_{i=1}^{I} \frac{n_i}{2} \log w_i - \\ - \frac{1}{2\sigma^2} \sum_{i=1}^{I} \frac{1}{w_i} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 + \sum_{i=1}^{I} n_i \log p_i.$$

Derivováním této funkce dostáváme

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mu_i} = \frac{1}{\sigma^2 w_i} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i), \quad i \in \{1, \ldots, I\}, \tag{3.17}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{I} \frac{1}{w_i} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2, \tag{3.18}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial w_i} = -\frac{n_i}{2w_i} + \frac{1}{2\sigma^2 w_i^2} \sum_{j=1}^{n_i} \left(Y_{ij} - \mu_i \right)^2, \quad i \in \{1, \ldots, I\}. \tag{3.19}$$

Položíme-li výraz (3.17) roven nule, tak dostaneme ten samý odhad pro $\boldsymbol{\mu}$ jako v předešlém případě, tedy odhad $\widehat{\boldsymbol{\mu}}_n = \widetilde{\boldsymbol{\mu}}_n$, viz (3.15). Jelikož předpokládáme znalost $w_1 = 1$, tak můžeme položit $\widehat{w}_1 = 1$. Rovnice (3.19) pro $i = 1$ je tedy nadbytečná. Ovšem vyjdeme-li z ní, tak dostaneme

$$\widehat{\sigma}_n^2 = \frac{1}{n_1} \sum_{j=1}^{n_1} \left(Y_{1j} - \overline{Y}_{1\bullet} \right)^2 = \frac{RSS_1}{n_1}.$$

Potom ze vztahu (3.19) pro každé $i \in \{2, \ldots, I\}$ dostaneme, že musí platit

$$n_i \widehat{w}_i = \frac{1}{\widehat{\sigma}_n^2} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet} \right)^2,$$

tedy jednoduchou úpravou dostáváme, že platí

$$\widehat{\sigma}_n^2 \widehat{w}_i = \frac{RSS_i}{n_i}, \quad \forall i \in \{1, \dots, I\}. \tag{3.20}$$

Odtud bychom dostali, že

$$\widehat{w}_i = \frac{RSS_i}{RSS_1} \frac{n_1}{n_i}, \quad \forall i \in \{1, \dots, I\}.$$

Zbývá ověřit, že takové odhady také nulují výraz (3.18), aby byla splněna celá soustava věrohodnostních rovnic. Jelikož

$$\begin{aligned} \sum_{i=1}^{I} \frac{1}{\widehat{w}_i} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 &= \sum_{i=1}^{I} \frac{RSS_i}{\widehat{w}_i} = \sum_{i=1}^{I} \frac{n_i}{n_1} RSS_1 \\ &= RSS_1 \frac{n_1 + \dots + n_I}{n_1} = \frac{nRSS_1}{n_1} = n\widehat{\sigma}_n^2, \end{aligned}$$

tak dostáváme, že opravdu

$$-\frac{n}{2\widehat{\sigma}_n^2} + \frac{1}{2\widehat{\sigma}_n^4} \sum_{i=1}^{I} \frac{1}{\widehat{w}_i} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 = -\frac{n}{2\widehat{\sigma}_n^2} + \frac{1}{2\widehat{\sigma}_n^4} n\widehat{\sigma}_n^2 = 0.$$

Nyní již jsme schopni určit testovou statistiku

$$LR = 2\left(\ell(\widehat{\boldsymbol{\theta}}_n) - \ell(\widetilde{\boldsymbol{\theta}}_n)\right).$$

Díky výše uvedeným vzorcům (3.11), (3.15), (3.16) a (3.20) platí:

$$\begin{aligned} \ell(\widehat{\boldsymbol{\theta}}_n) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\widehat{\sigma}_n^2 - \sum_{i=1}^{I} \frac{n_i}{2}\log\widehat{w}_i - \frac{1}{2\widehat{\sigma}_n^2} \sum_{i=1}^{I} \frac{RSS_i}{\widehat{w}_i} + \ell_2(\widehat{\boldsymbol{p}}_n) \\ &= -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{I} \frac{n_i}{2}\log\left(\widehat{\sigma}_n^2\widehat{w}_i\right) - \frac{1}{2}\sum_{i=1}^{I} \frac{RSS_i}{\widehat{\sigma}_n^2\widehat{w}_i} + \ell_2(\widehat{\boldsymbol{p}}_n) \\ &= -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{I} \frac{n_i}{2}\log\left(\frac{RSS_i}{n_i}\right) - \frac{1}{2}\sum_{i=1}^{I} \frac{n_i RSS_i}{RSS_i} + \ell_2(\widehat{\boldsymbol{p}}_n) \\ &= -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{I} \frac{n_i}{2}\log\left(\frac{RSS_i}{n_i}\right) - \frac{n}{2} + \ell_2(\widehat{\boldsymbol{p}}_n), \\ \ell(\widetilde{\boldsymbol{\theta}}_n) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\widetilde{\sigma}_n^2 - \frac{1}{2\widetilde{\sigma}_n^2}\sum_{i=1}^{I} RSS_i + \ell_2(\widetilde{\boldsymbol{p}}_n) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{1}{n}\sum_{i=1}^{I} RSS_i\right) - \frac{n}{2} + \ell_2(\widetilde{\boldsymbol{p}}_n). \end{aligned}$$

Nyní si už jen uvědomíme, že jsme odvodili, že $\widetilde{\boldsymbol{p}}_n = \widehat{\boldsymbol{p}}_n$, viz (3.12), a odečteme dvojnásobek výrazů výše, čímž dostaneme statistiku

$$\begin{aligned} LR &= n\log\left(\frac{1}{n}\sum_{i=1}^{I} RSS_i\right) - \sum_{i=1}^{I} n_i \log\left(\frac{RSS_i}{n_i}\right) \\ &= n\left[\log\left(\frac{1}{n}\sum_{i=1}^{I} RSS_i\right) - \sum_{i=1}^{I} \frac{n_i}{n}\log\left(\frac{RSS_i}{n_i}\right)\right]. \end{aligned} \tag{3.21}$$

Vidíme tedy, že naše testová statistika je založená na rozdílu logaritmu celkového reziduálního součtu čtverců děleného n a součtu vážených logaritmů reziduálních součtů čtverců dělených n_i v rámci i-té skupiny.

Zbývá ověřit předpoklady věty 7, které, jak víme, platí pro normální i multinomické rozdělení. Proto již není těžké odvodit, že jsou splněny i v tomto modelu.

Podle věty 7 má tedy naše statistika asymptotické rozdělení chí kvadrát o takovém počtu stupňů volnosti, kolik bylo cílových parametrů. Jelikož jsme předpokládali znalost $w_1 = 1$, tak mluvíme jen o zbylých $w_2, \ldots, w_I$. Tedy počet stupňů volnosti je $I - 1$. Celkově tedy

$$LR \xrightarrow{\mathsf{D}} \chi^2_{I-1}, \quad n \to \infty.$$

□

Vidíme, že výsledná statistika LR (viz (3.21)) je velmi podobná statistice B (viz (3.2)), dokonce mají i stejné asymptotické rozdělení za platnosti nulové hypotézy. Náš test tedy pro obě statistiky zamítne na asymptotické hladině významnosti $\alpha \in (0, 1)$ hypotézu $H_0 : w_1 = w_2 = \ldots = w_I = 1$ právě tehdy, když realizovaná hodnota statistiky B nebo LR převýší $\chi^2_{I-1}(1 - \alpha)$. Jak moc se ale použití statistik B a LR liší, shrneme v následující podkapitole.

3.2 Porovnání testových statistik

Nejprve se podívejme, v čem si jsou statistiky (3.2) a (3.21) podobné. Uvědomme si, že statistika B se také sestává z jednotlivých $RSS_i, i \in \{1, \ldots, I\}$. Jednotlivé odhady rozptylu v i-té skupině, kde $i \in \{1, \ldots, I\}$, vypadají takto:

$$S_i^2 = \frac{RSS_i}{n_i - 1}, \qquad \widehat{\sigma}_n^2 \widehat{w}_i = \frac{RSS_i}{n_i}.$$

Liší se tedy jen ve jmenovateli, kde v prvním případě se RSS_i dělí počtem stupňů volnosti $n_i - 1$, kdežto v té druhé jen počtem pozorování v i-té skupině.

Celkové odhady jsou potom také podobného tvaru

$$S^2 = \frac{1}{n - I} \sum_{i=1}^{I} RSS_i, \qquad \widetilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{I} RSS_i.$$

Ještě jednou uvádím statistiky, aby byl vidět jejich rozdíl:

$$\begin{aligned} LR &= n \left[\log \left(\frac{1}{n} \sum_{i=1}^{I} RSS_i \right) - \sum_{i=1}^{I} \frac{n_i}{n} \log \left(\frac{RSS_i}{n_i} \right) \right], \\ B &= \frac{n - I}{c} \left[\log \left(\frac{1}{n - I} \sum_{i=1}^{I} RSS_i \right) - \sum_{i=1}^{I} \frac{n_i - 1}{n - I} \log \left(\frac{RSS_i}{n_i - 1} \right) \right], \qquad (3.22) \\ &\text{kde } c = 1 + \frac{1}{3(I-1)} \left(\sum_{i=1}^{I} \frac{1}{n_i - 1} - \frac{1}{n - I} \right). \end{aligned}$$

Bartlettova statistika tedy používá počty stupňů volnosti namísto počtu pozorování a navíc je podělená konstantou c, která zpravidla bývá jen nepatrně větší než 1.

Všechny tyto úpravy způsobují, že statistika B má lepší asymptotické vlastnosti než LR, proto se také na rozdíl od LR v praxi používá k testování homoskedasticity. Toto tvrzení prozkoumáme dále v numerických simulacích v kapitole 5.

Jak již bylo řečeno, statistika B má za platnosti nulové hypotézy asymptoticky rozdělení χ^2_{I-1}. Udává se, že tato vlastnost lze použít k testování, pokud platí $n_i \geq 7, i \in \{1, \ldots, I\}$.

Tento Bartlettův test se zdá být při splnění předpokladu normality tím nejsilnějším z dostupných testů. Ovšem je velmi citlivý na porušení předpokladu o normálním rozdělení. Existují proto i další možné modifikace uvedené například v Zvára (2008, str. 120).

Kapitola 4

Breusch-Paganův test

Uvažme nyní následující normální heteroskedastický lineární model

$$(Y, \boldsymbol{X}) \sim \mathsf{NLM}\left(\boldsymbol{X}^\top \boldsymbol{\beta}, \sigma^2 \exp\left(\boldsymbol{Z}^\top \boldsymbol{\tau}\right)\right),$$

kde $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^\top \in \mathbb{R}^{k+1}$, $\sigma^2 > 0$ a $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_q)^\top \in \mathbb{R}^q$ jsou neznámé parametry a $k \in \mathbb{N}_0, q \in \mathbb{N}$. Dále vektor $\boldsymbol{Z}$ je nějakou známou transformací vektoru $\boldsymbol{X}$, tedy $\boldsymbol{Z} = g(\boldsymbol{X})$ pro známou funkci $g : \mathbb{R}^{k+1} \to \mathbb{R}^q$. Lze také místo $\exp\left(\boldsymbol{Z}^\top \boldsymbol{\tau}\right)$ uvažovat obecně nějakou diferencovatelnou kladnou reálnou funkci h od skalárního součinu $\boldsymbol{Z}^\top \boldsymbol{\tau}$ s vlastností $h(0) = 1$, zde volíme $h = \exp$ pro jednoduchost.

Mějme k dispozici náhodný výběr $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ z rozdělení $(Y, \boldsymbol{X})$. Na základě tohoto náhodného výběru máme sestavit test o homoskedasticitě. Jak je vidět z naší parametrizace, tak toto nastává, pokud $\boldsymbol{\tau} = \boldsymbol{0}^q$.

Jak už jsme ospravedlnili v sekci 2.1, lze pracovat jen s podmíněným rozdělením $Y|\boldsymbol{X}$, které je dle předpokladu normální. Předpokládejme tedy, že

$$\mathbb{X} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix} = (\boldsymbol{x}_{\bullet 0}, \ldots, \boldsymbol{x}_{\bullet k}), \quad \mathbb{Z} = \begin{pmatrix} \boldsymbol{z}_1^\top \\ \vdots \\ \boldsymbol{z}_n^\top \end{pmatrix} = \begin{pmatrix} g(\boldsymbol{x}_1)^\top \\ \vdots \\ g(\boldsymbol{x}_n)^\top \end{pmatrix} = (\boldsymbol{z}_{\bullet 1}, \ldots, \boldsymbol{z}_{\bullet q})$$

jsou matice reálných konstant s plnou sloupcovou hodností. Dále označme

$$\mathbb{W}(\boldsymbol{\tau}) = \mathsf{diag}\left(w_1(\boldsymbol{\tau}), \ldots, w_n(\boldsymbol{\tau})\right) = \mathsf{diag}\left(\exp\left(\boldsymbol{z}_1^\top \boldsymbol{\tau}\right), \ldots, \exp\left(\boldsymbol{z}_n^\top \boldsymbol{\tau}\right)\right).$$

Vzhledem k našemu předpokladu má náhodný vektor $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ při pevných $\mathbb{X}$ a $\mathbb{Z}$ rozdělení

$$\boldsymbol{Y} \sim \mathsf{N}_n\left(\mathbb{X}\boldsymbol{\beta}, \sigma^2 \mathbb{W}(\boldsymbol{\tau})\right). \tag{4.1}$$

Chceme tedy odvodit test pro následující hypotézu

$$H_0 : \boldsymbol{\tau} = \boldsymbol{0}^q \quad \text{proti alternativě} \quad H_1 : \boldsymbol{\tau} \neq \boldsymbol{0}^q. \tag{4.2}$$

Jelikož máme neznámé parametry $\boldsymbol{\tau}$, σ^2 a $\boldsymbol{\beta}$ a chceme se věnovat testování $\boldsymbol{\tau}$, tak použijeme teorie testů s rušivými parametry uvedené v podkapitole 2.2. Označme $\boldsymbol{\theta} = (\boldsymbol{\tau}, \sigma^2, \boldsymbol{\beta})$, příslušný parametrický prostor je tedy

$$\Theta = \mathbb{R}^q \times (0, \infty) \times \mathbb{R}^{k+1}.$$

Rušivé parametry tedy budou $\boldsymbol{\psi} = (\sigma^2, \beta_0, \ldots, \beta_k)^\top$ a cílovým parametrem bude $\boldsymbol{\tau}$. Naším cílem bude odvodit skórový test založený na testové statistice LM (2.1).

Buďte

$$\mathbb{A} = \mathbb{I}_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top, \quad \widetilde{\sigma}_n^2 = \frac{RSS}{n}, \quad \boldsymbol{v} = \begin{pmatrix} u_1^2 - \widetilde{\sigma}_n^2 \\ \vdots \\ u_n^2 - \widetilde{\sigma}_n^2 \end{pmatrix},$$

kde $\boldsymbol{u}$ je vektor reziduí, RSS je reziduální součet čtverců a $\mathbb{1}_n$ značí n-složkový sloupcový vektor $\mathbb{1}_n = (1, 1, \ldots, 1)^\top$.

Tvrzení 9. *V modelu popsaném výše za platnosti $H_0 : \boldsymbol{\tau} = \mathbf{0}^q$ statistika*

$$LM = \frac{1}{2\left(\widetilde{\sigma}_n^2\right)^2}\boldsymbol{v}^\top\mathbb{Z}\left(\mathbb{Z}^\top\mathbb{A}\mathbb{Z}\right)^{-1}\mathbb{Z}^\top\boldsymbol{v} \xrightarrow{D} \chi_q^2, \qquad n \to \infty.$$

Důkaz. Potřebujeme tedy nalézt Fisherovu informační matici a její inverzi, dále maximálně věrohodné odhady $\widetilde{\boldsymbol{\beta}}_n$ a $\widetilde{\sigma}_n^2$ za platnosti nulové hypotézy (4.2) a ověřit předpoklady věty 7.

Začněme se sdruženou hustotou $f_n\left(\boldsymbol{y}|\mathbb{X};\boldsymbol{\theta}\right)$ vektoru $\boldsymbol{Y}$ při dané realizaci $\mathbb{X}$ vzhledem k Lebesgueově míře λ^n. Platí

$$\begin{aligned} f_n(\boldsymbol{y}|\mathbb{X};\boldsymbol{\theta}) &= \frac{(2\pi)^{-\frac{n}{2}}}{\sqrt{\det\left(\sigma^2\mathbb{W}(\boldsymbol{\tau})\right)}}\exp\left\{-\frac{1}{2}\big(\boldsymbol{y}-\mathbb{X}\boldsymbol{\beta}\big)^\top\frac{\mathbb{W}^{-1}(\boldsymbol{\tau})}{\sigma^2}\big(\boldsymbol{y}-\mathbb{X}\boldsymbol{\beta}\big)\right\} \\ &= \frac{(2\pi)^{-\frac{n}{2}}}{(\sigma^2)^{\frac{n}{2}}\left(\prod_{i=1}^n w_i(\boldsymbol{\tau})\right)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\sum_{i=1}^n\frac{\left(y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2}{\sigma^2 w_i(\boldsymbol{\tau})}\right\} \\ &= \frac{(2\pi)^{-\frac{n}{2}}}{(\sigma^2)^{\frac{n}{2}}\exp\left\{\frac{1}{2}\sum_{i=1}^n\boldsymbol{z}_i^\top\boldsymbol{\tau}\right\}}\exp\left\{-\frac{1}{2}\sum_{i=1}^n\frac{\left(y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2}{\sigma^2\exp\left(\boldsymbol{z}_i^\top\boldsymbol{\tau}\right)}\right\}. \end{aligned}$$

První rovnost je zřejmá z vlastnosti našeho modelu (4.1) a definice mnohorozměrného normálního rozdělení. Druhá rovnost plyne z diagonality matice $\sigma^2\mathbb{W}(\boldsymbol{\tau})$, neboť pak její determinant je jen součin prvků na diagonále a inverzní matice má na diagonále jen převrácené prvky. Třetí rovnost nám plyne z vlastnosti prvků $w_i(\boldsymbol{\tau})$ a exponenciály.

Uvědomíme si, že systém hustot $\mathcal{F}_{q+k+2} = \{f(y|\boldsymbol{X};\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ je regulární, viz definice 8. Všechny parametry pocházejí z nějakého otevřeného intervalu v parametrickém prostoru Θ. Navíc $M = \{y \in \mathbb{R} : f(y|\boldsymbol{X};\boldsymbol{\theta}) > 0\}$ nezávisí na parametru $\boldsymbol{\theta}$. Díky vlastnostem normálního rozdělení jsou také splněny všechny ostatní předpoklady a také předpoklad z věty 5, tedy lze Fisherovu informační matici počítat pomocí střední hodnoty z druhých parciálních derivací logaritmické věrohodnosti ℓ.

Dále tedy budeme pracovat s přirozeným logaritmem hustoty. Platí

$$\ell(\boldsymbol{\theta}) = \log f_n(\boldsymbol{Y}|\mathbb{X};\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^n\boldsymbol{z}_i^\top\boldsymbol{\tau} - \frac{1}{2\sigma^2}\sum_{i=1}^n\frac{\left(Y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2}{\exp(\boldsymbol{z}_i^\top\boldsymbol{\tau})}.$$

Dalším krokem je spočtení všech derivací podle neznámých parametrů.

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \tau_m} = -\frac{1}{2}\sum_{i=1}^{n} z_{im} + \frac{1}{2\sigma^2}\sum_{i=1}^{n} \frac{\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})} z_{im}, \quad m \in \{1, \ldots, q\}, \tag{4.3}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} \frac{\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})}, \tag{4.4}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_j} = \frac{1}{\sigma^2}\sum_{i=1}^{n} \frac{x_{ij}\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})}, \quad j \in \{0, \ldots, k\}. \tag{4.5}$$

Dále si můžeme vyřešit soustavu věrohodnostních rovnic a nalézt si tak maximálně věrohodné odhady. Ovšem pro naše potřeby stačí jen nalézt odhady σ^2 a $\boldsymbol{\beta}$ za platnosti nulové hypotézy (4.2). Takže položíme výrazy z rovnic (4.4) a (4.5) rovny nule, dále položíme $\boldsymbol{\tau} = \mathbf{0}^q$ a nalezneme odhady $\widetilde{\sigma}_n^2$ a $\widetilde{\boldsymbol{\beta}}_n$.

V soustavě věrohodnostních rovnic tedy platí

$$\frac{1}{2\left(\widetilde{\sigma}_n^2\right)^2}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}_n\right)^2 = \frac{n}{2\widetilde{\sigma}_n^2} \iff \widetilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}_n\right)^2.$$

$$\begin{aligned}
\forall j \in \{0, \ldots, k\} : \sum_{i=1}^{n} x_{ij}\left(Y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}_n\right) = 0 &\iff \forall j : \boldsymbol{x}_{\bullet j}^\top\left(\boldsymbol{Y} - \mathbb{X}\widetilde{\boldsymbol{\beta}}_n\right) = 0 \\
&\iff \mathbb{X}^\top\left(\boldsymbol{Y} - \mathbb{X}\widetilde{\boldsymbol{\beta}}_n\right) = \mathbf{0}^{k+1} \\
&\iff \mathbb{X}^\top\mathbb{X}\widetilde{\boldsymbol{\beta}}_n = \mathbb{X}^\top\boldsymbol{Y} \\
&\iff \widetilde{\boldsymbol{\beta}}_n = \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\boldsymbol{Y}.
\end{aligned}$$

Zavedeme-li si podobně jako v sekci 1.2 vektor $\widetilde{\boldsymbol{Y}} = \mathbb{X}\widetilde{\boldsymbol{\beta}}_n$, dostaneme odhady tvaru

$$\widetilde{\boldsymbol{\beta}}_n = \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\boldsymbol{Y} \quad \text{a} \quad \widetilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widetilde{Y}_i\right)^2 = \frac{RSS}{n} \tag{4.6}$$

jako v klasickém modelu $\boldsymbol{Y} \sim \mathsf{N}_n\left(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbb{I}_n\right)$.

Pokračujme dále ve výpočtech druhých derivací funkce ℓ. Dostáváme

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \tau_{m_1}\partial \tau_{m_2}} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} \frac{\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})} z_{im_1} z_{im_2}, \quad m_1, m_2 \in \{1, \ldots, q\},$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \left(\sigma^2\right)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n} \frac{\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})},$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_{j_1}\partial \beta_{j_2}} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} \frac{x_{ij_1}x_{ij_2}}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})}, \quad j_1, j_2 \in \{0, \ldots, k\},$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \sigma^2 \partial \beta_j} = -\frac{1}{\sigma^4}\sum_{i=1}^{n} \frac{x_{ij}\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})}, \quad j \in \{0, \ldots, k\},$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \sigma^2 \partial \tau_m} = -\frac{1}{2\sigma^4}\sum_{i=1}^{n} \frac{\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})} z_{im}, \quad m \in \{1, \ldots, q\},$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_j \partial \tau_m} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} \frac{x_{ij}\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})} z_{im}, \quad j \in \{0, \ldots, k\}, m \in \{1, \ldots, q\}.$$

Využijme tedy těchto druhých derivací k tomu, abychom podle věty 5 spočetli Fisherovu informační matici za platnosti naší hypotézy (4.2). Přičemž tento předpoklad je v naší parametrizaci nutný jen u jedné rovnosti, proto jsem nad onu rovnost uvedl symbol H_0. Připomenu-li, že $\mathsf{E}\, Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ a $\mathsf{var}\, Y_i = \sigma^2 \exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})$, tak snadno dostaneme, že

$$\begin{aligned}
\mathsf{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \tau_{m_1} \partial \tau_{m_2}}\right] &= \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{\mathsf{var}\, Y_i}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})} z_{im_1} z_{im_2} = \frac{1}{2} \sum_{i=1}^{n} z_{im_1} z_{im_2} = \frac{\boldsymbol{z}_{\bullet m_1}^\top \boldsymbol{z}_{\bullet m_2}}{2}, \\
\mathsf{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \left(\sigma^2\right)^2}\right] &= -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} \sum_{i=1}^{n} \frac{\mathsf{var}\, Y_i}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})} = -\frac{n}{2\sigma^4} + \frac{n\sigma^2}{\sigma^6} = \frac{n}{2\sigma^4}, \\
\mathsf{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_{j_1} \partial \beta_{j_2}}\right] &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{x_{ij_1} x_{ij_2}}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})} \overset{H_0}{=} \frac{\boldsymbol{x}_{\bullet j_1}^\top \boldsymbol{x}_{\bullet j_2}}{\sigma^2}, \\
\mathsf{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \sigma^2 \partial \beta_j}\right] &= \frac{1}{\sigma^4} \sum_{i=1}^{n} \frac{x_{ij}\, \mathsf{E}\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})} = 0, \\
\mathsf{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \sigma^2 \partial \tau_m}\right] &= \frac{1}{2\sigma^4} \sum_{i=1}^{n} \frac{\mathsf{var}\, Y_i}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})} z_{im} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} z_{im} = \frac{\boldsymbol{z}_{\bullet m}^\top \mathbb{1}_n}{2\sigma^2}, \\
\mathsf{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_j \partial \tau_m}\right] &= \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{x_{ij}\, \mathsf{E}\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\tau})} z_{im} = 0.
\end{aligned}$$

Nyní již máme vše potřebné pro sestavení Fisherovy informační matice. Bude to matice řádu $q + 1 + (k+1)$, kterou zapíšu blokově tak, aby cílový parametr $\boldsymbol{\tau}$ byl v levém horním rohu. Matice má tedy tvar

$$\mathbb{J}^{(n)}\left(\boldsymbol{\tau}, \sigma^2, \boldsymbol{\beta}\right) \overset{H_0}{=} \begin{pmatrix} \dfrac{\mathbb{Z}^\top \mathbb{Z}}{2} & \dfrac{\mathbb{Z}^\top \mathbb{1}_n}{2\sigma^2} & \mathbb{O}^{q \times (k+1)} \\ \dfrac{\mathbb{1}_n^\top \mathbb{Z}}{2\sigma^2} & \dfrac{n}{2\sigma^4} & \mathbf{0}^{1 \times (k+1)} \\ \mathbb{O}^{(k+1) \times q} & \mathbf{0}^{(k+1) \times 1} & \dfrac{\mathbb{X}^\top \mathbb{X}}{\sigma^2} \end{pmatrix}. \tag{4.7}$$

Matici (4.7) si rozdělme na bloky a používejme dále značení zavedené v lemmatu 6. V našem případě máme takovouto situaci:

$$\mathbb{J}_{11}^{(n)} = \frac{\mathbb{Z}^\top \mathbb{Z}}{2}, \qquad \mathbb{J}_{12}^{(n)} = \begin{pmatrix} \dfrac{\mathbb{Z}^\top \mathbb{1}_n}{2\sigma^2} & \mathbb{O}^{q \times (k+1)} \end{pmatrix},$$

$$\mathbb{J}_{21}^{(n)} = \begin{pmatrix} \dfrac{\mathbb{1}_n^\top \mathbb{Z}}{2\sigma^2} \\ \mathbb{O}^{(k+1) \times q} \end{pmatrix}, \qquad \mathbb{J}_{22}^{(n)} = \begin{pmatrix} \dfrac{n}{2\sigma^4} & \mathbf{0}^{1 \times (k+1)} \\ \mathbf{0}^{(k+1) \times 1} & \dfrac{\mathbb{X}^\top \mathbb{X}}{\sigma^2} \end{pmatrix}.$$

Povšimneme si, že díky našim předpokladům o plných sloupcových hodnostech matic $\mathbb{X}$ a $\mathbb{Z}$ jsou matice $\mathbb{Z}^\top \mathbb{Z}$ a $\mathbb{X}^\top \mathbb{X}$ regulární, a tedy jsou regulární i $\mathbb{J}_{11}^{(n)}$ a $\mathbb{J}_{22}^{(n)}$.

Pro odvození testové statistiky budeme potřebovat $\mathbb{J}_{11.2}^{(n)}$ a následně pracovat s její inverzní maticí. Pokračujme tedy určením této matice. Využijeme přitom blokové diagonality matice $\mathbb{J}_{22}^{(n)}$, kterou zinvertujeme „po blocích“ (na tuto matici lze aplikovat lemma 6).

Postupnými úpravami dostáváme

$$\begin{aligned}
\mathbb{J}_{11.2}^{(n)} &= \mathbb{J}_{11}^{(n)} - \mathbb{J}_{12}^{(n)} \left(\mathbb{J}_{22}^{(n)}\right)^{-1} \mathbb{J}_{21}^{(n)} \\
&= \frac{\mathbb{Z}^\top \mathbb{Z}}{2} - \begin{pmatrix} \frac{\mathbb{Z}^\top \mathbb{1}_n}{2\sigma^2} & \mathbb{O}^{q\times(k+1)} \end{pmatrix} \begin{pmatrix} \frac{2\sigma^4}{n} & \mathbf{0}^{1\times(k+1)} \\ \mathbf{0}^{(k+1)\times 1} & \sigma^2 \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \end{pmatrix} \begin{pmatrix} \frac{\mathbb{1}_n^\top \mathbb{Z}}{2\sigma^2} \\ \mathbb{O}^{(k+1)\times q} \end{pmatrix} \\
&= \frac{\mathbb{Z}^\top \mathbb{Z}}{2} - \frac{\mathbb{Z}^\top \mathbb{1}_n}{2\sigma^2} \cdot \frac{2\sigma^4}{n} \cdot \frac{\mathbb{1}_n^\top \mathbb{Z}}{2\sigma^2} \\
&= \frac{1}{2}\mathbb{Z}^\top \left(\mathbb{I}_n - \frac{1}{n}\mathbb{1}_n \mathbb{1}_n\right)^\top \mathbb{Z} = \frac{1}{2}\mathbb{Z}^\top \mathbb{A}\mathbb{Z}.
\end{aligned}$$

Je zřejmé, že matice $\mathbb{A}$, kterou jsme si zavedli již před tvrzením 9, je symetrická ($\mathbb{A} = \mathbb{A}^\top$) a idempotentní ($\mathbb{A}^2 = \mathbb{A}$).

Označme dále $\boldsymbol{U}_1\left(\boldsymbol{\tau}, \sigma^2, \boldsymbol{\beta}\right)$ vektor parciálních derivací funkce ℓ podle cílových proměnných $\boldsymbol{\tau}$. Platí

$$\boldsymbol{U}_1\left(\boldsymbol{\tau}, \sigma^2, \boldsymbol{\beta}\right) = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\tau}, \sigma^2, \boldsymbol{\beta})}{\partial \tau_1} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\tau}, \sigma^2, \boldsymbol{\beta})}{\partial \tau_q} \end{pmatrix} \underset{H_0}{\overset{(4.3)}{=}} \begin{pmatrix} \frac{1}{2\sigma^2} \sum_{i=1}^n \left[\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2 - \sigma^2\right] z_{i1} \\ \vdots \\ \frac{1}{2\sigma^2} \sum_{i=1}^n \left[\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2 - \sigma^2\right] z_{iq} \end{pmatrix}.$$

Nás bude dle (2.1) zajímat hodnota tohoto vektoru pro $\boldsymbol{\tau} = \mathbf{0}^q$ a maximálně věrohodné odhady $\widetilde{\sigma}_n^2$ a $\widetilde{\boldsymbol{\beta}}_n$, což dle (4.6) dává

$$\boldsymbol{U}_1\left(\mathbf{0}^q, \widetilde{\sigma}_n^2, \widetilde{\boldsymbol{\beta}}_n\right) = \begin{pmatrix} \frac{1}{2\widetilde{\sigma}_n^2} \sum_{i=1}^n \left(u_i^2 - \widetilde{\sigma}_n^2\right) z_{i1} \\ \vdots \\ \frac{1}{2\widetilde{\sigma}_n^2} \sum_{i=1}^n \left(u_i^2 - \widetilde{\sigma}_n^2\right) z_{iq} \end{pmatrix},$$

kde $u_i = Y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}_n = Y_i - \boldsymbol{x}_i^\top \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \boldsymbol{Y} = Y_i - \widetilde{Y}_i$, $i \in \{1, \ldots, n\}$ jsou rezidua tvořící dohromady vektor $\boldsymbol{u}$. Připomeňme, že druhá mocnina eukleidovské normy tohoto vektoru je právě reziduální součet čtverců, neboli $RSS = \|\boldsymbol{u}\|^2$.

Vektor $\boldsymbol{U}_1$ lze dále upravit do jednoduššího tvaru. Ukážeme si dva možné způsoby, jak se na toto dívat. V prvním případě si definujme vektor $\boldsymbol{v}$ následovně

$$v_i := u_i^2 - \widetilde{\sigma}_n^2 = u_i^2 - \frac{RSS}{n} = \frac{1}{n} \sum_{j=1}^n \left(u_i^2 - u_j^2\right), i \in \{1, \ldots, n\}.$$

Vidíme, že i-tá složka vektoru $\boldsymbol{v}$ je průměr rozdílů čtverce rezidua u_i se všemi ostatními. Pak lze vektor $\boldsymbol{U}_1$ zapsat jako

$$\boldsymbol{U}_1\left(\mathbf{0}^q, \widetilde{\sigma}_n^2, \widetilde{\boldsymbol{\beta}}_n\right) = \begin{pmatrix} \frac{1}{2\widetilde{\sigma}_n^2} \sum_{i=1}^n v_i z_{i1} \\ \vdots \\ \frac{1}{2\widetilde{\sigma}_n^2} \sum_{i=1}^n v_i z_{iq} \end{pmatrix} = \begin{pmatrix} \frac{1}{2\widetilde{\sigma}_n^2} \boldsymbol{z}_{\bullet 1}^\top \boldsymbol{v} \\ \vdots \\ \frac{1}{2\widetilde{\sigma}_n^2} \boldsymbol{z}_{\bullet q}^\top \boldsymbol{v} \end{pmatrix} = \frac{\mathbb{Z}^\top \boldsymbol{v}}{2\widetilde{\sigma}_n^2} = \frac{n\mathbb{Z}^\top \boldsymbol{v}}{2RSS}. \tag{4.8}$$

Druhá možnost je definovat si pomocné veličiny

$$\overline{z}_m := \frac{1}{n}\sum_{i=1}^{n} z_{im} = \frac{1}{n}\boldsymbol{z}_{\bullet m}^{\top}\mathbb{1}_n,\ m \in \{1,\dots,q\},$$

které dohromady tvoří sloupcový vektor $\overline{\boldsymbol{z}} = \frac{1}{n}\mathbb{Z}^{\top}\mathbb{1}_n$. Potom platí

$$\begin{aligned}\sum_{i=1}^{n}\left(u_i^2 - \widetilde{\sigma}_n^2\right) z_{im} &= \sum_{i=1}^{n} u_i^2 z_{im} - \frac{RSS}{n}\sum_{i=1}^{n} z_{im} \\ &= \sum_{i=1}^{n} u_i^2 z_{im} - RSS\overline{z}_m = \sum_{i=1}^{n} u_i^2\left(z_{im} - \overline{z}_m\right).\end{aligned}$$

A tedy se potom dá vektor $\boldsymbol{U}_1$ zapsat jako

$$\boldsymbol{U}_1\left(\mathbf{0}^q, \widetilde{\sigma}_n^2, \widetilde{\boldsymbol{\beta}}_n\right) = \frac{1}{2\widetilde{\sigma}_n^2}\sum_{i=1}^{n} u_i^2\left(\boldsymbol{z}_i - \overline{\boldsymbol{z}}\right). \tag{4.9}$$

Pomocí těchto průměrů $\overline{z}_m$ lze také upravit naši matici $\mathbb{J}_{11.2}$, a to do tvaru

$$\mathbb{J}_{11.2}^{(n)} = \frac{1}{2}\mathbb{Z}^{\top}\mathbb{A}\mathbb{Z} = \frac{1}{2}\left(\mathbb{Z} - \mathbb{1}_n\overline{\boldsymbol{z}}^{\top}\right)^{\top}\left(\mathbb{Z} - \mathbb{1}_n\overline{\boldsymbol{z}}^{\top}\right). \tag{4.10}$$

O platnosti této rovnosti se jednoduše přesvědčíme pomocí následujících úprav:

$$\begin{aligned}\left(\mathbb{Z} - \mathbb{1}_n\overline{\boldsymbol{z}}^{\top}\right)^{\top}\left(\mathbb{Z} - \mathbb{1}_n\overline{\boldsymbol{z}}^{\top}\right) &= \left(\mathbb{Z} - \mathbb{1}_n\frac{1}{n}\left(\mathbb{Z}^{\top}\mathbb{1}_n\right)^{\top}\right)^{\top}\left(\mathbb{Z} - \mathbb{1}_n\frac{1}{n}\left(\mathbb{Z}^{\top}\mathbb{1}_n\right)^{\top}\right) \\ &= \left(\mathbb{Z} - \frac{\mathbb{1}_n\mathbb{1}_n^{\top}}{n}\mathbb{Z}\right)^{\top}\left(\mathbb{Z} - \frac{\mathbb{1}_n\mathbb{1}_n^{\top}}{n}\mathbb{Z}\right) \\ &= \left(\mathbb{A}\mathbb{Z}\right)^{\top}\left(\mathbb{A}\mathbb{Z}\right) = \mathbb{Z}^{\top}\mathbb{A}^2\mathbb{Z} = \mathbb{Z}^{\top}\mathbb{A}\mathbb{Z}.\end{aligned}$$

Nyní už si můžeme definovat testovou statistiku LM podle (2.1) následovně

$$LM := \left[\boldsymbol{U}_1\left(\mathbf{0}^q, \widetilde{\sigma}_n^2, \widetilde{\boldsymbol{\beta}}_n\right)\right]^{\top}\left[\mathbb{J}_{11.2}^{(n)}\left(\mathbf{0}^q, \widetilde{\sigma}_n^2, \widetilde{\boldsymbol{\beta}}_n\right)\right]^{-1}\boldsymbol{U}_1\left(\mathbf{0}^q, \widetilde{\sigma}_n^2, \widetilde{\boldsymbol{\beta}}_n\right).$$

Zde použijeme našich úvah výše, abychom viděli, jak testová statistika LM vypadá. Kombinací dvojic rovnic (4.8), (4.10) a (4.9), (4.10) dostaneme, že LM má tvar

$$\begin{aligned}LM &= \left(\frac{\mathbb{Z}^{\top}\boldsymbol{v}}{2\widetilde{\sigma}_n^2}\right)^{\top}\left(\frac{1}{2}\mathbb{Z}^{\top}\mathbb{A}\mathbb{Z}\right)^{-1}\frac{\mathbb{Z}^{\top}\boldsymbol{v}}{2\widetilde{\sigma}_n^2} = \frac{1}{2\left(\widetilde{\sigma}_n^2\right)^2}\boldsymbol{v}^{\top}\mathbb{Z}\left(\mathbb{Z}^{\top}\mathbb{A}\mathbb{Z}\right)^{-1}\mathbb{Z}^{\top}\boldsymbol{v} \\ &= \frac{1}{2\left(\widetilde{\sigma}_n^2\right)^2}\left[\sum_{i=1}^{n} u_i^2\left(\boldsymbol{z}_i - \overline{\boldsymbol{z}}\right)\right]^{\top}\left[\left(\mathbb{Z} - \mathbb{1}_n\overline{\boldsymbol{z}}^{\top}\right)^{\top}\left(\mathbb{Z} - \mathbb{1}_n\overline{\boldsymbol{z}}^{\top}\right)\right]^{-1}\left[\sum_{i=1}^{n} u_i^2\left(\boldsymbol{z}_i - \overline{\boldsymbol{z}}\right)\right].\end{aligned} \tag{4.11}$$

Zbývá už jen ověřit předpoklady věty 7. Už víme, že systém hustot $\mathcal{F}_{q+k+2}$ je regulární a že skutečná hodnota parametrů leží v nějakém otevřeném intervalu obsaženém v Θ. Dále je vzhledem k naší parametrizaci splněn předpoklad (O2) z vlastnosti hustoty normálního rozdělení. Existence třetích derivací (A1) plyne

jednoduše dalším derivováním druhých derivací funkce ℓ, v čemž nám nic nebrání. Předpoklad (A2) jsme již oprávněně používali při výpočtu Fisherovy informační matice. Předpoklad (A3) plyne opět z vlastnosti normálního rozdělení. Z rovnosti (4.7) plyne, že Fisherova matice je spojitá v každém $\boldsymbol{\theta} \in \Theta$, tedy i ve skutečné hodnotě tohoto parametru. Tím jsme již ověřili předpoklady.

Věta 7 mi tedy dává, že za platnosti nulové hypotézy $H_0 : \boldsymbol{\tau} = \mathbf{0}^q$ platí

$$LM \xrightarrow{\mathsf{D}} \chi^2_q, \qquad n \to \infty, \tag{4.12}$$

kde počet stupňů volnosti q odpovídá počtu cílových parametrů.

□

Náš skórový test tedy na asymptotické hladině významnosti $\alpha \in (0, 1)$ zamítne hypotézu H_0 právě tehdy, když $LM \geq \chi^2_q(1 - \alpha)$.

Tento test navrhli v roce 1979 Breusch a Pagan ve svém článku Breusch a Pagan (1979). Ukázalo se ovšem, že tento test je velmi citlivý na porušení předpokladu normálního rozdělení. Proto o dva roky později navrhl Koenker v článku Koenker (1981) úpravu, která tuto testovou statistiku *studentizovala*, takže se dá použít v případě, že si nejsme jisti s platností předpokladu normality.

Úprava spočívá v tom, že v testové statistice

$$LM = \frac{1}{2\left(\widetilde{\sigma}^2_n\right)^2} \boldsymbol{v}^\top \mathbb{Z} \left(\mathbb{Z}^\top \mathbb{A} \mathbb{Z}\right)^{-1} \mathbb{Z}^\top \boldsymbol{v}$$

nahradíme $\left(\widetilde{\sigma}^2_n\right)^2$ odhadem rozptylu veličin ε_i^2 pomocí

$$\widetilde{\varphi}_n = \frac{1}{n} \sum_{i=1}^{n} \left(u_i^2 - \widetilde{\sigma}^2_n\right)^2 = \frac{1}{n} \sum_{i=1}^{n} v_i^2 = \frac{\|\boldsymbol{v}\|^2}{n}.$$

Kapitola 5

Numerické studie

V této kapitole se na základě počítačových simulací v prostředí **R** (R Core Team, 2016) podíváme, jaké vlastnosti mají testy odvozené v kapitolách 3 a 4 v závislosti na rozsahu náhodného výběru n a na splnění předpokladu normality. Veškeré testy budeme provádět na asymptotické hladině 5 %.

5.1 Experimentální porovnání statistik B a LR

V této části se vrátíme zpět k testům homoskedasticity v modelu analýzy rozptylu jednoduchého třídění s obecně různými rozptyly v jednotlivých skupinách uvedeným v kapitole 3, které byly založeny na dvojici statistik

$$LR = n\left[\log\left(\frac{1}{n}\sum_{i=1}^{I} RSS_i\right) - \sum_{i=1}^{I}\frac{n_i}{n}\log\left(\frac{RSS_i}{n_i}\right)\right],$$

$$B = \frac{n-I}{c}\left[\log\left(\frac{1}{n-I}\sum_{i=1}^{I} RSS_i\right) - \sum_{i=1}^{I}\frac{n_i-1}{n-I}\log\left(\frac{RSS_i}{n_i-1}\right)\right],$$

$$\text{kde } c = 1 + \frac{1}{3(I-1)}\left(\sum_{i=1}^{I}\frac{1}{n_i-1} - \frac{1}{n-I}\right).$$

Obě statistiky by se při dostatečně velkém n měly řídit rozdělením χ^2_{I-1}, kde I je počet skupin. Na konkrétním příkladě tedy prozkoumáme, jak velké n je zapotřebí, abychom toto mohli tvrdit. Dále se podíváme, zda testy založené na těchto statistikách dodržují předepsanou hladinu. Nakonec prozkoumáme jejich schopnost rozpoznat neplatnou hypotézu, tedy experimentálně prozkoumáme jejich sílu. U těchto testů nebudeme zkoumat jejich chování při nesplnění předpokladu normality.

Naše simulace provedeme vždy pro $I = 5$ skupin. V podkapitole 3.1 jsme zavedli model tak, že přiřazování napozorovaných hodnot veličiny Y do skupin probíhá náhodně. Regresor určující příslušnost do skupiny je generován z rozdělení $\mathsf{Mult}_5(n, \boldsymbol{p})$, kde $\boldsymbol{p}$ si nastavíme na hodnoty

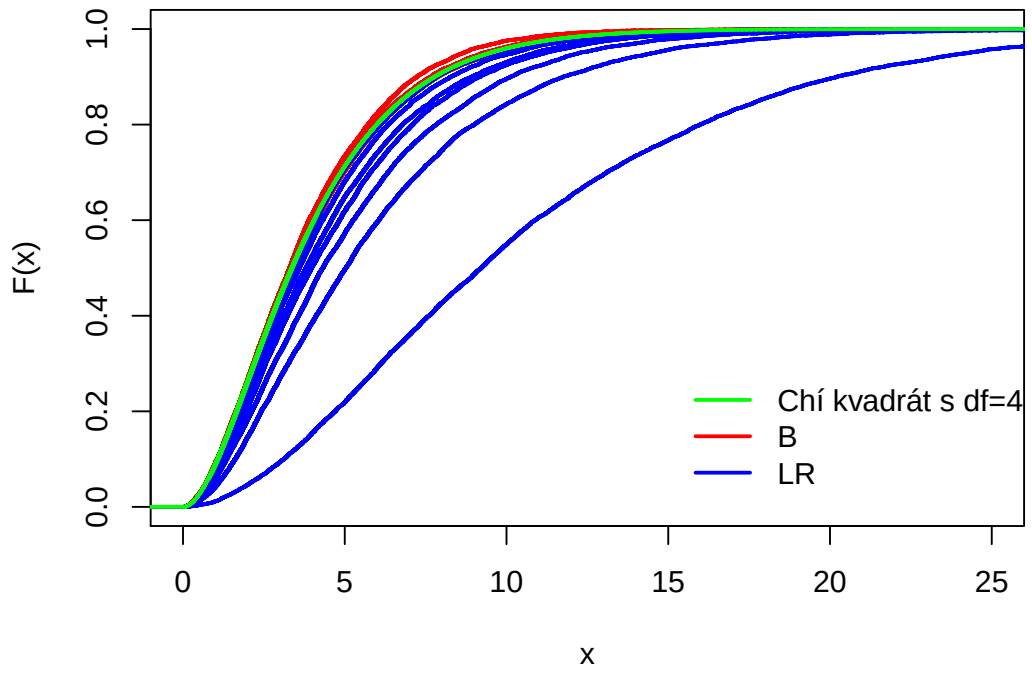
$$\boldsymbol{p} = \left(\frac{1}{15}, \frac{2}{15}, \frac{3}{15}, \frac{4}{15}, \frac{5}{15}\right)^\top.$$

Pro výpočet statistiky B musíme zajistit, aby v každé skupině byly alespoň dvě pozorování.

Nejprve začneme s případem, kdy je hypotéza homoskedasticity platná. Všechna pozorování y tedy budeme generovat z normálního rozdělení se stejnou hodnotou rozptylu $\sigma^2 = 1$. Střední hodnoty v jednotlivých skupinách nastavíme na různé hodnoty, v tomto případě nastavíme na hodnoty $\boldsymbol{\mu} = (-1, 0, 1, 2, 3)^\top$. Dále si určíme jednotlivé rozsahy výběrů $n \geq 5 \cdot 2$, pro které budeme chtít generovat data. Pro každé takové n potom 10 000-krát nasimulujeme náhodný výběr z tohoto rozdělení. Pro jednotlivou simulaci napočítáme hodnoty testových statistik B a LR a podíváme se, zda překročily kritickou hodnotu $\chi^2_4(0.95) \doteq 9.4877$. Statistiku LR spočteme jednoduše z našeho vzorce (3.21), na výpočet statistiky B lze použít funkce `bartlett.test` z balíčku `stats`.

Naše simulace provedeme pro $n \in \{10, 30, 50, 80, 100, 200, 400\}$. Výsledky experimentů jsou shrnuty na obrázku 5.1 a v tabulce 5.1. Obrázek 5.1 se zaměřuje na zjištění skutečného rozdělení jednotlivých statistik za nulové hypotézy. Můžeme si všimnout, že empirická distribuční funkce pro statistiku B se téměř překrývá s distribuční funkcí rozdělení χ^2_4, a to už i pro $n = 10$. Na druhou stranu je vidět, že empirická distribuční funkce statistiky LR se pro nízké hodnoty n výrazně odlišuje od distribuční funkce rozdělení χ^2_4, ale s rostoucím n se k této funkci blíží a při $n = 400$ už s ní také splývá.

Tabulka 5.1 uvádí, jaký byl podíl zamítnutí nulové hypotézy v 10 000 opakováních při jednotlivých rozsazích n. Vidíme, že Bartlettova statistika už při $n = 30$ hladiny 0.05 téměř dosáhla. Na druhou stranu statistika LR zamítala pro nízká n daleko častěji a hladině 0.05 se přiblížila až pro $n = 400$.



Obrázek 5.1: Grafy empirických distribučních funkcí statistik B a LR za platnosti hypotézy při rostoucím rozsahu výběru.

n	B	LR
10	0.0328	0.4815
30	0.0464	0.1783
50	0.0497	0.1228
80	0.0516	0.0899
100	0.0529	0.0822
200	0.0527	0.0632
400	0.0471	0.0528

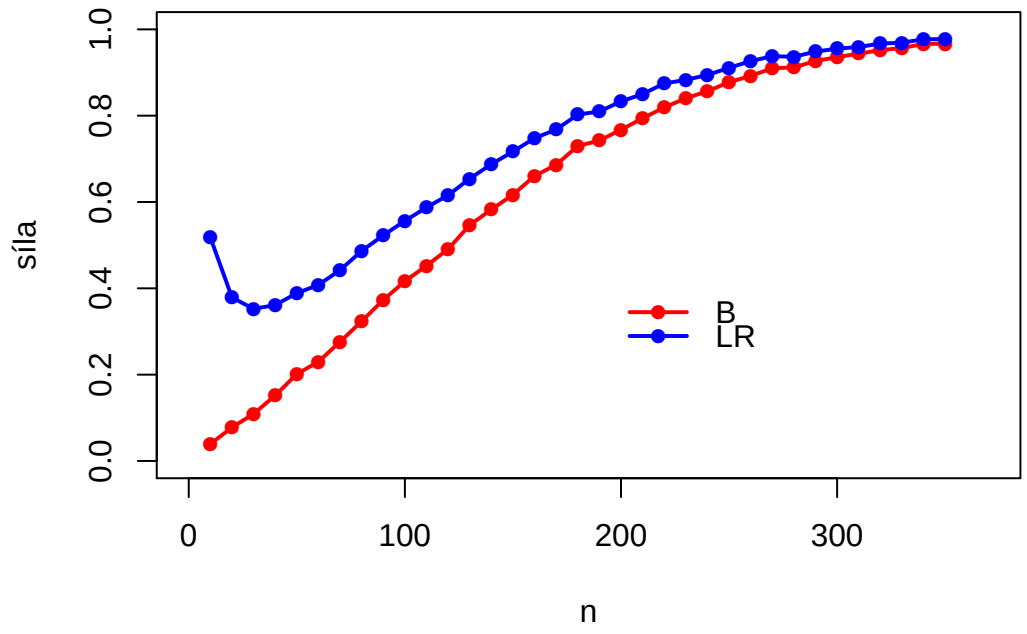
Tabulka 5.1: Naměřené hladiny statistik B a LR za platnosti nulové hypotézy.

Konvergence statistiky LR k limitnímu rozdělení χ^2_4 je viditelně pomalejší než v případě statistiky B. Toto je v souladu se zjištěními prezentovanými v literatuře, které jsme uvedli v kapitole 3.

Nyní budeme zkoumat sílu těchto statistik. Jediné, co na našem současném modelu musíme změnit, je platnost homoskedasticity. Budeme tedy generovat pozorování stejným postupem, jediné, co změníme, je nastavení rozptylu v jednotlivých skupinách. Nastavme tedy parametr směrodatné odchylky v jednotlivých skupinách na $\boldsymbol{\sigma} = (1, 1.2, 1.4, 1.6, 1.8)^\top$.

Abychom prozkoumali sílu těchto statistik, provedeme podrobnější simulace. Budeme uvažovat rozsahy náhodného výběru z množiny

$$N = \{10, 20, 30, 40, 50, \ldots, 340, 350\}.$$



Obrázek 5.2: Experimentální síla statistik B a LR při různém rozsahu výběru.

Pro každé $n \in N$ provedeme 10 000 simulací a zaznamenáme, kolikrát jsme se (správně) rozhodli zamítnout nulovou hypotézu homoskedasticity. Naše výsledky jsou zaznamenány na obrázku 5.2. Můžeme vidět, že u obou statistik dochází s rostoucím n ke zvětšování podílu zamítnutých hypotéz, dokonce se pro vysoké n blíží k jedné. Statistika LR zjevně zamítá neplatnou hypotézu homoskedasticity častěji než Bartlettova statistika, ovšem zásadnější je pro nás dodržení hladiny. Tu LR v našem modelu dodržuje až při n blízkém 400, viz tabulka 5.1, jenže pro takové n už jsou síly B a LR srovnatelné.

Na tomto modelu jsme se tedy přesvědčili, že statistika B má skutečně lepší asymptotické vlastnosti než námi odvozená statistika LR.

V příloze pod položkou 1) uvádím skript, který jsem stvořil při provádění simulací. Manipulací s tímto skriptem lze nejen zreprodukovat naše výsledky, ale také zkoumat dále vlastnosti statistik B a LR za volby jiných parametrů.

5.2 Breusch-Paganův test

Nyní budeme vyšetřovat vlastnosti testové statistiky LM odvozené v kapitole 4. V tomto případě budeme zkoumat, zda naše statistika dodržuje předepsanou hladinu, už se ale nebudeme zajímat o její sílu. Navíc prověříme, jak (ne)dodržuje předepsanou hladinu, není-li splněn předpoklad o normálním rozdělení. V obou případech budeme navíc zkoumat vliv *studentizace* této statistiky, která byla navržena v práci Koenker (1981), viz poznámka v závěru kapitoly 4. Naše simulace provedeme na následujícím modelu.

Pro zvolený rozsah výběru n budeme uvažovat matici konstant $\mathbb{X}$ o třech sloupcích o délce n, přičemž první z nich bude obsahovat pouze prvky 1, druhý bude tvořen posloupností n čísel tvořící ekvidistantní dělení na intervalu $[0, 1]$ a třetí sloupec bude střídavě obsahovat prvky 0 a 1. Neboli pro n sudé

$$\mathbb{X}^\top = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & \dots & 1 & 1 \\ 0 & \frac{1}{n-1} & \frac{2}{n-1} & \frac{3}{n-1} & \frac{4}{n-1} & \dots & \frac{n-2}{n-1} & 1 \\ 0 & 1 & 0 & 1 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

Sice hodláme naše data generovat za předpokladu pevného rozptylu, ale abychom mohli testovat, tak musíme určit na jakých pomocných regresorech $\mathbb{Z}$ by mohl obecně roztyl záviset. My zvolíme za pomocné regresory prostřední sloupec matice $\mathbb{X}$, tedy vektor

$$\boldsymbol{z} = \left(0, \frac{1}{n-1}, \frac{2}{n-1}, \frac{3}{n-1}, \frac{4}{n-1}, \dots, \frac{n-2}{n-1}, 1\right)^\top.$$

Tedy parametrizace rozptylu by za platnosti alternativy u veličiny Y_i vypadala jako $\sigma^2 \exp(z_i \cdot \tau), i \in \{1, \dots, n\}$, kde $\tau \in \mathbb{R}$ je neznámý jednorozměrný parametr, o němž budeme testovat, zda se rovná nule. Proto asymptotické rozdělení statistiky LM by mělo být χ^2_1. Příslušná kritická hodnota pro asymptotickou hladinu 5 % je $\chi^2_1(0.95) \doteq 3.8415$. Dále nastavíme hodnoty ostatních parametrů na

$$\boldsymbol{\beta} = (0, 1, 1)^\top, \qquad \sigma = 2.$$

Nyní budeme moct přistoupit ke generování veličin $y_i, i \in \{1, \ldots, n\}$, které budeme provádět následovně

$$y_i := \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, i \in \{1, \ldots, n\},$$

kde $\boldsymbol{x}_i$ a $\boldsymbol{\beta}$ již známe, ale veličiny ε_i si nagenerujeme. Budeme uvažovat 3 případy, podle toho, z jakého rozdělení ε_i pochází. Nejprve podle předpokladu provedeme pro $\varepsilon_i \sim \mathsf{N}(0, 2^2)$. Pro rozpor s předpokladem normality použijeme rozdělení $\mathsf{LN}(0, 2^2)$ (logaritmicko-normální) a t_3 (studentovo rozdělení se 3 stupni volnosti). Jakmile budeme mít všechny tyto veličiny nagenerovány, tak můžeme přistoupit k samotnému testování. Využijeme k tomu balíčku `lmtest` (Zeileis a Hothorn, 2002), konkrétně funkce `bptest`, která zvládá napočítat, jak obyčejnou statistiku LM, tak i její studentizovanou verzi.

Pro každé $n \in \{30, 50, 80, 100, 150, 200, 300, 400\}$ nagenerujeme 10 000 náhodných výběrů o rozsahu n pro každé z těch tří rozdělení a pro každý náhodný výběr z daného rozdělení napočítáme, jak studentizovanou, tak obecnou verzi statistiky LM. Zaznamenáme si, kolikrát jsme celkově přesáhli kritickou hodnotu $\chi^2_1(0.95) \doteq 3.8415$ v jednotlivých případech. Naše výsledky shrnuje tabulka 5.2.

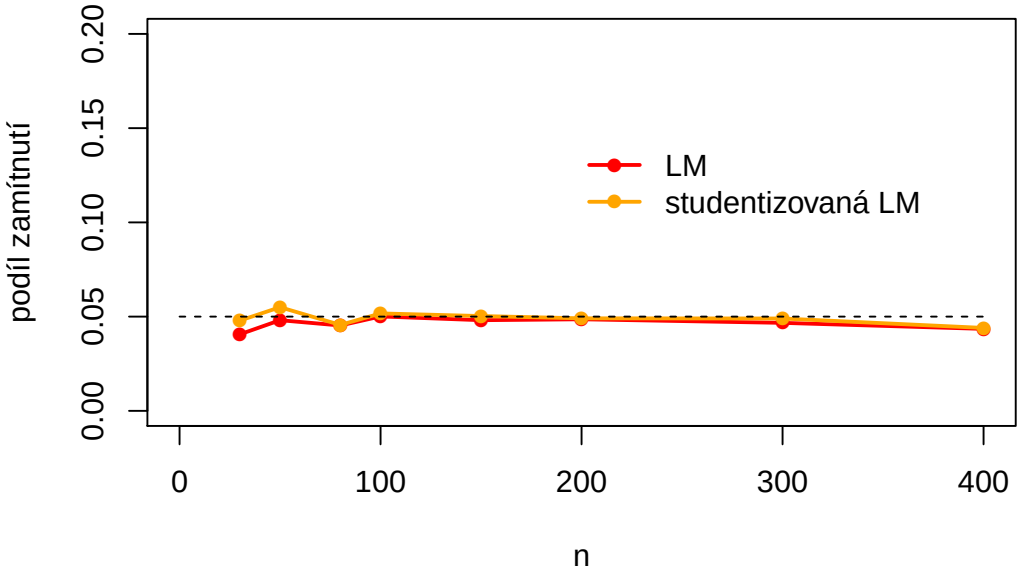
Pro jednotlivá rozdělení si tato data zakreslíme do grafu, abychom názorně viděli naše výsledky. Nejprve se podívejme na výsledky normálního rozdělení, viz obrázek 5.3. Vidíme, že v tomto případě statistika LM v obou případech vesměs dodržuje předepsanou asymptotickou hladinu 0.05.

Pro logaritmicko normální rozdělení to už neplatí. Na obrázku 5.4 můžeme vidět, že nestudentizovaná statistika překračuje kritickou hodnotu ve více jak polovině pozorování, a navíc s větším rozsahem výběru dokonce dosažená hladina roste. Na druhou stranu studentizovaná statistika už předepsanou hladinu 0.05 nepřekračuje, ale z tabulky 5.2 vidíme, že tato dosažená hladina se spíše blíží hodnotě 0.025.

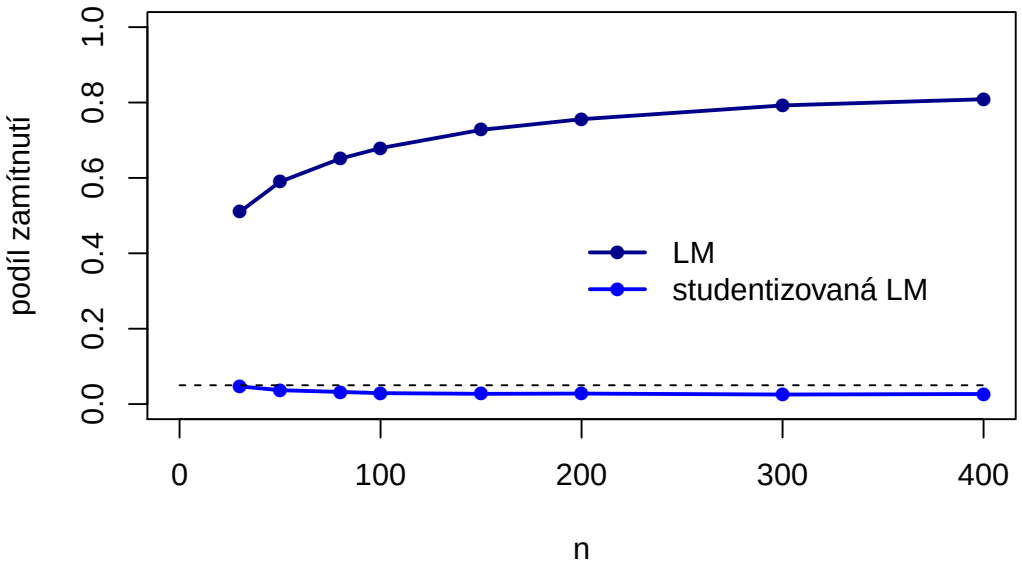
Studentovo rozdělení t_3 je na tom velmi podobně jako logaritmicko-normální rozdělení. Na obrázku 5.5 vidíme, že dosažená hladina u nestudentizované statistiky s větším n roste zhruba k hodnotě 0.45. Studentizovaná statistika už předepsanou hladinu 0.05 nepřekračuje, ale také je vidět, že se spíše blíží hodnotě 0.04 než hodnotě 0.05.

	Nestudentizovaná LM			Studentizovaná LM		
n	N	LN	t_3	N	LN	t_3
30	0.0406	0.5100	0.1844	0.0479	0.0469	0.0465
50	0.0481	0.5898	0.2440	0.0550	0.0369	0.0417
80	0.0453	0.6512	0.2947	0.0456	0.0322	0.0409
100	0.0501	0.6791	0.3145	0.0517	0.0290	0.0406
150	0.0481	0.7275	0.3505	0.0503	0.0272	0.0408
200	0.0486	0.7558	0.3866	0.0491	0.0279	0.0421
300	0.0467	0.7923	0.4158	0.0490	0.0254	0.0431
400	0.0435	0.8087	0.4396	0.0440	0.0268	0.0376

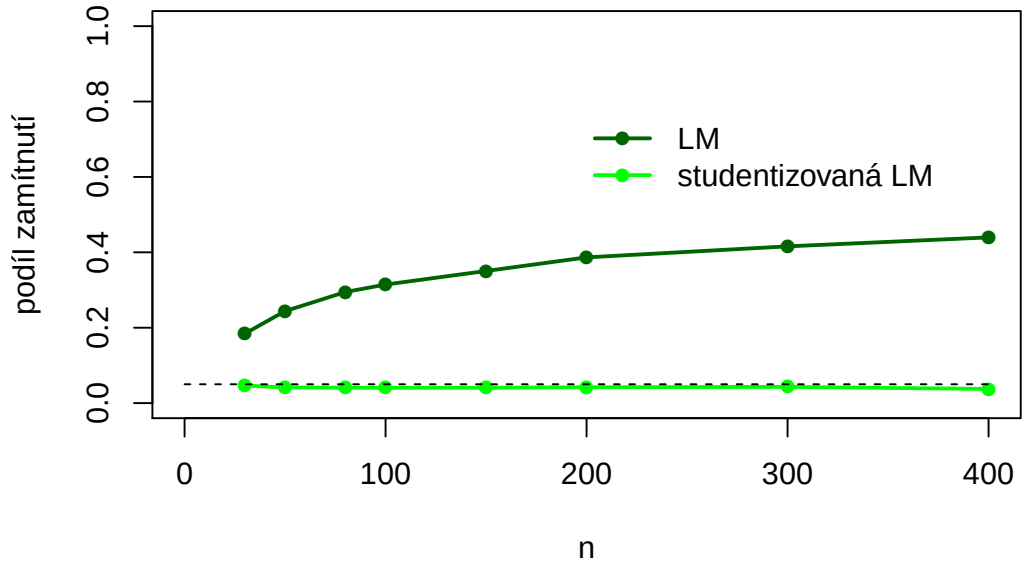
Tabulka 5.2: Dosažené hladiny testů pro jednotlivá rozdělení v závislosti na rozsahu výběru n.



Obrázek 5.3: Graf podílu zamítnutí při N rozdělení v závislosti na n.



Obrázek 5.4: Graf podílu zamítnutí při LN rozdělení v závislosti na n.

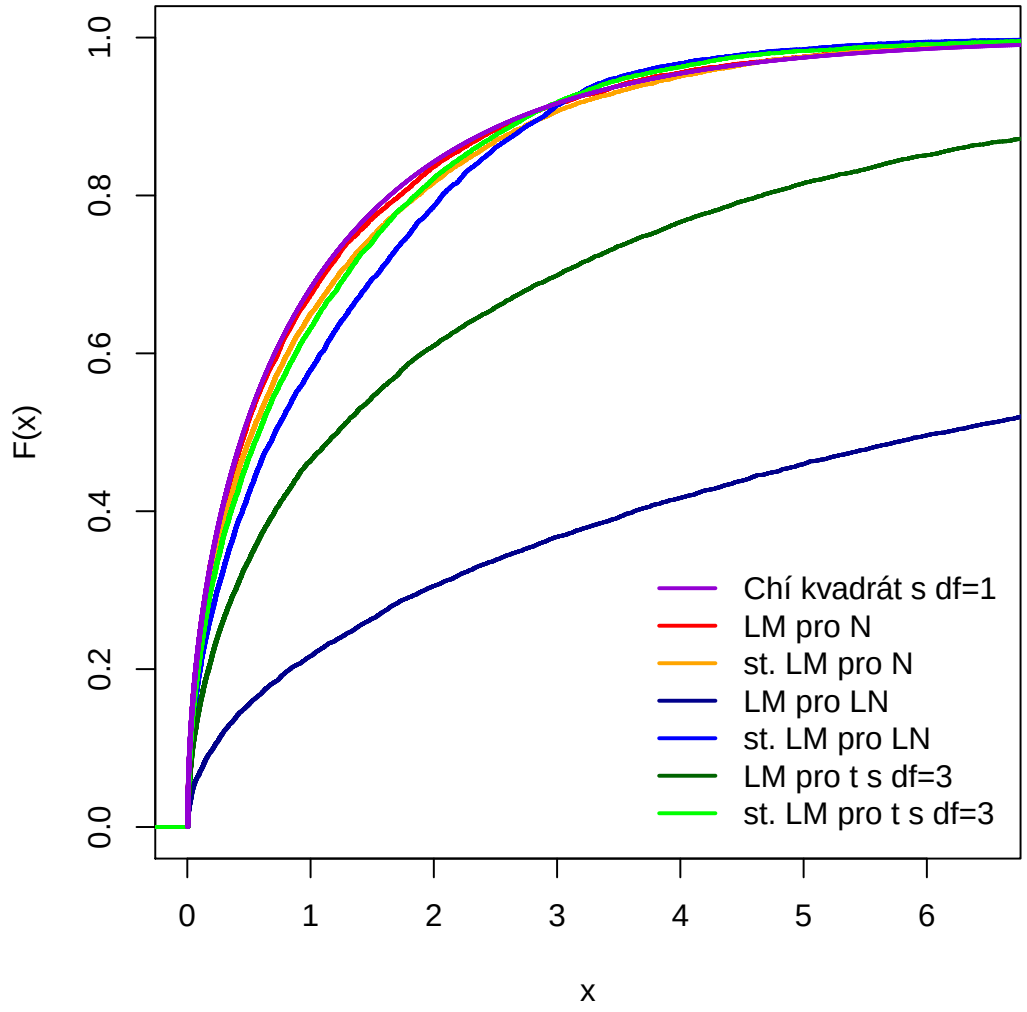


Obrázek 5.5: Graf podílu zamítnutí při $\mathtt{t}_3$ rozdělení v závislosti na n.

Můžeme se také podívat, jak vypadají empirické distribuční funkce těchto statistik v našich šesti případech. Tyto funkce si ale pro přehlednost vykreslíme pro jedinou hodnotu n. Na obrázku 5.6, který je vykreslen pro $n = 50$, vidíme, že pro data vygenerovaná z normálního rozdělení se graf empirické distribuční funkce statistiky LM podobá grafu distribuční funkce rozdělení χ^2_1. Pro nenormální data se studentizovanou statistikou už pozorujeme drobné odchylky a pro nestudentizovanou vidíme značný rozdíl.

Těmito simulacemi jsme se tedy přesvědčili, že za platného předpokladu normálního rozdělení naše testování homoskedasticity přibližně dodržuje předepsanou hladinu. Pokud tento předpoklad splněn není, tak LM sama o sobě dává nežádoucí výsledky, ale po její studentizaci jsou výsledky příznivější.

Mezi přílohami pod položkou 2) je uveden také vypracovaný skript pro prostředí **R**. Jeho modifikací lze zkoumat vlastnosti Breusch-Paganova testu i pro jiné parametry, než s jakými jsme pracovali zde.



Obrázek 5.6: Grafy empirických distribučních funkcí statistiky LM pro různá rozdělení při rozsahu výběru $n = 50$.

Závěr

Práce představila klasický lineární model a zobecnila jej na model heteroskedastický. K nezávisle proměnným (regresorům) bylo v těchto modelech přistupováno jako k náhodným vektorům. Následně jsme si připomněli zásadní poznatky z teorie maximální věrohodnosti a ukázali si, jak je aplikovat v lineárním modelu. Zejména jsme se potom zajímali o testy s rušivými parametry, které z teorie maximální věrohodnosti vycházejí. Na základě těchto teoretických podkladů jsme si odvodili dva základní testy předpokladu homoskedasticity. Nejprve jsme si zobecnili model analýzy rozptylu jednoduchého třídění ve smyslu heteroskedastického lineárního modelu a odvodili si test založený na věrohodnostním poměru pro shodnost rozptylů. Naši výslednou statistiku jsme porovnali s lehce modifikovanou statistikou navrženou Bartlettem. Na základě provedených simulací jsme se přesvědčili, že tato modifikace skutečně disponuje lepšími asymptotickými vlastnostmi než námi odvozená statistika. Ve druhém případě jsme zkoumali obecný normální heteroskedastický model, kde rozptyl chyb tohoto modelu závisel na doprovodných veličinách. Zde jsme po dlouhé řadě výpočtů došli ke skórové testové statistice a uvedli jsme její studentizovanou modifikaci. V následných numerických studiích jsme si ověřili její asymptotické vlastnosti a zkoumali, jak se tato statistika chová, není-li splněn předpoklad normálního rozdělení. Ukázalo se, že v tomto případě má její studentizovaná verze mnohem příznivější vlastnosti.

Literatura

Anděl, J. (2007). *Základy matematické statistiky.* Druhé opravené vydání. Matfyzpress, Praha. ISBN 978-80-7378-162-0.

Bartlett, M. S. (1937). Properties of Sufficiency and Statistical Tests. *Proceedings of the Royal Statistical Society*, **A 160**, 268–282.

Breusch, T. S. a Pagan, R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, **47**(5), 1287–1294.

Koenker, R. (1981). A Note on Studentizing a Test for Heteroscedasticity. *Journal of Econometrics*, **17**(1), 107–112.

R Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Zeileis, A. a Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, **2**(3), 7–10. URL `http://CRAN.R-project.org/doc/Rnews/`.

Zvára, K. (2008). *Regrese.* Druhé opravené vydání. Matfyzpress, Praha. ISBN 978-80-7378-041-8.

Seznam obrázků

Seznam tabulek

Přílohy

V kapitole 5 jsme pracovali s výpočetním prostředím **R**. Příslušná implementace našich provedených simulací (zvlášť pro Bartlettův test a zvlášť pro Breusch-Paganův test) je uvedena v následujících dvou skriptech, které jsou k dispozici ke stažení ze Studijního informačního systému.

Názvy jednotlivých souborů:

1) `skript-Bartlett.R`

2) `skript-Breusch-Pagan.R`