Statistika

(D360P03Z, D360P03U)

Karel Zvára

6. prosince 2004

připomenutí příkladu: klesá potratovost?

Y_i	24.7	25.7	31.6	24.3	26.8	30.6	21.1	23.5	26.9	22.5	23.1	24.9
Z_i	23.1	23.6	27.9	22.2	23.4	27.9	21.5	26.0	24.3	23.9	21.2	25.7
X_i	1.6	2.1	3.7	2.1	3.4	2.7	-0.4	-2.5	2.6	-1.4	1.9	-0.8
R_i^+	4	6	12	7	11	10	1	8	9	3	5	2

- použijeme údaje z 12 okresů v letech 2000 (Y_i) a 2001 (Z_i)
- hypotéza H_0 : v obou letech potratovost stejná, rozdíly dány náhodným kolísáním; H_1 : potratovost klesá (jednostranná alt.)
- za H_0 by rozdíly měly kolísat **symetricky kolem nuly**
- za H_1 by měly převládat kladné rozdíly, spíše velké
- průměrné pořadí z 8 kladných rozdílů: 8 (součet 64)
  průměrné pořadí ze 4 záporných rozdílů 3,5 (součet 14)

párový Wilcoxonův (Wilcoxon signed rank) test

- nechť $(Y_1, Z_1) \ldots, (Y_n, Z_n)$ nezávislé dvojice, $X_i = Y_i - Z_i$
- $\mathrm{H}_0 : Y_i, Z_i$ mají stejné rozdělení (populace jsou stejné)
- mají-li Y_i, Z_i stejné rozdělení, pak rozdíly $X_i = Y_i - Z_i$ jsou symetricky rozděleny kolem nuly
- postup
  - vyloučit nulové hodnoty X_i (tedy shodné hodnoty Y_i, Z_i), podle toho případně zmenšit n
  - určit pořadí R_i^+ **absolutních hodnot** $|X_i| = |Y_i - Z_i|$
  - určit W součet pořadí původně kladných hodnot X_i
  - podle W rozhodnout

rozhodování

- na základě centrální limitní věty lze použít

$$Z = \frac{W - \mathsf{E}\, W}{\mathsf{S.E.}(W)} = \frac{W - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

- hypotézu o shodě zamítneme, bude-li $|Z| \geq z(\alpha/2)$
- při jednostranné alternativě porovnat Z a $z(\alpha)$
- pro malý počet dvojic (do deseti) raději použít tabulky
- příklad ($W = 64, n = 12$, jinou metodou přesně je $p = 2{,}6$ %))

$$Z = \frac{64 - 12 \cdot 13/4}{\sqrt{12 \cdot 13 \cdot 25/24}} = 1{,}961 > 1{,}645 = z(0{,}05), p = 2{,}5 \text{ \%}$$

poznámky k výpočtu

- nezapomenout vyloučit nulové rozdíly
- shodným absolutním hodnotám rozdílům přiřadíme jejich průměrné pořadí:
- Excel 2000 řeší problém shod nestandardně (bohužel)
- jednoduchá ukázka

X_i	4	-2	5	2	-6	-4	2	7
$\|X_i\|$	4	2	5	2	6	4	2	7
R_i^+	4,5	2	6	2	7	4,5	2	8
Excel	4	1	6	1	7	4	1	8

párový znaménkový (sign) test

- hodnotí pouze **počet** kladných a záporných rozdílů, nezáleží na tom, jak jsou rozdíly veliké (slabší test než Wilcoxonův)
- $\mathrm{H}_0 : Y_i, Z_i$ mají stejné rozdělení; za hypotézy očekáváme, že počty kladných a záporných X_i jsou podobné
- označme Y počet kladných X_i z celkem n nenulových, za hypotézy $Y \sim \mathrm{bi}(n, 1/2)$
- přibližné rozhodování (centrální limitní věta)

$$Z = \frac{Y - n/2}{\sqrt{n/4}} = \frac{2Y - n}{\sqrt{n}}, \text{ zamítat pro } |Z| \geq z(\alpha/2)$$

- při jednostranné alternativě porovnáme Z a $z(\alpha)$

poznámky

- pro znaménkový test není třeba znát hodnoty Y_i, Z_i, stačí vědět, která z možností $Y_i > Z_i$, $Y_i < Z_i$, $Y_i = Z_i$ nastala
- náš příklad o možném poklesu potratovosti ($n = 12, Y = 8$)

$$Z = \frac{2 \cdot 8 - 12}{\sqrt{12}} = 1{,}155, \quad p = \mathsf{P}(Z > 1{,}155) = 1 - \Phi(1{,}155) = 0{,}124$$

- při malých hodnotách n (do 30) se doporučuje Yatesova korekce

$$Z_{\mathsf{Yates}} = \frac{|Y - n/2| - 1/2}{\sqrt{n/4}} \operatorname{sign}(Y - n/2) = \frac{|2Y - n| - 1}{\sqrt{n}} \operatorname{sign}(2Y - n)$$

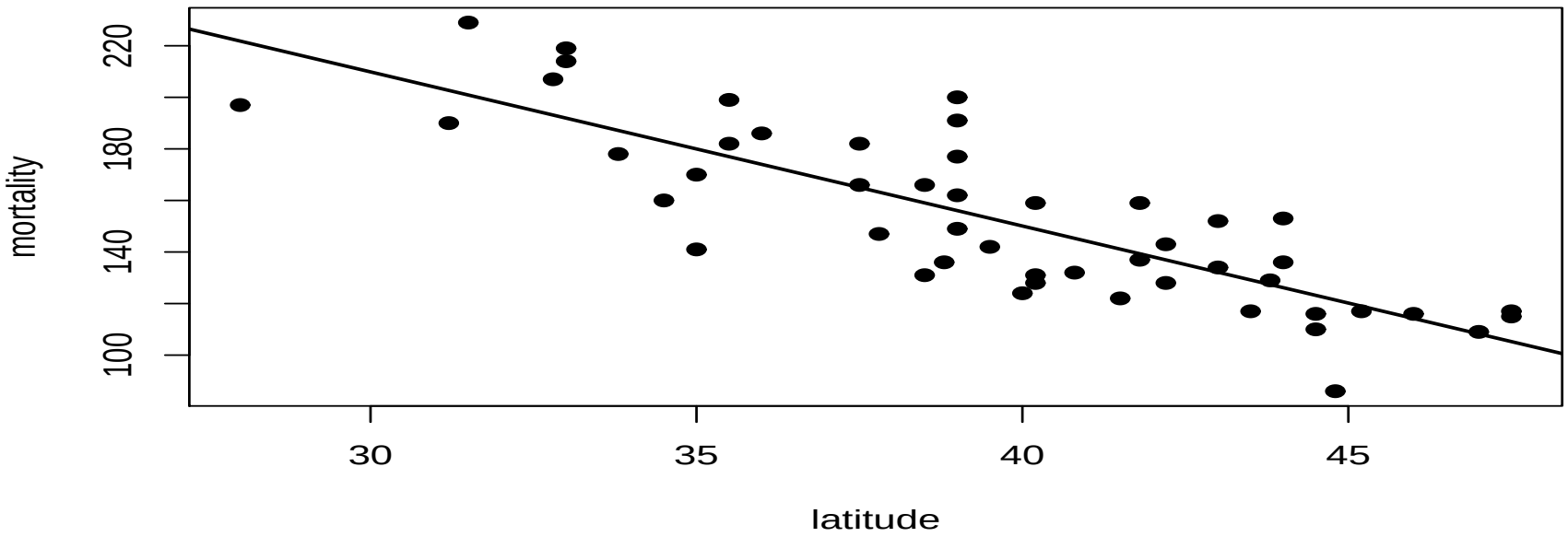
- náš příklad (Yatesova korekce, jiným způsobem přesně p =0,194)

$$Z = \frac{|2 \cdot 8 - 12| - 1}{\sqrt{12}} \cdot 1 = 0{,}866, \quad p = 1 - \Phi(0{,}866) = 0{,}193$$

Regrese

- na rozdíl od korelace (síla závislosti) hledáme tvar (způsob) závislosti, zajímá nás také průkaznost závislosti
- snažíme se z daných hodnot **regresorů (nezávisle proměnných)** předpovědět hodnoty **závisle proměnné (odezvy)**
- snažíme se variabilitu (kolísání hodnot) odezvy vysvětlit kolísáním regresorů
- prvně v tomto smyslu F. Galton (1886) při vyšetřování závislosti výšky synů na průměrné výšce rodičů: synové rodičů o dva palce vyšších než průměr všech rodičů byli v průměru jen o palec vyšší než průměr synů; dvoupalcová odchylka se nereprodukovala celá, byl patrný návrat (**regres**) k průměru

příklad: souvisí úmrtnost se zeměpisnou šířkou?



- úmrtnost na melanom na 10 000 000 obyvatel v státech USA

regresní přímka

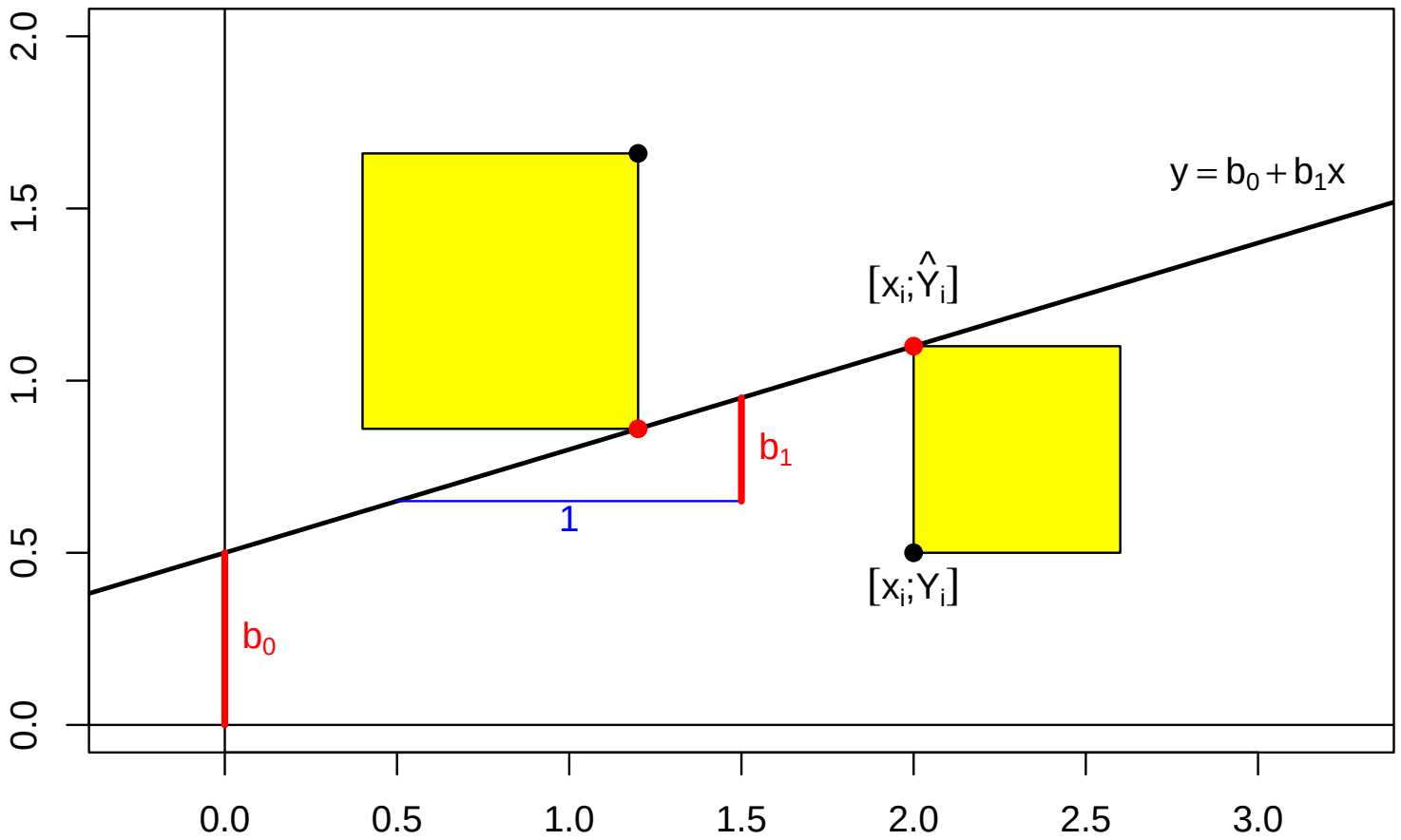
- chování Y (úmrtnost, mortality) vysvětlit lineární závislostí na x (zeměpisná šířka, latitude)
- každé zem. šířce odpovídá jakási střední úmrtnost, ta závisí na zeměpisné šířce lineárně

$$\mathsf{E}\, Y_i = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, n$$

- parametry β_0, β_1 odhadneme **metodou nejmenších čtverců** minimalizací přes β_0, β_1 součtu čtverců „svislých“ odchylek

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

- výsledné minimum (pro b_0, b_1) – **reziduální součet čtverců** S_e


y = b0 + b1x
[xi;Ŷi]
[xi;Yi]
b0
b1
1
0.0
0.5
1.0
1.5
2.0
2.5
3.0

náš příklad

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	389,19	23,81	16,34	<0,001
latitude	− 5,98	0,60	− 9,99	<0,001

- odhad závislosti est(mortality) = 389,2 − 5,98 latitude
- s každým stupněm sev. šířky klesá úmrtnost v průměru o 6 osob na 10 000 000 obyvatel
- na rovníku by úmrtnost měla být 389 jednotek, ale je to extrapolace mimo rozmezí známých hodnot − velmi nejisté
- závislost je průkazná, neboť v řádku pro x (latitude) je p <0,001

obecně

- odhadnutá závislost $y = b_0 + b_1 x$, modelová $y = \beta_0 + \beta_1 x$
- závislost na x prokazujeme testováním hypotézy $\mathsf{H}_0 : \beta_1 = 0$ proti oboustranné alternativě pomocí

$$T = \frac{b_1}{\mathsf{S.E.}(b_1)} = \frac{b_1}{s}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \text{zamítáme pokud } |T| \geq t_{n-2}(\alpha)$$

- **reziduální součet čtverců – nevysvětlená** variabilita odezvy $S_e = \sum_{i=1}^{n}(Y_i - (b_0 + b_1 x_i))^2$ reziduální rozptyl $s^2 = S_e/(n-2)$
- **koeficient determinace** ukazuje, jaký díl variability odezvy jsme závislostí vysvětlili

$$R^2 = 1 - \frac{S_e}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

náš příklad a tabulka analýzy rozptylu

variabilita	st. vol. f	součet čtverců SS	prům. čtverec MS	F	p
model	1	36 464,20	36 464,20	99,797	<0,001
reziduální	47	17 173,07	365,38		
celkem	48	53 637,27			

- kolísání úmrtnosti vysvětlíme závislostí z 68 %

$$R^2 = 1 - \frac{17173{,}07}{53637{,}27} = \frac{36464{,}20}{53637{,}27} = 0{,}680$$

- na 30. stupni očekáváme úmrtnost $389{,}19 - 5{,}98 \cdot 30 = 209{,}86$,
  na 40. stupni očekáváme úmrtnost $389{,}19 - 5{,}98 \cdot 40 = 150{,}08$

můžeme predikci zlepšit?

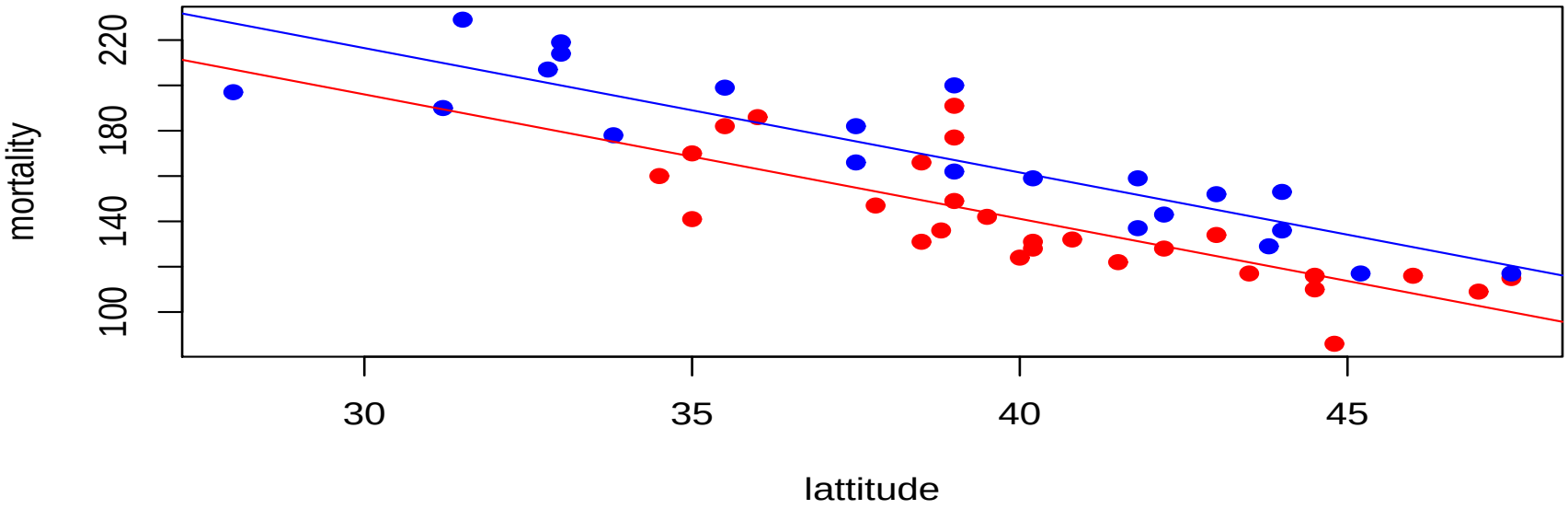
koef.	odhad	stř. chyba	t-stat.	p
abs. člen	401,17	28,04	14,31	$<$0,001
latitude	− 5,93	0,60	− 9,82	$<$0,001
longitude	0,15	0,19	0,82	0,418

- není průkazné, že by koeficient u longitude byl nenulový (nezamítneme hypotézu, že koeficient je nulový)
- koeficient determinace R^2 =0,684 (původně 0,680)

můžeme predikci zlepšit?

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	360,69	21,50	16,78	$<$0,001
ocean	20,43	4,83	4,23	$<$0,001
latitude	− 5,49	0,53	− 10,44	$<$0,001

- koeficient determinace R^2=0,770
- při „stěhování“ z vnitrozemí k oceánu po rovnoběžce roste úmrtnost v průměru o 20 osob na 10 milionů obyvatel
- je to ekvivalentní vnitrozemskému stěhování o 20,43/5,49 = 3,72 stupňů na jih
- na každý stupeň stěhování na sever klesá úmrtnost o 5,5, pokud se nezmění vztah k oceánu



- vnitrozemské státy: y=360,69-5,49 x
  přímořské státy: y=(360,69+20,43)-5,49 x =381,12-5,49 x
- lze ověřit, že přímky mohou být rovnoběžné (p =99,6 %)

pozor na interpretaci odhadů (příklad)

- závisí procento tuku dospělého muže na výšce?
  pokud ano, tak s výškou roste nebo klesá?
- závisí na tom, jak se na úlohu díváme, co bereme v úvahu
- est(fat) $= -$ 55,91 $+$ 0,391 height $\qquad R^2 = 14{,}0\ \%$
- est(fat) $=$ 13,29 $-$ 0,273 height $+$ 0,627 weight $\quad R^2 = 69{,}5\ \%$
- ve všech případech jsou koeficienty na 5% hladině průkazně nenulové
- rozdíl je v kvalitě vyrovnání, ale zejména v interpretaci
- průměrná změna procenta tuku při jednotkové změně výšky
  (a **nezměněné hmotnosti** pro druhý model)

regrese v MS Excelu 2000

	Excel 2000	označení
absolutní člen	Hranice	b_0
odhad	Koeficienty	b_i
střední chyba odhadu	Chyba střední hodnoty	S.E.(b_j)
koeficient (mnohonásobné) korelace	Násobné R	$\sqrt{R^2}$
koeficient determinace	Hodnota spolehlivosti R	R^2
adjustovaný koef. det.	Nastavená hodnota spol. R	R^2_{adj}
resid. směr. odchylka	Chyba střední hodnoty	s
počet pozorování	Pozorování	n
počet st. volnosti	Rozdíl	

obecné předpoklady

- **tvar závislosti**: známe jak vysvětlovaná veličina závisí na vysvětlujících
- **homoskedasticita**: pro všechny kombinace hodnot vysvětlujících veličin je rozptyl vysvětlované veličiny konstantní
- **nezávislost**: náhodné složky vysvětlovaných veličin jsou nezávislé
- **normalita**: náhodná složka má normální rozdělení
- předpoklady lze ověřovat (regresní diagnostika)
- někdy pomohou transformace