POZN: korelace x příčinnost

a) $x \to y$

$x \leftarrow y$

Příčinná spojitost (Causal link)

i když je příčinná spojitost mezi x a y, korelace samotná nám neříká, zda x ovlivňuje y nebo naopak

b)
$$\begin{tikzcd} & z \arrow[dl] \arrow[dr] & \\ x & & y \end{tikzcd}$$

Skrytá příčinnost (Hidden cause)

skrytá rel. z ovlivňuje x i y, což způsobuje jejich korelovanost

c)
$$\begin{tikzcd} & z \arrow[d] \\ x \arrow[r] & y \end{tikzcd}$$

~~Alt~~ Confounding factor

skrytá proměnná z i x ovlivňují y, výsledky tedy závisí i na z

d) $x \qquad y$

Coincidence

korelace je náhodná

POZN: korelace x příčinnost

a) $x \rightarrow y$

$x \leftarrow y$

Příčinná spojitost (Causal link)

i když je příčinná spojitost mezi x a y korelace samotná není určena, zda x ovlivňuje y nebo naopak

b)
$$\begin{array}{ccc} & z & \\ \swarrow & & \searrow \\ x & & y \end{array}$$

Skrytá příčinnost (Hidden cause)

skrytá vel. z ovlivňuje x i y, což způsobuje jejich korelovanost

c)
$$\begin{array}{ccc} z & & \\ & \searrow & \\ x & \rightarrow & y \end{array}$$

~~Ab~~ Confounding factor

skrytá proměnná z i x ovlivňují y, výsledky tedy závisí i na z

d) $x \quad y$

Coincidence

korelace je náhodná

Regrese bez počátku

existují případy, kdy přípustný model vyžaduje $\beta_0 = 0$, tj.

$$\underline{Y_i = \beta_1 x_i + e_i \,, \quad i = 1, \dots, n}$$

např.

1) je to předem známo na základě např. fyzikálních úvah

$(\; EY_0 = \beta_0 = 0)$

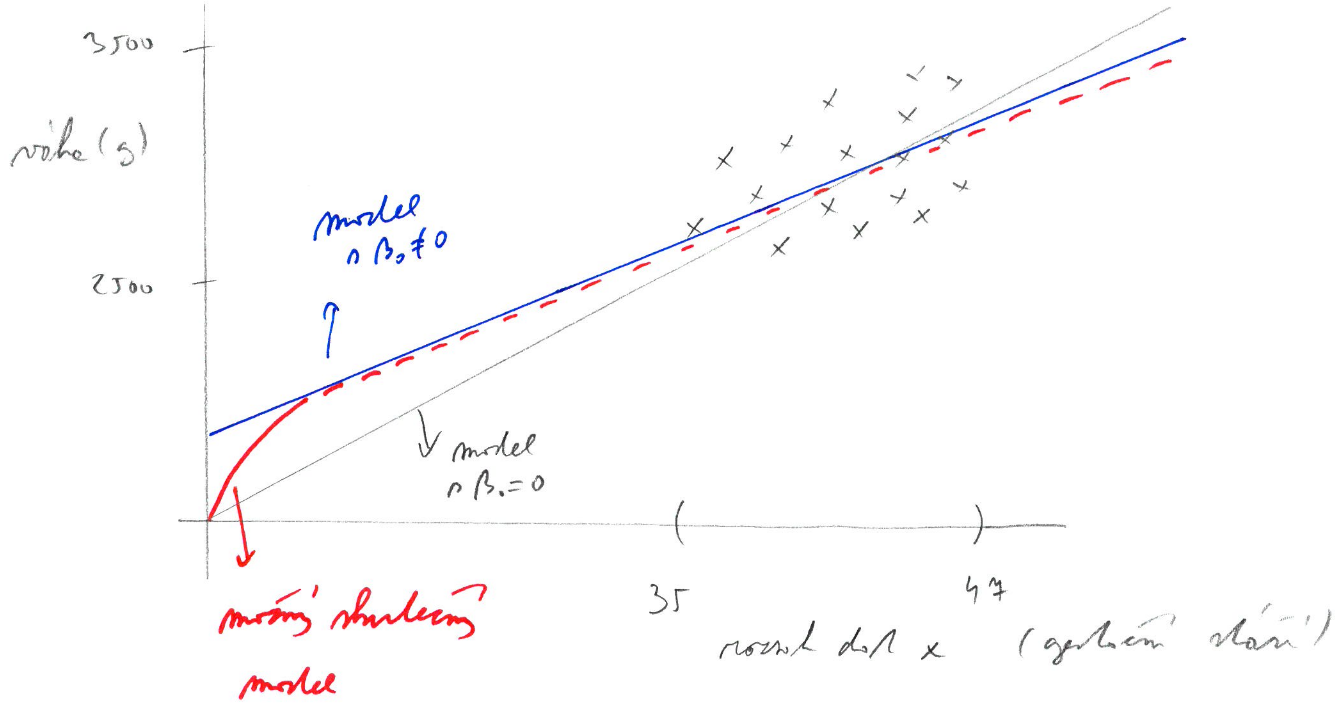
potom nemá smysl odhadovat β_0, protože to obecně snižuje přesnost odhadu σ^2 a tedy i β_1

2) na základě předpokládáme, že $\beta_0 \neq 0$ a t-test nezamítne $H_0 : \beta_0 = 0$, potom β_0 může být z modelu odstraněno

POZN: v praktických situacích si často nemůžeme být jisti, že model platí i blízko počátku. Část statistiků trvá na přítomnosti interceptu v modelu, i když je nevýznamný.

Položit $\beta_0 = 0$ apriorně, může být chybné i když $E(Y_0) = 0$.

Pokud totiž nevíme jistě, že model je lineární na okolí 0, volba $\beta_0 = 0$ může vést k vychýlenému odhadu β_1, pokud jsou nezávislé proměnné daleko od $x = 0$



Odhady a testy v případě $\beta_0 = 0$

LSE par. β_1 dostaneme minimalizací $S = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$

ve tvaru $$\hat\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

pokud $e_1, \dots, e_n$ i.i.d. $N(0, \sigma^2)$, pak $E\hat\beta_1 = \beta_1$ a $\operatorname{Var}\hat\beta_1 = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$

tedy $\hat\beta_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum x_i^2}\right)$ a

$$s_n^2 = \frac{\sum_{i=1}^{n} (y_i - \hat y_i)^2}{n-1} = \frac{SSE}{n-1}$$ je nestranný odhad σ^2

dále $\frac{SSE}{\sigma^2} \sim \chi^2(n-1)$ a nezávisí na $\hat\beta_1$

$H_0: \beta_1 = 0$ lze otestovat pomocí stat. $T = \frac{\hat\beta_1}{\frac{s_n}{\sqrt{\sum x_i^2}}} \sim t(n-1)$

$100(1-\alpha)\%$ IS pro β_1 je $\left(\hat\beta_1 \pm t_{1-\frac{\alpha}{2}}(n-1)\dfrac{s_n}{\sqrt{\sum x_i^2}}\right)$

- zbytek vše podobné jako pro případ $\beta_1 \neq 0$,
- rozdíl je ale v tabulce ANOVA a v míře dobré shody, problém je, že neplatí rozklad $\underline{SST = SSR + SSE}$ neboť součet reziduí $\sum_{i=1}^{n}(y_i - \hat y_i)$ nemusí být 0 a tedy

  $\overline{\hat y}_n \neq \bar y_n$
- odvodíme nový rozklad, který platí v obou případech, dokážeme ho ale jen pro $\beta_0 = 0$

<u>Věta:</u> V modelu s $\beta_0 = 0$ platí

$$\underline{\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat y_i^2 + \sum_{i=1}^{n} (y_i - \hat y_i)^2}$$

Dk:

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (y_i - \hat y_i + \hat y_i)^2 = \sum_{i=1}^{n} (y_i - \hat y_i)^2 + \sum_{i=1}^{n} \hat y_i^2 + 2\underbrace{\sum_{i=1}^{n}(y_i - \hat y_i)\hat y_i}_{\overset{?}{=}0}$$

z rovnice $\cancel{\frac{\partial S}{\partial \beta_1}} = \sum_{i=1}^{n} (y_i - \underbrace{\hat\beta_1 x_i}_{=\hat y_i}) x_i = 0 \quad | \cdot \hat\beta_1$

$\frac{\partial S}{\partial \beta_1} = 0$ dostaneme

$$\sum_{i=1}^{n} (y_i - \hat y_i) \cdot \hat y_i = 0 . \qquad \text{Q.E.D.}$$

Pokud vezmeme $\sum y_i^2$ jako míru variability v datech, analogie R^2 def. bude

$$\underline{R^2 = \frac{\sum_{i=1}^{n} \hat y_i^2}{\sum_{i=1}^{n} y_i^2}}$$

$$1 - R^2 = \frac{\sum y_i^2 - \sum \hat{y}_i^2}{\sum y_i^2} = \frac{\sum \hat{e}_i^2}{\sum y_i^2}$$

definujme
$$F = \frac{(n-1)R^2}{1-R^2}$$
potom

$$F = \frac{\sum \hat{y}_i^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \frac{\hat{\beta}_1^2 \sum x_i^2}{s_n^2} = T^2$$

vztah mezi R^2, F a T^2 je tedy stejný jako pro $\beta_0 \neq 0$.

POZN: tato definice R^2 se ale v praxi moc nepoužívá, protože neumožňuje přímé srovnání modelu bez a s interceptem

$\beta_0 = 0$:
$$R^2 = 1 - \frac{SSE}{\sum_{i=1}^{n} y_i^2}$$

$\beta_0 \neq 0$:
$$R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y}_n)^2}$$

obecně ale $\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 < \sum_{i=1}^{n} y_i^2$, R^2 v modelu s $\beta_0 = 0$

tedy bude větší než R^2 modelu s $\beta_0 \neq 0$ i když jsou jejich SSE srovnatelné

- definice vhodné R^2 pro $\beta_0 = 0$ vyvolává jistou kontroverzi a existuje několik verzí
- možná volba je $R^2 = (\rho(y, \hat{y}))^2$ $\quad \hat{y} = (\hat{y}_1, \dots, \hat{y}_n)$

protože tato vlastnost platí i pro případ $\beta_0 \neq 0$

- další možnost je srovnat modely pomocí hodnot s_n^2 (preferuje se model s nižší hodnotou s_n^2)

Tabulka ANOVA pro $\beta_0 = 0$

Source	df	SS	MS	F
Regression	1	$SSR = \sum \hat{y}_i^2$	$MSR = \frac{SSR}{1}$	$\frac{SSR}{s_n^2}$
Residual	$n-1$	$SSE = \sum (y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-1} = s_n^2$	
Total	n	$SST = \sum y_i^2$		

$$R^2 = \rho^2(y, \hat{y})$$

Př. –

Predikce

jakmile máme model, často bývá cílem odhadnout hodnoty vel. Y_0 pro nové x_0, které není v původních datech

budeme uvažovat dva typy predikce:

1) predikce střední hodnoty $\mu_0 = E(Y_0)$ v bodě x_0

2) predikce hodnoty nového pozorování Y_0 v bodě x_0

pro oba typy použijeme bodový odhad

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

intervalové odhady se ale budou lišit

ad 1) protože je $\mu_0 = \beta_0 + \beta_1 x_0$ vlastní parametr, lze pro něj odvodit IS (za předpokladu normality chyb)

spočteme $Var(\hat{Y}_0)$

dosazením odhadů $\underset{\bar{y} - \hat{\beta}_1 \bar{x}}{\hat{\beta}_0}, \hat{\beta}_1$ dostaneme $\hat{Y}_0 = \bar{y} + \hat{\beta}_1 (x_0 - \bar{x})$

a $$\underline{Var\,\hat{Y}_0} = Var(\bar{Y}) + (x_0 - \bar{x})^2\, Var(\hat{\beta}_1) + 2(x_0 - \bar{x}) \underbrace{Cov(\bar{Y}, \hat{\beta}_1)}_{=0 \text{ ukázali jsme v dk. věty}}$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{S_{xx}} = \underline{\sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}$$

nahrazením σ^2 statistikou s_n^2 dostaneme odhad $Var(\hat{Y}_0)$ ve tvaru

$$\hat{\sigma}^2(\hat{Y}_0) = s_n^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

$\hat{\sigma}(\hat{Y}_0)$ se obvykle nazývá <u>standardní chyba predikce v bodě x_0</u>

Jsou-li $e_1, \dots, e_n$ i.i.d. $N(0, \sigma^2)$, platí

$$\hat{Y}_0 \sim N\Big(\mu_0, \underbrace{\sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}_{\sigma^2(\hat{Y}_0)}\Big) \quad \text{a tedy}$$

$$\underline{\frac{\hat{Y}_0 - \mu_0}{\sigma(\hat{Y}_0)} \sim N(0,1)}$$

Celkem:

$$\underline{T} = \frac{\dfrac{\hat{Y}_0 - \mu_0}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}}}{\sqrt{\frac{(n-2) s_n^2}{\sigma^2} \cdot \frac{1}{n-2}}} = \frac{\hat{Y}_0 - \mu_0}{\sqrt{s_n^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}} = \underline{\frac{\hat{Y}_0 - \mu_0}{\hat{\sigma}^2(\hat{Y}_0)} \sim t(n-2)}$$

Vyjádření získané $100(1-\alpha)\%$ IS pro μ_0 ve tvaru

$$\hat{Y}_0 \pm t_{1-\frac{\alpha}{2}}(n-2)\cdot \hat{\sigma}(\hat{Y}_0)$$

POZN: z tvaru IS je vidět, že bude nejužší pro $x_0 = \bar{x}$

a s rostoucí vzdáleností $|x_0 - \bar{x}|$ se prodlužuje

– spec. čím dále jsme od oblasti, kde ~~jsou~~ máme data x, tím méně spolehlivá je naše predikce

– je třeba opatrnosti při predikci hodnot y mimo interval $(\min x_i, \max x_i)$

ad 2) intervalové odhady pro Y_0 nejsou IS, protože Y_0 není parametr; říká se jim intervaly predikce

potřebujeme rozptyl $Y_0 - \hat{Y}_0$, protože je nové pozorování Y_0 nezávislé na Y_i, ~~...~~ $i = 1, \dots, n$, platí

$$\mathrm{Var}(Y_0 - \hat{Y}_0) = \underbrace{\mathrm{Var}\, Y_0}_{=\sigma^2} + \mathrm{Var}\, \hat{Y}_0 + 0 =$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

odhad tohoto rozptylu bude s_f^2, kde

$$s_f = s_n \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Za předpokladu normality chyb platí

$$\underline{T} = \frac{Y_0 - \hat{Y}_0}{s_n \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} = \underline{\frac{Y_0 - \hat{Y}_0}{s_f} \sim t(n-2)}$$

Vyjádřením T získáme $\underline{100(1-\alpha)\%\ \text{interval predikce pro } Y_0}$

$$\underline{\hat{Y}_0 \pm t_{1-\frac{\alpha}{2}}(n-2)\, s_f}$$

POZN: přesnost predikce

a) roste s rostoucím n a rostoucím rozsahem x měřeným pomocí S_{xx}

b) klesá s rostoucím $|x_0 - \bar{x}|$

- pokud můžeme předem zvolit x_i, $i = 1, \dots, n$, lze přesnost predikce zvýšit ~~volbou~~ volbou dostatečně rozptýlených hodnot x, to ale může zvyšovat R^2 a nakonec vést k horšímu modelu

⇒ základní rozpor v regresní analýze

- dobrý model nemusí poskytovat dobré predikce
- dobré predikce mohou vycházet z méně přesných modelů.

POZN: odvozené výsledky platí za předpokladu normality chyb, protože ale za podmínek regularity jsou odhady $\hat{\beta}_0, \hat{\beta}_1$ as. normální, IS pro $E(Y_0)$ budou fungovat (použitelné) pro velká n, IP pro Y_0 ale závisí na normalitě chyb i pro velká n, mohou tedy být nepřesné pro nenormální chyby

Př. Ex 3.20

Ověření adekvátnosti modelu

- je to důležitá součást analýzy
- měla by být provedena dříve, než budeme interpretovat parametry modelu nebo přijímat nějaké závěry založené na modelu

- všechny výsledky ~~...~~ týkající se β_0, β_1 byly odvozeny za předpokladu linearity modelu a některé za předpokladu normality chyb

Bylo by dobré mít testy ověřující linearitu

Základní procedury:

1) prozkoumání scatter plotu dvojic (x_i, y_i)

např.

může indikovat, že lepší model bude

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$$

nicméně scatter plot může být zavádějící, pokud je ~~odchyl~~ odchylka od linearity způsobena spíše chybějící proměnnou než polynomiální závislostí na x

2) analýza hodnot testovacích statistik

např. • malá hodnota R^2 společně s významnou hodnotou t-statistiky pro par. β_1 obecně naznačuje, že skutečný model obsahuje i jiné proměnné x

- velká hodnota R^2 a významná t-statistika
  ale samo o sobě neznamená, že je model lineární

3) obrázky reziduí – efektivní diagnostický nástroj

rezidua odhadují "kolik variability v datech zůstane po odstranění lineární části v x", dá se tedy očekávat, že jejich hodnoty budou užitečné pro detekci odchylek od normality

POZN: analýza scatter plotu a obrázku reziduí je dost subjektivní, bylo by dobré mít nějaký objektivní analytický nástroj pro ověření linearity modelu

bohužel nejsou třeba žádné k dispozici a pro většinu dat jsou v praxi nejvíce využívané metody 1) – 3)

- jinak je tomu u navrženého experimentu typu industriálních nebo klinických studií, kde existuje doporučený analytický test „lack of fit" test (LOFT)

předpoklád, že máme více pozorování pro jedno x_i

ad 3) Analýza reziduí

intuitivně, pokud je náš model správný, měla by se rezidua chovat jako náhodný výběr z $N(0, \sigma^2)$

pokud se bude zdát, že se tak nechovají, ~~tedy~~ ~~může~~ bude to znamenat neadekvátnost modelu

ukázeme grafické nástroje, zkoumat ale vlastnosti ~~náhodné~~ reziduí

Věta: Nechť $\hat{e}_i$ jsou rezidua modelu (*) odhadnutého metodou nejmenších čtverců.
Potom platí:

1) $E(\hat{e}_i) = 0 \qquad i = 1, \dots, n$

2) $Var(\hat{e}_i) = \sigma^2_{\hat{e}_i} = \sigma^2 \left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right] \approx \sigma^2$ pro velká n

3) $Cov(\hat{e}_i, \hat{e}_j) = -\sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_i)(\bar{x} - x_j)}{S_{xx}} \right]$

4) $Cov(\hat{e}_i, \hat{y}_i) = 0 \qquad i = 1, \dots, n$

5) Pokud jsou $e_1, \dots, e_n$ i.i.d. $N(0, \sigma^2)$, platí

$$\hat{z}_i = \frac{\hat{e}_i}{\sigma_{\hat{e}_i}} \sim N(0,1).$$

Dk: 1) $\hat{e}_i = y_i - \hat{y}_i$, tedy $E(\hat{e}_i) = Ey_i - E\hat{y}_i$, ale

$E\hat{y}_i = E(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \beta_0 + \beta_1 x_i = E y_i$

nestranné odhady

$\sigma^2 \left[\frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}} \right]$

$\|$

2) $Var\, \hat{e}_i = Var(y_i - \hat{y}_i) = Var\, y_i + Var\, \hat{y}_i - 2\, Cov(y_i, \hat{y}_i) =$

$\|$

$\sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right]$

dk. věty o nestrannosti $\hat{\sigma}^2$

$= \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}} \right]$

3) $$\mathrm{Cov}(\hat{e}_i, \hat{e}_j) = \mathrm{Cov}(Y_i - \hat{Y}_i, Y_j - \hat{Y}_j) = \underbrace{\mathrm{Cov}(Y_i, Y_j)}_{=0} - \mathrm{Cov}(Y_i, \hat{Y}_j)$$
$$- \mathrm{Cov}(Y_i, \hat{Y}_j) + \mathrm{Cov}(\hat{Y}_i, \hat{Y}_j)$$

$$\underline{\mathrm{Cov}(\hat{Y}_i, \hat{Y}_j)} = \mathrm{Cov}(\hat{\beta}_0 + x_i \hat{\beta}_1, \hat{\beta}_0 + x_j \hat{\beta}_1) =$$
$$= \mathrm{Var}(\hat{\beta}_0) + (x_i + x_j)\,\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) + x_i x_j \mathrm{Var}(\hat{\beta}_1) = *$$

viz výše:
$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right), \quad \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}}, \quad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$* = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} - \frac{(x_i + x_j)\bar{x}}{S_{xx}} + \frac{x_i x_j}{S_{xx}}\right) = \underline{\sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}\right)}$$

podobně bychom dostali

$$\mathrm{Cov}(Y_i, \hat{Y}_j) + \mathrm{Cov}(Y_j, \hat{Y}_j) = 2\sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}\right)$$

tedy
$$\mathrm{Cov}(\hat{e}_i, \hat{e}_j) = -\sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}\right).$$

4) $$\mathrm{Cov}(\hat{e}_i, \hat{Y}_i) = \mathrm{Cov}(Y_i - \hat{Y}_i, \hat{Y}_i) = \mathrm{Cov}(Y_i, \hat{Y}_i) - \mathrm{Var}(\hat{Y}_i) = 0$$
$$\mathrm{Cov}(Y_i, \hat{Y}_i) = \sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right), \qquad \mathrm{Var}(\hat{Y}_i) = \mathrm{Cov}(\hat{Y}_i, \hat{Y}_i) = \sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)$$

5) $e_i \sim N(0, \sigma^2) \Rightarrow \hat{e}_i \sim N(\cdot, \cdot)$ protože $\hat{e}_i$ je ~~[illegible]~~ LK $Y_1, \dots, Y_n$

$$\left.\begin{aligned} 1) &\Rightarrow E\hat{e}_i = 0 \\ 2) &\Rightarrow \mathrm{Var}\,\hat{e}_i = \sigma^2_{\hat{e}_i} \end{aligned}\right\} \Rightarrow \frac{\hat{e}_i}{\sigma_{\hat{e}_i}} \sim N(0,1)$$

Q.E.D.

Pozn: z bodu 3) věty plyne, že $\mathrm{Cov}(\hat{e}_i, \hat{e}_j) \approx 0$ pro velká n

pokud jsou tedy e_i i.i.d. $N(0, \sigma^2)$, měla by se standardizovaná residua $\hat{z}_i = \frac{\hat{e}_i}{\sigma_{\hat{e}_i}}$ chovat pro velká n jako náhodný výběr z $N(0,1)$ rozdělení

v praxi ale budeme potřebovat odhad σ^2 pro výpočet $\hat{z}_i$

nejznámnější procedura: odhadnout σ^2 pomocí s_n^2, potom

$$\hat{r}_i = \frac{\hat{e}_i}{s_n \sqrt{1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)}}$$ studentizovaná residua

opět, pro velká n by se $\hat{r}_i$ měla chovat jako n. v. z $N(0,1)$

Pozn: $\hat{e}_i$ se užívají pro grafickou analýzu

jiná třída residuí – PRESS residua – neporušitelné metody zkoumání residuí

ozn. $\hat{\beta}_{0(-i)}$, $\hat{\beta}_{1(-i)}$ odhady parametrů β_0, β_1 pokud je vynecháno i-té pozorování

i-té PRESS residuum je def. jako

$$\hat{e}_{(-i)} = \hat{y}_i - \hat{y}_{(-i)}, \quad \text{kde} \quad \hat{y}_{(-i)} = \hat{\beta}_{0(-i)} + x_i \hat{\beta}_{1(-i)}$$

podrobněji se jim budeme věnovat později

Grafy reziduí

- **histogram reziduí** (náhled normality reziduí)

- **kvantilový graf** (Q-Q plot) studentizovaných reziduí

  seřadíme dle velikosti: $\hat{r}_{(1)} \le \hat{r}_{(2)} \le \dots \le \hat{r}_{(n)}$

  a vyneseme oproti: $\Phi^{-1}\left((i-\frac{1}{2})\frac{1}{n}\right)$, $i = 1, \dots, n$

  body by měly ležet přibližně na přímce

  ($E(e_{(i)}) \approx \Phi^{-1}((i-\frac{1}{2})\frac{1}{n})$ pro normální chyby)

  použití: ověření normality, detekce odlehlých pozorování

(obr. 3.6 str. 107 GLM)

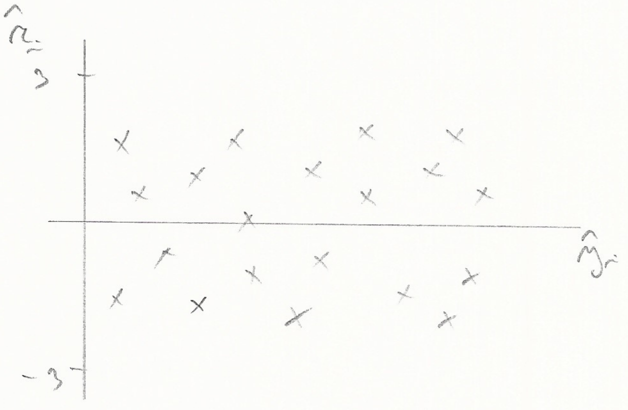
- **studentizované reziduum × jednotlivým vysvětlujícím proměnným x**

  $\hat{r}_i$ nezávisí na σ, graf $\hat{r}_i \times x_i$ lze použít pro detekci nelinearity nebo nekonstantního rozptylu

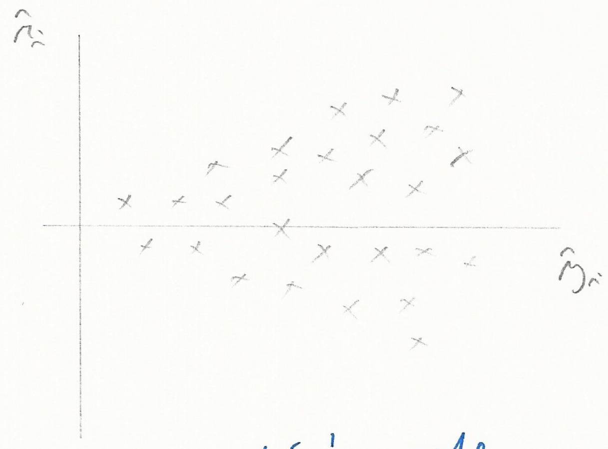
- **studentizované reziduum $\hat{r}_i$ × predikovaným hodnotám $\hat{y}_i$**

  $Cov(\hat{e}_i, \hat{y}_i) = 0$, tedy $\hat{e}_i (\hat{r}_i)$ a $\hat{y}_i$ by měly být nekorelované pokud platí model (*)

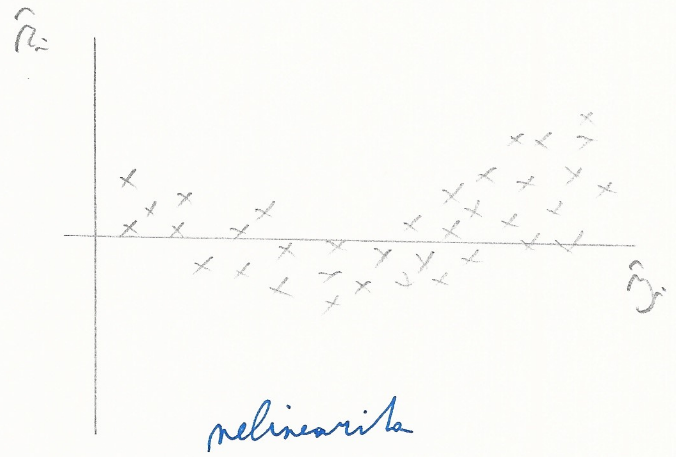
  tzn. graf $\hat{r}_i \times \hat{y}_i$ by měl být náhodně rozptýlený kolem osy x, navíc $\hat{r}_i$ by měla ležet v $(-3, 3)$ ($\hat{r}_i \approx N(0,1)$)



OK



nestejný rozptyl



nelinearita

- studentizovaná rezidua × pořadí pozorování

možná detekce řádové korelace mezi poz.