Jednorozměrná lin. regrese

předp., že sledujeme dvě veličiny x a y mezi kterými ex. lin. závislost

$$y = \beta_0 + \beta_1 x \qquad \beta_0, \beta_1 \text{ neznámé}$$

provede se experiment a zjišťují se hodnoty dvojic (x, y)

často je měření x prakticky zcela přesné (např. x se nastavuje na předem dané úrovně)

měření y se měří s chybou

chyba může být náhodná a y tedy budeme chápat jako náh. vel. a budeme značit Y

pro dvojice $(x_1, Y_1), \dots, (x_n, Y_n)$ se zavádí model

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + e_i,} \quad (*) \; i = 1, \dots, n$$

kde:

- Y_i - vysvětlovaná (závislá) proměnná
- x_i - vysvětlující (nezávislá) proměnná
  prediktor, regresor
- β_0, β_1 - neznámé regresní parametry
- e_i - náhodný šum, (náh. chyba)

  budeme předpokládat, že e_i jsou nezávislé (někdy bude stačit nekorelované)

  a $e_i \sim (0, \sigma^2)$

  tzn. $E e_i = 0$, $Var(e_i) = \sigma^2 \;\; \forall i$ (homoskedasticita) měřením získáme data $(x_1, y_1), \dots, (x_n, y_n)$

a cílem stat. analýzy je určit, zda model $(*)$ je schopen popsat pozorovanou variabilitu v y

první krok: odhad neznámých parametrů $\beta_0, \beta_1, \sigma^2$

(proložení dat přímkou)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

a porovnání $\hat{y}_i$ a y_i $\forall i$ umožňuje posoudit adekvátnost modelu

proložení dat přímkou: několik způsobů, zásadní bude souvislost rozdělení e_i a testy $\forall i$

i když apriori není zřejmé proč znát rozdělení a ne β_0, β_1

Dvě možnosti:

1) odhadnout β_0, β_1 pomocí metody nezávisející na rozdělení chyb

2) udělat vhodný předpoklad o rozdělení chyb, odhadnout β_0, β_1 a potom ověřit předpoklad

POZN: spec. důležitý případ je $e_i \sim N(0, \sigma^2)$ jež umožňuje získání MLE odhadů β_0, β_1 vede na metodu nejmenších čtverců, která může být použita bez ohledu na rozdělení chyb

Odhady parametrů

a) předpokládáme, že $e_1, \dots, e_n$ i.i.d. $N(0, \sigma^2)$

tzn. $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ a $Y_1, \dots, Y_n$ nezávislé

+ alternativní zápis modelu
$EY_i = \beta_0 + \beta_1 x_i = \mu_i$

MLE odhady:

věr. fce.: $$L = L(\beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}$$

$$l = \ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

pro pevné $\sigma^2 > 0$ je maximalizace l $\Leftrightarrow$ minimalizaci S

$$\underline{S} = S(\beta_0, \beta_1) = \underline{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}$$

proto metoda nejmenších čtverců

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\, x_i = 0$$

1. rovnice: $\sum_{i=1}^{n} y_i - \beta_1 \sum x_i = n\beta_0 \implies \underline{\beta_0} = \frac{1}{n}\sum y_i - \beta_1 \frac{1}{n}\sum x_i =$

$= \underline{\bar{y}_n - \beta_1 \bar{x}_n}$

2. rovnice: $\sum y_i x_i - \beta_0 \sum x_i - \beta_1 \sum x_i^2 = 0$

$\sum y_i x_i - \bar{y}\sum x_i + \beta_1 \bar{x}\sum x_i - \beta_1 \sum x_i^2 = 0$

$$\hat\beta_1 = \frac{\sum_{i=1}^n x_i y_i - n\,\bar x_n \bar y_n}{\sum x_i^2 - n\,\bar x_n^2} \qquad \hat\beta_0 = \bar y_n - \hat\beta_1 \bar x_n$$

---

odhad σ^2:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 = 0$$

a tedy $$\hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat\beta_0 - \hat\beta_1 x_i)^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat y_i)^2 = \frac{1}{n}\,SSE$$

sum of the squares of errors

kde $\hat y_i = \hat\beta_0 + \hat\beta_1 x_i$ je „predikce" modelu

odhad $E Y_i$

rozdíl $\hat e_i = y_i - \hat y_i$ i-té reziduum

velikost reziduí indikuje, jak dobře odhadnutá přímka odpovídá datům

rezidua jsou vlastně odhady chyb e_i, jejich analýza hraje významnou roli v ověření předpokladů rozdělení chyb

Pozn.: pro odhad σ^2 se používá častěji statistika

$$s_n^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat y_i)^2 = \frac{1}{n-2}\,SSE$$

která je nestranným odhadem σ^2 (pro lib. rozdělení e_i)
zatímco σ^2_{MLE} je vychýlený odhad – pro normální rozdělení chyb

odhad σ: $$s_n = \sqrt{\frac{1}{n-2}\sum (y_i - \hat y_i)^2}$$ standardní chyba regrese (standard error)

není nestranný

b) bez předpokladu normality chyb

tzn. $e_1, \dots, e_n$ nekorelované, $e_i \sim (0, \sigma^2)$

pro odhad β_0, β_1 lze použít minimalizaci S

což je rozumná procedura, která si nezachovává geometrickou interpretaci

nechť $y = \beta_0 + \beta_1 x$ je rovnice nějaké přímky

potom $y_i - (\beta_0 - \beta_1 x_i)$ je vertikální vzdálenost bodu (x_i, y_i) od přímky a

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

je míra, jak dobře přímka prokládá data

dává smysl vybrat takovou přímku, která minimalizuje S

minimalizace S získáme stejné odhady $\hat{\beta}_0, \hat{\beta}_1$ jako v MLE pro normální data

těm se ale označují **least squares estimators (LSE)** odhady met. nejmenších čtverců

existuje více měr vhodnosti přímky

použití LSE pro lib. rozdělení chyb má dvě zdůvodnění:

1) pro normální ~~data~~ chyby LSE splývá s MLE

2) LSE odhad je navíc BLUE (best linear unbiased estimator)

jak ukázáno v Gauss-Markov theorem

Př. nechť $e_1, \dots, e_n$ jsou i.i.d. s hustotou

$$f(\varepsilon) = \frac{1}{2} e^{-|\varepsilon|}$$

(Laplace distr. double exponential)

potom hustota Y_i je

$$f_{Y_i}(y_i) = \frac{1}{2} e^{-|y_i - \beta_0 - \beta_1 x_i|}$$

a věr. fce

$$L = \frac{1}{2^n} e^{-\sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i|}$$

$$l = -n \log 2 - \sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i|$$

MLE par. β_0, β_1 dostaneme minimalizací

$$A = \underline{\sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i|}$$

MAD
minimum absolute deviations

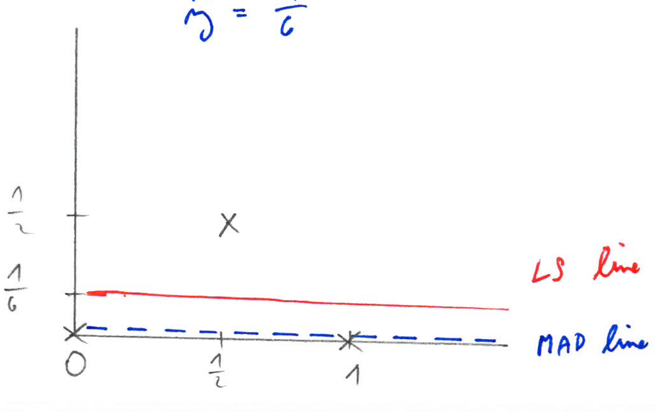
odhady budou jiné než LSE

uvažujme 3 body: $(0,0)$, $(1,0)$, $(\frac{1}{2}, \frac{1}{2})$

MLE: $\hat\beta_0 = \hat\beta_1 = 0$ $\quad A = \frac{1}{2}$ $\quad \hat y = 0$

LSE: $\bar x = \frac{1}{2}$ $\quad \bar y = \frac{1}{6}$ $\quad \sum x_i^2 = \frac{5}{4}$ $\quad \sum x_i y_i = \frac{1}{4}$ $\quad \hat\beta_1 = \dfrac{\frac{1}{4} - 3 \cdot \frac{1}{2} \cdot \frac{1}{6}}{\frac{5}{4} - 3\,\frac{1}{4}} = 0$

$\hat y = \frac{1}{6}$ $\quad \hat\beta_0 = \frac{1}{6}$



POZN: (před σ_i) i když s_n^2 je nestranný odhad σ^2

s_n je vychýlený odhad σ !

je to obecná vlastnost odhadů (nestranných) rozptylu, neboť

s^2 nestranný odhad σ^2 $\Rightarrow$ $Es \le \sigma$

uvažujme n. vel. X pro kterou $\operatorname{Var} X < +\infty$

$EX^2 = \operatorname{Var} X + (EX)^2$ dosazením $X = s$ dostaneme

$$Es^2 = \operatorname{Var} s + (Es)^2$$
$$Es^2 = \sigma^2 \Rightarrow (Es)^2 \le \sigma^2 \qquad Es \le \sigma$$

a rovnost nastává $\Leftrightarrow$ $\operatorname{Var} s = 0$

např. pro normální chyby je $s_n^2 \propto \chi^2$ $\Rightarrow$ $Es_n < \sigma$

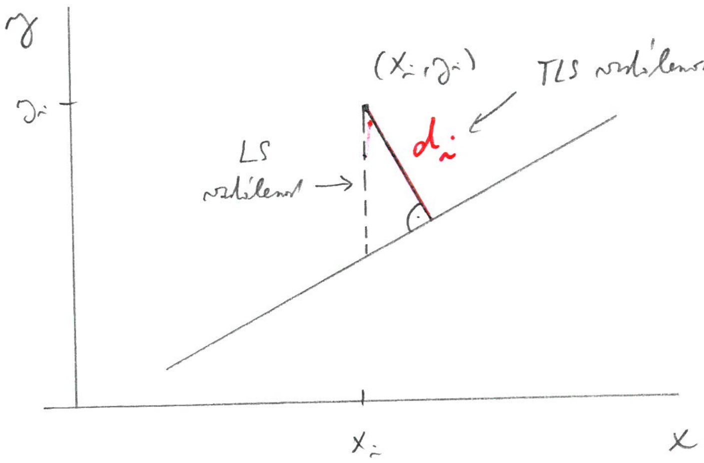
POZN: předp. jsme, že hodnoty x_i jsou dány přesně, což nemusí být vždy pravda

často obě veličiny (x, y) jsou měřeny nepřesně

$\longrightarrow$ EIV models "errors in variables"

v těchto modelech jsou často preferovány jiné odhady než LSE

populární metoda: total least squares (orthogonal least squares)



minimalizuje $\sum_{i=1}^n d_i^2$

neupřednostňuje veličinu x, ale přistupuje k x_i a y_i rovnocenně

POZN: v literatuře se někdy x uvažují jako realizace náh. vel. (ne vždy se x nastavují předem nebo je jasně dané (třeba pohlaví – jin

model má potom tvar

$$E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i \qquad Var(Y_i \mid X_i) = \sigma^2$$

pro většinu výsledků prezentovaných v této přednášce ale není podstatné, zda je x chápáno jako pevné nebo náhodné

Důkazy většinou fungují i pro podmíněné výrazy (E, Var, ...) při dané hodnotě x místo nepodmíněných.

Nicméně větší pozornost je třeba u odvozování asymptotických rozdělení odhadů.

Vlastnosti odhadů $\hat{\beta}_0, \hat{\beta}_1, s_n^2$

Věta: Nechť $\hat{\beta}_0, \hat{\beta}_1$ jsou LSE par. β_0, β_1 v lin. modelu

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \dots, n,$$

kde e_i jsou nezávislé náhodné veličiny (postačí nekorelovanost) se stejným rozptylem σ^2. Potom platí

1) $E(\hat{\beta}_1) = \beta_1$, $\quad E(\hat{\beta}_0) = \beta_0$ $\qquad$ (nestranné odhady)

2) $\operatorname{Var}(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$, kde $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$

3) $\operatorname{Var}(\hat{\beta}_0) = \sigma^2 \left(\dfrac{1}{n} + \dfrac{\bar{x}_n^2}{S_{xx}} \right)$

4) pokud navíc platí $e_i \sim N(0, \sigma^2)$, $i = 1, \dots, n$, potom

$$\hat{\beta}_j \sim N(\beta_j, \operatorname{Var}(\hat{\beta}_j)), \quad j = 0, 1.$$

Dk: 1) upravíme $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{y}_n \bar{x}_n}{\sum x_i^2 - n \bar{x}_n^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} =$$

$$= \frac{1}{S_{xx}} \cdot \Big(\sum_{i=1}^{n} (x_i - \bar{x}_n) y_i - \bar{y} \underbrace{\sum (x_i - \bar{x}_n)}_{= \sum_{i=1}^{n} x_i - n \bar{x}_n = 0} \Big) =$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}_n) y_i$$

$$\underline{E(\hat\beta_1)} = E\left(\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar x_n)\,Y_i\right) = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar x_n)\,EY_i$$

$$= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar x_n)(\beta_0+\beta_1 x_i) = \frac{\beta_0}{S_{xx}}\underbrace{\sum_{i=1}^{n}(x_i-\bar x_n)}_{=0} + \frac{\beta_1}{S_{xx}}\underbrace{\sum_{i=1}^{n}(x_i-\bar x_n)x_i}_{=S_{xx}}$$

$$\sum(x_i-\bar x_n)(x_i-\bar x_n) = \sum(x_i-\bar x_n)x_i + 0$$

$$\underline{= \beta_1}$$

$$\underline{E(\hat\beta_0)} = E(\bar Y_n - \hat\beta_1\bar x_n) = E\bar Y_n - \bar x_n E\hat\beta_1 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\underbrace{EY_i}_{=\beta_0+\beta_1 x_i} - \bar x_n\beta_1 = \beta_0 + \frac{\beta_1}{n}\sum x_i - \bar x_n\beta_1 = \underline{\beta_0}$$

2) $$\operatorname{Var}(\hat\beta_1) = \operatorname{Var}\left(\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar x_n)\,Y_i\right) = \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar x_n)^2\underbrace{\operatorname{Var}Y_i}_{\sigma^2}$$

$$= \frac{\sigma^2 S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$$

3) $$\operatorname{Var}(\hat\beta_0) = \operatorname{Var}(\bar Y - \hat\beta_1\bar x_n) = \operatorname{Var}\bar Y + \bar x^2\operatorname{Var}\hat\beta_1 - 2\bar x\operatorname{cov}(\bar Y,\hat\beta_1)$$

$$= \frac{\sigma^2}{n} + \frac{\bar x_n^2\sigma^2}{S_{xx}} - 2\bar x_n\underbrace{\operatorname{cov}(\bar Y,\hat\beta_1)}_{=0}$$ dále ukázat

$$\operatorname{Cov}(\bar Y,\hat\beta_1) = \operatorname{Cov}\left[\bar Y, \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar x_n)\,Y_i\right] = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar x_n)\operatorname{Cov}(\bar Y, Y_i)$$

$$\operatorname{Cov}(\bar Y, Y_i) = \operatorname{Cov}\left(\frac{1}{n}\sum_{j=1}^{n}Y_j, Y_i\right) = \frac{1}{n}\sum_{j=1}^{n}\underbrace{\operatorname{Cov}(Y_j,Y_i)}_{=0\ i\neq j} = \frac{1}{n}\underbrace{\operatorname{Cov}(Y_i,Y_i)}_{\operatorname{Var}Y_i} = \frac{\sigma^2}{n}$$

$$\Rightarrow \operatorname{Cov}(\bar Y,\hat\beta_1) = 0 = \frac{\sigma^2}{n S_{xx}}\underbrace{\sum_{i=1}^{n}(x_i-\bar x_n)}_{=0}$$

4) protože $\hat{\beta}_1 = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}_n) Y_i$

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \bar{x}_n$$

je $\hat{\beta}_0$ i $\hat{\beta}_1$ LK nezávislých normálních náh. vel. Y_i

$\Rightarrow$ mají normální rozdělení, E a Var jsme spočetli

Q.E.D.

Věta: Za předpokladů předchozí věty platí

$$\underline{E(s_n^2) = \sigma^2}.$$

(s_n^2 je nestranný odhad σ^2)

Dk:

$$E(s_n^2) = \frac{1}{n-2} E \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \frac{1}{n-2} \underbrace{\sum_{i=1}^{n} E(Y_i - \hat{Y}_i)^2}_{A}$$

protože $E(\hat{Y}_i) = E(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \beta_0 + \beta_1 x_i = E Y_i$

platí, že $\underline{E(Y_i - \hat{Y}_i)^2 = Var(Y_i - \hat{Y}_i)}$

$$= E(Y_i - \hat{Y}_i)^2 - \big(\underbrace{E(Y_i - \hat{Y}_i)}_{=0}\big)^2$$

$$A = \sum_{i=1}^{n} Var(Y_i - \hat{Y}_i) = \sum_{i=1}^{n} \big[\underbrace{Var(Y_i)}_{=\sigma^2} + Var(\hat{Y}_i) - 2\, Cov(Y_i, \hat{Y}_i)\big]$$

$$= n\sigma^2 + \sum_{i=1}^{n} Var(\hat{Y}_i) - 2 \sum_{i=1}^{n} Cov(Y_i, \hat{Y}_i) \quad (\#)$$

$$Var\, \hat{Y}_i = Var(\hat{\beta}_0 + \hat{\beta}_1 x_i) = Var\, \hat{\beta}_0 + x_i^2\, Var\, \hat{\beta}_1 + 2 x_i\, Cov(\hat{\beta}_0, \hat{\beta}_1)$$

$$\underline{Cov(\hat{\beta}_0, \hat{\beta}_1)} = Cov(\bar{Y}_n - \hat{\beta}_1 \bar{x}_n, \hat{\beta}_1) = \underbrace{Cov(\bar{Y}_n, \hat{\beta}_1)}_{=0 \text{ viz dříve}} - \bar{x}_n \underbrace{Var(\hat{\beta}_1)}_{=\frac{\sigma^2}{S_{xx}}} = \underline{-\frac{\sigma^2 \bar{x}_n}{S_{xx}}}$$

a tedy

$$\operatorname{Var}\hat{y}_i = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}} + x_i^2\frac{1}{S_{xx}} - \frac{2x_i\bar{x}_n}{S_{xx}}\right] = \sigma^2\left[\frac{1}{n} + \frac{(x_i-\bar{x}_n)^2}{S_{xx}}\right]$$

a $$\sum_{i=1}^{n}\operatorname{Var}\hat{y}_i = \sigma^2 + \frac{\sigma^2}{S_{xx}}\underbrace{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}_{=S_{xx}} = 2\sigma^2$$

$$\operatorname{Cov}(y_i,\hat{y}_i) = \operatorname{Cov}(y_i,\hat{\beta}_0+\hat{\beta}_1x_i) = \operatorname{Cov}(y_i,\hat{\beta}_0) + x_i\operatorname{Cov}(y_i,\hat{\beta}_1)$$

$$\operatorname{Cov}(y_i,\hat{\beta}_1) = \frac{1}{S_{xx}}\sum_{j=1}^{n}(x_j-\bar{x}_n)\underbrace{\operatorname{Cov}(y_i,y_j)}_{=0,\ i\neq j} = \frac{\sigma^2(x_i-\bar{x}_n)}{S_{xx}}$$

$$\operatorname{Cov}(y_i,\hat{\beta}_0) = \operatorname{Cov}(y_i,\bar{y}_n-\bar{x}_n\hat{\beta}_1) = \operatorname{Cov}(y_i,\bar{y}) - \bar{x}_n\operatorname{Cov}(y_i,\hat{\beta}_1)$$

$$= \frac{\sigma^2}{n} - \frac{\bar{x}_n\sigma^2(x_i-\bar{x}_n)}{S_{xx}}$$

a tak $$\operatorname{Cov}(y_i,\hat{y}_i) = \frac{\sigma^2}{n} - \frac{\bar{x}_n\sigma^2(x_i-\bar{x}_n)}{S_{xx}} + \frac{\sigma^2x_i(x_i-\bar{x}_n)}{S_{xx}}$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}(x_i-\bar{x}_n)^2$$

$$\sum_{i=1}^{n}\operatorname{Cov}(y_i,\hat{y}_i) = \sigma^2 + \frac{\sigma^2}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 = 2\sigma^2$$

dosazení do (#)

$$A = n\sigma^2 + 2\sigma^2 - 4\sigma^2 = (n-2)\sigma^2$$

celkem $$E(s_n^2) = \frac{1}{n-2}A = \sigma^2.$$

Q.E.D.

Tvrzení: Nechť platí předpoklady věty 1 a nechť $e_1, \dots, e_n$ i.i.d. $N(0, \sigma^2)$. Potom platí:

a) $$\frac{(n-2)\, s_n^2}{\sigma^2} \sim \chi^2(n-2)$$

b) s_n^2 je nezávislé na $\hat{\beta}_0$ a $\hat{\beta}_1$

Dk: vyplyne z obecnějších tvrzení pro vícerozměrnou regresi

POZN: spočetli jsme

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}} \right] \qquad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

|| ozn $\sigma^2(\hat{\beta}_0)$ || $\sigma^2(\hat{\beta}_1)$

nestranné odhady jsou:

$$\hat{\sigma}^2(\hat{\beta}_0) = s_n^2 \left[\frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}} \right] = s_n^2\, \delta_0$$

$$\hat{\sigma}^2(\hat{\beta}_1) = \frac{s_n^2}{S_{xx}} = s_n^2\, \delta_1$$

(variance multiplication factors)

odhad směrodatné odchylky vel. $\hat{\beta}_0, \hat{\beta}_1$ pak jsou

$$\hat{\sigma}(\hat{\beta}_0) = s_n \sqrt{\delta_0} \quad \text{a} \quad \hat{\sigma}(\hat{\beta}_1) = s_n \sqrt{\delta_1}$$

standardní chyby odhadů $\hat{\beta}_0, \hat{\beta}_1$

hrají zásadní roli při konstrukci IS a TH.

Gauss - Markov theorem

- chyby normální $\Rightarrow$ LSE pro $\hat{\beta}_0, \hat{\beta}_1$ je MLE těchto parametrů (eficientní odhad)

- pokud nejsou chyby normální, jaké je opodstatnění pro použití LSE?

ukážeme, že LSE jsou **BLUE** (best linear unbiased estimators) tedy lin. nestranné odhady s minimálním rozptylem

- je ale třeba poznamenat, že mohou ex. nelineární nebo vychýlené odhady par. β_0, β_1, které jsou eficientnější než LSE, pokud se rozdělení chyb liší výrazně od normálního (tím se zabývá robustní regresní analýza)

Uvažujme model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, \dots, n \qquad (*)$$

Def: Lineární odhad parametru β_j je statistika tvaru

$$\hat{\beta}_j = \sum_{i=1}^{n} c_{ij} Y_i,$$

kde c_i jsou dané reálné konstanty, $(i = 1, \dots, n.)$

Věta (Gauss - Markov): Nechť $e_1, \dots, e_n$ v modelu $(*)$ jsou nekorelované a mají stejný rozptyl $Var(e_i) = \sigma^2$, $i = 1, \dots, n$. Potom LSE $\hat{\beta}_j$ $(j = 0, 1)$ je BLUE parametru β_j.

Dk: ukážeme pro β_1, pro β_0 je důkaz podobný

nechť $\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$, pak

$$\operatorname{Var} \hat{\beta}_1 = \sum_{i=1}^{n} c_i^2 \operatorname{Var} Y_i = \sigma^2 \sum_{i=1}^{n} c_i^2$$

nekorelované Y_i stejný rozptyl

aby byl $\hat{\beta}_1$ nestranný, musí platit $E\hat{\beta}_1 = \beta_1$, tedy

$$E\hat{\beta}_1 = \sum_{i=1}^{n} c_i \underbrace{E Y_i}_{=\beta_0 + \beta_1 x_i} = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \overset{!}{=} \beta_1$$

protože to musí platit pro lib. β_0, β_1 dostáváme

$$\sum_{i=1}^{n} c_i = 0 \quad \text{a} \quad \sum_{i=1}^{n} c_i x_i = 1$$

hledání lineárního, nestranného odhadu β_1 je tedy redukováno

na minimalizaci $\sum_{i=1}^{n} c_i^2$ za podmínek $\sum c_i = 0$ $\sum c_i x_i = 1$

vyřešit

Lagrangeova funkce: $L = \sum_{i=1}^{n} c_i^2 - 2\lambda_1 \left(\sum_{i=1}^{n} c_i\right) - 2\lambda_2 \left(\sum_{i=1}^{n} c_i x_i - 1\right)$

$$\frac{\partial L}{\partial c_i} = 2c_i - 2\lambda_1 - 2\lambda_2 x_i = 0 \qquad i = 1, \dots, n$$

$$\frac{\partial L}{\partial \lambda_1} = -2\left(\sum c_i\right) = 0$$

$$\frac{\partial L}{\partial \lambda_2} = -2\left(\sum c_i x_i - 1\right) = 0$$

sečteme prvních n rovnic

$$\underbrace{\sum_{i=1}^{n} c_i}_{=0} - n\lambda_1 - \lambda_2 \sum_{i=1}^{n} x_i = 0 \Rightarrow \underline{n\lambda_1 + \lambda_2 \sum x_i = 0}$$

$$\Downarrow$$

$$\underline{\lambda_1 = -\lambda_2 \overline{x}_n}$$

sečteme prvních n rovnic vynásobených x_i

$$\underbrace{\sum_{i=1}^{n} c_i x_i}_{=1} - \lambda_1 \sum x_i - \lambda_2 \sum x_i^2 = 0$$

$$\Rightarrow \underline{\lambda_1 \sum x_i + \lambda_2 \sum x_i^2 = 1}$$

$$-\lambda_2 \overline{x}_n \cdot n\overline{x}_n + \lambda_2 \sum x_i^2 = 1$$

$$\lambda_2 \underbrace{\left(\sum x_i^2 - n\overline{x}_n^2\right)}_{=S_{xx}} = 1 \Rightarrow \underline{\lambda_2} = \frac{1}{S_{xx}}$$

$$\underline{\lambda_1} = -\frac{\overline{x}_n}{S_{xx}}$$

dosadíme za λ_1, λ_2

$$c_i + \frac{\overline{x}_n}{S_{xx}} - \frac{x_i}{S_{xx}} = 0 \Rightarrow \underline{c_i = \frac{x_i - \overline{x}_n}{S_{xx}}}$$

a $\underline{\hat{\beta}_1 = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}_n) Y_i}$ což je LSE.

<u>POZN:</u> ukázali jsme pouze, že to je stac. bod, že je tam i minimum ukážeme v obecnější větě u vícerozměrné regrese.

IS pro β_0, β_1

- IS poskytují jistou „míru přesnosti" bodových odhadů
- pro jejich konstrukci potřebujeme znát rozdělení f. bodového odhadu
- budeme tedy uvažovat normalitu chyb
- spočtené IS se ale často používají i když rozdělení chyb není normální, jejich použití se zdůvodňuje tím, že LSE odhady par. β jsou lineární funkcí Y_i, $i=1,\dots,n$ což umožňuje aplikovat CLT a dostat asymptotickou normalitu odhadů $\hat\beta_0, \hat\beta_1$

---

uvažujme model $Y_i = \beta_0 + \beta_1 x_i + e_i$, e_i i.i.d. $N(0,\sigma^2)$

víme $\hat\beta_i \sim N(\beta_i, \sigma^2(\hat\beta_i))$, $\frac{(n-2)s_n^2}{\sigma^2} \sim \chi^2(n-1)$

a nezávisí na $\hat\beta_0, \hat\beta_1$

POZN: $X \sim N(0,1)$, $Y \sim \chi^2(n)$, X, Y nez. $\Rightarrow \frac{X}{\sqrt{\frac{Y}{n}}} \sim t(n)$

tedy
$$T_i = \frac{\frac{\hat\beta_i - \beta_i}{\sigma(\hat\beta_i)}}{\frac{s_n}{\sigma}} = \frac{\hat\beta_i - \beta_i}{\hat\sigma(\hat\beta_i)} \sim t(n-2), \quad i = 0,1$$

neboť $\sigma(\hat\beta_i) = \sigma\sqrt{d_i}$ a $\hat\sigma(\hat\beta_i) = s_n\sqrt{d_i}$.

tzn. $$P\left[-t_{1-\frac{\alpha}{2}}(n-2) \le \frac{\hat{\beta}_i-\beta_i}{\hat{\sigma}(\hat{\beta}_i)} \le t_{1-\frac{\alpha}{2}}(n-2)\right] = 1-\alpha$$

a vyjádřením β_i dostaneme

$$P\left[\hat{\beta}_i - t_{1-\frac{\alpha}{2}}(n-2)\,\hat{\sigma}(\hat{\beta}_i) \le \beta_i \le \hat{\beta}_i + t_{1-\frac{\alpha}{2}}(n-2)\,\hat{\sigma}(\hat{\beta}_i)\right] = 1-\alpha$$

a tedy

$$\left(\hat{\beta}_i \pm t_{1-\frac{\alpha}{2}}(n-2)\,\hat{\sigma}(\hat{\beta}_i)\right)$$ je $100(1-\alpha)\%$ IS pro β_i

$i=0,1$

dosazením za $\hat{\sigma}(\hat{\beta}_i)$ dostaneme:

$100(1-\alpha)\%$ IS pro β_0: $\left(\hat{\beta}_0 \pm t_{1-\frac{\alpha}{2}}(n-2)\cdot s_n\sqrt{\frac{1}{n}+\frac{\bar{x}_n^2}{S_{xx}}}\right)$

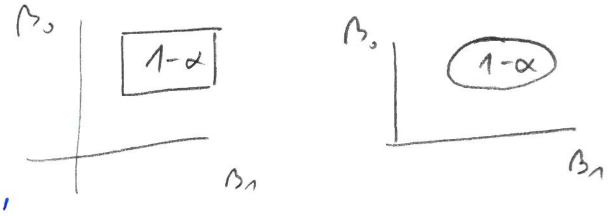
$100(1-\alpha)\%$ IS pro β_1: $\left(\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2}}(n-2)\, s_n \frac{1}{\sqrt{S_{xx}}}\right)$

POZN: ze tvaru IS lze pozorovat, že IS pro β_0 bude ve většině praktických případů širší než IS pro β_1

tzn. směrnice je obecně odhadnuta s větší přesností než absolutní člen (intercept).

Př.: —

POZN: někdy se konstruují simultánní IS pro oba parametry



zmíníme podrobněji u vícerozměrné regrese

TH pro β_0, β_1

chtěli bychom ověřit platnost předpokladu lin. vztahu mezi x a y

předpokládejme nyní, že model je lineární a že x je jediná dostupná vysvětlující proměnná

otázkou je, zda je x užitečná ve vysvětlení variability v y

chceme tedy ~~otestovat dvěma modely~~ rozhodnout mezi dvěma modely:

$$\underline{Y_i = \beta_0 + e_i} \quad \text{a} \quad \underline{Y_i = \beta_0 + \beta_1 x_i + e_i}$$

tzn. otestovat hypotézu $\underline{H_0 : \beta_1 = 0 \times H_1 : \beta_1 \neq 0}$

pokud nezamítneme H_0, závěr bude, že x nevysvětluje nic z variability y a není v modelu významná

zamítneme H_0, znamená, že x je významná

POZN: tyto závěry jsou ale správné pouze za předpokladu, že model je lineární!

- nezamítnutí H_0 ~~může~~ nemusí znamenat, že x není užitečná, může to pouze indikovat, že vztah mezi y a x není lineární
- zamítnutí H_0 naopak říká, že existuje lineární trend mezi y a x, ale mohou tam být i jiné typy závislostí

pro konstrukci testu využijeme odvozené IS

$H_0: \theta = \theta_0 \times H_1: \theta \neq \theta_0$ $\quad (\underline{\theta}, \bar{\theta})$ je $100(1-\alpha)\%$ IS pro θ

pak $\quad W = \{x \mid \theta_0 \notin (\underline{\theta}, \bar{\theta})\}$ je kritický obor testu na hladině α

$H_0: \beta_1 = 0$ zamítneme, pokud $\quad 0 \notin \left(\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2}}(n-2) \dfrac{s_n}{\sqrt{S_{xx}}}\right)$

tzn. buď $\quad \hat{\beta}_1 + t_{1-\frac{\alpha}{2}}(n-2) \dfrac{s_n}{\sqrt{S_{xx}}} < 0 \iff \hat{\beta}_1 \dfrac{\sqrt{S_{xx}}}{s_n} < -t_{1-\frac{\alpha}{2}}(n-2)$

nebo $\quad \hat{\beta}_1 - t_{1-\frac{\alpha}{2}}(n-2) \dfrac{s_n}{\sqrt{S_{xx}}} > 0 \iff \hat{\beta}_1 \dfrac{\sqrt{S_{xx}}}{s_n} > t_{1-\frac{\alpha}{2}}(n-2)$

zapíšeme dohromady

$$|T_n| = |\hat{\beta}_1| \frac{\sqrt{S_{xx}}}{s_n} > t_{1-\frac{\alpha}{2}}(n-2)$$

POZN: intuitivní interpretace:

$|T_n| = |\hat{\beta}_1| \dfrac{\sqrt{S_{xx}}}{s_n} = \dfrac{|\hat{\beta}_1|}{\hat{\sigma}(\hat{\beta}_1)}$ převrácená hodnota relativní chyby $\dfrac{\hat{\sigma}(\hat{\beta}_1)}{|\hat{\beta}_1|}$

pokud je β_1 dobře odhadnuto, očekáváme malý rozptyl $\hat{\sigma}(\hat{\beta}_1)$ vzhledem k β_1, tedy T bude velké

t-test tedy ~~tak~~ říká, že zamítneme H_0 pokud ~~je parametr dobře odhadnut~~ je relativní chyba odhadu malá

POZN: někdy dopředu známe kandidáta b_1 jako hodnotu parametru β_1 a chtěli bychom testovat

$$H_0: \beta_1 = b_1 \quad \times \quad H_1: \beta_1 \neq b_1$$

test bude zamítnout H_0 pokud

$$|\beta_1 - b_1| \cdot \frac{\sqrt{S_{xx}}}{s_n} > t_{1-\frac{\alpha}{2}}(n-2)$$

Test významnosti interceptu:

otázka je, zda přímka prochází počátkem $(0,0)$, tedy

$$H_0: \beta_0 = 0 \quad \times \quad H_1: \beta_0 \neq 0$$

nezamítnutí H_0 znamená, že jednodušší model $y = \beta_1 x + e$ lépe popisuje data než $y = \beta_0 + \beta_1 x + e$

H_0 zamítneme, pokud

$$T_n = \frac{|\hat{\beta}_0|}{\hat{\sigma}(\hat{\beta}_0)} = |\hat{\beta}_0| \frac{1}{s_n \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} > t_{1-\frac{\alpha}{2}}(n-2)$$

ANOVA přístup pro testování

odvodili jsme t-test pro test významnosti koeficientů; nyní odvodíme ekvivalentní F-test, který může být zobecněn na test celkové významnosti vícerozměrného regresního modelu (testy významnosti jednotlivých koeficientů mohou být taktéž zavádějící)

<u>myšlenka metody</u> (analýza rozptylu ANOVA) je určit, kolik variability v pozorováních $(y_1, \dots, y_n)$ je „vysvětleno" regresním modelem (přímkou).

míra variability v datech: <u>total sum of squares</u>

celkový součet čtverců

$$SST = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

pokud regresní přímka $y = \hat{\beta}_0 + \hat{\beta}_1 x$ dobře prokládá data, tedy $\hat{y}_i \approx y_i$

bude platit
$$\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}}_n)^2 \approx \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

ukážeme, že $\bar{\hat{y}}_n = \bar{y}_n$ a tak

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}}_n)^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2 = SSR$$

regression sum of squares

<u>regresní součet čtverců</u>

podíl

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_n)^2}$$

tedy vyjadřuje ~~poměr~~ podíl variability v $(y_1, \dots, y_n)$ vysvětlené regresním modelem

R^2 - koeficient determinace (coefficient of determination)

(pro dobrý model by měla být hodnota $R^2 \approx 1$)

- ukážeme, že R^2 je kvadrát výběrového korelačního koeficientu mezi x a y, což dává stat. R^2 význam míry „dobré shody"

- pokud bychom znali rozdělení f. stat. R^2 nabízí se její použít pro test $H_0 : \beta_1 = 0$, kterou bychom zamítli, pokud bude $R^2 \approx 1$
protože každá monotónní funkce R^2 vede na ekvivalentní test, budeme uvažovat statistiku (proč uvidíme v násl. větě)

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

Věta: Předpokládejme, že $SST \neq 0$. Potom platí

1) $0 \leq R^2 \leq 1$

2) $R^2 = 1 - \frac{SSE}{SST}$ $\left(SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right.$ - residuální součet čtverců$\left.\right)$

3) $R^2 = 1 \iff \hat{y}_i = y_i \ \forall i \in \hat{n}$ (všechna data leží na přímce)

4) pokud ozn. $x = (x_1, \dots, x_n)$ a $y = (y_1, \dots, y_n)$, potom

$$R^2 = \rho^2(x, y),$$

kde $\rho^2(x,y) = \dfrac{\left(\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)\right)^2}{S_{xx}\, S_{yy}}$ je druhá mocnina výběrového korelačního koeficientu vektorů x, y

5) $$F = \frac{SSR}{s_n^2} = T_n^2$$

6) pokud jsou chyby $e_1, \dots, e_n$ i.i.d. $N(0, \sigma^2)$ a $\beta_1 = 0$ (platí $H_0: \beta_1 = 0$) v modelu, potom

$$F \sim F(1, n-2)$$

POZN: • 5) a 6) $\Rightarrow$ použití test. statistiky T_n, R^2 nebo F vede na ekvivalentní test významnosti regrese

• R^2 ~~poskytuje~~ poskytuje hrubou představu o kvalitě modelu, čím je blíže 1, tím lépe přímka prokládá data (nicméně je třeba jisté obezřetnosti, jak uvidíme později)

• F lze chápat jako statistiku pro test významnosti velkých hodnot R^2

---

důkaz něj bude založen na následujícím rozkladu

$$\sum_{i=1}^{n} (y_i - \bar{y}_n)^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

neboli $$SST = SSR + SSE$$

pro důkaz potřebujeme lemma

Lemma: Nechť $\hat{e}_i = y_i - \hat{y}_i$ jsou residua, kde $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ a $\hat{\beta}_0, \hat{\beta}_1$ jsou LSE. Potom

1) $\sum_{i=1}^{n} \hat{e}_i = 0$

2) $\overline{\hat{y}}_n = \bar{y}_n$

3) $\sum_{i=1}^{n} \hat{e}_i \hat{y}_i = 0$

Dk: 1) z rovnice $\frac{\partial S}{\partial \beta_0} = 0$ dostaneme

$$0 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} \hat{e}_i$$

2) z bodu 1) plyne, že $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$, podělením n dostaneme tvrzení

3) z rovnice $\frac{\partial S}{\partial \beta_1} = 0$ dostaneme

$$0 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = \sum_{i=1}^{n} \hat{e}_i x_i$$

a tedy

$$\sum_{i=1}^{n} \hat{e}_i \hat{y}_i = \sum_{i=1}^{n} \hat{e}_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \sum_{i=1}^{n} \hat{e}_i \hat{\beta}_0 + \sum_{i=1}^{n} x_i \hat{e}_i \hat{\beta}_1 =$$

$$= \hat{\beta}_0 \underbrace{\sum_{i=1}^{n} \hat{e}_i}_{=0} + \hat{\beta}_1 \underbrace{\sum_{i=1}^{n} x_i \hat{e}_i}_{=0} = 0 .$$

Q.E.D.

Ukážeme platnost rozkladu:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 = \sum_{i=1}^{n} \left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}_n) \right]^2 =$$

$$= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y}_n)^2 + 2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}_n)$$

$$= SSE + SSR + 0$$

neboť

$$\sum_{i=1}^{n}\underbrace{(y_i-\hat{y}_i)}_{=\hat{e}_i}(\hat{y}_i-\bar{y}_n) = \underbrace{\sum_{i=1}^{n}\hat{e}_i\hat{y}_i}_{=0} - \bar{y}_n\underbrace{\sum_{i=1}^{n}\hat{e}_i}_{=0} = 0 .$$

Důkaz věty:

1) protože $SST = SSE + SSR$ platí

$$0 \le R^2 = \frac{SSR}{SST} \le \frac{SST}{SST} = 1$$

2) $SSR = SST - SSE \quad \Rightarrow \quad R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$

3) z bodu 2 plyne, že $R^2 = 1 \iff SSE = 0$

$$SSE = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = 0 \iff y_i = \hat{y}_i \quad \forall i \in \hat{n}$$

4) $$\hat{y}_i = \underset{\bar{y}_n - \hat{\beta}_1\bar{x}_n}{\overset{\|}{\hat{\beta}_0}} + \hat{\beta}_1 x_i = \bar{y}_n - \hat{\beta}_1(\bar{x}_n - x_i)$$

takže $\underline{SSR} = \sum_{i=1}^{n}(\hat{y}_i-\bar{y}_n)^2 = \hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 = \underline{\hat{\beta}_1^2 S_{xx}}$

protože $\hat{\beta}_1 = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)$, dostaneme

$$\rho^2(x,y) = \frac{\left[\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)\right]^2}{S_{xx}S_{yy}} = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} = \frac{SSR}{SST} = R^2$$

neboť $S_{yy} = \sum_{i=1}^{n}(y_i-\bar{y}_n)^2 = SST$.

5) z def. F

$$F = \frac{(n-2)R^2}{1-R^2} = \frac{(n-2)\frac{SSR}{SST}}{\frac{SSE}{SST}} = \frac{SSR}{\frac{SSE}{n-2}} = \frac{SSR}{s_n^2}$$

protože $T_n = \hat{\beta}_1 \frac{\sqrt{S_{xx}}}{s_n}$ (strana 20),

$$T_*^2 = \frac{\hat{\beta}_1^2}{\cancel{\frac{s_n^2}{S_{xx}}}} = \frac{\hat{\beta}_1^2 S_{xx}}{s_n^2} = \frac{SSR}{s_n^2} = F$$

$$\left(\frac{N(0,1)}{\sqrt{\frac{\chi^2(n-2)}{n-2}}}\right)^2 = \frac{\chi^2(1)}{\frac{\chi^2(n-1)}{n-2}} \sim F(1,n-$$

6) $T_* \sim t(n-2)$ (str. 17)

$$\Downarrow$$

$$F = T^2 \sim F(1, n-2)$$

Q.E.D.

Výsledky se většinou uvádí v tabulce ANOVA:

Source	df	SS	MS	F
Regression	1	SSR	MSR = SSR	$\frac{MSR}{MSE}$
Residual	$n-2$	SSE	$MSE = \frac{SSE}{n-2} = s_n^2$	
Total	$n-1$	SST		

$$R^2 = \frac{SSR}{SST}$$

source – zdroj součtu čtverců

df – počet stupňů volnosti příslušný danému součtu čtverců

SS – součet čtverců

MS – „mean squares" $MS = \frac{SS}{df}$ F – podíl $\frac{MSR}{MSE}$

POZN: $H_0 : \beta_1 = 0$ je zamítnuta, pokud

$$F > F_{1-\alpha}(1, n-2)$$

v tomto jednorozměrném případě je to ekvivalentní t-testu (neboť $F = T^2$)

Věta: Nechť $e_1, \dots, e_n$ jsou i.i.d. $N(0, \sigma^2)$. Za platnosti $H_0 : \beta_1 = 0$ je splněno:

$$\frac{SSR}{\sigma^2} \sim \chi^2(1), \quad \frac{SSE}{\sigma^2} \sim \chi^2(n-2), \quad \frac{SST}{\sigma^2} \sim \chi^2(n-1)$$

POZN: proto se v tabulce ANOVA uvádí df po řadě $1, n-2, n-1$

používají se i v případě jiného rozdělení chyb

představit si lze takto:

- $SSE = \sum \hat{e}_i^2$, ale n residuí $\hat{e}_1, \dots, \hat{e}_n$ mají 2 podmínky

  $\sum \hat{e}_i = 0$ a $\sum x_i \hat{e}_i = 0$ $\Rightarrow$ $n-2$ stupně volnosti

- $SST = \sum (y_i - \bar{y}_n)^2$ $\quad y_i - \bar{y}_n$ musí splňovat $\sum (y_i - \bar{y}_n) = 0$

  $\Rightarrow$ $n-1$ stupeň volnosti

- $SSR = SST - SSE$, počet stupňů volnosti je $(n-1) - (n-2) = 1$

Dk: v předchozím dk jsme ukázali $SSR = \hat{\beta}_1^2 S_{xx}$, tedy

$$\frac{SSR}{\sigma^2} = \left(\frac{\hat{\beta}_1 \sqrt{S_{xx}}}{\sigma} \right)^2,$$ víme $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$ a tedy

$$(\hat{\beta}_1 - \beta_1) \frac{\sqrt{S_{xx}}}{\sigma} \sim N(0,1)$$

pro $\beta_1 = 0$: $\quad \hat{\beta}_1 \frac{\sqrt{S_{xx}}}{\sigma} \sim N(0,1) \Rightarrow \frac{SSR}{\sigma^2} \sim \chi^2(1)$

$$\frac{SSE}{\sigma^2} = \frac{(n-2)\,s_n^2}{\sigma^2} \sim \chi^2(n-2)$$ viz dříve

(ukážeme později)

a nezávisí na $\hat{\beta}_1$

$$\Rightarrow \quad \frac{SSR}{\sigma^2} \text{ a } \frac{SSE}{\sigma^2} \text{ jsou nezávislé}$$

protože
$$\frac{SST}{\sigma^2} = \frac{SSR}{\sigma^2} + \frac{SSE}{\sigma^2} \Rightarrow \frac{SST}{\sigma^2} \sim \chi^2(n-1).$$

$$\chi^2(1) + \chi^2(n-2) = \chi^2(n-1)$$

nez.

Q.E.D.

Př. –

POZN: R^2 statistika – pozor na zjednodušené hodnocení kvality modelu

- nízké hodnoty R^2 nemusí znamenat, že regresní model není vyjímečný, v datech jen může být velké množství nevysvětlitelné náhodné ~~[illegible]~~ variability

  např. opakované hodnoty regresoru x snižují hodnotu R^2 oproti modelům s různými x

- velké hodnoty R^2 mohou být způsobeny velkým měřítkem dat (S_{xx} velké)

– platí totiž
$$E(R^2) \approx \frac{\beta_1^2 S_{xx}}{\beta_1^2 S_{xx} + \sigma^2}$$

  což je rostoucí funkce S_{xx}

– velký rozptyl $(x_1, \dots, x_n)$ může mít za následek velké R^2 přitom nic neříká o kvalitě modelu

– $E(R^2)$ je také rostoucí funkcí β_1^2

  modely v „velkém“ měřítku tedy budou mít obecně větší R^2 než modely s „malou“ směrnicí

Při hodnocení kvality modelu potřebujeme více kritérií

mezi ně patří např.:

1) „velké" R^2

2) „velké" F nebo $|T|$ hodnoty

3) „malé" hodnoty s_m^2 vzhledem k $\bar{y}_m$

další kritéria později

Př. Velká hodnota R^2 indikuje přibližně lineární vztah mezi x a y, ale vysoký stupeň korelace nemusí znamenat příčinný vztah.

data: 1924 – 1937 $\quad$ y_i – počet mentálních onemocnění na 100 000 obyvatel Anglie

(Ex 3.17 p. 89) $\quad$ x_i – počet rádií v populaci

model $y_i = \beta_0 + \beta_1 x_i + e_i$

$\hat{\beta}_0 = 4{,}5822$, $\hat{\beta}_1 = 2{,}2042$ $\quad$ $R^2 = 0{,}984$

tzn. velmi významný lineární vztah mezi x a y

závěr by mohl být, že rádia způsobují mentální onemocnění i když by to mohla být pravda, nabízí se mnohem věrohodnější vysvětlení; a to že x i y rostou lineárně s časem, tzn. y roste lineárně s x

rádia byla s časem dostupnější, lepší diagnostické procedury umožňovaly identifikovat více lidí s mentálními problémy