01RAD - přednáška 3, 1.10.2024

ANOVA přístup pro testování

- odvodili jsme t-test významnosti koeficientů, odvodíme ekvivalentní F-test, který může být zobecněn na test celkové významnosti vícerozměrného regresního modelu
- myšlenka metody (ANOVA): určit, kolik variability v pozorováních $(y_1, y_2, ..., y_n)$ je „vysvětleno" regresním modelem
- míra variability v datech: total sum of squares (SST) - celkový součet čtverců

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

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

$$\sum_{i=1}^{n} (\widehat{y}_i - \overline{\widehat{y}}_n)^2 \approx \sum_{i=1}^{n} (y_i - \overline{y})^2$$

ukážeme, že $\overline{\widehat{y}} = \overline{y}$ a tak

$$\sum_{i=1}^{n} (\widehat{y}_i - \overline{\widehat{y}})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = SSR \quad \text{regression sum of squares - regresní součet čtverců}$$

- Podíl
$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$
vyjadřuje podíl variability v $(y_1, \dots, y_n)$ vysvětlené regresním modelem
- R^2 - koeficient determinace (coefficient of determination) (pro dobrý model $R^2 \approx 1$)
- ukážeme, že $R^2 = \varrho^2(\boldsymbol{x}, \boldsymbol{y})$, tzn. R^2 je míra „dobré shody"
- R^2 by šla použít pro test $H_0: \ \beta_1 = 0$ (zamítnutí pokud $R^2 \approx 1$)
- každá monotonní funkce R^2 vede na ekvivalentní test, budeme uvažovat statistiku
$$F = \frac{(n-2)R^2}{1-R^2}$$

Věta 2.4

Předpokládejme, že v modelu $(*)$ je splněno $SST \neq 0$. Potom platí

1) $0 \leq R^2 \leq 1$,
2) $R^2 = 1 - \dfrac{SSE}{SST}$, kde $SSE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ je reziduální součet čtverců,
3) $R^2 = 1 \iff \widehat{y}_i = y_i,\ \forall i \in \hat{n}$ (všechna data leží na přímce),
4) pokud označíme $\boldsymbol{x} = (x_1, \dots, x_n)$ a $\boldsymbol{y} = (y_1, \dots, y_n)$, potom

$$R^2 = \varrho^2(\boldsymbol{x}, \boldsymbol{y}), \qquad \text{kde } \varrho(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sqrt{S_{xx}S_{yy}}},$$

5) $F = \dfrac{(n-2)R^2}{1-R^2} = \dfrac{SSR}{s_n^2} = T_1^2$,
6) pokud jsou chyby $e_1, \dots, e_n$ $i.i.d.$ $N(0, \sigma^2)$ a platí $H_0:\ \beta_1 = 0$, potom $F \sim F(1, n-2)$.

Poznámka 2.10

- Z bodů 5) a 6) vyplývá, že použití lib. stat. T_1, R^2 nebo F vede na ekvivalentní test významnosti regrese.
- R^2 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 věty bude založen na rozkladu

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2, \quad \text{neboli} \quad SST = SSR + SSE$$

Lemma 2.1

Nechť $\widehat{e}_i = y_i - \widehat{y}_i$ značí rezidua, kde $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ a $\widehat{\beta}_0, \widehat{\beta}_1$ jsou LSE v modelu $(*)$. Potom platí

1) $\sum_{i=1}^{n} \widehat{e}_i = 0$, 2) $\overline{\widehat{y}} = \overline{y}$, 3) $\sum_{i=1}^{n} \widehat{e}_i \widehat{y}_i = 0$.

Důkaz.

$\square$

Důkaz věty 2.4.

Tabulka ANOVA

Výsledky se většinou uvádí ve formě tabulky ANOVA:

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

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

Source – zdroj součtu čtverců, df – počet stupňů volnosti, SS – součet čtverců
MS – "mean squares", průměr čtverců, $(\mathrm{MS} = \frac{\mathrm{SS}}{\mathrm{df}})$

Poznámka 2.11

$$H_0 : \ \beta_1 = 0 \quad \text{je zamítnuta, pokud} \quad F > F_{1-\alpha}(1, n-2)$$

(v jednorozměrném případě je to ekvivalentní t-testu, neboť $F = T_1^2$).

VĚTA 2.5

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

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

POZNÁMKA 2.12

Proto se v tabulce ANOVA uvádí df po řadě $1, n-2, n-1$.

Používají se však i v případě jiného rozdělení chyb. Představit si je lze takto:

1. $SSE = \sum_{i=1}^{n} \widehat{e}_i^2$, na n-reziduí $\widehat{e}_1, \dots, \widehat{e}_n$ máme 2 podm. $\sum_{i=1}^{n} \widehat{e}_i = 0$ a $\sum_{i=1}^{n} x_i \widehat{e}_i = 0 \Rightarrow n-2$ stupňů vol.
2. $SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$, $y_i - \overline{y}$ musí splňovat $\sum_{i=1}^{n} (y_i - \overline{y}_n) = 0 \Rightarrow n-1$ stupňů volnosti
3. $SSR = SST - SSE$, počet stupňů volnosti je $(n-1) - (n-2) = 1$

Důkaz.

Příklad 2.2 (Měření rychlosti zvuku v závislosti na teplotě)

teplota	-20	0	20	50	100
rychlost (m/s)	323	327	340	364	386

```
mod <- lm(rychlost~teplota)

anova(mod)

## Analysis of Variance Table
##
## Response: rychlost
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## teplota    1 2773.14 2773.14   146.3 0.001216 **
## Residuals  3   56.86   18.95
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Připomenutí: $s_n^2 = 18.95, \qquad T_1^2 = (12.096)^2 = 146.3\,.$

Poznámka 2.13 (R^2 statistika - pozor na zjednodušené hodnocení kvality modelu!)

- nízké hodnoty R^2 nemusí znamenat, že regresní model není významný

  v datech jen může být velké množství nevysvětlitelné náhodné 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} je velké)
  - platí totiž

    $$E(R^2) \approx \frac{\beta_1^2 S_{xx}}{\beta_1^2 S_{xx} + \sigma^2} \qquad \text{(rostoucí funkce } S_{xx}\text{)}$$

  - velký rozptyl $(x_1, \ldots, x_n)$ může mít za následek velké R^2, přitom nic neříká o kvalitě modelu
  - $E(R^2)$ je také rostoucí funkcí β_1^2, modely s „velkou“ směrnicí tedy budou mít obecně větší R^2, než modely s „malou“ směrnicí

Při hodnocení kvality modelu potřebujeme více kritérií. Mezi ně patří například

1) „velká“ hodnota R^2, 2) „velké“ hodnoty statistik F nebo $|T_1|$,

3) „malé“ hodnoty s_n^2 vzhledem k $\overline{y}$ (další kritéria později)

Příklad 2.3

- velká hodnota R^2 indikuje přibližně lineární vztah mezi x a y
- vysoký stupeň korelace ale nemusí znamenat příčinný vztah!

Data: 1924-1937

y_i - počet mentálních onemocnění na 100000 obyvatel Anglie.

x_i - počet rádií v populaci.

Model: $y_i = \beta_0 + \beta_1 x_i + e_i$

$$\widehat{\beta}_0 = 4.5822, \qquad \widehat{\beta}_1 = 2.2042, \qquad R^2 = 0.984$$

tzn. velmi významný lineární vztah mezi x a y

Závěr: rádia způsobují mentální onemocnění (???)

věrohodnější vysvětlení: 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.)

Poznámka 2.14 (korelace $\times$ příčinnost)

a)

$$\begin{array}{ccc} x & \longrightarrow & y \\ x & \longleftarrow & y \end{array}$$

Causal link (příčinná spojitost)

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

b)

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

Hidden cause (skrytá příčinnost)

skrytá veličina z ovlivňuje x i y, což způsobuje jejich korelovanost

c)

$$\begin{tikzcd} & z \arrow[dr] & \\ x \arrow[rr] & & y \end{tikzcd}$$

Confounding factor (zavádějící faktor)

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

d)

$$x \qquad\qquad y$$

Coincidence (shoda okolností)

korelace je náhodná

01RAD - přednáška 4, 1.10.2024

Regresní model procházející počátkem

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

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

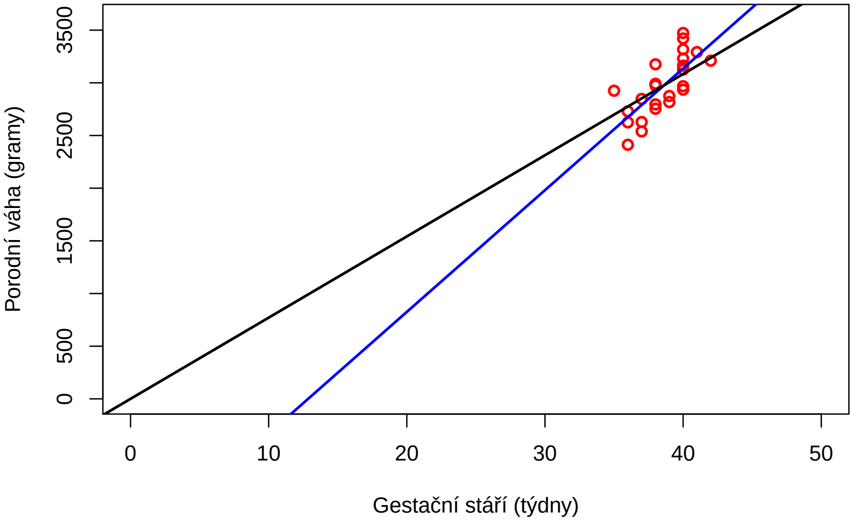
např.

- je to předem známo na základě fyzikálních úvah ($\mathrm{E} Y_0 = \beta_0 = 0$), potom nemá smysl odhadovat β_0, obecně to snižuje přesnost odhadu σ^2 a tím i β_1
- na začátku předpokládáme $\beta_0 \neq 0$ a t-test nezamítne $H_0 : \beta_0 = 0$, potom β_0 může být z modelu odstraněno

Poznámka 2.15

- v praxi často není jisté, 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 $\mathrm{E} Y_0 = 0$, pokud totiž nevíme jistě, že model je lineární na okolí 0, volba $\beta_0 = 0$ může vést k vychýleným odhadům β_1, pokud jsou nezávislé proměnné daleko od 0

Příklad: porodní váha a gestační stáří



Odhady a testy v případě $\beta_0 = 0$

- LSE parametru β_1 dostaneme minimalizací $S = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$ ve tvaru $\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$
- pokud $e_1, \ldots, e_n$ i.i.d. $N(0, \sigma^2)$, potom

$$\mathrm{E}\widehat{\beta}_1 = \beta_1, \quad \mathrm{Var}\widehat{\beta}_1 = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}, \quad \widehat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}),$$

$$s_n^2 = \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n-1} = \frac{SSE}{n-1} \text{ je nestranný odhad } \sigma^2, \quad \frac{SSE}{\sigma^2} \sim \chi^2(n-1) \text{ a nezávisí na } \widehat{\beta}_1$$

- $H_0 : \beta_1 = 0$ lze otestovat pomocí $T = \frac{\widehat{\beta}_1}{\frac{s_n}{\sqrt{\sum_{i=1}^{n} x_i^2}}} \sim t(n-1)$

- $100(1-\alpha)\%$ IS pro β_1 je $\left(\widehat{\beta}_1 \pm t_{1-\alpha/2}(n-1)\, s_n / \sqrt{\sum_{i=1}^{n} x_i^2}\right)$

Poznámka 2.16

- zatím vše podobné jako pro případ $\beta_1 \neq 0$
- rozdíl ale bude v tabulce ANOVA a R^2 statistice, neplatí totiž rozklad $SST = SSR + SSE$
- odvodíme nový rozklad (platí v obou modelech, dokážeme jen pro $\beta_0 = 0$)

Věta 2.6

V modelu s $\beta_0 = 0$ platí $\displaystyle \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \widehat{y}_i^2 + \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \,.$

Důkaz.

$\square$

Pokud vezmeme $\sum_{i=1}^{n} y_i^2$ jako míru variability v datech, analogie R^2 bude: $R^2 = \dfrac{\sum_{i=1}^{n} \widehat{y}_i^2}{\sum_{i=1}^{n} y_i^2}$

potom

$$1 - R^2 = \frac{\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} \widehat{y}_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\sum_{i=1}^{n} \widehat{e}_i^2}{\sum_{i=1}^{n} y_i^2}, \quad \text{a definicí} \quad F = \frac{(n-1)R^2}{1-R^2}$$

dostaneme

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

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

Poznámka 2.17

Tato definice R^2 se ale moc nepoužívá, protože neumožňuje přímé srovnání modelů *bez* a *s* interceptem

$$\beta_0 = 0: \quad R^2 = 1 - \frac{SSE}{\sum_{i=1}^{n} y_i^2} \qquad\qquad \beta_0 \neq 0: \quad R^2 = 1 - \frac{SSE}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

obecně ale $\sum_{i=1}^{n} (y_i - \overline{y})^2 < \sum_{i=1}^{n} y_i^2$, R^2 modelu s $\beta_0 = 0$ tak 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 ex. několik verzí
- možná volba je

$$R^2 = \varrho^2(\boldsymbol{y}, \widehat{\boldsymbol{y}}),$$

  kde $\widehat{\boldsymbol{y}} = (\widehat{y}_1, \dots, \widehat{y}_n)$ (platí i pro $\beta_0 \neq 0$)

- další možnost: srovnat modely pomocí s_n^2

Tabulka ANOVA pro $\beta_0 = 0$.

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

$$R^2 = \varrho^2(\boldsymbol{y}, \widehat{\boldsymbol{y}})$$

Příklad 2.4 (Porodní váha a gestační stáří)

```
mod <- lm(Weight ~ Age)
summary(mod)
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1485.0      852.6  -1.742   0.0955 .
## Age            115.5       22.1   5.228 3.04e-05 ***

## Residual standard error: 192.6 on 22 degrees of freedom
## Multiple R-squared:  0.554,        Adjusted R-squared:  0.5338
## F-statistic: 27.33 on 1 and 22 DF,  p-value: 3.04e-05
```

```
y.mod <- predict(mod)
cor(y.mod,Weight)^2

## 0.5540268
```

Model bez interceptu:

```
mod.bez <- lm(Weight ~ Age - 1)
summary(mod.bez)

## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## Age   77.081      1.063   72.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 200.9 on 23 degrees of freedom
## Multiple R-squared:  0.9956,        Adjusted R-squared:  0.9955
## F-statistic:  5258 on 1 and 23 DF,  p-value: < 2.2e-16

anova(mod.bez)

## Analysis of Variance Table
##
## Response: Weight
##           Df    Sum Sq   Mean Sq F value    Pr(>F)
## Age        1 212270368 212270368  5257.6 < 2.2e-16 ***
## Residuals 23    928596     40374

y.mod.bez <- predict(mod.bez)

cor(y.mod.bez,Weight)^2

## 0.5540268
```

Predikce

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

Budeme uvažovat dva typy predikce:

1) predikce střední hodnoty $\mu_0 = \mathrm{E}[Y_0]$ v bodě x_0,
2) predikce hodnoty nového pozorování Y_0 v bodě x_0.

Pro oba typy použijeme bodový odhad

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0,$$

intervalové odhady se ale budou lišit.

Ad 1) $\mu_0 = \beta_0 + \beta_1 x_0$ je vlastně parametr, lze pro něj odvodit IS (za předpokladu normality chyb).

Spočteme $\mathrm{Var}(\widehat{Y}_0)$:

Shrnutí: 100$(1-\alpha)$% IS pro μ_0

$$\widehat{Y}_0 \pm t_{1-\frac{\alpha}{2}}(n-2) \cdot \widehat{\sigma}(\widehat{Y}_0), \quad \text{kde} \quad \widehat{\sigma}^2(\widehat{Y}_0) = s_n^2\left[\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right]$$

$\widehat{\sigma}(\widehat{Y}_0)$ se obvykle nazývá standardní chyba predikce v bodě x_0

Poznámka 2.18

Z tvaru IS je vidět, že bude nejkratší pro $x_0 = \overline{x}$ a s rostoucí vzdáleností $|x_0 - \overline{x}|$ se prodlužuje.

- Speciálně potom čím dále jsme od oblasti, kde jsou naše data x, tím méně spolehlivé jsou 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 $\longleftrightarrow$ intervaly predikce.

Potřebujeme rozptyl $Y_0 - \widehat{Y}_0$:

Shrnutí: $100(1-\alpha)\%$ interval predikce pro Y_0

$$\widehat{Y}_0 \pm t_{1-\frac{\alpha}{2}}(n-2) \cdot s_p, \quad \text{kde} \quad s_p = s_n \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

Poznámka 2.19

Přesnost predikce

a) roste s rostoucím n a rostoucím rozsahem x naměřeným pomocí S_{xx},
b) klesá s rostoucím $|x_0 - \overline{x}|$.

Pokud je možno předem zvolit $x_1, \ldots x_n$, lze přesnost predikce zvýšit volbou dostatečně rozptýlených hodnot x.

To ale může zvyšovat R^2 a někdy vést k horšímu modelu.

$\Rightarrow$ základní rozpor v regresní analýze:

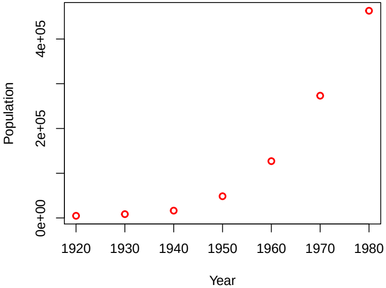
- dobrý model nemusí poskytovat dobré predikce,
- dobré predikce mohou vycházet z méně přesných modelů.

Poznámka 2.20

- odvozené výsledky platí za předpokladu normality chyb
- za podmínek regularity jsou ale odhady $\widehat{\beta}_0, \widehat{\beta}_1$ asymptoticky normální, tzn. IS pro $\mathrm{E}\,Y_0$ budou použitelné pro velká n i pro nenormální chyby
- IP pro Y_0 ale závisí na normalitě chyb i pro velká n, tzn. mohou tedy být nepřesné pro nenormální chyby

Příklad 2.5 (Clark County population data)

x	Year	Population
0	1920	4859
1	1930	8539
2	1940	16414
3	1950	48589
4	1960	127016
5	1970	273288
6	1980	463087



1. lineární model pro původní data: $Y_i = \beta_0 + \beta_1 x_i + e_i$

$$\widehat{\beta}_0 = -81\,328,\ p_{val} = 0.22, \quad \widehat{\beta}_1 = 71\,957,\ p_{val} = 0.007, \quad R^2 = 0.80, \quad F = 19.96$$

x	Year	Population	Fitted.value
0	1920	4859	-81328
1	1930	8539	-9371
2	1940	16414	62585
3	1950	48589	134542
4	1960	127016	206498
5	1970	273288	278455
6	1980	463087	350411

Predikce pro rok 1990:

$$\widehat{y}_{1990} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot 7 = 422\,368$$

a) 95% IS: $(237\,233, 607\,502)$

b) 95% IP: $(135\,559, 709\,177)$

Skutečná hodnota v roce 1990: 768 203

2. lineární model pro log-transformovaná data: $log(Y_i) = \beta_0 + \beta_1 x_i + e_i$

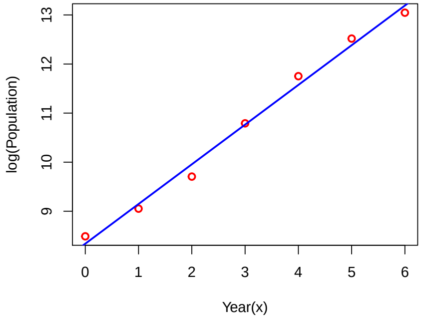
$$\widehat{\beta}_0 = 8.33, \; p_{val} < 10^{-4}, \quad \widehat{\beta}_1 = 0.809, \; p_{val} < 10^{-4}, \quad R^2 = 0.991, \quad F = 550.9$$

Predikce pro rok 1990 na log. škále:

$$\widehat{\log y}_{1990} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot 7 = 14.0004$$

a) 95% IS: $(13.604, 14.397)$

b) 95% IP: $(13.387, 14.614)$



Predikce pro rok 1990 na původní škále:

$$\widehat{y}_{1990} = 1\,203\,161$$

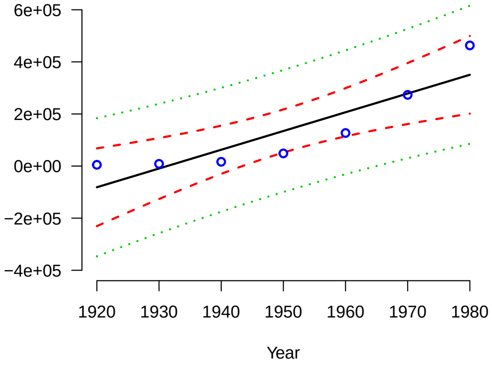
a) 95% IS: $(809\,576, 1\,788\,092)$

b) 95% IP: $(651\,269, 2\,222\,733)$

Skutečná hodnota v roce 1990: 768 203

Intervaly predikce v R

```
mod.lin <- lm(pop ~ year)

new <- data.frame(year=seq(0,6,0.1))

# predikce y
y.hat <- predict(lm(pop ~ year), new)

# 95% intervaly spolehlivosti
CI<-predict(mod.lin, new, interval = "confidence")

# 95% intervaly predikce
PI<-predict(mod.lin, new, interval = "prediction")

# obrazek
matplot(new$year, cbind(CI, PI[,-1]),
        lty = c(1,2,2,3,3), col = c(1,2,2,3,3),
        type = "l", lwd=2, ylab="",xlab = "Year",
        axes = FALSE, ylim=c(-400000,600000))
axis(side=1, at=0:6, labels=Year)
axis(side=2, las=2)
points(pop ~ year, col="blue", lwd=2)
```



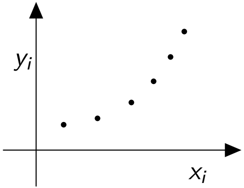
Ověření adekvátnosti modelu

- důležitá součást analýzy
- mělo by předcházet interpretaci modelu případně přijímání závěrů založených na modelu
- výsledky odvozeny za předpokladu linearity modelu,případně normality chyb

Základní procedury:

1) Prozkoumání scatter plotu dvojic (x_i, y_i)

   Např.



může indikovat, že lepší model bude

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

(může být zavádějící)

2) Analýza hodnot testovacích statistik

- např. malá hodnota R^2 společně s významnou hodnotou t-statistiky pro parametr β_1 naznačuje, že skutečný model obsahuje i jiné proměnné x
- velká hodnota R^2 a významná t-satistika 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 očekávat, že budou užitečné pro detekci odchylek od normality

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 tak nechovají, může to znamenat neadekvátnost modelu
- ukážeme grafické nástroje, začneme ale vlastnostmi reziduí

Věta 2.7

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

1) $E(\widehat{e}_i) = 0, \quad i = 1, \ldots, n$
2) $\mathrm{Var}(\widehat{e}_i) = \sigma^2_{\widehat{e_i}} = \sigma^2 \left[1 - \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}}\right)\right]$ $\quad (\approx \sigma^2$ pro velká $n)$
3) $\mathrm{Cov}(\widehat{e}_i, \widehat{e}_j) = -\sigma^2 \left[\frac{1}{n} + \frac{(\overline{x} - x_i)(\overline{x} - x_j)}{S_{xx}}\right]$
4) $\mathrm{Cov}(\widehat{e}_i, \widehat{Y}_i) = 0, \quad i = 1, \ldots, n$
5) Pokud jsou $e_1, \ldots, e_n$ i.i.d. $N(0, \sigma^2)$, potom platí: $\widehat{Z}_i = \dfrac{\widehat{e}_i}{\sigma_{\widehat{e_i}}} \sim N(0, 1).$

Důkaz.

Poznámka 2.21

- bod 3) věty $\Rightarrow$ $\mathrm{Cov}(\widehat{e}_i, \widehat{e}_j) \approx 0$ pro velké n
- pokud jsou tedy e_i i.i.d. $N(0, \sigma^2)$, měla by se *standardizovaná rezidua* $\widehat{Z}_i = \frac{\widehat{e}_i}{\sigma_{\widehat{e}_i}}$ chovat pro velké n jako náhodný výběr z $N(0, 1)$ rozdělení
- budeme potřebovat odhad σ^2 pro výpočet $\widehat{Z}_i$
- <u>nejznámější procedura:</u> odhadnout σ^2 pomocí s_n^2, potom

$$\widehat{r}_i = \frac{\widehat{e}_i}{s_n \sqrt{1 - \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}} \right)}} \qquad \text{studentizovaná rezidua}$$

- pro velká n by se opět měla $\widehat{r}_i$ chovat jako náh. výběr z $N(0, 1)$.

Poznámka 2.22

- $\widehat{e}_i, \widehat{r}_i$ se užívají pro grafickou analýzu
- jiná třída reziduí – PRESS rezidua (negrafické metody zkoumání reziduí):

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

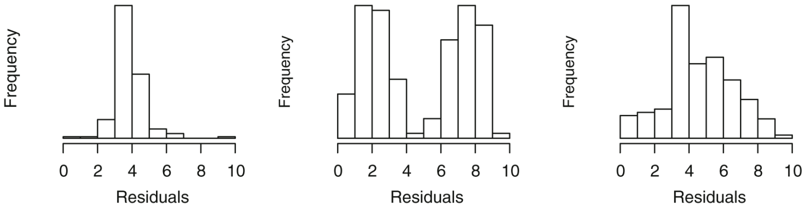
  pak i-té PRESS reziduum je definováno jako

$$\widehat{e}_{(-i)} = y_i - \widehat{y}_{(-i)}, \quad \text{kde} \;\; \widehat{y}_{(-i)} = \widehat{\beta}_{0(-i)} + x_i \widehat{\beta}_{1(-i)}.$$

  (podrobněji se jim budeme věnovat později)

Grafy reziduí

1) histogram reziduí

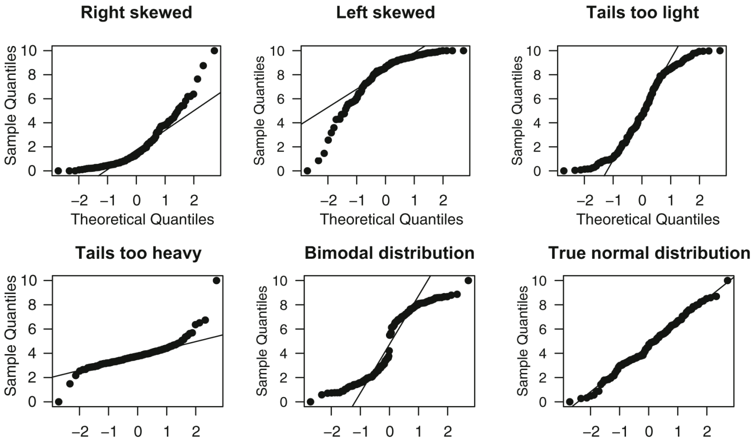


2) kvantilový graf (Q-Q plot) studentizovaných reziduí

- seřadíme dle velikosti:

$$\widehat{r}_{(1)} \leq \widehat{r}_{(2)} \leq \cdots \leq \widehat{r}_{(n)} \quad \text{a vyneseme oproti} \quad \Phi^{-1}\left((i - \tfrac{1}{2})\frac{1}{n}\right), i = 1, \ldots, n$$

- body by měly ležet přibližně na přímce $(\mathrm{E}(r_{(i)}) \approx \Phi^{-1}\left((i - \frac{1}{2})\frac{1}{n}\right)$ pro normální chyby)
- použití: ověření normality, detekce odlehlých pozorování



3) studentizovaná rezidua vs. jednotlivé vysvětlující proměnné x

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

4) studentizovaná rezidua $\widehat{r}_i$ vs. predikované hodnoty $\widehat{y}_i$

- $\mathrm{Cov}(\widehat{e}_i, \widehat{Y}_i) = 0$, tedy $\widehat{r}_i$ a $\widehat{Y}_i$ by měly být nekorelované, pokud platí model $(*)$
- tzn. graf $\widehat{r}_i \times \widehat{y}_i$ by měl být náhodně rozptýlený kolem osy x
- navíc $\widehat{r}_i$ by měla ležet v $(-3, 3)$ $(\widehat{r}_i \approx N(0, 1))$

5) studentizovaná rezidua vs. pořadí pozorování

možná detekce řadové korelace mezi pozorováními