01RAD - přednáška 7, 22.10.2024

Věta 3.9

Nechť v modelu $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$ tvaru $(**)$ platí $h(\boldsymbol{X}) = m + 1$ a $e_1, \ldots, e_n$ jsou i.i.d. $N(0, \sigma^2)$. Pokud $\boldsymbol{\beta}_s = \mathbf{0}$, tj. $\beta_1 = \beta_2 = \cdots = \beta_m = 0$, potom

$$F \sim F(m, n - m - 1).$$

Důkaz.

TEST: zamítnout $H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0$ pokud $F > \mathrm{F}_{1-\alpha}(m, n-m-1)$

POZNÁMKA: Odvozeno pro $e_i \sim N(0, \sigma^2)$. Obecně se používá, i když to nevíme, pro velké n může být často zdůvodněno pomocí CLV.

Tabulka ANOVA

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

Koeficient (vícenásobné) determinace R^2:

Podobně jako u jednorozměrné regrese, lze F-test chápat jako test významnosti $\boldsymbol{R}^2$, definovaného jako

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (SST = SSR + SSE)$$

Poznámka 3.6

- R^2 je možno zvětšovat přidáváním nových proměnných x, i když jsou statisticky nevýznamné (Pro n LN proměnných x a n pozorování dostaneme "perfect fit") (overfitting)
- Vysvětlení: $R^2 = 1 - \frac{SSE}{SST}$, kde
  - SST je pevně dáno daty y, ale SSE může být snížena přidáním proměnných x
  - minimalizujeme totiž $(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)$ přes větší množinu β
  - to znamená, že $\frac{SSE}{SST}$ je nerostoucí funkce a R^2 je neklesající funkce počtu proměnných

Upravený koeficient determinace (adjusted coefficient of determination):

$$\overline{R}^2 = R^2_{adj} = 1 - \frac{\dfrac{SSE}{n-m-1}}{\dfrac{SST}{n-1}} = 1 - \frac{n-1}{n-m-1}\frac{SSE}{SST}.$$

PŘÍKLAD 3.3 (Porodní váha a gestační stáří)

```
mod <- lm(Weight ~  Age + Sex + Age:Sex)
summary(mod)

##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1268.67    1114.64  -1.138 0.268492
## Age           111.98      29.05   3.855 0.000986 ***
## Sexgirl      -872.99    1611.33  -0.542 0.593952
## Age:Sexgirl    18.42      41.76   0.441 0.663893
## ---
## Residual standard error: 180.6 on 20 degrees of freedom
## Multiple R-squared:  0.6435,        Adjusted R-squared:   0.59
## F-statistic: 12.03 on 3 and 20 DF,  p-value: 0.000101

SSR <- sum((predict(mod)-mean(Weight))^2)
SSE <- sum((predict(mod)-(Weight))^2)
F <- (SSR/3)/(SSE/(24-3-1)); F

## 12.03152

anova(mod)

## Analysis of Variance Table
## Response: Weight
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Age        1 1013799 1013799 31.0779 1.862e-05 ***
## Sex        1  157304  157304  4.8221   0.04006 *
## Age:Sex    1    6346    6346  0.1945   0.66389
## Residuals 20  652425   32621
## ---
```

3.4 Intervaly spolehlivosti a t-testy pro parametry

- pokud se model ukáže jako významný, bude nás zajímat, které koeficienty přispívají
- lze použít IS a TH stejně, jako u jednorozměrné regrese
- výsledky jsou odvozeny pro normální chyby
- v praxi se používají i pro jiné typy chyb

Pro konstrukci použijeme dokázanou vlastnost

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{s_n \sqrt{v_j}} \sim t(n-m-1), \quad \text{kde} \quad v_j = (\boldsymbol{X}^T \boldsymbol{X})^{-1}_{jj}$$

$100(1-\alpha)\%$ IS pro β_j je

$$\left(\widehat{\beta}_j - t_{1-\frac{\alpha}{2}}(n-m-1) s_n \sqrt{v_j}, \widehat{\beta}_j + t_{1-\frac{\alpha}{2}}(n-m-1) s_n \sqrt{v_j}\right)$$

S pomocí IS lze odvodit kritický obor pro test $H_0 : \beta_j = b_j \quad \times \quad H_1 : \beta_j \neq b_j$ ve tvaru

$$\frac{|\widehat{\beta}_j - b_j|}{s_n\sqrt{v_j}} > t_{1-\frac{\alpha}{2}}(n-m-1).$$

Pro $b_j = 0$ dostaneme test významnosti β_j, tzn. $H_0 : \; \beta_j = 0$ zamítneme, pokud

$$\frac{|\widehat{\beta}_j|}{s_n\sqrt{v_j}} > t_{1-\frac{\alpha}{2}}(n-m-1).$$

Poznámka 3.7

- pokud nejsou porušeny předpoklady modelu nebo není přítomna kolinearita, lze zvážit odstranění všech nevýznamných proměnných (dle t-testu)
- v případě kolinearity, může být model významný (dle celkového F-testu), ale všechny nebo téměř všechny proměnné se mohou jevit jako nevýznamné (dle t-testů)
- naopak, pokud má model velký počet možných proměnných, některé proměnné se mohou jevit významné, i když jsou náhodným šumem
- při použití t-testů je třeba postupovat obezřetně

Příklad 3.4 (Porodní váha a gestační stáří)

```
mod <- lm(Weight ~  Age + Sex + Age:Sex)
summary(mod)

##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1268.67    1114.64  -1.138 0.268492
## Age           111.98      29.05   3.855 0.000986 ***
## Sexgirl      -872.99    1611.33  -0.542 0.593952
## Age:Sexgirl    18.42      41.76   0.441 0.663893
## ---
## Residual standard error: 180.6 on 20 degrees of freedom
## Multiple R-squared:  0.6435,        Adjusted R-squared:   0.59
## F-statistic: 12.03 on 3 and 20 DF,  p-value: 0.000101
```

```
vif(mod)
## Age     Sex       Age:Sex
## 1.96444 477.55172 483.47306
```

```
mod0 <- lm(Weight ~  Age + Sex)
summary(mod0)

##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1610.28     786.08  -2.049   0.0532 .
## Age           120.89      20.46   5.908 7.28e-06 ***
## Sexgirl      -163.04      72.81  -2.239   0.0361 *
## ---
## Residual standard error: 177.1 on 21 degrees of freedom
## Multiple R-squared:   0.64,        Adjusted R-squared:  0.6057
## F-statistic: 18.67 on 2 and 21 DF,  p-value: 2.194e-05
```

```
vif(mod0)
## Age      Sex
## 1.013904 1.013904
```

Příklad 3.5 (Data cement)

	x1	x2	x3	x4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

```
mod <- lm(y ~ x1 + x2 + x3 + x4)
summary(mod)

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.4054    70.0710   0.891   0.3991
## x1            1.5511     0.7448   2.083   0.0708 .
## x2            0.5102     0.7238   0.705   0.5009
## x3            0.1019     0.7547   0.135   0.8959
## x4           -0.1441     0.7091  -0.203   0.8441
## ---
## Residual standard error: 2.446 on 8 degrees of freedom
## Multiple R-squared:  0.9824,	Adjusted R-squared:  0.9736
## F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07

vif(mod)
##        x1        x2        x3        x4
##  38.49621 254.42317  46.86839 282.51286

mod1 <- lm(y ~ x1)
summary(mod1)

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  81.4793     4.9273   16.54 4.07e-09 ***
## x1            1.8687     0.5264    3.55  0.00455 **
## ---
## Residual standard error: 10.73 on 11 degrees of freedom
## Multiple R-squared:  0.5339,	Adjusted R-squared:  0.4916
## F-statistic:  12.6 on 1 and 11 DF,  p-value: 0.004552
```

```
mod12 <- lm(y ~ x1 + x2)
summary(mod12)

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 52.57735    2.28617   23.00 5.46e-10 ***
## x1           1.46831    0.12130   12.11 2.69e-07 ***
## x2           0.66225    0.04585   14.44 5.03e-08 ***
## ---
## Residual standard error: 2.406 on 10 degrees of freedom
## Multiple R-squared:  0.9787,       Adjusted R-squared:  0.9744
## F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-09

vif(mod12)
##       x1       x2
## 1.055129 1.055129
```

Poznámka 3.8

- statistiky $\mathrm{F}, \boldsymbol{R}^2$ a t jsou užitečné pro rozkrytí efektů jednotlivých proměnných
- nemohou být ale používány úplně automaticky

3.5 Obecná lineární hypotéza

Hypotézy uvažované v F-testu a t-testech jsou speciálním případem obecné lineární hypotézy

$$H_0 : \boldsymbol{C}\beta = \boldsymbol{b} \qquad \times \qquad H_1 : \boldsymbol{C}\beta \neq \boldsymbol{b},$$

kde $\boldsymbol{C} \in \mathbb{R}^{r\times(m+1)}$ a $h(\boldsymbol{C}) = r \quad$ (tzn. $r \leq m+1$)

Rovnice $\boldsymbol{C}\beta = \boldsymbol{b}$ reprezentuje r lineárně nezávislých podmínek

$$\sum_{j=0}^{m} c_{ij}\beta_j = b_i, \qquad i = 1, \dots, r.$$

POZNÁMKA 3.9

a) volba $\boldsymbol{b} = (0, \dots, 0)^T$ a $\boldsymbol{C} = \left(\begin{array}{c|cccc} 0 & 1 & 0 & \dots & 0 \\ \hline 0 & 0 & 1 & & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & 1 \end{array}\right)_{m\times(m+1)}$ vede na test

$$H_0 : \ \boldsymbol{C}\beta = \boldsymbol{0} \qquad \Leftrightarrow \qquad H_0 : \ \beta_1 = \beta_2 = \dots = \beta_m = 0,$$

b) volba $\boldsymbol{b} = \mathbf{0}$ a $\boldsymbol{C} = (0, \dots, 0, 1, 0, \dots, 0)$ vede na test $\quad H_0: \ \beta_j = 0,$

c) v modelu $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$ chceme testovat zároveň, že $\beta_2 = 0$ a $\beta_3 = \beta_4$,

to lze udělat volbou $\boldsymbol{C} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}, \ \boldsymbol{b} = (0, 0)^T.$

Pro test H_0 naladíme 2 modely:

1) plný model (full model) - bez podmínek na $\boldsymbol{C}\beta$,

2) redukovaný model (reduced model) - za předpokladu, že platí $H_0: \ \boldsymbol{C}\beta = \boldsymbol{b}$.

Označme příslušné reziduální součty čtverců SSE_F a SSE_R $\quad (SSE_F \leq SSE_R)$

- pokud neplatí H_0, dá se očekávat, že $\Delta SSE = SSE_R - SSE_F$ bude významně větší, než náhodná chyba σ^2
- H_0 tedy budeme zamítat, pokud $\frac{\Delta SSE}{s_n^2}$ bude velké
- zobecnění F-testu, za platnosti H_0 ukázeme pro normální chyby vztah

$$F = \frac{\frac{\Delta SSE}{r}}{s_n^2} \sim \mathrm{F}(r, n - m - 1)$$

Příklad 3.6

Uvažujme F-test pro $H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0$ v plném modelu $(**)$.