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

$$\sum_{i=1}^{n} (\widehat{y}_i - \overline{\widehat{y}})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = SSR$$ 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 $(y_1, ..., y_n)$ 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 $\boldsymbol{x} = (x_1, ..., x_n)$ a $\boldsymbol{y} = (y_1, ..., y_n)$, 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 $e_1, ..., e_n$ *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.

□

Důkaz věty 2.4.

Tabulka ANOVA

Výsledky se většinou uvádí ve formě tabulky ANOVA:

Source	df	SS	MS	F
Regression	1	SSR	$\mathrm{MSR}=\frac{SSR}{1}$	$\frac{\mathrm{MSR}}{\mathrm{MSE}}$
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 $e_1, ..., e_n$ *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 $\widehat{e}_1, ..., \widehat{e}_n$ 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

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

  - velký rozptyl $(x_1, \ldots, x_n)$ může mít za následek velké R^2, přitom nic neříká o kvalitě modelu
  - $\mathrm{E}(R^2)$ 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

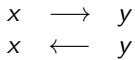
Závěr: rádia způsobují mentální onemocnění (???)

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 × příčinnost)

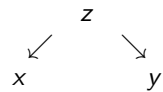
a)



Causal link (příčinná spojitost)

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

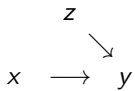
b)



Hidden cause (skrytá příčinnost)

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

c)



Confounding factor (zavádějící faktor)

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

d)

Coincidence (shoda okolností)

korelace je náhodná