Cvičení z biostatistiky 10

Poslední úprava dokumentu: 24. dubna 2024.

Lineární regrese

Prostudujte si, prosím, teorii k lineární regresi. Můžete využít buď slidy z přednášky, nebo texty *Lineární regrese I* (dále jako *LR1*) a Lineární regrese II (*LR2*), které je k dispozici na mých stránkách, či ve složce `V:/turcicova`.

1 Úvod

Datový soubor zaci.RData obsahuje údaje o prospěchu v 7. a 8. třídě u 119 dětí (58 dívek a 61 chlapců), hodnotě jejich IQ a obvodu hlavy. Najdeme zde následující proměnné:

`iq`	hodnota IQ (není známa u 8 dětí);
`zn7`	průměrná známka na pololetním vysvědčení v 7. třídě;
`zn8`	průměrná známka na pololetním vysvědčení v 8. třídě;
`Gender`	pohlaví (M = *chlapec*, F = *dívka*);
`obvod_hlavy`	obvod hlavy v cm.

1) Načtěte data do RStudia:.

```
load("zaci.RData")
```

Ujistěte se, že se všechny proměnné načetly tak, jak mají a že proměnná `Gender` se načetla jako faktora. Pokud ne, napravte to.

```
class(zaci$Gender)
zaci$Gender <- as.factor(zaci$Gender)
save(zaci, file = "zaci.RData")
attach(zaci)
```

✧ Prohlédněte si data, abyste si ověřili, zda se vše v pořádku načetlo.

```
View(zaci)
```

2) Prohlédněte si základní popisné statistiky pro proměnné z datového souboru.

```
summary(zaci)
```

2 Jednoduchá regrese

Závislost IQ na známkách ze 7. třídy

Jelikož hodnota IQ se zjišťuje na základě složitých psychologických testů, bylo by šikovné umět ji alespoň orientačně odhadnout na základě školních známek. Pro modelování IQ si zvolíme lineární regresní model a jako vysvětlující proměnnou použijeme známky z pololetí 7. třídy.

Označme jako Y_i IQ náhodně vybraného dítěte a jako x_i jeho průměrnou známku na pololetním vysvědčení v 7. třídě. Odhadněme pomocí regresní přímky, tj. pomocí modelu

$$\mathsf{E}_{x_i}(Y_i) = \beta_0 + \beta_1 x_i$$

závislost IQ na známkách ze 7. třídy.

3) Odhadněme parametry navrženého modelu.

```
mod1 <- lm(iq ~ zn7)
summary(mod1)
```

✧ Co znamenají a jak se interpretují jednotlivé údaje ve výstupu?

```
Call:
lm(formula = IQ ~ ZN7)

Residuals:
    Min      1Q  Median      3Q     Max
-31.779  -7.533  -0.100   7.900  34.873

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  137.742      2.560   53.80   <2e-16 ***
ZN7          -15.567      1.269  -12.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.24 on 179 degrees of freedom
Multiple R-squared:  0.4569,    Adjusted R-squared:  0.4538
F-statistic: 150.6 on 1 and 179 DF,  p-value: < 2.2e-16
```

popisné statistiky pro rezidua

odhad regresních koeficientů

chyby odhadu

t-statistika pro test nulovosti příslušného koeficientu (text LM2 vzorec (9))

=b0

=b1

p-hodnota testu nulovosti příslušného koeficientu

odmocnina z reziduálního rozptylu

koeficient determinace (text LM2, vzorec (12))

4) Spočtěme intervaly spolehlivosti pro regresní koeficienty.

```
confint(mod1, level=0.95)
```

✧ Jaká je interpretace těchto intervalů spolehlivosti?

5) Nakresleme odhadnutou regresní přímku do bodového grafu (barvami zde dále odlišíme chlapce a dívky).

```
COL <- c(F="pink3", M="skyblue")
PCH <- c(F=6, M=17)
plot(iq ~ zn7, col=COL[Gender], pch=PCH[Gender])
abline(mod1, col="darkgreen")
```

6) Grafické ověření předpokladů:

```
par(mfrow=c(2, 2))
plot(mod1, ask=FALSE)
par(mfrow=c(1, 1))
```

7) Ověření homoskedasticity (konstantnost chybového rozptylu) testem:

```
library(lmtest)          # knihovna obsahující Breusch-Paganův test
bptest(mod1)
```

8) Ověřme normalitu chyb pomocí Shapiro-Wilkova testu aplikovaného na (standardizovaná) rezidua.

```
shapiro.test(rstandard(mod1))
```

9) Pomocí lineární regrese prozkoumejte závislost IQ na známkách z 8. třídy. Nezapomeňte ověřit veškeré předpoklady.

3 Mnohonásobná regrese

Známky v 7. či 8. třídě nemusejí být jedinou veličinou, která má na IQ vliv a z níž by bylo možné ho odhadnout. Zkusme nyní vytvořit model mnohonásobné regrese, do něhož zařadíme veškeré dostupné veličiny a pomocí testu významnosti zjistíme, které z nich jsou pro predikci IQ opravdu důležité.

Musíme ale dát pozor, protože regresory (nezávisle proměnné) by měly být pokud možno vzájemně nezávislé. V praxi by jejich korelace neměla být vyšší než 0.8. Spočtěme si hodnoty korelačních koeficientů mezi jednotlivými proměnnými.

```
cor(zaci[,c("zn7", "zn8","obvod_hlavy")], use="complete.obs")
```

Jak vidíme z korelační matice výše, známky ze 7. a 8. třídy jsou silně korelovány, a tudíž do modelu zařadíme pouze jednu z nich (např. `zn7`). Jinak bychom v modelu měli problém s multikolinearitou regresorů (do modelu bychom dávali dvakrát skoro tutéž informaci).

✧ Argumentem `use="complete.obs"` říkáme R, aby k výpočtu korelace použil jen kompletní dvojice (to se hodí v případě chybějících pozorování), jinak bychom mohli místo hodnoty korelace dostat *NA* (*not available*).

✧ Jaká je interpretace spočtených korelačních koeficientů?
Koeficienty měří sílu lineární závislosti příslušné dvojice veličin.

Závislost IQ na známkách ze 7. třídy, pohlaví a obvodu hlavy

Označme si nyní jako Y_i IQ i-tého dítěte, jako x_i jeho průměrnou známku na pololetním vysvědčení v 7. třídě, jako v_i obvod jeho hlavy a jako z_i jeho pohlaví, přičemž $z_i = 0$ pro dívky a $z_i = 1$ pro chlapce. (Toto kódování bude v souladu s tím, jak kategorie veličiny Gender vnímá R: `F` je v abecedě před `M`, proto kategorii `F` bude vnímat jako základní, kategorie `M` bude následovat po ní.)

Náš model má tedy tvar

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1\, x_i + \beta_2\, z_i + \beta_3\, v_i$$

neboli

$$Y_i = \beta_0 + \beta_1\, x_i + \beta_2\, z_i + \beta_3\, v_i + e_i,$$

kde $i = 1, 2, \ldots, 119$. Předpokládáme opět, že $e_i \sim N(0, \sigma^2)$.

1) Odhadněme parametry modelu.

```
mod3 <- lm(iq ~ zn7 + Gender + obvod_hlavy)
summary(mod3)
```

Náš odhadnutý regresní model má tvar

$$Y_i = 159.1 - 17.9\, x_i + 4.5\, z_i - 0.4\, v_i + u_i, \quad i = 1, 2, \ldots, 119.$$

Odhadem společného rozptylu σ^2 náhodných chyb je reziduální rozptyl $s^2 = 10.81^2 = 116.86$.

2) Jaký podíl variability IQ tento model vysvětluje?
Koeficient determinace je roven 0.495, tedy závislostí na známkách v 7. třídě jsme vysvětlili 49.5 % variability IQ.

3) Jaká je interpretace odhadnutých regresních koeficientů?

b_0 = odhad absolutního členu pro regresní přímku dívek
(teoreticky by to byl odhad střední hodnoty IQ pro dívku, které měla na vysvědčení 0, ale zde to nemá praktický význam...)

$b_0 + b_2$ =odhad absolutního členu regresní přímky chlapců

b_1 =odhad změny IQ při jednotkové změně ve známkách a shodném pohlaví i obvodu hlavy

b_3 =odhad změny IQ při změně obvodu hlavy o 1 cm a shodné známce i pohlaví

4) Jsou všechny regresory v modelu významné?

Na 5% hladině lze z modelu vyloučit obvod hlavy, neboť data neprokázala, že by tento regresor byl v modelu významný (nezamítáme hypotézy $H_0 : \beta_3 = 0$). Další proměnné (byť se v současné chvíli taky jeví nevýznamné) v modelu ponecháme, neboť jejich významnost se může po vyřazení veličiny `obvod_hlavy` změnit.

5) Ověřme předpoklady použitého modelu.

```
plot(mod3)
bptest(mod3)
shapiro.test(rstandard(mod3))
```

- ✧ tvar závislosti: na bodovém grafu reziduí a vyhlazených hodnot $\hat{Y}_i$ (1. ze 4 diagnostických grafů) vidíme, že rezidua rovnoměrně poskakují kolem 0, tedy tvar závislosti se zdá být v pořádku
- ✧ homoskedasticita (konstantnost chybového rozptylu): na stejném grafu dále pozorujeme, že rezidua tvoří kolem 0 pás konstantní tloušťky, což nasvědčuje tomu, že homoskedasticita je splněna. Nezamítá ji ani Breusch-Paganův test (p-hodnota = 0.6647).
- ✧ normalita náhodných chyb $e_1, e_2, \ldots, e_n$: body na normálním diagrmu reziduí (2. ze 4 diagnostických grafů) sledují přímku, což nasvědčuje tomu, že normální rozdělení můžeme předpokládat. Normalitu nezamítá ani Shapiro-Wilkův test (p-hodnota = 0.4986).
- ✧ na žádném z grafů nepozorujeme odlehlá nebo vlivná pozorování (která by mohla mít nepříznivý vliv na odhad modelu)

6) Jelikož obvod hlavy významně nepřispívá k vysvětlení variability IQ, vytvoříme podmodel modelu `mod3`, který bude obsahovat pouze známky ze 7. třídy a pohlaví.

Podmodel - závislost IQ na zn7 a pohlaví

Náš podmodel je tedy regresní model tvaru:

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1\, x_i + \beta_2\, z_i$$

neboli

$$Y_i = \beta_0 + \beta_1\, x_i + \beta_2\, z_i + e_i,$$

kde $i = 1, 2, \ldots, 119$. Předpokládáme opět, že $e_i \sim N(0, \sigma^2)$ a x_i značí známky v pololetí 7. třídy a z_i pohlaví i-tého dítěte. Jelikož $z_i = 0$ nebo 1, jde vlastně o dvě rovnoběžné přímky, kde β_2 představuje opravu absolutního členu pro chlapce.

1) Odhadněme parametry modelu.

```
mod2 <- lm(iq ~ zn7 + Gender)
summary(mod2)
```

Náš odhadnutý regresní model má tvar

$$Y_i = 138.1 - 17.9\, x_i + 4.5\, z_i + u_i, \quad i = 1, 2, \ldots, 119.$$

Odhadem společného rozptylu σ^2 náhodných chyb je reziduální rozptyl $s^2 = 10.77^2 = 115.99$.

✧ Jsou oba regresory v modelu významné?

Pohlaví je stále těsně významné. P-hodnota testu $H_0 : \beta_2 = 0$ (proti alternativě $H_1 : \beta_2 \neq 0$) klesla ještě o trochu níže pod 0.05. Není ale výjimkou, že při odebrání jednoho nevýznamného regresoru z modelu se stane nějaký další regresor významný, ačkoli v obecnějším modelu byl nevýznamný. Na 5% hladině nelze z modelu vyloučit ani pohlaví, ani známku. Obě proměnné průkazně přispívají k vysvětlení variability IQ dětí.

✧ Jaká je interpretace odhadnutých regresních koeficientů?

b_0 = odhad absolutního členu pro regresní přímku dívek
$b_0 + b_2$ =odhad absolutního členu regresní přímky chlapců
b_1 =odhad změny IQ při jednotkové změně ve známkách a shodném pohlaví

✧ Jak vypadá předpověď IQ pro chlapce a dívky?

$$\text{chlapci: } \hat{Y}_i = b_0 + b_2 + b_1 x_i = 142.6 - 17.9\, x_i$$
$$\text{dívky: } \hat{Y}_i = b_0 + b_1 x_i = 138.1 - 17.9\, x_i$$

✧ Jaký je rozdíl ve známce u chlapce a dívky při stejném IQ?

Hledáme, co musí platit mezi známkami chlapce a dívky (tj. vztah mezi nějakým x^{ch} a x^d), pro něž platí

$$\text{IQ}^{ch} = \text{IQ}^d$$
$$b_0 + b_1 x^{ch} + b_2 = b_0 + b_1 x^d$$

Úpravou rovnice zjistíme, že

$$x^d = x^{ch} + \frac{b_2}{b_1} = x^{ch} - 0.25$$

a tedy chlapec má při stejné hodnotě IQ jako dívka o 0.25 horší (= vyšší) známku.

✧ Čím se liší předpověď IQ pro dva chlapce při zn7 = 2 a zn7 = 1?

Stačí dosadit tyto známky do odhadnutého modelu pro chlapce a porovnat výsledek. Vidíme, že

$$b_0 + b_1 \cdot 2 + b_2 < b_0 + b_1 \cdot 1 + b_2$$

neboť $b_1 < 0$. A tudíž IQ chlapce s horší známkou je podle našeho modelu o 17.9 nižší (neboť levá a pravá strana nerovnice se liší právě o b_1, které je -17.9).

✧ Je průkazné, že znalost pohlaví přispívá (při známé známce) k předpovědi IQ?

Ano. Na hladině 5% jsme zamítli hypotézu $H_0 : \beta_2 = 0$ ve prospěch alternativy $H_1 : \beta_2 \neq 0$. P-hodnota příslušného testu vyšla 0.0424, což je méně než 0.05.

✧ Je průkazné, že znalost známky přispívá (při známém pohlaví) k předpovědi IQ?

Ano. Na hladině 5% jsme zamítli hypotézu $H_0 : \beta_1 = 0$ ve prospěch alternativy $H_1 : \beta_1 \neq 0$. P-hodnota příslušného testu vyšla menší než $2 \cdot 10^{-16}$, což je méně než 0.05.

2) Nakresleme obě odhadnuté (rovnoběžné) přímky.

```
(beta <- coef(mod2))                    # odhady regres. koef.
(koefF <- beta[1:2])                    # abs. clen a smernice pro F
(koefM <- c(beta[1]+beta[3], beta[2]))  # abs. clen a smernice pro M

plot(iq ~ zn7, col=COL[Gender], pch=PCH[Gender])
abline(koefF, col="red")
abline(koefM, col="darkblue")
```

3) Ověřme předpoklady použitého modelu.

```
plot(mod2)
bptest(mod2)
shapiro.test(rstandard(mod2))
```

✧ tvar závislosti: na bodovém grafu reziduí a vyhlazených hodnot $\hat{Y}_i$ (1. ze 4 diagnostických grafů) vidíme, že rezidua rovnoměrně poskakují kolem 0, tedy tvar závislosti se zdá být v pořádku

✧ homoskedasticita (konstantnost chybového rozptylu): na stejném grafu dále pozorujeme, že rezidua tvoří kolem 0 pás konstantní tloušťky, což nasvědčuje tomu, že homoskedasticita je splněna. Nezamítá ji ani Breusch-Paganův test (p-hodnota = 0.8084).

✧ normalita náhodných chyb $e_1, e_2, \ldots, e_n$: body na normálním diagrmu reziduí (2. ze 4 diagnostických grafů) sledují přímku, což nasvědčuje tomu, že normální rozdělení můžeme předpokládat. Normalitu nezamítá ani Shapiro-Wilkův test (p-hodnota = 0.5232).

✧ na žádném z grafů nepozorujeme odlehlá nebo vlivná pozorování (která by mohla mít nepříznivý vliv na odhad modelu)

4 Mnohonásobná regrese - model s interakcemi

Nyní se nabízí otázka: Je tento model opravdu nejlepší možný? Nemůžou známky interagovat s pohlavím, tj. jsou přímky skutečně rovnoběžné? Pro ilustraci do hry opět vrátíme i obvod hlavy a vyjdeme z našeho obecného modelu. Odhadněme model, v kterém připustíme různoběžné přímky. To jest, odhadněme regresní model:

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1\, x_i + \beta_2\, z_i + \beta_3\, x_i\, z_i + \beta_4\, v_i,$$

který lze také ekvivalentně zapsat jako

$$Y_i = \beta_0 + \beta_1\, x_i + \beta_2\, z_i + \beta_3\, x_i\, z_i + \beta_4\, v_i + e_i,$$

kde $i = 1, 2, \ldots, 119$. Opět předpokládáme, že $e_i \sim N(0, \sigma^2)$. Do modelu jsme nyní přidali tzv. **interakci**, konkrétně se jedná o interakci regresorů zn7 a Gender. Interakci do modelu přidáváme, pokud se jednodušší tvary modelu nezdají být uspokojivé a máme podezření, že by kombinace regresorů mohla přinést něco nového. Jelikož $z_i = 0$ nebo 1, tak efekt interakce se zde projevuje ve

formě korekce směrnice pro chlapce. Obecně lze ale uvažovat interakce libovolných dvou regresorů, ne nutně musí být jeden z nich kategoriální, tj. teoreticky bychom mohli uvažovat interakci známek s obvodem hlavy, ale v tomto případě by to nedávalo dobrý smysl.

✧ Jaká je role jednotlivých regresních koeficientů v modelu?

$$\begin{aligned}
\beta_0 &= \text{absolutní člen pro dívky}\\
\beta_1 &= \text{směrnice pro } \texttt{zn7} \text{ pro dívky}\\
\beta_2 &= \text{oprava absolutního členu pro chlapce}\\
\beta_3 &= \text{oprava směrnice pro } \texttt{zn7} \text{ pro chlapce}\\
\beta_4 &= \text{směrnice pro } \texttt{obvod_hlavy}
\end{aligned}$$

Náš model je tedy vlastně spojením těchto dvou modelů:

$$\begin{aligned}
\text{chlapci: } Y_i &= \beta_0 + \beta_2 + (\beta_1 + \beta_3)x_i + \beta_4 v_i + e_i, \quad i = 1, 2, \ldots, 61\\
\text{dívky: } Y_j &= \beta_0 + \beta_1 x_j + \beta_4 v_j + e_j, \quad j = 1, 2, \ldots, 58
\end{aligned}$$

1) Odhadněme parametry modelu:

```
mod3i <- lm(iq ~ zn7 + Gender + zn7:Gender + obvod_hlavy)
summary(mod3i)
```

✧ Jaký je odhad jednotlivých parametrů v modelu?

$$\begin{aligned}
b_0 &= 160.1 \pm 54.7\\
b_1 &= -21.7 \pm 3.9\\
b_2 &= -3.1 \pm 7.4\\
b_3 &= 4.8 \pm 4.4\\
b_4 &= -0.3 \pm 1.0\\
s^2 &= 11.26^2
\end{aligned}$$

Všimněte si, že koeficienty β_2, β_3 a β_4 byly odhadnuty s chybou, která je stejně velká (nebo i větší) než samotný odhad. To koresponduje s faktem, že příslušné regresory nevyšly v modelu významné.

✧ Jaká je interpretace odhadnutých regresních koeficientů?

b_0 =odhad absolutního členu pro dívky

$b_0 + b_3$ =odhad absolutního členu pro chlapce

b_1 =odhad navýšení IQ dívky, která má o 1 lepší známku než jiná dívka, ale stejný obvod hlavy

$b_1 + b_2$ =odhad navýšení IQ chlapce, který má o 1 lepší známku než jiný chlapec, ale stejný obvod hlavy

b_4 =odhad navýšení IQ dítěte, které má o 1 cm větší obvod hlavy než jiné dítě, při stejných známkách i pohlaví

2) Ověřte předpoklady použitého modelu.

```
plot(mod3i)
bptest(mod3i)
shapiro.test(rstandard(mod3i))
```

Z provedených testů a diagnostických grafů nic nenasvědčuje tomu, že by předpoklady nebyly splněny. Získané výsledky lze tedy považovat za platné, a to včetně testů významnosti jednotlivých regresních koeficientů (to bude důležité pro následující odstavec).

3) Jsou všechny regresory v modelu významné? Jsou přímky významně různoběžné? Jaký je náš výsledný model?

V modelu je řada nevýznamných regresorů. P-hodnotu větší než 0.05 u testu nulovosti mají koeficienty β_2, β_3 i β_4. Opět ale pozor, neodebírejte z modelu **současně** všechny regresory, které vám vyšly nevýznamné! Odeberte pouze nejsložitější část modelu (v tomto případě interakci) a znovu odhadněte všechny parametry (dojde tak i k přepočítání příslušných p-hodnot). Přímky se tedy nezdají být významně různoběžné, protože p-hodnota u testu hypotézy $H_0 : \beta_3 = 0$ (proti alternativě $H_1 : \beta_3 \neq 0$) je větší než 0.05. Nezamítáme tedy, že interakce zn7 a Gender v modelu není významná.

Z modelu tedy odebereme tu nejsložitější část - interakci. Tím se dostáváme zpět k modelu `mod3` a jelikož (jak jsme zjistili dříve v Sekci 3) `obvod_hlavy` v něm není významný, opět ho z modelu odebereme. Tím se dostanete zpět k mod2 (dvě rovnoběžné přímky), kde, jak víme, již všechny přítomné regresory jsou významné. **Model mod2 bude tedy náš výsledný model.**