Korelační analýza, lineární regrese
16.11.2012

Úvodní nastavení.
– Otevřete si R Studio. Z internetu si stáhněte data Brain.txt .

Korelační analýza.

1. Načtěte si data Brain.txt. Datový soubor se vztahuje k výzkumné studii zabývající se souvislostí mezi velikostí mozku a inteligencí. Pro každou osobu jsou uvedeny 3 různé hodnoty IQ, velikost mozku, pohlaví, výška a hmotnost.

pohlavi	pohlaví osoby (Zena/Muz)
pIQ	hodnota IQ související s přestavivostí
vIQ	hodnota verbálního IQ
fsIQ	hodnota celkového IQ
hmotnost	hmotnost v kg
vyska	výška v cm
xMRI	velikost mozku (total pixel Count from the 18 MRI scans)

   Prohlédněte si jednotlivé proměnné z dat.

2. Zajímá nás, jak spolu souvisí IQ představivosti a verbální.
   (a) Podíváme se na obrázek, ze kterého uvidíme tuto závislost:

   ```
   plot(vIQ~pIQ)
   ```

   (b) Číselně lze tuto závislost popsat pomocí korelačního koeficientu, který je mírou lineární závislosti dvou veličin:

   ```
   cor(vIQ,pIQ)
   ```

   (c) Významnost korelačního koeficientu i otestujeme

   ```
   cor.test(vIQ,pIQ)
   ```

   (d) Výše uvedený test funguje pro normálně rozdělená data. Mají naše data normální rozdělení?
   (e) V případě, kdy nemáme normálně rozdělená data, je vhodnější použít tzv. Spearmanův korelační koeficient:

   ```
   cor.test(vIQ,pIQ,method="spearman")
   ```

3. Samostatně se podívejte na vztah výšky a celkového IQ: Vykreslete si příslušný obrázek a číselně tento vztah popište a otestujte.

4. Totéž proveďte pro velikost mozku a výšku.

Lineární regrese — regresní přímka

1. Připomeňte si algebraický zápis přímky v rovině a jeho vlastnosti.
2. Uvažujme závislost hmotnosti na výšce. Znázorněte si graf této závislosti.
3. Odhadneme model lineární závislosti hmotnosti na výšce.

   ```
   model=lm(hmotnost~vyska)

   summary(model)
   ```

   – Jaká je interpretace odhadnutých dvou parametrů?
   – Jaká je interpretace koeficientu determinace?
   – Lze prohlásit, že jsme prokázali závislost hmotnosti na výšce?
4. Graficky si proložení přímkou znázorníme následovně:

   ```
   plot(hmotnost~vyska)
   abline(model,col="red")
   ```
5. Předpokladem lineární regrese je, že náhodné chyby jsou nezávislé, normálně rozdělené s konstantním rozptylem.

   ```
   plot(model)

   r=resid(model)
   shapiro.test(r)
   ```
6. Předpovězte na základě modelu hmotnost osoby s 174 cm.

   ```
   predict(model,newdata=data.frame(vyska=174))

   # nebo "rucne"
   coef(model)[1]+coef(model)[2]*174
   ```
7. Pro předpověď lze konstruovat i interval spolehlivosti:

   ```
   predict(model,newdata=data.frame(vyska=174),interval="predict")

   predict(model,newdata=data.frame(vyska=174),interval="confidence")
   ```

Samostatná práce

1. Odhadněte model lineární závislosti velikosti mozku na výšce.
2. Jakým způsobem závisí průměrná velikost mozku na výšce? Jak se liší průměrná velikost mozku pro osoby, které se výškově liší o 5 cm?
3. Předpovězte velikost svého mozku.

Korelace, regrese

Korelace
- popisuje vzájemnou závislost dvou veličin X a Y pomocí korelačního koeficientu
- symetrická v X a Y

Lineární regrese
- popisuje vztah závisle proměnné Y na nezávisle proměnné x
- nesymetrická v x a Y
- lze použít pro předpovídání Y pro dané x
- lze uvažovat i více nezávisle proměnných (mnohonásobná regrese)

Korelační analýza

- Teoretická hodnota
$$\rho_{XY} = \frac{\mathsf{Cov}\,(X,Y)}{\sqrt{\mathsf{Var}\,(X)}\sqrt{\mathsf{Var}\,(Y)}} \in [-1,1]$$
  - míra lineární závislosti mezi X a Y
  - pro nezávislé veličiny $\rho_{XY} = 0$
  - $\rho_{XY} = \pm 1$ právě tehdy, když je jedna veličina lineární funkcí druhé, tj. $Y = aX + b$
- Výběrový korelační koeficient (též Pearsonův)
$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_j - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \in [-1,1]$$
  - odhaduje ρ na základě dat, přesnost závisí na počtu dat n
  - $r = \pm 1$ právě tehdy, když data (X_i, Y_i) leží přesně na přímce
  - test $H_0 : \rho = 0$ pro X, Y normálně rozdělené založen na $\sqrt{n-2}\dfrac{r}{\sqrt{1-r^2}}$
- Spearmanův korelační koeficient r_S
  - měří sílu monotónní závislosti
  - korelační koeficient spočítaný pro pořadí X a pořadí Y
  - je roven ± 1 právě tehdy, když je jedna veličina monotonní funkcí druhé
  - test významnosti bez předpokladu normálního rozdělení

Regresní přímka Předpokládá, že střední hodnota závisle proměnné Y je lineární funkcí nezávisle proměnné x,

$$\mathsf{E}Y = \beta_0 + \beta_1 x.$$

V praxi máme k dispozici data $(x_1, Y_1), \ldots, (x_n, Y_n)$, pro která platí

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

kde e_i jsou náhodné chyby s normálním rozdělením, nulovou střední hodnotou a shodným rozptylem. Koeficienty závislosti β_0, β_1 odhadneme z naměřených dat metodou nejmenších čtverců,

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \to \min .$$

Tyto odhady označíme $\hat{\beta}_0, \hat{\beta}_1$.

- Test $H_0 : \beta_1 = 0$ se provádí posouzením $\dfrac{\hat{\beta}_1}{S.E.(\hat{\beta}_1)}$.
- Koeficient determinace je podíl vysvětlené variability vůči celkové variabilitě

  $$R^2 = \frac{\text{vysvětlená variabilita}}{\text{celková variabilita}} \in [0, 1].$$

  Čím vyšší číslo, tím „lepší model“. Koeficient determinace ukazuje, zda má smysl provádět předpovědi pomocí modelu.
  V případě regresní přímky $R^2 = r^2$, kde r je korelační koeficient mezi Y a x.