SEMINARSKA NALOGA IZ STATISTIKE

UL FMF, Matematika — univerzitetni študij

Nace Kapus

2017/18

1. (a) Verjetnost, da posamezen anketiranec odgovori z DA, je enaka:

$$\begin{aligned} P(\text{DA}) &= P(\text{DA na delikatno vprašanje})P(\text{delikatno vprašanje}) \\ &+ P(\text{DA na nevtralno vprašanje})P(\text{nevtralno vprašanje}) \\ &= pq + (1-p)r := s. \end{aligned}$$

Da je ta verjetnost enaka za vse anketirance, sledi iz naših predpostavk.

(b) Označimo z I_i indikator dogodka, da i-ti anketiranec odgovori DA. Potem lahko S zapišemo kot vsoto indikatorjev na sledeč način:

$$S = \frac{1}{n}\sum_{i=1}^{n} I_i.$$

Računamo:

$$\begin{aligned} E(S) = E\left(\frac{1}{n}\sum_{i=1}^{n} I_i\right) &= \frac{1}{n}\sum_{i=1}^{n} E(I_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} s = \frac{1}{n}\cdot ns \\ &= s \end{aligned}$$

Torej je S res nepristranska cenilka za s.

(c)

$$\begin{aligned} \text{var}(S) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} I_i\right) &= \frac{1}{n^2}\,\text{var}\left(\sum_{i=1}^{n} I_i\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n}\text{var}(I_i) + \sum_{\substack{i,j=1 \\ i\neq j}}^{n}\text{cov}(I_i, I_j)\right) \end{aligned}$$

Ker so I_i indikatorji oziroma Bernoullijeve slučajne spremenljivke, zanje velja: $E(I_i) = P(I_i) = s$ in $\text{var}(I_i) = s(1-s)$. Izračunajmo še $\text{cov}(I_i, I_j)$:

$$\text{cov}(I_i, I_j) = E(I_iI_j) - E(I_i)E(I_j)$$

Naj bo N velikost populacije. $E(I_iI_j) = P(I_i = 1, I_j = 1)$, kar se lahko zgodi na več načinov:

- obema je dodeljeno delikatno vprašanje in oba imata lastnost A
- obema je dodeljeno nevtralno vprašanje in oba imata lastnost B
- nekdo dobi delikatno vprašanje in ima lastnost A, drugi dobi nevtralno vprašanje in ima lastnost B

$$\begin{aligned}
E(I_iI_j) &= P(I_i = 1, I_j = 1) \\
&= q \cdot \frac{qN-1}{N-1} \cdot p^2 \\
&+ (1-q) \cdot \frac{(1-q)N-1}{N-1} \cdot (1-p)^2 \\
&+ 2q \cdot \frac{(1-q)N}{N-1} \cdot p(1-p) \\
&= \frac{N}{N-1}s^2 - \frac{qp^2 + (1-q)(1-p)^2}{N-1}
\end{aligned}$$

Če povzamemo:

$$\begin{aligned}
\mathrm{var}(S) &= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathrm{var}(I_i) + \sum_{\substack{i,j=1\\ i\neq j}}^{n} \mathrm{cov}(I_i, I_j)\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} s(1-s) + \sum_{\substack{i,j=1\\ i\neq j}}^{n} \left(\frac{N}{N-1}s^2 - \frac{qp^2 + (1-q)(1-p)^2}{N-1} - s^2\right)\right) \\
&= \frac{1}{n}\left(s(1-s) + (n-1)\left(\frac{N}{N-1}s^2 - \frac{qp^2 + (1-q)(1-p)^2}{N-1} - s^2\right)\right)
\end{aligned}$$

(d) Kadar gre $N \longrightarrow \infty$, gre $\mathrm{cov}(I_i, I_j)$ proti 0. V tem primeru velja, da je

$$\mathrm{var}(S) \sim \frac{s(1-s)}{n}.$$

(e) Nepristranska cenilka za q obstaja v primeru, ko je p različen od 0 (vsaj kdo dobi delikatno vprašanje). Iz enakosti

$$s = pq + (1-p)r$$

lahko izrazimo cenilko za q, in sicer:

$$\hat{q} = \frac{S-(1-p)r}{p}.$$

(f) Iz prejšnje točke sledi:

$$\mathrm{var}(\hat{q}) = \mathrm{var}\left(\frac{S}{p}\right) \sim \frac{s(1-s)}{np^2}.$$

2. Povprečni dohodek izračunamo iz dohodkov družin, katere smo izbrali s slučajnim vzorčenjem. Če bi vzorčili ponovno, bi (najverjetneje) izbrali druge družine ter posledično dobili drugačen rezultat. Torej je to res slučajna spremenljivka.

3.

4.

5. (a)

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \quad / : \rho_i \\ \frac{Y_i}{\rho_i} &= \beta_0 \frac{1}{\rho_i} + \beta_1 \frac{x_i}{\rho_i} + \frac{\varepsilon_i}{\rho_i} \\ Z_i &= \beta_0 c_i + \beta_1 d_i + e_i \end{aligned}$$

Da nov model zadošča predpostavkam standardne linearne regresije, mora zanj veljati: $E(e_i) = 0$, $\mathrm{var}(e_i) = \sigma^2$, ter $\mathrm{cov}(e_i, e_j) = 0$ za $i \neq j$. Računamo:

$$\begin{aligned} E(e_i) &= E\left(\frac{\varepsilon_i}{\rho_i}\right) = \frac{1}{\rho_i} \cdot 0 = 0 \\ \mathrm{var}(e_i) &= \mathrm{var}\left(\frac{\epsilon_i}{\rho_i}\right) = \frac{1}{\rho_i^2}\,\mathrm{var}(\varepsilon_i) = \frac{\rho_i^2 \sigma^2}{\rho_i^2} = \sigma^2 \\ \mathrm{cov}(e_i, e_j) &= \mathrm{cov}\left(\frac{\varepsilon_i}{\rho_i}, \frac{\varepsilon_j}{\rho_j}\right) = \frac{1}{\rho_i \rho_j}\,\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0 \end{aligned}$$

Matrika modela je enaka

$$X = \begin{bmatrix} \frac{1}{\rho_1} & \frac{x_1}{\rho_1} \\ \frac{1}{\rho_2} & \frac{x_2}{\rho_2} \\ \vdots & \vdots \\ \frac{1}{\rho_n} & \frac{x_n}{\rho_n} \end{bmatrix}$$

(b) V novem modelu minimiziramo količino

$$\sum_{i=1}^{n}(z_i - \beta_o c_i - \beta_1 d_i)^2.$$

Računamo:

$$\sum_{i=1}^{n}(z_i - \beta_0 c_i - \beta_1 d_i)^2 = \sum_{i=1}^{n}\left(\frac{Y_i}{\rho_i} - \beta_0\frac{1}{\rho_i} - \beta_1\frac{x_i}{\rho_i}\right)^2 = \sum_{i=1}^{n}\frac{(Y_i - \beta_0 - \beta_1 x_i)^2}{\rho_i^2}$$

Komentar: funkcija, ki jo želimo minimizirati v novem modelu je zelo podobna funkciji izhodiščnega modela, pojavijo se le faktorji $\frac{1}{\rho_i^2}$. To pomeni, da imajo v funkciji, ki jo minimiziramo, večji pomen opažanja, ki imajo manjše variance in obratno.

(c) Ker imamo standarden linearni regresijski model vemo:

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (X^T X)^{-1} X^T z,$$

pri čemer je

$$X = \begin{bmatrix} \frac{1}{\rho_1} & \frac{x_1}{\rho_1} \\ \frac{1}{\rho_2} & \frac{x_2}{\rho_2} \\ \vdots & \vdots \\ \frac{1}{\rho_n} & \frac{x_n}{\rho_n} \end{bmatrix} in\ z = \begin{bmatrix} \frac{y_1}{\rho_1} \\ \frac{y_2}{\rho_2} \\ \vdots \\ \frac{y_n}{\rho_n} \end{bmatrix}.$$

Izračunamo:

$$X^T X = \begin{bmatrix} \sum_{i=1}^n \frac{1}{\rho_i^2} & \sum_{i=1}^n \frac{x_i}{\rho_i^2} \\ \sum_{i=1}^n \frac{x_i}{\rho_i^2} & \sum_{i=1}^n \frac{x_i^2}{\rho_i^2} \end{bmatrix},\ X^T z = \begin{bmatrix} \sum_{i=1}^n \frac{y_i}{\rho_i^2} \\ \sum_{i=1}^n \frac{x_i y_i}{\rho_i^2} \end{bmatrix}.$$

Od tod sledi:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n \frac{x_i^2}{\rho_i^2}\sum_{i=1}^n \frac{y_i}{\rho_i^2} - \sum_{i=1}^n \frac{x_i}{\rho_i^2}\sum_{i=1}^n \frac{x_i y_i}{\rho_i^2}}{\sum_{i=1}^n \frac{1}{\rho_i^2}\sum_{i=1}^n \frac{x_i^2}{\rho_i^2} - \sum_{i=1}^n \frac{x_i}{\rho_i^2}\sum_{i=1}^n \frac{x_i}{\rho_i^2}}$$

$$\hat{\beta}_1 = \frac{-\sum_{i=1}^n \frac{x_i}{\rho_i^2}\sum_{i=1}^n \frac{y_i}{\rho_i^2} + \sum_{i=1}^n \frac{1}{\rho_i^2}\sum_{i=1}^n \frac{x_i y_i}{\rho_i^2}}{\sum_{i=1}^n \frac{1}{\rho_i^2}\sum_{i=1}^n \frac{x_i^2}{\rho_i^2} - \sum_{i=1}^n \frac{x_i}{\rho_i^2}\sum_{i=1}^n \frac{x_i}{\rho_i^2}}.$$

(d) Po formuli iz predavanj:

$$\text{var}\left(\overset{\wedge}{\beta_0}\right) = \sigma^2\left[\left(X^TX\right)^{-1}\right]_{11} = \sigma^2 \cdot \frac{\sum_{i=1}^{1}\frac{x_i^2}{\rho_i^2}}{\sum_{i=1}^{n}\frac{1}{\rho_i^2}\sum_{i=1}^{n}\frac{x_i^2}{\rho_i^2} - \sum_{i=1}^{n}\frac{x_i}{\rho_i^2}\sum_{i=1}^{n}\frac{x_i}{\rho_i^2}}$$

$$var\left(\overset{\wedge}{\beta_1}\right) = \sigma^2\left[\left(X^TX\right)^{-1}\right]_{22} = \sigma^2 \cdot \frac{\sum_{i=1}^{1}\frac{1}{\rho_i^2}}{\sum_{i=1}^{n}\frac{1}{\rho_i^2}\sum_{i=1}^{n}\frac{x_i^2}{\rho_i^2} - \sum_{i=1}^{n}\frac{x_i}{\rho_i^2}\sum_{i=1}^{n}\frac{x_i}{\rho_i^2}}$$