# Examen

Soluții

*27 Iunie 2018*

**Exercițiul 1**

12p

1. Reamintim că estimatorii lui $\alpha$ și $\beta$ obținuți prin metoda celor mai mici pătrate sunt dați de

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{și} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

iar variantele lor sunt

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{și} \quad Var(\hat{\alpha}) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

4p

a) Dacă $\beta$ este cunoscut iar $\alpha$ este necunoscut atunci estimatorul $\tilde{\alpha}$ corespunde la valoarea care minimizează funcția

$$S(\alpha) = \sum_{i=1}^{n}\left[y_i - (\alpha + \beta x_i)\right]^2.$$

Rezolvând ecuația $S'(\alpha) = 0$ obținem

$$S'(\alpha) = -2\sum_{i=1}^{n}\left[y_i - (\alpha + \beta x_i)\right] = 0 \iff \tilde{\alpha} = \hat{y} - \beta\bar{x}.$$

4p

b) Pentru a calcula varianța lui $\tilde{\alpha}$ să observăm că, folosind relația $y_i = \alpha + \beta x_i$,

$$\tilde{\alpha} = \hat{y} - \beta\bar{x} = \underbrace{\left(\alpha + \beta\bar{x} + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right)}_{\hat{y}} - \beta\bar{x} = \alpha + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$$

și cum variabilele aleatoare $\varepsilon_i$ sunt centrate, necorelate și de varianță $\sigma^2$ găsim că

$$Var(\tilde{\alpha}) = \frac{\sigma^2}{n}.$$

Cum $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \leq \sum_{i=1}^{n} x_i^2$ deducem că $Var(\tilde{\alpha}) \leq Var(\hat{\alpha})$.

4p

c) Presupunând acum că $\alpha$ este cunoscut și $\beta$ este necunoscut avem că estimatorul $\tilde{\beta}$ corespunde la valoarea care minimizează funcția

$$S(\beta) = \sum_{i=1}^{n}\left[y_i - (\alpha + \beta x_i)\right]^2.$$

Rezolvând ecuația $S'(\beta) = 0$ găsim

$$S'(\beta) = -2\sum_{i=1}^{n} x_i\left[y_i - (\alpha + \beta x_i)\right] = 0 \iff \tilde{\beta} = \frac{\sum_{i=1}^{n} x_i(y_i - \alpha)}{\sum_{i=1}^{n} x_i^2}.$$

În mod similar cu punctul anterior putem rescrie $\tilde{\beta}$ prin

$$\tilde{\beta} = \frac{\sum_{i=1}^{n} x_i(y_i - \alpha)}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i(\overbrace{\alpha + \beta x_i + \varepsilon_i}^{y_i} - \alpha)}{\sum_{i=1}^{n} x_i^2} = \beta + \frac{\sum_{i=1}^{n} x_i\varepsilon_i}{\sum_{i=1}^{n} x_i^2},$$

ceea ce conduce la $Var(\tilde{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$.

Remarcăm că $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \leq \sum_{i=1}^{n} x_i^2$ ceea ce arată că $Var(\tilde{\beta}) \leq Var(\hat{\beta})$.

8p

2. Avem modelele

$$\begin{aligned} y_i &= \alpha + \beta x_i + \varepsilon_i \\ y_i &= \alpha' + \beta'(x_i - \bar{x}) + \varepsilon_i \\ &= \alpha' + \beta' z_i + \varepsilon_i \end{aligned}$$

4p

a) Deoarece $\bar{z} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ deducem că

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} z_i(y_i - \bar{y})}{\sum_{i=1}^{n} z_i^2} = \hat{\beta}'.$$

De asemenea avem $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ și respectiv $\hat{\alpha}' = \bar{y} - \hat{\beta}'\bar{z} = \bar{y}$ ceea ce arată că $\hat{\alpha}' \neq \hat{\alpha}$. Mai mult observăm că $\hat{\alpha}' \sim \mathcal{N}\left(\alpha + \beta\bar{z}, \frac{\sigma^2}{n}\right) = \mathcal{N}\left(\alpha, \frac{\sigma^2}{n}\right)$.

4p

b) Pentru a verifica necorelarea dintre $\hat{\alpha}'$ și $\hat{\beta}'$ să observăm că din

$$\hat{\alpha}' = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\beta}' = \frac{\sum_{i=1}^{n} z_i(y_i - \bar{y})}{\sum_{i=1}^{n} z_i^2} = \sum_{j=1}^{n}\left(\frac{z_j}{\sum_{i=1}^{n} z_i^2}\right) y_j$$

avem

$$Cov\left(\hat{\alpha}', \hat{\beta}'\right) = -\sigma^2\sum_{j=1}^{n}\frac{1}{n}\left(\frac{z_j}{\sum_{i=1}^{n} z_i^2}\right) = 0.$$

Altfel se poate consulta exercițiul 1.4 din Seminarul 2.

### Exercițiul 2

5p

1. Pentru a determina valorile lipsă vom folosi proprietatea de simetrie a matricei $\boldsymbol{X}^\intercal\boldsymbol{X}$, i.e. $\boldsymbol{X}^\intercal\boldsymbol{X} = (\boldsymbol{X}^\intercal\boldsymbol{X})^\intercal$ :

$$\boldsymbol{X}^\intercal\boldsymbol{X} = \begin{pmatrix} ? & ? & 9791.6 \\ ? & 3306476 & ? \\ ? & 471237.9 & 67660 \end{pmatrix} = \underbrace{\begin{pmatrix} ? & ? & ? \\ ? & 3306476 & 471237.9 \\ 9791.6 & ? & 67660 \end{pmatrix}}_{(\boldsymbol{X}^\intercal\boldsymbol{X})^\intercal}$$

ceea ce conduce la $\boldsymbol{X}^\intercal\boldsymbol{X} = \begin{pmatrix} ? & ? & 9791.6 \\ ? & 3306476 & 471237.9 \\ 9791.6 & 471237.9 & 67660 \end{pmatrix}$. De asemenea putem observa că matricea de design este $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{x}, \boldsymbol{z})$ unde $\boldsymbol{z} = \sqrt{\boldsymbol{x}} = (\sqrt{x_1}, \cdots, \sqrt{x_n})$ ceea ce conduce la

$$\boldsymbol{X}^\intercal\boldsymbol{X} = \begin{pmatrix} \boldsymbol{1}^\intercal \\ \boldsymbol{x}^\intercal \\ \boldsymbol{z}^\intercal \end{pmatrix} \begin{pmatrix} \boldsymbol{1} & \boldsymbol{x} & \boldsymbol{z} \end{pmatrix} = \begin{pmatrix} \boldsymbol{1}^\intercal\boldsymbol{1} & \boldsymbol{1}^\intercal\boldsymbol{x} & \boldsymbol{1}^\intercal\boldsymbol{z} \\ \boldsymbol{x}^\intercal\boldsymbol{1} & \boldsymbol{x}^\intercal\boldsymbol{x} & \boldsymbol{x}^\intercal\boldsymbol{z} \\ \boldsymbol{z}^\intercal\boldsymbol{1} & \boldsymbol{z}^\intercal\boldsymbol{x} & \boldsymbol{z}^\intercal\boldsymbol{z} \end{pmatrix} = \begin{pmatrix} n & n\bar{x} & n\bar{z} \\ n\bar{x} & \boldsymbol{x}^\intercal\boldsymbol{x} & \boldsymbol{x}^\intercal\boldsymbol{z} \\ n\bar{z} & \boldsymbol{z}^\intercal\boldsymbol{x} & \boldsymbol{z}^\intercal\boldsymbol{z} \end{pmatrix}$$

și cum $\boldsymbol{z}^\intercal\boldsymbol{z} = n\bar{x}$, iar $n = 1429$ rezultă că

$$\boldsymbol{X}^\intercal\boldsymbol{X} = \begin{pmatrix} 1429 & 67660 & 9791.6 \\ 67660 & 3306476 & 471237.9 \\ 9791.6 & 471237.9 & 67660 \end{pmatrix}.$$

5p

2. Valoarea diametrului mediu este dată de

$$\bar{x} = \frac{(\boldsymbol{X}^\intercal\boldsymbol{X})_{1,2}}{n} = \frac{67660}{1429} \approx 47.3\,cm.$$

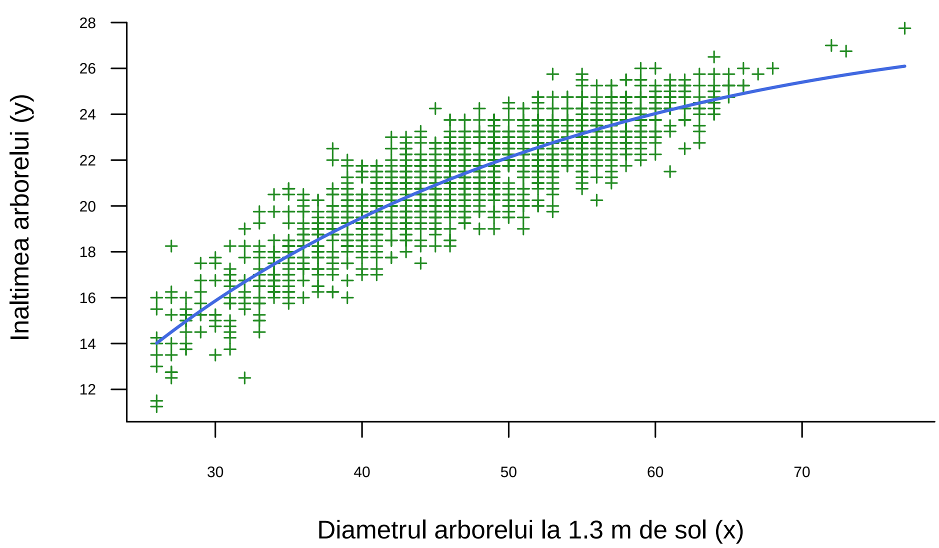
Media empirică a observațiilor $y_i$ se poate deduce din prima componentă a vectorului $\boldsymbol{X}^\intercal\boldsymbol{Y}$

$$\bar{y} = \frac{(\boldsymbol{X}^\intercal\boldsymbol{Y})_1}{n} = \frac{30312.5}{1429} = 21.21\,m.$$

5p

3. Estimatorul lui $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\intercal$ obținut prin metoda celor mai mici pătrate este:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\intercal\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\intercal\boldsymbol{Y} = \begin{pmatrix} -24.352 \\ -0.482 \\ 9.986 \end{pmatrix}$$




5p

4. Estimatorul nedeplasat $\hat{\sigma}^2$ a lui $\sigma^2$ este (a se vedea exercițiul 2.5 din Seminarul 2 și figura din secțiunea 2.3) dat de expresia

$$\hat{\sigma}^2 = \frac{\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{n-(p+1)} = \frac{\|\boldsymbol{Y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{n-(p+1)}$$

și cum

$$\|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 = \hat{\boldsymbol{\beta}}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X}\underbrace{(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{Y}}_{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}}^\intercal \boldsymbol{X}^\intercal \boldsymbol{Y}$$

găsim că

$$\hat{\sigma}^2 = \frac{\boldsymbol{Y}^\intercal\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^\intercal \boldsymbol{X}^\intercal \boldsymbol{Y}}{n-(p+1)} = \frac{651857.9 - 650017.2}{1429-(2+1)} \approx 1.29.$$

Deoarece

$$T_2 = \frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}\sqrt{[(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}]_{3,3}}} = \frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_{\hat{\beta}_2}} \sim t_{n-(p+1)} = t_{n-3}$$

găsim că un interval de încredere de nivel de încredere $\alpha = 95\%$ este

$$I(\beta_2) = \left[\hat{\beta}_2 - t_{n-3}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}\sqrt{[(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}]_{3,3}}, \hat{\beta}_2 + t_{n-3}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}\sqrt{[(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}]_{3,3}}\right]$$

și ținând cont că $t_{n-3}(0.975) = t_{1426}(0.975) \approx 1.96$ (din Teorema Limită Centrală) obținem

$$I(\beta_2) \approx [8.456, 11.517].$$

5p

5. Vrem să testăm ipotezele $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ la un nivel de semnificație de 10%. Sub ipoteza nulă avem că

$$\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{[(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}]_{2,2}}} =\sim t_{n-(p+1)} = t_{n-3} \approx \mathcal{N}(0,1)$$

prin urmare este suficient să comparăm valoarea absolută a statisticii de test cu cea a cuantilei de ordin 0.95 pentru repartiția normală (adică 1.645):

$$\left|\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}\right| = \frac{|-0.482|}{\sqrt{1.29}\times\sqrt{0.002}} \approx 8.33 > 1.645$$

În consecință respingem ipoteza nulă $H_0 : \beta_1 = 0$.

5p

6. Notând cu $\boldsymbol{x}_{n+1}^\intercal = (1, x_{n+1}, \sqrt{x_{n+1}}) = (1, 49, 7)$, valoarea prezisă $y_{n+1}$ a variabilei răspuns este

$$y_{n+1} = \boldsymbol{x}_{n+1}^\intercal\hat{\boldsymbol{\beta}} \approx 21.89\, m$$

iar un interval de predicție este

$$I(y_{n+1}) = \left[\boldsymbol{x}_{n+1}^\intercal\hat{\boldsymbol{\beta}} - t_{n-(p+1)}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}\sqrt{1+\boldsymbol{x}_{n+1}^\intercal(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}}, \boldsymbol{x}_{n+1}^\intercal\hat{\boldsymbol{\beta}} + t_{n-(p+1)}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}\sqrt{1+\boldsymbol{x}_{n+1}^\intercal(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}}\right]$$

care revine la $I(y_{n+1}) \approx [19.66, 24.12]$.

În mod similar găsim un interval de predicție și pentru $y_{n+1}$ atunci când $\boldsymbol{x}_{n+1}^{\intercal} = (1, x_{n+1}, \sqrt{x_{n+1}}) = (1, 25, 5)$: $I(y_{n+1}) \approx [11.25, 15.76]$.