Exerciții de seminar 2

Regresie

Obiectivul acestui seminar este de a prezenta câteva exerciții de regresie liniară.

1 Regresie liniară simplă

1.1 Metoada celor mai mici pătrate

Ex. 1.1 Arătați că estimatorii $\hat{\beta}_0$ și $\hat{\beta}_1$ obținuți prin metoda celor mai mici pătrate, adică valorile coeficienților β_0 și β_1 care minimizează funcția

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

sunt dați de expresiile

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{și} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Trebuie să determinăm

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1) \in \mathbb{R} \times \mathbb{R}}{\arg\min}\ S(\beta_0, \beta_1)$$

și observând că funcția $S(\beta_0, \beta_1)$ este convexă ea admite un punct de minim. Acesta se obține ca soluție a sistemului $\nabla S = 0$ de ecuați normale,

$$\begin{cases} \frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

Din prima ecuație obținem prin sumare $n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$ ceea ce conduce la $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

A doua ecuație conduce la

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

și înlocuind β_0 cu expresia obținută anterior, obținem soluția

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

De asemenea, se poate verifica că $S(\beta_0, \beta_1)$ se scrie sub forma

$$
\begin{aligned}
S(\beta_0, \beta_1) = n\left[\beta_0 - (\bar{y} - \beta_1\bar{x})\right]^2 + \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\beta_1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^2 \\
+ \left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]\left[1 - \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2\sum_{i=1}^{n}(y_i - \bar{y})^2}\right]
\end{aligned}
$$

care justifică în egală măsură soluția obținută anterior.

Ex. 1.2

Arătați că estimatorii obținuți prin metoda celor mai mici pătrate, $\hat{\beta}_0$ și $\hat{\beta}_1$, sunt estimatori nedeplasați.

Coeficienții $\hat{\beta}_0$ și $\hat{\beta}_1$ obținuți prin metoda celor mai mici pătrate sunt dați de $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ și $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ (acești sunt variabile aleatoare deoarece sunt funcții de Y_i care sunt variabile aleatoare). Înlocuind în expresia lui $\hat{\beta}_1$ pe y_i cu $\beta_0 + \beta_1 x_i + \varepsilon_i$ avem

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\beta_0 \overbrace{\sum_{i=1}^{n}(x_i - \bar{x})}^{=0} + \beta_1\sum_{i=1}^{n}(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\beta_1\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.
\end{aligned}
$$

Conform ipotezei modelului de regresie liniară simplă, $\mathbb{E}[\varepsilon_i] = 0$, prin urmare $\mathbb{E}[\hat{\beta}_1] = \beta_1$ ceea ce arată că $\hat{\beta}_1$ este un estimator nedeplasat pentru β_1.

În mod similar,

$$
\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\bar{y}] - \bar{x}\mathbb{E}[\hat{\beta}_1] = \beta_0 + \bar{x}\beta_1 - \bar{x}\beta_1 = \beta_0
$$

ceea ce arată că $\hat{\beta}_0$ este un estimator nedeplasat pentru β_0.

Ex. 1.3

Calculați matricea de varianță-covarianță a estimatorilor $\hat{\beta}_0$ și $\hat{\beta}_1$.

Notăm cu $W = \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{pmatrix}$ matricea de varianță-covarianță a estimatorilor $\hat{\beta}_0$ și $\hat{\beta}_1$.

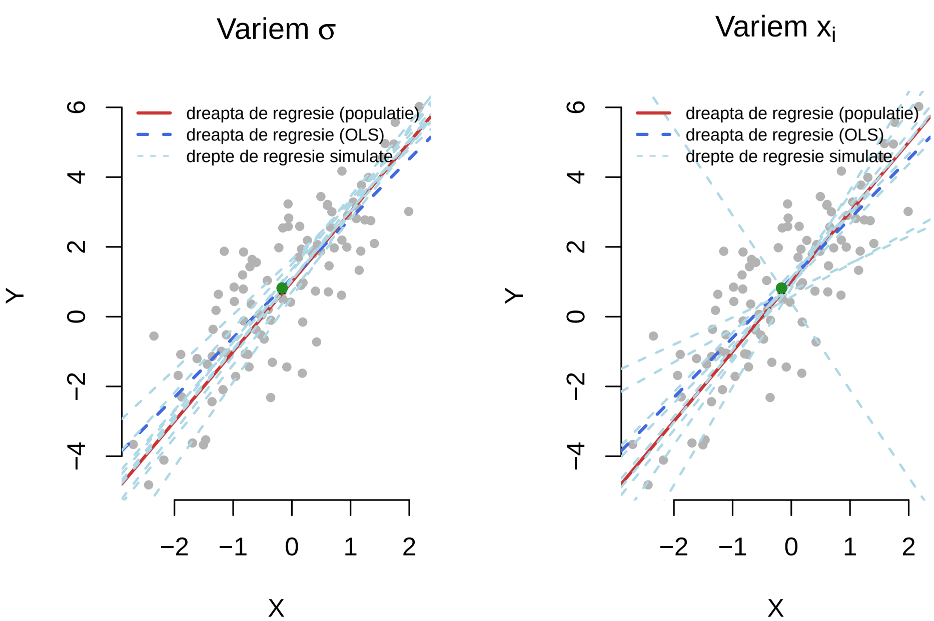
Avem, folosind expresia lui $\hat{\beta}_1$ determinată la punctul anterior și homoscedasticitatea și necorelarea erorilor $Cov(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$, că

$$
\begin{aligned}
Var(\hat{\beta}_1) &= Var\left(\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\\
&= \frac{Var\left(\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i\right)}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} = \frac{\sum_{i,j} (x_i - \bar{x})(x_j - \bar{x})Cov(\varepsilon_i, \varepsilon_j)}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2}\\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})^2\sigma^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}
$$

Din expresia $Var(\hat{\beta}_1)$ observăm că dacă σ^2 este mică (cu alte cuvinte y_i sunt aproape de dreapta de regresie) atunci estimarea este mai precisă. De asemenea, se constată că pe măsură ce valorile x_i sunt mai dispersate în jurul valorii medii $\bar{x}$ estimarea coeficientului $\hat{\beta}_1$ este mai precisă ($Var(\hat{\beta}_1)$ este mai mică). Acest fenomen se poate observa și în figura de mai jos în care am generat 100 de valori aleatoare X și 100 de valori pentru Y după modelul

$$y = 1 + 2x + \varepsilon$$

cu $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Dreapta roșie descrie adevărata relație $f(x) = 1 + 2x$ în populație iar dreapta albastră reprezintă dreapta de regresie calculată cu ajutorul metodei celor mai mici pătrate (OLS). Dreptele albastru deschis au fost generate tot cu ajutorul metodei celor mai mici pătrate atunci când variem σ^2 (în figura din stânga) și respectiv pe x_i în jurul lui $\bar{x}$ (în figura din dreapta).



Pentru a determina $Var(\hat{\beta}_0)$, vom folosi relația $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ ceea ce conduce la

$$
\begin{aligned}
Var(\hat{\beta}_0) &= Var(\bar{y} - \hat{\beta}_1\bar{x}) = Var(\bar{y}) - 2Cov(\bar{y}, \hat{\beta}_1\bar{x}) + Var(\hat{\beta}_1\bar{x})\\
&= Var\left(\frac{1}{n}\sum_{i=1}^n y_i\right) - 2\bar{x}Cov(\bar{y}, \hat{\beta}_1) + \bar{x}^2 Var(\hat{\beta}_1)\\
&= \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - 2\bar{x}Cov(\bar{y}, \hat{\beta}_1).
\end{aligned}
$$

Pentru $Cov(\bar{y}, \hat{\beta}_1)$ avem (ținând cont de faptul că β_0, β_1 și x_i sunt constante)

$$\begin{aligned}
Cov(\bar{y}, \hat{\beta}_1) &= Cov\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \beta_1 + \frac{\sum_{j=1}^{n}(x_j - \bar{x})\varepsilon_j}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right) = \frac{1}{n}\sum_{i=1}^{n} Cov\left(\beta_0 + \beta_1 x_i + \varepsilon_i, \beta_1 + \frac{\sum_{j=1}^{n}(x_j - \bar{x})\varepsilon_j}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} Cov\left(\varepsilon_i, \frac{\sum_{j=1}^{n}(x_j - \bar{x})\varepsilon_j}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sum_{j=1}^{n}(x_j - \bar{x})^2} Cov\left(\varepsilon_i, \sum_{j=1}^{n}(x_j - \bar{x})\varepsilon_j\right) \\
&= \frac{1}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\sum_{i=1}^{n}\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x}) Cov(\varepsilon_i, \varepsilon_i = j) = \frac{1}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\sum_{i=1}^{n}\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})\delta_{ij}\sigma^2 \\
&= \frac{\sigma^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\frac{1}{n}\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})}_{=0} = 0
\end{aligned}$$

prin urmare

$$Var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Calculul covarianței dintre $\hat{\beta}_0$ și $\hat{\beta}_1$ rezultă aplicând relațiile de mai sus

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = Cov(\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1) = Cov(\bar{y}, \hat{\beta}_1) - \bar{x}Var(\hat{\beta}_1) = -\frac{\sigma^2\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Observăm că $Cov(\hat{\beta}_0, \hat{\beta}_1) \leq 0$ iar intuitiv, cum dreapta de regresie (bazată pe estimatorii obținuți prin metoda celor mai mici pătrate) $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}$ trece prin centrul de greutate al datelor $(\bar{x}, \bar{y})$, dacă presupunem $\bar{x} > 0$ remarcăm că atunci când creștem panta (creștem $\hat{\beta}_1$) ordonata la origine scade (scade $\hat{\beta}_0$) și reciproc.

Matricea de varianță-covarianță a estimatorilor $\hat{\beta}_0$ și $\hat{\beta}_1$ devine

$$W = \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{pmatrix} = \begin{pmatrix} \frac{\sigma^2\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} & -\frac{\sigma^2\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ -\frac{\sigma^2\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} & \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{pmatrix}.$$

Ex. 1.4 Arătați că în cadrul modelului de regresie liniară simplă, suma valorilor reziduale este nulă.

Observăm, folosind definiția $\hat{\varepsilon}_i = y_i - \hat{y}_i$, că

$$\begin{aligned}
\sum_{i=1}^{n}\hat{\varepsilon}_i &= \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1) \\
&= \sum_{i=1}^{n}\left[y_i - \underbrace{(\bar{y} - \bar{x}\hat{\beta}_1)}_{=\hat{\beta}_0} - x_i\hat{\beta}_1\right] = \sum_{i=1}^{n}(y_i - \bar{y}) - \hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x}) = 0
\end{aligned}$$

Ex. 1.5

Arătați că în modelul de regresie liniară simplă statistica $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ este un estimator nedeplasat pentru σ^2.

Ținând cont de faptul că $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ și $\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{\varepsilon}$ (prin însumarea după i a relațiilor $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$) găsim că

$$\begin{aligned}
\hat{\varepsilon}_i = y_i - \hat{y}_i &= (\beta_0 + \beta_1 x_i + \varepsilon_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= (\underbrace{\bar{y} - \beta_1\bar{x} - \bar{\varepsilon}}_{=\beta_0} + \beta_1 x_i + \varepsilon_i) - (\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i) \\
&= (\beta_1 - \hat{\beta}_1)(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})
\end{aligned}$$

și prin dezvoltarea binomului și utilizând relația $\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ găsim

$$\begin{aligned}
\sum_{i=1}^{n}\hat{\varepsilon}_i^2 &= (\beta_1 - \hat{\beta}_1)^2\sum_{i=1}^{n}(x_i - \bar{x}) + \sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 + 2(\beta_1 - \hat{\beta}_1)\sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) \\
&= (\beta_1 - \hat{\beta}_1)^2\sum_{i=1}^{n}(x_i - \bar{x}) + \sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 + 2(\beta_1 - \hat{\beta}_1)\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i - 2(\beta_1 - \hat{\beta}_1)\bar{\varepsilon}\sum_{i=1}^{n}(x_i - \bar{x}) \\
&= (\beta_1 - \hat{\beta}_1)^2\sum_{i=1}^{n}(x_i - \bar{x}) + \sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 - 2(\beta_1 - \hat{\beta}_1)^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 - (\beta_1 - \hat{\beta}_1)^2\sum_{i=1}^{n}(x_i - \bar{x})^2.
\end{aligned}$$

Luând media găsim că

$$\mathbb{E}\left(\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right) = \mathbb{E}\left(\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2\right) - \sum_{i=1}^{n}(x_i - \bar{x})^2 Var(\hat{\beta}_1) = (n-1)\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

unde am folosit că $\mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2\right) = \sigma^2$ (deoarece $Var(\varepsilon_i) = \sigma^2$).

Concluzionăm că $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ este un estimator nedeplasat pentru σ^2.

Ex. 1.6

Fie x_{n+1} o nouă valoare pentru variabila X și ne propunem să prezicem valoarea y_{n+1} conform modelului

$$y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}$$

cu $\mathbb{E}[\varepsilon_{n+1}] = 0$, $Var(\varepsilon_{n+1}) = \sigma^2$ și $Cov(\varepsilon_{n+1}, \varepsilon_i) = 0$ pentru $i = 1, \dots, n$.

Arătați că varianța răspunsului mediu prezis este

$$Var(\hat{y}_{n+1}) = \sigma^2\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

iar varianța erorii de predicție $\hat{\varepsilon}_{n+1}$ satisface $\mathbb{E}[\hat{\varepsilon}_{n+1}] = 0$ și

$$Var(\hat{\varepsilon}_{n+1}) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right].$$

Cum $\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$ avem

$$\begin{aligned}
Var(\hat{y}_{n+1}) &= Var(\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}) = Var(\hat{\beta}_0) + 2Cov(\hat{\beta}_0, \hat{\beta}_1) + x_{n+1}^2 Var(\hat{\beta}_1) \\
&= \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} - 2\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\sigma^2 x_{n+1}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\left[\frac{1}{n}\sum_{i=1}^n x_i^2 - 2x_{n+1}\bar{x} + x_{n+1}^2\right] \\
&= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\left[\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 + \bar{x}^2 - 2x_{n+1}\bar{x} + x_{n+1}^2\right] \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right].
\end{aligned}$$

Constatăm că atunci când x_{n+1} este departe de valoarea medie $\bar{x}$ răspunsul mediu are o variabilitate mai mare.

Pentru a obține varianța erorii de predicție $\hat{\varepsilon}_{n+1} = y_{n+1} - \hat{y}_{n+1}$ să observăm că y_{n+1} depinde doar de ε_{n+1} pe când $\hat{y}_{n+1}$ depinde de ε_i, $i \in \{1, 2, \dots, n\}$. Din necorelarea erorilor deducem că

$$Var(\hat{\varepsilon}_{n+1}) = Var(y_{n+1} - \hat{y}_{n+1}) = Var(y_{n+1}) + Var(\hat{y}_{n+1}) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right].$$

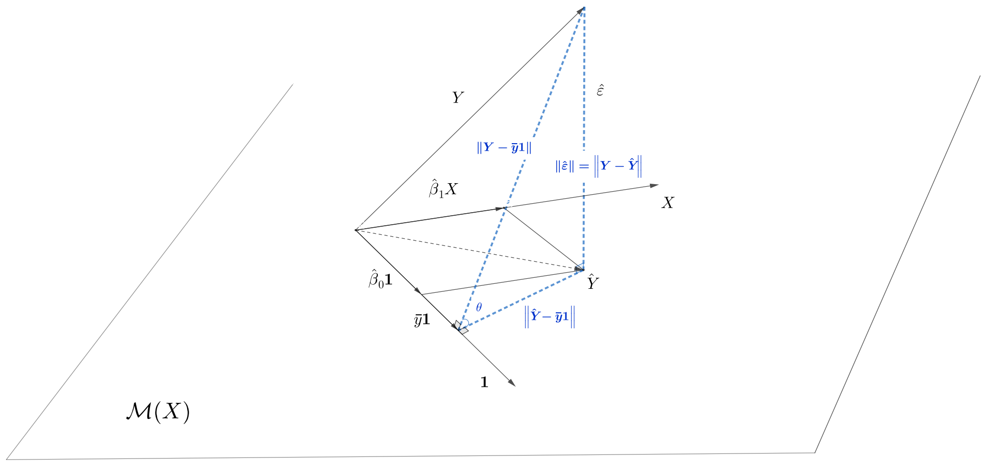
1.2 Coeficientul de determinare R^2 și coeficientul de corelație

În această secțiune încercăm să abordăm problema de regresie liniară simplă într-un context geometric. Din punct de vedere vectorial dispunem de doi vectori: vectorul $X = (x_1, x_2, \dots, x_n)^\intercal$ a celor n observații ale variabilei explicative și vectorul $Y = (y_1, y_2, \dots, y_n)^\intercal$ compus din cele n observații ale variabilei răspuns, pe care vrem să o explicăm. Cei doi vectori aparțin spațiului $\mathbb{R}^n$.

Fie $\mathbf{1} = (1, 1, \dots, 1)^\intercal \in \mathbb{R}^n$ și $\mathcal{M}(X)$ subspațiul liniar din $\mathbb{R}^n$ de dimensiune 2 generat de vectorii $\{\mathbf{1}, X\}$ (acești vectori nu sunt coliniari deoarece X conține cel puțin două elemente distincte). Notăm cu $\hat{Y}$ proiecția ortogonală a lui Y pe subspațiul $\mathcal{M}(X)$ și cum $\{\mathbf{1}, X\}$ formează o bază în $\mathcal{M}(X)$ deducem că există $\hat{\beta}_0, \hat{\beta}_1 \in \mathbb{R}$ astfel ca $\hat{Y} = \hat{\beta}_0\mathbf{1} + \hat{\beta}_1 X$. Cum, din definiția priecției ortogonale, $\hat{Y}$ este unicul vector din $\mathcal{M}(X)$ care minimizează distanța euclidiană (deci și pătratul ei)

$$\left\|Y - \hat{Y}\right\| = \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

deducem că $\hat{\beta}_0, \hat{\beta}_1$ coincid cu valorile obținute prin metoda celor mai mici pătrate. Astfel coeficienții $\hat{\beta}_0$ și $\hat{\beta}_1$ se reprezintă coordonatele proiecției ortogonale a lui Y pe subspațiul generat de vectorii $\{\mathbf{1}, X\}$ (a se vedea figura de mai jos).



Observăm că, în general, vectorii $\{\mathbf{1}, X\}$ nu formează o bază ortogonală în $\mathcal{M}(X)$ (cu excepția cazului în care $\langle \mathbf{1}, X\rangle = n\bar{x} = 0$) prin urmare $\hat{\beta}_0\mathbf{1}$ nu este proiecția ortogonală a lui Y pe $\mathbf{1}$ (aceasta este $\frac{\langle Y,\mathbf{1}\rangle}{\|\mathbf{1}\|^2}\mathbf{1} = \bar{y}\mathbf{1}$) iar $\hat{\beta}_1 X$ nu este proiecția ortogonală a lui Y pe X (aceasta fiind $\frac{\langle Y,X\rangle}{\|X\|^2}X$).

Fie $\hat{\varepsilon} = Y - \hat{Y} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \dots, \hat{\varepsilon}_n)^\intercal$ vectorul valorilor reziduale. Aplicând Teorema lui Pitagora (în triunghiul albastru) rezultă (descompunerea ANOVA pentru regresie) că

$$\|Y - \bar{y}\mathbf{1}\|^2 = \left\|\hat{Y} - \bar{y}\mathbf{1}\right\|^2 + \left\|\underbrace{\hat{\varepsilon}}_{Y-\hat{Y}}\right\|^2$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(\underbrace{\hat{\varepsilon}_i}_{y_i-\hat{y}_i})^2$$

$$SS_T = SS_{reg} + RSS$$

Din definiția coeficientului de determinare R^2 avem că

$$R^2 = \frac{SS_{reg}}{SS_T} = \frac{\left\|\hat{Y} - \bar{y}\mathbf{1}\right\|^2}{\|Y - \bar{y}\mathbf{1}\|^2} = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{y}\mathbf{1}\|^2}$$

și conform figurii de mai sus $R^2 = \cos^2(\theta)$. Prin urmare dacă $R^2 = 1$, atunci $\theta = 0$ și $Y \in \mathcal{M}(X)$, deci $y_i = \beta_0 + \beta_1 x_i$, $i \in \{1, 2, \dots, n\}$ (punctele eșantionului sunt perfect aliniate) iar dacă $R^2 = 0$, deducem că $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 0$, deci $\hat{y}_i = \bar{y}$ (modelul liniar nu este adaptat în acest caz, nu putem explica mai bine decât media).

Ex. 1.7

Arătați că

$$R^2 = r_{xy}^2 = r_{y\hat{y}}^2$$

unde r_{xy} este coeficientul de corelație empiric dintre x și y.

Din definiția coeficientului de determinare și folosind coeficienții $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ și $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2}$ obținuți prin metoda celor mai mici pătrate avem

$$\begin{aligned}
R^2 &= \frac{\left\|\hat{Y} - \bar{y}\mathbf{1}\right\|^2}{\|Y - \bar{y}\mathbf{1}\|^2} = \frac{\sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1\bar{x} - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}\\
&= \frac{\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}\\
&= \frac{\left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} = r_{xy}^2.
\end{aligned}$$

Pentru a verifica a doua parte, $R^2 = r_{y\hat{y}}^2$, să observăm că

$$r_{y\hat{y}}^2 = \frac{\left[\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})\right]^2}{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^n (y_i - \bar{y})^2}$$

iar $\bar{\hat{y}} = \frac{\sum_{i=1}^n \hat{y}_i}{n} = \hat{\beta}_0 + \hat{\beta}_1\bar{x} = \bar{y}$, prin urmare

$$r_{y\hat{y}}^2 = \frac{\left[\sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \bar{y})\right]^2}{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \sum_{i=1}^n (y_i - \bar{y})^2}.$$

De asemenea

$$\begin{aligned}
\sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \bar{y}) &= \sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i + \hat{y}_i - \bar{y})\\
&= \sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2
\end{aligned}$$

și cum

$$\begin{aligned}
\sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) &= \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\\
&= \sum_{i=1}^n (\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i - \bar{y})[(y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})]\\
&= \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2\\
&= \underbrace{\frac{S_{xy}}{S_{xx}}}_{\hat{\beta}_1} S_{xy} - \frac{S_{xy}^2}{S_{xx}^2} S_{xx} = 0
\end{aligned}$$

deducem că $r_{y\hat{y}}^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = R^2$.

1.3 Aplicații numerice

Ex. 1.8

Tabelul de mai prezintă o serie de date privind greutatea taților și respectiv a fiului lor cel mare

$Tata:$	65	63	67	64	68	62	70	66	68	67	69	71
$Fiu:$	68	66	68	65	69	66	68	65	71	67	68	70

Obținem următoarele rezultate numerice

$$\sum_{i=1}^{12} t_i = 800 \quad \sum_{i=1}^{12} t_i^2 = 53418 \quad \sum_{i=1}^{12} t_i f_i = 54107 \quad \sum_{i=1}^{12} f_i = 811 \quad \sum_{i=1}^{12} f_i^2 = 54849.$$

1. Determinați dreapta obținută prin metoda celor mai mici pătrate a greutății fiilor în funcie de greutatea taților.
2. Determinați dreapta obținută prin metoda celor mai mici pătrate a greutății taților în funcie de greutatea fiilor.
3. Arătați că produsul pantelor celor două drepte este egal cu pătratul coeficientului de corelație empirică dintre t_i și f_i (sau coeficientul de determinare).

1. Dreapta de regresie a greutății fiilor în funcție de greutatea taților este $f = \hat{\alpha}_0 + \hat{\alpha}_1 t$ unde coeficienții sunt dați de

$$\hat{\alpha}_0 = \bar{f} - \hat{\alpha}_1 \bar{t}, \quad \hat{\alpha}_1 = \frac{\sum_{i=1}^{12}(t_i - \bar{t})(f_i - \bar{f})}{\sum_{i=1}^{12}(t_i - \bar{t})^2}$$

Pentru datele din problema noastră coeficienții sunt $\hat{\alpha}_0 = 35.8$ și $\hat{\alpha}_1 = 0.48$ iar dreapta de regresie este $f = 35.8 + 0.48t$ (a se vedea figura din stânga).

2. Dreapta de regresie a greutății taților în funcție de greutatea fiilor este $t = \hat{\beta}_0 + \hat{\beta}_1 f$ unde coeficienții sunt dați de

$$\hat{\beta}_0 = \bar{t} - \hat{\beta}_1 \bar{f}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{12}(f_i - \bar{f})(t_i - \bar{t})}{\sum_{i=1}^{12}(f_i - \bar{f})^2}$$

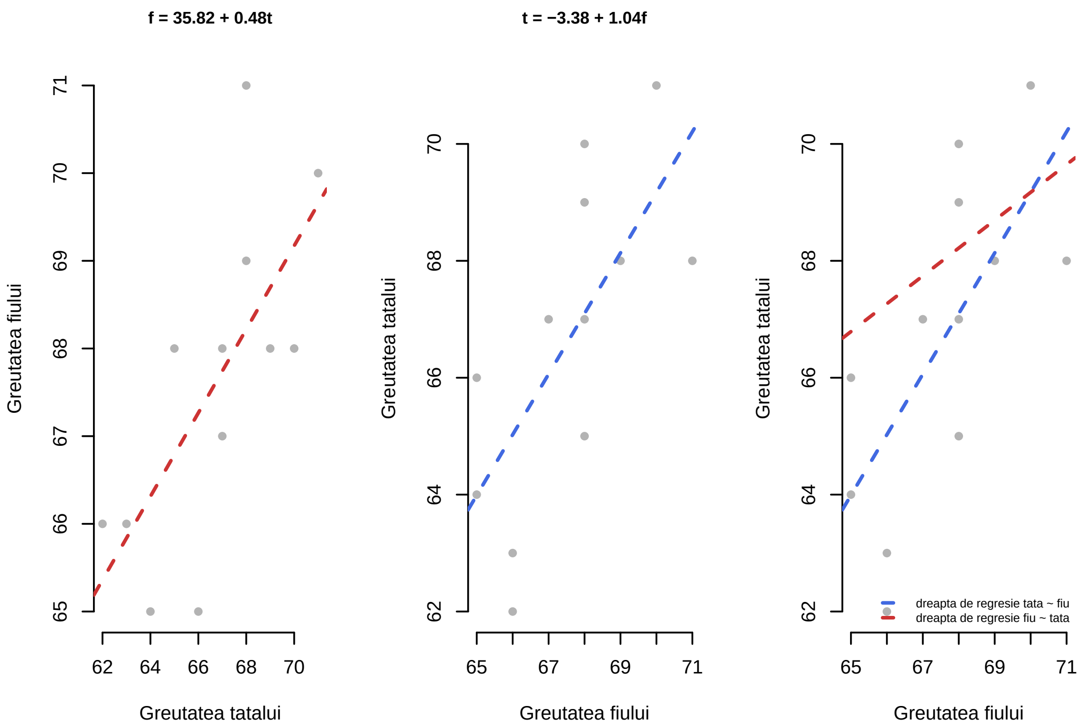
În cazul problemei, coeficienții sunt $\hat{\beta}_0 = -3.38$ și $\hat{\beta}_1 = 1.03$ iar dreapta de regresie este $t = -3.38 + 1.03f$ (a se vedea figura din mijloc).

3. Produsul pantelor celor două drepte este

$$\hat{\alpha}_1 \hat{\beta}_1 = \frac{\left[\sum_{i=1}^{12}(f_i - \bar{f})(t_i - \bar{t})\right]^2}{\sum_{i=1}^{12}(f_i - \bar{f})^2 \sum_{i=1}^{12}(t_i - \bar{t})^2}$$

și conform exercițiului 2 și a definiției coeficientului de determinare avem

$$\hat{\alpha}_1 \hat{\beta}_1 = r_{f,t}^2 = R^2.$$



Ex. ??: Dorim să exprimăm înălțimea y (măsurată în picioare) a unui arbore în funcție de diametrul său x (exprimat în centimetri) la înălțimea de 1m30 de la sol. Pentru aceasta dispunem de 20 de măsurători (x_i, y_i) =(diametru, înălțime) și în urma calculelor am obținut rezultatele următoare: $\bar{x} = 4.53$, $\bar{y} = 8.65$ și

$$\frac{1}{20}\sum_{i=1}^{20}(x_i-\bar{x})^2 = 10.97 \quad \frac{1}{20}\sum_{i=1}^{20}(y_i-\bar{y})^2 = 2.24 \quad \frac{1}{20}\sum_{i=1}^{20}(x_i-\bar{x})(y_i-\bar{y}) = 3.77.$$

.enumienumi.Notăm cu $y = \hat{\beta}_0 + \hat{\beta}_1 x$ dreapta de regresie. Calculați coeficienții $\hat{\beta}_0$ și $\hat{\beta}_1$. Dați și calculați o măsură care descrie calitatea concordanței datelor cu modelul propus. Să presupunem că abaterile standard pentru estimatorii $\hat{\beta}_0$ și $\hat{\beta}_1$ sunt $\hat{\sigma}_0 = 1.62$ și respectiv $\hat{\sigma}_1 = 0.05$. Presupunem că erorile ε_i sunt variabile aleatoare independente repartizare normal de medie 0 și variențe egale. Vrem să testăm ipotezele $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ pentru $j = 0, 1$. De ce acest test este interesant în contextul problemei noastre?

1. Estimatorii coeficienților dreptei de regresie $y = \hat{\beta}_0 + \hat{\beta}_1 x$ sunt dați de

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \approx 0.344$$

și respectiv

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \approx 7.09$$

2. Pentru a măsura calitatea concordanței datelor la modelul de regresie vom folosi coeficientul de determinare R^2. Am văzut că acesta corespunde pătratului coeficientului de corelație empirică:

$$R^2 = r_{x,y}^2 = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}\right)^2 \approx 0.58.$$

Observăm că modelul de regresie liniară simplă explică un pic mai mult de jumătate din variabilitatea datelor.

3. Sub ipoteza modelului condiționat normal (erorile ε_i sunt variabile aleatoare independente repartizare normal de medie 0 și varianțe egale) avem că $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2_{\hat{\beta}_j})$ și înlocuind variantele $\sigma^2_{\hat{\beta}_j}$ cu estimatorii $\hat{\sigma}_j^2$, deducem că $\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim t_{n-2}$.

Prin urmare, sub H_0 avem că

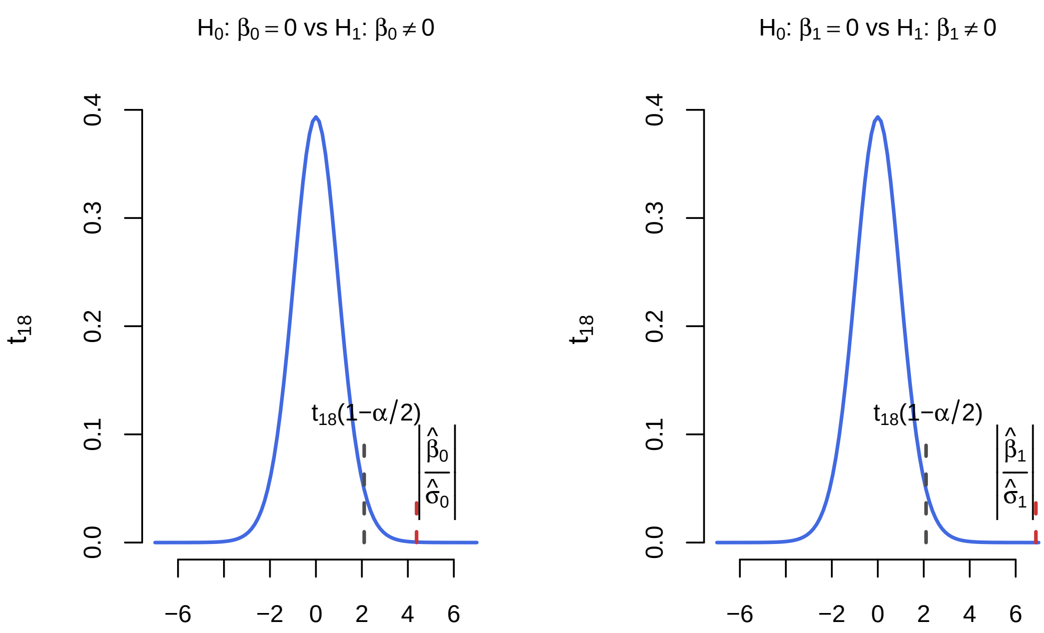
$$\frac{\hat{\beta}_0}{\hat{\sigma}_0} \sim t_{18},$$

iar pentru un prag de semnificație $1 - \alpha = 95\%$, ținând seama că $\left|\frac{\hat{\beta}_0}{\hat{\sigma}_0}\right| \approx 4.38$ și că $t_{18}(1 - \alpha/2) \approx 2.1$, concluzionăm că respingem ipoteza nulă.

În mod similar, pentru $\hat{\beta}_1$ găsim că

$$\left|\frac{\hat{\beta}_1}{\hat{\sigma}_1}\right| \approx 6.88 > 2.1$$

de unde respingem ipoteza nulă $H_0 : \beta_1 = 0$ în acest caz de asemenea.



2 Regresie liniară multiplă

Principiul problemei de regresie este de a modela o variabilă y, numită variabilă răspuns sau variabilă dependentă (cea pe care vrem să o explicăm), cu ajutorul unei funcții care depinde de un anumit număr de variabile $\boldsymbol{x} = (x_1, \dots, x_p)^\intercal$, numite variabile explicative sau predictori sau independente (vom folosi în general primele două denumiri)

$$y \approx g(\boldsymbol{x}) = g(x_1, \dots, x_p).$$

Astfel având dat un eșantion de talie n de $(p+1)$-upluri $(\boldsymbol{x}, y)$, $n > p$, ne propunem să determinăm g. Aproximarea, $\approx$, din relația anterioară se poate descrie matematic sub forma

$$g = \arg\min_{f \in \mathcal{G}} \sum_{i=1}^{n} L\left(y_i - f(\boldsymbol{x}_i)\right)$$

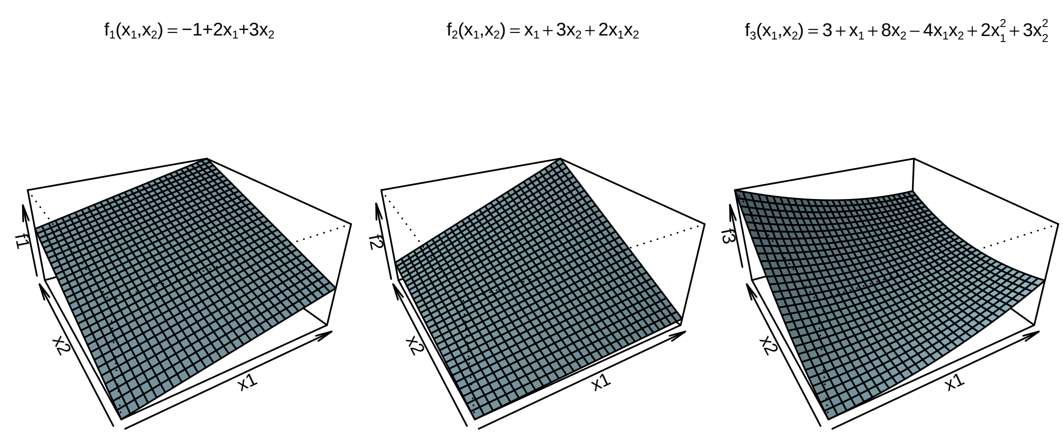
unde $L(\cdot)$ se numește funcție de cost sau de pierdere iar $\mathcal{G}$ este o clasă de funcții dată.

În general funcția de cost poate lua multe forme, e.g. Loss function, dar de cele mai multe ori vom întâlni două: funcția de cost absolut $L(x) = |x|$ sau funcția de cost pătratic $L(x) = x^2$.

În ceea ce privește clasa de funcții $\mathcal{G}$ vom considera funcțiile liniare

$$\mathcal{G} = \left\{ f : \mathbb{R}^p \to \mathbb{R} \,|\, f(\boldsymbol{x}) = \begin{pmatrix} 1 & \boldsymbol{x} \end{pmatrix} \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \right\}.$$

Atunci când vorbim de regresie liniară ne referim la liniaritatea în parametrii β_j și nu în variabilele explicative, e.g. $f_1(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, $f_2(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 x_2}_{x_3}$ sau $f_3(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 x_2}_{x_3} + \beta_4 \underbrace{x_1^2}_{x_4} + \beta_5 \underbrace{x_2^3}_{x_5}$ sunt toate funcții liniare în $\boldsymbol{\beta}$.



2.1 Modelare

Modelul de regresie liniară multiplă reprezintă o extensie a modelului de regresie liniară simplă atunci când numărul variabilelor explicative este finit (pentru mai multe detalii privind analiza de regresie se pot consulta monografiile [?], [?] sau [?]). Presupunem că datele verifică modelul

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

unde

- x_{ij} sunt valori cunoscute și nu sunt aleatoare
- parametrii β_j sunt necunoscuți și nu sunt aleatori
- ε_i sunt variabile aleatoare necunoscute

Scris sub formă compactă modelul devine

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

unde

- **X** se numește *matricea de design*

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}_{n\times(p+1)}$$

- **Y** este *vectorul răspuns* și este un vector aleator, $\boldsymbol{\beta}$ este *vectorul parametrilor* sau coeficienților și este necunoscut iar $\boldsymbol{\varepsilon}$ este *vectorul erorilor* și este un vector aleator

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}_{n\times 1}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1} \quad \text{și} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}.$$

Ipotezele modelului de regresie liniară multiplă sunt:

$$\begin{aligned} &\mathcal{H}_1 : rang(\boldsymbol{X}) = p + 1 \\ &\mathcal{H}_2 : \mathbb{E}[\boldsymbol{\varepsilon}] = 0,\, Var(\boldsymbol{\varepsilon}) = \sigma^2 I_n \end{aligned}$$

Prima ipoteză ne spune că matricea de design $\boldsymbol{X}$ are coloanele liniar independente iar a doua ipoteză se referă la centralitatea erorilor (medie nulă), homoscedasticitatea (aceeași varianță) și necorelarea acestora.

2.2 Metoda celor mai mici pătrate

În această secțiune vom prezenta estimatorul lui $\boldsymbol{\beta}$ obținut prin metoda celor mai mici pătrate, metodă care folosește ca funcție de cost L funcția $L(x) = x^2$. Estimatorul $\hat{\boldsymbol{\beta}}$ obținut prin metoda celor mai mici pătrate este definit prin

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]^2 = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2 = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

unde am folosit convenția $x_{i0} = 1$, $i = 1, 2, \dots, n$.

Arătați că dacă ipoteza $\mathcal{H}_1$ este adevărată atunci estimatorul $\hat{\boldsymbol{\beta}}$ obținut prin metoda celor mai mici pătrate este

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

Ne propunem să prezentăm mai multe metode de calcul pentru estimatorul $\hat{\boldsymbol{\beta}}$.

a) Metoda geometrică

Din punct de vedere geometric, ne plasăm în spațiul variabilelor $\mathbb{R}^n$ (am presupus că $n > p$)). Vectorul variabilelor răspuns $\boldsymbol{Y} \in \mathbb{R}^n$ iar matricea de design $\boldsymbol{X}$ poate fi văzută ca fiind formată din $p+1$ vectori coloană, $\boldsymbol{X} = \begin{bmatrix} \underbrace{\boldsymbol{X}_0}_{\mathbf{1}=(1,\cdots,1)} & |\boldsymbol{X}_1| \cdots |\boldsymbol{X}_p \end{bmatrix}$.

Începem prin a reaminti câteva noțiuni de algebră liniară, pentru mai multe detalii se poate consulta monografia [**?**]. Fiind dată matricea (de design) $\boldsymbol{X} \in \mathcal{M}_{n,p+1}(\mathbb{R})$ putem defini nucleul matricei

$$\ker(\boldsymbol{X}) = \left\{\boldsymbol{u} \in \mathbb{R}^{p+1} \,|\, \boldsymbol{X}\boldsymbol{u} = 0\right\}$$

ca fiind subspațiul lui $\mathbb{R}^{p+1}$ care conține vectorii ortogonali pe liniile din matricea $\boldsymbol{X}$. De asemenea, imaginea matricei $\boldsymbol{X}$ este subspațiul vectorilor din $\mathbb{R}^n$ care se pot scrie ca o combinație liniară de coloanele matricei $\boldsymbol{X}$ și este definit prin

$$\mathrm{Im}(\boldsymbol{X}) = \left\{\boldsymbol{v} \in \mathbb{R}^n \,|\, \exists \boldsymbol{u} \in \mathbb{R}^{p+1} \text{ a.î. } \boldsymbol{X}\boldsymbol{u} = \boldsymbol{v}\right\}.$$

Se poate arăta (a se vedea [**?**, Capitolul 1]) că $\dim \ker(\boldsymbol{X}) = p + 1 - \mathrm{rang}(\boldsymbol{X})$ și că $\dim \mathrm{Im}(\boldsymbol{X}) = \mathrm{rang}(\boldsymbol{X})$ ceea ce conduce la $\dim \ker(\boldsymbol{X}) + \dim \mathrm{Im}(\boldsymbol{X}) = p + 1$ (caz particular al *teoremei lui Grassman*).

Reamintim că doi vectori $\boldsymbol{u}$ și $\boldsymbol{v}$ sunt ortogonali dacă $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$, iar în acest caz scriem $\boldsymbol{u} \perp \boldsymbol{v}$, și că două subspații V și W sunt ortogonale dacă $\forall \boldsymbol{v} \in V$ și respectiv $\forall \boldsymbol{w} \in W$ avem $\boldsymbol{v} \perp \boldsymbol{w}$ și notăm $V \perp W$. De asemenea, spațiul ortogonal al lui V este spațiul $V^\perp$ care conține toți vectorii ortogonali pe V și are ca proprietăți: $V \cap V^\perp = \{0\}$ și $\left(V^\perp\right)^\perp = V$. Se poate arăta cu ușurință că dacă V este un subspațiu a lui $W \subset \mathbb{R}^n$ atunci pentru orice vector $\boldsymbol{w} \in W$ avem că $\boldsymbol{w} = \boldsymbol{v} + \boldsymbol{v}^\perp$ unde $\boldsymbol{v} \in V$ și $\boldsymbol{v}^\perp \in V^\perp$ iar descompunerea se face în mod unic (acest rezultat se mai scrie și sub forma $W = V \oplus V^\perp$). Se numește *proiecție ortogonală*[1] a lui $\boldsymbol{w}$ pe V de-a lungul lui $V^\perp$ endomorfismul $p_V : W \to W$ definit prin $p_V(\boldsymbol{w}) = \boldsymbol{v}$. Se poate arăta cu ușurință că dacă V este un subspațiu vectorial al lui W atunci au loc următoarele proprietăți:

i) pentru orice $\boldsymbol{w} \in W$ avem: $\boldsymbol{w} = p_V(\boldsymbol{w}) + p_{V^\perp}(\boldsymbol{w})$, $\|p_V(\boldsymbol{w})\| \leq \|\boldsymbol{w}\|$ iar $\boldsymbol{w} - p_V(\boldsymbol{w}) \perp \boldsymbol{v}$, $\forall \boldsymbol{v} \in V$

ii) pentru orice $\boldsymbol{w} \in W$ are loc $\|\boldsymbol{w} - p_V(\boldsymbol{w})\| = \inf_{\boldsymbol{v} \in V} \|\boldsymbol{w} - \boldsymbol{v}\|$

iii) dacă p este o proiecție ortogonală atunci $p \circ p = p$

iv) dacă p este un endomorfism care verifică $p \circ p = p$ și în plus $\mathrm{Im}(p) \perp \ker(p)$ atunci p este proiecția ortogonală pe $\mathrm{Im}(p)$ de-a lungul lui $\ker(p)$

v) dacă $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_k\}$ este o bază ortonormală în V atunci $p_V(\boldsymbol{w}) = \sum_{i=1}^k \langle \boldsymbol{w}, \boldsymbol{e}_i \rangle \boldsymbol{e}_i$

[1] În general, numim *proiecție* un endomorfism $p : W \to W$ cu proprietatea că există o descompunere în sumă directă $W = V_1 \oplus V_2$ (și spunem proiecție pe V_1 de-a lungul lui V_2) astfel încât $p(\boldsymbol{w}) = \boldsymbol{v}_1$, $\forall \boldsymbol{w} \in W$ cu $\boldsymbol{w} = \boldsymbol{v}_1 + \boldsymbol{v}_2$, $\boldsymbol{v}_1 \in V_1$ și $\boldsymbol{v}_2 \in V_2$.

Spunem că o matrice pătratică $P \in \mathcal{M}_n(\mathbb{R})$ este o matrice de proiecție dacă este idempotentă $P^2 = P$ (numele vine de la faptul că pentru $\boldsymbol{x} \in \mathbb{R}^n$ aplicația liniară $P\boldsymbol{x}$ este proiecția lui $\boldsymbol{x}$ pe $\text{Im}(P)$ de-a lungul lui $\ker(P)$ - a se vedea [**?**, Capitolul 1, Secțiunea 5.2] sau [**?**, Capitolul 2]). Dacă în plus matricea P este simetrică, i.e. $P = P^\intercal$, atunci $P\boldsymbol{x}$ este proiecția ortogonală a lui $\boldsymbol{x}$ pe $V = \text{Im}(P)$ de-a lungul lui $V^\perp = \ker(P)$, cu alte cuvinte în descompunerea $\boldsymbol{x} = P\boldsymbol{x} + (I - P)\boldsymbol{x}$ vectorii $P\boldsymbol{x}$ și respectiv $(I - P)\boldsymbol{x}$ sunt ortogonali [**?**, Capitolul 2, Secțiunea 2.2]. Prin urmare matricea P este o matrice de proiecție ortogonală dacă are loc relația $\boldsymbol{v} \perp \boldsymbol{v} - P\boldsymbol{v}$ adică $\langle \boldsymbol{v}, \boldsymbol{v} - P\boldsymbol{v}\rangle = 0$ pentru toți $\boldsymbol{v}$.

Dacă $\boldsymbol{X} \in \mathcal{M}_{m,k}(\mathbb{R})$ este o matrice de $\text{rang}(\boldsymbol{X}) = k$ $(m \geq k)$, deci $\boldsymbol{X}^\intercal\boldsymbol{X}$ este inversabilă, atunci pentru a determina matricea de proiecție ortogonală P_X pe subspațiul imagine $\text{Im}(\boldsymbol{X})$ să observăm că dacă $\boldsymbol{v} \in \text{Im}(\boldsymbol{X})$ atunci $\boldsymbol{v} = \boldsymbol{X}\boldsymbol{\alpha}$ și cum $P_X\boldsymbol{v} = \boldsymbol{v}$ deducem că $P_X\boldsymbol{v} = \boldsymbol{X}\underbrace{\left(\boldsymbol{X}^\intercal\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\intercal\boldsymbol{v}}_{=\boldsymbol{\alpha}}$. În plus dacă $\boldsymbol{v}^\perp \in \text{Im}(\boldsymbol{X})^\perp = \ker(\boldsymbol{X}^\intercal)$ atunci $\boldsymbol{X}^\intercal\boldsymbol{v}^\perp = 0$ prin urmare $\boldsymbol{X}\left(\boldsymbol{X}^\intercal\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\intercal\boldsymbol{v}^\perp = 0$ și astfel, pentru $\boldsymbol{w} = \boldsymbol{v} + \boldsymbol{v}^\perp$ arbitrar, avem $P_X\boldsymbol{w} = \boldsymbol{X}\left(\boldsymbol{X}^\intercal\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\intercal\boldsymbol{w}$ de unde găsim că

$$P_X = \boldsymbol{X}\left(\boldsymbol{X}^\intercal\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\intercal.$$

În mod similar se arată că matricea de proiecție ortogonală pe $\ker(\boldsymbol{X})$ este $P_{X\perp} = I - P_X$. Nu este dificil de văzut că cele două matrice, P_X și respectiv $P_{X\perp}$, sunt idempotente. De asemenea, ținând seama că valorile proprii ale unei matrice idempotente sunt 0 sau 1 concluzionăm că $\text{rang}(P_X) = \text{Tr}(P_X)$.

În cazul particular în care $\boldsymbol{X} = \boldsymbol{v}$ avem

$$P_v = \boldsymbol{v}\left(\boldsymbol{v}^\intercal\boldsymbol{v}\right)^{-1}\boldsymbol{v}^\intercal = \frac{\boldsymbol{v}\boldsymbol{v}^\intercal}{\|\boldsymbol{v}\|^2}$$

prin urmare proiecția vectorului $\boldsymbol{u}$ pe $\boldsymbol{v}$ este $P_v\boldsymbol{u} = \frac{\boldsymbol{v}\boldsymbol{v}^\intercal}{\|\boldsymbol{v}\|^2}\boldsymbol{u} = \frac{\langle\boldsymbol{u},\boldsymbol{v}\rangle}{\|\boldsymbol{v}\|^2}\boldsymbol{v}$.

Pentru problema noastră, notăm cu $\mathcal{M}(X) = \text{Im}(\boldsymbol{X})$ subspațiul imagine și conform ipotezei $\mathcal{H}_1$ avem că $\text{rang}(\boldsymbol{X}) = p + 1$, deci $\dim\mathcal{M}(X) = p + 1$. Din definiția spațiului imagine avem că toți vectorii din $\mathcal{M}(X)$ sunt de forma $\boldsymbol{X}\boldsymbol{\alpha}$, cu $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_p) \in \mathbb{R}^{p+1}$

$$\boldsymbol{X}\boldsymbol{\alpha} = \sum_{i=0}^{p}\alpha_i\boldsymbol{X}_i.$$

Conform modelului de regresie, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, vectorul răspuns $\boldsymbol{Y}$ este suma dintre un element din $\mathcal{M}(X)$ și un element din $\mathbb{R}^n$ care nu are niciun motiv să aparțină tot lui $\mathcal{M}(X)$. Astfel, problema minimizării funcției $S(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ revine la a găsi acel vector din $\mathcal{M}(X)$ care este cel mai aproape de $\boldsymbol{Y}$ în sensul distanței euclidiene, cu alte cuvinte de a determina vectorul proiecției ortogonale a lui $\boldsymbol{Y}$ pe $\mathcal{M}(X)$ (a se vedea figura de mai jos).

Proiecția ortogonală a lui $\boldsymbol{Y}$ pe $\mathcal{M}(X)$ se notează cu $\hat{\boldsymbol{Y}} = P_X\boldsymbol{Y}$, unde P_X este matricea de proiecție ortogonală pe $\mathcal{M}(X)$, iar spațiul ortogonal $\mathcal{M}(X)^\perp$ se mai numește și spațiul reziduurilor și are dimensiunea $\dim\mathcal{M}(X)^\perp = n - (p + 1)$ (*teorema lui Grassman*). Să remarcăm că $\hat{\boldsymbol{Y}} \in \mathcal{M}(X)$ deci putem scrie $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ unde $\hat{\boldsymbol{\beta}}$ reprezintă estimatorul obținut prin metoda celor mai mici pătrate iar elementele lui $\hat{\boldsymbol{\beta}}$ sunt coordonatele vectorului $\hat{\boldsymbol{Y}}$ în baza $\{\boldsymbol{X}_0, \boldsymbol{X}_1, \dots, \boldsymbol{X}_p\}$ a spațiului $\mathcal{M}(X)$. Cum reperul $\{\boldsymbol{X}_0, \boldsymbol{X}_1, \dots, \boldsymbol{X}_p\}$ nu este neapărat ortogonal, elementele $\hat{\beta}_j$ nu sunt neapărat coordonatele proiecției lui $\boldsymbol{Y}$ pe $\boldsymbol{X}_j$, aceasta din urmă fiind dată de

$$
\begin{aligned}
P_{X_j}\boldsymbol{Y} &= P_{X_j}P_X\boldsymbol{Y} = P_{X_j}\sum_{i=0}^{p}\hat{\beta}_i\boldsymbol{X}_i \\
&= \sum_{i=0}^{p}\hat{\beta}_i P_{X_j}\boldsymbol{X}_i = \hat{\beta}_j\boldsymbol{X}_j + \sum_{i\neq j}\hat{\beta}_i P_{X_j}\boldsymbol{X}_i
\end{aligned}
$$

În cazul în care $\boldsymbol{X}_j$ este ortogonal pe $\boldsymbol{X}_i$, $i \neq j$, atunci $P_{X_j}\boldsymbol{Y} = \hat{\beta}_j\boldsymbol{X}_j$ iar dacă $\langle \boldsymbol{X}_i, \boldsymbol{X}_j\rangle = 0$ pentru toți $i \neq j$ atunci matricea $\boldsymbol{X}^\intercal\boldsymbol{X} = \mathrm{diag}\left(\|\boldsymbol{X}_0\|^2, \|\boldsymbol{X}_1\|^2, \dots, \|\boldsymbol{X}_p\|^2\right)$.

O altă metodă (tot prin proiecții) de a arăta că $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{Y}$ se bazează pe observația că proiecția ortogonală $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ este definită ca fiind unicul vector pentru care $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$ este ortogonal pe $\mathcal{M}(X)$. Cum $\mathcal{M}(X)$ este generat de $\{\boldsymbol{X}_0, \boldsymbol{X}_1, \dots, \boldsymbol{X}_p\}$ putem spune că $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$ este ortogonal pe fiecare $\boldsymbol{X}_i$:

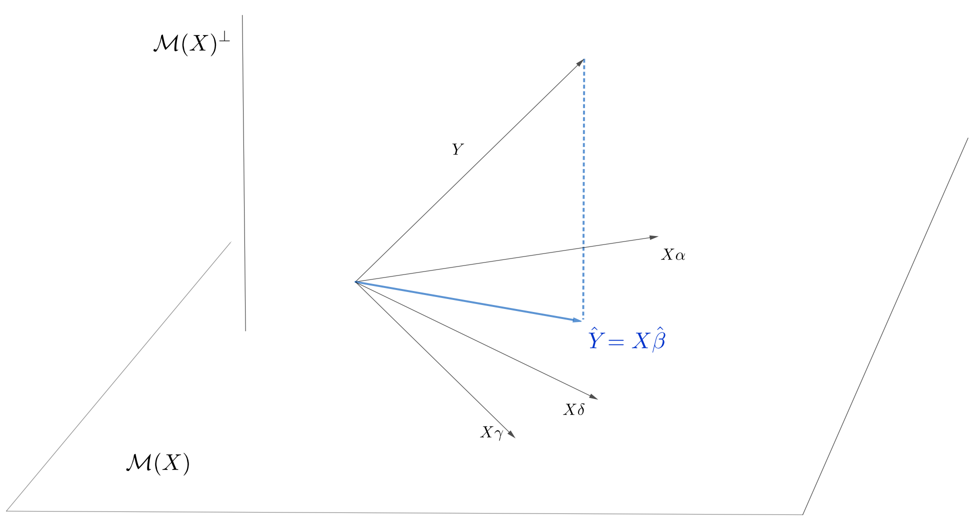
$$
\begin{cases}
\langle \boldsymbol{X}_0, \boldsymbol{Y} - \hat{\boldsymbol{Y}}\rangle = 0 \\
\langle \boldsymbol{X}_1, \boldsymbol{Y} - \hat{\boldsymbol{Y}}\rangle = 0 \\
\vdots \\
\langle \boldsymbol{X}_p, \boldsymbol{Y} - \hat{\boldsymbol{Y}}\rangle = 0
\end{cases}
\iff
\begin{cases}
\langle \boldsymbol{X}_0, \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\rangle = 0 \\
\langle \boldsymbol{X}_1, \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\rangle = 0 \\
\vdots \\
\langle \boldsymbol{X}_p, \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\rangle = 0
\end{cases}
\iff \boldsymbol{X}^\intercal\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = 0
$$

de unde găsim *sistemul de ecuații normale*

$$
\boldsymbol{X}^\intercal\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\intercal\boldsymbol{Y}
$$

care, atunci când ipoteza $\mathcal{H}_1$ este adevărată - matricea $\boldsymbol{X}^\intercal\boldsymbol{X}$ este inversabilă, revine la

$$
\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{Y}.
$$



Notând cu $P_{X^\perp} = I_n - P_X$ matricea de proiecție ortogonală pe $\mathcal{M}^\perp(X)$, unde $P_X = \boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal$ este matricea de proiecție ortogonală pe $\mathcal{M}(X)$, observăm că descompunerea

$$\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right) = P_X\boldsymbol{Y} + (I - P_X)\boldsymbol{Y} = P_X\boldsymbol{Y} + P_{X^\perp}\boldsymbol{Y}$$

nu este altceva decât descompunerea ortogonală a lui $\boldsymbol{Y}$ pe $\mathcal{M}(X)$ și respectiv $\mathcal{M}^\perp(X)$. De asemenea este de remarcat faptul că în ceea ce privește notația folosită în literatura de statistică de specialitate, matricea de proiecție ortogonală P_X se mai notează și cu H (care vine de la *hat*, $\hat{\boldsymbol{Y}} = H\boldsymbol{Y}$).

b) Metoda analitică

O a doua metodă este metoda analitică. Problema noastră cere să căutăm vectorul $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ care minimizează funcția

$$\begin{aligned} S(\boldsymbol{\beta}) &= \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\intercal (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{Y}^\intercal\boldsymbol{Y} - \boldsymbol{Y}^\intercal\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\intercal\boldsymbol{X}^\intercal\boldsymbol{Y} + \boldsymbol{\beta}^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{Y}^\intercal\boldsymbol{X}\boldsymbol{\beta} + \|\boldsymbol{Y}\|^2 \end{aligned}$$

unde am folosit faptul că $\boldsymbol{Y}^\intercal\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{Y}^\intercal\boldsymbol{X}\boldsymbol{\beta})^\intercal = \boldsymbol{\beta}^\intercal\boldsymbol{X}^\intercal\boldsymbol{Y}$ (sunt elemente de dimensiune 1×1). Pentru aceasta trebuie să rezolvăm ecuația $\nabla S(\boldsymbol{\beta}) = 0$ și să verificăm că soluția este într-adevăr punct de minim.

Reamintim că dacă $f : \mathbb{R}^k \to \mathbb{R}$ atunci

$$\nabla f = \frac{\partial f}{\partial \boldsymbol{x}^\intercal} = \begin{pmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_k} \end{pmatrix}.$$

În particular, pentru $f(\boldsymbol{x}) = \boldsymbol{a}^\intercal\boldsymbol{x}$ (formă liniară) avem

$$\nabla f = \frac{\partial \boldsymbol{a}^\intercal\boldsymbol{x}}{\partial \boldsymbol{x}^\intercal} = \begin{pmatrix} \frac{\partial \sum_{i=1}^k a_i x_i}{\partial x_1} & \frac{\partial \sum_{i=1}^k a_i x_i}{\partial x_2} & \dots & \frac{\partial \sum_{i=1}^k a_i x_i}{\partial x_k} \end{pmatrix} = \boldsymbol{a}^\intercal$$

iar $\frac{\partial \boldsymbol{a}^\intercal\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}$.

În cazul în care f este o aplicație liniară, $f(\boldsymbol{x}) = A\boldsymbol{x}$ unde $A \in \mathcal{M}_{m,k}(\mathbb{R})$, atunci

$$A\boldsymbol{x} = \begin{pmatrix} \sum_{j=1}^k a_{1j}x_j \\ \sum_{j=1}^k a_{2j}x_j \\ \vdots \\ \sum_{j=1}^k a_{mj}x_j \end{pmatrix}, \qquad \frac{\partial A\boldsymbol{x}}{\partial x_j} = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}$$

și

$$\frac{\partial A\boldsymbol{x}}{\partial \boldsymbol{x}^\intercal} = \left(\begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix}, \dots, \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}, \dots, \begin{pmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{mk} \end{pmatrix} \right) = \begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{ik} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mj} & \cdots & a_{mk} \end{pmatrix} = A.$$

În mod similar se poate verifica și relația $\frac{\partial \boldsymbol{x}^\intercal A^\intercal}{\partial \boldsymbol{x}} = A^\intercal$.

Dacă f este o formă pătratică $f(\boldsymbol{x}) = \boldsymbol{x}^\intercal A\boldsymbol{x}$ cu $A \in \mathcal{M}_k(\mathbb{R})$ (matrice pătratică de ordin k), atunci

$$\boldsymbol{x}^\intercal A\boldsymbol{x} = \sum_{i=1}^k \sum_{j=1}^k a_{ij}x_ix_j = \sum_{i=1}^k a_{ii}x_i^2 + \sum_{i=1}^k \sum_{\substack{j=1 \\ j\neq i}}^k a_{ij}x_ix_j$$

de unde

$$\frac{\partial \boldsymbol{x}^\intercal A\boldsymbol{x}}{\partial x_r} = 2a_{rr}x_r + \sum_{j\neq r} a_{rj}x_j + \sum_{i\neq r} a_{ir}x_i = \sum_{j=1}^{k} a_{rj}x_j + \sum_{i=1}^{k} a_{ir}x_i$$

prin urmare

$$\frac{\partial \boldsymbol{x}^\intercal A\boldsymbol{x}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial \boldsymbol{x}^\intercal A\boldsymbol{x}}{\partial x_1} \\ \vdots \\ \frac{\partial \boldsymbol{x}^\intercal A\boldsymbol{x}}{\partial x_k} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{k} a_{1j}x_j + \sum_{i=1}^{k} a_{i1}x_i \\ \vdots \\ \sum_{j=1}^{k} a_{kj}x_j + \sum_{i=1}^{k} a_{ik}x_i \end{pmatrix} = A\boldsymbol{x} + A^\intercal\boldsymbol{x}.$$

De asemenea putem observa că $\frac{\partial \boldsymbol{x}^\intercal A\boldsymbol{x}}{\partial \boldsymbol{x}^\intercal} = (A\boldsymbol{x} + A^\intercal\boldsymbol{x})^\intercal = \boldsymbol{x}^\intercal A^\intercal + \boldsymbol{x}^\intercal A$.

În plus, dacă $A \in \mathcal{M}_k(\mathbb{R})$ este o matrice simetrică ($A^\intercal = A$) atunci

$$\frac{\partial \boldsymbol{x}^\intercal A\boldsymbol{x}}{\partial \boldsymbol{x}} = 2A\boldsymbol{x}, \quad \frac{\partial \boldsymbol{x}^\intercal A\boldsymbol{x}}{\partial \boldsymbol{x}^\intercal} = 2\boldsymbol{x}^\intercal A^\intercal.$$

Revenind la problema noastră observăm că $S(\boldsymbol{\beta})$ este pătratică în $\boldsymbol{\beta}$ iar matricea $\boldsymbol{X}^\intercal\boldsymbol{X}$ este simetrică, prin urmare

$$\nabla S(\boldsymbol{\beta}) = \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\intercal} = \frac{\partial}{\partial \boldsymbol{\beta}^\intercal}\left(\boldsymbol{\beta}^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{Y}^\intercal\boldsymbol{X}\boldsymbol{\beta} + \|\boldsymbol{Y}\|^2\right) = 2\boldsymbol{\beta}^\intercal(\boldsymbol{X}^\intercal\boldsymbol{X}) - 2\boldsymbol{Y}^\intercal\boldsymbol{X} = 0$$

este echivalentă cu *sistem de ecuații normale* (prin trecerea la transpusă)

$$(\boldsymbol{X}^\intercal\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}^\intercal\boldsymbol{Y}$$

care atunci când ipoteza $\mathcal{H}_1$ este adevărată, ceea ce conduce la inversabilitatea matricei $\boldsymbol{X}^\intercal\boldsymbol{X}$ (are valori proprii nenule), revine la

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{Y}.$$

Pentru a arăta că $\hat{\boldsymbol{\beta}}$ este într-adevăr un punct de minim pentru $S(\boldsymbol{\beta})$ trebuie să arătăm că matricea hessiană este pozitiv definită. Matricea hessiană, ținând cont de simetria matricii $\boldsymbol{X}^\intercal\boldsymbol{X}$, este

$$\frac{\partial^2 S(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\intercal} = \frac{\partial}{\partial\boldsymbol{\beta}}\left(\frac{\partial S(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\intercal}\right) = \frac{\partial}{\partial\boldsymbol{\beta}}\left(2\boldsymbol{\beta}^\intercal(\boldsymbol{X}^\intercal\boldsymbol{X}) - 2\boldsymbol{Y}^\intercal\boldsymbol{X}\right) = 2\boldsymbol{X}^\intercal\boldsymbol{X}$$

iar pentru $\boldsymbol{u} \in \mathbb{R}^{p+1}\backslash\{0\}$ avem

$$\boldsymbol{u}^\intercal\frac{\partial^2 S(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\intercal}\boldsymbol{u} = \boldsymbol{u}^\intercal\left(2\boldsymbol{X}^\intercal\boldsymbol{X}\right)\boldsymbol{u} = \langle\boldsymbol{X}\boldsymbol{u}, \boldsymbol{X}\boldsymbol{u}\rangle = 2\|\boldsymbol{X}\boldsymbol{u}\|^2 > 0$$

deci $\boldsymbol{X}^\intercal\boldsymbol{X}$ este pozitiv definită prin urmare și $\frac{\partial^2 S(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\intercal}$ este pozitiv definită.

Ex: ?? Arătați că dacă ipotezele $\mathcal{H}_1$ și $\mathcal{H}_2$ sunt adevărate atunci estimatorul $\hat{\boldsymbol{\beta}}$ obținut prin metoda celor mai mici pătrate este nedeplasat, i.e. $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ iar matricea de varianță-covarianță $Var\left(\hat{\boldsymbol{\beta}}\right)$ este

$$Var\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\boldsymbol{X}^\intercal \boldsymbol{X}\right)^{-1}.$$

Pentru început, deoarece vorbim de operații cu vectori aleatori sau matrice aleatoare, vom reaminti câteva proprietăți de calcul a acestora (acestea generalizează noțiunile corespunzătoare din cazul variabilelor aleatoare). Reamintim că fiind dată matricea $\boldsymbol{Z} = (Z_{ij})_{i,j}$ unde Z_{ij}, $1 \leq i \leq m$, $1 \leq j \leq k$ sunt variabile aleatoare de medie $\mathbb{E}[Z_{ij}]$, media matricei $\mathbb{E}[\boldsymbol{Z}]$ este definită ca matricea mediilor $(\mathbb{E}[Z_{ij}])_{i,j}$. În plus dacă $\boldsymbol{A} \in \mathcal{M}_{l,m}(\mathbb{R})$, $\boldsymbol{B} \in \mathcal{M}_{k,q}(\mathbb{R})$ și $\boldsymbol{C} \in \mathcal{M}_{l,q}(\mathbb{R})$ sunt trei matrice cu coeficienți constanți atunci

$$\mathbb{E}\left[\boldsymbol{A}\boldsymbol{Z}\boldsymbol{B} + \boldsymbol{C}\right] = \boldsymbol{A}\mathbb{E}[\boldsymbol{Z}]\boldsymbol{B} + \boldsymbol{C}.$$

În mod similar, dacă $\boldsymbol{Z}$ și $\boldsymbol{T}$ sunt doi vectori aleatori de dimensiune $m \times 1$ și respectiv $k \times 1$ atunci covarianța acestora este definită prin $Cov(\boldsymbol{Z}, \boldsymbol{T}) = \left(Cov(Z_i, T_j)\right)_{i,j}$ și se poate verifica relația

$$Cov(\boldsymbol{Z}, \boldsymbol{T}) = \mathbb{E}\left[\left(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}]\right)\left(\boldsymbol{T} - \mathbb{E}[\boldsymbol{T}]\right)^\intercal\right] = \mathbb{E}\left[\boldsymbol{Z}\boldsymbol{T}^\intercal\right] - \mathbb{E}\left[\boldsymbol{Z}\right]\mathbb{E}\left[\boldsymbol{T}\right]^\intercal = Cov(\boldsymbol{T}, \boldsymbol{Z})^\intercal.$$

În particular pentru $\boldsymbol{T} = \boldsymbol{Z}$ avem matricea de varianță-covarianță $Var(\boldsymbol{Z}) = Cov(\boldsymbol{Z}, \boldsymbol{Z})$ care, în conformitate cu relația de mai sus, este egală cu

$$Var(\boldsymbol{Z}) = \mathbb{E}\left[\left(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}]\right)\left(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}]\right)^\intercal\right] = \mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^\intercal\right] - \mathbb{E}\left[\boldsymbol{Z}\right]\mathbb{E}\left[\boldsymbol{Z}\right]^\intercal.$$

Un calcul direct arată că dacă $\boldsymbol{A} \in \mathcal{M}_{l,m}(\mathbb{R})$ și $\boldsymbol{B} \in \mathcal{M}_{q,k}(\mathbb{R})$ sunt două matrice cu coeficienți constanți atunci

$$Cov\left(\boldsymbol{A}\boldsymbol{Z}, \boldsymbol{B}\boldsymbol{T}\right) = \boldsymbol{A}Cov\left(\boldsymbol{Z}, \boldsymbol{T}\right)\boldsymbol{B}^\intercal$$

iar atunci când $\boldsymbol{T} = \boldsymbol{Z}$ și $\boldsymbol{A} = \boldsymbol{B}$ găsim că $Var(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}Var(\boldsymbol{Z})\boldsymbol{A}^\intercal$. De asemenea, dacă $\boldsymbol{Z}$, $\boldsymbol{T}$, $\boldsymbol{U}$ și $\boldsymbol{V}$ sunt vectori aleatori de dimensiune $m \times 1$ iar $a, b, c, d \in \mathbb{R}$ sunt constante reale atunci

$$Cov(a\boldsymbol{Z} + b\boldsymbol{T}, c\boldsymbol{U} + d\boldsymbol{V}) = acCov(\boldsymbol{Z}, \boldsymbol{U}) + adCov(\boldsymbol{Z}, \boldsymbol{V}) + bcCov(\boldsymbol{T}, \boldsymbol{U}) + bdCov(\boldsymbol{T}, \boldsymbol{V})$$

și respectiv

$$Var(a\boldsymbol{Z} + b\boldsymbol{T}) = a^2 Var(\boldsymbol{Z}) + ab\left(Cov(\boldsymbol{Z}, \boldsymbol{T}) + Cov(\boldsymbol{T}, \boldsymbol{Z})\right) + b^2 Var(\boldsymbol{T}).$$

Pentru a verifica nedeplasarea estimatorului $\hat{\boldsymbol{\beta}}$ obținut prin metoda celor mai mici pătrate să notăm că, în contextul ipotezei $\mathcal{H}_1$, acesta este $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{Y}$. Astfel putem scrie

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}\left[(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{Y}\right] = (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\mathbb{E}[\boldsymbol{Y}] = (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\mathbb{E}[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}]$$

și cum $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ conform $\mathcal{H}_2$ deducem că

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \underbrace{(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{X}}_{=I_{p+1}}\boldsymbol{\beta} + \underbrace{\mathbb{E}[\boldsymbol{\varepsilon}]}_{=0} = \boldsymbol{\beta}.$$