

Exerciții de seminar 2

Regresie

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

1 Regresie liniară simplă

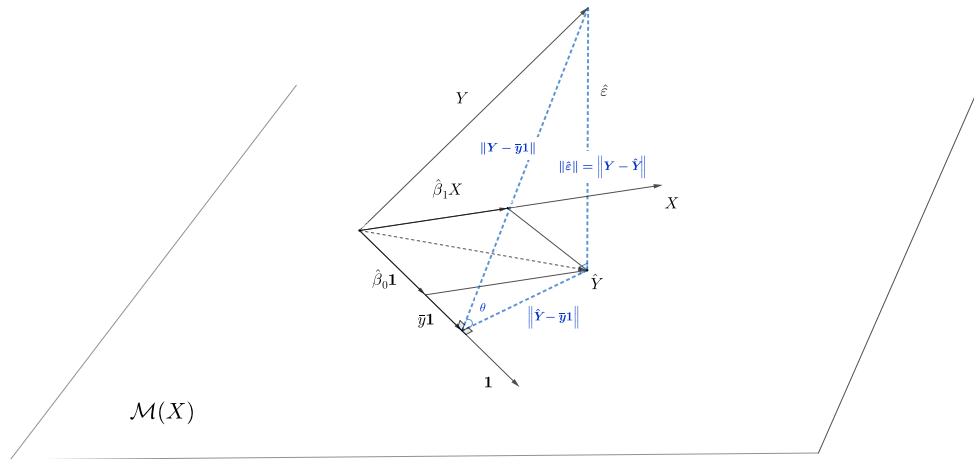
1.1 Interpretări geometrice

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 $X = (x_1, x_2, \dots, x_n)^\top$ a celor n observații ale variabilei explicative și vectorul $Y = (y_1, y_2, \dots, y_n)^\top$ compus din cele n observații ale variabilei răspuns, pe care vrem să o explicăm. Cei doi vectori aparțin spațiului \mathbb{R}^n .

Fie $\mathbf{1} = (1, 1, \dots, 1)^\top \in \mathbb{R}^n$ 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 proiecției ortogonale, \hat{Y} este unicul vector din $\mathcal{M}(X)$ care minimizează distanța euclidiană (deci și pătratul ei)

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

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

$$\begin{aligned} \|Y - \bar{y}\mathbf{1}\|^2 &= \|\hat{Y} - \bar{y}\mathbf{1}\|^2 + \left\| \underbrace{\hat{\varepsilon}}_{Y - \hat{Y}} \right\|^2 \\ \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \underbrace{(\hat{\varepsilon}_i)^2}_{y_i - \hat{y}_i} \\ SS_T &= SS_{reg} + RSS \end{aligned}$$

Din definiția coeficientului de determinare R^2 avem că

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

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 linear nu este adaptat în acest caz, nu putem explica mai bine decât media).

1.2 Exercițiul 1



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


$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\underbrace{\sum_{i=1}^n (x_i - \bar{x})\beta_0}_0 + \beta_1 \sum_{i=1}^n (x_i - \bar{x})x_i + \sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{aligned}$$

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 .

 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 $\hat{\beta}_1$.

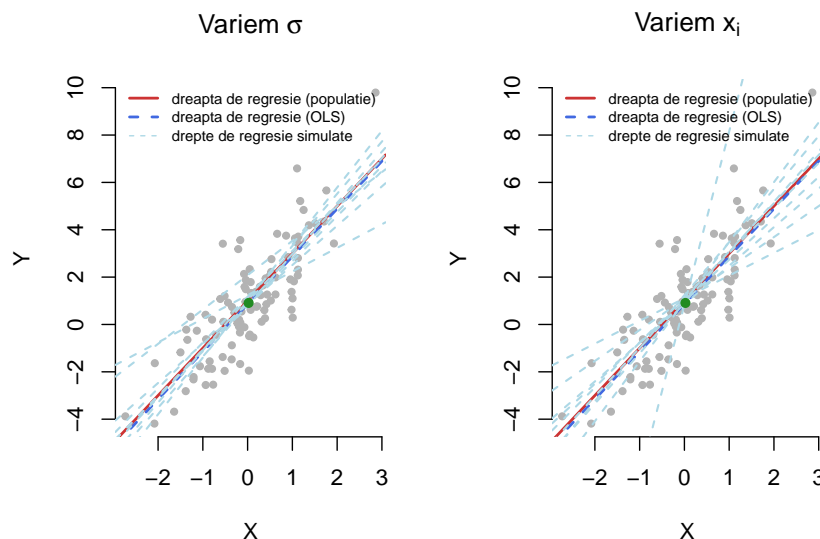
Avem, folosind expresia lui $\hat{\beta}_1$ determinată la punctul anterior și homoscedasticitatea și necorelarea erorilor $Cov(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$, că

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

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 albastre deschise 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)

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

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

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



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

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



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

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

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

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

Luând media găsim că

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

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

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



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

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

1.3 Exercițiul 2: Coeficientul de determinare R^2 și coeficientul de corelație



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

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

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

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

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

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

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