

Regresie liniară simplă. Regresie multiplă

Conf. dr. habil. Eduard Rothenstein

1 Motivație

Regresia este o metodă statistică utilizată pentru descrierea naturii relației între variabile, stabilind modul prin care o variabilă (aceasta va fi numită *variabilă dependentă*, sau *variabila prezisă*) depinde de altă variabilă, sau de alte variabile (aceasta va fi *variabilă independentă*, ce poate fi manipulată și care se mai regăsește în literatura de specialitate și sub numele de variabilă predictor, stimul sau variabilă comandată). *Analiza regresională* cuprinde tehnici de modelare și analiză a relației dintre aceste două categorii de variabile. De asemenea, abordează chestiuni precum predicția valorilor viitoare ale variabilei răspuns pornind de la o variabilă dată sau mai multe variabile stimul. Se poate preciza care dintre variabilele de plecare sunt importante în prezicerea variabilei răspuns.

Există multe variabile X și Y care ar părea să fie legate unde de cealaltă, într-un mod care nu este determinist. Un exemplu poate fi dat de X = media din cei patru ani de liceu și Y = media obținută la facultate. Valoarea lui Y nu poate fi determinată doar din cunoașterea lui X , doi elevi putând obține aceeași valoare pentru X dar valori diferite pentru Y . Cu toate acestea, există o tendință pentru acei studenți care au note liceale mici de a obține note scăzute la facultate. Cunoașterea rezultatelor liceale ar trebui să ne ajute să putem prezice cum va face față acea persoană la facultate. Alte exemple de variabile înrudite într-o manieră nedeterministă includ X = vârsta unui copil și Y = conținutul vocabularului său, X = capacitatea unui motor și Y = eficiența combustibilului pentru un automobil echipat cu acest motor.

Principiul după care se poate obține variabila dependentă în funcție de variabilele independente e asemănător celui întâlnit în Principiul programării dinamice al lui Bellman. Etapa decizională constă într-o stare de intrare (sau mai multe), o stare de ieșire, o decizie și, eventual, o funcție utilitate. În cazul analizei regresionale, datele de intrare sunt informațiile $x_1, x_2, \dots, x_m$, care sunt prelucrate (în timpul prelucrării apar anumiți parametri decizionali, $\beta_1, \beta_2, \dots, \beta_k$), iar rezultatul final de ieșire din sistem este înregistrat într-o singură variabilă răspuns, y . De asemenea, pe parcursul prelucrării datelor, sau după aceasta, pot apărea distorsiuni în sistem, de care putem ține cont dacă introducem un nou parametru, care să cuantifice eroarea ce poate apărea la observarea variabilei y . Se stabilește astfel o legătură între o variabilă dependentă, y , și una sau mai multe variabile independente, $x_1, x_2, \dots, x_m$, care, în cele mai multe cazuri, poate fi scrisă sub forma matematică generală

$$y = f(x_1, x_2, \dots, x_m; \beta_1, \beta_2, \dots, \beta_k) + \varepsilon, \quad (1)$$

unde $\beta_1, \beta_2, \dots, \beta_k$ sunt parametri reali necunoscuți *a priori* (denumiți *parametri de regresie*) și ε este o perturbație aleatoare. De regulă, ε este o eroare de măsură, considerată modelată printr-o variabilă aleatoare normală de medie zero. Funcția f se numește *funcție de regresie*. Dacă aceasta nu este cunoscută *a priori*, atunci poate fi greu de determinat iar utilizatorul analizei regresionale va trebui să o intuiască sau să o aproximeze utilizând metode de tip *trial and error* (prin încercări). Dacă avem doar o variabilă independentă (un singur x), atunci spunem că avem o *regresie simplă*. *Regresia multiplă* face referire la situația în care avem multe variabile independente, ecuația (1) putând fi scrisă sub forma vectorială $y = f(x, \beta) + \varepsilon$.

Pentru a o analiză completă a regresiei (1), va trebui să intuim forma funcției f și apoi să determinăm (aproximăm) valorile parametrilor de regresie. În acest scop, un experimentalist va face un număr suficient de observații (experimente statistice), în urma cărora va aproxima aceste valori. Dacă notăm cu n numărul de experimente efectuate, atunci le putem contabiliza pe acestea în următorul sistem stochastic de ecuații:

$$y_i = f(x, \beta) + \varepsilon_i, \quad i = 1, 2, \dots, n. \quad (2)$$

În ipoteze uzuale, erorile sunt variabile aleatoare identic repartizate $\mathcal{N}(0, \sigma^2)$, independente stochastic două câte două. Astfel, sistemul (2) are necunoscutele $\{\beta_j\}_{j=1,2,\dots,k}$ și σ , deci, în total $k + 1$ necunoscute.

În cazul în care numărul de experimente este mai mic decât numărul parametrilor ce trebuie aproximați ($n \leq k$), atunci nu avem suficiente informații pentru a determina aproximațiile. Dacă $n = k + 1$, atunci problema se reduce la a rezolva n ecuații cu n necunoscute. În cel de-al treilea caz posibil, $n > k + 1$, atunci avem un sistem compatibil, cu valori nedeterminate.

În funcție de forma funcției de regresie f , putem avea:

- *regresie liniară simplă*, în cazul în care avem doar o variabilă independentă și $f(x, \beta) = \beta_0 + \beta_1 x$.
- *regresie liniară multiplă*, dacă $f(x, \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$.

- *regresie liniară multiplă, pentru două variabile, cu interacțiuni* dacă $f(x, \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2$.
- *regresie polinomială*, dacă $f(x, \beta) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_k x^k$. Avem de a face cu *regresie pătratică* pentru $k = 2$, *regresie cubică* pentru $k = 3$.
- *regresie exponențială*, când $f(x, \beta) = \beta_0 e^{\beta_1 x}$.
- *regresie logaritmică*, dacă $f(x, \beta) = \beta_0 \cdot \log_{\beta_1} x$.
- *regresie logistică*, dacă $f(x, \beta) = e^{\beta_0 + \beta_1 x} / (1 + e^{\beta_0 + \beta_1 x})$.

În cadrul analizei regresionale, se cunosc datele de intrare, $\{x_i\}_i$, și căutăm să estimăm parametrii de regresie $\{\beta_j\}_j$ și deviația standard a erorilor, σ . Utilizăm pentru aceasta abordările:

Dacă funcția de regresie f este cunoscută (intuită):

metoda verosimilității maxime

metoda celor mai mici pătrate

metoda lui Bayes

Dacă f nu este cunoscută (intuită):

metoda celor mai mici pătrate

metoda minimax

2 Regresia liniară simplă

Este cel mai simplu tip de regresie, în care avem o singură variabilă independentă, x , și variabila dependentă y . Să presupunem că se cunoaște familia de date bidimensionale $\{(x_i, y_i)\}_{i=1,2,\dots,n}$. Reprezentăm grafic aceste date într-un sistem xOy și observăm o dependență aproape liniară a lui y de x . Dacă valoarea coeficientului de corelație liniară, r , este aproape de 1 sau -1 (indicând o corelație liniară strânsă, pozitivă sau negativă), atunci se pune problema identificării unei relații numerice exacte între x și y de forma

$$y = \beta_0 + \beta_1 x. \quad (3)$$

Această dreaptă se numește dreapta de regresie a lui y în raport cu x . De cele mai multe ori, datele reale nu urmează o relație perfectă de legătură, situație în care parametrii din dependența liniară trebuie a fi estimați. Așadar, va trebui să ținem cont și de eventualele perturbații din sistem. Putem presupune astfel că dependența lui y de x este de forma

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad (4)$$

cu $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Plecând de la observațiile $\{(x_i, y_i)\}_{i=1,2,\dots,n}$, trebuie să găsim o dreaptă ce se apropie cel mai mult de aceste date statistice. Cu alte cuvinte, va trebui să estimăm valorile parametrilor de regresie β_0 și β_1 . Înlocuind datele bidimensionale în (4), avem următorul sistem:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (5)$$

unde $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, pentru fiecare i și sunt independente stochastic.

Deoarece

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i), \quad i = 1, 2, \dots, n,$$

putem interpreta ε_i ca fiind erorile de aproximare a valorilor observate y_i cu cele prezise de dreapta de regresie (adică de valorile $\beta_0 + \beta_1 x_i$). Ținând cont că $\varepsilon_i \sim \mathcal{N}(0, \sigma)$ și β_0, β_1 sunt valori deterministe, din (5) rezultă că (notăm cu Y_i variabila de selecție corespunzătoare valorii empirice y_i):

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \quad \text{pentru fiecare } i,$$

Pentru estimarea parametrilor β_0, β_1 și σ vom aplica metoda verosimilității maxime. Considerăm, pentru aceasta, funcția de verosimilitate:

$$\mathcal{L}(\beta_0, \beta_1, \sigma) = \prod_{i=1}^n f_{Y_i}(y_i) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right).$$

Problema de maximizare ce trebuie rezolvată este următoarea:

$$\max_{\beta_0, \beta_1, \sigma} \mathcal{L}(\beta_0, \beta_1, \sigma).$$

Cum $\mathcal{L}$ și

$$\ln \mathcal{L}(\beta_0, \beta_1, \sigma) = -n \ln \sigma - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

au aceleași puncte de maxim, condițiile pentru obținerea punctelor critice (impuse pentru $\ln \mathcal{L}$) sunt:

$$\begin{cases} \frac{\partial \ln \mathcal{L}}{\partial \beta_0} = \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0, \\ \frac{\partial \ln \mathcal{L}}{\partial \beta_1} = \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0, \\ \frac{\partial \ln \mathcal{L}}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = 0. \end{cases}$$

Prin urmare, primele două ecuații conduc la

$$\begin{cases} -n\beta_0 - \beta_1 \sum_{i=1}^n x_i = -\sum_{i=1}^n y_i \\ -\beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 = -\sum_{i=1}^n x_i y_i \end{cases} \iff \begin{cases} n\beta_0 \sum_{i=1}^n x_i + \beta_1 \left(\sum_{i=1}^n x_i \right)^2 = \sum_{i=1}^n y_i \sum_{i=1}^n x_i \\ n\beta_0 \sum_{i=1}^n x_i + \beta_1 n \sum_{i=1}^n x_i^2 = n \sum_{i=1}^n x_i y_i \end{cases}$$

Scădem ecuațiile și deducem că

$$\begin{aligned} \beta_1 \left(n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right) &= n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \\ \iff \beta_1 \left(n \sum_{i=1}^n x_i^2 - n^2 \bar{x}^2 \right) &= n \sum_{i=1}^n x_i y_i - n^2 \bar{x} \bar{y} \iff \beta_1 \left(\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right) = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \\ \iff \beta_1 s_{xx} &= s_{xy} \end{aligned} \quad (6)$$

Obținem că, în urma rezolvării primelor două ecuații în raport cu β_0 și β_1 , găsim estimațiile:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} \quad \text{și} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad (7)$$

unde

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Pentru a obține ultima implicație din formula (6) utilizăm următoarea reprezentare echivalentă a cantităților s_{xx} și s_{xy} :

$$\begin{cases} s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2 \\ s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \bar{x} \sum_{i=1}^n y_i - \bar{y} \sum_{i=1}^n x_i + n\bar{x}\bar{y} = \sum_{i=1}^n x_i y_i - 2n\bar{x}\bar{y} + n\bar{x}\bar{y} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \end{cases}$$

Remarcăm că $s_{xx} = (n-1)s_x^2$ și $s_{xy} = (n-1)s_{x,y}$ unde s_x este deviația standard empirică modificată a setului de date $\{x_i\}_{i=1}^n$, iar $s_{x,y}$ este covarianța (corelația) empirică modificată a a setului de date perechi $\{(x_i, y_i)\}_{i=1}^n$. Prin urmare, pentru evaluarea pantei β_1 , avem formula echivalentă:

$$\hat{\beta}_1 = \frac{\text{cov}_e(x, y)}{s_x^2} = r_{xy} \frac{s_y}{s_x}, \quad \text{unde} \quad r_{xy} = \frac{\text{cov}_e(x, y)}{s_x s_y}.$$

Matricea Hessiană calculată în acest punct critic $(\hat{\beta}_0, \hat{\beta}_1)$ este negativ definită, deci tripletul critic este punct de maxim. Astfel, dreapta de regresie a lui y în raport cu x este aproximată de dreapta:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} - \frac{s_{xy}}{s_{xx}} \bar{x} + \frac{s_{xy}}{s_{xx}} x, \quad \text{sau, echivalent,} \quad y = \bar{y} + \frac{s_{xy}}{s_{xx}} (x - \bar{x}). \quad (8)$$

Din ultima condiție de extrem, găsim că o estimație pentru dispersia erorilor, σ^2 este:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2. \quad (9)$$

Însă, estimația pentru σ^2 dată prin formula (9) este una deplasată. În practică, în locul acestei estimații se utilizează următoarea estimație nedeplasată:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2. \quad (10)$$

Observația 2.1 În formulele anterioare (7), (8), (9), (10) am obținut estimări ale estimatorilor. Pentru a trece de la estimări la estimatori, care sunt statistici (deci variabile aleatoare) valorile estimate se înlocuiesc cu variabilele de selecție, respectiv caracteristicile din care provin. Mai precis, **caracterul determinist al valorilor independente x_i și cel aleator al valorilor y_i** duce la înlocuirea lor în formula estimatorului cu Y_i , iar $\bar{y}$ trebuie înlocuit cu "media de selecție" $\bar{Y}$. Vom obține astfel:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 (Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \frac{1}{n} \left(\sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n x_i \right), \quad \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2.$$

Cu toate aceste elemente obținute, păstrând convenția din literatura de specialitate, notăm statisticile ale căror realizări sunt $\hat{\beta}_0, \hat{\beta}_1, y_i$ tot cu aceleași simboluri, neomițând caracterul aleator al unor cantități, atunci când este cazul.

Sistematizând notațiile și cantitățile utilizate, obținem:

1. *dreapta de regresie*, $y = \beta_0 + \beta_1 x$, este dreapta ce determină dependența liniară a lui y de valorile lui x , pentru întreaga populație de date (dacă aceasta există);
2. *aproximarea dreptei de regresie* (en., *fitting line*), $y = \hat{\beta}_0 + \hat{\beta}_1 x$, este dreapta care se apropie cel mai mult (în sensul metodei celor mai mici pătrate) de datele experimentale (de selecție) $\{(x_i, y_i)\}_i$. Această dreaptă este o aproximare a dreptei de regresie;
3. valorile y_i se numesc *valori observate*, iar valorile $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, 2, \dots, n$ se numesc *valori prezise*;
4. valorile $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ se numesc *reziduuri* sau *valori reziduale*. Un reziduu măsoară deviația unui punct observat de la valoarea prezisă de estimarea dreptei de regresie;
5. *suma pătratelor erorilor*, $\sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$, se notează de obicei prin SSE (*sum of squared errors*).
6. *eroarea medie pătratică sau reziduală* este $MSE = \frac{SSE}{n-2}$ (mean squared error). După cum se poate observa din relația (10), $MSE = \hat{\sigma}^2$ este un estimator nedeplasat pentru dispersia erorilor, σ^2 .
7. rădăcina pătrată a MSE este $\hat{\sigma}$ și se numește *eroarea standard a regresiei*;

Un **rezultat fundamental** pentru estimarea dispersiei erorilor este dat de următorul comportament al sumei pătratelor erorilor, SSE, furnizat de formula de la punctul (5) anterior.

Teorema 1 Repartiția variabilei aleatoare SSE / σ^2 este următoarea:

$$\frac{SSE}{\sigma^2} = (n-2) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2, 1) = \chi^2(n-2).$$

Demonstrație. Variabilele aleatoare

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \quad \forall i \in \{1, \dots, n\}.$$

Varianța reziduală pe care o utilizăm este cea cu caracter aleator, nedeplasată, mai precis, cea dată de formula (10)

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2.$$

Prin urmare, din formula de mai sus, substituim pe $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, după cum am obținut în formula (7), și avem

$$\begin{aligned} (n-2)\hat{\sigma}^2 &= \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) y_i \stackrel{(7)}{=} \sum_{i=1}^n \left\{ (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right\} y_i = \sum_{i=1}^n (y_i - \bar{y}) y_i - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) y_i \quad (11) \\ &= \sum_{i=1}^n (y_i - \bar{y}) (y_i - \bar{y} + \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y} + \bar{y}) \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y}) + \underbrace{\bar{y} \sum_{i=1}^n (y_i - \bar{y})}_{=n\bar{y}-n\bar{y}=0} - \hat{\beta}_1 \underbrace{\bar{y} \sum_{i=1}^n (x_i - \bar{x})}_{=n\bar{x}-n\bar{x}=0} \\ &= s_{yy} - \hat{\beta}_1 s_{xy} \stackrel{(7)}{=} s_{yy} - \hat{\beta}_1 \left(\hat{\beta}_1 s_{xx} \right) = s_{yy} - \hat{\beta}_1^2 s_{xx}. \end{aligned}$$

Considerăm acum **cantitatea**

$$\tilde{\beta}_0 := \beta_0 + \beta_1 \bar{x},$$

repartiția vectorului aleator $Y = (Y_1, \dots, Y_n)$, $f_Y : \mathbb{R}^n \rightarrow [0, 1]$,

$$\begin{aligned} f_Y(y_1, \dots, y_n) &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 \bar{x} - \beta_1 (x_i - \bar{x}))^2 \right) \quad (12) \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \tilde{\beta}_0 - \beta_1 (x_i - \bar{x}))^2 \right) \end{aligned}$$

și **transformarea ortogonală** $(y_1, \dots, y_n) \rightarrow (z_1, \dots, z_n)$ dată de

$$\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = Q \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{x_1 - \bar{x}}{\sqrt{s_{xx}}} & \frac{x_2 - \bar{x}}{\sqrt{s_{xx}}} & \cdots & \frac{x_n - \bar{x}}{\sqrt{s_{xx}}} \\ \vdots & \vdots & \cdots & \vdots \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Matricea Q este o matrice ortogonală, cu primele două rânduri fixate. Obținem că

$$\left\{ \begin{aligned} z_1 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n y_i = \sqrt{n} \bar{y} \quad \Longleftrightarrow \quad \bar{y} = \frac{z_1}{\sqrt{n}} \quad \text{și} \\ z_2 &= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sqrt{s_{xx}}} = \frac{\sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{s_{xx}}} + \frac{\bar{y} \sum_{i=1}^n (x_i - \bar{x})}{\sqrt{s_{xx}}} = \frac{s_{xy}}{\sqrt{s_{xx}}} + \frac{\bar{y} \left(\sum_{i=1}^n x_i - n\bar{x} \right)}{\sqrt{s_{xx}}} \\ &= \frac{s_{xy}}{\sqrt{s_{xx}}} = \underbrace{\hat{\beta}_1 s_{xx}}_{=s_{xy}} \frac{1}{\sqrt{s_{xx}}} = \hat{\beta}_1 \sqrt{s_{xx}}. \end{aligned} \right.$$

Se poate demonstra că transformarea ortogonală Q implică faptul că

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n z_i^2, \quad (13)$$

ceea ce conduce la următoarea reprezentare a sumei din termenul exponențial definit de densitatea de repartiție

vectorială (12):

$$\begin{aligned}
\sum_{i=1}^n \left(y_i - \tilde{\beta}_0 - \beta_1 (x_i - \bar{x}) \right)^2 &= \sum_{i=1}^n y_i^2 + \sum_{i=1}^n \tilde{\beta}_0^2 + \sum_{i=1}^n \beta_1^2 (x_i - \bar{x})^2 - 2\tilde{\beta}_0 \sum_{i=1}^n y_i \\
&\quad - 2\beta_1 \sum_{i=1}^n (x_i - \bar{x}) y_i - 2\tilde{\beta}_0 \beta_1 \sum_{i=1}^n (x_i - \bar{x}) \\
&= \sum_{i=1}^n y_i^2 + n\tilde{\beta}_0^2 + \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 - 2\tilde{\beta}_0 n\bar{y} - 2\beta_1 \sum_{i=1}^n (x_i - \bar{x}) y_i \quad (14) \\
&\stackrel{(13)}{=} \sum_{i=1}^n z_i^2 + n\tilde{\beta}_0^2 + \beta_1^2 s_{xx} - 2\tilde{\beta}_0 \sqrt{n} z_1 - 2\beta_1 z_2 \sqrt{s_{xx}} \\
&= \left(z_1 - \tilde{\beta}_0 \sqrt{n} \right)^2 + \left(z_2 - \beta_1 \sqrt{s_{xx}} \right)^2 + \sum_{i=3}^n z_i^2.
\end{aligned}$$

Folosim reprezentarea anterioară (14) și deducem că, pentru $(z_1, \dots, z_n) \in \mathbb{R}^n$, densitatea vectorului aleator transformat $Z = (Z_1, \dots, Z_n)$ devine $f_Z : \mathbb{R}^n \rightarrow [0, 1]$,

$$f_Z(z_1, \dots, z_n) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp \left[-\frac{1}{2\sigma^2} \left(\left(z_1 - \tilde{\beta}_0 \sqrt{n} \right)^2 + \left(z_2 - \beta_1 \sqrt{s_{xx}} \right)^2 + \sum_{i=3}^n z_i^2 \right) \right].$$

Conform caracterizării repartițiilor Gaussiene multi-dimensionale, rezultă că variabilele aleatoare asociate transformării, $Z_1, Z_2, \dots, Z_n$ sunt variabile independente, de tip normal. Mai precis, ele au următoarele repartiții:

$$Z_1 \sim \mathcal{N}(\tilde{\beta}_0 \sqrt{n}, \sigma^2), \quad Z_2 \sim \mathcal{N}(\beta_1 \sqrt{s_{xx}}, \sigma^2), \quad Z_i \sim \mathcal{N}(0, \sigma^2), \quad \text{pentru } i \in \{3, 4, \dots, n\}.$$

Pentru a încheia demonstrația, folosim (11) și deducem:

$$\begin{aligned}
(n-2)\hat{\sigma}^2 &= s_{yy} - \hat{\beta}_1^2 s_{xx} = \sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\beta}_1^2 s_{xx} \\
&= \sum_{i=1}^n y_i^2 + \sum_{i=1}^n \bar{y}^2 - 2\bar{y} \sum_{i=1}^n y_i - \hat{\beta}_1^2 s_{xx} = \sum_{i=1}^n y_i^2 + n\bar{y}^2 - 2n\bar{y} - \hat{\beta}_1^2 s_{xx} \\
&= \sum_{i=1}^n y_i^2 - n\bar{y}^2 - \hat{\beta}_1^2 s_{xx} \stackrel{(13)}{=} \sum_{i=1}^n z_i^2 - z_1^2 - z_2^2 = \sum_{i=3}^n z_i^2.
\end{aligned}$$

Independența variabilelor aleatoare $Z_3, Z_4, \dots, Z_n \sim \mathcal{N}(0, \sigma^2)$ conduce la

$$\sum_{i=3}^n \frac{Z_i^2}{\sigma^2} \sim \chi^2(n-2, 1) = \chi^2(n-2), \quad \text{ceea ce este echivalent cu} \quad \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Singura chestiune care mai trebuie justificată este afirmația anterioară că $\sum_{i=3}^n Z_i^2 / \sigma^2$ este repartizată χ^2 cu $(n-2)$ grade de libertate (ale variabilei SSE). Prezentăm raționamentul care conduce la această concluzie. Demonstrațiile riguroase pentru fiecare afirmație se regăsesc în cele patru rezultate care urmează. Avem că

$$Z_i \sim \mathcal{N}(0, \sigma^2) \implies Z_i^2 \sim \chi^2(1, \sigma), \quad \forall i \in \{3, 4, \dots, n\}.$$

Prin urmare,

$$\frac{Z_i^2}{\sigma^2} \sim \chi^2\left(1, \sigma \frac{1}{\sigma}\right) = \chi^2(1, 1) = \chi^2(1) \quad \forall i \in \{3, 4, \dots, n\}.$$

În consecință, independența variabilelor aleatoare Z_i conduce la faptul că

$$\sum_{i=3}^n \frac{Z_i^2}{\sigma^2} \sim \chi^2(n-2, 1) = \chi^2(n-2),$$

iar demonstrația este, în acest moment, încheiată.

Lema 2.1 Pentru orice $a > 0$,

$$X \sim \chi^2(n, \sigma) \quad \text{dacă și numai dacă} \quad aX \sim \chi^2(n, \sqrt{a}\sigma), \quad (15)$$

unde $n \in \mathbb{N}^*$ și $\sigma > 0$.

Demonstrație. Avem, pentru orice $x \leq 0$, $F_{aX}(x) = 0$ și pentru orice $x > 0$,

$$F_{aX}(x) = \mathbb{P}(aX \leq x) = \mathbb{P}(X \leq x/a) = F_X(x/a).$$

Deci

$$f_{aX}(x) = (F_{aX}(x))' = (F_X(x/a))' = f_X\left(\frac{x}{a}\right) \frac{1}{a} = \frac{1}{2^{\frac{n}{2}} (\sqrt{a}\sigma)^n \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2(\sqrt{a}\sigma)^2}\right),$$

adică $aX \sim \chi^2(n, \sqrt{a}\sigma)$.

Lema 2.2 Dacă X, Y sunt două variabile aleatoare independente, distribuite normal, de tipul $\mathcal{N}(0, \sigma^2)$, unde $\sigma > 0$, atunci

$$X^2 \sim \chi^2(1, \sigma) \quad \text{și} \quad (X^2 + Y^2) \sim \chi^2(2, \sigma).$$

Demonstrație. Avem, pentru orice $y \leq 0$, $F_{X^2}(y) = 0$ și pentru orice $y > 0$,

$$F_{X^2}(y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

Deci

$$\begin{aligned} f_{X^2}(y) &= (F_{X^2}(y))' = (F_X(\sqrt{y}) - F_X(-\sqrt{y}))' = f_X(\sqrt{y}) \frac{1}{2\sqrt{y}} + f_X(-\sqrt{y}) \frac{1}{2\sqrt{y}} \\ &= f_X(\sqrt{y}) \frac{1}{\sqrt{y}} = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{(\sqrt{y})^2}{2\sigma^2}\right) \frac{1}{\sqrt{y}} = \frac{1}{\sqrt{2\pi}\sigma} y^{\frac{1}{2}-1} \exp\left(-\frac{y}{2\sigma^2}\right), \end{aligned}$$

adică X^2 corespunde unei variabile aleatoare distribuite $\chi^2(1, \sigma)$. Dacă $X, Y \sim \mathcal{N}(0, \sigma^2)$, atunci $X^2, Y^2 \sim \chi^2(1, \sigma)$ și, prin urmare, $(X^2 + Y^2) \sim \chi^2(1 + 1, \sigma)$.

Rezultatul se poate generaliza la cazul a n variabile aleatoare independente.

Lema 2.3 Dacă $X_i, i = \overline{1, n}$, sunt variabile aleatoare de selecție corespunzătoare unei selecții de volum n asupra caracteristicii $X \sim \mathcal{N}(0, \sigma^2)$, unde $\sigma > 0$, atunci

$$\sum_{i=1}^n X_i^2 \sim \chi^2(n, \sigma).$$

Demonstrație. Conform rezultatului anterior, $Y_k = \frac{1}{\sigma^2} X_k^2 \sim \chi^2(1)$, pentru orice $k = \overline{1, n}$,

$$f_{Y_k}(x) = \frac{1}{\sqrt{2\pi}x} e^{-x/2} \mathbf{1}_{(0, +\infty)}(x)$$

și atunci funcția sa caracteristică este $\varphi_{Y_k} : \mathbb{R} \rightarrow \mathbb{C}$,

$$\varphi_{Y_k}(t) = \mathbb{E}(e^{itY_k}) = \int_0^{+\infty} e^{itx} \frac{1}{\sqrt{2\pi}x} e^{-x/2} dx = (1 - 2it)^{-1/2}.$$

Independența variabilelor aleatoare $Y_k, k = \overline{1, n}$, conduce la următoarea funcție caracteristică pentru $\sum_{k=1}^n Y_k$:

$$\varphi_{\sum_{k=1}^n Y_k}(t) = \prod_{k=1}^n (1 - 2it)^{-1/2} = (1 - 2it)^{-n/2}, \quad t \in \mathbb{R},$$

adică $\sum_{k=1}^n Y_k \sim \chi^2(n) = \chi^2(n, 1)$. De aici rezultă că

$$\sum_{k=1}^n Y_k = \sum_{k=1}^n \frac{1}{\sigma^2} X_k^2 = \frac{1}{\sigma^2} \sum_{k=1}^n X_k^2 \sim \chi^2(n, 1), \quad \text{adică} \quad \sum_{k=1}^n X_k^2 \sim \chi^2(n, \sigma),$$

demonstrația fiind, astfel, încheiată.

Lema 2.4 Considerăm $X_i, i = \overline{1, n}$, variabile aleatoare de selecție corespunzătoare unei selecții de volum n asupra caracteristicii $X \sim \mathcal{N}(\mu, \sigma^2)$, unde $\sigma > 0$.

(a) Dacă media caracteristicii este cunoscută, atunci:

$$H^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \sim \chi^2(n, 1) = \chi^2(n).$$

(b) Dacă media caracteristicii este necunoscută, considerăm media de selecție $\bar{X} = (\sum_{i=1}^n X_i) / n$ și vom avea:

$$\sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2(n-1, \sigma) \quad \text{sau, echivalent} \quad \chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2(n-1, 1) = \chi^2(n-1).$$

Demonstrație. Avem că suma $\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$ și apoi $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. Prin urmare, deducem că

$$(X_i - \mu) \sim \mathcal{N}(0, \sigma^2) \quad \text{și} \quad (\bar{X} - \mu) \sim \mathcal{N}(0, \sigma^2/n).$$

În consecință, $(X_i - \mu)^2 \sim \chi^2(1, \sigma)$ ceea ce conduce la

$$\sum_{i=1}^n (X_i - \mu)^2 \sim \chi^2(n, \sigma).$$

Obținem că

$$H^2 = \frac{n}{\sigma^2} \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \sim \chi^2(n, 1) = \chi^2(n)$$

De asemenea,

$$(\bar{X} - \mu)^2 \sim \chi^2(1, \sigma/\sqrt{n}) \quad \text{și} \quad n(\bar{X} - \mu)^2 \sim \chi^2(1, \sigma).$$

Pe de altă parte, avem că

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n [(X_i - \mu) - (\bar{X} - \mu)]^2 = \sum_{i=1}^n [(X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2] \\ &= \sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + \sum_{i=1}^n (\bar{X} - \mu)^2 = \sum_{i=1}^n (X_i - \mu)^2 - 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2. \end{aligned}$$

Deci

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \sim \chi^2(n, \sigma) - \chi^2(1, \sigma) = \chi^2(n-1, \sigma),$$

ceea ce conduce la

$$\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2(n-1, 1) = \chi^2(n-1).$$

Determinăm repartiția dispersiei de selecție modificată astfel:

$$\chi^2 = \frac{n}{\sigma^2} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n}{\sigma^2} S^2 = \frac{n}{\sigma^2} \frac{n-1}{n} (S^*)^2 = \frac{n-1}{\sigma^2} (S^*)^2.$$

Rezultă că $(n-1)(S^*)^2 \sim \chi^2(n-1, \sigma)$ și, în mod similar, $n(S^*)^2 \sim \chi^2(n, \sigma)$.

Estimația dispersiei este o măsură a gradului de împrăștiere a punctelor (x, y) în jurul dreptei de regresie. Valorile din formulele (7) și (10):

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{și} \quad \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \quad (16)$$

sunt doar estimății ale parametrilor necunoscuți și nu valorile lor exacte.

Dacă deviația standard σ ar fi cunoscută *a priori*, atunci putem estima parametrii β_0 și β_1 prin **Metoda celor mai mici pătrate** (primele formulări ale problemei îi aparțin lui Gauss). Mai precis, estimațiile lor vor fi acele valori care minimizează suma pătratelor erorilor SSE:

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

Suma pătratelor deviațiilor observațiilor de la dreapta de regresie este înmagazinată în funcția

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

iar condițiile ce determină punctele critice sunt:

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

Rezolvând acest sistem de ecuații algebrice în raport cu β_0 și β_1 , găsim drept puncte critice soluțiile $\hat{\beta}_0$ și, respectiv, $\hat{\beta}_1$ de mai sus. Într-adevăr,

$$\begin{cases} n\beta_0 + n\bar{x}\beta_1 = n\bar{y} \\ n\bar{x}\beta_0 + \left(\sum_{i=1}^n x_i^2\right)\beta_1 = \sum_{i=1}^n x_i y_i \end{cases} \iff \begin{cases} n\bar{x}\beta_0 + n\bar{x}^2\beta_1 = n\bar{y}\bar{x} \\ n\bar{x}\beta_0 + \left(\sum_{i=1}^n x_i^2\right)\beta_1 = \sum_{i=1}^n x_i y_i \end{cases}$$

Deci, scăzând ecuațiile, obținem că

$$\beta_1 = \frac{n\bar{y}\bar{x} - \sum_{i=1}^n x_i y_i}{n\bar{x}^2 - \sum_{i=1}^n x_i^2} \quad (17)$$

Dar, pe de altă parte, dacă prelucrăm în formula (16) expresia lui $\hat{\beta}_1$, obținem forma echivalentă

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

adică exact formula obținută și în (17). Regăsim, în mod similar, și forma pentru β_0 identică cu cea din (16).

Pentru matricea Hessiană

$$H = \begin{pmatrix} \frac{\partial^2 \Lambda}{\partial \beta_0^2} & \frac{\partial^2 \Lambda}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \Lambda}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \Lambda}{\partial \beta_1^2} \end{pmatrix} = \begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum_{i=1}^n x_i^2 \end{pmatrix}$$

avem $2n > 0$ și, pentru $n \geq 3$,

$$\begin{aligned}\det(H) &= 4n \sum_{i=1}^n x_i^2 - 4n^2 \bar{x}^2 = 4 \left(n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right) \\ &= 4 \left((n-1) \sum_{i=1}^n x_i^2 - 2 \sum_{1 \leq i < j \leq n} x_i x_j \right) \geq 0.\end{aligned}$$

Deci H este pozitiv definită, adică punctele critice obținute $(\hat{\beta}_0, \hat{\beta}_1)$ sunt puncte de minimum pentru funcția Λ . Sistemul de ecuații liniare

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

poartă denumirea de *ecuațiile normale ale celor mai mici pătrate*.

În concluzie, se observă că, dacă erorile ε_i sunt variabile aleatoare identic Gaussian repartizate și independente stochastic în ansamblu, **metoda verosimilității maxime este, în fapt, echivalentă cu metoda celor mai mici pătrate**.

Așa cum am descris anterior, un reziduu (sau *măsură reziduală*) măsoară deviația unui punct observat de la valoarea prezisă de estimarea dreptei de regresie. În spiritul acestei observații, putem să furnizăm un rezultat mai precis privitor la comportamentul dispersiei (varianței) acestei cantități estimate $\hat{\varepsilon}_i = y_i - \hat{y}_i$.

Propoziția 2.1 Pentru fiecare $i \in \{1, \dots, n\}$,

$$D^2(\hat{\varepsilon}_i) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{s_{xx}} \right).$$

Demonstrație. Pentru regresia liniară

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

$D^2(\varepsilon_i) = \sigma^2$, iar dispersia estimatorului $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^t$ este

$$D^2(\hat{\beta}) = \sigma^2 (X^t X)^{-1},$$

unde

$$X^t X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}.$$

Prin urmare,

$$(X^t X)^{-1} = \left(n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right)^{-1} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

În ceea ce privește $\det(X^t X)$, obținem

$$n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 = n \sum_{i=1}^n x_i^2 - n^2 \bar{x}^2 = n \sum_{i=1}^n (x_i^2 - \bar{x}^2) = n s_{xx},$$

ultima egalitate având loc deoarece

$$s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2 = \sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n (x_i^2 - \bar{x}^2).$$

Obținem deci

$$\frac{1}{\sigma^2} D^2(\hat{\beta}) = (X^t X)^{-1} = \frac{1}{n s_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} = \frac{1}{s_{xx}} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}. \quad (18)$$

Folosind forma echivalentă determinată anterior pentru cantitatea s_{xx} ,

$$s_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2 \iff \sum_{i=1}^n x_i^2 = s_{xx} + n\bar{x}^2,$$

identificăm din scrierea matriceală (18) componentele următoare:

$$\begin{cases} D^2(\hat{\beta}_0) &= \frac{\sigma^2}{s_{xx}} \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) = \frac{\sigma^2}{n} \frac{s_{xx} + n\bar{x}^2}{s_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right), \\ D^2(\hat{\beta}_1) &= \frac{\sigma^2}{s_{xx}} \quad \text{și} \\ cov(\hat{\beta}_0, \hat{\beta}_1) &= -\frac{\sigma^2 \bar{x}}{s_{xx}}. \end{cases} \quad (19)$$

Valoarea reziduală cu indicele i , fixat, este diferența dintre valoarea observată y_i și cea prezisă $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$:

$$\hat{\varepsilon}_i := y_i - \hat{y}_i = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_i + \varepsilon_i,$$

iar dispersia sa se obține astfel:

$$\begin{aligned} D^2(\hat{\varepsilon}_i) &= D^2 \left((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_i + \varepsilon_i \right) \\ &= D^2(\varepsilon_i) + D^2(\hat{\beta}_0) + x_i^2 D^2(\hat{\beta}_1) + 2x_i cov(\hat{\beta}_0, \hat{\beta}_1) + 2cov((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_i, \varepsilon_i) \\ &\stackrel{(19)}{=} \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right) + x_i^2 \frac{\sigma^2}{s_{xx}} - 2x_i \frac{\sigma^2 \bar{x}}{s_{xx}} + 2cov((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_i, \varepsilon_i) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} + \frac{x_i^2}{s_{xx}} - 2 \frac{\bar{x}x_i}{s_{xx}} \right) + 2cov((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_i, \varepsilon_i) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}} \right) + 2cov((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_i, \varepsilon_i). \end{aligned} \quad (20)$$

Pentru ultimul termen, ce conține covarianța procedăm astfel. Fixăm indicele i și obținem, deoarece $\mathbb{E}(\varepsilon_i) = 0$

și, conform cu (7) $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$,

$$\begin{aligned}
2\text{cov}((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_i, \varepsilon_i) &= 2\mathbb{E}\left(\varepsilon_i \left((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_i\right)\right) \\
&\stackrel{(7)}{=} 2\beta_0\mathbb{E}(\varepsilon_i) - 2\mathbb{E}(\varepsilon_i \hat{\beta}_0) + 2\beta_1 x_i \mathbb{E}(\varepsilon_i) - 2x_i \mathbb{E}(\hat{\beta}_1 \varepsilon_i) = -2\mathbb{E}(\varepsilon_i(\bar{y} - \hat{\beta}_1 \bar{x})) - 2x_i \mathbb{E}(\hat{\beta}_1 \varepsilon_i) \\
&= -2\mathbb{E}(\bar{y}\varepsilon_i) - 2(x_i - \bar{x})\mathbb{E}(\hat{\beta}_1 \varepsilon_i) = -2\frac{\sigma^2}{n} - 2(x_i - \bar{x})\mathbb{E}\left(\frac{\sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{s_{xx}}\varepsilon_i\right) \\
&= -2\frac{\sigma^2}{n} - 2\frac{x_i - \bar{x}}{s_{xx}}\left(\sum_{j=1}^n (x_j - \bar{x})\mathbb{E}(y_j \varepsilon_i - \bar{y}\varepsilon_i)\right) \\
&= -2\frac{\sigma^2}{n} - 2\frac{x_i - \bar{x}}{s_{xx}}\left(-\frac{\sigma^2}{n}\sum_{j \neq i} (x_j - \bar{x}) + (x_i - \bar{x})\sigma^2\left(1 - \frac{1}{n}\right)\right) \\
&= -2\frac{\sigma^2}{n} - 2\frac{x_i - \bar{x}}{s_{xx}}\left(-\frac{\sigma^2}{n}\underbrace{\sum_{j=1}^n (x_j - \bar{x})}_{=0} + (x_i - \bar{x})\sigma^2\right) \\
&= -2\frac{\sigma^2}{n} - 2\frac{x_i - \bar{x}}{s_{xx}}(x_i - \bar{x})\sigma^2 = -2\frac{\sigma^2}{n} - 2\sigma^2\frac{(x_i - \bar{x})^2}{s_{xx}} = \sigma^2\left(-\frac{2}{n} - \frac{2(x_i - \bar{x})^2}{s_{xx}}\right).
\end{aligned} \tag{21}$$

La rândul al treilea din (21) am folosit formula $\mathbb{E}(\bar{y}\varepsilon_i) = \sigma^2/n$. Aceasta are loc datorită următorului raționament:

$$\mathbb{E}(\bar{y}\varepsilon_i) = \mathbb{E}\left(\left(\frac{1}{n}\sum_{j=1}^n y_j\right)\varepsilon_i\right) = \frac{1}{n}\mathbb{E}\left(\sum_{j=1}^n (\beta_0 + \beta_1 x_j + \varepsilon_j)\varepsilon_i\right) = \frac{1}{n}\sum_{j=1}^n \text{cov}(\varepsilon_j \varepsilon_i) = \frac{1}{n}D^2(\varepsilon_i) = \frac{\sigma^2}{n}.$$

Inserăm acum formula (21) în formula (20), ce dă dispersia rezidului, și obținem

$$D^2(\hat{\varepsilon}_i) = \sigma^2\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{s_{xx}}\right),$$

demonstrația fiind încheiată.

2.1 Intervale de încredere pentru parametrii de regresie

Reamintim că, în ipoteza că erorile ε din modelul

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

sunt variabile aleatoare de tip $\mathcal{N}(0, \sigma^2)$, iar valorile x_i sunt fixate (determinate), Y va fi o variabilă aleatoare cu media $\mu_{Y|x} = \beta_0 + \beta_1 x$ și dispersia $\text{Var}(Y) = D^2(Y) = D^2(\varepsilon) = \sigma^2$. Prin urmare, valorile estimate $\hat{\beta}_0$ și $\hat{\beta}_1$ ale parametrilor de regresie depind de valorile observate y_i .

Pentru a decide dacă valorile calculate pe baza datelor experimentale $\{(x_i, y_i)\}_{i=1}^n$ pot fi considerate valorile potrivite pentru întreaga populație, se vor utiliza teste statistice pentru testarea valorilor ambilor parametri, β_0 și β_1 , însă **cel mai uzual test este testul pentru verificarea valorii pantei drepte de regresie**, β_1 .

Arătăm, pentru început că $\hat{\beta}_0$ și $\hat{\beta}_1$ sunt estimatori nedeplasați pentru β_0 și, respectiv, β_1 (cititorul interesat poate consulta Montgomery, Runger [11]). Printr-un abuz de notație, vom utiliza, ca și până în acest moment, aceleași notații ($\hat{\beta}_0$ și $\hat{\beta}_1$, ca și valorile estimate) pentru cele două statistici care reprezintă estimatorii. Obținem, deoarece

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \frac{1}{n}\sum_{i=1}^n \varepsilon_i,$$

iar x_i sunt deterministe și y_i valori ale unei variabile aleatoare,

$$\mathbb{E}(\hat{\beta}_1) = \mathbb{E}\left(\frac{s_{xy}}{s_{xx}}\right) = \mathbb{E}\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \frac{\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\beta_1 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1,$$

deoarece

$$\mathbb{E}(y_i - \bar{y}) = \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 \bar{x}) = \beta_1 (x_i - \bar{x}).$$

Cum

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

putem concluda că **cei doi estimatori sunt nedeplasați**:

$$\mathbb{E}(\hat{\beta}_1) = \beta_1 \quad \text{și} \quad \mathbb{E}(\hat{\beta}_0) = \beta_0.$$

Ne concentrăm acum pe determinarea dispersiilor acestor estimatori, abordând acum fiecare estimator în parte (demonstrația fiind, oarecum, mai simplă), spre deosebire de abordarea vectorială a lor de la finalul secțiunii anterioare. Avem, folosind reprezentarea pe care deja am utilizat-o pentru $\hat{\beta}_1$,

$$D^2(\hat{\beta}_1) = D^2\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \frac{1}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} \sum_{i=1}^n (x_i - \bar{x})^2 D^2(y_i) = \frac{\sigma^2}{s_{xx}}$$

Pentru $D^2(\hat{\beta}_0)$ procedăm astfel:

$$\begin{aligned} D^2(\hat{\beta}_0) &= D^2(\bar{y} - \hat{\beta}_1 \bar{x}) = D^2(\bar{y}) - 2\bar{x} \text{cov}(\bar{y}, \hat{\beta}_1) + \bar{x}^2 D^2(\hat{\beta}_1) \\ &= D^2\left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i\right) - 2\bar{x} \text{cov}\left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i, \frac{\sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 x_i + \varepsilon_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) + \frac{\sigma^2 \bar{x}^2}{s_{xx}} \\ &= \frac{\sigma^2}{n} - 2\bar{x} \text{cov}\left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i, \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) + \frac{\sigma^2 \bar{x}^2}{s_{xx}} = \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right) - \frac{2\bar{x}}{n \sum_{i=1}^n (x_i - \bar{x})^2} \text{cov}\left(\sum_{i=1}^n \varepsilon_i, \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i\right) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right) \end{aligned}$$

Am obținut deci că:

$$D^2(\hat{\beta}_1) = \frac{\sigma^2}{s_{xx}} \quad \text{și} \quad D^2(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right). \quad (22)$$

Este important de remarcat faptul că, deoarece $\hat{\beta}_1$ (și, în consecință, și $\hat{\beta}_0$) sunt transformări liniare ale unor variabile aleatoare independente, repartizate normal, atunci și acești estimatori vor avea repartiții de tip Gaussian. Această observație, împreună cu formulele mediilor și dispersiilor lor, determinate anterior, conduc către:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right) \quad \text{și} \quad \hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)\right).$$

Ținând cont că estimatorii $\hat{\beta}_0$ și $\hat{\beta}_1$ sunt nedeplasați, de relațiile (22), și de estimatorul $\hat{\sigma}^2$ pentru σ^2 , demonstrăm următoarele repartiții pentru fracțiile de mai jos:

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}} \sim t(n-2) \quad \text{și} \quad T_2 = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}} \sim t(n-2), \quad (23)$$

unde, ținând cont de formula lui $\hat{\sigma}^2$ dată de (10),

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}.$$

Pentru obținerea rezultatelor din (23) trebuie parcurși doi pași. Primul constă în determinarea repartiției numărătorului și numitorului, după care determinăm repartiția cântului celor două. Deoarece $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma^2/s_{xx})$ și $\hat{\beta}_0 \sim \mathcal{N}(\beta_0, \sigma^2(1/n + \bar{x}^2/s_{xx}))$, obținem că $\hat{\beta}_1 - \beta_1 \sim \mathcal{N}(0, \sigma^2/s_{xx})$, iar $\hat{\beta}_0 - \beta_0 \sim \mathcal{N}(0, \sigma^2(1/n + \bar{x}^2/s_{xx}))$.

În ceea ce privește numitorul, au loc echivalențele:

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2, 1) \iff \frac{(n-2)\hat{\sigma}^2}{s_{xx}} \sim \chi^2(n-2, 1) \iff \frac{(n-2)\hat{\sigma}^2}{s_{xx}} \sim \chi^2\left(n-2, \frac{\sigma}{\sqrt{s_{xx}}}\right)$$

Pentru cel de al doilea pas, repartiția statisticii T_1 se deduce folosind următorul rezultat. Într-o manieră absolut similară determinăm distribuția lui T_2 .

Lema 2.5 Dacă $X \sim \mathcal{N}(0, \sigma^2)$ și $Y \sim \chi^2(n, \sigma)$, unde $n \in \mathbb{N}^*$ și $\sigma > 0$, sunt două variabile aleatoare independente, atunci distribuția

$$T = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t(n).$$

Demonstrație. Vectorul aleator (X, Y) are densitatea de repartiție, pentru $x \in \mathbb{R}$, $y \geq 0$,

$$f_{(X,Y)}(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \cdot \frac{1}{2^{n/2}\sigma^n \Gamma(n/2)} y^{\frac{n}{2}-1} \exp\left(-\frac{y}{2\sigma^2}\right) = \frac{1}{\sqrt{\pi}\sigma^{n+1} 2^{\frac{n+1}{2}} \Gamma(n/2)} y^{\frac{n}{2}-1} \exp\left(-\frac{x^2+y}{2\sigma^2}\right).$$

Să considerăm transformarea

$$\begin{cases} u = \frac{x}{\sqrt{\frac{y}{n}}}, \\ v = y, \end{cases} \quad \text{cu inversa} \quad \begin{cases} x = \frac{u\sqrt{v}}{\sqrt{n}}, \\ y = v, \end{cases} \quad \text{și Jacobianul } J(u, v) = \sqrt{v}/\sqrt{n}.$$

Obținem astfel densitatea de repartiție a vectorului aleator $f_{(U,V)}$:

$$f_{(U,V)}(u, v) = f_{(X,Y)}\left(\frac{u\sqrt{v}}{\sqrt{n}}, v\right) \left| \frac{\sqrt{v}}{\sqrt{n}} \right| = \frac{v^{\frac{n-1}{2}} \exp\left(-\frac{v}{2\sigma^2} \left(\frac{u^2}{n} + 1\right)\right)}{\sqrt{n\pi}\sigma^{n+1} 2^{\frac{n+1}{2}} \Gamma(n/2)}.$$

Densitatea marginală este dată de (avem $v = y \geq 0$)

$$f_U(u) = \frac{1}{\sqrt{n\pi}\sigma^{n+1} 2^{\frac{n+1}{2}} \Gamma(n/2)} \int_0^{+\infty} v^{\frac{n-1}{2}} \exp\left(-\frac{v}{2\sigma^2} \left(\frac{u^2}{n} + 1\right)\right) dv.$$

Facem substituția $\frac{v}{2\sigma^2} \left(\frac{u^2}{n} + 1\right) = v'$ deci $dv = 2\sigma^2 dv' / \left(\frac{u^2}{n} + 1\right)$ și

$$\begin{aligned} f_U(u) &= \frac{1}{\sqrt{n\pi}\sigma^{n+1} 2^{\frac{n+1}{2}} \Gamma(n/2)} \int_0^{+\infty} \left(\frac{2\sigma^2}{\frac{u^2}{n} + 1}\right)^{\frac{n-1}{2}} (v')^{\frac{n-1}{2}} e^{-v'} \frac{2\sigma^2}{\frac{u^2}{n} + 1} dv' \\ &= \frac{1}{\sqrt{n\pi}\Gamma(n/2)} \left(\frac{u^2}{n} + 1\right)^{-\frac{n+1}{2}} \int_0^{+\infty} v^{\frac{n-1}{2}} e^{-v} dv = \frac{1}{\sqrt{n\pi}\Gamma(n/2)} \left(\frac{u^2}{n} + 1\right)^{-\frac{n+1}{2}} \Gamma\left(\frac{n+1}{2}\right). \end{aligned}$$

Deci U urmează distribuția Student cu n grade de libertate.

Formulele (23) sunt cruciale în construcția intervalelor de încredere pentru parametrii β_1 și β_0 . Pentru un nivel de încredere $\delta > 0$, unul de semnificație α și $t \geq 0$, avem

$$\alpha = 1 - \delta = 1 - \mathbb{P}(|T_1| < t) = \mathbb{P}\left(\left|\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}}\right| \geq t\right) = 2\mathbb{P}\left(\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}} \geq t\right)$$

Punctul critic reprezentat de cuantila $t_{\alpha/2, n-2}$ se va citi din tabelul repartiției student cu $n-2$ grade de libertate. Inegalitatea

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}} \geq t_{\alpha/2, n-2} \quad \text{implică} \quad \beta_1 \leq \hat{\beta}_1 - t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}.$$

Obținem intervalul de încredere la nivelul de semnificație α (pentru parametrul β_1):

$$\left(\hat{\beta}_1 - t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}, \hat{\beta}_1 + t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{s_{xx}}} \right). \quad (24)$$

Similar, un interval de încredere pentru parametrul β_0 , la nivelul de semnificație α , este:

$$\left(\hat{\beta}_0 - t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}, \hat{\beta}_0 + t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}} \right) \quad (25)$$

Observația 2.2 În general, dispersia σ^2 a erorilor de regresie ε_i nu este cunoscută a priori. În cazul în care aceasta este cunoscută, atunci în loc de (23) am avea:

$$Z_1 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{s_{xx}}}} \sim \mathcal{N}(0, 1) \quad \text{și} \quad Z_2 = \frac{\hat{\beta}_0 - \beta_0}{\sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}} \sim \mathcal{N}(0, 1). \quad (26)$$

În acest caz, intervalele de încredere pentru β_0 și β_1 vor fi similare cu cele din relațiile (25) și (24), cu diferența că $t_{\alpha/2, n-2}$ este înlocuit prin $z_{1-\alpha/2}$. Într-adevăr, pentru $z > 0$, avem:

$$\delta = \mathbb{P}(|Z_1| < z) = \mathbb{P}(-z < Z_1 < z) = \Phi(z) - \Phi(-z) = \Phi(z) - (1 - \Phi(z)) = 2\Phi(z) - 1,$$

unde Φ se numește funcția lui Laplace și reprezintă funcția de repartiție pentru variabile aleatoare repartizate Gaussian standard:

$$\Phi : \mathbb{R} \rightarrow [0, 1], \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy.$$

Prin urmare, obținem $\Phi(z) = (1 + \delta)/2 = (1 + 1 - \alpha)/2 = 1 - \alpha/2$. Cuantila corespunzătoare se identifică din tabelele de valori ale distribuției $\mathcal{N}(0, 1)$, iar valoarea critică pentru testul statistic va fi $z_{1-\alpha/2}$.

2.2 Verificarea ipotezelor statistice pentru β_1 (slope) și pentru β_0 (intercept)

Prezentăm testele statistice care verifică dacă β_1 (respectiv β_0) iau o valoare dată β_{10} (respectiv β_0^*) sau nu, la un nivel de semnificație α . Presupunem că dispersia erorilor de regresie este necunoscută. Formulăm ipoteza nulă versus ipoteza alternativă bilaterală:

$$(H_0) : \beta_1 = \beta_{10} \quad \text{versus} \quad (H_1) : \beta_1 \neq \beta_{10}.$$

Revenim la statistica T_1 , introdusă de formula (23):

$$T_1 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}} \sim t(n-2).$$

Urmăm în continuare etapele clasice ale verificării ipotezelor unui test statistic.

1. Calculăm valoarea empirică observată $t_0 = (\hat{\beta}_1 - \beta_{10})\sqrt{s_{xx}}/\hat{\sigma}$.
2. Identificăm cuantila de ordin $\alpha/2$ pentru repartiția $t(n-2)$, $t_{\alpha/2, n-2}$.
3. Decizia finală este:

$$\begin{cases} \text{Dacă } |t_0| < t_{\alpha/2, n-2}, & \text{atunci ipoteza nulă } (H_0) \text{ este acceptată.} \\ \text{Dacă } |t_0| \geq t_{\alpha/2, n-2}, & \text{atunci ipoteza alternativă bilaterală } (H_1) \text{ este acceptată.} \end{cases}$$

Intervalul de acceptare este $(-t_{\alpha/2, n-2}, t_{\alpha/2, n-2})$, iar $|t| \geq t_{\alpha/2, n-2}$ este regiunea critică. Mai precis,

$$\mathcal{V} = \left\{ \{(x_k, y_k)\}_{k \in \{1, 2, \dots, n\}} : |T_1|_{\{(x_k, y_k)\}_{k \in \{1, 2, \dots, n\}}} \geq t_{\alpha/2, n-2} \right\}.$$

Observația 2.3 (1) O ipoteză alternativă poate fi considerată și una dintre următoarele:

$$(H_1)_s : \beta_1 < \beta_{10} \text{ (ipoteza alternativă stânga)} \quad \text{sau} \quad (H_1)_d : \beta_1 > \beta_{10} \text{ (ipoteza alternativă dreaptă)}.$$

(2) Testul cel mai popular pentru β_1 este pentru ipoteza nulă (H_0) : $\beta_1 = 0$ (adică $\beta_{10} = 0$). Ipoteza alternativă $\beta_1 \neq 0$ reprezintă faptul că între x și y există o dependență liniară. Acest test verifică semnificația pantei drepte de regresie. Dacă ipoteza nulă este respinsă, atunci panta drepte este semnificativ diferită de zero.

Testul statistic pentru β_0 se abordează într-o manieră similară. De asemenea, presupunem că dispersia σ^2 a erorilor de regresie ε_i este necunoscută. Testăm

$$(H_0) : \beta_0 = \beta_0^*, \quad \text{versus} \quad (H_1) : \beta_0 \neq \beta_0^*.$$

Considerăm statistica T_2 , introdusă de formula (23):

$$T_2 = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}} \sim t(n-2).$$

Urmăm în continuare etapele clasice ale verificării ipotezelor unui test statistic.

1. Calculăm valoarea observată $t_0 = (\hat{\beta}_0 - \beta_0^*) / (\hat{\sigma} \sqrt{1/n + \bar{x}^2/s_{xx}})$.
2. Identificăm cuantila de ordin $\alpha/2$ pentru repartiția $t(n-2)$, $t_{\alpha/2, n-2}$.
3. Decizia finală este:

$$\begin{cases} \text{Dacă } |t_0| < t_{\alpha/2, n-2}, & \text{atunci ipoteza nulă } (H_0) \text{ este acceptată.} \\ \text{Dacă } |t_0| \geq t_{\alpha/2, n-2}, & \text{atunci ipoteza alternativă bilaterală } (H_1) \text{ este acceptată.} \end{cases}$$

Intervalul de acceptare este $(-t_{\alpha/2, n-2}, t_{\alpha/2, n-2})$, iar $|t| \geq t_{\alpha/2, n-2}$ este regiunea critică.

Observația 2.4 De asemenea, există teste unilaterale și pentru testarea valorii lui β_0 .

Observația 2.5 Dacă dispersia erorilor, σ^2 , este cunoscută a priori, atunci putem utiliza statisticile repartizate normal Z_1 și Z_2 , definite de (26), statistici care permit folosirea testului z (ca și la determinarea intervalelor de încredere) pentru testarea ipotezelor de mai sus, atât pentru parametrul β_0 , cât și pentru β_1 .

2.3 Predicție prin regresie

În anumite cazuri, putem folosi regresia în predicția unor valori ale variabilei dependente. De exemplu, putem prezice temperatura într-un anumit oraș plecând de la observațiile temperaturilor din orașele învecinate. Regresia poate fi utilizată pentru predicție după cum urmează. Să presupunem că datele pe care le deținem, $\{(x_i, y_i)\}_{i \in \{1, 2, \dots, n\}}$, pot fi modelate de o dreaptă de regresie de forma (3). Dat fiind o valoarea x_p ce nu se află printre valorile x_i , dar este o valoare cuprinsă între valorile extreme ale variabilei independente, $x_{\min}$ și $x_{\max}$, dorim să prezicem valoarea răspuns,

$$y_p = \beta_0 + \beta_1 x_p + \varepsilon_p.$$

Dacă $\hat{\beta}_0$ și $\hat{\beta}_1$ sunt estimațiile pentru parametrii de regresie β_0 , respectiv, β_1 , atunci valoarea prezisă pentru y_p corespunzătoare unui x_p observat va fi o valoare $\hat{y}_p$ de pe dreapta de regresie, dată de formula:

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p. \quad (27)$$

Folosind estimările din relația (20), putem afirma că statistica $\hat{Y}_p$ asociată acestei valori empirice urmează o repartiție normală,

$$\hat{Y}_p \sim \mathcal{N} \left(\hat{\beta}_0 + \hat{\beta}_1 x_p, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{s_{xx}} \right) \right).$$

Urmând pașii secțiunii anterioare, deducem că o predicție pentru intervalul de încredere corespunzător lui y pentru un x_p dat, la nivelul de semnificație α ($x_p \in [x_{\min}, x_{\max}]$) este:

$$\left(\hat{y}_p - t_{\alpha/2, n-2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{s_{xx}}}, \hat{y}_p + t_{\alpha/2, n-2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{s_{xx}}} \right). \quad (28)$$

- Observația 2.6** 1. Este important ca x_p să fie o valoare cuprinsă între $x_{\min}$ și $x_{\max}$. Dacă se folosește formula (27) și pentru valori ale lui x în afara intervalului valorilor predictor pentru x , atunci erorile de aproximarea a lui y cu $\hat{y}_p$ pot fi foarte mari. În practică, pentru estimarea de valori viitoare ale variabilei dependente folosind valori ale variabilelor independente ce ies din domeniu, se utilizează termenul de prognoză. Aceasta este folosită în analiza seriilor de timp.
2. Valoarea prezisă $\hat{y}_p$ nu este una stabilită cu exactitate, ci este doar o medie așteptată a valorilor lui y pentru un x_p dat. În cazul în care coeficientul de determinare $R^2 = 1$, atunci valoarea pentru y va fi prezisă fără eroare, deoarece toate punctele se află pe dreapta de regresie. În general, perechile $(x_i, y_i)_i$ se află împrăștiat în jurul dreptei de regresie.
3. Valoarea $\hat{y}_p$ este determinată doar pe baza selecției date, de aceea, pentru a verifica dacă această valoare poate fi extrapolată la întreaga populație este nevoie de inferență statistică, adică de un test statistic pentru verificarea acceptării ipotezei nule. Un astfel de test, similar celor pentru testarea parametrilor β_0 și β_1 este prezentat succint. Vom testa: Prezentăm, în continuare, un test ce compară valoarea $\hat{y}_p$ cu o constantă dată. Testăm

$$(H_0) : \hat{y}_p = y_0, \quad \text{versus} \quad (H_1) : \hat{y}_p \neq y_0.$$

Etapele testului sunt cele deja discutate:

1. Estimăm valoarea observată, pe baza datelor empirice, $\hat{y}_p$ utilizând formula (27).
2. Considerăm statistica

$$T = \frac{\hat{y}_p - y}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{s_{xx}}}} \sim t(n-2)$$

pe care o evaluăm în datele empirice $t_0 = (\hat{y}_p - y_0) / \left(\hat{\sigma} \sqrt{1 + 1/n + (x_p - \bar{x})^2 / s_{xx}} \right)$.

3. Decizia finală este:

$$\begin{cases} \text{Dacă } |t_0| < t_{\alpha/2, n-2}, & \text{atunci ipoteza nulă } (H_0) \text{ este acceptată.} \\ \text{Dacă } |t_0| \geq t_{\alpha/2, n-2}, & \text{atunci ipoteza alternativă bilaterală } (H_1) \text{ este acceptată.} \end{cases}$$

Intervalul de acceptare este $(-t_{\alpha/2, n-2}, t_{\alpha/2, n-2})$, iar $|t| \geq t_{\alpha/2, n-2}$ este regiunea critică.

2.4 Validitatea modelului de regresie liniară simplă

Presupunem ca X și Y sunt două variabile de interes, pentru care se dorește a determina o relație liniară de forma

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Pentru a determina oportunitatea unei astfel de legături, se culeg date relativ la aceste variabile. Considerăm că aceste observații sunt $\{(x_i, y_i)\}_{i=1,2,\dots,n}$. Pe baza acestor date se poate aproxima dreapta de regresie liniară (dacă există) astfel:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X, \quad \text{unde: } \hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} \quad \text{și} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \text{unde}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Pentru a verifica dacă modelul de regresie liniară este unul valid, se pot folosi mai multe metode, dintre care amintim cele mai uzuale:

- coeficientul de determinare R^2 . Acest coeficient se calculează folosind următoarea formulă:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}},$$

unde

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

Aici, SST reprezintă suma totală a pătratelor (*the total sum of squares*). În analiza regresională, coeficientul R^2 este o statistica folosită în a determina cât de bine pot fi estimate valorile lui y pe baza modelului de regresie. Valorile lui R^2 sunt între 0 și 1 și, pentru a avea un model destul de bun, ar fi necesar un coeficient de determinare aproape de 1. Totuși, este posibil ca R^2 să aibă valori mai mari ca 1 în cazul în care modelul de regresie nu este unul liniar. În cazul regresiei liniare simple, $R^2 = r^2$, adică pătratul coeficientului de corelație Pearson.

- grafice:
 - y_i versus x_i : Din această figură (scatter plot) ne putem da seama de oportunitatea modelării datelor observate folosind un model de regresie liniară simplă. Această figură ar trebui făcută înainte de aproximarea dreptei de regresie. Pentru a putea utiliza un model de regresie liniară simplă, valorile reprezentate ar trebui să fie apropiate de o anumită dreaptă.
 - grafic ce indică normalitatea reziduurilor $\hat{\varepsilon}_i$: Acest grafic reprezintă probabilitățile de normalitate ale erorilor versus cuantilele repartiției $\mathcal{N}(0, 1)$. Dacă modelul este valid, atunci valorile reprezentate în figură vor fi cât mai apropiate de prima bisectoare.
 - $\hat{y}_i$ versus y_i : Dacă modelul este valid, atunci valorile reprezentate în figură vor fi cât mai apropiate de prima bisectoare.
 - $\hat{\varepsilon}_i$ versus x_i : Dacă modelul este valid, atunci valorile reprezentate în figură nu ar avea nicio tendință clară.
 - $\hat{\varepsilon}_i$ versus y_i : Dacă modelul este valid, atunci valorile reprezentate în figură nu ar avea nicio tendință clară.
- test de utilitate a modelului: Se testează ipoteza $(H_0) : \beta_1 = 0$ versus. ipoteza $(H_1) : \beta_1 \neq 0$. După cum am văzut mai sus, acceptarea ipotezei alternative indică faptul că modelul liniar simplu este valid.
- test pentru semnificația parametrilor modelului: Putem testa o valoare anume a pantei dreptei de regresie folosind ipoteza nulă $(H_0) : \beta_1 = \beta_{10}$ vs. ipoteza alternativă $(H_1) : \beta_1 \neq \beta_{10}$. Panta dreptei de regresie este importantă în a determina magnitudinea variației variabilei răspuns la o variație de o unitate a variabilei stimul.

Dacă modelul de regresie liniară simplă nu este unul valid, atunci:

- Este posibil ca Y să nu depindă liniar de X . Acest fapt poate fi observat de la început, din diagrama scatter plot ce reprezintă y_i vs. x_i . Pentru modele neliniare, se poate încerca o transformare a variabilelor X și Y astfel încât modelul liniar pentru variabilele transformate să fie unul aplicabil (nu merge întotdeauna).
- Se poate întâmpla ca reziduurile $\hat{\varepsilon}_i$ să prezinte o dependență clară de x_i (fapt ce poate fi observat dintr-o reprezentare $\hat{\varepsilon}_i$ vs. x_i), așadar aplicabilitatea modelului de regresie liniară este inoportună.
- Dacă reziduurile nu sunt normale, modelul liniar de regresie nu este oportun.
- Există posibilitatea ca datele observate $\{(x_i, y_i)\}_{i=1,2,\dots,n}$ să conțină valori aberante. Este important de a înțelege aceste valori și, în caz că nu sunt semnificative, pot fi șterse din setul de date care este supus analizei de regresie.
- În multe cazuri, o singură variabilă predictor X nu poate explica o singură variabilă Y , cazuri în care se apelează la o regresie multiplă (se iau în considerare și alte variabile predictor).

3 Regresie liniară multiplă

Regresia multiplă ia în considerare cel puțin doi predictorii pentru a determina valorile unei variabile de interes Y . În continuare, vom considera cazul unui număr de k predictorii independenți, notați $X_1, X_2, \dots, X_k$. Pentru o variabilă de interes Y se dorește a determina o relație liniară (un hiperplan) de forma

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon, \quad (29)$$

unde β_j ($j = 1, 2, \dots, n$) sunt niște constante reale, iar $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ este eroarea predicției. Pentru un j fixat, coeficientul β_j reprezintă variația în variabila Y rezultată în urma variației predictorului X_j cu o unitate, în timp ce ceilalți predictorii sunt menținuți la valori fixate. Coeficientul β_0 se numește *intercept* pentru hiperplanul de regresie

Pentru a determina oportunitatea unei astfel de legături, se culeg date relativ la aceste variabile. Fie aceste observații $\{(x_{1i}, x_{2i}, \dots, x_{ki}, y_i)\}_{i \in \{1, 2, \dots, n\}}$, unde x_{ji} denotă observația de rang i pentru variabila X_j , pentru fiecare $j = 1, 2, \dots, k$ și $i = 1, 2, \dots, n$. Pentru un model bun de regresie este necesar un număr suficient de mare de observații. Volumul de observații n ar trebui să fie mai mare (uneori mult mai mare) decât numărul de parametri ce urmează a fi estimați ($k + 2$ parametri, β_i și σ). Pe baza acestor date se poate aproxima suprafața

de regresie liniară (dacă există, prin metoda celor mai mici pătrate sau prin metoda verosimilității maxime), în scopul de a determina un estimator de forma:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k, \quad (30)$$

unde $\hat{\beta}_j$ ($j = 1, 2, \dots, n$) sunt estimatori pentru parametri reali β_j .

3.1 Metoda celor mai mici pătrate pentru estimarea parametrilor

Presupunem că avem la dispoziție $n > k$ observații dintr-o colectivitate statistică și notăm cu x_{ij} observația cu indicele i (sau nivelul variabilei x_j). De regulă, datele unei probleme de regresie multiplă se înregistrează într-un tabel de forma:

y	x_1	x_2	$\dots$	x_k
y_1	x_{11}	x_{12}	$\dots$	x_{1k}
y_2	x_{21}	x_{22}	$\dots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\dots$	x_{nk}

Fiecare observație $(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)$, $i = 1, 2, \dots, n$ și $n > k$ verifică ecuația

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \varepsilon_i.$$

Funcția celor mai mici pătrate se definește ca fiind:

$$\mathcal{L} = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2,$$

scopul fiind acela de a rezolva problema $\min_{\beta_0, \dots, \beta_k} \mathcal{L}$. Punctele critice se obțin prin rezolvarea sistemului:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \beta_0} \Big|_{\hat{\beta}_0, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij} \right) \\ \frac{\partial \mathcal{L}}{\partial \beta_j} \Big|_{\hat{\beta}_0, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij} \right) x_{ij}, \quad j = 1, 2, \dots, k \end{cases} \quad (31)$$

Rezolvăm sistemul (31) și obținem *ecuațiile normale ale celor mai mici pătrate*:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n x_{ik} = \sum_{i=1}^n y_i \\ \hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1} x_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n x_{i1} x_{ik} = \sum_{i=1}^n x_{i1} y_i \\ \vdots \\ \hat{\beta}_0 \sum_{i=1}^n x_{ik} + \hat{\beta}_1 \sum_{i=1}^n x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{ik} x_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n x_{ik}^2 = \sum_{i=1}^n x_{ik} y_i \end{cases} \quad (32)$$

Sistemul (32) are $p = k + 1$ ecuații normale, câte una pentru fiecare coeficient necunoscut de regresie. Soluțiile acestor ecuații vor reprezenta estimațiile estimatorilor, pe care îi vom nota tot cu simbolurile $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$. Sistemul de ecuații liniare anterior se poate rezolva prin metode algebrice sau numeric.

Regresia liniară multiplă suportă și o **abordare matriceală**. Presupunem că există k variabile regresionale și n observații, $\{(x_{1i}, x_{2i}, \dots, x_{ki}, y_i)\}_{i \in \{1, 2, \dots, n\}}$, iar relațiile de regresie sunt:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

Scrierea echivalentă matriceală este:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

unde

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Scopul este acela de a determina vectorul estimațiilor celor mai mici pătrate $\hat{\boldsymbol{\beta}}$ care minimizează

$$\mathcal{L} = \sum_{i=1}^n \varepsilon_i^2 = \boldsymbol{\varepsilon}^t \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad \text{adică rezolvăm ecuația matriceală} \quad \frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = \mathbf{0} \text{ (vectorul nul).}$$

Ecuația matriceală ce trebuie rezolvată este

$$\mathbf{X}^t \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^t \mathbf{y}, \quad \text{cu soluția} \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y}.$$

Există $p = k + 1$ ecuații normale, cu p necunoscute, reprezentate de estimațiile parametrilor. Matricea $\mathbf{X}^t \mathbf{X}$ este presupusă a fi nesingulară, iar rezolvarea sistemului se realizează în practică pe calculator.

Similar cu cazul regresiei liniare simple, vom utiliza terminologia următoare:

- Hipersuprafața

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$$

este *aproximarea suprafeței de regresie*. Ea este suprafața care se apropie cel mai mult (în sensul metodei celor mai mici pătrate) de datele experimentale. Această suprafață este o aproximare a suprafeței de regresie;

- valorile y_i se numesc *valori observate*, iar valorile $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_k x_{ki}$, unde $i = 1, 2, \dots, n$, se numesc *valori prezise* ($i = 1, 2, \dots, n$). Forma echivalentă matriceală este dată de $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.
- valorile $\hat{\varepsilon}_i = y_i - \hat{y}_i$ se numesc *reziduuri*. Un reziduu măsoară deviația unui punct observat de la valoarea prezisă de estimarea hipersuprafeței de regresie. Se presupune că aceste reziduuri sunt independente între ele și sunt repartizate $\mathcal{N}(0, \sigma^2)$; forma matriceală a reziduurilor este reprezentată prin formula $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.
- la fel ca și în cazul regresiei liniare simple, este importantă evaluarea dispersiei σ^2 a termenilor ce reprezintă erorile. Definim *suma pătratelor erorilor*,

$$\sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \dots - \hat{\beta}_k x_{ki} \right)^2,$$

se notează tot prin SSE;

- dacă în cazul regresiei liniare simple se împărțea SSE la $(n - 2)$, ce reprezenta numărul gradelor de libertate, iar 2 era de fapt numărul parametrilor din ecuația de regresie, acum definim *eroarea medie pătratică sau reziduală* prin $\text{MSE} = \frac{\text{SSE}}{n - p} = \frac{\text{SSE}}{n - k - 1}$. Statistica $\text{MSE} = \hat{\sigma}$ este un estimator pentru dispersia erorilor, σ^2 .
- rădăcina pătrată a MSE este $\hat{\sigma}$ este și se numește *eroarea standard a regresiei*;
- se poate demonstra, prin exact aceeași metodă utilizată în cazul regresiei liniare simple pentru distribuția $\frac{\text{SSE}}{\sigma^2} = (n - 2) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n - 2, 1) = \chi^2(n - 2)$, că

$$\frac{\text{SSE}}{\sigma^2} = (n - k - 1) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n - k - 1, 1) = \chi^2(n - k - 1).$$

Acest rezultat fundamental asigură, la fel ca și la regresia liniară, o metodă de obținere a unui interval de încredere pentru valoarea reală a lui σ^2 .

În ceea ce privește analiza unor proprietăți ale estimațiilor celor mai mici pătrate, presupunem, ca de obicei, că termenii ce dau erorile $\varepsilon_1, \dots, \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ sunt variabile aleatoare independente stochastice. În aceste condiții, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ sunt estimatori nedeplasați ai coeficienților de regresie $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. Într-adevăr, folosind scrierea matriceală avem, deoarece $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$:

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbb{E}((\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y}) = \mathbb{E}((\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})) = \mathbb{E}((\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \boldsymbol{\varepsilon}) = \boldsymbol{\beta}.$$

Matricea $\text{cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}$, de dimensiune $p \times p$, reprezintă matricea de covarianță a coeficienților de regresie $\hat{\beta}$. Elementele diagonale ale acestei matrice sunt dispersiile estimatorilor $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$, iar restul elementelor sunt covarianțele efective ale perechilor de câte doi estimatori cu indici distincți. Spre exemplificare, dacă $k = 2$, atunci

$$\mathbf{C} = (\mathbf{X}^t \mathbf{X})^{-1} = \begin{pmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{pmatrix} \quad \text{est simetrică deoarece} \quad (\mathbf{X}^t \mathbf{X})^{-1} \quad \text{are această proprietate.}$$

Mai precis, $D^2(\hat{\beta}_j) = \sigma^2 C_{jj}$, $j \in \{0, 1, 2\}$ și $\text{cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}$, pentru $i \neq j$.

3.2 Teste statistice regresia liniară multiplă

Există diverse teste de testare a ipotezelor în cazul regresiei liniare multiple pentru stabilirea acurateții modelului. La fel ca și în cazul regresiei liniare simple, trebuie impuse ipoteze de normalitate și independență stochastică pentru erorile de măsurare ε_i .

3.2.1 Test statistic pentru semnificația regresiei

Acest test statistic constă în determinarea existenței unei relații liniare între variabila de răspuns y și o mulțime de variabile independente $x_1, x_2, \dots, x_k$. Ipotezele formulate sunt:

$$(H_0) : \beta_1 = \beta_2 = \dots = \beta_k = 0, \quad \text{versus} \quad (H_1) : \beta_j \neq 0, \text{ pentru cel puțin un indice } j.$$

Respingerea ipotezei nule implică faptul că cel puțin une dintre variabilele stimul $x_1, x_2, \dots, x_k$ contribuie semnificativ la model. Acceptarea ipotezei alternative indică faptul că modelul de regresie liniară multiplă este semnificativ. Statistica de test în acest caz este:

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} = \frac{\text{SSR}/(k \cdot \text{SST})}{\text{SSE}/((n - k - 1) \cdot \text{SST})} = \frac{\text{SSR}/k}{\text{SSE}/(n - k - 1)} = \frac{\text{MSR}}{\text{MSE}}, \quad \text{unde}$$

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad \text{SSE} = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \dots - \hat{\beta}_k x_{ki} \right)^2,$$

$$\text{SSR} = \text{SST} - \text{SSE}, \quad \text{MSR} = \frac{\text{SSR}}{k}.$$

Dacă ipoteza (H_0) este adevărată, atunci

$$\frac{\text{SSR}}{\sigma^2} \sim \chi^2(k), \quad \text{iar} \quad \frac{\text{SSE}}{\sigma^2} = (n - k - 1) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n - k - 1),$$

iar variabilele SSR și SSE sunt independente stochastic. Din acest motiv, pentru statistica de test pentru verificarea ipotezei nule avem repartiția Snedecor-Fisher:

$$F = \frac{\frac{\text{SSR}}{\sigma^2} \cdot \frac{1}{k}}{\frac{\text{SSE}}{\sigma^2} \cdot \frac{1}{n - k - 1}} = \frac{\text{MSR}}{\text{MSE}} \sim F(k, n - k - 1) \quad (33)$$

Justificarea pentru afirmația (33) este oferită de următorul rezultat.

Lema 3.1 Dacă $X \sim \chi^2(m, 1) = \chi^2(m)$ și $Y \sim \chi^2(n, 1) = \chi^2(n)$, unde $m, n \in \mathbb{N}^*$, sunt două variabile aleatoare independente, atunci distribuția

$$F = \frac{X/m}{Y/n} \sim F(m, n).$$

Demonstrație. Să calculăm mai întâi funcția de repartiție a v.a. X/Y , $F_{X/Y} : \mathbb{R} \rightarrow [0, 1]$,

$$F_{\frac{X}{Y}}(u) = \iint_{\mathcal{D}} f_X(x) f_Y(y) dx dy,$$

unde $\mathcal{D} = \{(x, y) \in \mathbb{R}^2 : x \geq 0, y > 0, x/y \leq u\}$. Notăm

$$C_1 = \frac{1}{2^{\frac{n}{2}} \sigma^n \Gamma(n/2)}, \quad C_2 = \frac{1}{2^{\frac{m}{2}} \sigma^m \Gamma(m/2)}.$$

Explicitând și domeniul $\mathcal{D}$ obținem $\mathcal{D} = \{(x, y) : x \geq 0, y \geq x/u\}$, deci

$$F_{\frac{X}{Y}}(u) = C_1 C_2 \int_0^{+\infty} \left(\int_{x/u}^{+\infty} x^{\frac{n}{2}-1} y^{\frac{m}{2}-1} \exp\left(-\frac{x+y}{2\sigma^2}\right) dy \right) dx.$$

Derivând funcția de repartiție de mai sus, obținem:

$$f_{\frac{X}{Y}}(u) = C_1 C_2 \int_0^{+\infty} (-) \left(\frac{x}{u}\right)' x^{\frac{n}{2}-1} \left(\frac{x}{u}\right)^{\frac{m}{2}-1} \exp\left(-\frac{x+\frac{x}{u}}{2\sigma^2}\right) dx = \frac{C_1 C_2}{u^{1+m/2}} \int_0^{+\infty} x^{\frac{n+m}{2}-1} \exp\left(-\frac{x}{2\sigma^2} \left(1 + \frac{1}{u}\right)\right) dx.$$

În această integrală facem substituția $\frac{x}{2\sigma^2} \left(1 + \frac{1}{u}\right) = x'$, cu $dx = \frac{2\sigma^2}{1+1/u} dx'$, deci

$$\begin{aligned} f_{\frac{X}{Y}}(u) &= \frac{C_1 C_2}{u^{1+m/2}} \int_0^{+\infty} \left(\frac{2\sigma^2}{1+1/u} x' \right)^{\frac{n+m}{2}-1} e^{-x'} \frac{2\sigma^2}{1+1/u} dx' = \frac{C_1 C_2}{u^{1+m/2}} \frac{(2\sigma^2)^{\frac{n+m}{2}}}{(1+u)^{\frac{n+m}{2}} u^{-\frac{n+m}{2}}} \int_0^{+\infty} x^{\frac{n+m}{2}-1} e^{-x} dx \\ &= C_1 C_2 \frac{u^{n/2-1}}{(1+u)^{\frac{n+m}{2}}} (2\sigma^2)^{\frac{n+m}{2}} \Gamma\left(\frac{n+m}{2}\right). \end{aligned}$$

Înlocuind cantitățile C_1 și C_2 în formula de mai sus obținem:

$$f_{\frac{X}{Y}}(u) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{m}{2}\right)} \frac{u^{n/2-1}}{(1+u)^{\frac{n+m}{2}}}.$$

Se notează acum $V = \frac{m}{n} U$ și determinăm densitatea variabilei aleatoare V :

$$f_V(v) = f_U\left(\frac{n}{m}v\right) \frac{n}{m}.$$

Folosind și legătura dintre funcțiile Gamma și Beta, concluzionăm că

$$\begin{aligned} f_V(v) &= \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{m}{2}\right)} \frac{\left(\frac{n}{m}v\right)^{n/2-1} \frac{n}{m}}{\left(1 + \frac{n}{m}v\right)^{\frac{n+m}{2}}} = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{m}{2}\right)} \left(\frac{n}{m}\right)^{n/2} v^{\frac{n}{2}-1} \left(1 + \frac{n}{m}v\right)^{-\frac{n+m}{2}} \\ &= \frac{\left(\frac{n}{m}\right)^{n/2}}{\beta\left(\frac{n}{2}, \frac{m}{2}\right)} v^{\frac{n}{2}-1} \left(1 + \frac{n}{m}v\right)^{-\frac{n+m}{2}}, \quad v \geq 0. \end{aligned}$$

Relația din enunț se poate obține și calculând direct densitatea de repartiție $f_{X/Y}$ cu ajutorul formulei cântului a două variabile aleatoare independente, repartizate $\chi^2(m)$, respectiv $\chi^2(n)$.

Revenim acum la verificarea ipotezelor testului statistic F . Determinăm din tabelul cuantilelor repartiției Snedecor-Fisher valoarea critică $f_{\alpha, k, n-k-1}$. Pentru o valoare observată $f_0 \geq f_{\alpha, k, n-k-1}$, respingem ipoteza nulă. În caz contrar, o acceptăm.

O formulă de calcul pentru SSE este dată de:

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \varepsilon_i^2 = \mathbf{e}^t \mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\beta})^t (\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{y}^t \mathbf{y} - \hat{\beta}^t \mathbf{X}^t \mathbf{y}. \quad (34)$$

Pentru calcularea statisticii SSR, deoarece $\text{SST} = \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2 / n$, atunci (34) devine:

$$\text{SSE} = \mathbf{y}^t \mathbf{y} - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 - \left(\hat{\beta}^t \mathbf{X}^t \mathbf{y} - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right) = \text{SST} - \text{SSR}, \quad \text{adică} \quad \text{SSR} = \hat{\beta}^t \mathbf{X}^t \mathbf{y} - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2.$$

O măsură a semnificației regresiei multiple poate fi și *coeficientul de determinare multiplă* R^2 , definit prin:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}}.$$

Valorile lui R^2 sunt între 0 și 1 și, pentru a avea un model destul de bun, ar fi un coeficient de determinare aproape de 1. Mulți statisticieni preferă utilizarea unui alt instrument de măsură pentru verificarea validității regresiei multiple utilizate, și anume *coeficientul de determinare multiplu ajustat*, $\text{adj } R^2$. Acesta este definit de formula:

$$\text{adj } R^2 = 1 - \frac{\text{MSE}}{\text{MST}} = \frac{\text{SSE} / (n - k - 1)}{\text{SST} / (n - 1)},$$

unde n este volumul datelor și k este numărul de variabile independente în modelul liniar (fără a considera constanta). Deoarece $\text{SSE} / (n - k - 1)$ este eroarea medie pătratică (reziduală), iar $\text{SST} / (n - 1)$ este o constantă, atunci $\text{adj } R^2$ va crește atunci când o variabilă este adăugată la model doar dacă noua variabilă reduce pătratul erorii.

3.2.2 Test statistic pentru semnificația individuală a coeficientului β_i din regresia multiplă

De multe ori, o problemă care poate apărea este verificarea ipotezelor pentru coeficienții individuali ai regresiei. Adăugarea unei variabile la modelul de regresie face ca suma pătratelor să crească și ca eroarea sumei pătratelor să descrească (acesta este motivul pentru care R^2 crește întotdeauna când o nouă variabilă este adăugată). Trebuie decis dacă creșterea sumei este suficient de mare pentru a justifica utilizarea unei noi variabile în model.

Mai jos prezentăm testul ce verifică dacă coeficientul de regresie β_i (i este un indice fixat între 1 și k) este semnificativ, la un nivel de semnificație α .

Testăm

$$(H_0) : \beta_i = 0, \quad \text{versus} \quad (H_1) : \beta_i \neq 0.$$

Dacă ipoteza nulă nu este respinsă, putem elimina variabila independentă x_i din model. Considerăm statistica test similară, atât ca formulă, cât și ca repartiție, celei din cazul regresiei liniare simple:

$$T = \frac{\hat{\beta}_i}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_{ki} - \bar{x}_k)^2}}} \sim t(n - k - 1).$$

Etapele testului sunt deja cunoscute. Calculăm valoarea observată a statisticii T , valoare pe care o vom nota cu t_0 . Identificăm cuantila de ordin $\alpha/2$ pentru repartiția $t(n - k - 1)$, $t_{\alpha/2, n-k-1}$. Decizia finală este:

$$\begin{cases} \text{Dacă } |t_0| < t_{\alpha/2, n-k-1}, & \text{atunci ipoteza nulă } (H_0) \text{ este acceptată.} \\ \text{Dacă } |t_0| \geq t_{\alpha/2, n-k-1}, & \text{atunci ipoteza alternativă bilaterală } (H_1) \text{ este acceptată.} \end{cases}$$

Intervalul de acceptare este $(-t_{\alpha/2, n-k-1}, t_{\alpha/2, n-k-1})$, iar $|t| \geq t_{\alpha/2, n-k-1}$ este regiunea critică.

Validitatea modelului de regresie liniară multiplă. Ca și în secțiunea precedentă, pentru a verifica dacă modelul de regresie liniară multiplă este unul valid, se pot folosi mai multe metode, printre care:

- coeficientul de determinare R^2 și coeficientul ajustat de determinare, $\text{adj } R^2$.
- grafic ce indică normalitatea reziduurilor $\hat{\varepsilon}_i$: Acest grafic reprezintă probabilitățile de normalitate ale erorilor versus cuantilele repartiției $\mathcal{N}(0, 1)$. Dacă modelul este valid, atunci valorile reprezentate în figură vor fi cât mai apropiate de prima bisectoare.
- $\hat{y}_i$ versus y_i : Dacă modelul este valid, atunci valorile reprezentate în figură vor fi cât mai apropiate de prima bisectoare.
- $\hat{y}_i$ versus y_i : Dacă modelul este valid, atunci valorile reprezentate în figură nu ar avea nicio tendință clară.

Reducerea unor tipuri de regresie la regresie liniară multiplă.

O *regresie polinomială simplă de ordin k* este de forma:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k + \varepsilon. \quad (35)$$

Ea poate fi redusă la una simplă multiplă dacă notăm $X_1 = X$, $X_2 = X^2, \dots, X_k = X^k$. Interpretările coeficienților β_j în cazul unei regresii polinomiale nu se mai potrivesc cu cele ale coeficienților din cazul regresiei multiple, fiind greu de determinat.

O *regresie multiplă cu interacțiuni de ordin 2* este de forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2 + \varepsilon.$$

Ea poate fi redusă la una simplă multiplă dacă notăm $X_1 X_2 = X_3$, $X_1^2 = X_4$, $X_2^2 = X_5$.

4 Regresie logistică

De multe ori este nevoie de a obține clasificări ale datelor în funcție de valorile observate pentru o anumită variabilă răspuns. Spre exemplu:

- Preziceri ale șanselor unei anumite tumori să devină malignă, sau să rămână benignă;
- Predicții pentru următorul președinte, bazate pe diverse măsurători politice, sociale sau istorice;
- Clasificarea unor plante în funcție de anumite caracteristici.

Astfel, variabila răspuns poate lua un număr discret de valori (categorii). O clasificare este gruparea datelor într-un număr discret de categorii, prin atribuirea unei valori răspuns corespunzătoare. O problemă statistică de clasificare constă în prezicerea valorii variabilei răspuns nominale pe baza unor observații asupra unui set de variabile independente. Pentru început, să presupunem că variabila răspuns Y poate lua doar două posibile valori 0 și 1, adică $Y \sim \mathcal{B}(1, p)$. Spre exemplu, variabila Y reprezintă decizia ca o anumită mașină să aibă nevoie de revizie. Valoarea $Y = 0$ reprezintă NU și $Y = 1$ reprezintă DA. Astfel, p reprezintă probabilitatea ca mașina să necesite revizie. În general, această probabilitate depinde de mai mulți factori, spre exemplu: X_1 = numărul de km parcursi, X_2 = vechimea mașinii, X_3 = timpul scurs de la ultima revizie. Pentru simplitate, ne limităm doar la acești trei factori. Însă, este clar, probabilitatea p nu poate depinde liniar de acești factori, și nici altă formă de regresie studiată până acum nu poate fi aplicată. Motivul este simplu: dacă am presupune că

$$p = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon,$$

atunci există posibilitatea că membrul din dreapta să nu aparțină intervalului $[0, 1]$. În consecință, este nevoie de o nouă dependență a probabilității p de acești predictorii. Un exemplu potrivit este cel dat de funcția logit, i.e., $f(x) = e^{a+bx} / (1 + e^{a+bx})$. Vom considera următorul model de regresie:

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}}, \quad (36)$$

numit regresie logistica multiplă. Aici, $p = p(\mathbf{X}) = P(Y = 1|\mathbf{X})$ este probabilitatea condiționată ca variabila răspuns Y să ia valoarea 1, știind că am observat datele $\mathbf{X} = (X_1, X_2, \dots, X_n)$, iar $1-p = 1-p(\mathbf{X}) = P(Y = 0|\mathbf{X})$ este probabilitatea ca variabila răspuns Y să ia valoarea 0, știind că am observat datele $\mathbf{X}$. În cazul unei singure variabile independente, X , regresia se va numi regresie logistica. Din relația (36), obținem:

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}.$$

Expresia $p/(1-p)$ se numește *cota de realizare a evenimentului*. Prin logaritmare, găsim că

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3. \quad (37)$$

Astfel, modelul poate fi privit ca un model de regresie liniară multiplă. Coeficienții β_i sunt ușor de interpretat. Spre exemplu, dacă în relația (37) variabila X_1 crește cu o unitate, menținând celelalte două variabile fixe, atunci logaritmul cotei se va modifica cu cantitatea β_1 . Dacă facem același lucru în relația (36), atunci o creștere cu o unitate a variabilei X_1 , ținând celelalte variabile fixe, va conduce la o modificare cu e^{β_1} a cotei pentru care $Y = 1$. Pe baza observațiilor se pot determina estimatori pentru parametri, iar pe baza acestor estimatori se estimează probabilitatea p . O metodă de estimare a parametrilor β_i este metoda verosimilității maxime. Generalizarea pentru k variabile independente, poate fi realizată.

4.1 Metoda verosimilității maxime

Dat fiind un eșantion $(x_{1i}, x_{2i}, x_{3i}, y_i)$, $i = 1, 2, \dots, n$, notăm cu $p(\mathbf{x}_i) = P(y_i = 1|\mathbf{x}_i)$. Presupunem că

$$\ln \left(\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)} \right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, \quad i = 1, 2, \dots, n,$$

echivalent cu

$$p(\mathbf{x}_i) = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}}}, \quad i = 1, 2, \dots, n.$$

Deoarece $Y \sim \mathcal{B}(1, p(\mathbf{x}))$, funcția de probabilitate este

$$f(y; p(\mathbf{x})) = p(\mathbf{x})^y \cdot (1 - p(\mathbf{x}))^{1-y}, \quad \text{unde } y \in \{0, 1\}.$$

Astfel, presupunând independența datelor observate, funcția de verosimilitate corespunzătoare eșantionului va fi

$$\mathcal{L}(\beta) = \prod_{i=1}^n p(\mathbf{x})^{y_i} (1 - p(\mathbf{x}))^{1-y_i}.$$

Estimatorii parametrilor $\beta_1, \beta_2, \beta_3$ obținuți prin metoda verosimilității maxime sunt valorile pentru care se obține maximumul acestei funcții. Maximizarea aceste funcții este echivalentă cu maximizarea logaritmului său. Astfel, estimatorii $\hat{\beta}$ sunt aleși astfel încât maximizează funcția

$$\begin{aligned} l(\beta) &= \sum_{i=1}^n (y_i \ln p(\mathbf{x}_i) + (1 - y_i) \ln (1 - p(\mathbf{x}_i))) = \sum_{i=1}^n \left(y_i \ln \left(\frac{p(\mathbf{x}_i)}{1 - p(\mathbf{x}_i)} \right) + \ln (1 - p(\mathbf{x}_i)) \right) \\ &= \sum_{i=1}^n (y_i (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}) - \ln (1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}})). \end{aligned}$$

Dacă am căuta punctele critice ale acestei funcții prin anularea derivatelor parțiale în raport cu $\beta_0, \beta_1, \beta_2$ și β_3 nu vom găsi soluții explicite. De aceea, pentru a maximiza această funcție se folosesc metode iterative numerice. Astfel, determinarea estimatorilor pentru parametrii de regresie logistică este o muncă mult mai dificilă decât în cazul regresiei liniare multiple, ce necesită implementarea de metode numerice potrivite pe un calculator. Putem determina chiar și intervale de încredere pentru parametrii de regresie. După determinarea estimatorilor parametrilor de regresie $\hat{\beta}$, următorul pas este prezicerea rezultatului pentru o nouă dată de intrare $\mathbf{x} = (x_1, x_2, x_3)$. Vom avea:

$$\widehat{p(\mathbf{x})} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3}}.$$

Pe baza acestei estimări, putem prezice clasa asociată astfel:

$$\widehat{y(\mathbf{x})} = \begin{cases} 1 & , \text{dacă } \widehat{p(\mathbf{x})} \geq 0.5, \\ 0 & , \text{dacă } \widehat{p(\mathbf{x})} < 0.5, \end{cases}$$

sau, în mod echivalent,

$$\widehat{y(\mathbf{x})} = \begin{cases} 1 & , \text{dacă } \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 \geq 0, \\ 0 & , \text{dacă } \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 < 0. \end{cases}$$

Mulțimea

$$\left\{ \mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3 : \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = 0 \right\}$$

se numește frontiera de decizie între clasele 0 și 1.

4.2 Regresie logistică multinomială

Presupunem că variabila nominală Y poate lua un set de $\kappa \geq 3$ valori distincte (sau valorile lui Y pot fi grupate în κ clase disjuncte), $1, 2, \dots, \kappa$. În mod similar, presupunând că logaritmul cotelor urmează un model linear de regresie multiplă, putem scrie

$$\begin{aligned} \ln \left(\frac{P(Y=1|\mathbf{X})}{P(Y=\kappa|\mathbf{X})} \right) &= \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3 = \beta_1^T X; \\ \ln \left(\frac{P(Y=2|\mathbf{X})}{P(Y=\kappa|\mathbf{X})} \right) &= \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3 = \beta_2^T X; \\ &\vdots \\ \ln \left(\frac{P(Y=\kappa-1|\mathbf{X})}{P(Y=\kappa|\mathbf{X})} \right) &= \beta_{\kappa-1,0} + \beta_{\kappa-1,1}X_1 + \beta_{\kappa-1,2}X_2 + \beta_{\kappa-1,3}X_3 = \beta_{\kappa-1}^T X. \end{aligned}$$

De aici, obținem că

$$\begin{aligned}
 P(Y = 1|\mathbf{X}) &= \frac{e^{\beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3}}{1 + \sum_{j=1}^{\kappa-1} e^{\beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \beta_{j3}X_3}} = \frac{e^{\beta_1^T X}}{1 + \sum_{j=1}^{\kappa-1} e^{\beta_j^T X}}; \\
 P(Y = 2|\mathbf{X}) &= \frac{e^{\beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3}}{1 + \sum_{j=1}^{\kappa-1} e^{\beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \beta_{j3}X_3}} = \frac{e^{\beta_2^T X}}{1 + \sum_{j=1}^{\kappa-1} e^{\beta_j^T X}}; \\
 &\vdots \\
 P(Y = \kappa - 1|\mathbf{X}) &= \frac{e^{\beta_{\kappa-1,0} + \beta_{\kappa-1,1}X_1 + \beta_{\kappa-1,2}X_2 + \beta_{\kappa-1,3}X_3}}{1 + \sum_{j=1}^{\kappa-1} e^{\beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \beta_{j3}X_3}} = \frac{e^{\beta_{\kappa-1}^T X}}{1 + \sum_{j=1}^{\kappa-1} e^{\beta_j^T X}}; \\
 P(Y = \kappa|\mathbf{X}) &= \frac{1}{1 + \sum_{j=1}^{\kappa-1} e^{\beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \beta_{j3}X_3}} = \frac{1}{1 + \sum_{j=1}^{\kappa-1} e^{\beta_j^T X}}.
 \end{aligned}$$

Funcția pentru care $f(\mathbf{z})_j = \frac{e^{z_j}}{1 + \sum_{k=1}^{\kappa-1} e^{z_k}}, j = 1, 2, \dots, \kappa - 1$ se numește funcția *softmax*. Este considerată a fi generalizarea funcției logit. Interpretarea coeficienților este similară cazului binomial. Estimările coeficienților se pot obține prin metoda verosimilității maxime, folosind metode numerice pe un computer. După determinarea estimatorilor parametrilor de regresie $\hat{\beta}$, următorul pas este prezicerea rezultatului pentru o nouă dată de intrare $\mathbf{x} = (x_1, x_2, x_3)$. Vom avea:

$$\widehat{p_1(\mathbf{x})} = \frac{e^{\hat{\beta}_1^T X}}{1 + \sum_{j=1}^{\kappa-1} e^{\hat{\beta}_j^T X}}, \quad \widehat{p_2(\mathbf{x})} = \frac{e^{\hat{\beta}_2^T X}}{1 + \sum_{j=1}^{\kappa-1} e^{\hat{\beta}_j^T X}}, \dots, \quad \widehat{p_\kappa(\mathbf{x})} = \frac{1}{1 + \sum_{j=1}^{\kappa-1} e^{\hat{\beta}_j^T X}}.$$

Pentru predicția claselor, se alege clasa cu acel indice j pentru care $j = \arg \max_{j \in \{1, \dots, \kappa\}} \widehat{p_j(\mathbf{x})}$. Aceasta înseamnă că dacă pentru datele observate găsim o predicție maximă, atunci variabila răspuns va fi încadrată în clasa respectivă.

5 Considerente asupra regresiei neliniare

Regresia liniară oferă un cadru de lucru eficient în diverse domenii din viața cotidiană. Cu toate acestea, ea nu este potrivită oricărei stuații întâlnite deoarece, de multe ori, nu se poate stabili o relație de legătură liniară între variabilele predictor și cele răspuns. Soluția problemei este oferită de *regresia neliniară*. Atunci când aplicăm metoda celor mai mici pătrate acestor modele, ecuațiile normale rezultate nu mai sunt de tip liniar și, în general, dificil de rezolvat. Abordarea uzuală constă în minimizarea directă a erorii reziduale prin intermediul unor proceduri iterative. Prezentăm în cele ce urmează o descriere succintă a tehnicii folosite în regresia neliniară.

Putem rescrie modelul de regresie liniară sub forma generală

$$Y = \mathbf{x}^t \beta + \varepsilon = f(\mathbf{x}, \beta) + \varepsilon$$

Cum $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, avem $\mathbb{E}(Y) = f(\mathbf{x}, \beta)$. Funcția $f(\mathbf{x}, \beta)$ va fi numită valoarea așteptată a modelului. Orice model pentru care această funcție este neliniară în parametri necunoscuți se va numi un *model de regresie neliniară* (de exemplu, $Y = \theta_1 e^{\theta_2 x} + \varepsilon$ este neliniară în θ_1 și θ_2).

În general, un model de regresie neliniară se va scrie sub forma generală

$$Y_i = f(\mathbf{x}_i, \theta) + \varepsilon_i, \quad \theta \in \mathbb{R}^p, \quad i = 1, 2, \dots, n.$$

Erorile în acest caz sunt presupune a avea aceeași distribuție ca și în cazul liniar. Prin $\mathbf{x}_i$ înțelegem vectorul $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ik})^t$, pentru fiecare $i = 1, 2, \dots, n$. Funcția celor mai mici pătrate asociate modelului este

$$\Lambda : \mathbb{R}^p \rightarrow \mathbb{R}_+, \quad \Lambda(\theta) = \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \theta))^2.$$

Pentru determinarea punctelor critice trebuie să rezolvăm sistemul de p ecuații normale asociate modelului de regresie neliniară:

$$\sum_{i=1}^n (y_i - f(\mathbf{x}_i, \theta)) \frac{\partial f(\mathbf{x}_i, \theta)}{\partial \theta_j} = 0, \quad \text{pentru } i = 1, 2, \dots, n.$$

Caracterul neliniar al funcțiilor implicate poate face ca rezolvarea sistemului și determinarea vectorului punctelor critice $\hat{\theta}$ să fie foarte greu de realizat. Pentru a exemplifica aceasta, să considerăm un simplu exemplu, în care $Y = \theta_1 e^{\theta_2 x} + \varepsilon$. Sistemul de rezolvat este:

$$\begin{cases} \sum_{i=1}^n (y_i - \hat{\theta}_1 e^{\hat{\theta}_2 x_i}) e^{\hat{\theta}_2 x_i} = 0 \\ \sum_{i=1}^n (y_i - \hat{\theta}_1 e^{\hat{\theta}_2 x_i}) \hat{\theta}_1 x_i e^{\hat{\theta}_2 x_i} = 0 \end{cases} \iff \begin{cases} \sum_{i=1}^n y_i e^{\hat{\theta}_2 x_i} - \hat{\theta}_1 \sum_{i=1}^n e^{2\hat{\theta}_2 x_i} = 0 \\ \sum_{i=1}^n y_i x_i e^{\hat{\theta}_2 x_i} - \hat{\theta}_1 \sum_{i=1}^n x_i e^{2\hat{\theta}_2 x_i} = 0 \end{cases}$$

Nu putem determina, prin metode clasice soluțiile, sistemului anterior. Din acest motiv, trebuie utilizate metode iterative de aproximare a soluțiilor. De asemenea, pot exista soluții multiple pentru sistemul obținut.

Instrumentele folosite. Metoda folosită în abordarea acestui tip de regresie o constituie **liniarizarea funcției neliniare**, urmată de **metode iterative de tip Gauss-Newton** pentru estimarea parametrilor. Liniarizarea se realizează prin dezvoltarea în serie Taylor a funcției $f(\mathbf{x}_i, \theta)$, în vecinătatea punctului critic $\theta'_0 = (\theta_{10}, \theta_{20}, \dots, \theta_{p0})^t$, dezvoltare din care păstrăm doar termenul liniar. Obținem:

$$f(\mathbf{x}_i, \theta) = f(\mathbf{x}_i, \theta_0) + \sum_{j=1}^p \left(\frac{\partial f(\mathbf{x}_i, \theta)}{\partial \theta_j} \right)_{\theta=\theta_0} \cdot (\theta_j - \theta_{j0}). \quad (38)$$

Notăm

$$f_i^0 = f(\mathbf{x}_i, \theta_0), \quad \beta_j^0 = \theta_j - \theta_{j0}, \quad Z_{ij}^0 = \left(\frac{\partial f(\mathbf{x}_i, \theta)}{\partial \theta_j} \right)_{\theta=\theta_0},$$

iar modelul de regresie neliniară se scrie sub forma (aproximantă) liniară, cu valoarea de start a parametrilor θ_0 :

$$y_i - f_i^0 = \sum_{j=1}^p \beta_j^0 Z_{ij}^0 + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

Ecuția (38) se scrie sub formă matriceală $\mathbf{y}_0 = \mathbf{Z}_0 \beta_0 + \varepsilon$, cu estimarea pentru β_0 :

$$\hat{\beta}_0 = (\mathbf{Z}_0^t \mathbf{Z}_0)^{-1} \mathbf{Z}_0^t \mathbf{y}_0 = (\mathbf{Z}_0^t \mathbf{Z}_0)^{-1} \mathbf{Z}_0^t (\mathbf{y} - \mathbf{f}_0).$$

Cum $\beta_0 = \theta - \theta_0$, definim $\hat{\theta}_1 = \hat{\beta}_0 + \theta_0$ ca fiind *estimările ajustate* ale lui θ . Termenul $\hat{\beta}_0$ mai este numit *vectorul incrementărilor*. Putem acum utiliza aceste estimări ajustate în (38) în locul lui θ_0 și obținem o nouă versiune a acestor estimări ajustate, notată cu $\hat{\theta}_2$. Algoritmul continuă în același mod. Iterația cu indicele k este

$$\begin{cases} \hat{\theta}_{k+1} = \hat{\theta}_k + \hat{\beta}_k = \hat{\theta}_k + (\mathbf{Z}_k^t \mathbf{Z}_k)^{-1} \mathbf{Z}_k^t (\mathbf{y} - \mathbf{f}_k), & \text{unde} \\ \mathbf{Z}_k = (Z_{ij}^k)_{ij}, \quad \mathbf{f}_k = (f_1^k, f_2^k, \dots, f_n^k)^t, \quad \hat{\theta}_k = (\theta_{1k}, \theta_{2k}, \dots, \theta_{pk})^t. \end{cases}$$

Algoritmul iterativ continuă până este atins un criteriu de convergență, precum

$$\left| \frac{\hat{\theta}_{j,k+1} - \hat{\theta}_{j,k}}{\hat{\theta}_{j,k}} \right| < \delta, \quad j = 1, 2, \dots, p,$$

unde δ este un prag suficient de mic (de exemplu, de ordinul lui 10^{-4}). Acești algoritmi sunt implementați în pachetele diverselor softuri matematice.

Bibliografie

- [1] Anderson, M., *A characterization of the multivariate normal distribution*, The Annals of Mathematical Statistics, vol. 42, no. 2, 824-827, 1971.

- [2] Benhamou, E.; Melot, V., *Seven proofs of the Pearson Chi-squared independence test and its graphical interpretation*, arXiv:1808.09171v3, 2018.
- [3] Berk, R., *Review 1922 of 'Invariance of Maximum Likelihood Estimators' by Peter W. Zehna*, *Mathematical Reviews*, 33, 342-343, 1967.
- [4] Devore, J; Berk, K., *Modern Mathematical Statistics with Applications*, 2nd Edition, Springer New York Dordrecht Heidelberg London, 2012.
- [5] Duret, R., *Probability: Theory and Examples*, 5th Edition, Cambridge Series in Statistical and Probabilistic Mathematics, 2014.
- [6] Gibbons Dickinson, J.; Chakraborti, S., *Nonparametric Statistical Inference*, Fourth Edition, Revised and Expanded, Marcel Dekker, INC., New York, Basel, 2003.
- [7] Kendall, M.G., *The Advanced Theory of Statistics*, Volume 1, Distribution Theory, London, Charles Griffin & Company, 1945 (Edition by Stuart, Alan, Ord, Keith, 2010).
- [8] Kendall, M.G.; Stuart, A., *The Advanced Theory of Statistics*, Volume 2, Inference and Relationships, Hafner Publishing Company, 1961 (Edition by Wiley, 2010).
- [9] Klenke, A., *Probability Theory: A Comprehensive Course*, 2nd Edition, Springer, 2014.
- [10] Kolmogorov, A. N., *Sulla Determinazione Empirica di Una Legge di Distribuzione*, *Giornale dell'Istituto Italiano degli Attuari*, 4. 83-91, 1933.
- [11] Montgomery, D; Runger, G, *Applied Statistics and Probability for Engineers*, 3rd Edition, John Wiley & Sons, Inc, 2003.
- [12] Owen, A, *Lectures on statistics*, Department of Statistics, Stanford University.
- [13] Stoleriu, I., *Statistică aplicată*, note de curs, 2019.
- [14] Wackerly, D.; Mendenhall, W.; Scheaffer, R., *Mathematical Statistics with Applications*, 7th Edition, Thomson Brooks/Cole, 2008.
- [15] Walck, C., *Handbook on Statistical distributions for experimentalists*, Particle Physics Group, University of Stockholm.
- [16] Watson, G.S., *Some recent results in chi-square goodness-of-fit tests*, *Biometrics*, 15, 440, 1959.