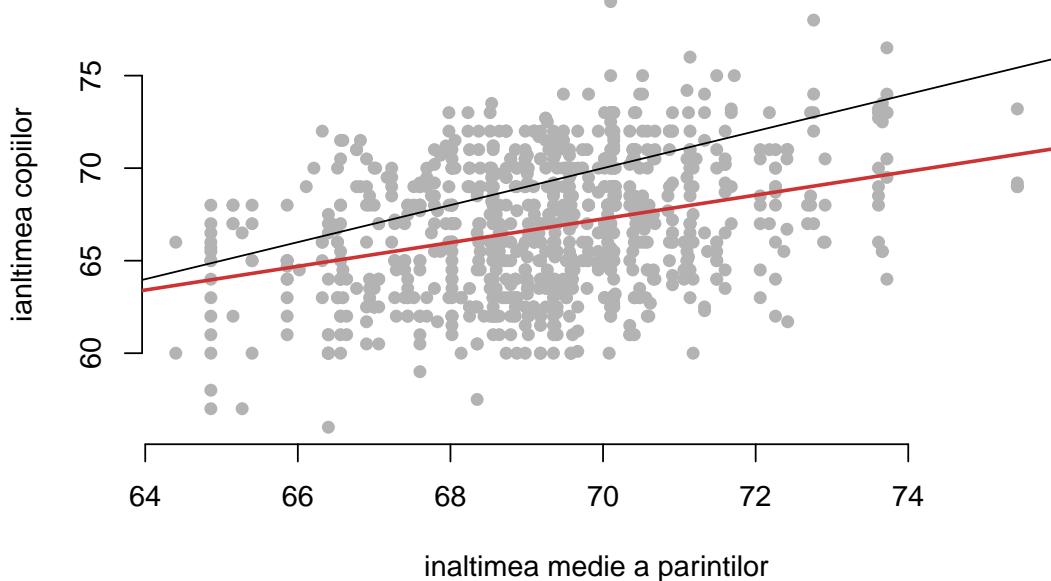


Regresie liniară

Acstea note de curs sunt destinate studenților de anul I de master de la Facultatea de Matematică și Informatică.

1 Introducere

Cuvântul “regresie” vine de la Sir Francis Galton (1822 - 1911) care, fiind interesat de problema transmiterii unui caracter ereditar de la părinți la copii, a strâns date despre înălțimea părinților și cea a copiilor lor ajunși adolescenți¹. Astfel a încercat să examineze relația dintre înălțimea copiilor și înălțimea medie a părinților (a ajustat diferențele naturale dintre sexe înmulțind înălțimea persoanelor de sex feminin cu un coeficient de 1.08).



A observat că între înălțimea copiilor (ajunși la vîrstă adultă) și înălțimea medie a părinților există o relație (aproximativ) liniară cu o pantă de $2/3$ (mai exact 0.6411904). Având o pantă mai mică de 1 , Galton a tras concluzia că acei copii care provin din familii cu părinți foarte înalți (sau scunzi) sunt în general mai scunzi (înalți) decât părinții lor. Astfel, oricare ar fi situația (familii cu părinți înalți ori scunzi), înălțimea copiilor tinde spre media populației, ceea ce Galton a numit *regresie* spre medie. Cu alte cuvinte, Galton a studiat modul în care *variabila explicativă* x = “înălțimea medie a părinților” influențează *variabila răspuns* y = “înălțimea copiilor” și a propus, ținând cont de faptul că relația dintre cele două variabile nu este exact liniară ci depinde de erori aleatoare, următorul model

¹Setul de date folosit în figura de mai jos poate fi descărcat de [aici](#)

$$y = \beta_0 + \beta_1 x + \varepsilon$$

în care componenta sistematică $\beta_0 + \beta_1 x$ este liniară și ε este eroarea aleatoare.

De manieră informală, un model explicativ este un model prin care o variabilă y este exprimată ca o funcție de una sau mai multe variabile, numite în cele ce urmează explicative. De manieră formală, încercăm să modelăm efectul unei variabile sau a mai multor variabile explicative x_1, \dots, x_k asupra variabilei de interes y . Variabila y se numește *variabilă răspuns* sau *variabilă dependentă* iar variabilele explicative se mai numesc și *covariabile*, *variabile independente* (termen pe care nu îl vom folosi), *factori* sau încă *regresori*. Într-un model de regresie, căutăm, de cele mai multe ori, să determinăm modul în care variabila răspuns evoluează, *în medie*, în funcție de variabilele explicative. O caracteristică principală a modelelor de regresie este că relația dintre y și covariabile nu se exprimă ca o funcție deterministă $f(x_1, \dots, x_k)$ ci prezintă erori aleatoare ceea ce sugerează că variabila răspuns este o variabilă aleatoare a cărei distribuție depinde de valorile variabilelor explicative. De exemplu, în cazul problemei lui Galton, chiar dacă știam cu exactitate care este înălțimea părinților nu puteam prezice exact înălțimea copiilor, ceea ce puteam face era să estimăm *înălțimea medie* a acestora și gradul de împrăștiere.

De manieră generală, modelul de regresie poate fi scris sub forma

$$y(x_1, \dots, x_k) = f(x_1, \dots, x_k) + \varepsilon(x_1, \dots, x_k)$$

unde, membrul stâng arată dependența variabilei răspuns de variabilele explicative $y(x_1, \dots, x_k)$ iar membrul drept este compus din doi termeni, *componenta sistematică* a modelului $f(x_1, \dots, x_k)$ care prezintă influența covariabilelor asupra valorii medii a variabilei dependente ($\mathbb{E}[y(x_1, \dots, x_k)] = f(x_1, \dots, x_k)$) și *componenta aleatoare*, $\varepsilon(x_1, \dots, x_k)$, numită și termen eroare care prezintă incertitudinea modelului. Cum $\mathbb{E}[y(x_1, \dots, x_k)] = f(x_1, \dots, x_k)$ avem că $\mathbb{E}[\varepsilon(x_1, \dots, x_k)] = 0$. În modelul clasic de regresie vom presupune că termenul eroare nu depinde de covariabile, prin urmare $\varepsilon(x_1, \dots, x_k) = \varepsilon$ și modelul se va scrie sub forma

$$y = f(x_1, \dots, x_k) + \varepsilon.$$

În funcție de modul în care sunt efectuate observațiile, covariabilele pot fi deterministe sau aleatoare. Pentru prima situație putem presupune că ne aflăm în contextul unui plan de experiență planificat, în care valorile covariabilelor sunt fixate înaintea derulării experimentului. De exemplu, să considerăm că ne aflăm în contextul unui experiment prin care un inginer agronom dorește să investigheze influența pe care o are cantitatea de îngrășământ (covariabilă măsurată în kg/hectar) asupra randamentului unei culturi de cereale (variabilă răspuns măsurată în tone/hectar). În acest context, suprafața cultivată se parceleză și pentru fiecare parcelă inginerul atribuie o anumită cantitate de îngrășământ prestatibilită: x_1, \dots, x_n ². Randamentul culturii de pe fiecare parcelă poate fi văzut ca o variabilă aleatoare care depinde de mai mulți factori, alții decât nivelul de îngrășământ (e.g. dăunători, umiditate în sol). Valorile observate y_1, \dots, y_n sunt văzute ca realizări ale unor variabile aleatoare Y_1, \dots, Y_n , unde Y_i reprezintă randamentul pentru nivelul de îngrășământ x_i și $\mathbb{E}[Y_i] = f(x_i)$. De cele mai multe ori, în schimb, nu ne aflăm în condițiile unui plan de experiență planificat ci în contextul unui experiment în care valorile covariabilelor nu sunt cunoscute înaintea efectuării experimentului. Să presupunem, spre exemplu, că ne aflăm în contextul unui sondaj efectuat pentru a investiga cum variază venitul (variabila răspuns) în funcție de vârstă populației (variabila explicativă). Astfel, unui individ ales la întâmplare din grupul său corespunde un cuplu de variabile aleatoare (X, Y) unde X este vârstă iar Y este valoarea venitului. În acest context, valorile observate pentru n indivizi sunt considerate ca realizări de variabile aleatoare și obiectivul este de a studia cum variază în medie venitul în funcție de vârstă, altfel spus funcția de regresie $f(x)$ este media condiționată a lui Y la valoare lui $X = x$, $\mathbb{E}[Y|X = x] = f(x)$.

Prin urmare în cazul modelului de regresie condiționat, componenta sistematică este dată de media condiționată la nivelul covariabilelor iar forma generală a modelului de regresie devine

²Aici trebuie avut grijă să nu se confundă valorile x_1, \dots, x_n care corespund unei singure covariabile cu n covariabile.

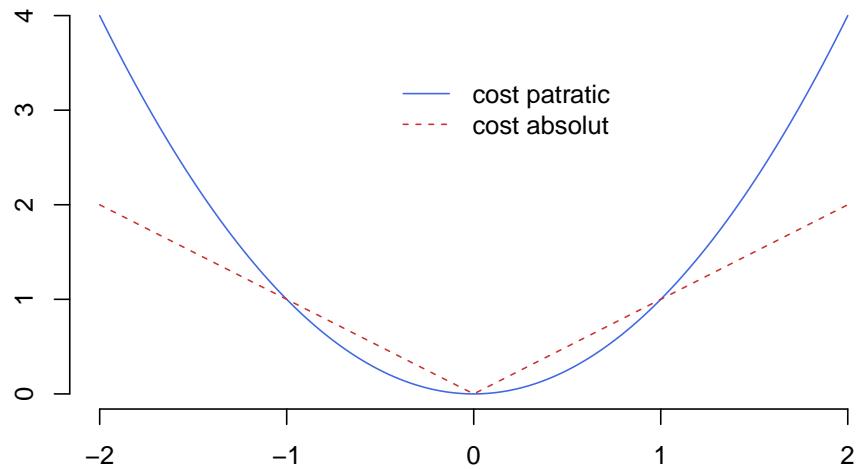
$$y = \mathbb{E}[y|x_1, \dots, x_k] + \varepsilon = f(x_1, \dots, x_k) + \varepsilon.$$

Modelele de regresie se pot clasifica, în funcție de forma variabilei răspuns, în modele univariate atunci când aceasta este o variabilă aleatoare și respectiv multivariate atunci când aceasta este un vector aleator iar în raport cu numărul de predictori în modele simple atunci când avem un singur predictor sau modele multiple atunci când intervin mai multe variabile explicative. Raportându-ne la forma componentei sistematice, putem avea modele liniare, în care parametrii care descriu forma lui f intră liniar (e.g. $f(x) = \beta_0 + \beta_1 x$, $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ sau $f(x) = \beta_0 + \beta_1 \log(x) + \beta_2 \cos(x)$), sau modele neliniare, în care parametrii nu apar liniar (e.g. $f(x) = \beta_0 + \beta_1 e^{\beta_3 x}$).

Scopul analizei de regresie este de a utiliza observațiile, datele, $y_i, x_{i1}, \dots, x_{ik}$, $i = 1, \dots, n$ în vederea estimării (aproximării) componentei sistematice f a modelului și de a o separa pe aceasta de componenta aleatoare ε . Problema matematică se scrie sub forma

$$\operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n L(y_i - f(x_{i1}, \dots, x_{ik})),$$

unde L se numește funcție de cost sau de pierdere (loss) iar \mathcal{F} este clasa de funcții în care presupunem că se regăsește adevarata componentă sistematică f , altfel spus dorim să determinăm acea funcție din clasa de funcții \mathcal{F} care minimizează costul. Cel mai des utilizate funcții de cost sunt costul pătratic, $L(u) = u^2$, și respectiv costul absolut $L(u) = |u|$.



În acest curs vom studia modelul clasic de regresie liniară în care componenta sistematică f face parte din clasa de funcții liniare $\mathcal{F} = \{f \mid f(x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k\}$, prin urmare media (condiționată) a lui y este o combinație liniară de covariabile iar variabila răspuns este continuă:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon.$$

Atunci când înlocuim cu observațiile obținem n ecuații

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

cu parametrii necunoscuți (sau coeficienții de regresie) β_0, \dots, β_k .

2 Seturi de date

În secțiunile care urmează vom prezenta o serie de seturi de date care vor fi utilizate pe parcursul acestor note pentru a ilustra noțiunile descrise.

2.1 Înălțimea arborilor de eucalipt

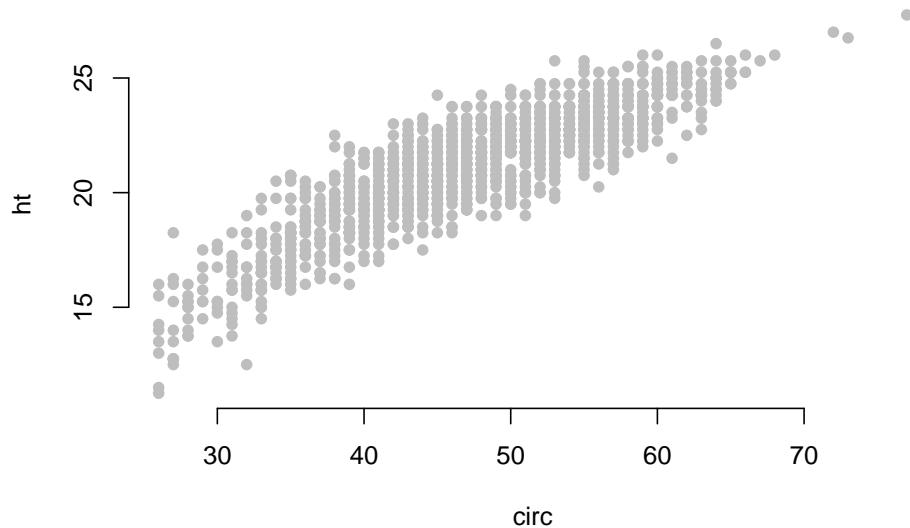
Setul de date **Eucalypt** (care poate fi descărcat de [aici](#)) face referire la înălțimea și circumferința (măsurată la 1m 30cm de sol) a 1429 arbori de eucalipt plantați într-o regiune experimentală din Franța. Cele două caracteristici sunt măsurate la vîrsta de maturitate a arborilor, anume la 6 ani. O imagine a primelor observații din setul de date este dată de tabelul de mai jos:

individ	ht	circ
1	18.25	36
2	19.75	42
3	16.50	33
4	18.25	39
5	19.50	43
6	16.25	34

Pentru a avea o imagine de ansamblu asupra datelor putem folosi funcția **summary** și aceasta întoarce:

ht	circ
Min. :11.25	Min. :26.00
1st Qu.:19.75	1st Qu.:42.00
Median :21.75	Median :48.00
Mean :21.21	Mean :47.35
3rd Qu.:23.00	3rd Qu.:54.00
Max. :27.75	Max. :77.00

Scopul este de a găsi o relație între înălțimea arborilor și circumferința acestora în vederea estimării volumului de lemn din zona studiată (volum calculat după o formulă de tip trunchi de con). Reprezentarea setului de date este dată în figura de mai jos:



2.2 Prețul chiriei locuințelor în München

Setul de date **Munchen** (care poate fi descărcat de [aici](#)) face referire la prețul net și respectiv prețul net pe metrul pătrat al chiriei unei locuințe din orașul München, Germania pentru anul 1999. Setul de date prezintă prețurile a mai multe de 3000 de apartamente împreună cu o serie de variabile explicative precum suprafața de locuit, anul de construcție a imobilului, etc. Aceste informații pot fi găsite în tabelul de mai jos:

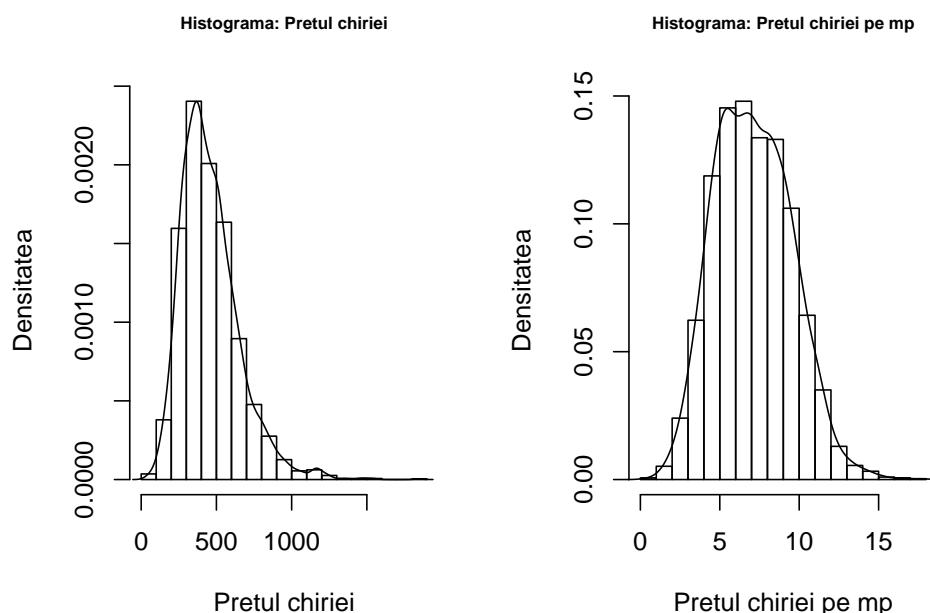
Variabilă	Descriere	Media/Frecvență	Abaterea standard	Min/Max
rent	Prețul lunar net al chiriei (în Euro)	459.44	195.66	40.51/1843.38
rentsqm	Prețul lunar net al chiriei pe m^2 (în Euro)	7.11	2.44	0.41/17.72
area	Suprafața de locuit în m^2	67.37	23.72	20/160
yearc	Anul de construcție	1956	22.31	1918/1997
location	Calitatea locației: 1 - medie, 2 - bună și 3- de top	58.91%, 39.26%, 2.53%		
bath	Calitatea băilor: 0 - standard, 1 - premium	93.8%, 6.2%		
kitchen	Calitatea bucătariei: 0 - standard, 1 - premium	95.75%, 4.25%		
cheating	Încălzire centralizată: 0 - fără încălzire, 1 - cu încălzire	10.42%, 89.58%		

2.3 Primii pași

Înainte de a începe analiza de regresie (sau orice altă analiză) este bine să înțelegem mai bine variabilele din setul de date cu care lucrăm. Pentru a realiza acest lucru, un prim pas constă în summarizarea variabilelor

atât prin statistici descriptive (calcularea mediilor, medianelor, a abaterilor standard, a valorilor minime și maxime, etc.) cât și prin tehnici de vizualizare (histograme, diagrame cu bare, boxplot, etc.). Atunci când lucrăm cu variabile cantitative, statisticile descriptive se rezumă la măsuri de locție (media, mediana, modul) și variație (abaterea standard, minimul, maximul) iar în cazul variabilelor calitative putem include frecvența de apariție a fiecărei categorii.

Dacă facem referire la setul de date a indicilor prețului chiriei în München, atunci observăm că prețul net lunar variază între 40 și 1843 de Euro având o medie de 459 de Euro. Figura de mai jos arată cum este repartizat prețul net lunar și respectiv prețul net lunar pe m^2 și constatăm că pentru majoritatea apartamentelor acesta variază între 50 și 1200 de Euro:



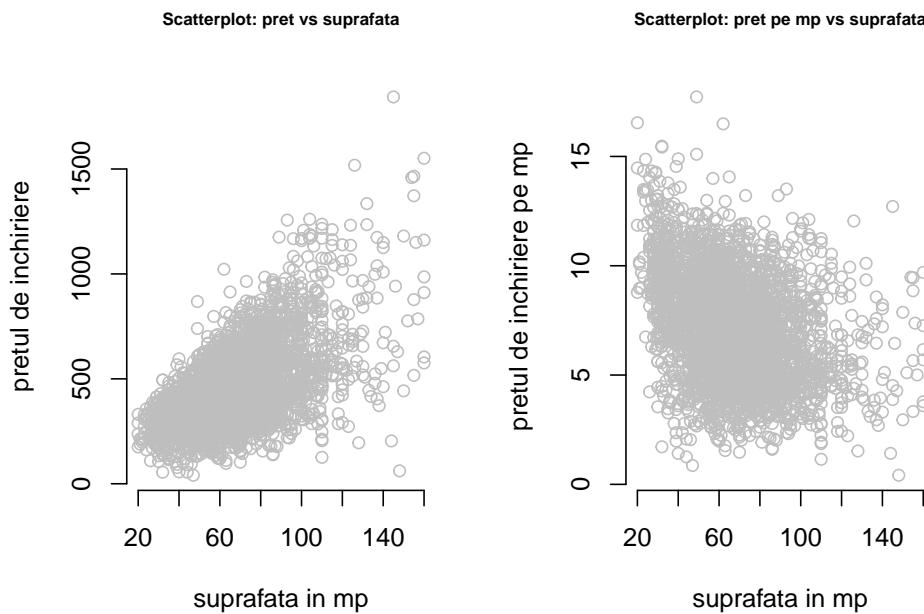
Ex. 2.1



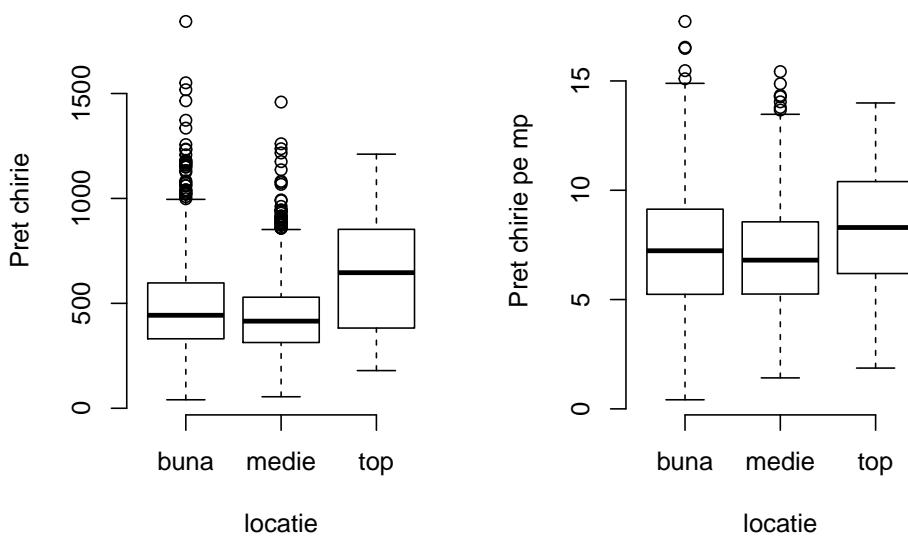
Ilustrati prin intermediul unei histograme (`hist()`) cum este repartizată suprafața de locuit (variabila `area`) și respectiv anul de construcție a imobilului (variabila `yearch`).

În cazul în care ne interesăm asupra relației dintre variabila răspuns și variabila explicativă putem folosi ca metodă grafică diagrama de împărăștiere (scatterplot) în situația în care covariabila este continuă sau boxplot-ul în situația în care covariabila este categorică.

De exemplu, figura de mai jos prezintă diagrama de împărăștiere dintre prețul lunar net sau prețul lunar net pe m^2 și suprafața de locuit. Dat fiind numărul mare de observații graficul este aglomerat și nu foarte informativ. Cu toate acestea constatăm o oarecare relație liniară între prețul lunar net și suprafața de locuit precum și că variabilitatea prețului crește odată cu suprafața.



Atunci când variabila explicativă este categorială este de preferat utilizarea boxplot-ului (diagramei cu mustați). Astfel, în figura următoare se poate observa că valoarea mediană a prețului lunar net al chiriei crește odată cu calitatea locației apartamentului.

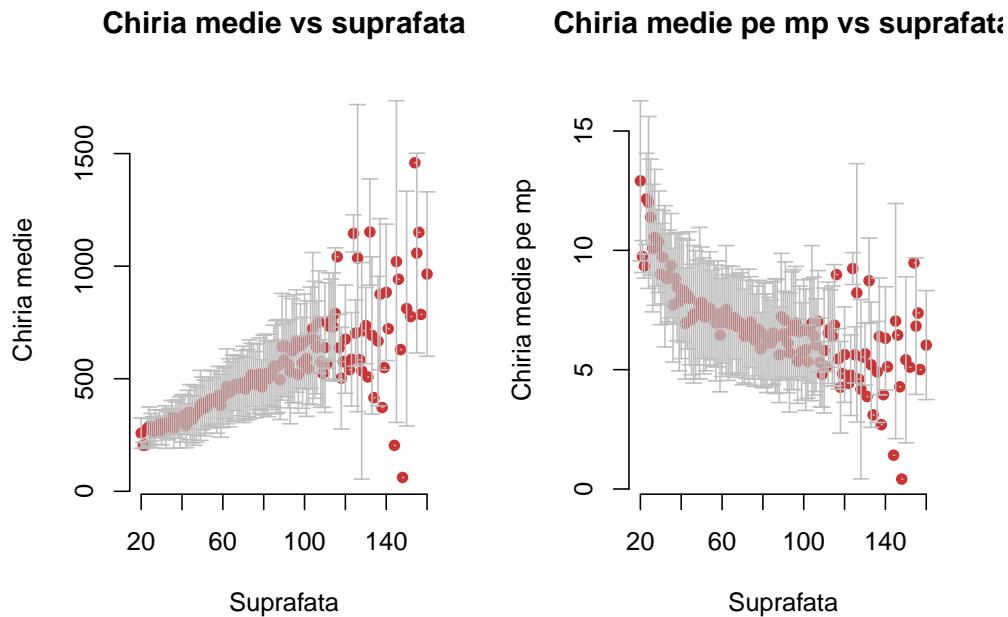


Ex. 2.2



Ilustrați prin intermediul unei diagrame de împărtiere relația dintre prețul lunar net sau prețul lunar net pe m^2 și anul de construcție a imobilului iar prin intermediul unui boxplot relația dintre prețul lunar net sau prețul lunar net pe m^2 și calitatea băii sau a bucătăriei.

Numărul mare de observații din setul de date face dificilă interpretarea diagramei de împrăștiere și în această situație o reprezentare grafică pe grupuri (clustere) de date este de preferat. De exemplu dacă numărul valorilor unice ale variabilei explicative este mic în raport cu numărul observațiilor atunci o idee ar fi să sumarizăm variaabilă răspuns pentru fiecare nivel (medie, medie - abatere standard, medie + abatere standard) și să ilustrăm doar datele sumarizate.



Functia care permite trasarea graficului anterior este:

```
plot.with.errorbars = function(x, y, err_low, err_up, ...) {
  ylim = c(min(err_low), max(err_up))

  plot(x, y, ylim=ylim,
        pch=16,
        bty = "n",
        col = "brown3",
        ...)
  arrows(x, err_low, x, err_up, length=0.05, angle=90, code=3,
         col = "grey")
}
```

3 Modelul de Regresie liniară simplă

În cele ce urmează vom considera cazul modelului de regresie liniară simplă

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

unde, în forma generală $y = f(x) + \varepsilon$, media condiționată (componenta sistematică) $\mathbb{E}[y|x] = f(x)$ este presupusă liniară. Specific, pentru un eșantion de n puncte (x_i, y_i) modelul de regresie liniară simplă se scrie sub forma

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

Modelul clasic de regresie presupune că termenii eroare sunt variabile aleatoare necorelate, centrate și de varianță constantă (homoscedasticitate), altfel spus aceștia îndeplinesc ipotezele:

$$(\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \mathbb{E}[\varepsilon_i] = 0 \text{ pentru toți indicii } i \\ (\mathcal{H}_2) : \text{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2 \text{ pentru toate perechile } (i, j) \end{cases}$$

unde $\delta_{ij} = 1$ dacă $i = j$ și $\delta_{ij} = 0$ altfel.

Pentru a determina coeficienții de regresie (β_0 și β_1) vom folosi ca funcție de pierdere, costul pătratic $L(u) = u^2$. În acest context, numim estimatori obținuți prin *metoda celor mai mici pătrate* (OLS - Ordinary Least Squares) valorile $\hat{\beta}_0$ și $\hat{\beta}_1$ care minimizează funcția (RSS - Residual Sum of Squares)

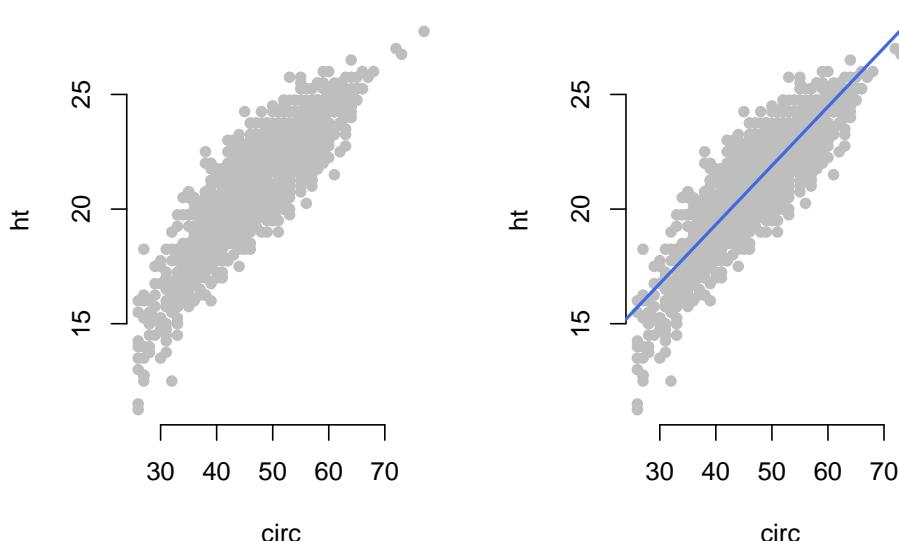
$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

altfel spus, dreapta de regresie obținută prin metoda celor mai mici pătrate minimizează distanțele verticale dintre punctele (x_i, y_i) și dreapta ajustată $y = \hat{\beta}_0 + \hat{\beta}_1 x$. Pentru unicitatea estimatorilor $\hat{\beta}_0$ și $\hat{\beta}_1$ vom presupune că setul de date conține cel puțin două puncte de abscise diferite, i.e. $x_i \neq x_j$.

a) *Exemplu - Înălțimea arborilor de eucalipt*

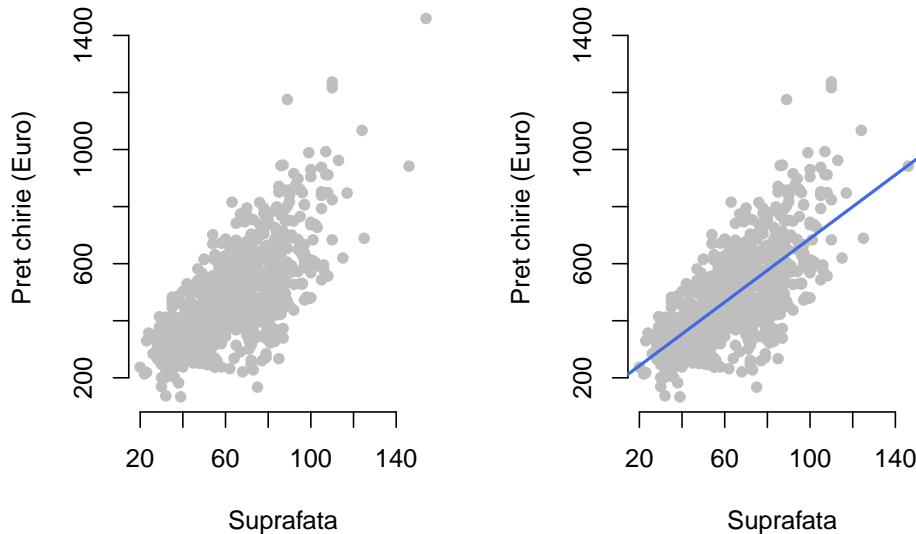
Ca prim exemplu, putem considera setul de date referitor la înălțimea și circumferința arborilor de eucalipt. Modelul de regresie liniară simplă prin care dorim să explicăm înălțimea (medie) a arborilor (variabila răspuns) în funcție de circumferința lor (variabila explicativă) este dat de

$$ht_i = \beta_0 + \beta_1 circ_i + \varepsilon_i, \quad i = 1, \dots, 1429.$$



b) *Exemplu - Prețul chiriielor în München*

Să considerăm acum setul de date referitor la prețul chiriilor în München pentru apartamentele dintr-o locație medie, construite după anul 1966. Diagrama de împrăștiere ilustrată în figura de mai jos, prezintă o relație aproximativ liniară între prețul net al chiriei (variabila răspuns) și suprafață (covariabila).



Modelul de regresie liniară simplă se scrie

$$pret_i = \beta_0 + \beta_1 suprafata_i + \varepsilon_i$$

ceea ce înseamnă că prețul de închiriere mediu este o funcție liniară de suprafață de locuit, i.e. $\mathbb{E}[pret|suprafata] = \beta_0 + \beta_1 suprafata$.

3.1 Metoda celor mai mici pătrate

3.1.1 Calculul estimatorilor prin metoda celor mai mici pătrate

Metoda celor mai mici pătrate este o metodă deterministă de calcul a estimatorilor coeficientilor dreptei de regresie, ipotezele făcute asupra termenilor eroare nu intervin în acest calcul. Acestea din urmă vor interveni atunci când vrem să explicităm proprietățile statistice ale acestor estimatori.

Prop. 3.1



Estimatorii $\hat{\beta}_0$ și $\hat{\beta}_1$ obținuți prin metoda celor mai mici pătrate, adică valorile coeficientilor β_0 și β_1 care minimizează funcția

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

sunt dați de expresiile

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

Trebuie să determinăm

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(\beta_0, \beta_1) \in \mathbb{R} \times \mathbb{R}} RSS(\beta_0, \beta_1)$$

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

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

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

A doua ecuație conduce la

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

și înlocuind β_0 cu expresia obținută anterior, obținem soluția

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

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

$$\begin{aligned} RSS(\beta_0, \beta_1) &= n [\beta_0 - (\bar{y} - \hat{\beta}_1 \bar{x})]^2 + \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \left[\beta_1 - \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 \\ &\quad + \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right] \left[1 - \frac{(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} \right] \end{aligned}$$

care justifică în egală măsură soluția obținută anterior. \square

Odată ce am determinat estimatorii $\hat{\beta}_0$ și $\hat{\beta}_1$ putem scrie dreapta de regresie sub forma

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

și, în acest context, dacă evaluăm dreapta în punctele x_i care au ajutat la estimarea parametrilor atunci obținem valorile ajustate (fitate) \hat{y}_i iar dacă evaluăm dreapta în alte puncte, valorile obținute se numesc valori prezise (valori previzionale). De asemenea, din $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ se remarcă faptul că dreapta de regresie trece prin punctul de coordonate (\bar{x}, \bar{y}) , centrul de greutate al norului de puncte.

Exemplu: prețul chiriilor în Munchen

Exp. 3.2

Putem ilustra modelul de regresie liniară simplă în contextul prețului chiriilor din Munchen pentru apartamentele construite după anul 1966 care se regăsesc într-o locție medie. Conform metodei celor mai mici pătrate, găsim că $\hat{\beta}_0 = 130.554$ și respectiv $\hat{\beta}_1 = 5.576$ ceea ce conduce la modelul

$$pret_i = 130.554 + 5.576 \text{suprafața}_i + \varepsilon_i.$$

Panta dreptei de regresie, coeficientul $\hat{\beta}_1 = 5.576$ poate fi interpretat în modul următor: dacă suprafața de locuit crește cu 1 m^2 atunci prețul chiriei crește în medie cu 5.576 Euro.

3.1.2 Proprietăți ale estimatorilor obținuți prin metoda celor mai mici pătrate

Sub ipotezele făcute asupra termenilor eroare (H_1 și H_2), de centrare, necorelare și homoscedasticitate putem prezenta o serie de proprietăți ale estimatorilor obținuți prin metoda celor mai mici pătrate.

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

Prop. 3.3

Coeficienții $\hat{\beta}_0$ și $\hat{\beta}_1$ obținuți prin metoda celor mai mici pătrate sunt dați de $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ și $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ (aceștia sunt variabile aleatoare deoarece sunt funcții de Y_i care sunt variabile aleatoare).

Înlocuind în expresia lui $\hat{\beta}_1$ pe y_i cu $\beta_0 + \beta_1 x_i + \varepsilon_i$ avem

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

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

În mod similar,

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

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

Putem de asemenea să determinăm varianța și covarianța estimatorilor $\hat{\beta}_0$ și $\hat{\beta}_1$.

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

Prop. 3.4

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

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

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

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

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

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

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

prin urmare

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

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

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

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

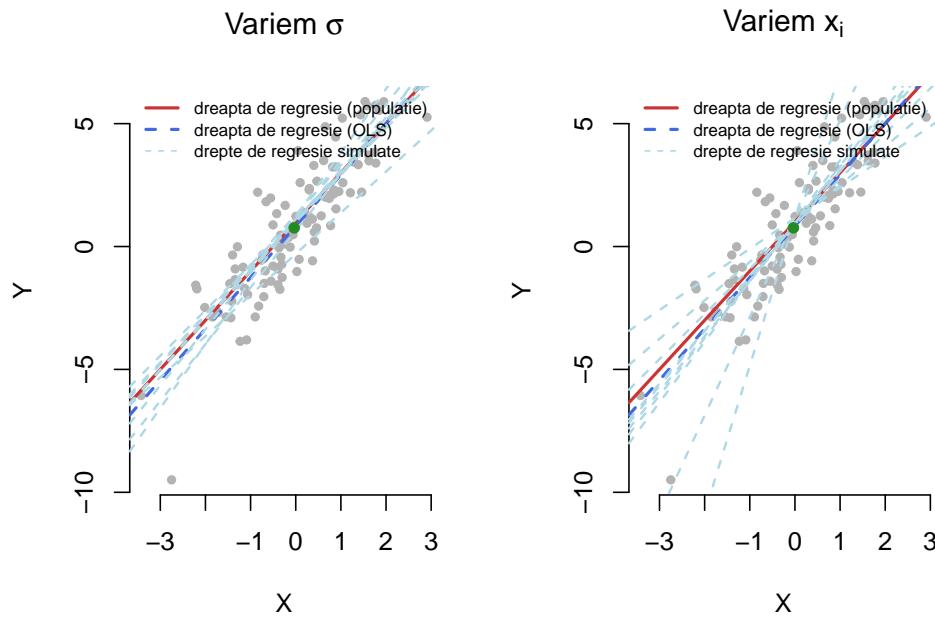
Matricea de varianță-covarianță a estimatorilor $\hat{\beta}_0$ și $\hat{\beta}_1$ devine

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

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

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

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



Rezultatul următor, cunoscut și sub numele de *Teorema Gauss-Markov*, afirmă că estimatorii obținuți prin metoda celor mai mici pătrate sunt optimali în clasa estimatorilor liniari și nedeplasați.

Prop. 3.5 În clasa estimatorilor nedeplasați și liniari în y , estimatorii $\hat{\beta}_0$ și $\hat{\beta}_1$ sunt de varianță minimală.

Începem prin a reaminti că un estimator este liniar în y dacă se poate scrie sub forma $\sum_{i=1}^n d_i y_i$ cu d_1, \dots, d_n constante. Să observăm că atât $\hat{\beta}_0$ cât și $\hat{\beta}_1$ sunt estimatori liniari în y_i , $\hat{\beta}_1 = \sum_{i=1}^n \lambda_i y_i$ unde $\lambda_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Fie $\tilde{\beta}_1$ un alt estimator liniar și nedeplasat pentru β_1 , cu alte cuvinte

$$\tilde{\beta}_1 = \underbrace{\sum_{i=1}^n d_i y_i}_{\text{liniaritate}} \quad \text{și} \quad \underbrace{\mathbb{E}[\tilde{\beta}_1] = \beta_1, \forall \beta_0, \beta_1}_{\text{nedeplasare}}.$$

Observăm că

$$\mathbb{E}[\tilde{\beta}_1] = \beta_0 \sum_{i=1}^n d_i + \beta_1 \sum_{i=1}^n d_i x_i + \sum_{i=1}^n d_i \underbrace{\mathbb{E}[\varepsilon_i]}_{=0} = \beta_0 \sum_{i=1}^n d_i + \beta_1 \sum_{i=1}^n d_i x_i$$

prin urmare, folosind proprietatea de nedeplasare, $\beta_0 \sum_{i=1}^n d_i + \beta_1 \sum_{i=1}^n d_i x_i = \beta_1$ pentru orice valori ale lui β_0 și β_1 ceea ce implica $\sum_{i=1}^n d_i = 0$ și respectiv $\sum_{i=1}^n d_i x_i = 1$.

Pentru a verifica inegalitatea $Var(\tilde{\beta}_1) \geq Var(\hat{\beta}_1)$, să notăm că

$$Var(\tilde{\beta}_1) = Var(\tilde{\beta}_1 - \hat{\beta}_1) + Var(\hat{\beta}_1) + 2Cov(\tilde{\beta}_1 - \hat{\beta}_1, \hat{\beta}_1)$$

dar

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

și tinând cont că $\sum_{i=1}^n d_i = 0$ și $\sum_{i=1}^n d_i x_i = 1$ rezultă că $Cov(\tilde{\beta}_1 - \hat{\beta}_1, \hat{\beta}_1) = 0$ ceea ce conduce la

$$Var(\tilde{\beta}_1) = Var(\tilde{\beta}_1 - \hat{\beta}_1) + Var(\hat{\beta}_1) \geq Var(\hat{\beta}_1) \quad \square$$

3.1.3 Valori reziduale

În modelul de regresie liniară simplă am estimat prin intermediul metodei celor mai mici pătrate atât ordonata la origine a dreptei de regresie, coeficientul $\hat{\beta}_0$, cât și panta acesteia, coeficientul $\hat{\beta}_1$. Definim *valorile reziduale* $\hat{\varepsilon}_i$ ca fiind diferența dintre ordonata observată a punctului și ordonata ajustată la dreapta de regresie, altfel spus

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$



În cadrul modelului de regresie liniară simplă, suma valorilor reziduale este nulă.

Prop. 3.6

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

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

Trebuie observat că atât varianțele cât și covarianța estimatorilor $\hat{\beta}_0$ și $\hat{\beta}_1$ depind de varianța termenului eroare σ^2 , care în general nu este cunoscută. În propoziția de mai jos este propus un estimator nedeplasat a lui σ^2 .

Prop. 3.7



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

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

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

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

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

Luând media găsim că

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

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

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

Exemplu: prețul chiriielor în München

Exp. 3.8

Observăm că pentru setul de date care face referire la prețul chiriielor în München, găsim că valoarea estimatorului varianței termenului eroare este $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2 = 1.5241065 \times 10^4$ iar matricea de varianță-covarianță a estimatorilor obținuți prin metoda celor mai mici pătrate este

$$W = \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{pmatrix} = \begin{pmatrix} \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} & -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{pmatrix} = \begin{pmatrix} 208.29 & -2.96 \\ -2.96 & 0.04 \end{pmatrix}.$$

3.1.4 Predicție

Unul dintre scopurile modelului de regresie este acela de a face predicție, cu alte cuvinte de a prezice valoarea variabilei răspuns y în raport cu o nouă observație a variabilei explicative x .

Prop. 3.9



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

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

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

Atunci varianța răspunsului mediu prezis este

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

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

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

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

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

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

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

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

3.2 Coeficientul de determinare R^2 și coeficientul de corelație

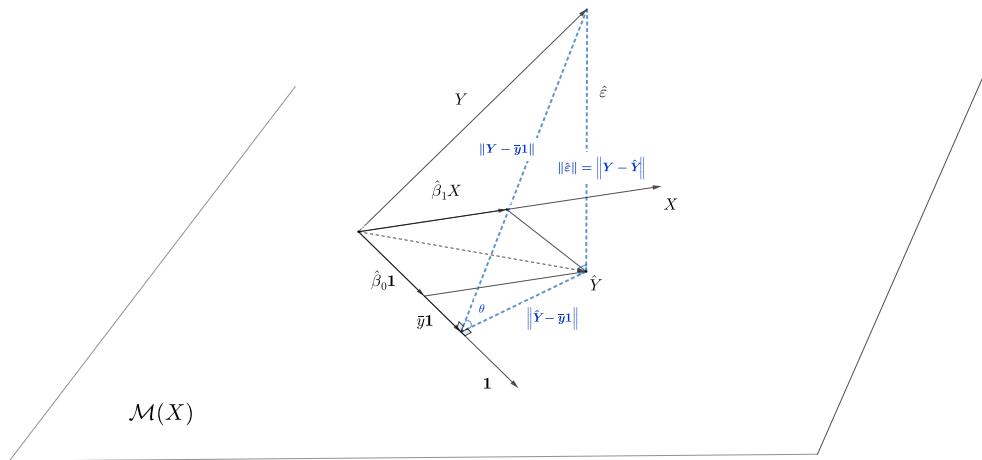
În această secțiune încercăm să abordăm problema de regresie liniară simplă într-un context geometric. Din punct de vedere vectorial dispunem de doi vectori: vectorul $X = (x_1, x_2, \dots, x_n)^\top$ a celor n observații ale

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

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

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

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



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

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

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

unde $SS_T = \sum_{i=1}^n (y_i - \bar{y})^2$ reprezintă suma abaterilor pătratice totale (Total Sum of Squares), $SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ reprezintă suma abaterilor pătratice explicate de modelul de regresie (Regression Sum of Squares) iar $RSS = \sum_{i=1}^n \hat{\varepsilon}_i^2$ reprezintă suma abaterilor pătratice reziduale (Residual Sum of Squares).

Coeficientul de determinare R^2 este definit prin

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

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



În modelul de regresie liniară simplă avem

Prop. 3.10

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

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

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

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

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

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

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

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

De asemenea

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

și cum

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

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

Exemplu: prețul chiriiilor în Munchen

Exp. 3.11 Pentru a vedea dacă modelul de regresie liniară simplă este potrivit în contextul setului de date referitor la prețul chiriiilor în Munchen, vom calcula coeficientul de determinare R^2 . Astfel obținem că $R^2 = 0.472$ ceea ce implică faptul că modelul nostru nu este foarte bine ajustat la date. Trebuie să ținem cont că modelul ales este unul simplu și de asemenea să remarcăm faptul că setul de date nu respectă întru totul ipoteza homoscedasticității erorilor, se observă că variabilitatea în prețul chiriiilor crește odată cu creșterea suprafetei de locuit. Această problemă va fi tratată într-o secțiune ulterioară care face referire la validarea ipotezelor modelului propus.

3.3 Inferență statistică - cazul erorilor gaussiene

Până în acest moment am determinat valorile estimatorilor coeficienților dreptei de regresie și varianțele acestor estimatori dar nu am vorbit despre modul în care aceștia sunt repartizați. Pentru a putea face acest lucru trebuie să adăugăm o ipoteză nouă asupra termenilor eroare ε_i ce face referire la modul de repartizare a acestora. Vom presupune că termenii eroare sunt normal repartizați și în acest caz ipotezele \mathcal{H}_1 și \mathcal{H}_2 se scriu sub forma

$$(\mathcal{H}) \left\{ \begin{array}{l} (\mathcal{H}_1) : \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \\ (\mathcal{H}_2) : \varepsilon_i \text{ independente} \end{array} \right.$$

Prin urmare modelul de regresie liniară simplă devine un model parametric guvernăt de parametrul $\theta = (\beta_0, \beta_1, \sigma^2)$ care ia valori în $\theta \in \Theta = \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+^*$. Cum cunoaștem repartițiile termenilor eroare putem determina și repartițiile variabilelor răspuns y_i ,

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

iar y_i sunt independente pentru că ε_i sunt. Astfel putem calcula funcția de verosimilitate și estimatorii de verosimilitate maximă pentru θ . Funcția de verosimilitate se scrie

$$\begin{aligned}
 L(\beta_0, \beta_1, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right] \\
 &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left[-\frac{1}{2\sigma^2} RSS(\beta_0, \beta_1) \right]
 \end{aligned}$$

iar logaritmul acesteia devine

$$\ell(\beta_0, \beta_1, \sigma^2) = \log L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} RSS(\beta_0, \beta_1).$$

Pentru a determina parametrii $(\beta_0, \beta_1, \sigma^2)$ care maximizează funcția $\ell(\beta_0, \beta_1, \sigma^2)$ să observăm pentru început că (β_0, β_1) intervin doar termenul $RSS(\beta_0, \beta_1)$ care trebuie astfel minimizat (apare cu semn negativ în expresia lui ℓ). Cum valorile (β_0, β_1) care minimizează funcția $RSS(\beta_0, \beta_1)$ sunt date de valorile obținute prin metoda celor mai mici pătrate (vezi Prop. 3.1) deducem că estimatorii de verosimilitate maximă pentru (β_0, β_1) sunt $(\hat{\beta}_0, \hat{\beta}_1)$. Pentru a determina estimatorul de verosimilitate maximă pentru σ^2 rămâne să găsim valoarea maximă a lui $L(\hat{\beta}_0, \hat{\beta}_1, \sigma^2)$ în raport cu σ^2 . Prin derivare obținem

$$\frac{\partial \log L(\hat{\beta}_0, \hat{\beta}_1, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} RSS(\hat{\beta}_0, \hat{\beta}_1) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

și rezolvând ecuația $\frac{\partial \log L(\hat{\beta}_0, \hat{\beta}_1, \sigma^2)}{\partial \sigma^2} = 0$ găsim

$$\sigma_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2.$$

Observăm că estimatorul de verosimilitate maximă pentru σ^2 este deplasat, $\mathbb{E}[\sigma_{MLE}^2] = \frac{n-2}{n}\sigma^2$, și diferă de estimatorul $\hat{\sigma}^2$ văzut anterior (cf. Prop. 3.7).

Pentru o mai bună înțelegere a conceptelor și a noțiunilor ce urmează a fi prezentate în cadrul acestei secțiuni puteți consulta Anexele.

3.3.1 Repartițiile estimatorilor și regiuni de încredere

Pentru a ușura scrierea rezultatelor din această secțiune vom folosi următoarele notații (a se vedea Prop. 3.4):

$$\begin{aligned} \sigma_{\hat{\beta}_0}^2 &= \sigma^2 \left(\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right), & \hat{\sigma}_{\hat{\beta}_0}^2 &= \hat{\sigma}^2 \left(\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right) \\ \sigma_{\hat{\beta}_1}^2 &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2}, & \hat{\sigma}_{\hat{\beta}_1}^2 &= \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2} \end{aligned}$$

unde $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2$. Observăm că $\sigma_{\hat{\beta}_0}^2, \sigma_{\hat{\beta}_1}^2$ sunt varianțele (teoretice) estimatorilor obținuți prin metoda celor mai mici pătrate iar $\hat{\sigma}_{\hat{\beta}_0}^2, \hat{\sigma}_{\hat{\beta}_1}^2$ sunt estimatorii acestora obținuți prin înlocuirea lui σ^2 cu estimatorul său nedeplasat $\hat{\sigma}^2$.

Prop. 3.12



Repartițiile estimatorilor obținuți prin metoda celor mai mici pătrate atunci când varianța σ^2 este cunoscută sunt:

1. $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma_{\hat{\beta}_j}^2)$ pentru $j = 0, 1$
2. $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim \mathcal{N}(\beta, \sigma^2 V)$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ și $V = \frac{1}{\sum(x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$
3. $\frac{(n-2)}{\sigma^2} \hat{\sigma}^2$ este repartizată χ^2 cu $n-2$ grade de libertate, i.e. $\frac{(n-2)}{\sigma^2} \hat{\sigma}^2 \sim \chi_{n-2}^2$
4. $(\hat{\beta}_0, \hat{\beta}_1)$ și $\hat{\sigma}^2$ sunt independente.

Demonstrația acestui rezultat se va face în contextul general al regresiei liniare multiple și va fi omis acum.

Cum, în general, nu cunoaștem valoarea lui σ^2 aceasta va fi estimată prin intermediul lui $\hat{\sigma}^2$. În acest context repartițiile estimatorilor obținuți prin metoda celor mai mici pătrate se vor modifica astfel:

Prop. 3.13



Repartițiile estimatorilor obținuți prin metoda celor mai mici pătrate atunci când varianța σ^2 nu este cunoscută verifică:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-2},$$

pentru $j = 0, 1$, unde t_{n-2} este repartitia Student cu $n - 2$ grade de libertate și respectiv

$$\frac{1}{2\hat{\sigma}^2}(\hat{\beta} - \beta)^\top V^{-1}(\hat{\beta} - \beta) \sim F_{2,n-2},$$

unde $F_{2,n-2}$ este repartitia Fisher-Snedecor cu 2 grade de libertate la numărător și $n - 2$ grade de libertate la numitor.

Acest rezultat ne permite să construim intervale și regiuni de încredere pentru estimatorii noștri. Următoarea propoziție prezintă intervalele și regiunile de încredere la un nivel de semnificație $1 - \alpha$:

Prop. 3.14



Avem următoarele intervale și regiuni de încredere pentru estimatorii obținuți prin metoda celor mai mici pătrate. Un interval de încredere bilateral de nivel $1 - \alpha$ pentru β_j , $j = 0, 1$ este

$$[\hat{\beta}_j - t_{n-2}(1 - \alpha/2)\hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + t_{n-2}(1 - \alpha/2)\hat{\sigma}_{\hat{\beta}_j}]$$

unde $t_{n-2}(1 - \alpha/2)$ este cuantila de ordin $1 - \alpha/2$ a repartitiei Student t_{n-2} .

O regiune de încredere de nivel $1 - \alpha$ pentru parametrii β este dată de

$$RC(\beta_0, \beta_1) = \left\{ \frac{1}{2\hat{\sigma}^2} \left[n(\hat{\beta}_0 - \beta_0)^2 + 2n\bar{x}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + \sum x_i^2(\hat{\beta}_1 - \beta_1)^2 \right] \leq f_{(2,n-2)}(1 - \alpha) \right\}$$

unde $f_{(2,n-2)}(1 - \alpha)$ este cuantila de ordin $1 - \alpha$ din repartitia Fisher-Snedecor $F_{2,n-2}$.

Un interval de încredere de nivel $1 - \alpha$ pentru σ^2 este

$$\frac{(n - 2)\hat{\sigma}^2}{\chi_{n-2}^2(1 - \alpha/2)}, \frac{(n - 2)\hat{\sigma}^2}{\chi_{n-2}^2(\alpha/2)}$$

unde $\chi_{n-2}^2(\alpha)$ este cuantila de ordin α a repartitiei χ_{n-2}^2 .

Regiunea de încredere din propoziția anterioară este o elipsă și face referire la parametrii de regresie $\beta = (\beta_0, \beta_1)^\top$ simultan luând în calcul și corelația dintre aceștia, spre deosebire de intervalele de încredere corespunzătoare.

Exemplu: prețul chirilor în Munchen

Ex. 3.15

Raportându-ne la setul de date referitor la prețul chirilor în Munchen avem următoarele intervale de încredere:

- pentru $\hat{\beta}_0$

$$[\hat{\beta}_0 - t_{n-2}(1 - \alpha/2)\hat{\sigma}_{\hat{\beta}_0}, \hat{\beta}_0 + t_{n-2}(1 - \alpha/2)\hat{\sigma}_{\hat{\beta}_0}] = [102.221, 158.886]$$

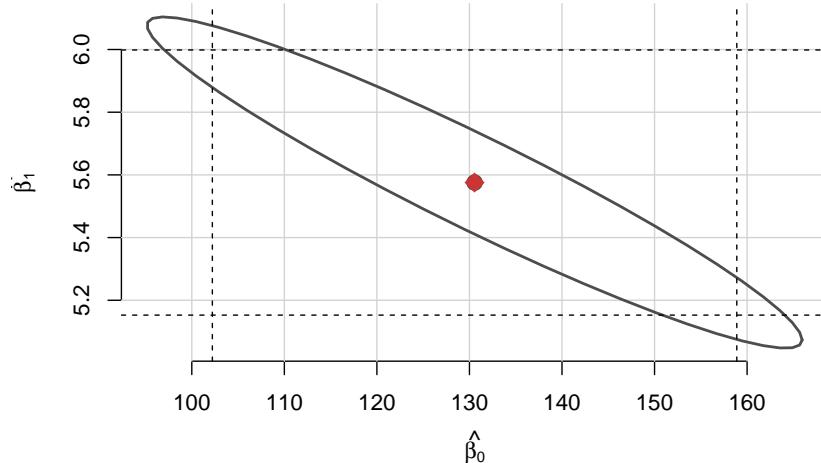
Observăm că intervalul de încredere pentru $\hat{\beta}_0$ este destul de larg (lungime 56.665), ceea ce este justificat dat fiind variabilitatea erorilor ($\hat{\sigma} = 123.455$) dar în special pentru că suprafețele sunt mai mari de zero.

- pentru $\hat{\beta}_1$

$$\left[\hat{\beta}_1 - t_{n-2}(1 - \alpha/2)\hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + t_{n-2}(1 - \alpha/2)\hat{\sigma}_{\hat{\beta}_1} \right] = [5.152, 5.998]$$

Pentru $\hat{\beta}_1$ intervalul de încredere este mult mai mic ceea ce arată că avem un efect al suprafetei asupra prețului net al apartamentelor.

Referitor la regiunea de încredere $RC(\beta_0, \beta_1)$ în figura de mai jos (realizată prin intermediul pachetului **ellipse**) este ilustrată regiunea $RC(\beta_0, \beta_1)$ împreună cu intervalele de încredere asociate pentru fiecare parametru:



Se poate observa că regiunea de încredere $RC(\beta_0, \beta_1)$ evidențiază corelația dintre cei doi parametrii.

3.3.2 Predicție

Vom începe prin a da intervalul de încredere pentru răspunsul mediu care reiese din prima parte a Prop. 3.9:

Prop. 3.16



Intervalul de încredere pentru $\mathbb{E}[y_i|x_i] = \beta_0 + \beta_1 x_i$ este

$$\left[\hat{y}_i - t_{n-2}(1 - \alpha/2)\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}}, \hat{y}_i + t_{n-2}(1 - \alpha/2)\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}} \right]$$

unde $t_{n-2}(1 - \alpha/2)$ este cantila de ordin $1 - \alpha/2$ a repartiției Student t_{n-2} .

Atunci când vorbim de predicție vom dori să determinăm un *interval de predicție* pentru valoarea rezultată din model. Se numește *interval de predicție* de nivel $1 - \alpha$ pentru o variabilă aleatoare neobservată Y date fiind datele X intervalul aleator $[A(X), B(X)]$ care verifică

$$\mathbb{P}_\theta(A(X) \leq Y \leq B(X)) \geq 1 - \alpha, \quad \forall \theta.$$

Se observă similaritatea dintre un interval de predicție și un interval de încredere, diferența fiind dată că un interval de predicție este un interval pentru o variabilă aleatoare și nu pentru un parametru (care este o

constantă). Având în vedere variabilitatea v.a. Y este de așteptat ca intervalul de predicție să fie mai mare dacă cel de încredere la același nivel de încredere.

În cazul modelului gaussian (termenii eroare sunt repartizați normal), observăm, ținând cont de liniaritatea lui \hat{y}_{n+1} în $\hat{\beta}_0, \hat{\beta}_1$ și ε_{n+1} , că

$$y_{n+1} - \hat{y}_{n+1} \sim \mathcal{N} \left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \right)$$

și cum varianta σ^2 nu este cunoscută în general aceasta este estimată prin $\hat{\sigma}^2$. De asemenea, notând că variabilele aleatoare $y_{n+1} - \hat{y}_{n+1}$ și $\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$ sunt independente avem următorul rezultat referitor la intervalul de predicție.

Prop. 3.17



Avem că

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}}} \sim t_{n-2}$$

ceea ce conduce la un interval de predicție de nivel $1 - \alpha$ pentru y_{n+1}

$$\left[\hat{y}_{n+1} \pm t_{n-2}(1 - \alpha/2) \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \right],$$

unde $t_{n-2}(1 - \alpha/2)$ este cuantila de ordin $1 - \alpha/2$ a repartiției Student t_{n-2} .

Observăm că pe măsură ce punctul de prezis admite o abscisă x_{n+1} mai depărtată de media \bar{x} cu atât intervalul de predicție este mai mare. Mai mult, putem constata că atunci când x_{n+1} variază, curba descrisă de capetele intervalului este o hiperbolă de axe $x = \bar{x}$ și $y = \hat{\beta}_0 + \hat{\beta}_1 x$. Pentru a observa acest lucru putem face schimbarea de variaibile $u = x - \bar{x}$ și $v = y - \hat{\beta}_0 + \hat{\beta}_1 x$ de unde concluzionăm că un punct de coordonate (u, v) se află în regiunea de mai sus dacă

$$\frac{u^2}{a^2} - \frac{v^2}{b^2} \leq 1$$

unde $a^2 = (1 + \frac{1}{n}) \sum (x_i - \bar{x})^2$ și $b^2 = (1 + \frac{1}{n})(t_{n-2}(1 - \alpha/2)\hat{\sigma})^2$. În mod similar se poate arăta și că forma curbei descrisă de capetele intervalului de încredere a răspunsului mediu este tot o hiperbolă.

De asemenea sunt multe circumstanțele în care am dori să avem intervalele de încredere pentru răspunsul mediu respectiv intervalele de predicție în mai mult de un punct, prin urmare ne aflăm în cadrul unei probleme de inferență simultană. O soluție la această problemă, în cazul în care avem m puncte, este data de inegalitatea lui Bonferroni ($\mathbb{P}(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n \mathbb{P}(A_i)$) care conduce la marginea

$$\hat{y}_i - t_{n-2} \left(1 - \frac{\alpha}{2m} \right) \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}} \leq \beta_0 + \beta_1 x_i \leq \hat{y}_i + t_{n-2} \left(1 - \frac{\alpha}{2m} \right) \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}}$$

pentru răspunsul mediu $\mathbb{E}[y_i|x_i] = \beta_0 + \beta_1 x_i$, $i \in \{n+1, \dots, n+m\}$, iar

$$\hat{y}_i - t_{n-2} \left(1 - \frac{\alpha}{2m} \right) \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}} \leq y_i \leq \hat{y}_i + t_{n-2} \left(1 - \frac{\alpha}{2m} \right) \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}}$$

pentru valoarea prezisă y_i , $i \in \{n+1, \dots, n+m\}$ conform modelului, unde $t_{n-2}(1 - \alpha/2m)$ este cuantila de ordin $1 - \alpha/2m$ a repartiției Student t_{n-2} . Marginea de mai sus este adevărată pentru m puncte dar dacă

am să găsim o relație similară pentru orice x atunci această metodă nu ar mai putea fi aplicată. Scheffe a propus următoarea margină numită și banda lui Scheffe:

Prop. 3.18



În contextul problemei de regresie liniare simple cu erori repartizare gaussiană are loc următoarea relație, pentru răspunsul mediu

$$\mathbb{P} \left(\hat{y} - M_\alpha \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_j - \bar{x})^2}} \leq \beta_0 + \beta_1 x \leq \hat{y} + M_\alpha \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_j - \bar{x})^2}}, \quad \forall x \right) \geq 1 - \alpha$$

și pentru valorile prezise

$$\mathbb{P} \left(\hat{y} - M_\alpha \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_j - \bar{x})^2}} \leq y \leq \hat{y} + M_\alpha \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_j - \bar{x})^2}}, \quad \forall x \right) \geq 1 - \alpha$$

unde $M_\alpha = \sqrt{2F_{2,n-2}(1-\alpha)}$ iar $F_{2,n-2}(1-\alpha)$ este cuantila de ordin $1 - \alpha$ a repartiției Fisher-Snedecor $F_{2,n-2}$.

Exemplu: prețul chiriilor în München

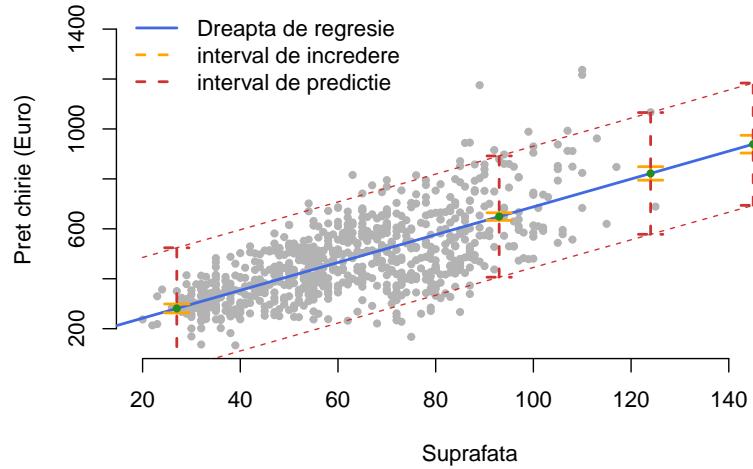
Exp. 3.19

Revenind la exemplul modelului de regresie liniară simplă introdus pe setul de date **München** să presupunem că ne interesăm la prețul de închiriere al apartamentelor care au suprafața de 27, 93, 124 și respectiv 145 de metrii pătrați. Tabelele următoare prezintă valorile prezise și intervalele de încredere (stânga) respectiv de predicție (dreapta) pentru noile suprafețe.

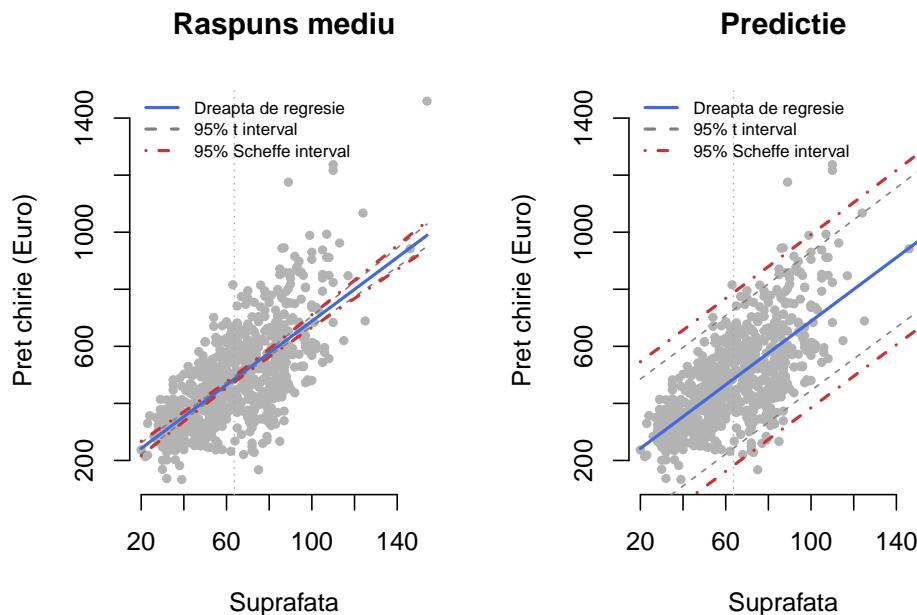
Suprafața	Valori prezise	CI_Inf	CI_Sup
27	281.0942	263.2530	298.9355
93	649.0816	633.8394	664.3237
124	821.9241	794.9069	848.9413
145	939.0110	903.4809	974.5411

Suprafața	Valori prezise	Pred_Inf	Pred_Sup
27	281.0942	38.08053	524.1079
93	649.0816	406.24485	891.9183
124	821.9241	578.06498	1065.7832
145	939.0110	694.06255	1183.9594

În figura de mai jos avem ilustrate intervalele de încredere pentru răspunsul mediu (cu portocaliu) și respectiv intervalele de predicție (cu roșu) pentru cele patru valori ale variabilei explicative. Se poate observa diferența mare între incertitudinea dată de cele două intervale.



Dacă facem referire la răspunsul mediu și respectiv la valorile prezise atunci putem trasa banda de încredere corespunzătoare intervalelor de încredere din Prop. 3.16 și 3.17 respectiv banda lui Scheffe din Prop. 3.18:



3.4 Aplicații numerice

Ex. 3.20



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

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

Obținem următoarele rezultate numerice

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

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

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

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

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

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

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

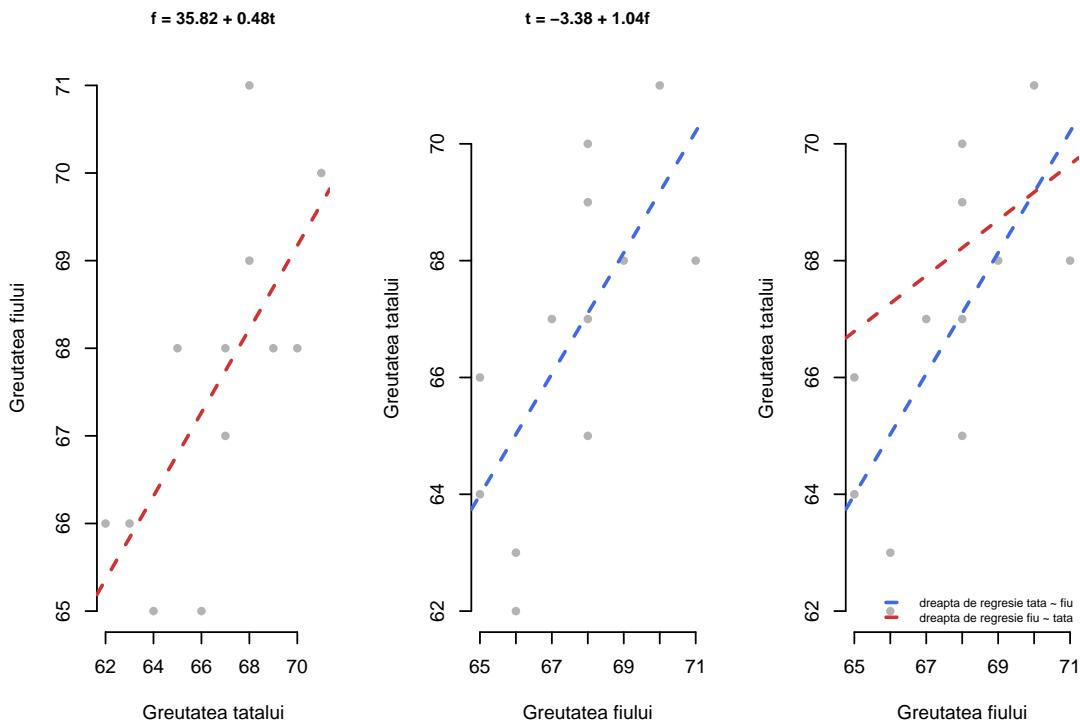
În cazul problemei, coeficienții sunt $\hat{\beta}_0 = -3.38$ și $\hat{\beta}_1 = 1.03$ iar dreapta de regresie este $t = -3.38 + 1.03f$ (a se vedea figura din mijloc).

3. Produsul pantelor celor două drepte este

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

și conform Prop. 3.10 și a definiției coeficientului de determinare avem

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



Ex. 3.21



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

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

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

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

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

și respectiv

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

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

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

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

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

Prin urmare, sub H_0 avem că

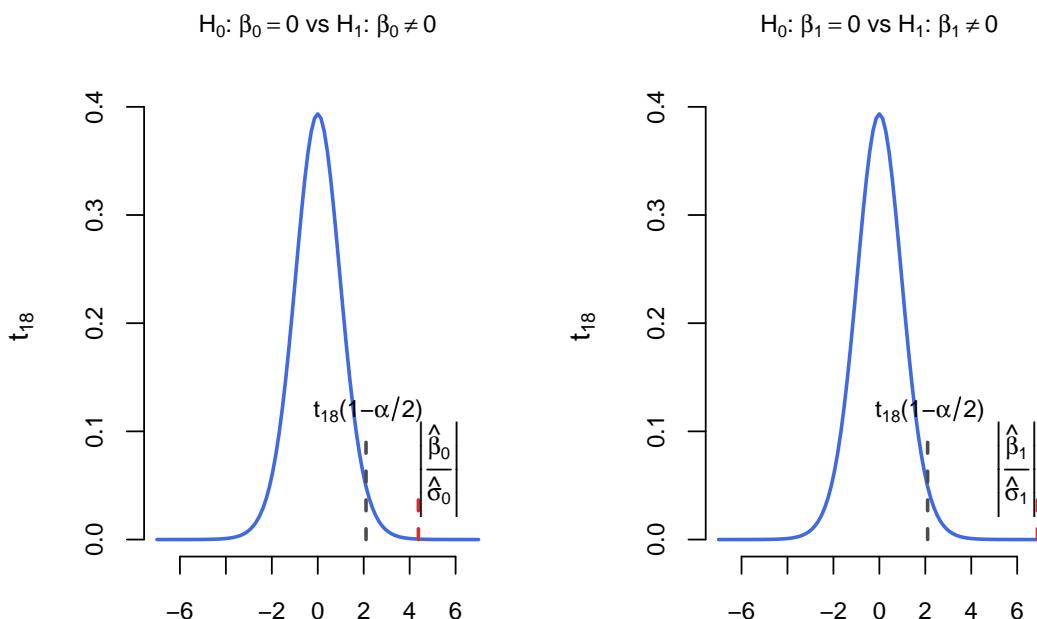
$$\frac{\hat{\beta}_0}{\hat{\sigma}_0} \sim t_{18},$$

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

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

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

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



4 Modelul de regresie liniară multiplă

4.1 Introducere

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

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

În realitate, relația de mai sus nu este exactă (i.e. \approx) deoarece se află sub influența unei erori aleatoare ε . Presupunând că erorile sunt aditive avem o relație de tipul

$$y = g(x_1, \dots, x_p) + \varepsilon.$$

Astfel având dat un eşantion de talie n de $(p+1)$ -upluri (\mathbf{x}_i, y_i) cu $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$, $i \in \{1, \dots, n\}$, $n > p$, ne propunem să determinăm g , i.e. să separăm componenta sistematică g (influența covariabilelor asupra răspunsului mediu) de termenul eroare ε (incertitudinea modelului). Aproximarea, \approx , din relația anterioară se poate descrie matematic sub forma

$$g = \arg \min_{f \in \mathcal{F}} \sum_{i=1}^n L(y_i - f(\mathbf{x}_i))$$

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

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

În ceea ce privește clasa de funcții \mathcal{F} ce descriu componenta sistematică a modelului, în contextul modelului liniar vom considera clasa de funcții liniare (i.e. componenta sistematică este o combinație liniară de covariabile)

$$\mathcal{F} = \{f : \mathbb{R}^p \rightarrow \mathbb{R} \mid f(\mathbf{x}) = (1 \quad \mathbf{x}) \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}.$$

Este important de menționat că atunci când vorbim de *regresie liniară* ne referim la liniaritatea în *parametrii modelului* β_j și nu în variabilele explicative.

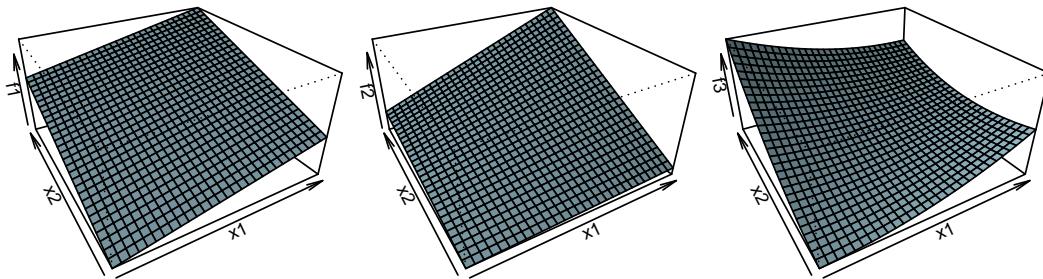
De exemplu, în contextul modelului de regresie liniară simplă, următoarele modele $y = \beta_0 + \beta_1 x^2 + \varepsilon$, $y = \beta_0 + \beta_1 \log(x) + \varepsilon$ și $y = \beta_0 + \beta_1 (x^3 - \cos(x)^2 + e^x) + \varepsilon$ sunt liniare chiar dacă variabila predictor x poate prezenta un efect neliniar asupra răspunsului mediu. Pe de altă parte, modelul $y = \beta_0 + \beta_1 x^{\beta_2} + \varepsilon$ nu este un model liniar.

De asemenea, $f_1(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, $f_2(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 x_2}_{x_3}$ sau $f_3(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 x_2}_{x_3} + \beta_4 \underbrace{x_1^2}_{x_4} + \beta_5 \underbrace{x_2^2}_{x_5}$ sunt toate funcții liniare în $\boldsymbol{\beta}$.

$$f_1(x_1, x_2) = -1 + 2x_1 + 3x_2$$

$$f_2(x_1, x_2) = x_1 + 3x_2 + 2x_1x_2$$

$$f_3(x_1, x_2) = 3 + x_1 + 8x_2 - 4x_1x_2 + 2x_1^2 + 3x_2^2$$



4.2 Modelare

Modelul de regresie liniară multiplă reprezintă o extensie a modelului de regresie liniară simplă atunci când numărul variabilelor explicative este finit (pentru mai multe detalii privind analiza de regresie se pot consulta monografile [Seber and Lee, 2003], [Rencher and Schaalje, 2008], [Weisberg, 2014] sau [Faraway, 2015]).

Pentru a estima parametrii necunoscute β , colectăm datele (y_i, \mathbf{x}_i) și presupunem că acestea verifică modelul

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

unde

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

Scris sub formă compactă (matriceală) modelul devine

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

unde

- \mathbf{X} se numește *matricea de design*

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

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

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1) \times 1} \quad \text{și} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n \times 1}.$$

Coloana cu termeni de 1 din matricea de design \mathbf{X} corespunde termenului liber (ordonatei la origine - intercept). Un model foarte simplu, numit și *modelul nul* este cel în care nu apar variabile explicative $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}_{n \times 1} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n \times 1}.$$

În cele ce urmează vom presupune că matricea de design \mathbf{X} are rangul $\text{rang}(\mathbf{X}) = p+1$ acest fapt implicând că vectorii coloană sunt liniar independenți. O condiție necesară pentru această ipoteză este ca talia eșantionului n să fie mai mare decât numărul de covariabile din model. Această presupunere este necesară pentru a estima în mod unic coeficienții modelului. În cazul în care matricea de design nu este de rang $p+1$ atunci ne confruntăm cu o problemă de identifiabilitate a modelului.

Def. 4.1



Un model de regresie liniară multiplă este definit de relația

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

unde

- \mathbf{Y} este un vector aleator de dimensiune n
- \mathbf{X} este o matrice de dimensiune $n \times (p+1)$ numită matrice de design
- $\boldsymbol{\beta}$ este un vector de dimensiune $p+1$ de parametrii necunoscuți ai modelului
- $\boldsymbol{\varepsilon}$ este vectorul aleator centrat al erorilor

Pentru a specifica în totalitate *modelul clasic de regresie liniară* trebuie să facem următoarele ipoteze referitoare la componenta sistematică a modelului și la vectorul eroare. În primul rând, dacă ne raportăm la tipul variabilelor explicative sau la tipul de experiment care a condus la observarea datelor (e.g. experiment planificat), putem distinge între două situații: covariabilele sunt deterministe, cum este cazul unui plan de experiență planificat, sau sunt aleatoare atunci când avem de-a face cu date observaționale. În cel de-al doilea caz, când regresorii sunt aleatori, datele trebuie înțelese ca realizări ale unui vector aleator iar modelul trebuie văzut ca unul condiționat (la matricea de design $\mathbf{X} = x$), a se vedea secțiunea de Introducere.

În al doilea rând, dacă facem referire la termenii eroare ε_i (cei care descriu incertitudinea modelului) aceștia trebuie să fie de medie 0, $\mathbb{E}[\varepsilon_i] = 0$ sau sub formă matriceală $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$, să aibă varianță constantă (homoscedasticitate) $\text{Var}(\varepsilon_i) = \sigma^2$ și să fie necorelate $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ pentru $i \neq j$. Ipoteza de homoscedasticitate și necorelare poate fi scrisă sub formă matriceală $\text{Cov}(\boldsymbol{\varepsilon}) = \text{Var}(\boldsymbol{\varepsilon}) = \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \sigma^2 I_n$. Să remarcăm că în contextul modelului condiționat aceste ipoteze se scriu sub formă $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$ și respectiv $\text{Cov}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2 I_n$ dar pentru ușurință vom elimina condiționarea din notație aceasta fiind subînțeleasă din context.

Sumarizând, ipotezele modelului de regresie liniară multiplă sunt:

$$(\mathcal{H}) \left\{ \begin{array}{l} \mathcal{H}_1 : \text{rang}(\mathbf{X}) = p+1 \\ \mathcal{H}_2 : \mathbb{E}[\boldsymbol{\varepsilon}] = 0, \text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n \end{array} \right.$$

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

Înainte de a trece la o discuție asupra ipotezelor modelului de regresie liniară multiplă vom introduce o serie de notații utile în cele ce vor urma. Pentru a face distincția dintre parametrii modelului β și σ^2 și estimatorii lor vom folosi, ca de obicei, notația cu căciulă, respectiv $\hat{\beta}$ și $\hat{\sigma}^2$, în general fiind imposibil să determinăm valorile adevărate ale parametrilor. Folosind $\hat{\beta}$ ca estimator al lui β putem determina cu ușurință estimatorul $\widehat{\mathbb{E}[y_i]}$ al mediei lui y_i ,

$$\widehat{\mathbb{E}[y_i]} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} = \mathbf{x}_i^T \hat{\beta},$$

unde am considerat $\mathbf{x}_i^T = (1, x_{i1}, \dots, x_{ip})$. Acest estimator, numit și valoare ajustată, se mai notează și cu \hat{y}_i , i.e. $\hat{y}_i = \widehat{\mathbb{E}[y_i]}$ iar eroare estimată, adică diferența dintre valoarea reală y_i și cea estimată \hat{y}_i , se numește valoare reziduală și se notează cu $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \mathbf{x}_i^T \hat{\beta}$. Este important de notat că valorile reziduale (reziduurile) $\hat{\varepsilon}_i$ nu sunt identice cu termenii eroare ε_i ei pot fi interpretați ca estimări (sau mai precis predicții) ai acestora, care, ca și vectorul de parametrii β , sunt necunoscuți. Reziduurile conțin variabilitatea din date (incertitudinea modelului) care nu poate fi explicată prin intermediul variabilelor explicative.

4.2.1 Liniaritatea efectelor variabilelor explicative

Chiar dacă într-o primă etapă ipoteza de liniaritate a efectelor covariabilelor pare o ipoteză restrictivă, am văzut că modelul de regresie liniară permite și relații neliniare. Spre exemplu, dacă luăm în considerare următorul model de regresie în care covariabila z_i admite un efect logaritmic,

$$y_i = \beta_0 + \beta_1 \log(z_i) + \varepsilon_i$$

atunci definind variabila $x_i = \log(z_i)$ regăsim un model de regresie liniară simplă $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Trebuie menționat că într-un model de regresie liniară putem avea relații neliniare între variabila răspuns și predictori atât timp cât este păstrată liniaritatea în parametrii modelului. Pentru modelare relațiilor neliniare vom prezenta în secțiunea [Interpretarea efectului variabilelor explicative](#) două abordări: una bazată pe transformarea variabilelor explicative și cealaltă prin intermediul polinoamelor.

Exemplu: înălțimea arborilor de eucalipt

Exp. 4.2

În acest exemplu facem referire la setul de date [Eucalypt](#) (care poate fi descărcat de [aici](#)) și dorim să modelăm relația dintre înălțimea (medie) arborilor de eucalipt și circumferința acestora. Modelul de regresie liniară simplă este dat de

$$ht_i = \beta_0 + \beta_1 circ_i + \varepsilon_i, \quad i = 1, \dots, 1429$$

care, prin aplicarea metodei celor mai mici pătrate, conduce la estimatorii $\hat{\beta}_0 = 9.037$ și respectiv $\hat{\beta}_1 = 0.257$ pentru coeficienții modelului și astfel

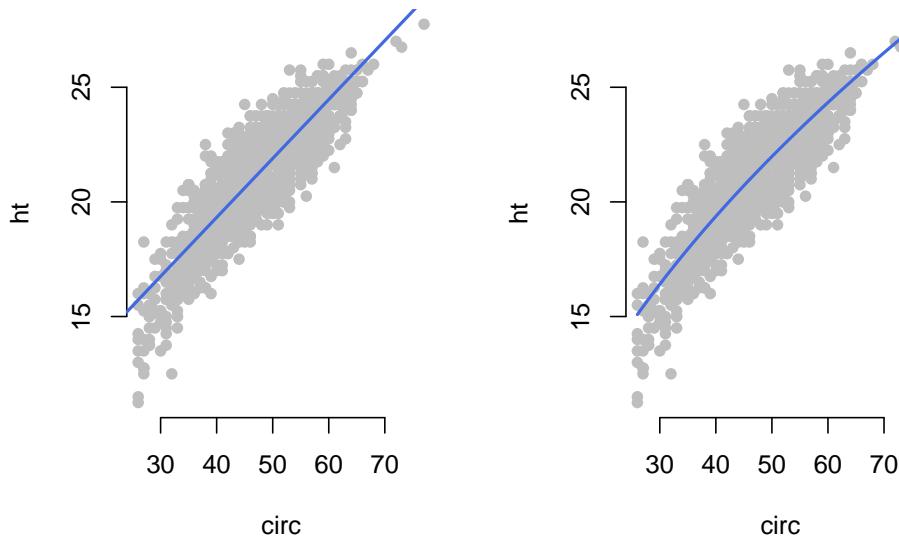
$$\hat{ht}_i = 9.037 + 0.257 circ_i, \quad i = 1, \dots, 1429.$$

Observăm că pentru o creștere în circumferință de un 1 cm avem o creștere medie în înălțime de 25.7 cm. De asemenea putem remarca din figura de mai jos imaginea din stânga, că pentru valori mici ale circumferinței arborilor acestea se regăsesc sub dreapta de regresie ceea ce sugerează o relație neliniară între înălțime și circumferință. Astfel, putem considera un model simplu de tipul

$$ht_i = \beta_0 + \beta_1 \sqrt{circ_i} + \varepsilon_i,$$

care ar conduce la o mai bună ajustare pe date (coeficientul de determinare $R^2 = 0.7820638$ este un pic mai mare față de cel din modelul de regresie simplă $R^2 = 0.7683202$) și la estimatorii $\hat{\beta}_0 = -2.73$ și respectiv $\hat{\beta}_1 = 3.494$,

$$\hat{ht}_i = -2.73 + 3.494\sqrt{circ_i}, \quad i = 1, \dots, 1429. \quad \square$$



Exemplu: prețul chiriilor din Munchen

Exp. 4.3

Raportându-ne la setul de date referitor la prețul chiriilor în München pentru apartamentele dintr-o locație medie, construite după anul 1966 (date ce pot fi descărcate de [aici](#)) observăm că modelul de regresie liniară simplă dintre prețul chiriilor pe metru pătrat și suprafață

$$pret_m^2_i = \beta_0 + \beta_1 suprafata_i + \varepsilon_i$$

devine în urma estimării

$$pret_m^2_i = 10.52 - 0.041 suprafata_i + \varepsilon_i$$

ceea ce ar sugera că o relație neliniară între variabila răspuns și variabila explicativă ar fi mai potrivită (același lucru se observă și din imaginea din stânga a figurii de mai jos). O mai bună ajustare pe date se poate obține definind variabila explicativă $x = \frac{1}{suprafata}$ care ar conduce la modelul de regresie

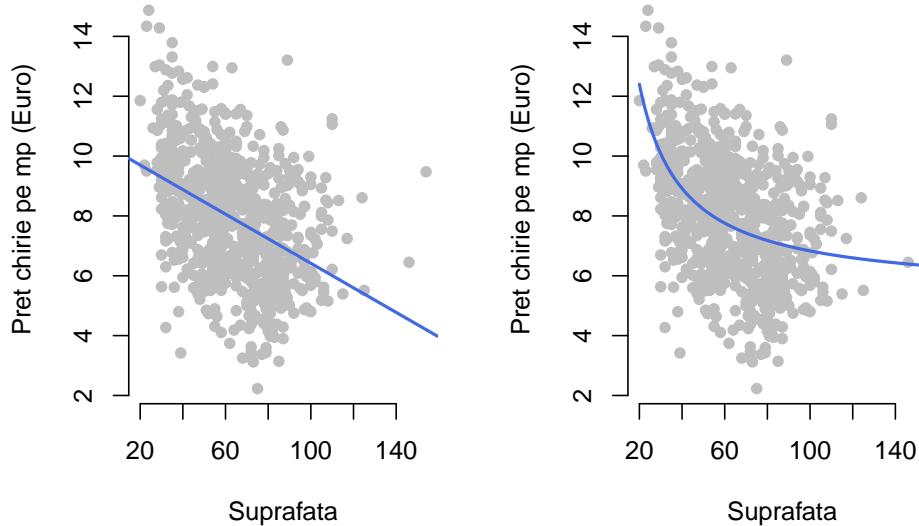
$$pret_m^2_i = \beta_0 + \beta_1 \frac{1}{suprafata_i} + \varepsilon_i.$$

Aplicând metoda celor mai mici pătrate obținem

$$\widehat{pret_m^2}_i = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{suprafata_i} = 5.438 + 139.269 \frac{1}{suprafata_i}$$

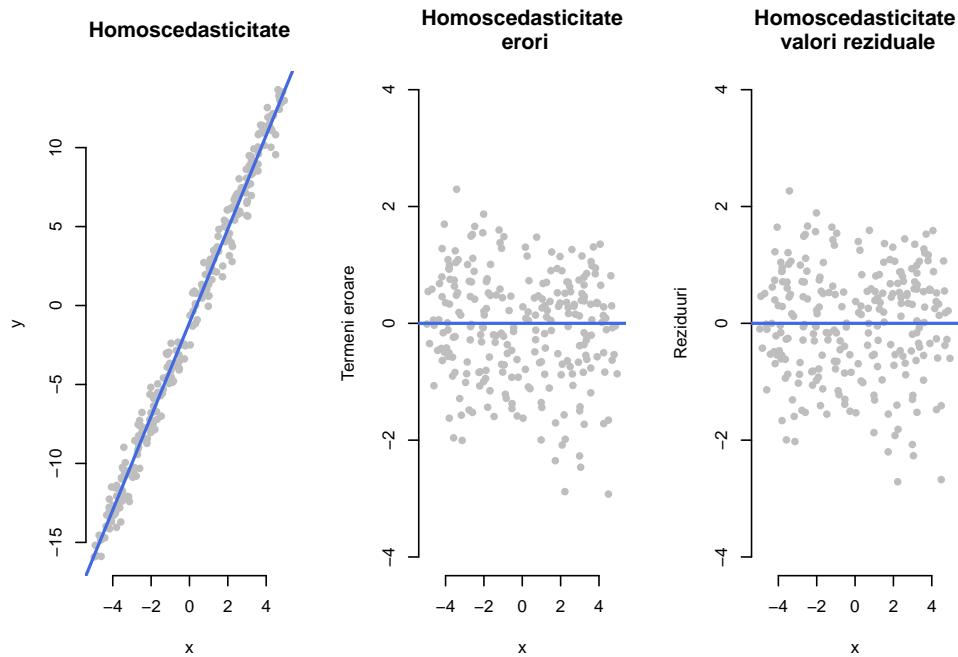
ceea ce arată că prețul mediu pe m^2 scade neliniar odată cu creșterea suprafeței apartamentului (a se vedea imaginea din dreapta din figura de mai jos în care putem constata o mai bună ajustare a modelului pe date). De exemplu, diferența în medie dintre prețul pe m^2 al unui apartament de $35 m^2$ față de unul de $36 m^2$ este

$$\widehat{\text{pret_}m^2}(35) - \widehat{\text{pret_}m^2}(36) = \frac{139.269}{35} - \frac{139.269}{36} \approx 0.110531 \text{ Euro} \quad \square$$

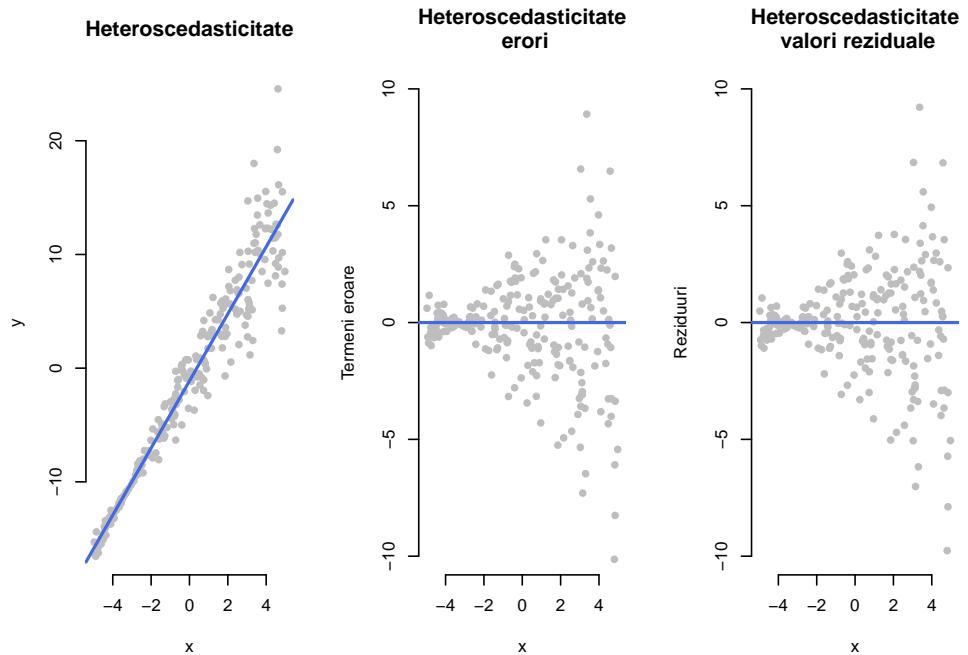


4.2.2 Homoscedasticitatea termenilor eroare

Proprietatea de homoscedasticitate a erorilor (ipoteza H_2) presupune că varianța lui ε_i nu variază în mod sistematic odată cu creșterea sau descreșterea valorilor variabilelor explicative. În figura de mai jos avem ilustrat un model simulat homoscedastic după relația $y_i \sim \mathcal{N}(-1 + 3x_i, 1)$ unde $x_i \sim \mathcal{U}[-5, 5]$ împreună cu dreapta de regresie (figura din stânga). Se poate constata că observațiile fluctuează constant în jurul dreptei de regresie (mediei), aceeași concluzie putând fi trasă și din examinarea graficului în care apar termenii eroare (figura din mijloc). Cum datele sunt simulate și cunoaștem adevăratul model (cunoaștem adevăratele valori ale coeficientilor), termenii eroare sunt $\varepsilon_i = y_i - \mathbf{x}_i^\top \boldsymbol{\beta}$ și se cunosc. În general, vom folosi valorile reziduale pentru a estima termenii eroare $\hat{\varepsilon}_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ și pentru a verifica proprietatea de homoscedasticitate a erorilor (figura din dreapta). Vom vedea, că atât termenii eroare cât și valorile reziduale au medie zero dar spre deosebire de termenii eroare, reziduurile pot avea varianțe diferite și în plus sunt corelate.



Un exemplu, des întâlnit în practică, de situație în care termenii eroare sunt heteroscedastici este ilustrat în figura următoare. Modelul folosit pentru generarea observațiilor este $y_i \sim \mathcal{N}(-1 + 3x_i, (1.2 + 0.5(x_i + 1))^2)$ unde $x_i \sim \mathcal{U}[-5, 5]$. Se poate constata că pe măsură ce x crește, variabilitatea observațiilor în jurul dreptei de regresie crește de asemenea (figura din stânga). Acest fenomen (de pănie) se poate observa cu mai mare ușurință atunci când trasăm graficul erorilor (respectiv valorilor reziduale).



Conform teoriei, dacă modelul de regresie este corect (componenta sistematică a modelului este cea reală) și ipotezele modelului sunt verificate atunci toate graficele care implică valorile reziduale ($\hat{\varepsilon}$) versus orice funcție ce depinde de covariabile (sau de o combinație liniară a acestora) ar trebui să ilustreze o variație

simetrică și constantă pe verticală ($\hat{\varepsilon}$), prin urmare este recomandată inspectarea a cât mai multor astfel de grafice. Cele mai uzuale grafice de reziduuri (*residual plots*) sunt cele care implică valorile reziduale $\hat{\varepsilon}_i$ și variabilele explicative x_i care sunt incluse în modelul propus (sau nu - în acest caz orice structură observată în grafic poate sugera includerea covariabilei în model) respectiv valorile ajustate \hat{y}_i . Acest din urmă grafic de diagnostic ($\hat{\varepsilon}_i$ vs \hat{y}_i) este și cel mai des folosit în practică deoarece conține informații de la toate covariabilele incluse în model [Faraway, 2015]. Aceste grafice permit de asemenea detectarea neliniarității în componenta sistematică, de structură, a modelului.

Este important de menționat că sunt multe situațiile (i.e. atunci când avem de-a face cu un model nelinier) în care nu putem trage o concluzie asupra ipotezelor modelului doar privind comportamentul valorilor reziduale. De cele mai multe ori trebuie să avem grijă ca modelul specificat să fie corect. Următorul exemplu sugerează acest fapt:

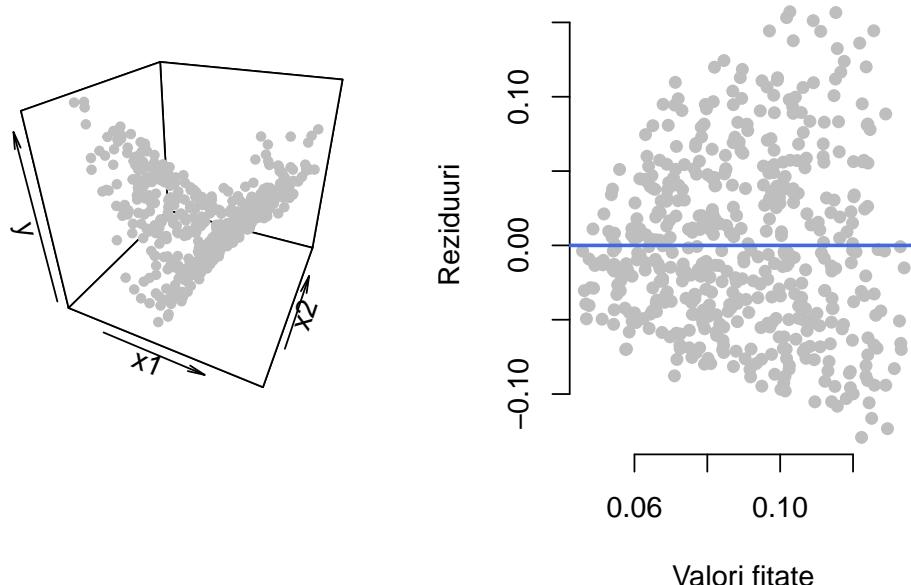
Exemplu: importanța corectitudinii modelului - date simulate

Exp. 4.4

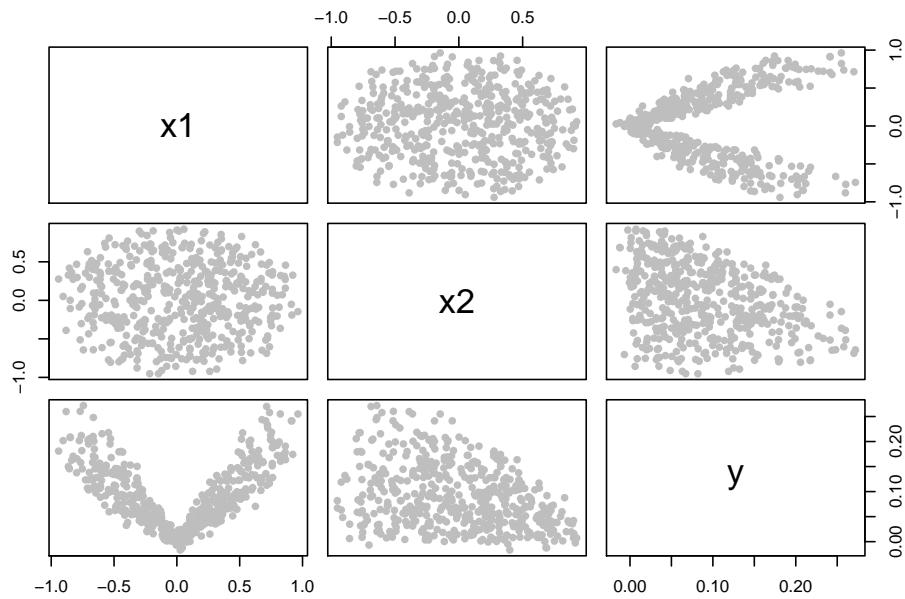
Acet exemplu este preluat din [Cook and Weisberg, 1999] și evidențiază importanța corectitudinii modelului atunci când este evaluat un grafic al reziduurilor. În figura de mai jos avem ilustrat grafic diagrama de împrăștiere a punctelor precum și evoluția valorilor reziduale în raport cu valorile ajustate pentru modelul de regresie liniară multiplă:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

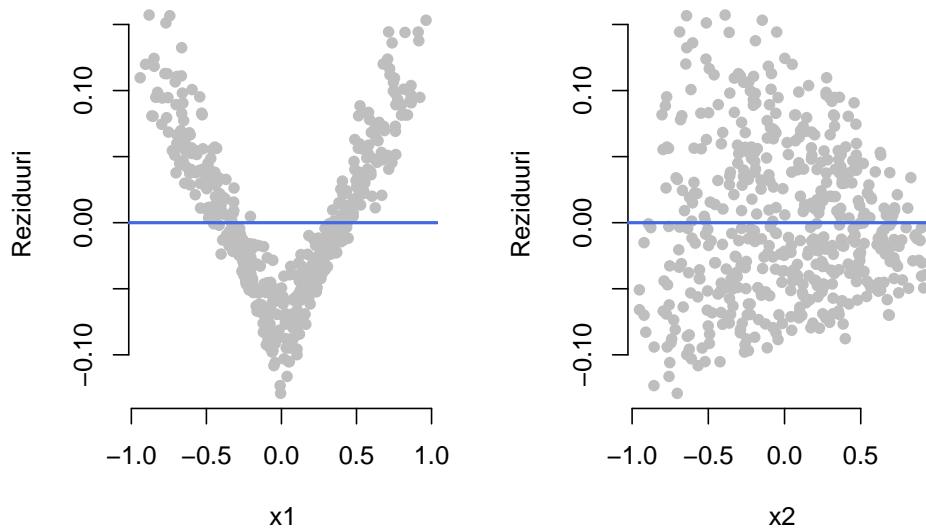
unde observațiile au fost simulate artificial ($n = 500$). Efectul de pâlnie al reziduurilor (figura din dreapta) sugerează heteroscedasticitatea termenilor eroare în modelul propus, atât timp cât acest model este valid.



Dacă ne uităm la diagramele de împrăștiere pentru fiecare variabilă în parte (matricea diagramelor de împrăștiere - scatterplot matrix) observăm că variabilele predictor x_1 și x_2 sunt necorelate (coeficientul de corelație este -0.022) și sunt generate sferic (mai exact repartitia comună a predictorilor face parte din clasa repartițiilor elitice simetrice - $\mathbb{E}[x_i|x_j] \approx \alpha_0 + \alpha_1 x_j$).



Problema care apare este că modelul de regresie este specificat greșit, componenta sistematică a modelului nu este liniară sau o funcție care să depindă de o combinație liniară de covariabile (i.e. $g(\mathbf{x}^T \boldsymbol{\beta})$), chiar dacă din graficul valorilor reziduale am crede că avem o problemă cu heteroscedasticitatea erorii. În acest exemplu particular, în care avem doar doi predictori, se putea observa, prin trasarea diagramei de împrăștiere în trei dimensiuni (figura de mai sus - stânga), că răspunsul mediu nu depinde liniar de x_1 și x_2 (în general există metode mai complexe de tipul *sliced inverse regression*). Acest lucru putea fi remarcat și prin trasarea valorilor reziduale în raport cu fiecare variabilă explicativă în parte.



Trebuie menționat că datele au fost generate după următorul model:

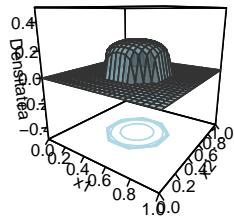
$$y = \underbrace{\frac{|x_1|}{2 + (1.5 + x_2)^2} + \varepsilon}_{= \mathbb{E}[y|x_1, x_2]}$$

unde termenul eroare $\varepsilon \sim \mathcal{N}(0, 0.01^2)$ are varianță constantă iar predictorii (x_1, x_2) au fost simulați dintr-o repartiție eliptică Pearson de tip II pe discul unitate în \mathbb{R}^2 ([Johnson, 2013]) având densitatea comună dată de

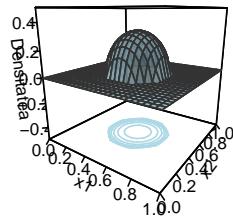
$$f_P(x_1, x_2) = \frac{\Gamma(m+2)}{\Gamma(m+1)\pi\sqrt{2}}(1-x_1^2-x_2^2)^m$$

unde $m > -1$ este un parametru de formă (pentru simulare am folosit $m = 0.5$). În figura de mai jos este ilustrată densitatea Pearson de tip II pentru mai multe valori ale parametrului de formă.

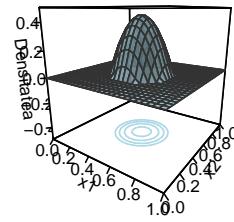
m = 0.2



m = 0.5



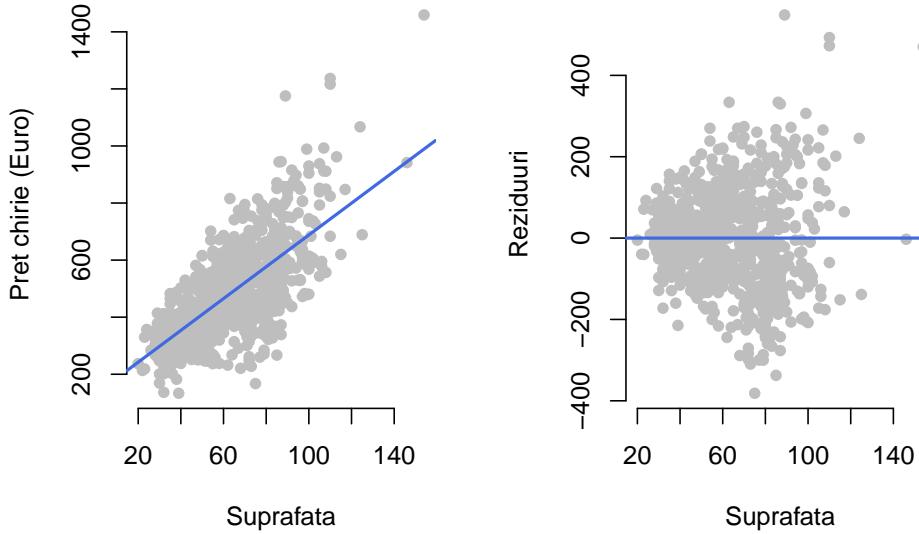
m = 1



Exemplu: prețul chiriilor din München

Exp. 4.5

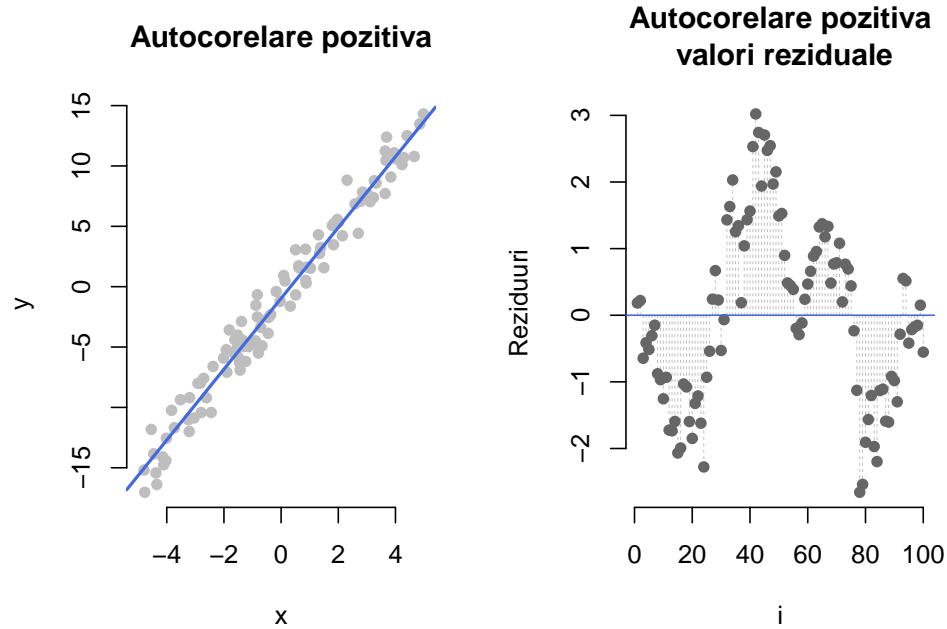
În cazul setului de date ce face referire la prețul chiriilor din München, dacă ne referim la relația dintre prețul net al chiriilor și suprafața de locuit (figura din stânga) constatăm o tendință de creștere a erorilor odată cu creșterea în suprafață a locuințelor (figura din stânga). Astfel, găsim că pentru suprafete mai mari avem o plajă mai largă de valori ale prețurilor chiriilor.



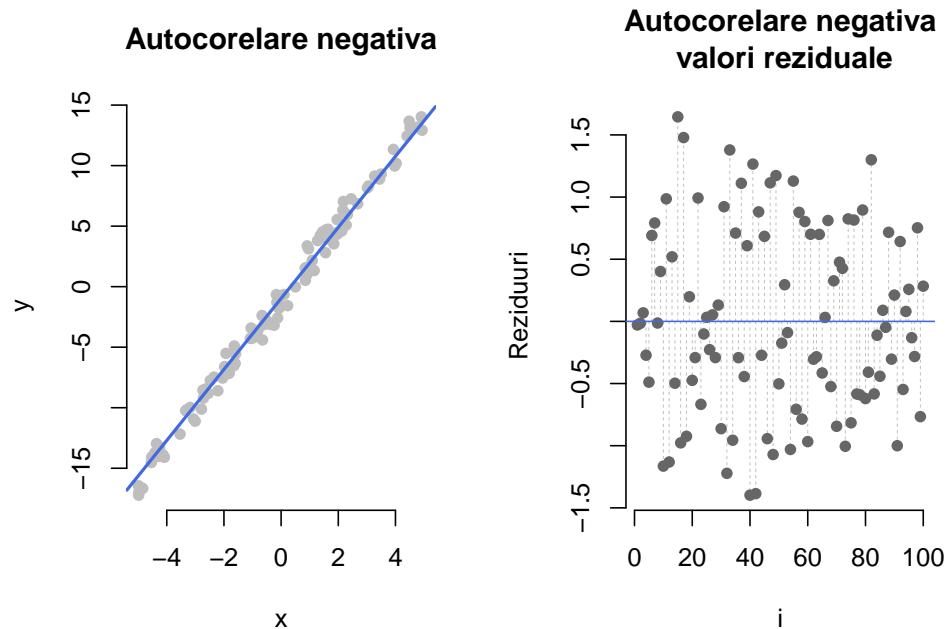
4.2.3 Necorelarea termenilor eroare

Ipoteza \mathcal{H}_2 presupune pe lângă homoscedasticitatea termenilor eroare și că aceștia sunt necorelați, i.e. $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ pentru $i \neq j$. Această ipoteză poate să nu fie conformă cu realitatea în special atunci când avem de-a face cu serii de timp sau date longitudinale (date care presupun eşantionarea acelorași unități de eşantionare la momente diferite de timp, e.g. urmărirea și eşantionarea lunară a unor caracteristici de interes - viremia - pentru o serie de pacienți cu o boală gravă în vederea caracterizării stării imunității acestora). În aceste situații se observă o autocorelare a termenilor eroare. De exemplu, o autocorelare de ordin unu a termenilor eroare implică o relație liniară între eroarea la momentul i și eroarea la momentul $i-1$, e.g. $\varepsilon_i = \rho\varepsilon_{i-1} + u_i$ unde u_i sunt variabile aleatoare independente și identic repartizate iar o autocorelare de ordin k se scrie sub forma $\varepsilon_i = \sum_{t=1}^k \rho_t \varepsilon_{i-t} + u_i$.

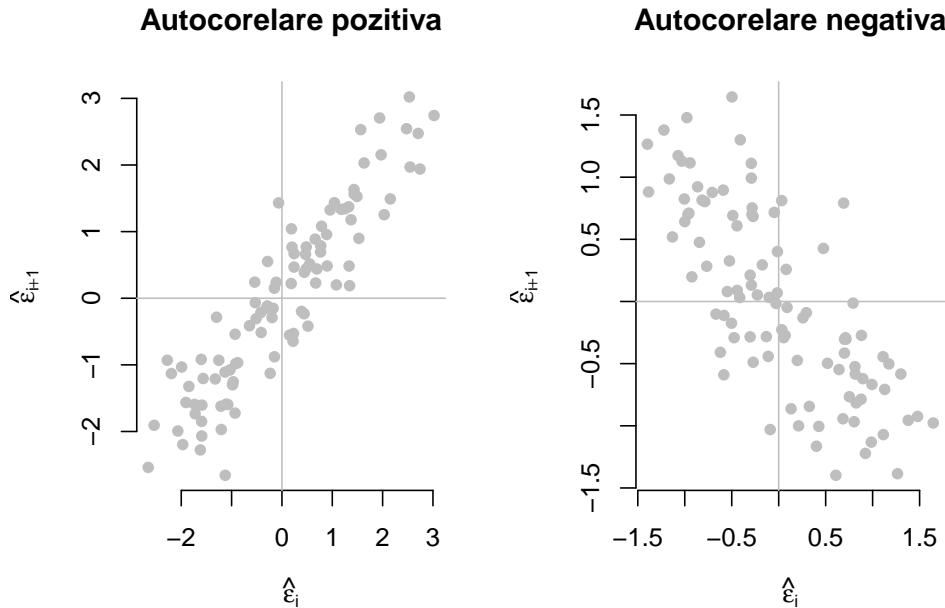
În figura de mai jos avem ilustrat un exemplu de model de regresie liniară simplă în care erorile prezintă o autocorelare (pozitivă) de ordinul 1. Modelul a fost generat după relația $y_i = -1 + 3x_i + \varepsilon_i$ unde $x_i \sim \mathcal{U}[-5, 5]$ iar $\varepsilon_i = 0.9\varepsilon_{i-1} + u_i$ cu $u_i \sim \mathcal{N}(0, 0.5^2)$. Se observă că autocorelarea termenilor eroare este *pozitivă* deoarece un termen eroare pozitiv (negativ) este mai probabil să fie urmat tot de un termen pozitiv (negativ) - figura din dreapta.



Următoarea figură ilustrează un exemplu clasic de autocorelare *negativă* a termenilor eroare. În această situație, termenii eroare pozitivi (negativi) sunt urmați cu precădere de termeni eroare negativi (pozitivi), prin urmare observăm frecvent o alternare a semnelor termenilor eroare. Modelul din figură este generat după relația $y_i = -1 + 3x_i + \varepsilon_i$ unde $x_i \sim \mathcal{U}[-5, 5]$ iar $\varepsilon_i = -0.9\varepsilon_{i-1} + u_i$ cu $u_i \sim \mathcal{N}(0, 0.5^2)$.



Dacă termenii eroare nu ar fi corelați ne-am aștepta ca graficul valorilor reziduale să prezinte o variație aleatoare în jurul dreptei orizontale $\varepsilon = 0$. O abordare alternativă de a investiga dacă erorile sunt corelate este de a trasa un grafic pentru perechile $(\hat{\varepsilon}_i, \hat{\varepsilon}_{i+1})$.

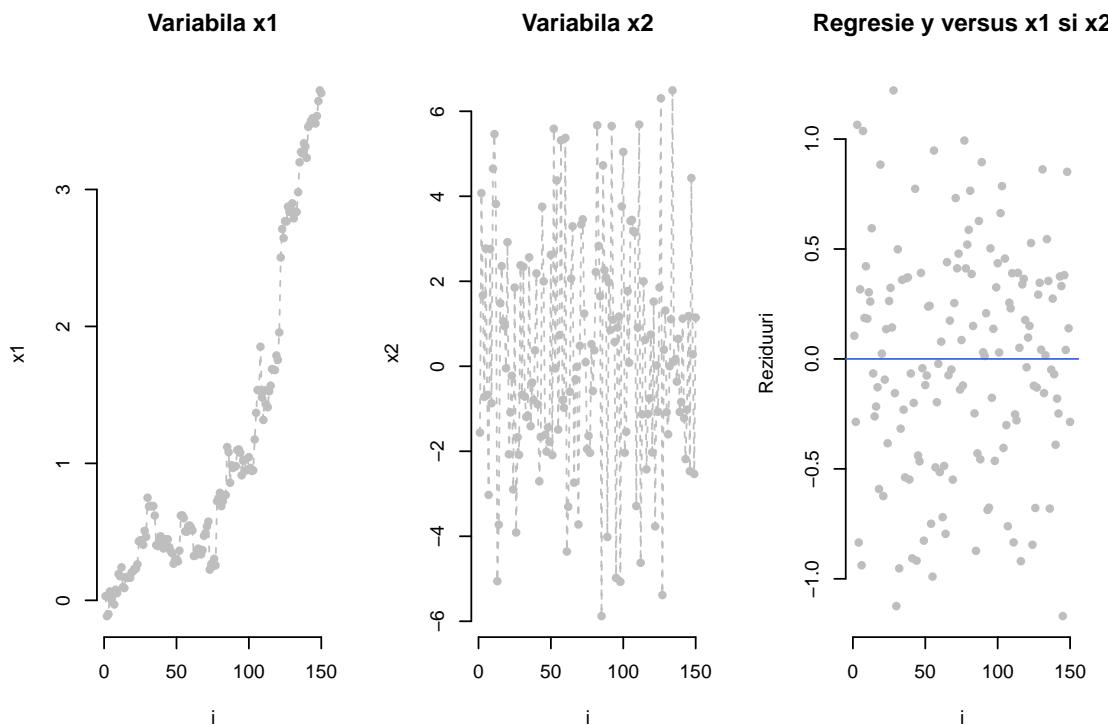


Autocorelarea erorilor apare de cele mai multe ori atunci când modelul de regresie nu este specificat corespunzător spre exemplu atunci când lipsește o variabilă explicativă din model sau efectul unei covariabile continue este neliniar. Cu toate acestea, sunt multe situațiile în care anumite variabile explicative care ar putea fi relevante pentru răspuns nu pot fi incluse în model deoarece nu au putut fi observate și în cazul în care acestea prezintă un caracter temporal ele pot induce o corelare a termenilor eroare. Vom ilustra acest fenomen printr-un exemplu simulat.

Exemplu: autocorelarea erorilor

Exp. 4.6

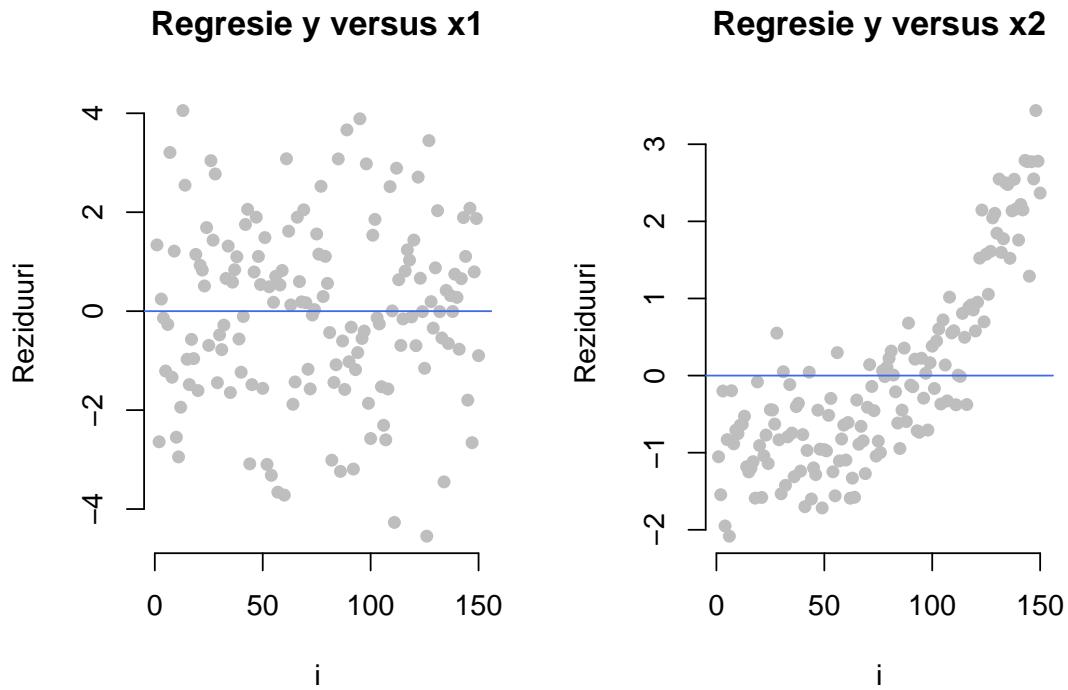
În figura de mai jos sunt ilustrate diagramele de împărtiere a variabilelor explicative x_1 (stânga) și respectiv x_2 (mijloc) unde se poate observa că prima covariabilă prezintă un caracter temporal ($x_{i1} = \sum_{k=1}^i a_k \sin(b_k)$ cu $a_k \sim \mathcal{N}(0, 0.7^2)$, $b_k \sim \mathcal{N}(0, 0.2^2)$ i.i.d și independente între ele) iar a doua fluctuează în jurul lui 0 ($x_2 \sim \mathcal{N}(0, 9)$). Variabila răspuns este generată după modelul $y_i = -1 + x_1 - 0.6x_2 + \varepsilon_i$ unde $\varepsilon_i \sim \mathcal{N}(0, 0.5^2)$ care verifică ipotezele unui model clasic de regresie liniară.



Modelul estimat obținut prin metoda celor mai mici pătrate este $\hat{y}_i = -0.997 + 1.031x_1 - 0.638x_2$ iar graficul valorilor reziduale (figura de mai sus dreapta) nu prezintă niciun indicu asupra autocorelării termenilor eroare.

Dacă presupunem că variabila predictor x_2 nu a fost observată atunci modelul de regresie liniară simplă estimat devine $\hat{y}_i = -1.237 + 1.064x_1$ iar graficul valorilor reziduale este ilustrat în figura de mai jos (stânga). Chiar dacă variabila x_2 lipsește din model, coeficienții modelului ajustat (redus) sunt apropiati de valorile adevărate care au fost folosite pentru generarea datelor iar graficul valorilor reziduale prezintă același caracter aleator al erorilor.

În cazul în care variabila predictor x_1 nu a fost observată atunci modelul ajutat devine $\hat{y}_i = 0.179 - 0.646x_1$ dar în acest caz graficul reziduurilor prezintă o autocorelare a erorilor (figura din dreapta). Motivul pentru care apare acest fenomen este că variabila explicativă x_1 prezenta un caracter temporal pe când covariabila x_2 nu iar efectele acestora, β_1x_1 respectiv β_2x_2 , au fost absorbite în termenul eroare. Mai exact, dacă variabila explicativă x_1 (respectiv x_2) este omisă din model atunci termenul eroare devine $\tilde{\varepsilon} = \beta_1x_1 + \varepsilon$ (respectiv $\tilde{\varepsilon} = \beta_2x_2 + \varepsilon$). Cum covariabila x_2 nu prezintă un trend temporal, efectul ei păstrează necorelarea termenului eroare. Pe de altă parte, atunci când efectul covariabilei x_1 , care prezintă un trend temporal, este absorbit în termenul eroare acest fapt este reflectat în autocorelarea valorilor reziduale.



4.3 Interpretarea efectului variabilelor explicative

În această secțiune ne propunem să explicăm și să interpretăm modul în care modelul ales și variabilele explicative influențează răspunsul mediu. Pentru aceasta vom considera situația în care covariabilele sunt continue, sunt discrete (categoriale) precum și modul în care acestea interacționează între ele.

4.3.1 Cazul variabilelor explicative continue

De cele mai multe ori atunci când avem de-a face doar cu variabile predictor continue este necesară ajustarea modelului de regresie astfel încât să se permită și relații/efekte neliniare între răspuns și covariabile. Am văzut în secțiunea [Liniaritatea efectelor variabilelor explicative](#) că acest lucru este permis în contextul modelului de regresie liniară, iar în această secțiune vom prezenta două abordări pentru obținerea acestui lucru: *transformarea covariabilelor* și *regresia polinomială*.

Transformarea variabilelor explicative permite înlăturarea efectelor neliniare în predictori și liniarizarea modelului. Dacă variabila explicativă z prezintă (aproximativ) un efect neliniar de tipul $\beta_1 \times f(z)$ unde f este o funcție cunoscută atunci modelul de regresie

$$y_i = \beta_0 + \beta_1 \times f(z_i) + \cdots + \varepsilon_i$$

poate fi transformat într-un model de regresie liniară

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

unde $x_i = f(z_i) - \frac{1}{n} \sum_{j=1}^n f(z_j)$. Operația de centrare face ca efectul estimat $\hat{\beta}_1 x$ să fie automat centrat în zero și permite o mai bună interpretare.

Printre cele mai eficiente transformări se numără transformările de tip puteri (power transformations) care au ca formă generală

$$x_i = f_\lambda(z_i) = \begin{cases} z_i^\lambda & , \lambda \neq 0 \\ \log(z_i) & , \lambda = 0 \end{cases}$$

unde $\lambda \in \Lambda = \{-1, -\frac{1}{2}, -\frac{1}{3}, 0, \frac{1}{3}, \frac{1}{2}, 1\}$ (scara puterilor). O inspecție vizuală prin trasarea matricei de diagrame de împrăștiere (scatterplot matrix) între predictori poate sugera ce transformări sunt necesare [Olive, 2017].

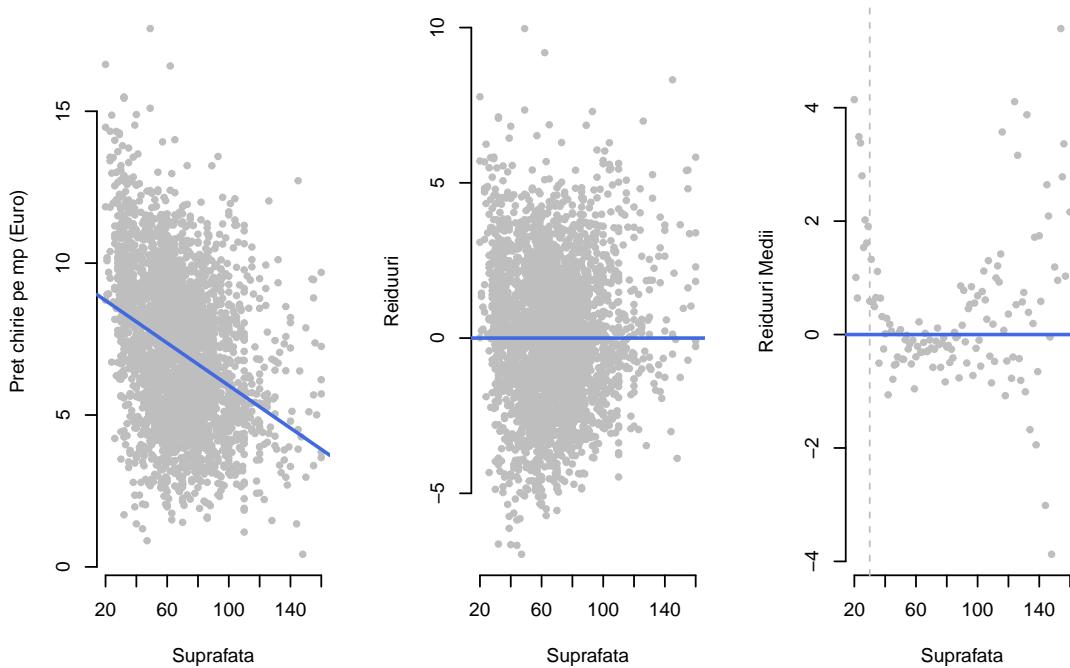
Exemplu: prețul chiriilor din Munchen - transformarea variabilelor

Exp. 4.7

Vom exemplifica efectul asupra răspunsului mediu obținut prin transformarea variabilelor explicative folosind, pentru început, întregul set de date ce face referire la prețul chiriilor în Munchen. Dacă ne uităm la modelul de regresie liniară simplă între prețul net pe metru pătrat al chiriei (ca variabilă răspuns) în funcție de suprafața locuinței de închiriat (variabila explicativă) găsim următoarea relație

$$\widehat{\text{pret_m}^2}_i = 9.469 - 0.035 \times \text{suprafata}_i.$$

În figura de mai jos avem ilustrat grafic modelul de regresie astfel: în subfigura din stânga avem diagrama de împrăștiere cu dreapta de regresie, în subfigura din mijloc avem graficul rezidualelor iar în subfigura din dreapta avem valorile reziduale medii calculate pentru fiecare valoare a suprafeței de locuit. Observăm că pentru apartamentele cu suprafață mică (suprafață mai mică de 30 metrii pătrați) rezidualele sunt predominant pozitive (mai exact peste 88% dintre acestea au reziduuri pozitive) ceea ce sugerează existența unei relații neliniare între preț pe metru pătrat și suprafață.



Specificarea acestei relații neliniare poate fi făcută, de exemplu, prin transformarea variabilei predictor conducând astfel la un model de regresie de tipul

$$pret_m^2_i = \beta_0 + \beta_1 \times f(suprafata_i) + \varepsilon_i$$

unde f este o funcție arbitrară care trebuie aleasă în avans (deterministă) și care nu este estimată de model.

Observăm că modelul de mai sus, chiar dacă reprezintă o generalizare a modelului de regresie liniară simplă considerat anterior, este tot un caz special de regresie liniară în care variabila predictor acum este $x_i = f(suprafata_i)$. Astfel, presupunând că prețul mediu net pe metrul pătrat variază invers proporțional cu suprafața de locuit atunci o transformare potrivită ar fi $f(suprafata_i) = \frac{1}{suprafata_i}$ și în această situație modelul ar deveni

$$pret_m^2_i = \beta_0 + \beta_1 \frac{1}{suprafata_i} + \varepsilon_i.$$

Trebuie menționat că această transformare nu este singura care poate fi luată în calcul, de exemplu am putea considera o transformare logaritmică $f(suprafata_i) = \log(suprafata_i)$, și în general alegem transformarea potrivită în funcție sau de procesul teoretic care a condus la generarea datelor (atunci când acesta este cunoscut sau se poate determina - experiment fizic) sau prin inspecție vizuală a diagramei de împrăștiere.

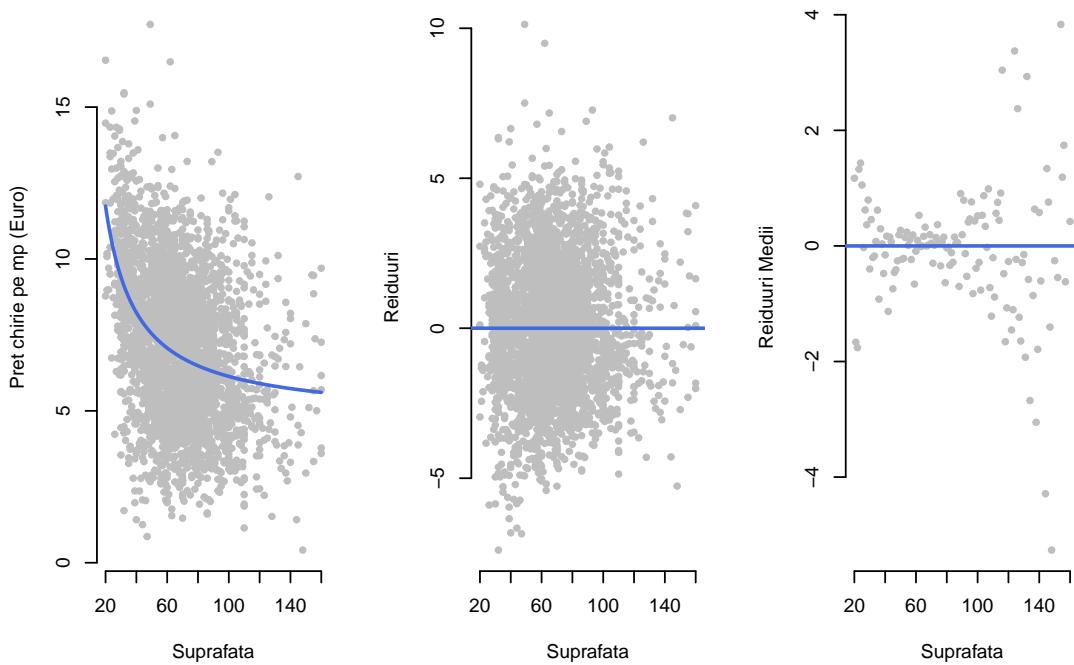
În urma estimării prin metoda celor mai mici pătrate (a se vedea secțiunea [Metoda celor mai mici pătrate](#)) obținem că prețul mediu net pe metrul pătrat este:

$$\widehat{pret_m^2}_i = 4.732 + 140.178 \frac{1}{suprafata_i}$$

iar matricea de design \mathbf{X} pentru acest model

$$\mathbf{X} = \begin{pmatrix} 1 & \frac{1}{suprafata_1} \\ 1 & \frac{1}{suprafata_2} \\ \vdots & \vdots \\ 1 & \frac{1}{suprafata_{3082}} \end{pmatrix} = \begin{pmatrix} 1 & 0.029 \\ 1 & 0.01 \\ \vdots & \vdots \\ 1 & 0.016 \end{pmatrix}.$$

Relația neliniară dintre răspunsul mediu și covariabila suprafața locuinței este ilustrată în figura de mai jos. Se constată că, per total, prețul mediu net al chiriei pe metrul pătrat descrește odată cu creșterea suprafetei de locuit dar această descreștere este mai abruptă pentru apartamentele cu suprafață mică (până în jurul valorii de 40 de m^2) după care aceasta devine aproape liniară cu o tendință de nivelare în jurul valorii de 100 de m^2 . Graficele valorilor reziduale (sau reziduale medii) sugerează o (mai) bună ajustare a modelului neliniar ales la date. \square



Exemplu: înălțimea arborilor de eucalipt - transformarea variabilelor

Exp. 4.8

În contextul setului de date **Eucalypt** am văzut că înălțimea (medie) a arborilor de eucalipt poate fi modelată prin intermediul modelului de regresie liniară simplă prin $\hat{ht}_i = 9.037 + 0.257 \circ c_i$ (a se vedea Exemplul 4.2) unde, observând că majoritatea arborilor a căror circumferință avea valori mai mici de 35 cm se aflau sub dreapta de regresie, am dedus că o relație nelinieră între înălțime și circumferință ar fi mai potrivită. Inspectând vizual diagrama de împrăștiere (figura de mai jos partea stângă) putem sugera o transformare de tipul $f(circ_i) = \sqrt{circ_i}$ ceea ce ar conduce, de exemplu, la modelul de regresie

$$ht_i = \beta_0 + \beta_1 \times circ_i + \beta_2 \times \sqrt{circ_i} + \varepsilon_i, \quad i = 1, \dots, 1429.$$

Estimând coeficienții din modelul de mai sus găsim că înălțimea medie a arborilor de eucalipt este

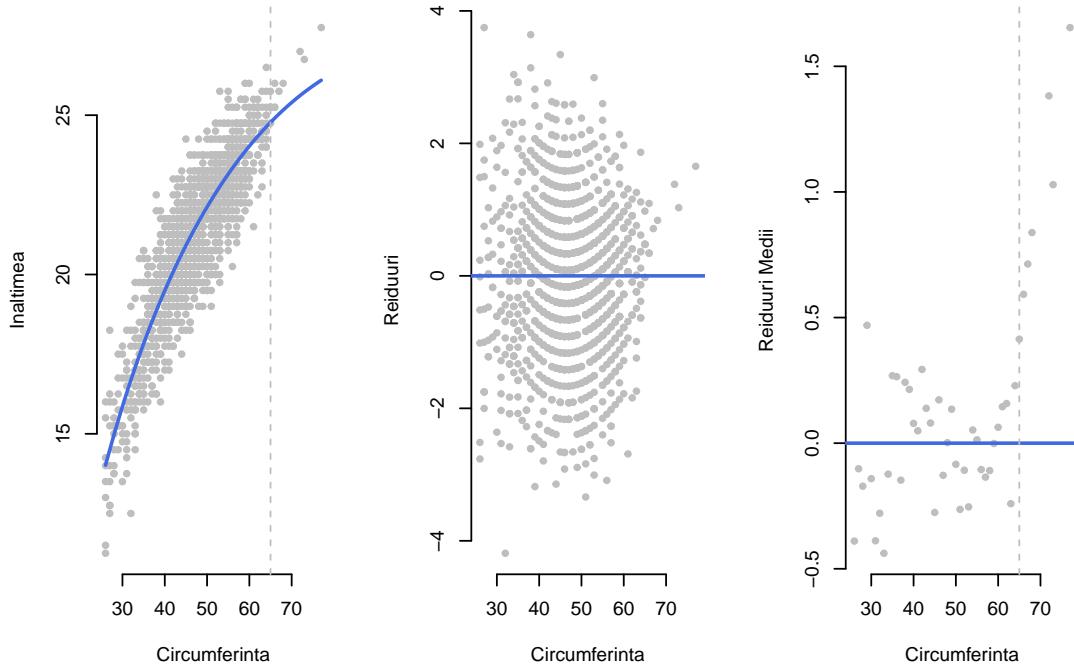
$$\widehat{ht}_i = -24.352 - 0.483 \times circ_i + 9.987 \times \sqrt{circ_i}$$

iar în acest caz matricea de design \mathbf{X} este

$$\mathbf{X} = \begin{pmatrix} 1 & circ_1 & \sqrt{circ_1} \\ 1 & circ_2 & \sqrt{circ_2} \\ \vdots & \vdots & \vdots \\ 1 & circ_{1429} & \sqrt{circ_{1429}} \end{pmatrix} = \begin{pmatrix} 1 & 36 & 6 \\ 1 & 42 & 6.481 \\ \vdots & \vdots & \vdots \\ 1 & 40 & 6.325 \end{pmatrix}.$$

În figura de mai jos avem diagrama de împrăștiere a datelor și curba de regresie care modelează înălțimea medie în funcție de circumferință conform modelului propus, graficul valorilor reziduale precum și graficul valorilor reziduale medii calculate pentru fiecare valoare a circumferinței. Putem observa că modelul propus este ajustat (fitat) mai bine pe date (coeficientul de determinare $R^2 = 0.7921904$ este un pic mai mare atât față de cel în care apare doar termenul $\sqrt{circ_i}$ care este $R^2 = 0.7820638$ cât și de modelul de regresie simplă $R^2 = 0.7683202$ - acest fenomen fiind absolut normal în contextul modelelor imbricate), în special pentru

arbori a căror circumferință este mai mică de 65 cm. Pentru arborii cu circumferințe mai mari modelul sugerează o înălțime medie mai mică decât cea măsurată. \square



În exemplul următor vom face referire la *regresia polinomială* care reprezintă o altă modalitate simplă de a specifica un efect neliniar al predictorilor asupra variabilei răspuns. În general, atunci când covariabila continuă z prezintă un efect neliniar (polinomial) de tipul $\beta_1 z + \beta_2 z^2 + \dots + \beta_k z^k$, modelul

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \dots + \beta_k z_i^k + \varepsilon_i$$

poate fi transformat în modelul de regresie liniară simplă

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

unde $x_{i1} = z_i$, $x_{i2} = z_i^2$, ..., $x_{ik} = z_i^k$. Este important de remarcat faptul că pentru a facilita interpretarea este necesară uneori centrarea (și uneori ortogonalizarea) vectorilor $\mathbf{X}_j = (x_{1j}, \dots, x_{nj})^\top$ în $\mathbf{X}_j - \bar{\mathbf{X}}_j$ unde $\bar{\mathbf{X}}_j = \left(\bar{x}_j = \frac{\sum_{i=1}^n x_{ij}}{n}, \dots, \bar{x}_j \right)^\top$.

Exemplu: prețul chirilor din Munchen - regresie polinomială

Exp. 4.9

Vom ilustra modelarea prin regresie polinomială folosind setul de date **Munchen**. Vom presupune că variabila predictor suprafața de locuit admite un efect polinomial. Pentru un efect polinomial de ordin doi obținem următorul model de regresie:

$$\text{pret_m}^2_i = \beta_0 + \beta_1 \times \text{suprafata}_i + \beta_2 \times \text{suprafata}_i^2 + \varepsilon_i$$

iar pentru un efect de ordin trei avem

$$pret_m^2_i = \beta_0 + \beta_1 \times suprafata_i + \beta_2 \times suprafata_i^2 + \beta_3 \times suprafata_i^3 + \varepsilon_i.$$

Bineînțeles, putem folosi și polinoame de grad superior dar în general preferăm pentru simplitate și interpretabilitate polinoame de grad mai mic. De asemenea, atunci când folosim polinoame de grad mai mare ca 3 observăm că rezultatul devine instabil în special la capetele domeniului covariabilei (a se vedea figura de mai jos dreapta unde folosim un polinom de grad 4). Notând cu $x_{i1} = suprafata_i$, $x_{i2} = suprafata_i^2$ și $x_{i3} = suprafata_i^3$ obținem modelul de regresie liniară

$$pret_m^2_i = \begin{cases} \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, & \text{polinom de grad 2} \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, & \text{polinom de grad 3} \end{cases}$$

Matricele de design corespunzătoare sunt

$$\mathbf{X} = \begin{pmatrix} 1 & 35 & 35^2 \\ 1 & 104 & 104^2 \\ \vdots & \vdots & \vdots \\ 1 & 62 & 62^2 \end{pmatrix} = \begin{pmatrix} 1 & 35 & 1225 & 1.0816 \times 10^4 \\ 1 & 104 & 1.0816 \times 10^4 & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 62 & 3844 & 2.38328 \times 10^5 \end{pmatrix}$$

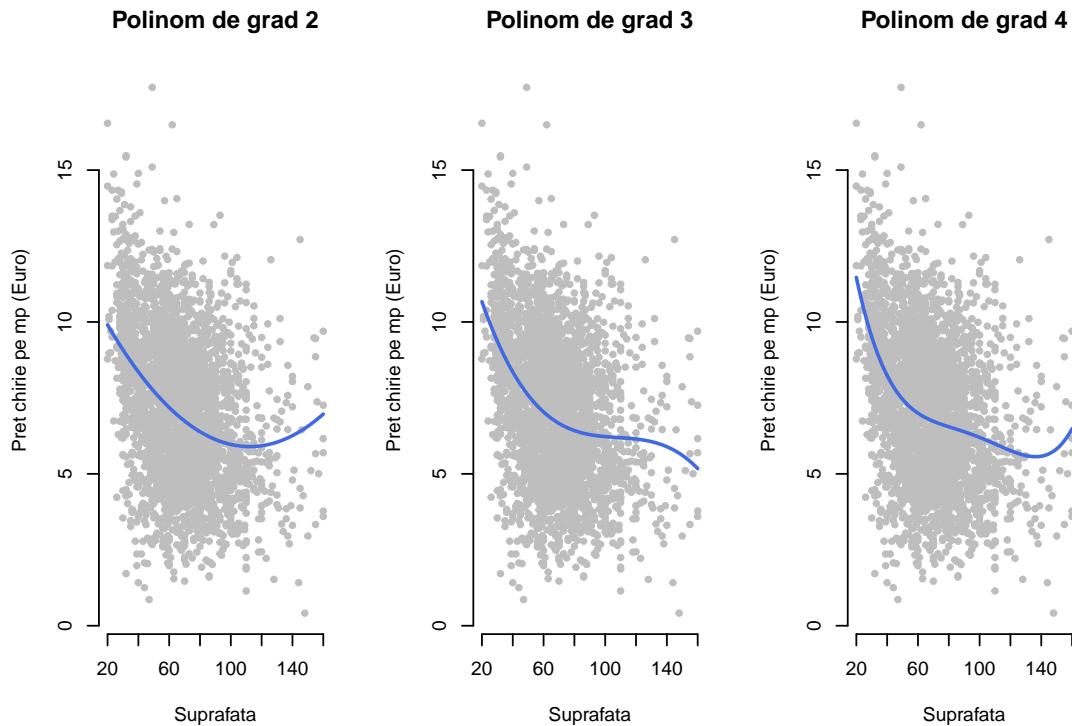
respectiv

$$\mathbf{X} = \begin{pmatrix} 1 & 35 & 35^2 & 35^3 \\ 1 & 104 & 104^2 & 104^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 62 & 62^2 & 62^3 \end{pmatrix} = \begin{pmatrix} 1 & 35 & 1225 & 4.2875 \times 10^4 \\ 1 & 104 & 1.0816 \times 10^4 & 1.124864 \times 10^6 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 62 & 3844 & 2.38328 \times 10^5 \end{pmatrix}$$

iar curbele de regresie estimate

$$\widehat{pret_m^2}_i = \begin{cases} 11.83 - 0.106 \times suprafata_i + 4.7 \times 10^{-4} \times suprafata_i^2, & \text{grad 2} \\ 14.28 - 0.218 \times suprafata_i + 0.002 \times suprafata_i^2 - 6 \times 10^{-6} \times suprafata_i^3, & \text{grad 3} \end{cases}$$

Figura de mai jos ne prezintă diagramele de împărăștiere împreună curbele de regresie specifice modelelor selectate. Observăm că rezultatele sunt similare cu cele obținute prin transformarea $f(suprafata_i) = \frac{1}{suprafata_i}$ (modelele polinomiale considerate fiind imbricate conduc la o creștere a coeficientului de determinare ca și în cazul Exemplului 4.8). \square



Dacă până acum am considerat doar modele în care apărarea o singură variabilă explicativă, exemplul de mai jos consideră o generalizare la doi sau mai mulți predictori.

Exemplu: prețul chiriielor din München - model aditiv

Ex. 4.10 În acest exemplu vom modela prețul mediu net al chiriei pe metrul pătrat adăugând pe lângă efectul invers proporțional dat de suprafața de locuit (a se vedea Exemplul 4.7) și efectul dat de anul de construcție al apartamentului (*an_con*). Dacă anul construcției prezintă un efect liniar atunci asupra prețului chiriei, obținem modelul (denumit modelul 1)

$$pret_m^2_i = \beta_0 + \beta_1 \times f(suprafata_i) + \beta_2 \times an_con_i + \varepsilon_i$$

unde $f(suprafata_i) = \frac{1}{suprafata_i}$ și care în urma estimării devine

$$\widehat{pret_m^2}_i = -65.406 + 119.361 \frac{1}{suprafata_i} + 0.036 \times an_con_i.$$

Alternativ, putem să considerăm că anul construcției locuinței prezintă un efect nelinier asupra prețului chiriei pe metrul pătrat, aspect pe care îl vom modela prin intermediul unui polinom de grad 3 și în acest caz avem modelul (denumit model 2)

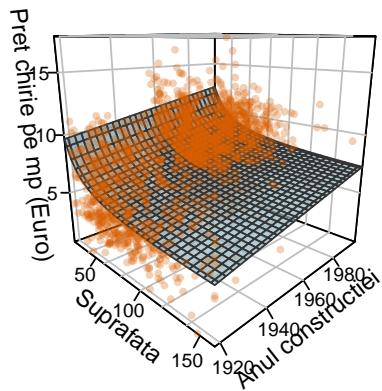
$$pret_m^2_i = \beta_0 + \beta_1 \times \frac{1}{suprafata_i} + \beta_2 \times an_con_i + \beta_3 \times an_con_i^2 + \beta_4 \times an_con_i^3 + \varepsilon_i.$$

Modelul estimat (răspunsul mediu estimat) în acest caz este

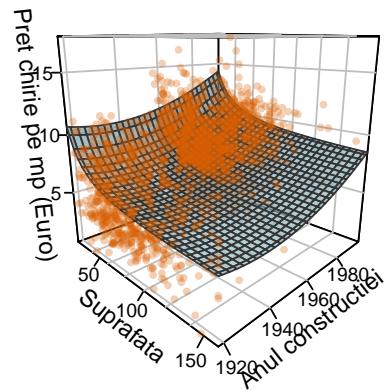
$$\widehat{pret_m^2}_i = 2.94193 \times 10^4 + 129.57 \times \frac{1}{suprafata_i} - 43.3 \times an_con_i + 0.021 \times an_con_i^2 - 3 \times 10^{-6} \times an_con_i^3$$

În figura de mai jos avem ilustrate diagramele de împrăștiere (3D) împreună cu suprafețele de regresie induse de modelele menționate mai sus. Ca și în cazul regresiei liniare simple, vizualizarea diagramelor de împrăștiere atunci când avem de-a face cu două variabile explicative este un instrument grafic care permite o mai bună înțelegere a relației dintre variabila răspuns și covariabilele implicate. Pentru trasarea în R a diagramelor de împrăștiere am folosit funcția `scatter3D` din pachetul `plot3D`.

Model 1



Model 2



Combinând în modelul 2 efectele variabilelor explicative *suprafața de locuit* și *anul de construcție a locuinței* în funcțiile f_1 și respectiv f_2 , obținem următorul model

$$\begin{aligned} \text{pret_m}^2_i &= \beta_0 + \underbrace{f_1(\text{suprafata}_i)}_{=\beta_1 \times \frac{1}{\text{suprafata}_i}} + \underbrace{f_2(\text{an_con}_i)}_{=\beta_2 \times \text{an_con}_i + \beta_3 \times \text{an_con}_i^2 + \beta_4 \times \text{an_con}_i^3} + \varepsilon_i. \end{aligned}$$

Funcțiile estimate sunt

$$\hat{f}_1(\text{suprafata}) = 129.57 \times \frac{1}{\text{suprafata}}$$

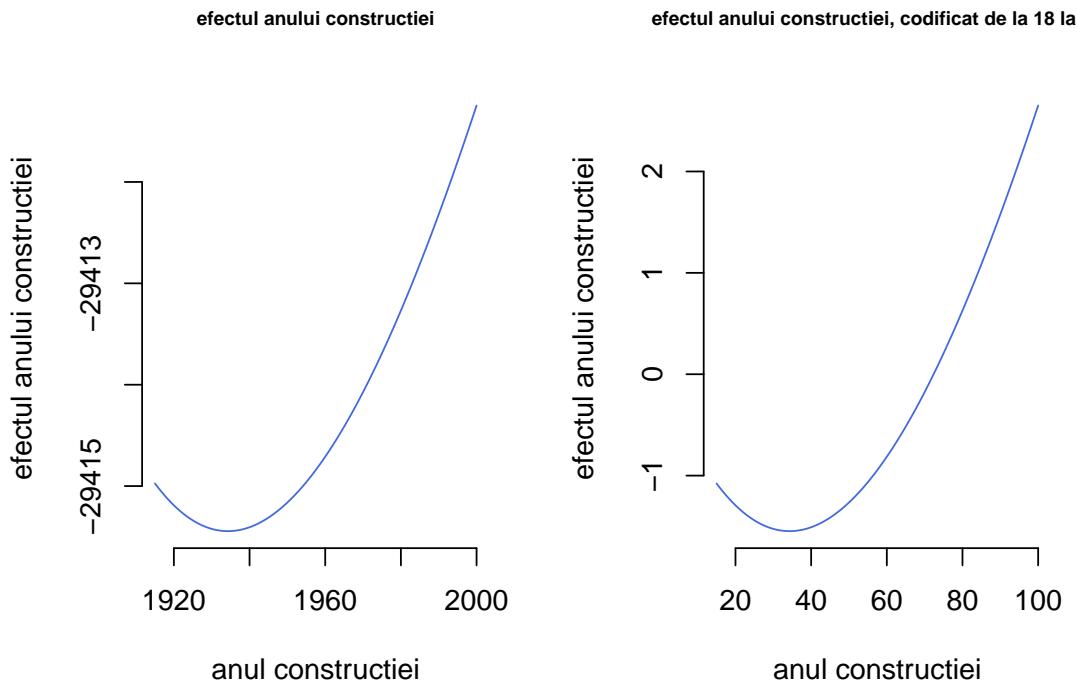
și respectiv

$$\hat{f}_2(\text{an_con}) = -43.3 \times \text{an_con}_i + 0.021 \times \text{an_con}_i^2 - 3 \times 10^{-6} \times \text{an_con}_i^3.$$

Efectul estimat $\hat{f}_1(\text{suprafata})$ și $\hat{f}_2(\text{an_con})$ al celor două covariabile poate fi interpretat mai ușor prin intermediul vizualizării. De exemplu, în figura de mai jos este prezentat efectul anului construcției asupra prețului chiriei pe metrul patrat. Observăm că, datorită scalei covariabilei *anul de construcție a locuinței* care ia valori între 1918 și 1998, efectul acestei covariabile este între -2.9415×10^4 și -2.9411×10^4 . Dacă în schimb am specifica anul de construcție ca având valori de la 18 la 98 (scăzând 1900) atunci am obține modelul

$$\widehat{\text{pret_m}^2}_i = 5.4 + 129.57 \times \frac{1}{\text{suprafata}_i} - 0.09 \times \text{an_con}_i + 0.002 \times \text{an_con}_i^2 - 3 \times 10^{-6} \times \text{an_con}_i^3$$

Chiar dacă parametrii noului model diferă de cei ai modelului 2, atunci când ne uităm la efectul anului de construcție (figura de mai jos dreapta) vedem că aceasta este doar translatată vertical. Acest fenomen este natural ținând cont de observația că adăugând un termen constant la \hat{f}_2 și scăzând același termen din $\hat{\beta}_0$ valoarea estimată a răspunsului mediu rămâne neschimbată (nivelul unei funcții neliniare nu este identificabil).



Observația anterioară, că nivelul unei funcții neliniare poate fi schimbat în mod arbitrar de exemplu prin transformarea variabilei predictor, conduce la recomandarea *centrării* covariabilelor (sau a funcțiilor de covariabile). Astfel, în cazul modelului propus putem înlocui variabila $\text{suprafata_inv} = \frac{1}{\text{suprafata}}$ cu $\text{suprafata_inv} - \overline{\text{suprafata_inv}}$ și respectiv variabilele $\text{an_con_j} = \text{an_con}^j$ cu $\text{an_con_j} - \overline{\text{an_con_j}}$.

În cazul modelării prin intermediul polinoamelor putem să folosim *polinoame ortogonale*³, unde baza uzuială 1, x, x^2, x^3, \dots va fi înlocuită de o bază de polinoame ortogonale ceea ce implică centralitatea și ortogonalitatea coloanelor matricei de design corespunzătoare. De asemenea este important de remarcat faptul că folosind o bază de polinoame ortogonale calculul estimatorilor prin metoda celor mai mici pătrate este mai stabil. În R putem construi polinoame ortogonale folosind funcția `poly()`.

Folosind variabilele *ortogonale* an_co_j , $j = 1, 2, 3$ și variabila centrată $\text{suprafata_inv_cen} = \text{suprafata_inv} - \overline{\text{suprafata_inv}}$ obținem modelul de regresile liniare

$$\widehat{\text{pret_m}^2}_i = 7.111 + 129.572 \times \text{suprafata_inv_cen} + 43.938 \times \text{an_co_1} + 27.539 \times \text{an_co_2} - 1.756 \times \text{an_co_3}$$

³Spunem că un sistem de polinoame $\{p_n(x)\}$, $n = 0, 1, \dots$ este ortogonal pe intervalul (a, b) dacă $\int_a^b p_n(x)p_m(x) dx = \delta_{nm}$ unde δ_{nm} este simbolul lui Kronecker. În general putem avea ortogonalitate în raport cu o funcție de pondere $w(x) \geq 0$ (continuă sau continuă pe porțiuni) dacă $\int_a^b p_n(x)p_m(x)w(x) dx = \delta_{nm}$.

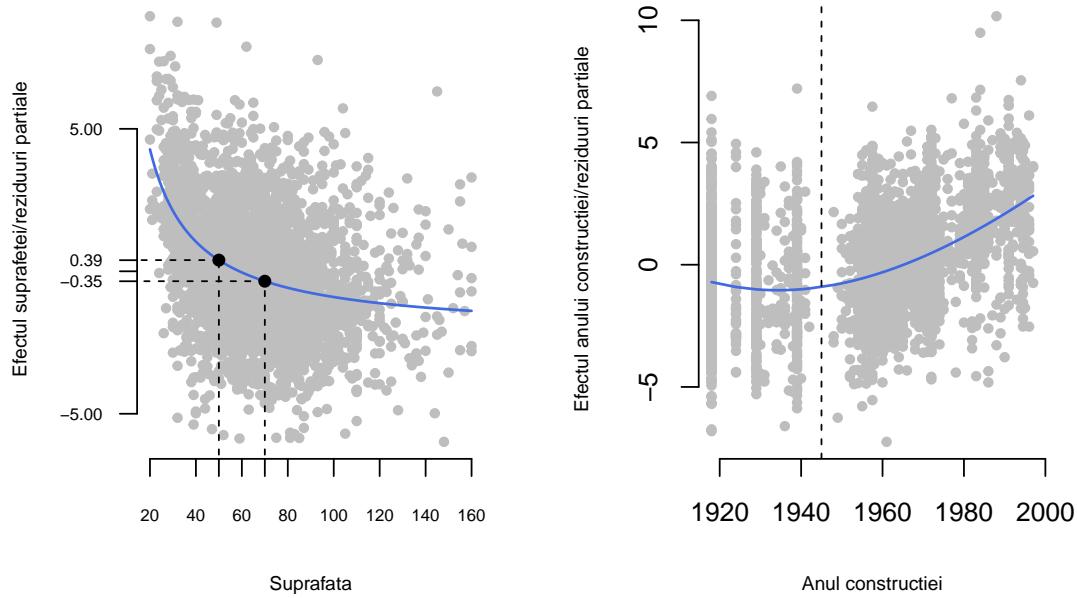
În figura următoare sunt ilustrate efectele covariabilelor *suprafața de locuit* și *anul de construcție a locuinței*, $\hat{f}_1(suprafata_i)$ și respectiv $\hat{f}_2(an_con_i)$, împreună cu reziduurile parțiale, definite prin

$$\hat{\varepsilon}_{suprafata,i} = pret_m^2_i - \hat{\beta}_0 - \hat{f}_2(an_con_i) = \hat{\varepsilon}_i + \hat{f}_1(suprafata_i)$$

și respectiv

$$\hat{\varepsilon}_{an_con,i} = pret_m^2_i - \hat{\beta}_0 - \hat{f}_1(suprafata_i) = \hat{\varepsilon}_i + \hat{f}_2(an_con_i).$$

Reziduurile parțiale pentru covariabila *suprafața de locuit*, $\hat{\varepsilon}_{suprafata,i}$, iau în calcul și variabilitatea indusă de suprafața de locuit atunci când toate celelalte efecte sunt înălțurate (în acest caz anul de construcție). În mod similar, pentru reziduurile parțiale asociate anului de construcție a locuinței - $\hat{\varepsilon}_{an_con,i}$, efectul suprafetei este eliminat dar nu și cel al anului de construcție. Trasând reziduurile parțiale putem investiga dintr-o perspectivă vizuală dacă modelul neliniar ales este potrivit sau nu.



Din punct de vedere al interpretabilității, efectul suprafetei de locuit (figura din stânga) specifică influența acestei covariabile asupra prețului mediu net al chiriei pe metrul pătrat atunci când celelalte covariabile sunt păstrate constante. Astfel pentru apartamentele cu o suprafață de 50 și respectiv 70 metrii pătrați obținem un efect estimat de $\hat{f}_1(50) \approx 0.39$ și respectiv $\hat{f}_1(70) \approx -0.35$ euro pe metrul pătrat. Acest fapt implică faptul că în medie prețul chiriei pe metrul pătrat pentru apartamentele de 70 metrii pătrați sunt mai ieftine cu 0.74 euro pe metrul pătrat față de cele cu o suprafață de 50 metrii pătrați, presupunând că ambele au fost construite în același an. De asemenea, dacă ne uităm la apartamentele cu suprafață de 60 și respectiv 80 metrii pătrați atunci diferența este de doar 0.54 euro fapt indicat de efectul aproape constant al covariabilei atunci când locuитеle au suprafață mare.

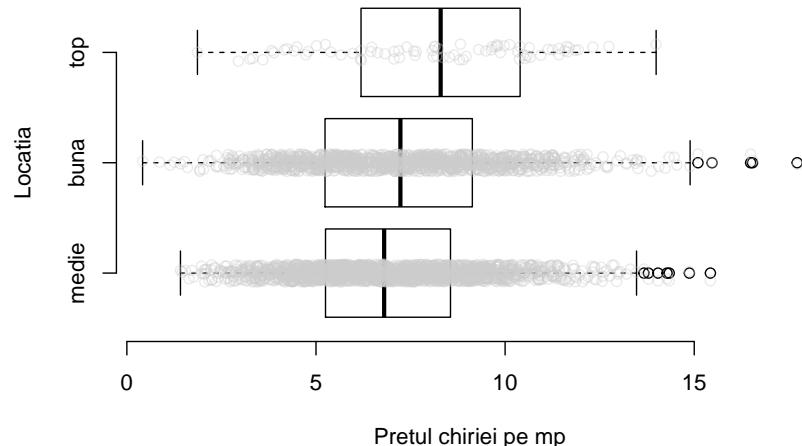
În același mod poate fi interpretat și graficul care prezintă efectul anului de construcție a locuinței. Putem observa (figura de mai sus dreapta) că pentru apartamentele, cu aceeași suprafață, construite înaintea Celui de-al Doilea Război Mondial efectul anului construcției asupra prețului mediu pe metrul pătrat al chiriei este constant pe când pentru apartamentele construite după 1945 tendința este de creștere în mod liniar.

4.3.2 Cazul variabilelor explicative discrete

În secțiunea precedentă am prezentat succint care este efectul induș de o variabilă explicativă continuă asupra variabilei răspuns. În această secțiune vom ilustra cazul variabilelor explicative discrete (categoriale).

Exemplu: prețul chiriilor din Munchen - covariabile categoriale

Pentru prezentarea metodologiei ne plasăm în contextul setului de date **Munchen** referitor la prețul chiriilor din orașul Munchen și investigăm modul în care locația imobilului (variabila **locatie** - **location**) influențează prețul chiriei locuinței. Variabila **locatie** admite trei categorii: 1 = locație medie, 2 = locație bună și 3 = locație de top.



Dacă am modelat prețul chiriei pe metrul pătrat în funcție de locație folosind un model de regresie liniară simplă și tratând covariabila **locatie** ca o variabilă continuă am avea

$$pret_m^2_i = \beta_0 + \beta_1 \times locatie_i + \varepsilon_i$$

care în urma estimării ar conduce la

$$\widehat{pret_m^2}_i = 6.544 + 0.393 \times locatie_i.$$

Observăm că, datorită modului în care am ales codificarea variabilei **locatie**, prețul mediu al chiriilor pe metrul pătrat pentru locuințele care se situează într-o locație bună este de două ori mai mare decât al celor care se află într-o locație medie (0.393 Euro față de 0.786 Euro) iar cel al locuințelor dintr-o locație de top este de trei ori mai mare. Bineînțeles că dacă am fi codificat variabila **locatie** prin 2 = locație medie, 6 = locație bună și 8 = locație de top atunci efectele ar fi fost diferite (apartamentele din locații de top ar fi fost de patru ori mai scumpe, în medie, față de cele din locație medie).

Prin urmare, constatăm că modul în care codificăm covariabila categorială influențează direct rezultatele și modul de interpretare. Problema care apare este că în cazul variabilelor categoriale distanțele dintre categorii nu au întotdeauna aceeași însemnatate, precum nici apartamentele dintr-o locație bună nu sunt de două ori (sau de trei ori) mai bune față de cele situate într-o locație medie. O modalitate de a corecta această problemă constă în introducerea unor noi covariabile, denumite variabile ajutătoare - *dummy*, care să permită estimarea efectului induș de fiecare categorie a variabilei explicative în cauză. Mai precis, pentru variabila **locatie**

introducem trei noi variabile ajutătoare **alotatie**, **glocatie** și **tlocatie** de tip variabilă indicator după cum locuința se află într-o locație medie (average), bună (good) respectiv de top astfel

$$\begin{aligned} alocatie &= \begin{cases} 1, & \text{locație} = 1 \text{ (locație medie)} \\ 0, & \text{altfel} \end{cases} \\ glocatie &= \begin{cases} 1, & \text{locație} = 2 \text{ (locație bună)} \\ 0, & \text{altfel} \end{cases} \\ tlocatie &= \begin{cases} 1, & \text{locație} = 3 \text{ (locație de top)} \\ 0, & \text{altfel} \end{cases} \end{aligned}$$

Folosind cele trei variabile ajutătoare pentru modelarea prețului chiriei pe metrul pătrat avem modelul de regresie

$$pret_m^2_i = \beta_0 + \beta_1 \times alocatie_i + \beta_2 \times glocatie_i + \beta_3 \times tlocatie_i + \varepsilon_i$$

care prezintă acum o nouă problemă, și anume că parametrii acestuia nu sunt identificabili, deci nu pot fi determinați în mod unic. Acest lucru se poate vedea imediat dacă ne uităm la efectele celor trei covariabile: apartamentele dintr-o locație medie impun un efect asupra prețului mediu al chiriei de $\beta_0 + \beta_1$ ($alocatie_i = 1$, $glocatie_i = 0$ și $tlocatie_i = 0$), iar cele care se situează în locații bune și de top prezintă un efect de $\beta_0 + \beta_2$ și respectiv $\beta_0 + \beta_3$. Termenul β_0 (ordonata la origine) apare în toate cele trei efecte astfel, adăugând o cantitate constantă la β_0 și scăzând-o apoi din coeficienții β_1 , β_2 și β_3 conduce la același efect asupra variabilei dependente. Același lucru se poate observa și dacă ne uitam la matricea de design \mathbf{X} care acum nu mai avea rang plin, vectorii coloană nu mai sunt liniar independenți, în fapt primul vector coloană este egal cu suma vectorilor coloană 2, 3 și 4.

Problema identifiabilității se poate rezolva în mai multe moduri dar aici vom considera cazul cel mai ușual și anume eliminarea din model a unei variabile ajutătoare din cele trei noi create. Prin urmare, dacă eliminăm variabila **alocatie** obținem modelul

$$pret_m^2_i = \beta_0 + \beta_1 \times glocatie_i + \beta_2 \times tlocatie_i + \varepsilon_i$$

care în urma estimării devine

$$\widehat{pret_m^2}_i = 6.957 + 0.316 \times glocatie_i + 1.216 \times tlocatie_i.$$

Trebuie să avem grijă ca atunci când interpretăm coeficienții modelului de mai sus și efectele induse să ținem seama de categoria (variabila ajutătoare) eliminată - în cazul nostru locație medie. Această categorie se numește *categorie de referință*. Astfel, pentru apartamentele situate într-o locație medie obținem un efect estimat $\hat{\beta}_0 = 6.957$, pentru cele din locație bună $\hat{\beta}_0 + \hat{\beta}_1 = 7.273$ iar pentru cele din locație de top efectul este $\hat{\beta}_0 + \hat{\beta}_3 = 8.173$. Constatăm că în comparație cu apartamentele aflate într-o locație medie, prețul mediu al chiriei pe metrul pătrat al apartamentelor situate într-o locație bună este cu 0.316 Euro mai mult iar cel al apartamentelor aflate într-o locație de top cu 1.216 Euro mai mult. \square

În general, atunci când vrem să modelăm efectul unei variabile predictor categoriale z cu c categorii, $z \in \{1, 2, \dots, c\}$, fixăm o categorie de referință (de obicei categoria cea mai frecventă sau categoria de control), în acest caz categoria c , și definim $c - 1$ variabile de tip *dummy*

$$x_{i1} = \begin{cases} 1, & z_i = 1, \\ 0, & \text{altfel} \end{cases} \quad \dots \quad x_{i,c-1} = \begin{cases} 1, & z_i = c - 1, \\ 0, & \text{altfel} \end{cases}$$

pe care să le includem ca variabile explicative în model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{i,c-1} x_{i,c-1} + \varepsilon_i.$$

Efectele estimate trebuie interpretate în raport cu categoria de referință. Este important de menționat că există mai multe scheme de codificare a variabilelor explicative categoriale (**simple coding**, **deviation** sau **effect coding**, **Helmert coding**, **Orthogonal polynomial coding**, etc.), cea de mai sus este cea mai des întâlnită.

Înainte de a încheia această secțiune vom ilustra o serie (patru) de metode prin care putem crea în R variabilele ajutătoare de tip *dummy*. Vom începe prin crearea efectivă a variabilelor folosind librăria **tidyverse** (sau funcțiile de bază):

```
# 1) creare variabilelor alocation, glocation si tlocation
munich = munich %>%
  mutate(alocation = ifelse(location == 1, 1, 0),
         glocation = ifelse(location == 2, 1, 0),
         tlocation = ifelse(location == 3, 1, 0))

# exemplu
lm(rentsqm ~ alocation + tlocation, data = munich)
lm(rentsqm ~ glocation + tlocation, data = munich)
```

O a doua metodă este de a folosi funcția **I()** (permite interpretarea aritmetică a expresiei din paranteză) în care specificăm categoria variabilei explicative care vrem să fie inclusă în model. Această metodă ne permite fixarea unei categorii de referință, prin eliminarea ei din model.

```
# 2) Folosirea functiei I()
lm(rentsqm ~ I(location == 1) + I(location == 3), data = munich)
lm(rentsqm ~ I(location == 2) + I(location == 3), data = munich)
```

În al treilea rând putem folosi funcția **factor()** care asigură programul R că avem de-a face cu o variabilă categorială, fiecare categorie numindu-se nivel (level).

```
# 3) Folosirea functiei factor
munich$locationf = factor(munich$location)

lm(rentsqm ~ locationf, data = munich)

# sau echivalent
lm(rentsqm ~ factor(location), data = munich)
```

Ultima funcție este funcția **contr.treatment()**, care în fapt se folosește în pereche cu funcția **contrasts()**. Funcția **contr.treatment()** ne permite specificarea categoriei de referință prin includerea opțiunii **base =**.

```
# 4) Folosirea functiei contr.treatment
munich$locationf = factor(munich$location)

contrasts(munich$locationf) = contr.treatment(3, base = 2)
# 3 - este numarul de niveluri iar base reprezinta nivelul de referinta
lm(rentsqm ~ locationf, data = munich)

contrasts(munich$locationf) = contr.treatment(3, base = 1)
lm(rentsqm ~ locationf, data = munich)
```

4.3.3 Interacția dintre două covariabile

Atunci când pentru fiecare variabilă explicativă efectul ei asupra răspunsului mediu nu depinde de nivelul celorlalte covariabile din model spunem că variabilele predictor au efect *aditiv* sau nu interacționează unele cu celealte. În cazul în care efectul unei covariabile depinde de nivelul cel puțin a uneia dintre celealte covariabile, spunem că există un efect de interacție între ele. Un exemplu simplu de model de regresie liniară neaditiv este

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

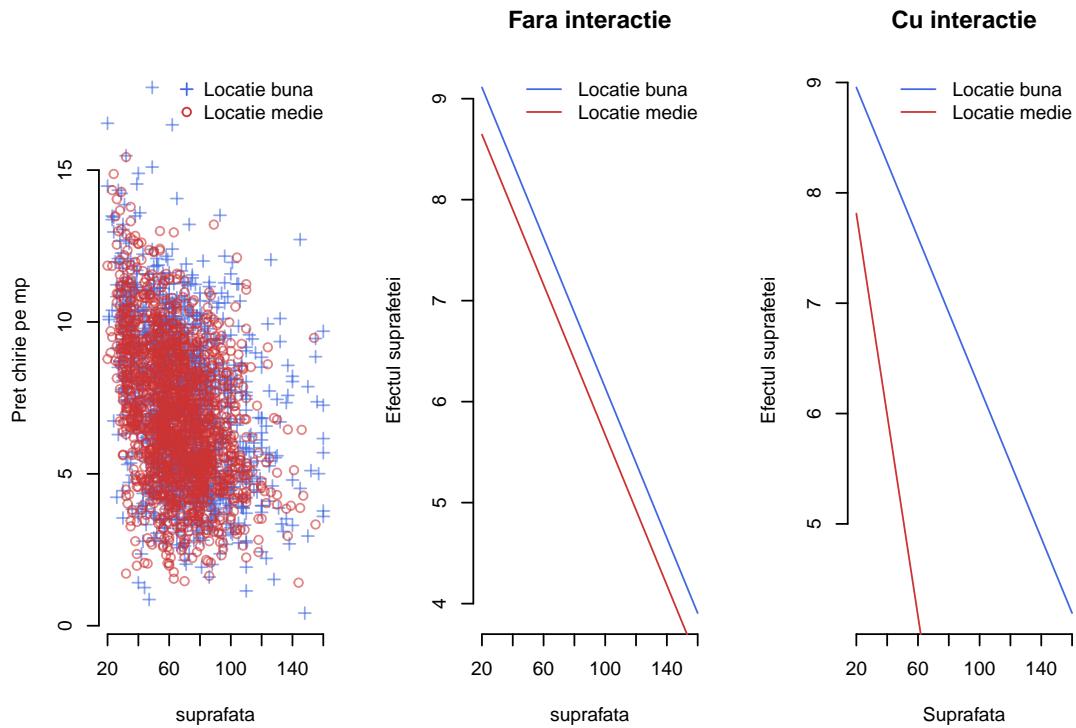
unde termenul $\beta_3 x_1 x_2$ se numește *termen de interacție* sau *interacția* dintre x_1 și x_2 (*interaction effect*). Cum termenii $\beta_1 x_1$ și $\beta_2 x_2$ depind doar de covariabilele x_1 și respectiv x_2 , se numesc *efecte principale* (*main effects*). Evident că modelul de mai sus este un caz particular de model de regresie liniară deoarece putem nota termenul $x_1 x_2$ cu x_3 . În general, produsul a două sau mai multe efecte principale diferite reprezintă o interacție.

Pentru a înțelege care este impactul termenului de interacție asupra răspunsului mediu să considerăm că variabila predictor x_1 se modifică prin cantitatea $\delta = 1$:

$$\mathbb{E}[y|x_1 + \delta, x_2] - \mathbb{E}[y|x_1, x_2] = \beta_1 \delta + \beta_3 \delta x_2 = \beta_1 + \beta_3 x_2$$

ceea ce arată că, în cazul în care $\beta_3 \neq 0$, răspunsul mediu se modifică prin $\beta_1 + \beta_3 x_2$ unități odată cu creșterea cu o unitate a lui x_1 și păstrarea la nivel constant a lui x_2 . Prin urmare dacă efectul indus asupra răspunsului mediu de schimbarea unei variabile explicative depinde de valoarea unei alte variabile explicative atunci avem de-a face cu un efect de interacție. De cele mai multe ori, atunci când semnele coeficienților β_1 și β_2 sunt la fel (negative sau pozitive) și semnul lui β_3 coincide cu acestea spunem că avem un efect de întărire (reinforcement effect) iar când semnul lui β_3 este contrar atunci avem efect de intersecție (intersection effect).

În figura următoare avem ilustrat, pentru setul de date **München** în care am eliminat proprietățile din locații de top, variația pretului mediu pe metrul pătrat în funcție de suprafața de locuit în situația în care apartamentele se găsesc într-o locație bună sau medie și atunci când avem un efect de interacție (figura din dreapta) între cele două covariabile și respectiv când nu avem acest efect (figura din stânga).



Ca regulă, în cazul unui model de regresie liniară care conține un termen de interacție sau o putere este recomandat ca acesta să conțină și efectele principale corespunzătoare care au alcătuit termenul de interacție sau putere. De exemplu, dacă în modelul de regresie apar termenii x_1^3 și respectiv $x_2x_3x_4$ atunci modelul ar trebui să conțină și termenii (efectele principale) x_1 , x_2 , x_3 și respectiv x_4 .

În subsecțiunile care urmează, vom trata separat cazurile în care efectul de interacție apare între două variabile categoriale sau între o variabilă categorială și una continuă.

4.3.3.1 Interacția dintre două variabile categoriale În această secțiune vom trata cazul în care în model apar două variabile predictor calitative x_1 și x_2 , categoriale, cu c și respectiv m categorii și suntem interesați de efectul interacției dintre cele două asupra răspunsului mediu. Cum putem modela efectul fiecărei variabile predictor categoriale prin intermediul a $c - 1$ și respectiv $m - 1$ variabile ajutătoare de tip indicator (*dummy*) este suficient să ne limităm atenția asupra covariabilelor binare. Într-adevăr, cele două variabile explicative x_1 și x_2 induc în modelul de regresie $cm - 1$ termeni, fiecare fiind o variabilă binară.

Exemplu: prețul chiriilor din Munchen - interacția dintre două covariabile categoriale

Exp. 4.12 În contextul setului de date **Munchen** să presupunem că vrem să investigăm relația dintre prețul mediu al chiriilor pe metrul pătrat și tipul de baie (0 = standard, 1 = premium), respectiv bucătărie (0 = standard, 1 = premium) cu care este dotat apartamentul. În acest caz, cele două covariabile binare $x_1 = bucatarie$ și $x_2 = baie$ conduc la modelul de regresie

$$pret_m^2_i = \beta_0 + \beta_1 \times bucatarie_i + \beta_2 \times baie_i + \beta_3 \times bucatarie_i \cdot baie_i + \varepsilon_i$$

care în urma estimării devine

$$\widehat{pret_m^2}_i = 7.024 + 1.419 \times bucatarie_i + 0.401 \times baie_i + 0.287 \times bucatarie_i \cdot baie_i$$

Astfel, coeficientul $\hat{\beta}_1 = 1.419$ măsoară efectul indus de o bucătărie premium asupra prețului mediu al chiriei pe metrul pătrat, $\hat{\beta}_2 = 0.401$ măsoară efectul indus de o baie premium iar $\hat{\beta}_3 = 0.287$ măsoară efectul adițional pentru locuințele care au și baie și bucătărie premium. Observăm că față de un apartament cu dotări standard (baie și bucătărie standard), apartamentele care au doar bucătărie premium sunt mai scumpe în medie cu 1.419 Euro pe metrul pătrat pe când cele care au doar baie premium au un preț mai ridicat cu 0.401 Euro. Având și baie și bucătărie premium face ca prețul mediu al chiriei pe metrul pătrat să crească cu 2.107 Euro.

Dacă ne interesăm acum la efectul interacției dintre covariabila locația apartamentului (*locatie* - 1 = medie, 2 = bună, 3 = de top) și covariabila tipul de bucătărie (*bucatarie* - 0 = standard sau 1 = premium) asupra prețului chiriei pe metrul pătrat, atunci avem modelul

$$\begin{aligned} \text{pret_m}^2_i &= \beta_0 + \beta_1 \times \text{alocatie}_i + \beta_2 \times \text{glocatie}_i + \beta_3 \times \text{bucatarie}_i + \\ &\quad + \beta_4 \times \text{alocatie}_i \cdot \text{bucatarie}_i + \beta_5 \times \text{glocatie}_i \cdot \text{bucatarie}_i + \varepsilon_i \end{aligned}$$

în care am considerat toate combinațiile posibile între valorile celor două variabile explicative (cu excepția categoriilor de referință - în acest caz locații de top). Efectul variabilei explicative *locatie* este explicat prin intermediul a două variabile ajutătoare binare *alocatie* și *glocatie* folosind ca și categorie de referință locațiile din zone de top. Interpretarea coeficienților este următoarea:

$$\begin{aligned} \beta_0 &: \text{efectul locație} = \text{de top și bucătărie} = \text{standard} \\ \beta_0 + \beta_1 &: \text{efectul locatie} = \text{medie și bucatarie} = \text{standard} \\ \beta_0 + \beta_2 &: \text{efectul locatie} = \text{bună și bucatarie} = \text{standard} \\ \beta_0 + \beta_3 &: \text{efectul locatie} = \text{de top și bucatarie} = \text{premium} \\ \beta_0 + \beta_1 + \beta_3 + \beta_4 &: \text{efectul locatie} = \text{medie și bucatarie} = \text{premium} \\ \beta_0 + \beta_2 + \beta_3 + \beta_5 &: \text{efectul locatie} = \text{bună și bucatarie} = \text{premium} \end{aligned}$$

Putem observa că, față de un apartament de top cu bucătărie standard, un apartament dintr-o locație medie cu același tip de bucătărie are un preț mediu al chiriei pe metrul pătrat cu $\hat{\beta}_1 = -1.279$ Euro mai mic pe când un apartament dintr-o zonă bună cu o bucătărie premium are un preț mediu cu $\hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_5 = 1.052$ Euro mai mare.

4.3.3.2 Interacția dintre o variabilă categorială și una continuă Să presupunem că vrem să modelăm efectul pe care îl are interacția dintre o covariabilă continuă x și una discretă (categorială) z cu c categorii asupra răspunsului mediu. Într-o primă etapă pentru a include efectul variabilei explicative z vom fixa o categorie ca și categorie de referință și vom introduce, ca și până acum, $c - 1$ variabile ajutătoare de tip indicator (*dummy*). Pentru a simplifica discuția vom considera că variabila categorială $z \in \{0, 1\}$ este o variabilă de tip indicator, binară, și vom presupune că efectele principale și efectele de interacție datorate covariabilelor x și z sunt liniare în raport cu răspunsul prin urmare avem un model de tipul

$$y = \beta_0 + \underbrace{\beta_1 x + \beta_2 z}_{\text{efekte principale}} + \underbrace{\beta_3 xz}_{\text{efect de interacție}} + \dots + \varepsilon$$

unde prin \dots înțelegem că în model ar putea intra și alți termeni care depind de alte covariabile. În general, ar fi trebuit să includem în model toți cei $2c - 1$ termeni: efecte principale x, z_1, \dots, z_{c-1} și interacții xz_1, \dots, xz_{c-1} .

În ceea ce privește interpretarea termenilor modelului, aceasta se poate face atât din perspectiva covariabilei continue cât și din cea a celei discrete. Astfel, în raport cu covariabila x , $\beta_1 x$ este efectul liniar al covariabilei x atunci când $z = 0$ iar $\beta_2 + (\beta_1 + \beta_3)x$ este efectul liniar indus de x atunci când $z = 1$. Alternativ, în raport cu z , $\beta_2 + \beta_3 x$ reprezintă efectul produs de diferența dintre observațiile pentru care $z = 1$ și $z = 0$. Această diferență depinde de valorile covariabilei x , nu este constantă ca și în cazul unui model fără interacții.

Exemplu: prețul chiriilor din Munchen - interacția dintre o covariabilă continuă și una discretă

Exp. 4.13 În acest exemplu vom ilustra efectul de interacție dintre o variabilă explicativă continuă și una discretă folosindu-ne de setul de date **Munchen**. Ne propunem să modelăm relația dintre prețul net al chiriei unui apartament în funcție de suprafața acestuia și tipul de locație în care se află. În acest context avem de-a face cu o variabilă predictor cu trei categorii (locație de top, bună și respectiv medie) pentru care vom introduce două variabile ajutătoare de tip indicator *alocatie* și *glocatie*, ținând ca nivel de referință categoria locațiilor de top. Presupunând că atât efectele principale cât și cele datorate interacțiilor intră liniar în model, avem

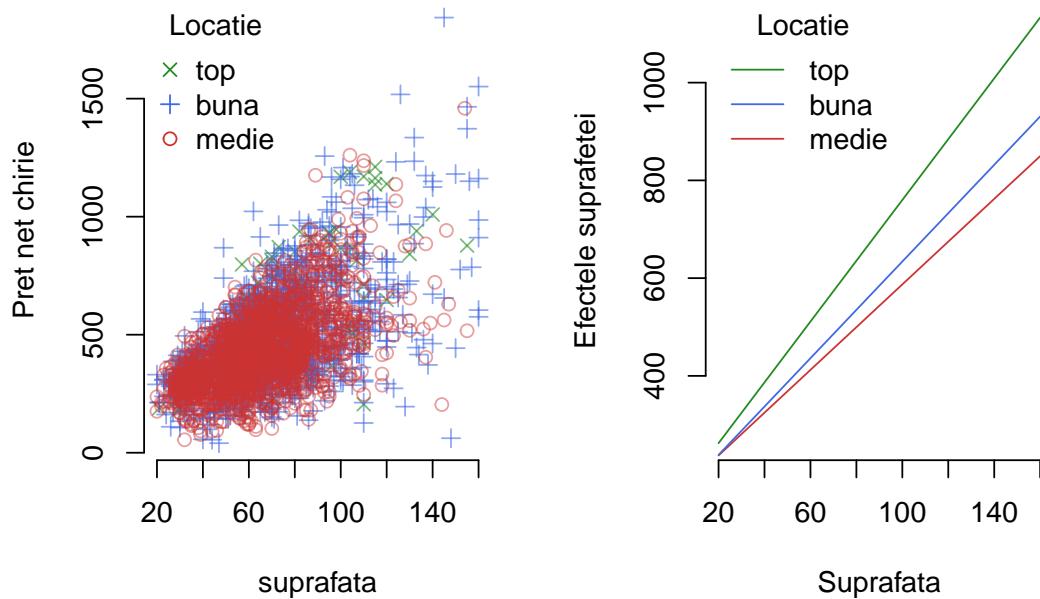
$$\begin{aligned} \text{pret}_i = & \beta_0 + \beta_1 \times \text{suprafata}_i + \beta_2 \times \text{alocatie}_i + \beta_3 \times \text{glocatie}_i + \\ & + \beta_4 \times \text{suprafata}_i \cdot \text{alocatie}_i + \beta_5 \times \text{suprafata}_i \cdot \text{glocatie}_i + \varepsilon_i \end{aligned}$$

unde prin estimare devine

$$\widehat{\text{pret}}_i = 137.924 + 6.22 \times \text{suprafata}_i + 11.824 \times \text{alocatie}_i + 1.473 \times \text{glocatie}_i + \\ - 1.846 \times \text{suprafata}_i \cdot \text{alocatie}_i - 1.274 \times \text{suprafata}_i \cdot \text{glocatie}_i.$$

Putem interpreta efectele după cum urmează (acestea pot fi vizualizate în figura de mai jos):

- $\beta_1 \times \text{suprafata}_i = 6.22 \times \text{suprafata}_i$ - efectul suprafeței de locuit pentru apartamentele situate într-o locație de top asupra prețului mediu net al chiriei (*alocation* = 0 și *glocation* = 0)
- $\beta_2 + (\beta_1 + \beta_4) \times \text{suprafata}_i = 11.824 + 4.374 \times \text{suprafata}_i$ - efectul suprafeței de locuit pentru apartamentele situate într-o locație medie
- $\beta_3 + (\beta_1 + \beta_5) \times \text{suprafata}_i = 1.473 + 4.946 \times \text{suprafata}_i$ - efectul suprafeței de locuit pentru apartamentele situate într-o locație bună
- $\beta_3 + \beta_5 \times \text{suprafata}_i = 1.473 - 1.274 \times \text{suprafata}_i$ - efectul diferență pentru apartamentele care se află într-o locație bună față de cele care se află într-o locație de top; coeficientul $\hat{\beta}_5 = -1.274$ negativ implică faptul că prețul mediu net al chiriilor pentru apartamentele aflate într-o zonă de top sunt întotdeauna mai mari decât cele dintr-o zonă bună, pentru fiecare metru pătrat adăugat suprafeței de locuit face ca diferența dintre prețul mediu net al chiriei pentru locații bune și să scadă cu 1.274 Euro față de cel al locuințelor situate în locații de top
- $\beta_2 + \beta_4 \times \text{suprafata}_i = 11.824 - 1.846 \times \text{suprafata}_i$ - efectul diferență pentru apartamentele care se află într-o locație medie față de cele care se află într-o locație de top
- $\beta_3 - \beta_2 + (\beta_5 - \beta_4) \times \text{suprafata}_i = -10.351 + 0.572 \times \text{suprafata}_i$ - efectul diferență pentru apartamentele care se află într-o locație bună față de cele care se află într-o locație medie \square



În exemplul anterior am presupus că efectul principal induș de covariabila continuă și efectul de interacție dintre covariabila continuă și cea categorială este liniar asupra răspunsului mediu dar putem include în model și efecte neliniare. Să presupunem că efectul principal induș de covariabila x este $f_1(x)$ și efectul de interacție este $f_2(x)z$ (pentru cazul de liniaritate aveam $f_1(x) = \beta_1x$ și respectiv $f_2(x) = \beta_2x$) ceea ce conduce la modelul

$$y = \beta_0 + \beta_1z + f_1(x) + \underbrace{f_2(x)z}_{\text{efect de interacție}} + \dots + \varepsilon$$

Ca și în secțiunea **Cazul variabilelor explicative continue**, neliniaritatea efectelor poate fi dobândită atât prin folosirea de transformări asupra covariabilelor cât și prin intermediul polinoamelor. De exemplu, putem considera că $f_1(x) = \beta_2x + \beta_3 \log(x)$ iar $f_2(x) = \beta_4x + \beta_5x^2$ și în acest caz modelul devine

$$y = \beta_0 + \beta_1z + \beta_2x + \beta_3 \log(x) + \beta_4xz + \beta_5x^2z + \dots + \varepsilon.$$

Pentru interpretarea modelului avem:

- $f_1(x)$ - efectul neliniar induș de covariabila continuă x atunci când $z = 0$
- $\beta_1 + f_1(x) + f_2(x)$ - efectul neliniar induș de covariabila continuă x atunci când $z = 1$; cele două curbe $f_1(x)$ și $\beta_1 + f_1(x) + f_2(x)$ prezintă intensitatea interacției, în cazul lipsei acesteia curbele fiind paralele (ar fi o translație cu β_1)
- $\beta_1 + f_2(x)$ - efectul diferență apărut pentru observațiile cu $z = 1$ în raport cu cele pentru care $z = 0$

Pentru a facilita interpretarea, efectul principal $f_1(x)$ trebuie centrat în zero, același lucru nefiind necesar pentru f_2 .

Exemplu: prețul chiriilor din München - interacția dintre o covariabilă continuă și una discretă - efect neliniar

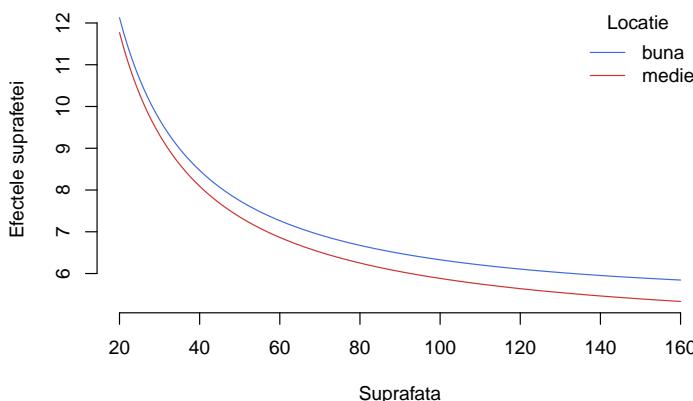
Exp. 4.14 Pentru a ilustra efectul neliniar de interacție dintre o covariabilă continuă și una discretă vom considera, în contextul setului de date **Munchen**, acele apartamente care se află în locații bune sau locații de tip mediu (vom folosi doar covariabila *glocation*). Dorim să modelăm relația dintre prețul chiriei pe metrul pătrat în funcție de suprafața de locuit și tipul de locație și vom folosi ca efect principal pentru suprafață transformarea $\frac{1}{x}$ ca și în Exemplul 4.7. Presupunând că avem un efect de interacție liniar între suprafață și tipul locației atunci avem modelul

$$pret_m^2_i = \beta_0 + \beta_1 \times glocation_i + f_1(suprafata_i) + f_2(suprafata_i) \cdot glocation_i + \varepsilon_i$$

unde $f_1(suprafata_i) = \beta_1 \left(\underbrace{\frac{1}{suprafata_i}}_{suprafata_inv_cen_i} - \overline{\frac{1}{suprafata_i}} \right)$ iar $f_2(suprafata_i) = \beta_3 \times suprafata_i$. În urma estimării găsim

$$\widehat{pret_m^2}_i = 6.918 + 0.334 \times glocation_i + 147.152 \times suprafata_inv_cen_i + 0.001suprafata_i \cdot glocation_i$$

Pentru interpretarea rezultatelor am inclus în figura de mai jos efectul estimat $\hat{f}_1(suprafata_i)$ a suprafeței de locuit în locații medii (cu roșu) și efectul estimat $\hat{\beta}_1 + \hat{f}_1(suprafata_i) + \hat{f}_2(suprafata_i)$ al suprafeței pentru apartamentele din locații bune (albastru).



Constatăm că efectele suprafeței pentru apartamentele situate în locații bune și medii sunt similare, chiar dacă cele două curbe nu sunt paralele (ceea ce implică un efect de interacție). Dacă suprafața de locuit crește atunci prețul mediu pe metrul pătrat scade pentru ambele locații. De asemenea, se observă că apartamentele situate în locații bune sunt întotdeauna mai scumpe decât cele din locații medii, indiferent de mărimea acestora. Diferența de preț este mai mică pentru apartamentele cu suprafață mai mică față de cele cu suprafață mai mare ceea ce sugerează că odată cu creșterea în suprafață găsim și o creștere în diferența medie de preț pe metrul pătrat între imobile situare în locații bune și medii, iar această diferență crește liniar.

4.4 Estimarea parametrilor

În această secțiune vom prezenta estimatorii parametrilor modelului de regresie liniară, β și σ^2 , precum și o serie de proprietăți statistice ale acestora fără a face ipoteze suplimentare asupra distribuției răspunsului. Înainte de a începe, vom reaminti că modelul de regresie liniară multiplă se poate scrie sub forma matricială

$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, unde \mathbf{Y} , \mathbf{X} , β , și ε sunt în conformitate cu Definiția 4.1. Ipotezele făcute asupra modelului sunt cele menționate în secțiunea Modelare și anume: $\mathcal{H}_1 : \text{rang}(\mathbf{X}) = p + 1$ ($n > p + 1$) și respectiv $\mathcal{H}_2 : \mathbb{E}[\varepsilon] = 0$, $\text{Var}(\varepsilon) = \sigma^2 I_n$.

4.4.1 Estimarea coeficientilor de regresie prin metoda celor mai mici pătrate

În conformitate cu metoda celor mai mici pătrate (ordinary least squares), coeficientii (necunoscuți) modelului de regresie β se estimează prin minimizarea costului total

$$\arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^n L \left(y_i - \underbrace{(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}_{f(\mathbf{x}_i)} \right)$$

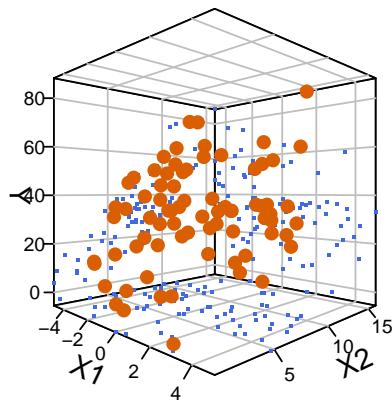
unde funcția de L este funcția de cost pătratic $L(u) = u^2$. Prin urmarea, estimatorul $\hat{\beta}$ obținut prin metoda celor mai mici pătrate este definit prin

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

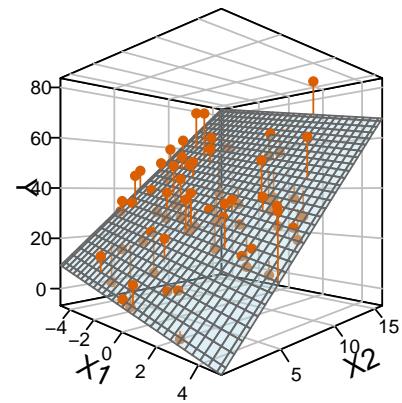
unde am folosit convenția $x_{i0} = 1$, $i = 1, 2, \dots, n$.

De exemplu, în cazul în care $p = 2$, în care avem doar două variabile explicative și putem afișa observațiile într-un cadru trei dimensional, figura de mai jos ilustrează, pentru un set de date simulat, diagrama de împrăștiere (cu proiecțiile pe fiecare plan - albastru) împreună cu planul de regresie obținut prin metoda celor mai mici pătrate (acel plan pentru care suma pătratelor distanțelor verticale până la plan este minimă).

Diagrama de împrăștiere



Planul de regresie



În propoziția următoare prezentăm forma estimatorului coeficienților de regresie obținut prin metoda celor mai mici pătrate:

Prop. 4.15



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

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

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

a) Metoda geometrică

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

Începem prin a reaminti câteva noțiuni de algebră liniară, pentru mai multe detalii se poate consulta monografia [Ornea and Turtoi, 2000]. Fiind dată matricea (de design) $\mathbf{X} \in \mathcal{M}_{n,p+1}(\mathbb{R})$ putem defini nucleul matricei

$$\ker(\mathbf{X}) = \{ \mathbf{u} \in \mathbb{R}^{p+1} \mid \mathbf{X}\mathbf{u} = 0 \}$$

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

$$\text{Im}(\mathbf{X}) = \{ \mathbf{v} \in \mathbb{R}^n \mid \exists \mathbf{u} \in \mathbb{R}^{p+1} \text{ a.î. } \mathbf{X}\mathbf{u} = \mathbf{v} \}.$$

Se poate arăta (a se vedea [Ornea and Turtoi, 2000, Capitolul 1]) că $\dim \ker(\mathbf{X}) = p + 1 - \text{rang}(\mathbf{X})$ și că $\dim \text{Im}(\mathbf{X}) = \text{rang}(\mathbf{X})$ ceea ce conduce la $\dim \ker(\mathbf{X}) + \dim \text{Im}(\mathbf{X}) = p + 1$ (caz particular al *teoremei lui Grassman*).

Reamintim că doi vectori \mathbf{u} și \mathbf{v} sunt ortogonali dacă $\langle \mathbf{u}, \mathbf{v} \rangle = 0$, iar în acest caz scriem $\mathbf{u} \perp \mathbf{v}$, și că două subspații V și W sunt ortogonale dacă $\forall \mathbf{v} \in V$ și respectiv $\forall \mathbf{w} \in W$ avem $\mathbf{v} \perp \mathbf{w}$ și notăm $V \perp W$. De asemenea, spațiul ortogonal al lui V este spațiul V^\perp care conține toți vectorii ortogonali pe V și are ca proprietăți: $V \cap V^\perp = \{0\}$ și $(V^\perp)^\perp = V$. Se poate arăta cu ușurință că dacă V este un subspațiu a lui $W \subset \mathbb{R}^n$ atunci pentru orice vector $\mathbf{w} \in W$ avem că $\mathbf{w} = \mathbf{v} + \mathbf{v}^\perp$ unde $\mathbf{v} \in V$ și $\mathbf{v}^\perp \in V^\perp$ iar descompunerea se face în mod unic (acest rezultat se mai scrie și sub forma $W = V \oplus V^\perp$). Se numește *proiecție ortogonală*⁴ a lui \mathbf{w} pe V de-a lungul lui V^\perp endomorfismul $p_V : W \rightarrow W$ definit prin $p_V(\mathbf{w}) = \mathbf{v}$. Se poate arăta cu ușurință că dacă V este un subspațiu vectorial al lui W atunci au loc următoarele proprietăți:

- i) pentru orice $\mathbf{w} \in W$ avem: $\mathbf{w} = p_V(\mathbf{w}) + p_{V^\perp}(\mathbf{w})$, $\|p_V(\mathbf{w})\| \leq \|\mathbf{w}\|$ iar $\mathbf{w} - p_V(\mathbf{w}) \perp \mathbf{v}$, $\forall \mathbf{v} \in V$
- ii) pentru orice $\mathbf{w} \in W$ are loc $\|\mathbf{w} - p_V(\mathbf{w})\| = \inf_{\mathbf{v} \in V} \|\mathbf{w} - \mathbf{v}\|$
- iii) dacă p este o proiecție ortogonală atunci $p \circ p = p$
- iv) dacă p este un endomorfism care verifică $p \circ p = p$ și în plus $\text{Im}(p) \perp \ker(p)$ atunci p este proiecția ortogonală pe $\text{Im}(p)$ de-a lungul lui $\ker(p)$

⁴În general, numim *proiecție* un endomorfism $p : W \rightarrow W$ cu proprietatea că există o descompunere în sumă directă $W = V_1 \oplus V_2$ (și spunem proiecție pe V_1 de-a lungul lui V_2) astfel încât $p(\mathbf{w}) = \mathbf{v}_1$, $\forall \mathbf{w} \in W$ cu $\mathbf{w} = \mathbf{v}_1 + \mathbf{v}_2$, $\mathbf{v}_1 \in V_1$ și $\mathbf{v}_2 \in V_2$.

v) dacă $\{e_1, e_2, \dots, e_k\}$ este o bază ortonormală în V atunci $p_V(\mathbf{w}) = \sum_{i=1}^k \langle \mathbf{w}, e_i \rangle e_i$

Spunem că o matrice pătratică $P \in \mathcal{M}_n(\mathbb{R})$ este o matrice de proiecție dacă este idempotentă $P^2 = P$ (numele vine de la faptul că pentru $\mathbf{x} \in \mathbb{R}^n$ aplicația liniară $P\mathbf{x}$ este proiecția lui \mathbf{x} pe $\text{Im}(P)$ de-a lungul lui $\ker(P)$ - a se vedea [Ornea and Turtoi, 2000, Capitolul 1, Secțiunea 5.2] sau [Yanai et al., 2011, Capitolul 2]). Dacă în plus matricea P este simetrică, i.e. $P = P^\top$, atunci $P\mathbf{x}$ este proiecția ortogonală a lui \mathbf{x} pe $V = \text{Im}(P)$ de-a lungul lui $V^\perp = \ker(P)$, cu alte cuvinte în descompunerea $\mathbf{x} = P\mathbf{x} + (I - P)\mathbf{x}$ vectorii $P\mathbf{x}$ și respectiv $(I - P)\mathbf{x}$ sunt ortogonali [Yanai et al., 2011, Capitolul 2, Secțiunea 2.2]. Prin urmare matricea P este o matrice de proiecție ortogonală dacă are loc relația $\mathbf{v} \perp \mathbf{v} - P\mathbf{v}$ adică $\langle \mathbf{v}, \mathbf{v} - P\mathbf{v} \rangle = 0$ pentru toți \mathbf{v} .

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

$$P_X = \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top.$$

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

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

$$P_v = \mathbf{v} (\mathbf{v}^\top \mathbf{v})^{-1} \mathbf{v}^\top = \frac{\mathbf{v} \mathbf{v}^\top}{\|\mathbf{v}\|^2}$$

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

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

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

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

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

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

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

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

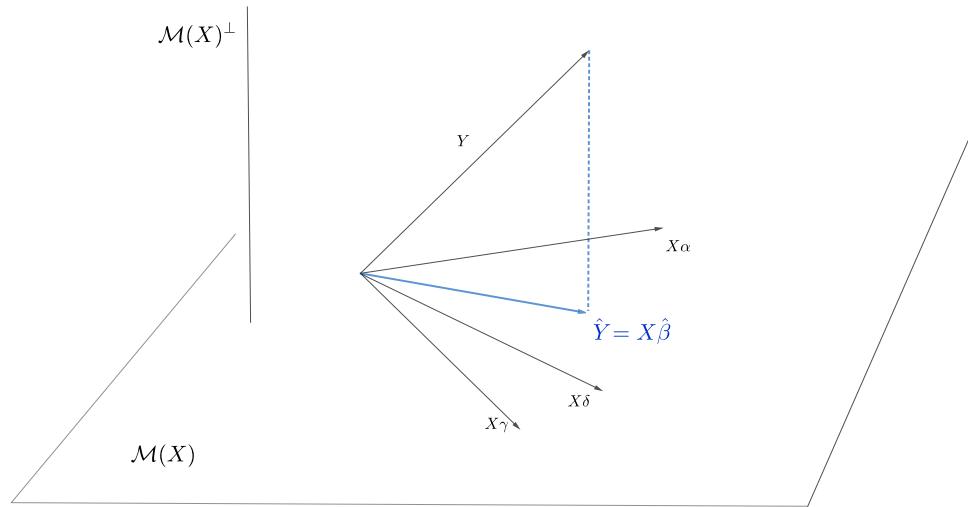
$$\left\{ \begin{array}{l} \langle \mathbf{X}_0, \mathbf{Y} - \hat{\mathbf{Y}} \rangle = 0 \\ \langle \mathbf{X}_1, \mathbf{Y} - \hat{\mathbf{Y}} \rangle = 0 \\ \vdots \\ \langle \mathbf{X}_p, \mathbf{Y} - \hat{\mathbf{Y}} \rangle = 0 \end{array} \right. \iff \left\{ \begin{array}{l} \langle \mathbf{X}_0, \mathbf{Y} - \mathbf{X} \hat{\beta} \rangle = 0 \\ \langle \mathbf{X}_1, \mathbf{Y} - \mathbf{X} \hat{\beta} \rangle = 0 \\ \vdots \\ \langle \mathbf{X}_p, \mathbf{Y} - \mathbf{X} \hat{\beta} \rangle = 0 \end{array} \right. \iff \mathbf{X}^\top (\mathbf{Y} - \mathbf{X} \hat{\beta}) = 0$$

de unde găsim *sistemul de ecuații normale*

$$\mathbf{X}^\top \mathbf{X} \hat{\beta} = \mathbf{X}^\top \mathbf{Y}$$

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

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$



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

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

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

b) Metoda analitică

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

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

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

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

$$\nabla f = \frac{\partial f}{\partial \mathbf{x}^\top} = \left(\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \dots \quad \frac{\partial f}{\partial x_k} \right).$$

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

$$\nabla f = \frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}^\top} = \left(\frac{\partial \sum_{i=1}^k a_i x_i}{\partial x_1} \quad \frac{\partial \sum_{i=1}^k a_i x_i}{\partial x_2} \quad \dots \quad \frac{\partial \sum_{i=1}^k a_i x_i}{\partial x_k} \right) = \mathbf{a}^\top$$

iar $\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$.

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

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

și

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

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

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

$$\mathbf{x}^\top A\mathbf{x} = \sum_{i=1}^k \sum_{j=1}^k a_{ij} x_i x_j = \sum_{i=1}^k a_{ii} x_i^2 + \sum_{i=1}^k \sum_{j=1, j \neq i}^k a_{ij} x_i x_j$$

de unde

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

prin urmare

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

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

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

$$\frac{\partial \mathbf{x}^\top A \mathbf{x}}{\partial \mathbf{x}} = 2A\mathbf{x}, \quad \frac{\partial \mathbf{x}^\top A \mathbf{x}}{\partial \mathbf{x}^\top} = 2\mathbf{x}^\top A^\top.$$

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

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

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

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

care atunci când ipoteza H_1 este adevarată, ceea ce conduce la inversabilitatea matricei $\mathbf{X}^\top \mathbf{X}$ (are valori proprii nenule), revine la

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

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

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

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

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

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

În cele ce urmează vom prezenta două cazuri particulare ale modelului de regresie liniară: modelul nul și respectiv modelul de regresie liniară simplă.

Exemplu: modelul nul

Exp. 4.16 În acest exemplu considerăm că modelul de regresie liniară nu conține niciun predictor ci doar termenul constant μ , prin urmare modelul se scrie

$$y_i = \mu + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

Acesta este un caz particular în care $p = 0$, $\mathbf{X} = \mathbf{1}$ iar $\boldsymbol{\beta} = \mu$. Cum $\mathbf{X}^\top \mathbf{X} = \mathbf{1}^\top \mathbf{1} = n$ găsim că estimatorul coeficientului de regresie obținut prin metoda celor mai mici pătrate este

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \frac{1}{n} \mathbf{1}^\top \mathbf{Y} = \bar{y}$$

Exemplu: modelul de regresie liniară simplă

Exp. 4.17 În contextul modelului de regresie liniară simplă

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

avem $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ iar $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$. Astfel găsim că

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \quad \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

iar $(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$ ceea ce conduce la estimatorul

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} (\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i) \\ -(\sum x_i)(\sum y_i) + n \sum x_i y_i \end{pmatrix}$$

care coincide (după mici manipulații algebrice) cu cel găsit anterior. \square

4.4.2 Proprietăți ale estimatorilor obținuți prin metoda celor mai mici pătrate

În această subsecțiune vom prezenta o serie de proprietăți statistice ale estimatorilor coeficienților de regresie obținuți prin metoda celor mai mici pătrate fără a face ipoteze suplimentare asupra distribuției răspunsului.

Prop. 4.18



În ipotezele \mathcal{H}_1 și \mathcal{H}_2 , estimatorul $\hat{\boldsymbol{\beta}}$ obținut prin metoda celor mai mici pătrate este nedeplasat, i.e. $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ iar matricea de varianță-covarianță $Var(\hat{\boldsymbol{\beta}})$ este

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$$

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

$$\mathbb{E}[\mathbf{A}\mathbf{Z}\mathbf{B} + \mathbf{C}] = \mathbf{A}\mathbb{E}[\mathbf{Z}]\mathbf{B} + \mathbf{C}.$$

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

$$Cov(\mathbf{Z}, \mathbf{T}) = \mathbb{E}[(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])(\mathbf{T} - \mathbb{E}[\mathbf{T}])^\top] = \mathbb{E}[\mathbf{Z}\mathbf{T}^\top] - \mathbb{E}[\mathbf{Z}]\mathbb{E}[\mathbf{T}]^\top = Cov(\mathbf{T}, \mathbf{Z})^\top.$$

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

$$Var(\mathbf{Z}) = \mathbb{E}[(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])^\top] = \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top] - \mathbb{E}[\mathbf{Z}]\mathbb{E}[\mathbf{Z}]^\top.$$

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

$$Cov(\mathbf{A}\mathbf{Z}, \mathbf{B}\mathbf{T}) = \mathbf{A}Cov(\mathbf{Z}, \mathbf{T})\mathbf{B}^\top$$

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

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

și respectiv

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

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

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\mathbf{Y}] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\mathbf{X}\beta + \varepsilon]$$

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

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

Pentru matricea de varianță-covarianță avem

$$\begin{aligned} Var(\hat{\beta}) &= Var((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Var(\mathbf{Y}) ((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top)^\top \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Var(\mathbf{Y}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \end{aligned}$$

unde în ultima egalitate am ținut cont de simetria matricei $\mathbf{X}^\top \mathbf{X}$: $((\mathbf{X}^\top \mathbf{X})^{-1})^\top = (\mathbf{X}^\top \mathbf{X})^{-1}$. Din modelul de regresie avem că $Var(\mathbf{Y}) = Var(\mathbf{X}\beta + \varepsilon) = Var(\varepsilon)$ și din ipoteza \mathcal{H}_2 avem $Var(\varepsilon) = \sigma^2 I_n$ prin urmare

$$Var(\hat{\beta}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \underbrace{Var(\mathbf{Y})}_{=\sigma^2 I_n} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} = \sigma^2 \underbrace{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}}_{=I_{p+1}} (\mathbf{X}^\top \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}. \square$$

Trebuie remarcat că atunci când variabilele explicative sunt ortogonale două câte două (și prin urmare și coloanele matricei de design), componentele vectorului β sunt necorelate deoarece matricea simetrică $\mathbf{X}^\top \mathbf{X}$ este o matrice diagonală.

Următoarea propoziție ne spune că estimatorul obținut prin metoda celor mai mici pătrate este optimal într-o anumită clasă de estimatori.

Prop. 4.19



Estimatorul $\hat{\beta}$ obținut prin metoda celor mai mici pătrate este estimatorul liniar, nedeplasat de varianță minimală (**BLUE**).

Rezultatul acestei propoziții este cunoscut și sub denumirea de **Teorema Gauss-Markov** (a se vedea de exemplu [Faraway, 2015, Capitolul 2]). Pentru a putea demonstra acest rezultat trebuie pentru început să remarcăm câteva aspecte: în primul rând, *liniaritatea* estimatorului se referă la *liniaritatea în raport cu \mathbf{Y}* , adică vorbim de clasa estimatorilor de forma \mathbf{AY} unde $\mathbf{A} \in \mathcal{M}_{p+1,n}(\mathbb{R})$; în al doilea rând pentru a determina *varianța minimală* avem nevoie de o relație parțială de ordine pe multimea matricelor simetrice, ori o asemenea relație există și spune că $\mathbf{S}_1 \leq \mathbf{S}_2$ atunci când $\mathbf{S} = \mathbf{S}_2 - \mathbf{S}_1$ este pozitiv semidefinită (valorile proprii ale lui \mathbf{S} sunt ≥ 0), cu alte cuvinte $\mathbf{S}_1 \leq \mathbf{S}_2$ dacă $\mathbf{x}^\top \mathbf{S}_1 \mathbf{x} \leq \mathbf{x}^\top \mathbf{S}_2 \mathbf{x}$ pentru orice \mathbf{x} .

În acest context, cum $\hat{\beta}$ este un estimator nedeplasat (conform exercițiului anterior) și liniar deoarece $\hat{\beta} = \underbrace{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}}_A$, considerând $\tilde{\beta}$ un alt estimator liniar și nedeplasat pentru β vrem să arătăm că $Var(\tilde{\beta}) \geq Var(\hat{\beta})$. Avem, conform proprietăților matricei de varianță-covarianță de mai sus, descompunerea

$$Var(\tilde{\beta}) = Var(\tilde{\beta} - \hat{\beta} + \hat{\beta}) = Var(\tilde{\beta} - \hat{\beta}) + Var(\hat{\beta}) + Cov(\tilde{\beta} - \hat{\beta}, \hat{\beta}) + Cov(\hat{\beta}, \tilde{\beta} - \hat{\beta}).$$

Cum matricele de varianță-covarianță sunt pozitiv semidefinite (pentru $\mathbf{a} \in \mathbb{R}^{p+1}$ avem $\mathbf{a}^\top Var(\tilde{\beta}) \mathbf{a} = Var(\mathbf{a}^\top \tilde{\beta}) \geq 0$, deoarece varianța este pozitivă) este suficient să arătăm că $Cov(\tilde{\beta} - \hat{\beta}, \hat{\beta}) = 0$. Liniaritatea estimatorului $\tilde{\beta}$ implică $\tilde{\beta} = \mathbf{AY}$ iar nedeplasarea acestuia, $\mathbb{E}[\tilde{\beta}] = \beta$ pentru orice β , conduce la $\mathbb{E}[\mathbf{AY}] = \mathbf{AX}\beta = \beta$ pentru orice β de unde $\mathbf{AX} = I_{p+1}$.

În final avem

$$\begin{aligned} Cov(\tilde{\beta} - \hat{\beta}, \hat{\beta}) &= Cov(\tilde{\beta}, \hat{\beta}) - Var(\hat{\beta}) = Cov(\mathbf{AY}, (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}) - Var(\hat{\beta}) \\ &= \mathbf{ACov}(\mathbf{Y}, \mathbf{Y}) ((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top)^\top - Var(\hat{\beta}) = \mathbf{A} \underbrace{Var(\mathbf{Y})}_{=\sigma^2 I_n} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} - \underbrace{Var(\hat{\beta})}_{=\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}} \\ &= \sigma^2 \underbrace{\mathbf{AX} (\mathbf{X}^\top \mathbf{X})^{-1}}_{=I_{p+1}} - \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} = 0. \square \end{aligned}$$

Trebuie remarcat că teorema Gauss-Markov poate fi extinsă la o combinație liniară $b_0\beta_0 + \dots + b_p\beta_p = \mathbf{b}^\top \beta$ a lui β , mai precis putem afirma că cel mai bun estimator liniar, nedeplasat și de varianță minimală pentru $\mathbf{b}^\top \beta$ este $\mathbf{b}^\top \hat{\beta}$ unde $\hat{\beta}$ este estimatorul obținut prin metode celor mai mici pătrate.

O altă proprietate interesantă a lui $\hat{\beta}$ este că valorile ajustate $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ (media (condiționată) a lui \mathbf{Y} , $\widehat{\mathbb{E}[\mathbf{Y}]}$) sunt invariante la simple schimbări liniare de scală a covariabilelor. Cu alte cuvinte, se poate arăta că dacă $\mathbf{Z}_j = c_j \mathbf{X}_j$, $j = 1, 2, \dots, p$ și c_j sunt coeficienți constanți, atunci $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{Z}\hat{\beta}_Z$ unde $\hat{\beta}_Z$ este estimatorul obținut prin metoda celor mai mici pătrate în modelul de regresie $\mathbf{Y} = \mathbf{Z}\beta_Z + \varepsilon$. Acest rezultat poate fi extins la o clasă mai largă de transformări liniare [Rencher and Schaalje, 2008].

4.4.3 Reziduuri și varianța reziduală

Am văzut (în metoda geometrică de determinare a lui β din Propoziția 4.15) că plecând de la estimatorul obținut prin metoda celor mai mici pătrate pentru coeficientii modelului de regresie liniară, putem estima media (conditionată) a vectorului răspuns \mathbf{Y} prin

$$\widehat{\mathbb{E}[\mathbf{Y}]} = \hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$$

ceea ce conduce la $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \underbrace{\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top}_{P_X \text{ (sau } H)} \mathbf{Y}$. Prin intermediul matricei de proiecție P_X putem exprima vectorul valorilor reziduale $\varepsilon_i = y_i - \hat{y}_i$ prin

$$\varepsilon = \mathbf{Y} - \hat{\mathbf{Y}} = (I - P_X)\mathbf{Y}.$$

Vectorul valorilor reziduale verifică următoarele proprietăți:

Prop. 4.20



Fie $\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$ vectorul valorilor reziduale. Atunci sub ipotezele \mathcal{H}_1 și \mathcal{H}_2 au loc relațiile

1. $\mathbb{E}[\hat{\varepsilon}] = 0$ și $Var(\hat{\varepsilon}) = \sigma^2 P_{X^\perp}$
2. $\mathbb{E}[\hat{\mathbf{Y}}] = \mathbf{X}\beta$ și $Var(\hat{\mathbf{Y}}) = \sigma^2 P_X$
3. $Cov(\hat{\varepsilon}, \hat{\mathbf{Y}}) = 0$

1. Am văzut că $\hat{\mathbf{Y}} = P_X \mathbf{Y}$ prin urmare ținând cont de modelul de regresie $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ găsim că

$$\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)^\top = \mathbf{Y} - \hat{\mathbf{Y}} = (I - P_X)\mathbf{Y} = P_{X^\perp}\mathbf{Y} = P_{X^\perp}(\mathbf{X}\beta + \varepsilon) = P_{X^\perp}\varepsilon$$

deoarece $\mathbf{X}\beta \in \mathcal{M}(X)$, deci $P_{X^\perp}\mathbf{X}\beta = 0$. Astfel $\hat{\varepsilon} \in \mathcal{M}(X)^\perp$ și

$$\mathbb{E}[\hat{\varepsilon}] = \mathbb{E}[P_{X^\perp}\varepsilon] = P_{X^\perp}\mathbb{E}[\varepsilon] = 0$$

iar, ținând seama de proprietatea de simetrie $P_{X^\perp}^\top = P_{X^\perp}$ și idempotență $P_{X^\perp}^2 = P_{X^\perp}$ a matricei de proiecție,

$$Var(\hat{\varepsilon}) = Var(P_{X^\perp}\varepsilon) = P_{X^\perp}Var(\varepsilon)P_{X^\perp}^\top = P_{X^\perp}\underbrace{Var(\varepsilon)}_{=\sigma^2 I_n}P_{X^\perp} = \sigma^2 P_{X^\perp}^2 = \sigma^2 P_{X^\perp}.$$

2. Din nedeplasarea lui $\hat{\beta}$ și definiția lui $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ avem

$$\mathbb{E}[\hat{\mathbf{Y}}] = \mathbb{E}[\mathbf{X}\hat{\beta}] = \mathbf{X}\mathbb{E}[\hat{\beta}] = \mathbf{X}\beta.$$

În mod similar

$$Var(\hat{\mathbf{Y}}) = Var(\mathbf{X}\hat{\beta}) = \mathbf{X}\underbrace{Var(\hat{\beta})}_{=\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}}\mathbf{X}^\top = \sigma^2 \underbrace{\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top}_{=P_X} = \sigma^2 P_X.$$

3. Pentru a calcula $Cov(\hat{\varepsilon}, \hat{\mathbf{Y}})$ folosim $Var(\hat{\varepsilon}) = \sigma^2 P_{X^\perp}$ și $Var(\mathbf{Y}) = Var(\varepsilon) = \sigma^2 I_n$ și avem

$$\begin{aligned} Cov(\hat{\varepsilon}, \hat{Y}) &= Cov(\hat{\varepsilon}, Y - \hat{\varepsilon}) = Cov(\hat{\varepsilon}, Y) - Var(\hat{\varepsilon}) = Cov(P_{X^\perp} Y, Y) - \sigma^2 P_{X^\perp} \\ &= P_{X^\perp} Var(Y) P_{X^\perp}^\top - \sigma^2 P_{X^\perp} = \sigma^2 P_{X^\perp} - \sigma^2 P_{X^\perp} = 0. \square \end{aligned}$$

Tinând cont de faptul că matricea de design \mathbf{X} are coloanele ortogonale pe vectorul valorilor reziduale $\hat{\varepsilon}$,

$$\mathbf{X}^\top \hat{\varepsilon} = \mathbf{X}^\top (I - P_X) \mathbf{Y} = \mathbf{X}^\top \mathbf{Y} - \mathbf{X}^\top \underbrace{\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top}_{P_X} \mathbf{Y} = \mathbf{X}^\top \mathbf{Y} - \mathbf{X}^\top \mathbf{Y} = \mathbf{0},$$

apar o serie de proprietăți interesante. Ca și în cazul modelului de regresie liniară simplă, se poate verifica cu ușurință că suma valorilor reziduale este nulă

$$\sum_{i=1}^n \hat{\varepsilon}_i = \mathbf{1}^\top \hat{\varepsilon} = 0,$$

unde în ultima egalitate am ținut cont că $\mathbf{1} \in \mathcal{M}(X)$ iar $\hat{\varepsilon} \in \mathcal{M}(X)^\perp$, ceea ce implică la rândul ei că media eșantionului valorilor ajustate este egală cu media eșantionului răspunsului observat, $\frac{1}{n} \sum_{i=1}^n \hat{y}_i = \bar{y}$. Aceasta din urmă observație conduce la rezultatul similar regresiei liniare simple, și anume că hiperplanul de regresie trece prin mijlocul norului de puncte, i.e. punctul $(\bar{x}_1, \dots, \bar{x}_p, \bar{y})$, mai exact avem (prin sumare)

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \dots + \hat{\beta}_p \bar{x}_p.$$

Următorul exemplu arată cum putem crea o matrice de design ortogonal, procedeul fiind cunoscut sub numele de procedeul Gram-Schmidt:

Exemplu: ortogonalizarea matricei de design

Exp. 4.21 O matrice de design ortogonală implică, pe lângă alte proprietăți, că varibilele explicative sunt necorelate prin urmare ortogonalizarea matricei de design se dovedește foarte utilă, de exemplu atunci când construim polinoame ortogonale (a se vedea Exemplul 4.10).

Fie \mathbf{X} matricea de design și \mathbf{X}_j coloana j a acesteia. Scopul este de a crea o nouă matrice $\tilde{\mathbf{X}}$ ale cărei coloane $\tilde{\mathbf{X}}_j$ sunt ortogonale. Astfel pentru fiecare $j = 2, \dots, p+1$ considerăm transformarea

$$\tilde{\mathbf{X}}_j = \mathbf{X}_j - \tilde{\mathbf{Z}}_j \left(\tilde{\mathbf{Z}}_j^\top \tilde{\mathbf{Z}}_j \right)^{-1} \tilde{\mathbf{Z}}_j^\top \mathbf{X}_j,$$

unde $\tilde{\mathbf{Z}}_j$ este matricea care conține primele $j-1$ coloane transformate $\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_{j-1}$, $\tilde{\mathbf{Z}}_j = \mathbf{1}$. Prima coloană a matricei de design \mathbf{X} rămâne netransformată prin urmare $\tilde{\mathbf{X}}_2$ nu este altceva decât \mathbf{X}_2 centralizat. Putem observa că vectorul coloană transformat $\tilde{\mathbf{X}}_j$ poate fi interpretat ca vectorul valorilor reziduale din modelul de regresie care consideră ca variabilă răspuns pe \mathbf{X}_j și ca variabile predictor pe $\mathbf{1}, \tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_{j-1}$. Datorită ortogonalității dintre valorile reziduale și coloanele matricei de design, găsim că $\tilde{\mathbf{X}}_j$ este ortogonal pe $\tilde{\mathbf{X}}_i$, $i = 1, \dots, j-1$. \square

Rezultatul următoarei propoziții ne spune că în cazul în care variabilele explicative sunt ortogonale, a efectua o regresie multiplă revine la a efectua p regresii simple.



Considerăm modelul de regresie liniară

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

Prop. 4.22

unde $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathcal{M}_{n,p}(\mathbb{R})$ este o matrice compusă din p vectori ortogonali, $\boldsymbol{\beta} \in \mathbb{R}^p$ iar $\boldsymbol{\varepsilon} \in \mathbb{R}^n$. Fie \mathbf{Z} matricea formată din primele q coloane ale lui \mathbf{X} și \mathbf{U} matricea formată din ultimele $p - q$ coloane ale lui \mathbf{X} . Prin metoda celor mai mici pătrate obținem

$$\begin{aligned}\hat{\mathbf{Y}}_X &= \hat{\beta}_1^X \mathbf{X}_1 + \cdots + \hat{\beta}_p^X \mathbf{X}_p \\ \hat{\mathbf{Y}}_Z &= \hat{\beta}_1^Z \mathbf{X}_1 + \cdots + \hat{\beta}_q^Z \mathbf{X}_q \\ \hat{\mathbf{Y}}_U &= \hat{\beta}_{q+1}^U \mathbf{X}_{q+1} + \cdots + \hat{\beta}_p^U \mathbf{X}_p\end{aligned}$$

1. Atunci $\|P_X \mathbf{Y}\|^2 = \|P_Z \mathbf{Y}\|^2 + \|P_U \mathbf{Y}\|^2$.
2. Pentru $i \in \{1, 2, \dots, p\}$ dat, avem că $\hat{\beta}_i^X = \hat{\beta}_i^Z$ dacă $i \leq q$ și $\hat{\beta}_i^X = \hat{\beta}_i^U$ altfel.

1. Cum $P_Z + P_{Z^\perp} = I$ putem scrie

$$\hat{\mathbf{Y}}_X = P_X \mathbf{Y} = (P_Z + P_{Z^\perp}) P_X \mathbf{Y} = P_Z P_X \mathbf{Y} + P_{Z^\perp} P_X \mathbf{Y},$$

și din $P_Z P_X = P_{Z \cap X} = P_Z$ găsim că $\hat{\mathbf{Y}}_X = P_Z \mathbf{Y} + P_{Z^\perp} P_X \mathbf{Y}$. De asemenea, să notăm că $P_{Z^\perp} P_X = P_{Z^\perp \cap X}$ și înținând cont de faptul că matricea \mathbf{X} are coloanele ortogonale avem $P_{Z^\perp \cap X} = P_U$ (în fapt ortogonalitatea coloanelor lui \mathbf{X} implică $\mathcal{M}(X) = \mathcal{M}(Z) \overset{\perp}{\oplus} \mathcal{M}(U)$, sumă directă de spații ortogonale). Prin urmare obținem descompunerea ortogonală

$$\hat{\mathbf{Y}}_X = P_Z \mathbf{Y} + P_U \mathbf{Y} = \hat{\mathbf{Y}}_Z + \hat{\mathbf{Y}}_U$$

și din Teorema lui Pitagora avem

$$\|P_X \mathbf{Y}\|^2 = \|P_Z \mathbf{Y}\|^2 + \|P_U \mathbf{Y}\|^2.$$

2. Vom arăta relația pentru $i \leq q$, cazul general fiind analog. Din formula generală avem că $\hat{\boldsymbol{\beta}}^X$ este

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

și cum coloanele lui \mathbf{X} sunt ortogonale deducem că matricea $\mathbf{X}^\top \mathbf{X}$ este o matrice diagonală, $\mathbf{X}^\top \mathbf{X} = \text{diag}(\|\mathbf{X}_1\|^2, \dots, \|\mathbf{X}_p\|^2)$. Notând de asemenea că $\mathbf{X}^\top \mathbf{Y}$ este un vector coloană cu elemente de tipul $\langle \mathbf{X}_i, \mathbf{Y} \rangle$, deducem că

$$\hat{\boldsymbol{\beta}}^X = \begin{pmatrix} \frac{\langle \mathbf{X}_1, \mathbf{Y} \rangle}{\|\mathbf{X}_1\|^2} & \dots & \frac{\langle \mathbf{X}_p, \mathbf{Y} \rangle}{\|\mathbf{X}_p\|^2} \end{pmatrix}^\top,$$

prin urmare $\hat{\beta}_i^X = \frac{\langle \mathbf{X}_i, \mathbf{Y} \rangle}{\|\mathbf{X}_i\|^2}$. Pentru $i \leq q$, coloana i a matricei \mathbf{Z} este $\mathbf{Z}_i = \mathbf{X}_i$ și aplicând raționamentul anterior găsim că

$$\hat{\beta}_i^Z = \frac{\langle \mathbf{Z}_i, \mathbf{Y} \rangle}{\|\mathbf{Z}_i\|^2} = \frac{\langle \mathbf{X}_i, \mathbf{Y} \rangle}{\|\mathbf{X}_i\|^2} = \hat{\beta}_i^X. \square$$

Ca și în cazul modelului de regresie liniară simplă, un estimator *natural* al varianței reziduale σ^2 ar fi

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n} = \frac{1}{n} \|\hat{\boldsymbol{\varepsilon}}\|^2,$$

care se dovedește a coincide cu estimatorul de verosimilitate maximă (a se vedea Propoziția 4.33). Chiar dacă acest estimator este deplasat putem determina cu ușurință unul nedeplasat după cum arată și rezultatul următor.

Prop. 4.23



Statistica $\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-(p+1)}$ este un estimator nedeplasat pentru σ^2 .

Ca să arătăm că $\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-(p+1)}$ este un estimator nedeplasat pentru σ^2 trebuie să calculăm $\mathbb{E}[\|\hat{\varepsilon}\|^2]$. Cum $\|\hat{\varepsilon}\|^2$ este un scalar atunci $\|\hat{\varepsilon}\|^2 = \text{Tr}(\|\hat{\varepsilon}\|^2)$ (este egal cu urma sa) prin urmare

$$\mathbb{E}[\|\hat{\varepsilon}\|^2] = \mathbb{E}[\text{Tr}(\|\hat{\varepsilon}\|^2)] = \mathbb{E}[\text{Tr}(\hat{\varepsilon}^\top \hat{\varepsilon})]$$

și cum pentru orice matrice \mathbf{A} urma verifică $\text{Tr}(\mathbf{A}^\top \mathbf{A}) = \text{Tr}(\mathbf{A} \mathbf{A}^\top) = \sum_{i,j} a_{ij}^2$ avem ($\mathbb{E}[\hat{\varepsilon}] = 0$)

$$\begin{aligned} \mathbb{E}[\|\hat{\varepsilon}\|^2] &= \mathbb{E}[\text{Tr}(\hat{\varepsilon}^\top \hat{\varepsilon})] = \mathbb{E}[\text{Tr}(\hat{\varepsilon} \hat{\varepsilon}^\top)] = \text{Tr}(\mathbb{E}[\hat{\varepsilon} \hat{\varepsilon}^\top]) = \text{Tr}(\mathbb{E}[\hat{\varepsilon} \hat{\varepsilon}^\top] - \mathbb{E}[\hat{\varepsilon}] \mathbb{E}[\hat{\varepsilon}]^\top) \\ &= \text{Tr}(Var(\hat{\varepsilon})) = \text{Tr}(\sigma^2 P_{X^\perp}) = \sigma^2 \text{Tr}(P_{X^\perp}). \end{aligned}$$

Cum $P_{X^\perp} = I - P_X$ este matrice de proiecție avem că $\text{Tr}(P_{X^\perp}) = \text{rang}(P_{X^\perp}) = n - (p + 1)$ (urma matricei de proiecție este egală cu dimensiunea spațiului pe care proiectăm) prin urmare

$$\mathbb{E}[\|\hat{\varepsilon}\|^2] = \sigma^2 [n - (p + 1)]$$

de unde $\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}\left[\frac{\|\hat{\varepsilon}\|^2}{n-(p+1)}\right] = \sigma^2$. \square

Trebuie menționat că estimatorul nedeplasat $\hat{\sigma}^2$ pentru σ^2 coincide, în cazul modelului (condiționat) normal (a se vedea secțiunea Modelul (condiționat) normal), cu estimatorul de verosimilitate maximă restrâns (REML - restricted maximum likelihood estimator) care este definit ca

$$\arg \max_{\sigma^2} L(\sigma^2; \mathbf{Y}),$$

unde $L(\sigma^2; \mathbf{Y}) = \int L(\boldsymbol{\beta}, \sigma^2; \mathbf{Y}) d\boldsymbol{\beta}$. În general acest estimator este mai puțin deplasat față de estimatorul de verosimilitate maximă și prin urmare este de preferat în practică.

Folosind estimatorul $\hat{\sigma}^2$ pentru σ^2 putem construi un estimator $\hat{\sigma}_{\hat{\beta}}^2$ pentru varianța $Var(\hat{\beta}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$:

$$\hat{\sigma}_{\hat{\beta}}^2 = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1} = \frac{\|\hat{\varepsilon}\|^2}{n - (p + 1)} (\mathbf{X}^\top \mathbf{X})^{-1}.$$

În particular, un estimator pentru abaterea standard a estimatorului coeficientului β_j , $\hat{\beta}_j$, este dat de elementul $j + 1$ de pe diagonala matricii $(\mathbf{X}^\top \mathbf{X})^{-1}$:

$$\hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma} \sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{j+1,j+1}}, \quad j = 0, 1, \dots, p.$$

În exemplul următor vom ilustra numeric aceste rezultate:

Exemplu: prețul chirilor din Munchen

Exp. 4.24 Să considerăm modelul propus în Exemplul 4.10 prin care exprimăm prețul mediu al chiriei pe metrul pătrat pentru locuințele din orașul Munchen în funcție de efectul invers proporțional dat de suprafața de locuit și de anul de construcție a apartamentului:

$$pret_m^2_i = \beta_0 + \beta_1 \times \frac{1}{suprafata_i} + \beta_2 \times an_con_i + \varepsilon_i.$$

Matricea de design este

$$\mathbf{X} = \begin{pmatrix} 1 & \frac{1}{suprafata_1} & an_con_1 \\ 1 & \frac{1}{suprafata_2} & an_con_2 \\ \vdots & \vdots & \vdots \\ 1 & \frac{1}{suprafata_{3082}} & an_con_{3082} \end{pmatrix} = \begin{pmatrix} 1 & 0.029 & 1939 \\ 1 & 0.01 & 1939 \\ \vdots & \vdots & \vdots \\ 1 & 0.016 & 1953 \end{pmatrix}$$

iar estimatorul coeficienților de regresie obținut prin metoda celor mai mici pătrate este

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} -65.406 \\ 119.361 \\ 0.036 \end{pmatrix}.$$

Varianța reziduală estimată $\hat{\sigma}^2$ este $\hat{\sigma}^2 = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n-(p+1)} = 4.3993026$ iar estimatorul matricei de varianță-covarianță $Var(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$ devine

$$\hat{\sigma}_{\hat{\boldsymbol{\beta}}}^2 = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1} = \begin{pmatrix} 11.23 & 2.808 & -0.006 \\ 2.808 & 31.767 & -0.002 \\ -0.006 & -0.002 & 0 \end{pmatrix}.$$

Astfel găsim că estimatorii pentru abaterile standard ale coeficienților de regresie sunt $\hat{\sigma}_{\hat{\beta}_0} = 3.3511$, $\hat{\sigma}_{\hat{\beta}_1} = 5.6362$ și respectiv $\hat{\sigma}_{\hat{\beta}_2} = 0.0017$. \square

4.4.4 Predicție

Unul dintre scopurile aplicării unui model de regresie este acela de a prezice valoarea variabilei răspuns atunci când avem de-a face cu un set nou de valori pentru variabilele explicative considerate în model. Să presupunem că $(x_{n+1,1}, \dots, x_{n+1,p})$ sunt valori ale variabilelor explicative ce corespund unei noi observații și dorim să prezicem y_{n+1} conform modelului de regresie liniară

$$y_{n+1} = \beta_0 + \beta_1 x_{n+1,1} + \dots + \beta_p x_{n+1,p} + \varepsilon_{n+1}$$

pentru care $\mathbb{E}[\varepsilon_{n+1}] = 0$, $Var(\varepsilon_{n+1}) = \sigma^2$ și $Cov(\varepsilon_{n+1}, \varepsilon_i) = 0$ pentru $i = 1, \dots, n$. Metoda naturală, este de a prezice valoarea corespunzătoare prin intermediul valorii ajustate $\hat{y}_{n+1} = \mathbf{x}_{n+1}^\top \hat{\boldsymbol{\beta}}$ unde $\mathbf{x}_{n+1}^\top = (1, x_{n+1,1}, \dots, x_{n+1,p})$ și în acest caz eroarea de predicție este $\hat{\varepsilon}_{n+1} = y_{n+1} - \hat{y}_{n+1}$.

Trebuie remarcat că atunci când vorbim de predicție putem să ne referim sau la *predicție asupra răspunsului mediu* sau la *predicție asupra unei noi valori*. În prima situație ne referim la valoarea ajustată $\mathbf{x}_{n+1}^\top \hat{\boldsymbol{\beta}}$ în care doar varianța lui $\hat{\boldsymbol{\beta}}$ este luată în calcul atunci când vorbim de variabilitatea predicției. Plasându-ne în contextul setului de date referitor la prețul chirilor din orașul Munchen, acest caz revine la a răspunde la întrebare: Cu cât se va închiria, în medie, un apartament al cărui caracteristici sunt $(x_{n+1,1}, \dots, x_{n+1,p})$ (suprafață, an de construcție, etc.)? A doua situație se referă la prețul de închiriere $\mathbf{x}_{n+1}^\top \hat{\boldsymbol{\beta}} + \varepsilon_{n+1}$ al unui nou apartament cu particularitățile $(x_{n+1,1}, \dots, x_{n+1,p})$. În acest caz, variabilitatea predicției vine din două surse: varianța lui $\hat{\boldsymbol{\beta}}$ și respectiv a termenului eroare ε_{n+1} .

Rezultatul următor descrie varianța erorii de predicție.

Prop. 4.25



Fie $(x_{n+1,1}, \dots, x_{n+1,p})$ o nouă observație și considerăm $\mathbf{x}_{n+1}^\top = (1, x_{n+1,1}, \dots, x_{n+1,p})$. Ne propunem să prezicem valoarea y_{n+1} conform modelului

$$y_{n+1} = \mathbf{x}_{n+1}^\top \boldsymbol{\beta} + \varepsilon_{n+1}$$

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

Atunci eroarea de predicție $\hat{\varepsilon}_{n+1} = y_{n+1} - \hat{y}_{n+1}$ verifică proprietățile

1. $\mathbb{E}[\hat{\varepsilon}_{n+1}] = 0$
2. $Var(\hat{\varepsilon}_{n+1}) = \sigma^2 (1 + \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1})$

1. Cum $\mathbb{E}[\varepsilon_{n+1}] = 0$ și ținând cont de nedeplasarea estimatorului $\hat{\boldsymbol{\beta}}$ avem

$$\mathbb{E}[\hat{\varepsilon}_{n+1}] = \mathbb{E}[y_{n+1} - \hat{y}_{n+1}] = \mathbb{E} \left[\underbrace{\mathbf{x}_{n+1}^\top \boldsymbol{\beta} + \varepsilon_{n+1}}_{y_{n+1}} - \underbrace{\mathbf{x}_{n+1}^\top \hat{\boldsymbol{\beta}}}_{\hat{y}_{n+1}} \right]$$

de unde $\mathbb{E}[\hat{\varepsilon}_{n+1}] = \mathbf{x}_{n+1}^\top (\boldsymbol{\beta} - \mathbb{E}[\hat{\boldsymbol{\beta}}]) + \mathbb{E}[\varepsilon_{n+1}] = 0$.

Cum $\hat{\boldsymbol{\beta}}$ depinde doar de variabilele $\{\varepsilon_1, \dots, \varepsilon_n\}$ iar $Cov(\varepsilon_{n+1}, \varepsilon_i) = 0$ pentru $i = 1, \dots, n$ deducem că

$$\begin{aligned} Var(\hat{\varepsilon}_{n+1}) &= Var(y_{n+1} - \hat{y}_{n+1}) = Var(\mathbf{x}_{n+1}^\top (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \varepsilon_{n+1}) \\ &= Var(\mathbf{x}_{n+1}^\top (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})) + Var(\varepsilon_{n+1}) = \mathbf{x}_{n+1}^\top Var(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \mathbf{x}_{n+1} + \sigma^2 \\ &= \mathbf{x}_{n+1}^\top Var(\hat{\boldsymbol{\beta}}) \mathbf{x}_{n+1} + \sigma^2 = \mathbf{x}_{n+1}^\top \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1} + \sigma^2 \\ &= \sigma^2 (1 + \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1}). \square \end{aligned}$$

Observăm astfel că incertitudinea erorii de predicție este egală cu suma incertitudinii datorate lui $\boldsymbol{\beta}$, $\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$, și cea datorată lui ε_{n+1} , σ^2 .

Rezultatul următor generalizează rezultatul similar din cadrul modelului de regresie liniară simplă.

Prop. 4.26



Fie modelul de regresie $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ unde matricea de design se scrie

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}_{n \times (p+1)} = \begin{bmatrix} \underbrace{\mathbf{X}_0}_{\mathbf{1} = (1, \dots, 1)} & |\mathbf{X}_1| \cdots |\mathbf{X}_p \end{bmatrix} = [\mathbf{1} | \mathbf{Z}]$$

cu \mathbf{Z} matricea de dimensiune $n \times p$ formată din coloanele $\{\mathbf{X}_1, \dots, \mathbf{X}_p\}$. În contextul propoziției precedente, fie $\mathbf{z}_{n+1}^\top = (x_{n+1,1}, \dots, x_{n+1,p})$ o nouă observație și considerăm $\mathbf{x}_{n+1}^\top = (1, \mathbf{z}_{n+1}^\top)$. Arătați că varianța erorii de predicție este

$$Var(\hat{\varepsilon}_{n+1}) = \sigma^2 \left[1 + \frac{1}{n} + \frac{1}{n} (\mathbf{z}_{n+1} - \bar{\mathbf{x}})^\top \boldsymbol{\Gamma}^{-1} (\mathbf{z}_{n+1} - \bar{\mathbf{x}}) \right]$$

unde $\boldsymbol{\Gamma} = \frac{1}{n} \mathbf{Z}^\top \mathbf{Z} - \bar{\mathbf{x}} \bar{\mathbf{x}}^\top$ este o matrice simetrică și pozitiv definită atunci când $\mathbf{X}^\top \mathbf{X}$ este inversabilă iar $\bar{\mathbf{x}}^\top = (\bar{\mathbf{x}}_1^\top, \dots, \bar{\mathbf{x}}_p^\top)$ cu $\bar{\mathbf{x}}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$.

Pentru început să observăm că matricea $\mathbf{X}^\top \mathbf{X}$ se scrie sub forma

$$\begin{aligned} \mathbf{X}^\top \mathbf{X} &= \begin{pmatrix} \mathbf{1}^\top \\ \mathbf{X}_1^\top \\ \vdots \\ \mathbf{X}_p^\top \end{pmatrix} \begin{bmatrix} 1 & \mathbf{X}_1 & \cdots & \mathbf{X}_p \end{bmatrix} = \begin{pmatrix} \underbrace{\mathbf{1}^\top \mathbf{1}}_{=n} & \underbrace{\mathbf{1}^\top \mathbf{X}_1}_{=n\bar{\mathbf{x}}_1} & \cdots & \underbrace{\mathbf{1}^\top \mathbf{X}_p}_{=n\bar{\mathbf{x}}_p} \\ \mathbf{X}_1^\top \mathbf{1} & \mathbf{X}_1^\top \mathbf{X}_1 & \cdots & \mathbf{X}_1^\top \mathbf{X}_p \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{X}_p^\top \mathbf{1} & \mathbf{X}_p^\top \mathbf{X}_1 & \cdots & \mathbf{X}_p^\top \mathbf{X}_p \end{pmatrix} = \begin{pmatrix} n & n\bar{\mathbf{x}}_1 & \cdots & n\bar{\mathbf{x}}_p \\ n\bar{\mathbf{x}}_1 & \mathbf{X}_1^\top \mathbf{X}_1 & \cdots & \mathbf{X}_1^\top \mathbf{X}_p \\ \cdots & \cdots & \cdots & \cdots \\ n\bar{\mathbf{x}}_p & \mathbf{X}_p^\top \mathbf{X}_1 & \cdots & \mathbf{X}_p^\top \mathbf{X}_p \end{pmatrix} \\ &= \begin{pmatrix} n & n\bar{\mathbf{x}}^\top \\ n\bar{\mathbf{x}} & \mathbf{Z}^\top \mathbf{Z} \end{pmatrix} = n \begin{pmatrix} 1 & \bar{\mathbf{x}}^\top \\ \bar{\mathbf{x}} & \frac{1}{n} \mathbf{Z}^\top \mathbf{Z} \end{pmatrix}. \end{aligned}$$

Reamintim, e.g. [Henderson and Searle, 1981] sau [Rencher and Schaalje, 2008, Capitolul 2], că dacă \mathbf{F} este o matrice pătrată inversabilă care se scrie sub formă de bloc de patru submatrice

$$\mathbf{F} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$$

cu \mathbf{A} inversabilă atunci matricea $\mathbf{Q} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ este inversabilă iar inversa matricii \mathbf{F} este

$$\mathbf{F}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1} \mathbf{B} \mathbf{Q} \mathbf{C} \mathbf{A}^{-1} & -\mathbf{A}^{-1} \mathbf{B} \mathbf{Q} \\ -\mathbf{Q} \mathbf{C} \mathbf{A}^{-1} & \mathbf{Q} \end{pmatrix}.$$

În cazul problemei noastre avem $\mathbf{A} = 1$, $\mathbf{B} = \bar{\mathbf{x}}^\top$, $\mathbf{C} = \bar{\mathbf{x}}$ și $\mathbf{D} = \frac{1}{n} \mathbf{Z}^\top \mathbf{Z}$ prin urmare

$$\mathbf{Q} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} = \frac{1}{n} \mathbf{Z}^\top \mathbf{Z} - \bar{\mathbf{x}} \bar{\mathbf{x}}^\top = \boldsymbol{\Gamma}$$

iar

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{n} \begin{pmatrix} 1 & \bar{\mathbf{x}}^\top \\ \bar{\mathbf{x}} & \frac{1}{n} \mathbf{Z}^\top \mathbf{Z} \end{pmatrix}^{-1} = \frac{1}{n} \begin{pmatrix} 1 + \bar{\mathbf{x}}^\top \boldsymbol{\Gamma}^{-1} \bar{\mathbf{x}} & -\bar{\mathbf{x}}^\top \boldsymbol{\Gamma}^{-1} \\ -\boldsymbol{\Gamma}^{-1} \bar{\mathbf{x}} & \boldsymbol{\Gamma}^{-1} \end{pmatrix}.$$

Dat fiind $\mathbf{x}_{n+1}^\top = (1, \mathbf{z}_{n+1}^\top)$ am văzut în Propoziția 4.25 că varianța erorii de predicție este dată de formula

$$Var(\hat{\varepsilon}_{n+1}) = \sigma^2 (1 + \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1}).$$

Tinând cont de scrierea sub formă de blocuri a matricei $(\mathbf{X}^\top \mathbf{X})^{-1}$ găsim că

$$\begin{aligned} \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1} &= (1 \ z_{n+1}^\top) \frac{1}{n} \begin{pmatrix} 1 + \bar{\mathbf{x}}^\top \boldsymbol{\Gamma}^{-1} \bar{\mathbf{x}} & -\bar{\mathbf{x}}^\top \boldsymbol{\Gamma}^{-1} \\ -\boldsymbol{\Gamma}^{-1} \bar{\mathbf{x}} & \boldsymbol{\Gamma}^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ \mathbf{z}_{n+1} \end{pmatrix} \\ &= \frac{1}{n} (1 + \bar{\mathbf{x}}^\top \boldsymbol{\Gamma}^{-1} \bar{\mathbf{x}} - \mathbf{z}_{n+1}^\top \boldsymbol{\Gamma}^{-1} \bar{\mathbf{x}} - \bar{\mathbf{x}}^\top \boldsymbol{\Gamma}^{-1} \mathbf{z}_{n+1} + \mathbf{z}_{n+1}^\top \boldsymbol{\Gamma}^{-1} \mathbf{z}_{n+1}). \end{aligned}$$

Notând cu \mathbf{Z}_c matricea (centrată) cu coloanele $\mathbf{X}_j - \bar{\mathbf{x}}_j \mathbf{1}$, $j = 1, \dots, p$ putem observa că

$$\boldsymbol{\Gamma} = \frac{1}{n} \mathbf{Z}^T \mathbf{Z} - \bar{\mathbf{x}} \bar{\mathbf{x}}^T = \frac{1}{n} \mathbf{Z}_c^T \mathbf{Z}_c$$

prin urmare $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^T$, deci $\boldsymbol{\Gamma}$ este simetrică. Astfel, folosind faptul că un scalar este egal cu transpusul său, avem

$$\mathbf{z}_{n+1}^T \boldsymbol{\Gamma}^{-1} \bar{\mathbf{x}} = (\mathbf{z}_{n+1}^T \boldsymbol{\Gamma}^{-1} \bar{\mathbf{x}})^T = \bar{\mathbf{x}}^T \boldsymbol{\Gamma}^{-1} \mathbf{z}_{n+1}$$

ceea ce conduce la

$$\begin{aligned} \mathbf{x}_{n+1}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_{n+1} &= \frac{1}{n} (1 + \bar{\mathbf{x}}^T \boldsymbol{\Gamma}^{-1} \bar{\mathbf{x}} - 2 \bar{\mathbf{x}}^T \boldsymbol{\Gamma}^{-1} \mathbf{z}_{n+1} + \mathbf{z}_{n+1}^T \boldsymbol{\Gamma}^{-1} \mathbf{z}_{n+1}) \\ &= \frac{1}{n} [1 + (\mathbf{z}_{n+1} - \bar{\mathbf{x}})^T \boldsymbol{\Gamma}^{-1} (\mathbf{z}_{n+1} - \bar{\mathbf{x}})] \end{aligned}$$

de unde găsim

$$Var(\hat{\varepsilon}_{n+1}) = \sigma^2 \left[1 + \frac{1}{n} + \frac{1}{n} (\mathbf{z}_{n+1} - \bar{\mathbf{x}})^T \boldsymbol{\Gamma}^{-1} (\mathbf{z}_{n+1} - \bar{\mathbf{x}}) \right].$$

Mai mult, să remarcăm că dacă $\mathbf{X}^T \mathbf{X}$ este inversabilă atunci matricea $\boldsymbol{\Gamma}$ este pozitiv definită. Într-adevăr am văzut că $\boldsymbol{\Gamma} = \frac{1}{n} \mathbf{Z}_c^T \mathbf{Z}_c$ iar pentru $\mathbf{u} \in \mathbb{R}^p$ putem scrie

$$\mathbf{u}^T \boldsymbol{\Gamma} \mathbf{u} = \frac{1}{n} \mathbf{u}^T \mathbf{Z}_c^T \mathbf{Z}_c \mathbf{u} = \frac{1}{n} \|\mathbf{Z}_c \mathbf{u}\|^2 \geq 0$$

deci $\boldsymbol{\Gamma}$ este pozitiv semidefinită. Avem egalitate $\frac{1}{n} \|\mathbf{Z}_c \mathbf{u}\|^2 = 0$ atunci când $\mathbf{Z}_c \mathbf{u} = 0$ sau, altfel spus, atunci când

$$u_1(\mathbf{X}_1 - \bar{\mathbf{x}}_1 \mathbf{1}) + \cdots + u_p(\mathbf{X}_p - \bar{\mathbf{x}}_p \mathbf{1}) = 0 \iff \sum_{j=1}^n u_j \mathbf{X}_j = \left(\sum_{j=1}^n u_j \bar{\mathbf{x}}_j \right) \mathbf{1}.$$

Această ultimă relație ne spune că prima coloană a lui \mathbf{X} se scrie ca o combinație liniară de celelalte, ceea ce înseamnă că $\text{rang}(\mathbf{X}) \leq p$ de unde concluzionăm că matricea $\mathbf{X}^T \mathbf{X}$ nu este inversabilă. \square

Este interesant de remarcat cum din faptul că $\boldsymbol{\Gamma}$ este simetrică și pozitiv definită deducem că $(\mathbf{z}_{n+1} - \bar{\mathbf{x}})^T \boldsymbol{\Gamma}^{-1} (\mathbf{z}_{n+1} - \bar{\mathbf{x}}) \geq 0$, cu egalitate dacă și numai dacă $\mathbf{z}_{n+1} = \bar{\mathbf{x}}$, adică $\mathbf{x}_{n+1}^T = (1, \mathbf{z}_{n+1}^T) = (1, \bar{\mathbf{x}}^T)$. În acest caz $Var(\hat{\varepsilon}_{n+1}) = \sigma^2 (1 + \frac{1}{n})$, rezultat care generalizează rezultatul corespunzător de la regresia liniară simplă: trebuie să ne plasăm în centrul norului de puncte ale variabilelor explicative pentru a prezice variabila răspuns cu mai multă precizie.

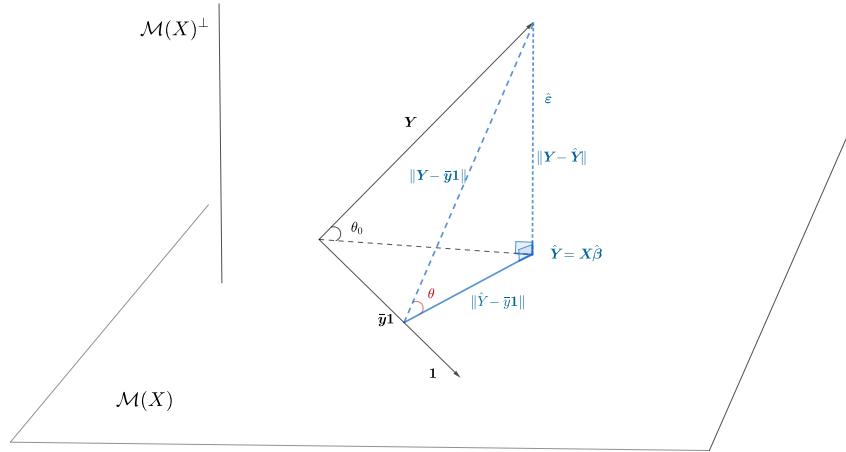
4.5 Interpretări geometrice - coeficientul de determinare

În această secțiune vom prezenta descompunerea ANOVA a modelului de regresie, vom defini, dintr-o perspectivă geometrică, coeficientul de determinare multiplă și vom investiga o serie de proprietăți ale acestuia.

Vom considera atât modelul de regresie liniară multiplă în care este inclus termenul liber cât și cel în care acesta nu apare. Fie astfel, modelul de regresie liniară $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ în care $\mathbf{Y} \in \mathbb{R}^n$ este vectorul răspuns, $\mathbf{X} = [\mathbf{1} | \mathbf{X}_1 | \cdots | \mathbf{X}_p] \in \mathcal{M}_{n,p+1}(\mathbb{R})$ sau $\mathbf{X} = [\mathbf{X}_1 | \cdots | \mathbf{X}_p] \in \mathcal{M}_{n,p}(\mathbb{R})$ (pentru modelul fără termenul liber)

este matricea de design ($\text{rang}(\mathbf{X}) = p + 1$ respectiv $\text{rang}(\mathbf{X}) = p$), $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ sau $\boldsymbol{\beta} \in \mathbb{R}^p$ este vectorul parametrilor după caz și $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ este vectorul termenilor eroare.

Notăm, ca și până acum, cu $\mathcal{M}(X)$ subspațiul generat de coloanele matricii \mathbf{X} , cu $\mathcal{M}(X)^\perp$ subspațiul ortogonal, cu $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ proiecția ortogonală a lui \mathbf{Y} pe $\mathcal{M}(X)$ și cu $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} \in \mathcal{M}(X)^\perp$ vectorul valorilor reziduale.



Dacă modelul de regresie considerat nu include termenul liber (β_0) atunci, conform figurii de mai sus, observăm că putem aplica *Teorema lui Pitagora* direct (în triunghiul determinat de \mathbf{Y} , $\hat{\mathbf{Y}}$ și $\hat{\boldsymbol{\varepsilon}}$) și obținem că

$$\begin{aligned}\|\mathbf{Y}\|^2 &= \|\hat{\mathbf{Y}}\|^2 + \|\hat{\boldsymbol{\varepsilon}}\|^2 \\ &= \|\mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \\ SS_T &= SS_{reg} + RSS.\end{aligned}$$

Pe de altă parte, atunci când modelul include și termenul liber (cum este cazul în cele mai multe aplicații), atunci tot prin aplicarea *Teoremei lui Pitagora*, dar de această dată în triunghiul albastru, are loc următoarea descompunere (descompunerea ANOVA pentru regresie)

$$\begin{aligned}\|\mathbf{Y} - \bar{\mathbf{y}}\mathbf{1}\|^2 &= \|\hat{\mathbf{Y}} - \bar{\mathbf{y}}\mathbf{1}\|^2 + \|\hat{\boldsymbol{\varepsilon}}\|^2 \\ SS_T &= SS_{reg} + RSS\end{aligned}$$

unde

- $SS_T = \|\mathbf{Y} - \bar{\mathbf{y}}\mathbf{1}\|^2$ se numește **suma abaterilor pătratice totale** (variația totală)
- $SS_{reg} = \|\hat{\mathbf{Y}} - \bar{\mathbf{y}}\mathbf{1}\|^2$ se numește **suma abaterilor pătratice de regresie** (variația explicată de model)
- $RSS = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \|\hat{\boldsymbol{\varepsilon}}\|^2$ se numește **suma abaterilor pătratice reziduale** (variația reziduală)

astfel

$$\text{Variația totală} = \text{Variația explicată de regresie} + \text{Variația reziduală}.$$

Def. 4.27



Coefficientul de determinare R^2 (în R are denumirea de **Multiple R-Squared**) este definit prin

$$R^2 = \cos^2(\theta_0) = \frac{\|\hat{\mathbf{Y}}\|^2}{\|\mathbf{Y}\|^2} = 1 - \frac{\|\hat{\mathbf{\epsilon}}\|^2}{\|\mathbf{Y}\|^2} = 1 - \frac{RSS}{SS_T}$$

atunci când termenul liber nu este inclus în model și prin

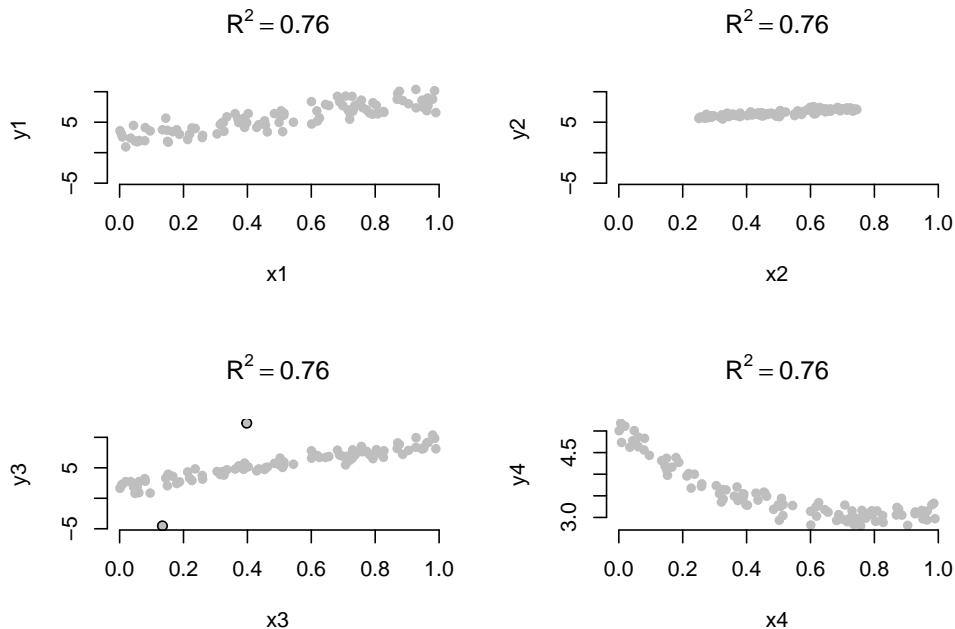
$$R^2 = \cos^2(\theta) = \frac{\text{Variația explicată de model}}{\text{Variația totală}} = \frac{\|\hat{\mathbf{Y}} - \bar{\mathbf{y}}\mathbf{1}\|^2}{\|\mathbf{Y} - \bar{\mathbf{y}}\mathbf{1}\|^2} = 1 - \frac{\|\hat{\mathbf{\epsilon}}\|^2}{\|\mathbf{Y} - \bar{\mathbf{y}}\mathbf{1}\|^2} = 1 - \frac{RSS}{SS_T}$$

atunci când termenul liber este inclus în model.

Din descompunerea ANOVA pentru regresie putem interpreta coefficientul de determinare (procentul din varianta totală explicată de model) astfel: cu cât este mai aproape R^2 de 1 cu atât este mai mică suma abaterilor pătratice reziduale $\|\hat{\mathbf{\epsilon}}\|^2 = \sum_{i=1}^n \hat{\epsilon}_i^2$, prin urmare cu atât este mai bună potrivirea modelului pe date (fitul) - cazul extrem în care avem $R^2 = 1$ implică $\|\hat{\mathbf{\epsilon}}\|^2 = 0$ ceea ce înseamnă că punctele se regăsesc în hiperplanul de regresie și avem potrivire perfectă; pe de altă parte, cu cât R^2 este mai aproape de 0, cu atât potrivirea pe date este slabă (fitul nu este prea bun) - cazul limită $R^2 = 0$ implică faptul că predicția răspunsului y_i este întotdeauna egală cu media \bar{y} răspunsului și nu depinde de valorile covariabilelor, ceea ce atată că acestea nu au niciun impact asupra explicării modelului.

Trebuie să ținem cont că suntem în ipoteza în care modelul liniar considerat este adecvat datelor, în caz contrar (i.e. atunci când există relații neliniare între covariabile și răspuns), în care modelul nu este specificat corespunzător, putem avea un coefficient de determinare apropiat de 0 și cu toate acestea covariabilele să aibă o mare putere explicativă asupra răspunsului. Prin urmare, folosirea coefficientului de determinare ca (singură) măsură de ajustare a modelului se poate dovedi înselătoare.

În figura următoare avem diagramele de împrăștiere a patru modele simulate care întorc aproximativ același coefficient de determinare $R^2 \approx 0.76$. Dacă în figura din stânga sus avem un exemplu de situație normală, în figura din dreapta sus variația reziduală este mai scăzută decât cea din primul grafic dar atunci și variația lui x este mai mică conducând la o valoare similară a lui R^2 . Figura din stânga jos arată o corelație puternică între răspuns și variabila explicativă (un coefficient de corelație de 0.954), cu excepția unei observații aberrante (outlier) iar figura din dreapta jos prezintă un model specificat incorect, relația fiind una pătratică.



Este important de remarcat și că în modelul de regresie liniară multiplă, ca și în cazul modelului de regresie liniară simplă, putem interpreta coeficientul de determinare ca fiind pătratul coeficientului de corelație dintre răspuns y_i și valorile ajustate \hat{y}_i : $R^2 = r_{\mathbf{Y}\hat{\mathbf{Y}}}^2$.

Atunci când vrem să comparăm diferite modele pe baza coeficientului de determinare, trebuie să avem grijă ca următoarele condiții să fie îndeplinite:

- fiecare model trebuie să aibă aceeași variabilă răspuns, nu putem compara modele în care primul are ca variabilă răspuns pe y iar al doilea pe $h(y)$, h o transformare
- fiecare model trebuie să aibă același număr de parametrii
- fiecare model trebuie să includă termenul liber

Astfel, în general, nu putem folosi coeficientul de determinare pentru compararea modelelor după cum se vede și în exemplul următor.

Exemplu: prețul chirilor din Munchen - compararea modelelor după R^2

Exp. 4.28 După cum am văzut în exemplele 4.7 și 4.9, putem explicita relația dintre prețul chiriei pe metrul pătrat al apartamentelor din Munchen în funcție de suprafața de locuit a acestora prin diferite modele de regresie liniară:

$$M_1 : \text{pret_m}^2_i = \beta_0 + \beta_1 \times \text{suprafata}_i + \varepsilon_i,$$

$$M_2 : \text{pret_m}^2_i = \beta_0 + \beta_1 \times \frac{1}{\text{suprafata}_i} + \varepsilon_i,$$

$$M_3 : \text{pret_m}^2_i = \beta_0 + \beta_1 \times \text{suprafata}_i + \beta_2 \times \text{suprafata}_i^2 + \varepsilon_i$$

$$M_4 : \text{pret_m}^2_i = \beta_0 + \beta_1 \times \text{suprafata}_i + \beta_2 \times \text{suprafata}_i^2 + \beta_3 \times \text{suprafata}_i^3 + \varepsilon_i$$

Tabelul de mai jos include modelele estimate împreună cu coeficienții de determinare corespunzători:

Modelul	Modelul estimat	Coeficientul de determinare R^2
M_1	$\widehat{\text{pret_m}^2}_i = 9.469140.178 \times \text{suprafata}_i$	0.116
M_2	$\widehat{\text{pret_m}^2}_i = 4.732 + 140.178 \times \frac{1}{\text{suprafata}_i}$	0.154

Modelul	Modelul estimat	Coefficientul de determinare R^2
M_3	$\widehat{\text{pret}_m}^2_i = 11.83 - 0.106 \times \text{suprafata}_i + 4.7 \times 10^{-4} \times \text{suprafata}_i^2$	0.143
M_4	$\widehat{\text{pret}_m}^2_i = 14.28 - 0.218 \times \text{suprafata}_i + 0.002 \times \text{suprafata}_i^2 - 6 \times 10^{-6} \times \text{suprafata}_i^3$	0.15

Se observă că pentru toate cele patru modele, coefficientul de determinare este relativ scăzut motivul principal fiind că, pe lângă variabilitatea mare a datelor, modelele nu includ o mare parte din covariabilele cu mare impact explicativ.

Dacă ne uităm la primele două modele, observăm că M_2 are un coefficient de determinare mai mare $R_{M_2}^2 = 0.154 \geq R_{M_1}^2 = 0.116$. Cum ambele modele conțin același număr de parametrii deducem că acest model este de preferat între cele două. Dacă ne uităm la modelele M_1 , M_3 și M_4 constatăm că acestea sunt *modele imbricate* (nested models) ceea ce înseamnă că modelul M_4 include atât modelul M_3 ($\beta_3 = 0$) cât și pe M_1 ($\beta_3 = \beta_2 = 0$) iar modelul M_3 îl conține pe M_1 ca și caz particular ($\beta_2 = 0$). După cum vom vedea în Propoziția 4.32, asta implică faptul că M_1 are cel mai mic coefficient de determinare între cele trei modele, M_3 are al doilea cel mai mic R^2 iar M_4 are cel mai mare R^2 . O comparație între cele trei modele, în funcție de R^2 , în acest caz nu este recomandată deoarece modelele au un număr diferit de parametrii. Este de asemenea important de remarcat faptul că un model cu un număr mai mare de parametrii nu implică neapărat că are și un coefficient de determinare mai mare față de un model cu un număr mai mic de parametrii. Acest fenomen se observă atunci când comparăm M_2 cu M_3 , $R_{M_2}^2 = 0.154 \geq R_{M_3}^2 = 0.143$. În acest caz, modelul M_2 este de preferat modelului M_3 deoarece conține un număr mai mic de parametrii și are un coefficient R^2 mai mare. \square

Cum coefficientul de determinare $R^2 = 1 - \frac{RSS}{SS_T} = 1 - \frac{\hat{\sigma}}{SS_T}[n - (p + 1)]$ observăm că acesta crește odată cu numărul de variabile explicative p . În cele ce urmează vom defini o variantă a coefficientului de determinare care să amelioreze această problemă. Cum R^2 nu ține cont de dimensiunea spațiului $\mathcal{M}(X)$ se definește *coefficientul de determinare ajustat* R_a^2 (în R are denumirea de Adjusted R-Squared) prin



Coefficientul de determinare ajustat R_a^2 este definit prin

Def. 4.29

$$R_a^2 = 1 - \frac{\frac{\|\hat{\varepsilon}\|^2}{n-(p+1)}}{\frac{\|Y\|^2}{n}} = 1 - \frac{n}{n-(p+1)} \frac{\|\hat{\varepsilon}\|^2}{\|Y\|^2} = 1 - \frac{n}{n-(p+1)}(1-R^2)$$

atunci când termenul liber nu este inclus în model și respectiv prin

$$R_a^2 = 1 - \frac{\frac{\|\hat{\varepsilon}\|^2}{n-(p+1)}}{\frac{\|Y-\bar{y}\mathbf{1}\|^2}{n-1}} = 1 - \frac{n-1}{n-(p+1)} \frac{\|\hat{\varepsilon}\|^2}{\|Y-\bar{y}\mathbf{1}\|^2} = 1 - \frac{n-1}{n-(p+1)}(1-R^2).$$

atunci când termenul liber este inclus în model.

Ajustarea corespunde la împărțirea normelor la pătrat cu dimensiunea subspațiului din care fac parte vectorii. Notăm de asemenea că R_a^2 nu mai măsoară proporția variabilității lui y explicate de modelul de regresie și în plus $R_a^2 \leq 1$ și poate lua chiar și valori negative.

Exemplu: compararea coeficienților de determinare R^2 și R_a^2

Exp. 4.30

Pentru a vedea diferența dintre cei doi coeficienți de determinare considerăm următorul studiu de simulare în care presupunem că generăm observațiile $x_{i1}, x_{i2}, x_{i3}, y_i$ pentru $i = 1, 2, \dots, n = 100$ conform modelului de regresie liniară

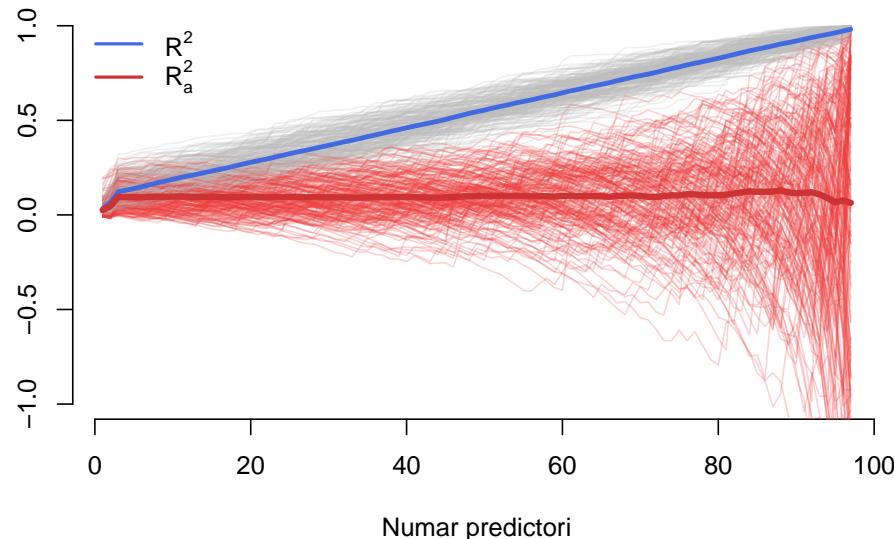
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,$$

unde $x_1, x_2, x_3 \sim \mathcal{N}(0, 1)$ și $\varepsilon_i \sim \mathcal{N}(0, 9)$. La aceste date adăugăm $p - 3 = 94$ variabile explicative fictive

$x_j \sim \mathcal{N}(0, 1)$ care sunt independente de y rezultând într-un total de $p = 97$ covariabile dintre care doar primele două (x_1 și x_2) sunt relevante în explicarea răspunsului. Calculăm coeficienții de determinare $R^2(j)$ și respectiv $R_a^2(j)$ pentru fiecare dintre modelele

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j + \varepsilon,$$

unde $j = 1, 2, \dots, p$ și trasăm pe același grafic curbele $(j, R^2(j))$ și $(j, R_a^2(j))$. Cum R^2 și R_a^2 sunt variabile aleatoare, repetăm experimentul de $M = 200$ ori. Rezultatele sunt ilustrate în figura de mai jos (curbele îngrosate sunt curbele medii). Putem observa că R^2 crește liniar odată cu numărul de predictori din model chiar dacă doar primii trei predictori sunt relevanți în explicarea lui y . Pe de altă parte R_a^2 crește în primele trei covariabile după care descrește încet arătând că, în medie, numărul optim de covariabile din model sunt în jurul lui $p = 3$. Cu toate acestea se constată că R_a^2 are o variabilitate mare atunci când p se apropie de $n - 3$, o consecință a factorului $n - 1$. \square



Următorul exercițiu este o aplicație numerică de calcul a coeficienților de determinare în contextul unui model de regresie cu două covariabile.

Ex. 4.31



Ne propunem să determinăm evoluția unei variabile răspuns y_i în funcție de două variabile explicative x_i și z_i . Fie $\mathbf{X} = (\mathbf{1}, \mathbf{x}, \mathbf{z})$ matricea de design.

1. Am obținut rezultatele următoare

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 25 & 0 & 0 \\ ? & 9.3 & 5.4 \\ ? & ? & 12.7 \end{pmatrix}$$

- Determinați valorile care lipsesc (?).
- Cât este n ?
- Calculați coeficientul de corelație dintre \mathbf{x} și \mathbf{z} .

2. Modelul de regresie liniară a lui \mathbf{Y} în funcție de $\mathbf{1}, \mathbf{x}, \mathbf{z}$ este

$$\mathbf{Y} = -1.6\mathbf{1} + 0.61\mathbf{x} + 0.46\mathbf{z} + \hat{\boldsymbol{\varepsilon}}, \quad \|\hat{\boldsymbol{\varepsilon}}\|^2 = 0.3$$

- a) Determinați $\bar{\mathbf{y}}$.
- b) Calculați suma abaterilor pătratice de regresie SS_{reg} , suma abaterilor pătratice totale SS_T , coeficientul de determinare și coeficientul de determinare ajustat.

1. a) Pentru a determina cele trei valori care lipsesc să notăm că matricea $\mathbf{X}^\top \mathbf{X}$ este simetrică, prin urmare $\mathbf{X}^\top \mathbf{X} = (\mathbf{X}^\top \mathbf{X})^\top$ de unde

$$\begin{pmatrix} 25 & 0 & 0 \\ ? & 9.3 & 5.4 \\ ? & ? & 12.7 \end{pmatrix} = \begin{pmatrix} 25 & ? & ? \\ 0 & 9.3 & ? \\ 0 & 5.4 & 12.7 \end{pmatrix}$$

ceea ce conduce la valorile 0, 0, 5.4.

b) Pentru a determina valoarea lui n să observăm că

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

ceea ce arată că $n = (\mathbf{X}^\top \mathbf{X})_{1,1} = 25$.

c) Coeficientul de corelație dintre \mathbf{x} și \mathbf{z} este dat de

$$r_{\mathbf{x}, \mathbf{z}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (z_i - \bar{z})^2}}$$

și cum, din punctul anterior, $\bar{x} = \frac{(\mathbf{X}^\top \mathbf{X})_{1,2}}{n} = 0$ și $\bar{z} = \frac{(\mathbf{X}^\top \mathbf{X})_{1,3}}{n} = 0$ deci

$$r_{\mathbf{x}, \mathbf{z}} = \frac{\sum_{i=1}^n x_i z_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n z_i^2}} = \frac{\mathbf{x}^\top \mathbf{z}}{\sqrt{\mathbf{x}^\top \mathbf{x}} \sqrt{\mathbf{z}^\top \mathbf{z}}} = \frac{(\mathbf{X}^\top \mathbf{X})_{2,3}}{\sqrt{(\mathbf{X}^\top \mathbf{X})_{2,2}} \sqrt{(\mathbf{X}^\top \mathbf{X})_{3,3}}} = \frac{5.4}{\sqrt{9.3} \sqrt{12.7}} \approx 0.5$$

2. a) Avem modelul

$$y_i = -1.6 + 0.61x_i + 0.46z_i + \hat{\varepsilon}_i$$

care prin sumare devine

$$\bar{y} = -1.6 + 0.61\bar{x} + 0.46\bar{z} + \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i.$$

Cum $\mathbf{1}$ aparține modelului, deci $\mathbf{1} \in \mathcal{M}(\mathbf{X})$ și ținând cont de faptul că $\hat{\boldsymbol{\varepsilon}} \in \mathcal{M}(X)^\perp$ deducem că $\langle \mathbf{1}, \hat{\boldsymbol{\varepsilon}} \rangle = 0$ sau $\frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i = 0$. Astfel

$$\bar{y} = -1.6 + 0.61 \underbrace{\bar{x}}_{=0} + 0.46 \underbrace{\bar{z}}_{=0} + \underbrace{\frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i}_{=0} = -1.6.$$

b) Avem că suma abaterilor pătratice explicate de model este dată de

$$SS_{reg} = \|\hat{\mathbf{Y}} - \bar{y}\mathbf{1}\|^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (0.61x_i + 0.46z_i)^2$$

adică

$$\begin{aligned} SS_{reg} &= \|\hat{\mathbf{Y}} - \bar{y}\mathbf{1}\|^2 = 0.61^2 \sum_{i=1}^n x_i^2 + 2 \times 0.61 \times 0.46 \sum_{i=1}^n x_i z_i + 0.46^2 \sum_{i=1}^n z_i^2 \\ &= 0.61^2 (\mathbf{X}^\top \mathbf{X})_{2,2} + 2 \times 0.61 \times 0.46 (\mathbf{X}^\top \mathbf{X})_{2,3} + 0.46^2 (\mathbf{X}^\top \mathbf{X})_{3,3} = 9.18. \end{aligned}$$

Suma abaterilor pătratice totale se determină folosind formula de descompunere a varianței:

$$\underbrace{SS_T}_{\|\mathbf{Y} - \bar{y}\mathbf{1}\|^2} = \underbrace{SS_{reg}}_{\|\hat{\mathbf{Y}} - \bar{y}\mathbf{1}\|^2} + \underbrace{RSS}_{\|\hat{\varepsilon}\|^2} = 9.18 + 0.3 = 9.48.$$

Coefficientul de determinare R^2 este

$$R^2 = \frac{SS_{reg}}{SS_T} = \frac{9.18}{9.48} \approx 0.968$$

ceea ce arată că aproximativ 97% din variabilitatea datelor este explicată de modelul de regresie iar coefficientul de determinare ajustat este

$$R_a^2 = 1 - \frac{n-1}{n-(p+1)} \frac{\|\hat{\varepsilon}\|^2}{\|\mathbf{Y} - \bar{y}\mathbf{1}\|^2} = 1 - \frac{n-1}{n-(p+1)} (1 - R^2) = 1 - \frac{25-1}{25-(2+1)} (1 - 0.968) \approx 0.965$$

ceea ce verifică relația generală $R_a^2 < R^2$. \square

Următorul rezultat justifică relația care există între coeficienții de determinare în cazul modelelor imbricate.

Prop. 4.32



Fie $\mathbf{Z} \in \mathcal{M}_{n,q}(\mathbb{R})$ o matrice de rang q și $\mathbf{X} \in \mathcal{M}_{n,p}(\mathbb{R})$ o matrice de rang p , $p > q$, compusă din cei q vectori coloană ai lui \mathbf{Z} și din alți $p-q$ vectori liniar independenti. Considerăm modelele următoare:

$$\begin{aligned} \mathbf{Y} &= \mathbf{Z}\beta + \varepsilon, \\ \mathbf{Y} &= \mathbf{X}\gamma + \eta \end{aligned}$$

Presupunem, pentru simplitate, că niciun model nu conține termenul liber (constanta). Notăm cu P_X și P_Z matricele de proiecție ortogonală pe spațiile $\mathcal{M}(X)$ și respectiv $\mathcal{M}(Z)$, generate de coloanele lui \mathbf{X} și \mathbf{Z} . De asemenea notăm cu $P_{X \cap Z^\perp}$ matricea de proiecție ortogonală pe $\mathcal{M}(X) \cap \mathcal{M}(Z)^\perp$, spațiul ortogonal al lui $\mathcal{M}(Z)$ din $\mathcal{M}(X)$, astfel

$$\mathbb{R}^n = \mathcal{M}(X) \oplus \mathcal{M}(X)^\perp = (\mathcal{M}(Z) \oplus (\mathcal{M}(X) \cap \mathcal{M}(Z)^\perp)) \oplus \mathcal{M}(X)^\perp$$

Atunci:

1. Are loc relația $\|P_X \mathbf{Y}\|^2 = \|P_Z \mathbf{Y}\|^2 + \|P_{X \cap Z^\perp} \mathbf{Y}\|^2$.
2. Între coeficienții de determinare ai celor două modele există inegalitatea $R_X^2 \geq R_Z^2$.

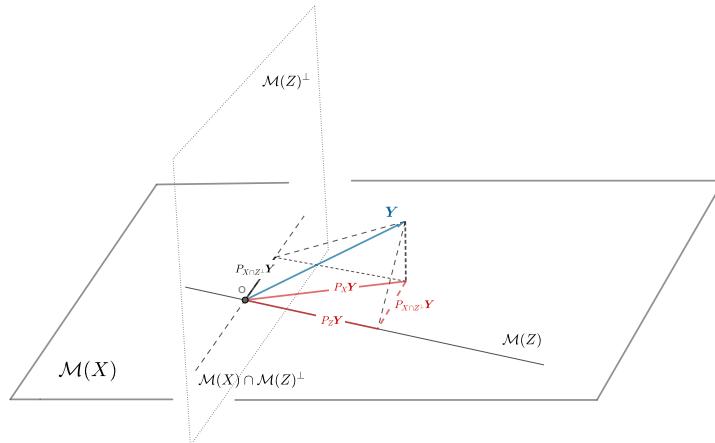
Această propoziție face referire la modelele imbricate - *nested models*.

1. Deoarece $\mathcal{M}(\mathbf{Z}) \subset \mathcal{M}(\mathbf{X})$ și aplicând *teorema celor trei perpendiculare*⁵ avem, vezi figura de mai jos,

$$\begin{aligned}\hat{\mathbf{Y}}_p &= P_X \mathbf{Y} = (\underbrace{P_Z + P_{Z^\perp}}_I) P_X \mathbf{Y} = P_Z P_X \mathbf{Y} + P_{Z^\perp} P_X \mathbf{Y} \\ &= P_X \mathbf{Y} + P_{X \cap Z^\perp} \mathbf{Y} = \hat{\mathbf{Y}}_q + P_{X \cap Z^\perp} \mathbf{Y}\end{aligned}$$

și din Teorema lui Pitagora (vezi triunghiul roșu) deducem că

$$\|P_X \mathbf{Y}\|^2 = \|P_Z \mathbf{Y}\|^2 + \|P_{X \cap Z^\perp} \mathbf{Y}\|^2.$$



2. Dacă modelul nu are termenul liber (constanta) atunci coeficientul de determinare este definit prin $R_X^2 = \frac{\|P_X \mathbf{Y}\|^2}{\|\mathbf{Y}\|^2}$ prin urmare

$$R_X^2 = \frac{\|P_X \mathbf{Y}\|^2}{\|\mathbf{Y}\|^2} = \frac{\|P_Z \mathbf{Y}\|^2 + \|P_{X \cap Z^\perp} \mathbf{Y}\|^2}{\|\mathbf{Y}\|^2} = R_Z^2 + \frac{\|P_{X \cap Z^\perp} \mathbf{Y}\|^2}{\|\mathbf{Y}\|^2} \geq R_Z^2. \square$$

Rezultatul de mai sus se generalizează și în cazul modelelor imbricate care conțin termenul constant și ne arată că modelul mai mare are un coeficient de determinare superior modelului mai mic. Altfel spus, dacă la un model dat adăugăm una sau mai multe variabile explicative ameliorăm variabilitatea explicată de model, chiar dacă variabilele explicative suplimentare nu sunt pertinente! În acest caz este de preferat coeficientul de determinare ajustat.

⁵ Teorema celor trei perpendiculare spune că dacă \mathbf{V} și \mathbf{W} sunt două subspații vectoriale astfel încât $\mathbf{V} \subset \mathbf{W}$ atunci $P_V P_W = P_W P_V = P_V$.

4.6 Modelul (condiționat) normal

Considerăm modelul de regresie $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ în care $\mathbf{Y} \in \mathbb{R}^n$ este vectorul răspuns, $\mathbf{X} = [\mathbf{1} | \mathbf{X}_1 | \cdots | \mathbf{X}_p] \in \mathcal{M}_{n,p+1}(\mathbb{R})$ este matricea de design, $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ este vectorul parametrilor și $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ este vectorul termenilor eroare. Până acum, ipotezele modelului erau

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

și proprietățile estimatorilor rezultați din metoda celor mai mici pătrate au fost obținute fără a face apel la repartitia termenilor eroare.

În această secțiune ne propunem să studiem proprietățile statistice ale modelului de regresie liniară plasându-ne într-un context parametric și anume în contextul modelului gaussian:

$$\begin{aligned}\mathcal{H}_1 : & \text{rang}(\mathbf{X}) = p + 1 \\ \mathcal{H}'_2 : & \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)\end{aligned}$$

Observăm că ipoteza \mathcal{H}'_2 este un caz particular al ipotezei \mathcal{H}_2 și în plus implică faptul că reziduurile sunt independente și identic repartizate condiție care nu era necesară până acum. Ipoteza de normalitate a erorilor ne permite să scriem funcția de verosimilitate asociată modelului, să deducem repartițiile estimatorilor propuși, să construim intervale de încredere și să testăm ipoteze statistice asupra parametrilor modelului.

4.6.1 Estimatori de verosimilitate maximă

odată cu ipoteza suplimentară de normalitate a termenilor eroare, putem determina estimatorii de verosimilitate maximă. Rezultatul următor face legătura dintre estimatorii obținuți prin metoda celor mai mici pătrate din capitolul anterior și cei obținuți prin metoda verosimilității maxime.

Prop. 4.33



Considerăm modelul de regresie liniară $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ sub ipotezele \mathcal{H}_1 și \mathcal{H}'_2 de mai sus. Atunci estimatorii de verosimilitate maximă pentru $\boldsymbol{\beta}$ și σ^2 sunt date de

$$\hat{\boldsymbol{\beta}}_{VM} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}, \quad \hat{\sigma}_{VM}^2 = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{VM}\|^2}{n} = \frac{n - (p + 1)}{n} \hat{\sigma}^2.$$

Începem prin a observa că din modelul de regresie

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

și conform ipotezei \mathcal{H}'_2 avem $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ prin urmare $y_i \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$ și în plus y_i sunt variabile aleatoare independente deoarece ε_i sunt independente. Astfel funcția de verosimilitate se scrie

$$\begin{aligned}L(\boldsymbol{\beta}, \sigma^2; \mathbf{Y}) &= \prod_{i=1}^n f_Y(y_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2}\end{aligned}$$

ceea ce conduce la logaritmul funcției de verosimilitate

$$\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{Y}) = \log L(\boldsymbol{\beta}, \sigma^2; \mathbf{Y}) = -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Vrem să determinăm

$$(\hat{\boldsymbol{\beta}}_{VM}, \hat{\sigma}_{VM}^2) = \arg \max_{(\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^{p+1} \times (0, \infty)} \ell(\boldsymbol{\beta}, \sigma^2; \mathbf{Y})$$

și observăm că pentru σ^2 fixat, a maximiza funcția $\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{Y})$ revine la a determina valoarea lui $\boldsymbol{\beta}$ care minimizează $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$, aceasta ne fiind alta decât $\hat{\boldsymbol{\beta}}$ obținut prin metoda celor mai mici pătrate. Astfel găsim că

$$\hat{\boldsymbol{\beta}}_{VM} = \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Pentru a determina $\hat{\sigma}_{VM}^2$ să observăm că problema revine la a determina soluția ecuației

$$\frac{\partial l(\hat{\boldsymbol{\beta}}_{VM}, \sigma^2; \mathbf{Y})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{VM}\|^2 = 0$$

care este $\hat{\sigma}_{VM}^2 = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{VM}\|^2}{n} = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n}$. Comparând acest rezultat cu cel obținut prin metoda celor mai mici pătrate observăm că

$$\hat{\sigma}_{VM}^2 = \frac{n - (p + 1)}{n} \hat{\sigma}^2,$$

deci estimatorul de verosimilitate maximă $\hat{\sigma}_{VM}^2$ a lui σ^2 este un estimator deplasat. \square

Se poate arăta că estimatorii de verosimilitate maximă $\hat{\boldsymbol{\beta}}_{VM} = \hat{\boldsymbol{\beta}}$ și $\hat{\sigma}_{VM}^2 = \frac{n - (p + 1)}{n} \hat{\sigma}^2$ și prin urmare și cei obținuți prin metoda celor mai mici pătrate sunt suficienți în sensul că niciun alt estimator nu poate să îmbunătățească estimarea lui $\boldsymbol{\beta}$ și σ^2 plecând de la eșantionul dat [Rencher and Schaalje, 2008]. Mai mult, dacă în ipotezele \mathcal{H}_1 și \mathcal{H}_2 estimatorul $\hat{\boldsymbol{\beta}}$ era de varianță minimală în clasa estimatorilor liniari nedeplasati (Teorema Gauss-Markov - Propoziția 4.19) în ipoteza suplimentară de normalitate \mathcal{H}'_2 acest rezultat se extinde la clasa tuturor estimatorilor nedeplasati. În același mod și $\hat{\sigma}^2$ are varianță minimală în clasa tuturor estimatorilor nedeplasati pentru σ^2 [Rencher and Schaalje, 2008].

În cele ce urmează vom considera estimatorul nedeplasat $\hat{\sigma}^2 = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n - (p + 1)} = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n - (p + 1)}$ ca estimator al lui σ^2 .

4.6.2 Repartițiile estimatorilor

Înainte de a investiga repartițiile estimatorilor $\hat{\boldsymbol{\beta}}$ și $\hat{\sigma}^2$ pentru modelul de regresie liniară multiplă sub ipoteza de normalitate, să reamintim câteva noțiuni legate de vectorii gaussiani (pentru mai multe detalii se poate consulta [Seber and Lee, 2003, Capitolul 2] sau [Jacod and Protter, 2003, Capitolul 16]).

Spunem că un vector $\mathbf{Y} = (Y_1, \dots, Y_n) \in \mathbb{R}^n$ este un vector gaussian dacă toate combinațiile liniare $\sum_{i=1}^n a_i Y_i$ sunt repartizate normal (posibil degenerat dacă toți coeficienții sunt nuli). Acest vector admite o medie $\boldsymbol{\mu} = \mathbb{E}[\mathbf{Y}]$ și o matrice de varianță-covarianță $\Sigma_{\mathbf{Y}} = \mathbb{E}[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^\top]$ care caracterizează complet repartitia lui \mathbf{Y} . În acest caz notăm $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma_{\mathbf{Y}})$. Un vector gaussian \mathbf{Y} admite o densitate de repartitie $f_{\mathbf{Y}}$ pe \mathbb{R}^n dacă și numai dacă matricea sa de varianță-covarianță $\Sigma_{\mathbf{Y}}$ este inversabilă și în acest caz

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\det \Sigma_{\mathbf{Y}}}} e^{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top \Sigma_{\mathbf{Y}}^{-1} (\mathbf{y} - \boldsymbol{\mu})}$$

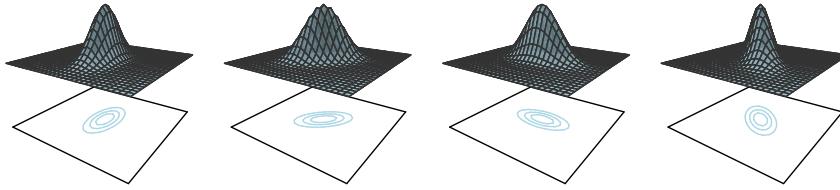
În cazul în care matricea Σ_Y nu este inversabilă înseamnă că Y ia valori într-un subspațiu de dimensiune $n_0 < n$ unde este repartizat ca un vector gaussian de dimensiune n_0 .

Figura de mai jos prezintă patru repartiții normale bivariate ($n = 2$) $Y \sim \mathcal{N}(\mu, \Sigma_Y)$, cu $\mu = (\mu_1, \mu_2)^\top$ iar $\Sigma_Y = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$, rotite cu unghiul α . În acest caz densitatea de repartiție este

$$f_Y(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right) + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2 \right]}$$

unde $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$ e

$$\alpha=0 \qquad \qquad \alpha=\frac{\pi}{4} \qquad \qquad \alpha=\frac{\pi}{2} \qquad \qquad \alpha=\frac{3\pi}{2}$$



Una dintre proprietățile importante ale vectorilor gaussiani este stabilitatea prin transformări affine: dacă $A \in \mathcal{M}_{m,n}(\mathbb{R})$ și $b \in \mathcal{M}_{m,1}(\mathbb{R})$ sunt respectiv o matrice și un vector de scalari atunci

$$Y \sim \mathcal{N}(\mu, \Sigma_Y) \implies AY + b \sim \mathcal{N}(A\mu + b, A\Sigma_Y A^\top)$$

De asemenea, plecând de la funcția caracteristică, se poate verifica proprietatea de independență a componentelor unui vector gaussian care ne spune că acestea sunt independente dacă și numai dacă matricea de variantă-covariantă este o matrice diagonală [Jacod and Protter, 2003].

Avem următorul rezultat care face legătura dintre repartiția normală și repartiția χ^2 :

Prop. 4.34



Fie $Y \sim \mathcal{N}(\mu, \Sigma_Y)$ un vector gaussian în \mathbb{R}^n . Dacă Σ_Y este o matrice inversabilă atunci

$$(Y - \mu)^\top \Sigma_Y^{-1} (Y - \mu) \sim \chi_n^2$$

repartiția χ^2 cu n grade de libertate.

Deoarece matricea de variantă-covariantă Σ_Y este o matrice simetrică și pozitiv definită atunci din Teorema de descompunere spectrală ea se poate descompune sub forma $\Sigma_Y = Q\Delta Q^\top$ unde $Q = [\mathbf{v}_1, \dots, \mathbf{v}_n]$ este o matrice ortogonală ($Q^\top = Q^{-1}$) formată din vectorii proprii $\mathbf{v}_1, \dots, \mathbf{v}_n$ corespunzători valorilor proprii $\lambda_1, \dots, \lambda_n$, iar Δ este matricea diagonală $\text{diag}(\lambda_1, \dots, \lambda_n)$. Dacă notăm cu $\Delta^{-\frac{1}{2}}$ matricea diagonală de coeficienți diagonali $\frac{1}{\sqrt{\lambda_1}}, \dots, \frac{1}{\sqrt{\lambda_n}}$ (aceste fracții există deoarece pozitiv definită matricei Σ_Y implică faptul că valorile proprii $\lambda_i > 0$, $i = 1, \dots, n$) atunci

$$\Sigma_Y = Q\Delta Q^\top \implies \Sigma_Y^{-1} = Q\Delta^{-1}Q^\top = \left(Q\Delta^{-\frac{1}{2}}Q^\top\right)\left(Q\Delta^{-\frac{1}{2}}Q^\top\right) = \Sigma_Y^{-\frac{1}{2}}\Sigma_Y^{-\frac{1}{2}}.$$

Prin urmare găsim că

$$(Y - \mu)^\top \Sigma_Y^{-1} (Y - \mu) = \left(\Sigma_Y^{-\frac{1}{2}}(Y - \mu)\right)^\top \left(\Sigma_Y^{-\frac{1}{2}}(Y - \mu)\right)$$

și cum vectorii gaussiani rămân gaussiani și prin aplicarea unor transformări affine (vezi rezultatul de mai sus) avem că

$$Y \sim \mathcal{N}(\mu, \Sigma_Y) \implies \Sigma_Y^{-\frac{1}{2}}(Y - \mu) \sim \mathcal{N}(0, I_n).$$

Astfel vectorul $V = [V_1, \dots, V_n]^\top = \Sigma_Y^{-\frac{1}{2}}(Y - \mu)$, care nu este altceva decât vectorul Y centrat și redus, este gaussian standard ($V_j \sim \mathcal{N}(0, 1)$ și V_i și V_j sunt independente) și

$$(Y - \mu)^\top \Sigma_Y^{-1} (Y - \mu) = \|V\|^2 = V_1^2 + \dots + V_n^2 \sim \chi_n^2. \square$$

Următorul rezultat, cunoscut sub denumirea de **Teorema lui Cochran** (a se vedea [Sen and Srivastava, 2012] sau [Greene, 2011] pentru o demonstrație a acestui rezultat), asigură că descompunerea unui vector gaussian în componente independente pe spații ortogonale ne dă vectori independenți a căror repartiție o putem explicita. Această teoremă poate fi văzută și ca o generalizare a Teoremei lui Pitagora:

Prop. 4.35



Fie $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, $\mathcal{M} \subset \mathbb{R}^n$ un subspațiu de dimensiune p , P matricea de proiecție ortogonală pe \mathcal{M} și $P_\perp = I_n - P$ matricea de proiecție ortogonală pe \mathcal{M}^\perp . Au loc următoarele proprietăți:

1. $PY \sim \mathcal{N}(P\mu, \sigma^2 P)$ și $P_\perp Y \sim \mathcal{N}(P_\perp \mu, \sigma^2 P_\perp)$
2. vectorii PY și $P_\perp Y = Y - PY$ sunt independenți
3. $\frac{\|P(Y - \mu)\|^2}{\sigma^2} \sim \chi_p^2$ și $\frac{\|P_\perp(Y - \mu)\|^2}{\sigma^2} \sim \chi_{n-p}^2$.

Acum avem instrumentele necesare pentru a deduce repartițiile estimatorilor în modelul gaussian de regresie liniară.

Prop. 4.36



Considerăm modelul de regresie liniară $Y = X\beta + \varepsilon$, sub ipotezele \mathcal{H}_1 și \mathcal{H}'_2 . Dacă presupunem că varianța σ^2 este cunoscută atunci au loc proprietățile

1. Vectorul $\hat{\beta}$ este un vector gaussian de medie β și matrice de varianță-covarianță $\sigma^2(X^\top X)^{-1}$, i.e. $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^\top X)^{-1})$
2. $\hat{\beta}$ și $\hat{\sigma}^2$ sunt independenți
3. $[n - (p + 1)]\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-(p+1)}^2$

1. Pentru a arăta că vectorul $\hat{\beta}$ este un vector gaussian vom folosi expresia acestuia obținută prin metoda celor mai mici pătrate

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y = (X^\top X)^{-1} X^\top (X\beta + \varepsilon) = \beta + (X^\top X)^{-1} X^\top \varepsilon.$$

Conform ipotezei \mathcal{H}'_2 , $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ este un vector gaussian prin urmare și $\hat{\beta} = \beta + \underbrace{(X^\top X)^{-1} X^\top}_{A} \varepsilon$ este tot un vector gaussian repartizat

$$\hat{\beta} = \beta + \underbrace{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top}_{\mathbf{A}} \varepsilon = \beta + \mathbf{A} \varepsilon \sim \mathcal{N}(\beta, \mathbf{A} \sigma^2 I_n \mathbf{A}^\top) = \mathcal{N}(\beta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$$

deoarece $\mathbf{A} \sigma^2 I_n \mathbf{A}^\top = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \sigma^2 I_n \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

2. Fie $\mathcal{M}(X)$ subspațiul lui \mathbb{R}^n generat de coloanele matricei \mathbf{X} și fie $P_X = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ matricea de proiecție ortogonală pe $\mathcal{M}(X)$. Avem

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \left(\underbrace{\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top}_{=P_X} \right) \mathbf{Y} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top P_X \mathbf{Y}$$

prin urmare $\hat{\beta}$ este un vector aleator ce depinde de $P_X \mathbf{Y}$. Observăm de asemenea că din definiția lui $\hat{\sigma}^2$,

$$\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n - (p + 1)} = \frac{\|\mathbf{Y} - P_X \mathbf{Y}\|^2}{n - (p + 1)}$$

acesta este o funcție de $\mathbf{Y} - P_X \mathbf{Y}$. Aplicând Teorema lui Cochran (Propoziția 4.35 de mai sus) observăm că vectorii $P_X \mathbf{Y}$ și $\mathbf{Y} - P_X \mathbf{Y}$ sunt independenți prin urmare și $\hat{\beta}$ și $\hat{\sigma}^2$ sunt independenți ca funcții de vectori independenți.

3. Notând cu P_{X^\perp} proiecția ortogonală pe subspațiul ortogonal $\mathcal{M}(X)^\perp$, subspațiu de dimensiune $n - (p + 1)$, avem

$$\hat{\varepsilon} = \mathbf{Y} - P_X \mathbf{Y} = P_{X^\perp} \mathbf{Y} = P_{X^\perp} \left(\underbrace{\mathbf{X} \beta}_{\in \mathcal{M}(X)} + \varepsilon \right) = P_{X^\perp} \varepsilon,$$

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

Din Teorema lui Cochran găsim că

$$(n - (p + 1)) \frac{\hat{\sigma}^2}{\sigma^2} = \frac{\|P_{X^\perp} \varepsilon\|^2}{\sigma^2} = \frac{\|P_{X^\perp} (\varepsilon - \mathbb{E}[\varepsilon])\|^2}{\sigma^2} \sim \chi^2_{n - (p + 1)}. \square$$

Putem observa că rezultatul anterior nu este suficient pentru a determina o regiune de încredere pentru coeficienții modelului de regresie β deoarece varianța σ^2 a fost presupusă cunoscută, fenomen care nu este aplicabil în caz general. Rezultatul următor vine să acopere acest deficit.

Prop. 4.37



Considerăm modelul de regresie liniară $\mathbf{Y} = \mathbf{X} \beta + \varepsilon$, sub ipotezele \mathcal{H}_1 și \mathcal{H}'_2 și presupunem că varianța σ^2 nu este cunoscută. Atunci au loc proprietățile

- pentru $j \in \{0, 1, \dots, p\}$, notând cu $[(\mathbf{X}^\top \mathbf{X})^{-1}]_{j+1, j+1}$ elementul $j+1$ de pe diagonala matricei $(\mathbf{X}^\top \mathbf{X})^{-1}$, avem

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 [(\mathbf{X}^\top \mathbf{X})^{-1}]_{j+1, j+1}), \quad T_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{j+1, j+1}}} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n - (p + 1)}$$

- fie $R \in \mathcal{M}_{q, p+1}(\mathbb{R})$ o matrice de rang q ($q \leq p + 1$) atunci

$$\frac{1}{q \hat{\sigma}^2} \left(R(\hat{\beta} - \beta) \right)^\top [R(\mathbf{X}^\top \mathbf{X})^{-1} R^\top]^{-1} R(\hat{\beta} - \beta) \sim F_{q, n - (p + 1)}$$

1. Din Propoziția 4.36 am văzut că $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$ prin urmare

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 [(\mathbf{X}^\top \mathbf{X})^{-1}]_{j+1,j+1})$$

ceea ce implica $\frac{\hat{\beta}_j - \beta_j}{\sigma \sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{j+1,j+1}}} \sim \mathcal{N}(0, 1)$.

Scriind

$$T_j = \frac{\frac{\hat{\beta}_j - \beta_j}{\sigma}}{\sqrt{\frac{(n-(p+1))\hat{\sigma}^2}{n-(p+1)}}} = \frac{\frac{\hat{\beta}_j - \beta_j}{\sigma \sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{j+1,j+1}}}}{\sqrt{\frac{(n-(p+1))\hat{\sigma}^2}{n-(p+1)}}}$$

și ținând cont de faptul că $(n-(p+1))\hat{\sigma}^2 \sim \chi^2_{n-(p+1)}$ iar $\hat{\beta}_j$ și $\hat{\sigma}^2$ sunt independente, deducem că $T_j \sim t_{n-(p+1)}$.

2. Să remarcăm că matricea $R(\mathbf{X}^\top \mathbf{X})^{-1}R^\top$ pătratică de ordin q este inversabilă deoarece matricea $(\mathbf{X}^\top \mathbf{X})^{-1}$ este de rang $p+1$, $p+1 \geq q$. Cum $\hat{\beta}$ este un vector gaussian deducem că și $R\hat{\beta}$ este un vector gaussian de medie și matrice de covarianță

$$R\hat{\beta} \sim \mathcal{N}(R\beta, \sigma^2 R(\mathbf{X}^\top \mathbf{X})^{-1}R^\top)$$

Astfel, conform Propoziției 4.34 privind legătura dintre repartiția normală și repartiția χ^2 , rezultă că

$$\frac{1}{\sigma^2} (R(\hat{\beta} - \beta))^\top [R(\mathbf{X}^\top \mathbf{X})^{-1}R^\top]^{-1} R(\hat{\beta} - \beta) \sim \chi_q^2.$$

În expresia de mai sus, înlocuim pe σ^2 cu $\hat{\sigma}^2$ și ținând cont că $\hat{\beta}$ și $\hat{\sigma}^2$ sunt independente iar $(n-(p+1))\hat{\sigma}^2 \sim \chi^2_{n-(p+1)}$ concluzionăm că

$$\frac{\frac{1}{\hat{\sigma}^2} (R(\hat{\beta} - \beta))^\top [R(\mathbf{X}^\top \mathbf{X})^{-1}R^\top]^{-1} R(\hat{\beta} - \beta)}{\frac{q}{(n-(p+1))\hat{\sigma}^2}} \sim \frac{\frac{\chi_q^2}{q}}{\frac{\chi^2_{n-(p+1)}}{n-(p+1)}} = F_{q, n-(p+1)}. \square$$

Exemplu: aplicație pentru regresia liniară simplă

Vom exemplifica rezultatul celui de-al doilea subpunkt al Propoziției 4.37 pentru cazul particular în care $p+1 = q = 2$ iar matricea $R = I_2$. Avem că

$$R(\hat{\beta} - \beta) = \begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \end{pmatrix}$$

și dacă termenul liber face parte din modelul de regresie atunci matricea $\mathbf{X}^\top \mathbf{X}$ devine (a se vedea și Exemplul 4.17)

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}.$$

Astfel găsim că $\frac{1}{q\hat{\sigma}^2} (R(\hat{\beta} - \beta))^\top [R(\mathbf{X}^\top \mathbf{X})^{-1}R^\top]^{-1} R(\hat{\beta} - \beta)$ se rescrie ca

$$\frac{1}{2\hat{\sigma}^2} \left[n(\hat{\beta}_0 - \beta_0)^2 + 2n\bar{x}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + \sum_{i=1}^n x_i^2(\hat{\beta}_1 - \beta_1)^2 \right] \sim F_{2,n-2}$$

proprietate care coincide cu rezultatul obținut în cazul regresiei liniare simple. Mai mult, dacă $p+1 = q$ și $R = I_{p+1}$ atunci repartiția distanței (ponderate de inversa matricei de covarianță) dintre estimatorul obținut prin metoda celor mai mici pătrate $\hat{\beta}$ și valoarea reală β este $F_{p+1,n-(p+1)}$:

$$\frac{1}{\hat{\sigma}^2(p+1)}(\hat{\beta} - \beta)^\top (X^\top X)^{-1}(\hat{\beta} - \beta) \sim F_{p+1,n-(p+1)}.$$

4.6.3 Intervale și regiuni de încredere

Rezultatele din propozițiile anterioare conduc la următoarele intervale (atunci când parametrii sunt considerați separati) și regiuni de încredere (atunci când luăm în calcul mai mulți parametrii simultan prin urmare ținem cont și de dependența dintre ei) pentru parametrii modelului de regresie liniară:

- 1) Un interval de încredere bilateral de nivel de încredere $1 - \alpha$ pentru parametrul β_j , $j \in \{0, 1, \dots, p\}$ este dat de

$$IC^{1-\alpha}(\beta_j) = \left[\hat{\beta}_j - t_{n-(p+1)} \left(1 - \frac{\alpha}{2}\right) \hat{\sigma} \sqrt{[(X^\top X)^{-1}]_{j+1,j+1}}, \hat{\beta}_j + t_{n-(p+1)} \left(1 - \frac{\alpha}{2}\right) \hat{\sigma} \sqrt{[(X^\top X)^{-1}]_{j+1,j+1}} \right]$$

unde $t_{n-(p+1)} \left(1 - \frac{\alpha}{2}\right)$ este cuantila de ordin $1 - \frac{\alpha}{2}$ a repartiției Student $t_{n-(p+1)}$.

- 2) Un interval de încredere bilateral de nivel de încredere $1 - \alpha$ pentru σ^2 este dat de

$$IC^{1-\alpha}(\sigma^2) = \left[\frac{[n - (p+1)]\hat{\sigma}^2}{\chi_{n-(p+1)}^2 \left(1 - \frac{\alpha}{2}\right)}, \frac{[n - (p+1)]\hat{\sigma}^2}{\chi_{n-(p+1)}^2 \left(\frac{\alpha}{2}\right)} \right]$$

unde $\chi_{n-(p+1)}^2 \left(1 - \frac{\alpha}{2}\right)$ și $\chi_{n-(p+1)}^2 \left(\frac{\alpha}{2}\right)$ sunt cuantilele de ordin $1 - \frac{\alpha}{2}$ și respectiv $\frac{\alpha}{2}$ a repartitiei $\chi_{n-(p+1)}^2$.

- 3) O regiune de încredere de nivel de încredere $1 - \alpha$ pentru q ($q \leq p+1$) parametrii β_j , notați $(\beta_{j_1}, \dots, \beta_{j_q})$, este dată de
 - atunci când σ este cunoscută, de

$$RC^{1-\alpha}(R\beta) = \left\{ R\beta \in \mathbb{R}^q \mid \frac{1}{\sigma^2} (R(\hat{\beta} - \beta))^\top [R(X^\top X)^{-1} R^\top]^{-1} R(\hat{\beta} - \beta) \leq \chi_q^2(1 - \alpha) \right\}$$

- atunci când σ este necunoscută, de

$$RC^{1-\alpha}(R\beta) = \left\{ R\beta \in \mathbb{R}^q \mid \frac{1}{q\hat{\sigma}^2} (R(\hat{\beta} - \beta))^\top [R(X^\top X)^{-1} R^\top]^{-1} R(\hat{\beta} - \beta) \leq f_{q,n-(p+1)}(1 - \alpha) \right\}$$

unde R este o matrice de dimensiune $q \times (p+1)$ cu toate elementele nule exceptând elementele R_{i,j_i} care sunt egale cu 1 iar $f_{q,n-(p+1)}(1 - \alpha)$ este cuantila de ordin $1 - \alpha$ a repartiției Fisher-Snedecor $F_{q,n-(p+1)}$ cu q grade de libertate la numărător și $n - (p+1)$ grade de libertate la numitor.

Exemplu: regiune de încredere caz particular

Ca ilustrare a regiunii de încredere de mai sus, să presupunem că $p \geq 2$ și $q = 2$ iar matricea R este

$$R = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$$

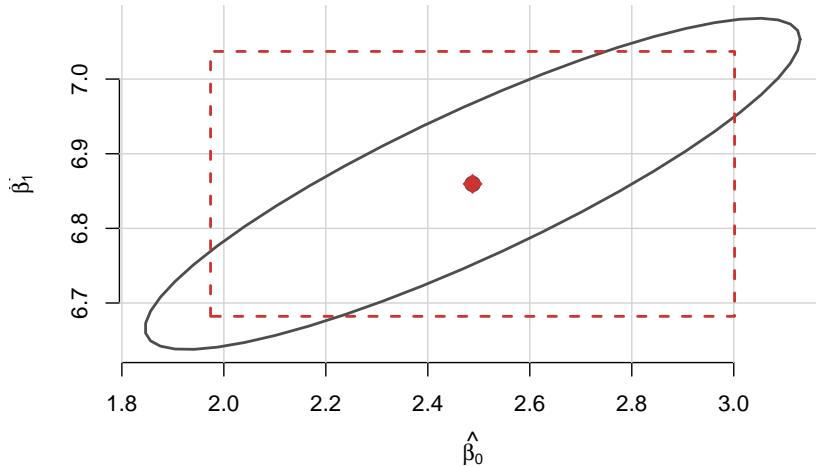
cea ce conduce la $R(\hat{\beta} - \beta) = \begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \end{pmatrix}$. Dacă presupunem că σ^2 este necunoscut atunci regiunea de încredere pentru (β_0, β_1) este dată de

$$RC^{1-\alpha}(\beta_0, \beta_1) = \left\{ (\beta_0, \beta_1) \in \mathbb{R}^2 \mid \frac{1}{2\hat{\sigma}^2} (\hat{\beta}_0 - \beta_0 \quad \hat{\beta}_1 - \beta_1) [R(\mathbf{X}^\top \mathbf{X})^{-1} R^\top]^{-1} \begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \end{pmatrix} \leq f_{2,n-(p+1)}(1 - \alpha) \right\}$$

și notând cu c_{ij} elementul de pe linia i coloana j a matricei $(\mathbf{X}^\top \mathbf{X})^{-1}$, găsim că

$$RC^{1-\alpha}(\beta_0, \beta_1) = \left\{ (\beta_0, \beta_1) \in \mathbb{R}^2 \mid \frac{c_{22}(\hat{\beta}_0 - \beta_0)^2 - 2c_{12}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + c_{11}(\hat{\beta}_1 - \beta_1)^2}{2\hat{\sigma}^2(c_{11}c_{22} - c_{12}^2)} \leq f_{2,n-(p+1)}(1 - \alpha) \right\}$$

cea ce arată că regiunea are formă de elipsă (vezi [Ornea and Turtoi, 2000, Capitolul 2, Secțiunea 16]). Figura următoare ilustrează această regiune pentru date simulate conform modelului de regresie liniară simplă $y_i = 3 + 7x_i - \varepsilon_i$, cu $x_i \sim \mathcal{U}[-5, 5]$ iar $\varepsilon_i \sim \mathcal{N}(0, 4)$:



Exemplu: prețul chirilor în Munchen

Ex. 4.40 În acest exemplu vom explicita relația dintre prețul chiriei pe metrul pătrat în raport cu inversul suprafetei de locuit și anul de construcție a imobilului folosind modelul de regresie prezentat în Exemplul 4.10 în care am considerat variabila centrată *suprafata_inv_cen* și variabilele ortogonale *an_co_j*

$$\widehat{\text{pret_m}^2}_i = \beta_0 + \beta_1 \times \text{suprafata_inv_cen} + \beta_2 \times \text{an_co_1} + \beta_3 \times \text{an_co_2} + \beta_4 \times \text{an_co_3}$$

Modelul estimat este

$$\widehat{\text{pret_m}^2}_i = 7.111 + 129.572 \times \text{suprafata_inv_cen} + 43.938 \times \text{an_co_1} + 27.539 \times \text{an_co_2} - 1.756 \times \text{an_co_3}$$

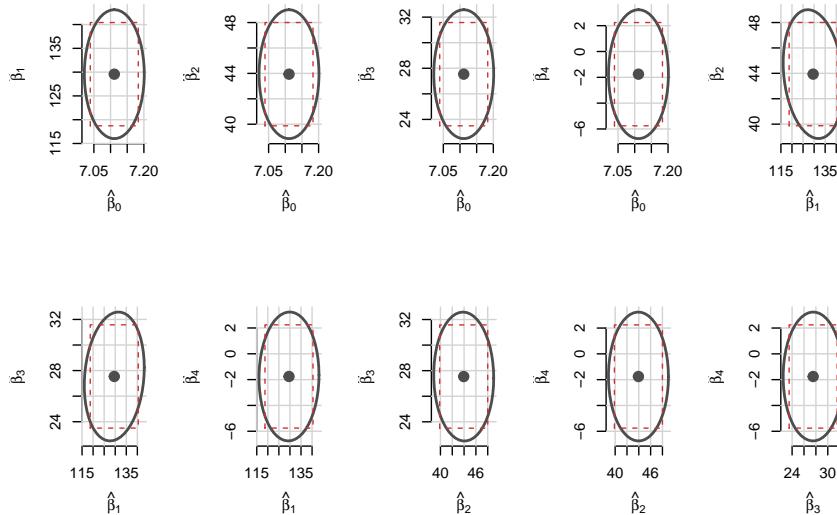
Intervalele de încredere pentru coeficienții modelului de regresie la un nivel de încredere de $1 - \alpha = 95\%$ sunt

$$\begin{aligned} IC(\hat{\beta}_0)^{1-\alpha} &= [7.039, 7.183], \\ IC(\hat{\beta}_1)^{1-\alpha} &= [118.717, 140.426], \\ IC(\hat{\beta}_2)^{1-\alpha} &= [39.875, 48.002], \\ IC(\hat{\beta}_3)^{1-\alpha} &= [23.501, 31.577], \\ IC(\hat{\beta}_4)^{1-\alpha} &= [-5.756, 2.244], \end{aligned}$$

iar intervalul de încredere pentru σ^2 la același nivel de încredere este

$$IC(\hat{\sigma}^2)^{1-\alpha} = [3.9591228, 4.3753046].$$

În figura de mai jos sunt prezentate regiunile de încredere pentru toate perechile de parametri ai modelului împreună cu intervalele de încredere corespunzătoare (elipse versus dreptunghiuri):



Atunci când vrem să facem o predicție dorim să evaluăm certitudinea cu care această predicție este efectuată. Având în vedere variabilitatea noii valori pe care vrem să o prezicem este de așteptat ca intervalul de predicție să fie mai mare decât cel de încredere la același nivel de încredere.

Prop. 4.41



Fie $(x_{n+1,1}, \dots, x_{n+1,p})$ o nouă observație și considerăm $\mathbf{x}_{n+1}^\top = (1, x_{n+1,1}, \dots, x_{n+1,p})$. Ne propunem să prezicem valoarea y_{n+1} conform modelului

$$y_{n+1} = \mathbf{x}_{n+1}^\top \boldsymbol{\beta} + \varepsilon_{n+1}$$

cu $\varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2)$ independentă de ε_i , $1 \leq i \leq n$.

Arătați că un interval de predicție de nivel de încredere $1 - \alpha$ pentru y_{n+1} este dat de

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

Plecând de la un eșantion de talie n , $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$, găsim estimatorul de verosimilitate maximă (același cu cel obținut prin metoda celor mai mici pătrate) $\hat{\beta}$ a lui β cu ajutorul căruia putem prezice valoarea \hat{y}_{n+1} după relația

$$\hat{y}_{n+1} = \mathbf{x}_{n+1}^\top \hat{\beta}.$$

Pentru a cuantifica eroarea de predicție $y_{n+1} - \hat{y}_{n+1}$ folosim următoarea descompunere

$$y_{n+1} - \hat{y}_{n+1} = \mathbf{x}_{n+1}^\top (\beta - \hat{\beta}) + \varepsilon_{n+1},$$

care este o sumă de două variabile gaussiene independente, ε_{n+1} și $\hat{\beta}$ care este un vector gaussian care depinde de ε_i , $1 \leq i \leq n$. Prin urmare $y_{n+1} - \hat{y}_{n+1}$ este un vector gaussian și ținând cont că matricea de variantă-covariantă $Var(\hat{\varepsilon}_{n+1}) = \sigma^2 (1 + \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1})$ avem

$$y_{n+1} - \hat{y}_{n+1} \sim \mathcal{N}(0, \sigma^2 (1 + \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1})).$$

Această relație se scrie sub forma

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\sigma \sqrt{(1 + \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1})}} \sim \mathcal{N}(0, 1)$$

și înlocuind σ cu estimatorul său $\hat{\sigma}$ obținem

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\hat{\sigma} \sqrt{(1 + \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1})}} = \frac{\frac{y_{n+1} - \hat{y}_{n+1}}{\sigma \sqrt{(1 + \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1})}}}{\frac{\sigma}{\hat{\sigma}}} = \frac{\frac{y_{n+1} - \hat{y}_{n+1}}{\sigma \sqrt{(1 + \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1})}}}{\sqrt{\frac{[n-(p+1)]\sigma}{n-(p+1)}}} \sim \frac{\mathcal{N}(0, 1)}{\sqrt{\frac{\chi^2_{n-(p+1)}}{n-(p+1)}}}.$$

Remarcăm că numărătorul și numitorul sunt variabile aleatoare independente deoarece $y_{n+1} - \hat{y}_{n+1} = \mathbf{x}_{n+1}^\top (\beta - \hat{\beta}) + \varepsilon_{n+1}$ este independentă de $\hat{\sigma}$ pentru că $\hat{\sigma}$ este independentă de $\hat{\beta}$ și de ε_{n+1} (estimatorul $\hat{\sigma}$ nu depinde decât de ε_i , $1 \leq i \leq n$). Prin urmare găsim că

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\hat{\sigma} \sqrt{(1 + \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1})}} \sim t_{n-(p+1)}$$

de unde rezultă intervalul de încredere dorit. \square

4.6.4 Testare de ipoteze statistice

În această secțiune vom prezenta o serie de teste statistice de semnificație pentru coeficienții necunoscuți β ai modelului de regresie liniară. Pentru aceasta ne vom plasa în contextul modelului de regresie liniară multiplă sub ipoteza de normalitate a termenilor eroare \mathcal{H}'_2 , chiar dacă mare parte din rezultatele ce vor fi prezentate sunt robuste la deviații mici de la această ipoteză în special atunci când talia eșantionului n este mare [Sen and Srivastava, 2012] sau [Rencher and Schaalje, 2008].

Vom începe prin a da un exemplu în care se justifică necesitatea testării de ipoteze statistice în contextul modelului de regresie.

Exemplu: prețul chirilor în München - testare de ipoteze statistice

Exp. 4.42 Ne plasăm în contextul setului de date referitor la prețul chiriiilor apartamentelor din orașul Munchen. În acest exemplu vom folosi atât datele din anul 1999 cât și pe cele din anul 2001 pentru a face o comparație între prețuri. Setul de date **Munchen2** conține $n = 4571$ observații referitoare la prețul chiriei apartamentelor pentru anii 1999 (3082 observații) și 2001 (1489 observații). Considerăm modelul de regresie

$$\begin{aligned} \widehat{\text{pret_m}^2}_i &= \beta_0 + \beta_1 \times \text{suprafata_inv_cen}_i + \beta_2 \times \text{an_co_1} + \beta_3 \times \text{an_co_2} + \beta_4 \times \text{an_co_3} \\ &\quad + \beta_5 \times \text{nbucatarie} + \beta_6 \times \text{pbucatarie} + \beta_7 \times \text{an01} + \varepsilon_i \end{aligned}$$

unde *suprafata_inv_cen* este covariabila dată de inversa suprafeței de locuit centrată în zero, *an_co_j* sunt variabilele rezultate prin aplicarea polinoamelor ortogonale de grad 3 (a se vedea Exemplul 4.10) anului de construcție, *nbucatarie* și *pbucatarie* sunt variabile indicator ajutătoare care specifică tipul de bucătărie al locuinței și sunt rezultate din variabila categorială **bucătărie** care avea nivelele *substandard* (categorie de referință), *standard/normală* și respectiv *premium*. Covariabila *an01* este o variabilă indicator care precizează dacă observația este din anul 1999 (*an01* = 0) sau 2001 (*an01* = 1).

Modelul estimat rezultat este

$$\begin{aligned} \widehat{\text{pret_m}^2}_i &= 6.932 + 123.77 \times \text{suprafata_inv_cen}_i + 49.373 \times \text{an_co_1} + 29.5 \times \text{an_co_2} \\ &\quad - 0.884 \times \text{an_co_3} + 1.043 \times \text{nbucatarie} + 1.302 \times \text{pbucatarie} - 0.185 \times \text{an01} \end{aligned}$$

ceea ce arată că prețul mediu net pe metrul pătrat al chiriei pentru apartamentele din anul 2001 scade cu aproape 0.185 euro față de cele din anul 1999. Se poate pune întrebarea dacă această descreștere este semnificativă (poate fi extrapolată la întreaga populație) sau este datorată variabilității eșantionării aleatoare, prin urmare vrem să verificăm dacă parametrul β_7 este diferit în mod semnificativ față de 0 ceea ce în termeni de teste statistice se scrie sub forma:

$$H_0 : \beta_7 = 0 \quad \text{versus} \quad H_1 : \beta_7 \neq 0.$$

Alte întrebări asupra acestui model pot să apară, de exemplu putem să verificăm dacă tipul de bucătărie este semnificativ în explicarea prețului chiriei pe metrul pătrat, adică dacă coeficienții β_5 și β_6 sunt simultan 0:

$$H_0 : \begin{pmatrix} \beta_5 \\ \beta_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{versus} \quad H_1 : \begin{pmatrix} \beta_5 \\ \beta_6 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

O altă întrebare, ținând seama că estimatorii coeficienților $\hat{\beta}_5 = 1.043$ și $\hat{\beta}_6 = 1.302$ nu sunt foarte diferenți, ar fi dacă este necesar să diferențiem între apartamentele cu bucătărie standard și cele cu bucătărie premium:

$$H_0 : \beta_5 = \beta_6 \quad \text{versus} \quad H_1 : \beta_5 \neq \beta_6. \square$$

Toate întrebările puse în Exemplul 4.42 sunt ilustrative pentru modelul de regresie în general și testele statistice de semnificație corespunzătoare pot fi scrise sub forma:

1. *Teste asupra modelelor imbricate* - în acest caz avem un test compus asupra subvectorului $\beta_0 = (\beta_0, \dots, \beta_{p_0})^\top$

$$H_0 : \beta_0 = \mathbf{0} \quad \text{versus} \quad H_1 : \beta_0 \neq \mathbf{0}.$$

2. *Test de semnificație asupra unui coeficient*

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0.$$

3. *Testul global asupra tuturor covariabilelor din model*

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad \text{versus} \quad H_1 : \exists j \in \{1, 2, \dots, p\} \text{ cu } \beta_j \neq 0.$$

4. Testul asupra unei ipoteze liniare generalizate

$$H_0 : R\beta = r \quad \text{versus} \quad H_1 : R\beta \neq r$$

unde R este o matrice dimensiune $q \times (p+1)$ cu $\text{rang}(R) = q \leq p+1$ iar r este un vector de dimensiune q . În secțiunile următoare vom încerca să construim pentru fiecare astfel de test statistica de test precum și regiunea critică corespunzătoare.

4.6.4.1 Teste asupra modelelor imbricate

Considerăm modelul de regresie liniară

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

sub ipotezele $\mathcal{H}_1 : \text{rang}(\mathbf{X}) = p+1$ și $\mathcal{H}'_2 : \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. În particular, acest model ne spune că $\mathbb{E}[\mathbf{Y}] = \mathbf{X}\beta \in \mathcal{M}(X)$ unde $\mathcal{M}(X)$ este subspațiul de dimensiune $p+1$ a lui \mathbb{R}^n generat de coloanele matricei de design \mathbf{X} .

Pentru început, în această secțiune ne propunem să testăm dacă ultimii q coeficienți, $q = p - p_0 \leq p+1$, sunt nuli, altfel spus vrem să testăm ipotezele

$$H_0 : \beta_{p_0+1} = \cdots = \beta_p = 0 \quad \text{versus} \quad H_1 : \exists j \in \{p_0+1, \dots, p\} \beta_j \neq 0$$

În termeni de model, ipoteza H_0 ne spune că modelul revine la

$$\mathbf{Y} = \mathbf{X}_0\beta_0 + \varepsilon_0$$

sub ipotezele $\mathcal{H}_1 : \text{rang}(\mathbf{X}_0) = p_0+1$ și $\mathcal{H}'_2 : \varepsilon_0 \sim \mathcal{N}(0, \sigma^2 I_n)$, unde matricea $\mathbf{X}_0 \in \mathcal{M}_{n,p_0+1}(\mathbb{R})$ este compusă din primele p_0+1 coloane ale lui \mathbf{X} iar $\beta_0 \in \mathcal{M}_{p_0+1,1}(\mathbb{R})$ (în acest model am inclus și termenul liber și acesta este motivul pentru care avem p_0+1 și nu p_0).

Fie $\mathcal{M}(X_0)$ subspațiul lui $\mathcal{M}(X)$ de dimensiune p_0 generat de coloanele lui \mathbf{X}_0 (deoarece $\text{rang}(\mathbf{X}) = p+1$ deducem că $\text{rang}(\mathbf{X}_0) = p_0+1$). Sub ipoteza H_0 avem că $\mathbb{E}[\mathbf{Y}] = \mathbf{X}_0\beta_0 \in \mathcal{M}(X_0)$.

Odată ce avem fixate ipotezele procedurii de testare, mai rămâne de a determina statistică de test corespunzătoare și regiunea critică a testului. Următoarea propoziție face lumină în acest sens folosind o abordare geometrică (pentru o versiune analitică se poate consulta Propoziția 4.48).



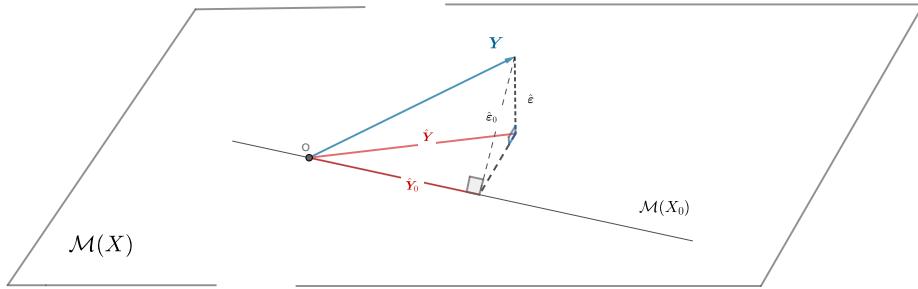
Sub ipoteza nulă H_0 statistică de test

Prop. 4.43

$$F = \frac{n - (p+1)}{q} \frac{\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|^2}{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2} = \frac{n - (p+1)}{q} \frac{RSS_0 - RSS}{RSS} \sim F_{q, n-(p+1)}$$

unde $q = p - p_0$ iar $F_{q, n-(p+1)}$ este repartiția lui Fisher cu $(q, n - (p+1))$ grade de libertate.

Vom prezenta mai jos o abordare din punct de vedere geometric a problemei de testare. Considerăm spațiul $\mathcal{M}(X_0) \subset \mathcal{M}(X)$ și am văzut că sub ipoteza nulă avem $\mathbb{E}[\mathbf{Y}] = \mathbf{X}_0\beta_0 \in \mathcal{M}(X_0)$. În această situație, metoda celor mai mici pătrate consistă în proiectarea lui \mathbf{Y} pe spațiul $\mathcal{M}(X_0)$ pentru a obține vectorul valorilor ajustate $\hat{\mathbf{Y}}_0$ (vezi figura de mai jos).



Ideea procedurii de testare, și prin urmare a deciziei de a respinge sau nu ipoteza nulă, este următoarea: dacă proiecția \hat{Y}_0 a lui \mathbf{Y} pe $\mathcal{M}(X_0)$ este aproape de proiecția \hat{Y} a lui \mathbf{Y} pe $\mathcal{M}(X)$ atunci nu respingem ipoteza nulă, în caz contrar o respingem în favoarea alternativei H_1 . De fapt, dacă informația adusă de cele două modele nu diferă foarte mult atunci este recomandat să păstrăm modelul mai simplu (principiul parcimoniei).

Pentru a cuantifica *apropierea* dintre cele două proiecții putem considera distanța euclidiană $\|\hat{Y} - \hat{Y}_0\|^2$ dintre acestea. Această distanță variază în funcție de datele și de unitățile de măsură a acestora și pentru a evita această problemă de scară se recomandă *standardizarea* acestei distanțe prin împărțirea ei la pătratul normei erorii reziduale $\|\hat{\epsilon}\|^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = [n - (p + 1)]\hat{\sigma}^2$. Cum vectorii aleatori $\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0$ și $\hat{\epsilon}$ nu aparțin în subspații de aceeași dimensiune trebuie să îi împărțim și prin gradele de libertate corespunzătoare, respectiv $q = p - p_0$ și $n - (p + 1)$. Obținem astfel statistică de test

$$F = \frac{\frac{\|\hat{Y} - \hat{Y}_0\|^2}{p - p_0}}{\frac{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2}{n - (p + 1)}} = \frac{n - (p + 1)}{p - p_0} \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2}.$$

Pentru a putea utiliza această statistică de test este necesar să determinăm repartiția sa sub ipoteza nulă. Notând cu P_X și respectiv P_{X_0} matricele de proiecție ortogonală pe $\mathcal{M}(X)$ și $\mathcal{M}(X_0)$, stim că

$$\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0 = P_X \mathbf{Y} - P_{X_0} \mathbf{Y}$$

și cum $\mathcal{M}(X_0) \subset \mathcal{M}(X)$, deci $P_{X_0} \mathbf{Y} = P_{X_0} P_X \mathbf{Y}$ și

$$\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0 = P_X \mathbf{Y} - P_{X_0} \mathbf{Y} = (I_n - P_{X_0}) P_X \mathbf{Y} = P_{X_0}^\perp P_X \mathbf{Y}.$$

Astfel am găsit că $\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0 \in \mathcal{M}(X_0)^\perp \cap \mathcal{M}(X)$ și cum $\mathbf{Y} - \hat{\mathbf{Y}} \in \mathcal{M}(X)^\perp$ găsim că $\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0 \perp \mathbf{Y} - \hat{\mathbf{Y}}$. Vectorii $\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0$ și $\mathbf{Y} - \hat{\mathbf{Y}}$ sunt elemente din spații ortogonale ceea ce implică $\text{Cov}(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0, \mathbf{Y} - \hat{\mathbf{Y}}) = 0$ și cum sunt și vectori gaussieni (conform ipotezei \mathcal{H}'_2) deducem că sunt și independenți ceea ce arată că numărătorul și numitorul lui F sunt independente.

Conform Teoremei lui Cochran (a se vedea Propoziția 4.35) avem pentru numitor

$$\frac{1}{\sigma^2} \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \frac{1}{\sigma^2} \|P_{X^\perp} \mathbf{Y}\|^2 = \frac{1}{\sigma^2} \|P_{X^\perp} (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\|^2 = \frac{1}{\sigma^2} \|P_{X^\perp} \boldsymbol{\varepsilon}\|^2 \sim \chi^2_{n-(p+1)}$$

iar pentru numărător

$$\frac{1}{\sigma^2} \|P_{X_0}^\perp P_X (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\|^2 \sim \chi^2_q$$

Sub H_0 , parametrul $\|P_{X_0}^\perp P_X \mathbf{X}\boldsymbol{\beta}\|^2$ este nul deoarece $\mathbf{X}\boldsymbol{\beta} \in \mathcal{M}(X_0)$ ceea ce conduce la

$$\frac{1}{\sigma^2} \|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|^2 = \frac{1}{\sigma^2} \|P_{X_0}^\perp P_X \mathbf{Y}\|^2 \stackrel{H_0}{=} \frac{1}{\sigma^2} \|P_{X_0}^\perp P_X (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\|^2 \sim \chi^2_q$$

prin urmare

$$F \xrightarrow{H_0} F_{q,n-(p+1)}.$$

Putem observa de asemenea că aplicând Teorema lui Pitagora (în triunghiul punctat cu negru)

$$\begin{aligned} \|\mathbf{Y} - \hat{\mathbf{Y}}_0\|^2 &= \|\mathbf{Y} - P_X \mathbf{Y} + P_X \mathbf{Y} - P_{X_0} \mathbf{Y}\|^2 = \|(I_n - P_X) \mathbf{Y} + P_{X_0} P_X \mathbf{Y} - P_{X_0} \mathbf{Y}\|^2 \\ &= \|P_{X^\perp} \mathbf{Y} + (I_n - P_{X_0}) P_X \mathbf{Y}\|^2 = \|P_{X^\perp} \mathbf{Y} + P_{X_0^\perp} P_X \mathbf{Y}\|^2 \\ &= \|P_{X^\perp} \mathbf{Y}\|^2 + \|P_{X_0^\perp} P_X \mathbf{Y}\|^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 + \|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|^2 \end{aligned}$$

altfel spus

$$\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}_0\|^2 - \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = RSS_0 - RSS$$

ceea ce justifică a doua scriere a statisticii de test. \square

Se poate arăta [Rencher and Schaalje, 2008, Teorema 8.3] că dacă termenul liber face parte din ambele modele (sau nu face parte din niciunul) atunci statistică de test precedentă se poate exprima în funcție de coeficienții de determinare R^2 și respectiv R_0^2 corespunzători modelelor considerate:

$$F = \frac{n - (p + 1)}{q} \frac{R^2 - R_0^2}{1 - R^2}$$

prin urmare dacă avem la dispoziție coeficienții de determinare ai celor două modele putem efectua testul de semnificație a lui Fisher pentru modele imbricate.

Exemplu: prețul chirilor în München - test asupra coeficienților

Exp. 4.44 În contextul modelului de regresie din Exemplul 4.42 ne propunem să testăm semnificația efectului bucătăriei asupra prețului pe metrul pătrat. Astfel considerăm ipotezele

$$H_0 : \begin{pmatrix} \beta_5 \\ \beta_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{versus} \quad H_1 : \begin{pmatrix} \beta_5 \\ \beta_6 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Pentru a calcula statistică de test (în R folosim comanda `linearHypothesis` din pachetul `car`) vom folosi expresia $F = \frac{n-(p+1) RSS_0 - RSS}{q} \frac{RSS_0 - RSS}{RSS}$ unde în cazul nostru avem $n - (p + 1) = 4565$ ($n = 4571$ și $p = 7$), $q = 2$ ($p_0 = 5$), $RSS_0 = 1.8236158 \times 10^4$ și $RSS = 1.7602848 \times 10^4$ ceea ce rezultă într-o valoare a statisticii de $F = 82.083$.

La un nivel de semnificativitate de $\alpha = 0.05$ găsim că cuantila repartiției Fisher este egală cu

$$f_{2,4563}(0.95) = 2.998$$

și cum $F = 82.083 > 2.998 = f_{2,4563}(0.95)$ putem respinge ipoteza nulă la acest prag de semnificație. Astfel putem concluziona că efectul tipului de bucatarie prezintă o influență semnificativă asupra prețului mediu net al chiriei pe metrul pătrat. Cu toate acestea astă nu înseamnă că ambii coeficienți sunt nenuli, unul ar putea fi nul pentru a respinge H_0 . \square

4.6.4.2 Testul Student de semnificație a unui coeficient În această secțiune vom considera un caz particular al testului pentru modele imbricate.

Prop. 4.45



Ne propunem să testăm ipoteza nulă $H_0 : \beta_j = 0$ versus ipoteza alternativă $H_1 : \beta_j \neq 0$. Statistica de test este

$$F = \frac{\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|^2}{\hat{\sigma}^2} = \frac{\hat{\beta}_j^2}{\hat{\sigma}_{\hat{\beta}_j}^2} \sim F_{1,n-(p+1)}(1-\alpha)$$

care este echivalentă cu statistica $T_j = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$.

Această problemă este un caz particular al rezultatului obținut anterior pentru $q = 1$. Avem statistică de test

$$F = \frac{n-(p+1)}{1} \frac{\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|^2}{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2} = \frac{\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|^2}{\hat{\sigma}^2}$$

u ajutorul căreia putem construi regiunea critică

$$C = \{\omega \mid F(\omega) > f_{1,n-(p+1)}(1-\alpha)\}.$$

În modelul redus, \mathbf{X}_0 este matricea $\mathbf{X} = [\mathbf{X}_0 | X_{j+1}]$ fără coloana $j+1$ (corespunzătoare coeficientului β_j) cu $j = 0, 1, \dots, p$, prin urmare

$$F = \frac{\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|^2}{\hat{\sigma}^2} = \frac{\|\mathbf{X}\hat{\beta} - P_{\mathbf{X}_0}\mathbf{X}\hat{\beta}\|^2}{\hat{\sigma}^2} = \frac{\|X_{j+1}\hat{\beta}_j - \hat{\beta}_j P_{\mathbf{X}_0}X_{j+1}\|^2}{\hat{\sigma}^2} = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2} X_{j+1}^\top (I - P_{\mathbf{X}_0}) X_{j+1}.$$

Ținând cont că $\mathbf{X} = [\mathbf{X}_0 | X_{j+1}]$ avem că

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} \mathbf{X}_0^\top \mathbf{X}_0 & \mathbf{X}_0^\top X_{j+1} \\ X_{j+1}^\top \mathbf{X}_0 & X_{j+1}^\top X_{j+1} \end{pmatrix}$$

și folosind formula de inversare a matricelor de tip bloc (a se vedea demonstrația Propoziției 4.26) găsim că

$$[(\mathbf{X}^\top \mathbf{X})^{-1}]_{j+1,j+1} = (X_{j+1}^\top X_{j+1} - X_{j+1}^\top \mathbf{X}_0 (\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top X_{j+1})^{-1} = (X_{j+1}^\top (I - P_{\mathbf{X}_0}) X_{j+1})^{-1}.$$

Se observă astfel că acest test este echivalent cu testul bazat pe statistică de test repartizată student cu $n - (p + 1)$ grade de libertate (mai exact $F^2 = T_j$)

$$T_j = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{j+1,j+1}}}.$$

Regiunea critică a testului este

$$C = \left\{ \omega \mid T_j(\omega) > t_{n-(p+1)} \left(1 - \frac{\alpha}{2} \right) \right\}.$$

Aceasta este forma sub care testul de semnificativitate pentru un coeficient apare în limbajul R. Mai mult, cu o mică modificare a statisticii de test se pot testa ipotezele un pic mai generale

$$H_0 : \beta_j = r_j \quad \text{versus} \quad H_1 : \beta_j \neq r_j$$

folosind $T_j = \frac{\hat{\beta}_j - r_j}{\hat{\sigma}_{\hat{\beta}_j}}$.

Exemplu: prețul chiriilor în München - test asupra coeficientilor

Exp. 4.46 În contextul modelului de regresie liniară prezentat în Exemplul 4.42 putem testa semnificativitatea coeficienților modelului. Tabelul de mai jos prezintă rezultatele obținute în mod standard atunci când se folosește funcția `lm()` din R:

Variabila	Coeficientul	Eroarea standard	t-statistica	p-valoarea
<i>intercept</i>	6.932	0.038	184.539	< 0.001
<i>suprafata_inv_cen</i>	123.77	4.429	27.945	< 0.001
<i>an_co_1</i>	49.373	2.026	24.373	< 0.001
<i>an_co_2</i>	29.5	2.025	14.567	< 0.001
<i>an_co_3</i>	-0.884	1.972	-0.448	0.654
<i>nbucatarie</i>	1.043	0.102	10.279	< 0.001
<i>pbucatarie</i>	1.302	0.152	8.552	< 0.001
<i>an01</i>	-0.185	0.062	-2.981	0.003

În tabel se regăsesc numele variabilelor care apar în model, valorile estimate ale coeficienților $\hat{\beta}_j$, erorile standard $\hat{\sigma}_{\hat{\beta}_j}$, valorile statisticilor de test T_j pentru $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ și p-valorile corespunzătoare (pragul minim de semnificație α pentru care respingem ipoteza nulă).

Observăm că scăderea de preț, în medie, cu 0.185 euro pentru anul 2001 este semnificativă din punct de vedere statistic comparativ cu anul 1999.

4.6.4.3 Testul lui Fisher global Testul global al lui Fisher răspunde la următoarea întrebare: este vreunul dintre predictori folositor în explicarea răspunsului?

Prop. 4.47



O procedură de testare pentru ipotezele

$$H_0 : \beta_1 = \dots = \beta_p = 0 \quad \text{versus} \quad H_1 : \exists j \in \{1, \dots, p\} \beta_j \neq 0$$

are la bază statistica de test

$$F = \frac{n - (p + 1)}{p} \frac{\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|^2}{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2} = \frac{n - (p + 1)}{p} \frac{R^2}{1 - R^2} \sim F_{p, n-(p+1)}.$$

Vrem să testăm dacă variabilele explicative au sau nu influență asupra variabilei răspuns, cu alte cuvinte vrem să testăm dacă toți coeficienții sunt nuli exceptând constanta (β_0) - acesta este testul care apare ca rezultat atunci când efectuăm sumarul modelului liniar în R (comanda `summary(model)`). Suntem în cazul particular în care $\hat{\mathbf{Y}}_0 = \bar{y}\mathbf{1}$ și obținem din Propoziția 4.43 statistică de test a lui Fisher (*testul global al lui Fisher*)

$$F = \frac{n - (p + 1)}{p} \frac{\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|^2}{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2} \sim F_{p, n-(p+1)}$$

sau exprimată în termeni de coeficient de determinare

$$F = \frac{n - (p + 1)}{p} \frac{R^2}{1 - R^2} \sim F_{p, n-(p+1)}.$$

4.6.4.4 Legătura cu testul bazat pe raportul de verosimilități În rezultatul următor facem legătura dintre testul propus în Propoziția 4.43 pentru modele imbricate și testul bazat pe raportul de verosimilități.



Vrem să testăm ipotezele

Prop. 4.48

$$H_0 : \beta_{p_0+1} = \dots = \beta_p = 0 \quad \text{versus} \quad H_1 : \exists j \in \{p_0 + 1, \dots, p\} \beta_j \neq 0$$

cu ajutorul testului bazat pe raportul de verosimilități. Arătați că acest test este echivalent cu testul F al lui Fisher.

Testul bazat pe raportul de verosimilitate este

$$\Lambda(\mathbf{y}) = \frac{\sup_{\theta \in \Theta_0} L(\theta | \mathbf{y})}{\sup_{\theta \in \Theta} L(\theta | \mathbf{y})},$$

unde Θ este spațiul parametrilor modelului, Θ_0 este spațiul parametrilor corespunzător ipotezei nule iar $L(\theta | \mathbf{x})$ este funcția de verosimilitate.

Observăm că spațiul parametrilor modelului este

$$\Theta = \{(\beta, \sigma^2) | \beta \in \mathbb{R}^{p+1}, \sigma^2 \in (0, \infty)\},$$

cel corespunzător ipotezei nule este

$$\Theta_0 = \{(\beta_0, \sigma^2) | \beta_0 \in \mathbb{R}^{p_0}, \sigma^2 \in (0, \infty)\}$$

iar funcția de verosimilitate este

$$L(\beta, \sigma^2; \mathbf{Y}) = \prod_{i=1}^n f_Y(y_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \beta)^2} = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\beta\|^2}.$$

Observăm că

$$\sup_{\theta \in \Theta} L(\theta | \mathbf{y}) = \sup_{\beta, \sigma^2} L(\beta, \sigma^2; \mathbf{Y}) = L(\hat{\beta}_{VM}, \hat{\sigma}_{VM}^2; \mathbf{Y})$$

unde $\hat{\beta}_{VM} = \hat{\beta}$ este estimatorul obținut prin metoda celor mai mici pătrate iar $\hat{\sigma}_{VM}^2 = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2}{n} = \frac{RSS}{n}$. Astfel găsim

$$\sup_{\theta \in \Theta} L(\theta | \mathbf{y}) = \sup_{\beta, \sigma^2} L(\beta, \sigma^2; \mathbf{Y}) = \left(\frac{n}{2\pi \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2} \right)^{\frac{n}{2}} e^{-\frac{n}{2}} = \left(\frac{n}{2\pi RSS} \right)^{\frac{n}{2}} e^{-\frac{n}{2}}.$$

Sub H_0 avem

$$\sup_{\theta \in \Theta_0} L(\theta | \mathbf{y}) = \sup_{\beta_0, \sigma^2 \in \Theta_0} L(\beta_0, \sigma^2; \mathbf{Y}) = L(\hat{\beta}_0, \hat{\sigma}_0^2; \mathbf{Y}) = \left(\frac{n}{2\pi RSS_0} \right)^{\frac{n}{2}} e^{-\frac{n}{2}}$$

unde $RSS_0 = \|\mathbf{Y} - \mathbf{X}_0\hat{\beta}_0\|^2$ iar $\hat{\sigma}_0^2 = \frac{\|\mathbf{Y} - \mathbf{X}_0\hat{\beta}_0\|^2}{n} = \frac{RSS_0}{n}$.

Prin urmare raportul de verosimilitate devine

$$\Lambda(\mathbf{y}) = \frac{\sup_{\theta \in \Theta_0} L(\theta | \mathbf{y})}{\sup_{\theta \in \Theta} L(\theta | \mathbf{y})} = \frac{\left(\frac{n}{2\pi RSS_0} \right)^{\frac{n}{2}} e^{-\frac{n}{2}}}{\left(\frac{n}{2\pi RSS} \right)^{\frac{n}{2}} e^{-\frac{n}{2}}} = \left(\frac{RSS_0}{RSS} \right)^{-\frac{n}{2}}$$

iar regiunea critică a testului se scrie (testul respinge ipoteza nulă atunci când statistica Λ este suficient de mică)

$$C = \{\mathbf{Y} \in \mathbb{R}^n \mid \Lambda(\mathbf{Y}) < \lambda_0\}.$$

Observăm că pentru $\lambda > 0$, funcția $h(\lambda) = \lambda^{-\frac{2}{n}} - 1$ este descrescătoare, adică $\lambda < \lambda_0$ implică $h(\lambda) > h(\lambda_0)$. Avem

$$h(\lambda) > h(\lambda_0) \iff \frac{RSS_0 - RSS}{RSS} > h(\lambda_0) \iff \frac{n - (p + 1)}{p - p_0} \frac{RSS_0 - RSS}{RSS} > f_0$$

unde f_0 se determină din condiția

$$\mathbb{P}_{H_0} \left(\frac{n - (p + 1)}{p - p_0} \frac{RSS_0 - RSS}{RSS} > f_0 \right) = \alpha$$

și cum $\frac{n - (p + 1)}{p - p_0} \frac{RSS_0 - RSS}{RSS} \sim F_{(p+1)-p_0, n-(p+1)}$ conform Propoziției 4.43 deducem că $f_0 = f_{p-p_0, n-(p+1)}(1-\alpha)$. Am găsit că regiunea critică a testului bazat pe raportul de verosimilitate este identică cu cea a testului F a lui Fisher (a se vedea Propoziția 4.43).

4.6.4.5 Testarea ipotezei liniare generalizate În această secțiune prezentăm un test statistic care generalizează testele menționate în secțiunile anterioare. Toate testele prezentate până acum pot fi văzute ca un caz particular al ipotezei liniare generale

$$H_0 : R\beta = r \quad \text{versus} \quad H_1 : R\beta \neq r$$

unde R este o matrice dimensiune $q \times (p + 1)$ cu $\text{rang}(R) = q \leq p + 1$ iar r este un vector de dimensiune q .

De exemplu, pentru testarea ipotezei din modelele imbricate, $H_0 : \beta_{p_0+1} = \dots = \beta_p = 0$ versus $H_1 : \exists j \in \{p_0 + 1, \dots, p\} \beta_j \neq 0$ putem alege

$$R = \begin{pmatrix} 0 & I_q \end{pmatrix} \quad \text{și} \quad r = 0_q$$

iar pentru testarea semnificativității coeficientului β_j alegem $R = (0, \dots, 0, 1, 0, \dots, 0)^\top$ unde 1 se află pe poziția $j + 1$ și $r = 0$ scalar. Mai mult, pentru testarea egalității a doi coeficienți $\beta_i = \beta_j$ alegem

$$R = (0, \dots, 0, \underbrace{1}_{i+1}, 0, \dots, 0, \underbrace{-1}_{j+1}, 0, \dots, 0).$$

Impunând restricția $R\beta = r$ revine la a impune q (rangul matricei R) restricții liniare asupra parametrilor ceea ce implică faptul că răspunsul mediu $\mathbb{E}[\mathbf{Y}]$ nu mai aparține spațiului generat de coloanele matricei de design, i.e. $\mathbb{E}[\mathbf{Y}] \notin \mathcal{M}(X)$, ci unui subspațiu generat de coloanele matricei \mathbf{X} care satisfac restricția $R\beta = 0$.

Def. 4.49



Se numește ipoteză liniară generală asociată modelului de regresie liniară multiplă, ipoteza H_0 de forma $R\beta = r$, unde R este o matrice de dimensiune $q \times (p+1)$ și de rang q iar r este un vector de dimensiune q .

Avem următorul rezultat a cărui demonstrație se poate consulta spre exemplu în [Rencher and Schaalje, 2008] sau [Sen and Srivastava, 2012]:

Prop. 4.50



Fie modelul de regresie liniară $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ sub ipotezele \mathcal{H}_1 și \mathcal{H}'_2 . Ne propunem să testăm ipoteza liniară generală $H_0 : R\beta = r$, unde R are rangul q versus alternativa $H_1 : R\beta \neq r$. Fie \mathcal{M}_0 subspațiul lui $\mathcal{M}(X)$ de dimensiune $p+1-q$ generat de restricția $R\beta = r$ (sub H_0) și $\mathcal{M}(X)$ subspațiul de dimensiune $p+1$ asociat lui H_1 .

Pentru a testa cele două ipoteze vom folosi statistica de test F care sub ipoteza nulă verifică

$$\begin{aligned} F &= \frac{\dim(\mathcal{M}(X)^\perp)}{\dim(\mathcal{M}_0^\perp \cap \mathcal{M}(X))} \frac{\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|^2}{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2} = \frac{n - (p+1)}{q} \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}_0\|^2 - \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2}{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2} \\ &= \frac{n - (p+1)}{q} \frac{RSS_0 - RSS}{RSS} \sim F_{q, n-(p+1)} \end{aligned}$$

În plus avem că $\hat{\mathbf{Y}}_0$ este

$$\mathbf{X}\hat{\beta}_0 = \mathbf{X}\hat{\beta} + \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}R^\top [R(\mathbf{X}^\top \mathbf{X})^{-1}R^\top]^{-1}(r - R\hat{\beta}).$$

Anexe

Repartiții derive din repartiția normală

În afară de repartiția normală, următoarele trei repartiții sunt des utilizate în inferență statistică a modelului clasic de regresie liniară: repartiția χ^2 , repartiția Student t și repartiția Fisher-Snedecor F .

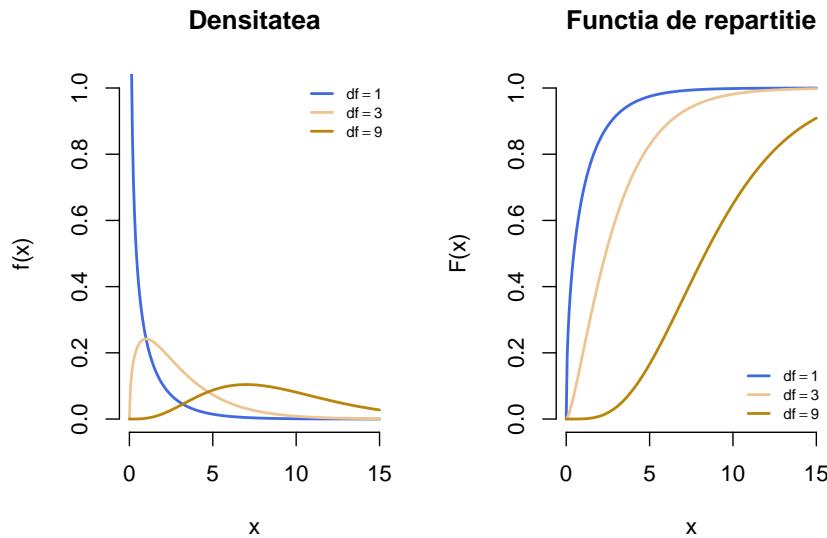
Repartiția χ^2

Fie X_1, \dots, X_n variabile aleatoare i.i.d. repartizate $\mathcal{N}(0, 1)$. Repartiția variabilei aleatoare $X = \sum_{i=1}^n X_i^2$ se numește repartiția χ^2 (Hi-pătrat) cu n grade de libertate și se notează cu $X \sim \chi_n^2$. Densitatea repartiției este

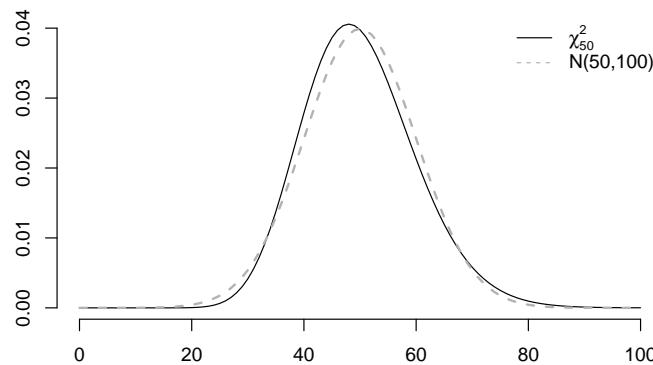
$$f(x) = \frac{1}{2^{n/2}\Gamma(n)} x^{n/2-1} e^{-x/2} \mathbf{1}_{\{y>0\}}$$

unde $\Gamma(\cdot)$ este funcția Gamma dată de $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du, x > 0$.

Pentru o v.a. $X \sim \chi_n^2$ avem că $E[X] = n$ și $Var(X) = 2n$.



Din Teorema Limită Centrală avem că pentru n suficient de mare, $X \approx \mathcal{N}(n, 2n)$ ceea ce sugerează că aproximativ 95% dintre valori se situează în intervalul $[n - 2\sqrt{2n}, n + 2\sqrt{2n}]$.

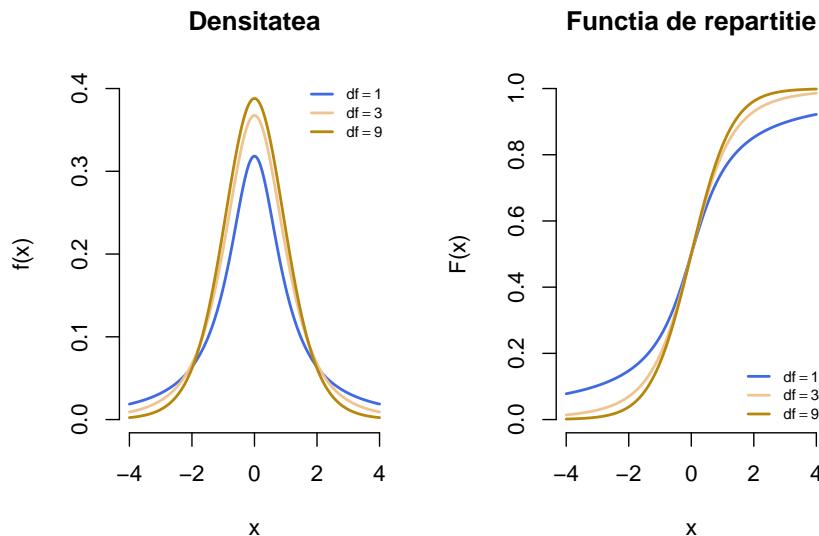


Repartiția t-Student

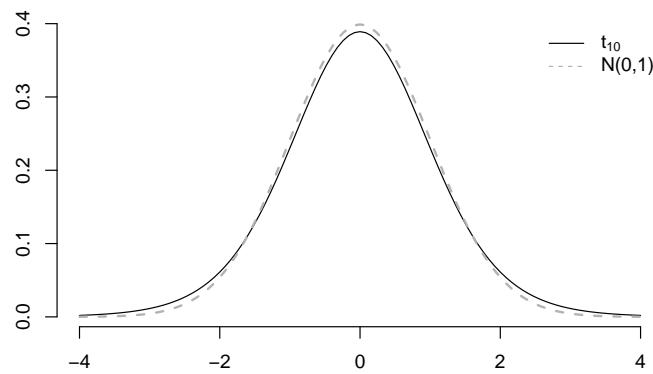
Fie U o variabilă aleatoare repartizată $\mathcal{N}(0, 1)$ și V o variabilă repartizată χ_n^2 , cu U și V independente. Repartiția variabilei aleatoare $T = \frac{U}{\sqrt{\frac{V}{n}}}$ se numește repartiția Student cu n grade de libertate și se notează cu $T \sim t_n$. Densitatea repartiției t_n este

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n} \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, x \in \mathbb{R}$$

Dacă $n = 1$ atunci variabila T este repartizată Cauchy (raport de două normale independente) și prin urmare nu are medie (evident nici varianță). Dacă $n = 2$ atunci T este de medie 0 dar de varianță infinită iar pentru $n \geq 3$, $\mathbb{E}[T] = 0$ și $Var(T) = \frac{n}{n-2}$.



Pentru n suficient de mare se poate arăta că $T \approx \mathcal{N}(0, 1)$ (de exemplu observând că, din Legea Numerelor Mari, numitorul tinde la 1 atunci când $n \rightarrow \infty$).

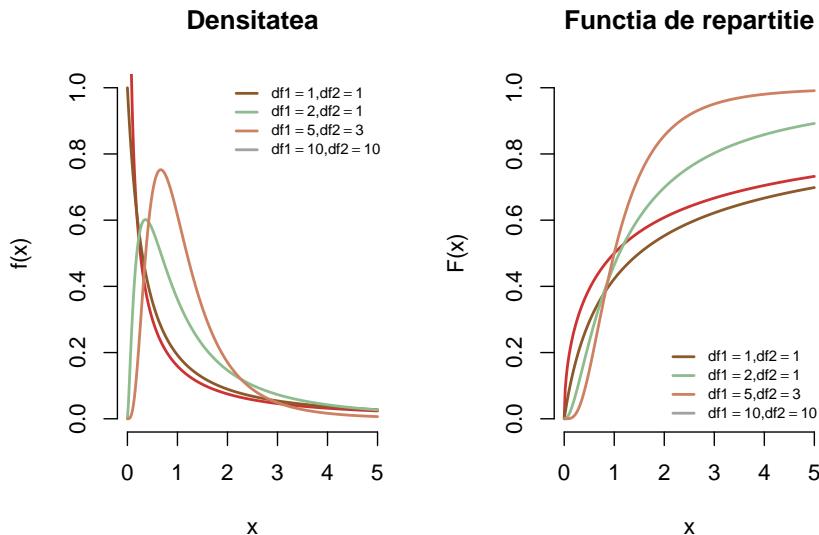


Repartiția Fisher-Snedecor

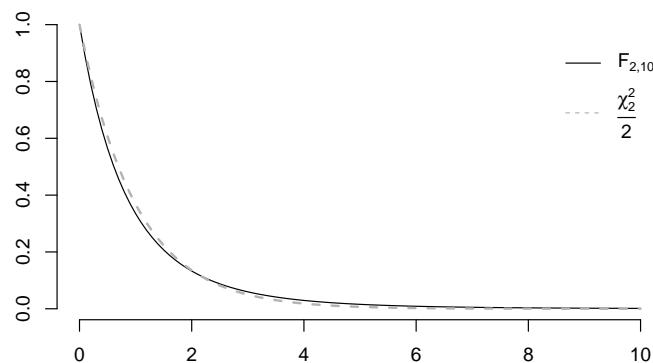
Fie U o variabilă aleatoare repartizată $\chi^2_{n_1}$ și V o variabilă aleatoare repartizată $\chi^2_{n_2}$, cu U și V independente. Repartiția variabilei aleatoare $F = \frac{U/n_1}{V/n_2}$ se numește repartiția Fisher-Snedecor cu n_1 grade de libertate la numărător și n_2 grade de libertate la numitor și se notează $F \sim F_{n_1, n_2}$. Densitatea de repartiție este

$$f_{n_1, n_2}(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{x^{\frac{n_1-2}{2}}}{\left(1 + \frac{n_1}{n_2}x\right)^{\frac{n_1+n_2}{2}}} \quad \text{dacă } x > 0 \quad (0 \text{ altfel})$$

Pentru $n_2 \geq 3$ media variabilei aleatoare F există și este egală cu $\frac{n_2}{n_2-2}$ iar pentru $n_2 \geq 5$ varianța există și este egală cu $\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$.



În plus dacă n_2 este mare atunci putem aproxima repartiția lui F cu $F \approx \frac{\chi_{n_1}^2}{n_1}$.



Intervale de încredere

Dacă în problema de estimare punctuală am găsit, plecând de la un eșantion generat dintr-o populație guvernată de un parametru θ , o valoare cât mai apropiată de parametrul necunoscut, în această secțiune ne propunem să determinăm un interval de valori care să aibă o probabilitate mare să conțină parametrul real.

Este clar că între estimarea punctuală și cea prin intervale de încredere va fi o strânsă legătură, cea de-a doua abordare bazându-se pe prima. Avem următoarea definiție

Def. 4.51



Fie X_1, \dots, X_n un eșantion de talie n dintr-o populație $F_\theta = \mathbb{P} \circ X^{-1}$, $\theta \in \Theta$ cu $X_i : \Omega \rightarrow (S, \mathcal{S})$. Fie $\alpha \in (0, 1)$ și funcțiile măsurabile $A_\alpha, B_\alpha : S^n \rightarrow \mathbb{R}$ care verifică

$$A_\alpha(x_1, \dots, x_n) \leq B_\alpha(x_1, \dots, x_n), \quad \forall (x_1, \dots, x_n) \in S^n.$$

Intervalul aleator $[A_\alpha(X_1, \dots, X_n), B_\alpha(X_1, \dots, X_n)]$ se numește interval de estimare pentru θ cu coeficientul de încredere $1 - \alpha$ dacă

$$\mathbb{P}_\theta([A_\alpha(X_1, \dots, X_n), B_\alpha(X_1, \dots, X_n)] \ni \theta) \geq 1 - \alpha, \quad \forall \theta \in \Theta.$$

Este important de notat că în definiția anterioară intervalul $[A_\alpha(X_1, \dots, X_n), B_\alpha(X_1, \dots, X_n)]$ este un interval aleator și **nu** parametrul θ este aleator.

Def. 4.52



Având dat un interval de estimare $[A_\alpha(X_1, \dots, X_n), B_\alpha(X_1, \dots, X_n)]$ pentru θ se numește *probabilitatea de acoperire* a acestuia

$$\mathbb{P}_\theta([A_\alpha(X_1, \dots, X_n), B_\alpha(X_1, \dots, X_n)] \ni \theta)$$

iar *coeficientul de încredere* al intervalului este definit prin

$$\inf_{\theta} \mathbb{P}_\theta([A_\alpha(X_1, \dots, X_n), B_\alpha(X_1, \dots, X_n)] \ni \theta) = 1 - \alpha$$

Trebuie menționat că atunci când avem observațiile x_1, \dots, x_n , intervalul de estimare realizat se numește *interval de încredere* de nivel de încredere $1 - \alpha$ și se notează cu

$$IC^{1-\alpha}(\theta) = [A_\alpha(x_1, \dots, x_n), B_\alpha(x_1, \dots, x_n)].$$

Chiar dacă noțiunea de *interval de încredere* și *interval de estimare* va fi folosită alternativ (prin abuz de limbaj) dar este important de notat că pentru un interval de încredere nu are sens să vorbim de $\mathbb{P}_\theta(IC^{1-\alpha}(\theta) \ni \theta)$ deoarece această probabilitate este 0 sau 1 după cum parametrul $\theta \in IC^{1-\alpha}(\theta)$.

Ca prim exemplu să presupunem că X_1, \dots, X_n este un eșantion de talie n dintr-o populație $\mathcal{N}(\theta, 1)$. Am văzut că $\bar{X}_n \sim \mathcal{N}(\theta, 1/n)$ sau încă $\frac{\bar{X}_n - \theta}{\sqrt{\frac{1}{n}}} \sim \mathcal{N}(0, 1)$. Fie z_1, z_2 cuantile ale repartiției normale standard pentru care are loc relația

$$\mathbb{P}_\theta \left(z_1 \leq \frac{\bar{X}_n - \theta}{\sqrt{\frac{1}{n}}} \leq z_2 \right) = \mathbb{P}_\theta \left(\bar{X}_n - \frac{z_2}{\sqrt{n}} \leq \theta \leq \bar{X}_n - \frac{z_1}{\sqrt{n}} \right) = 1 - \alpha$$

Atunci, pentru eșantionul realizat x_1, \dots, x_n , intervalul $IC^{1-\alpha}(\theta) = \left[\bar{x}_n - \frac{z_2}{\sqrt{n}}, \bar{x}_n - \frac{z_1}{\sqrt{n}} \right]$ este un interval de încredere de nivel de încredere $1 - \alpha$ pentru θ a cărui lungime este $l(IC^{1-\alpha}(\theta)) = \frac{1}{\sqrt{n}}(z_2 - z_1)$. Observăm că putem avea o infinitate de astfel de intervale dar dorim să determinăm acel interval care are lungimea cea mai mică. Dacă facem ipoteza suplimentară că $z_2 = z_2(z_1)$ (o funcție de z_1) atunci suntem în cadrul unei probleme de optimizare, mai precis avem problema (cu notațiile standard pentru funcția de repartitie și densitatea repartiției normale standard)

$$\begin{cases} \min \frac{1}{\sqrt{n}}(z_2 - z_1) \\ \Phi(z_2) - \Phi(z_1) = 1 - \alpha \end{cases}$$

care, prin derivare, conduce la relațiile $\phi(z_2) \frac{d}{dz_1} z_2 - \phi(z_1) = 0$ și $\frac{d}{dz_1} z_2 - 1 = 0$ ceea ce implică $\phi(z_2) = \phi(z_1)$. Din simetria repartiției normale deducem că $z_2 = z_{1-\alpha/2}$ și $z_1 = z_{\alpha/2} = -z_{1-\alpha/2}$ unde z_α este cuantila de ordin α a repartiției $\mathcal{N}(0, 1)$. Am obținut astfel intervalul de încredere pentru θ

$$IC^{1-\alpha}(\theta) = \left[\bar{x}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}}, \bar{x}_n + \frac{z_{1-\alpha/2}}{\sqrt{n}} \right]$$

Metoda pivotului de determinare a intervalelor de încredere

O primă metodă de determinare a intervalelor de încredere este *metoda pivotului*.

Def. 4.53



Fie X_1, \dots, X_n un eșantion de talie n dintr-o populație $F_\theta = \mathbb{P} \circ X^{-1}$, $\theta \in \Theta$. O funcție $g(x_1, \dots, x_n, \theta) : S^n \times \Omega \rightarrow \mathbb{R}$ se numește funcție pivot dacă verifică următoarele proprietăți:

- a) repartitia lui $g(X_1, \dots, X_n, \theta)$ nu depinde de θ
- b) pentru orice valori reale $u_1 \leq u_2$ și orice $(x_1, \dots, x_n) \in S^n$ inecuația $u_1 \leq g(x_1, \dots, x_n, \theta) \leq u_2$ se poate rezolva în θ (se poate pivota) conducând la o soluție de forma $a(x_1, \dots, x_n) \leq \theta \leq b(x_1, \dots, x_n)$.

Existența unei funcții pivot asigură o procedură de determinare a intervalelor de încredere de nivel de încredere dat.

Prop. 4.54



Fie X_1, \dots, X_n un eșantion de talie n dintr-o populație $F_\theta = \mathbb{P} \circ X^{-1}$, $\theta \in \Theta$ și $g : S^n \times \Omega \rightarrow \mathbb{R}$ o funcție care verifică:

- a) $g(x_1, \dots, x_n, \cdot)$ este continuă și strict monotonă ca funcție de θ
- b) $g(\cdot, \theta)$ este măsurabilă ca funcție de (x_1, \dots, x_n) pentru orice θ și variabila aleatoare $g(X_1, \dots, X_n, \theta)$ are o repartitie independentă de θ .

Atunci pentru orice $\alpha \in (0, 1)$ există un interval de încredere $IC^{1-\alpha}(\theta)$ de nivel $1 - \alpha$ pentru θ .

Pentru a exemplifica rezultatul anterior să considerăm că avem un eșantion $X_1, \dots, X_n \sim \mathcal{E}(\lambda)$. Cum variabila aleatoare $T_n = \sum_{i=1}^n X_i \sim \Gamma(n, \lambda)$, densitatea ei este

$$f_{T_n}(t) = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t} \mathbf{1}_{\{t \geq 0\}}$$

prin urmare variabila aleatoare λT_n are densitatea

$$f_{\lambda T_n}(t) = \frac{d}{dt} F_{\lambda T_n}(t) = \frac{d}{dt} \mathbb{P}(\lambda T_n \leq t) = \frac{d}{dt} F_{T_n}\left(\frac{t}{\lambda}\right) = \frac{1}{\lambda} f_{T_n}\left(\frac{t}{\lambda}\right) = \frac{1}{(n-1)!} t^{n-1} e^{-t} \mathbf{1}_{\{t \geq 0\}}$$

ceea ce implică $\lambda T_n \sim \Gamma(n, 1)$ și nu depinde de λ . Definim funcția

$$g(x_1, \dots, x_n, \lambda) = \lambda \sum_{i=1}^n x_i$$

care verifică proprietățile unei funcții pivot. Fie, de asemenea, u_1 și u_2 cuantilele de ordin $\alpha/2$ și respectiv $1 - \alpha/2$ a repartiției $\Gamma(n, 1)$ atunci

$$\mathbb{P}_\lambda(u_1 \leq g(X_1, \dots, X_n, \lambda) \leq u_2) = 1 - \alpha$$

determină intervalul de estimare (încredere) de nivel $1 - \alpha$ pentru λ

$$IC^{1-\alpha}(\theta) = \left[\frac{u_1}{\sum_{i=1}^n X_i}, \frac{u_2}{\sum_{i=1}^n X_i} \right].$$

Una dintre întrebările importante atunci când vorbim de intervale de încredere este cum alegem cuantilele u_1 și u_2 dat fiind că în general avem o infinitate de metode de alegere (ar trebui să verifice $F(u_2) - F(u_1) = 1 - \alpha$). Următorul rezultat răspunde parțial la această întrebare:

Def. 4.55



Spunem că o densitate de repartitie f este unimodală dacă există x^* astfel încât $f(x)$ este crescătoare pentru $x \leq x^*$ și descrescătoare pentru $x \geq x^*$.

Prop. 4.56



Fie $f(x)$ o densitate de repartitie unimodală. Dacă intervalul $[a, b]$ satisfacă:

- 1) $\int_a^b f(x) dx = 1 - \alpha$
- 2) $f(a) = f(b) > 0$
- 3) $a \leq x^* \leq b$ unde x^* este modul lui f

atunci $[a, b]$ este intervalul de lungime minimă care verifică $\int_a^b f(x) dx = 1 - \alpha$.

Metode asimptotice

În context general presupunem că avem un eșantion X_1, \dots, X_n un eșantion de talie n dintr-o populație $F_\theta = \mathbb{P} \circ X^{-1}$, $\theta \in \Theta$ și există un estimator T_n pentru θ care verifică o relație de tipul *Teoremei Limită Centrale*,

$$\frac{T_n - \theta}{s_n(\theta)} \xrightarrow{d} \mathcal{N}(0, 1)$$

unde $s_n(\cdot)$ este o funcție care depinde de θ , aleasă potrivit (de cele mai multe ori este chiar abaterea standard a lui T_n). Dacă funcția ($\ln \theta$) $\frac{T_n - \theta}{s_n(\theta)}$ se poate pivota atunci putem rezolva pentru a-l izola pe θ și astfel putem determina intervalul de încredere dorit. În caz contrar, vom folosi abordarea lui Wald (Wald plug-in) care ne spune că, în anumite condiții de regularitate, putem să-l înlocuim pe θ de la numitor cu estimatorul T_n , acesta fiind consistent (dat fiind relația de tip Teoremă Limită Centrală). Reamintim următorul rezultat, datorat lui Slutsky:

Prop. 4.57



Fie $(X_n)_n$ și $(Y_n)_n$ două șiruri de variabile aleatoare astfel încât $X_n \xrightarrow{d} X$ și $Y_n \xrightarrow{\mathbb{P}} a$, a este o constantă. Atunci

$$X_n + Y_n \xrightarrow[n \rightarrow \infty]{d} X + a \quad \text{și} \quad X_n Y_n \xrightarrow[n \rightarrow \infty]{d} aX$$

Dacă $s_n(\cdot)$ este continuă și $T_n \xrightarrow{\mathbb{P}} \theta$ atunci $s_n(T_n) \xrightarrow{\mathbb{P}} s_n(\theta)$ și din Teorema lui Slutsky avem

$$\frac{T_n - \theta}{s_n(T_n)} = \underbrace{\frac{T_n - \theta}{s_n(\theta)}}_{\xrightarrow{d} \mathcal{N}(0, 1)} \times \underbrace{\frac{s_n(\theta)}{s_n(T_n)}}_{\xrightarrow{\mathbb{P}} 1} \xrightarrow{d} \mathcal{N}(0, 1).$$

Astfel, pentru $\alpha \in (0, 1)$ fixat și $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$ și $z_{1-\frac{\alpha}{2}}$ cuantilele de ordin $\frac{\alpha}{2}$ și respectiv $1 - \frac{\alpha}{2}$ ale repartiției $\mathcal{N}(0, 1)$ avem

$$\mathbb{P}_\theta \left(z_{\frac{\alpha}{2}} \leq \frac{T_n - \theta}{s_n(T_n)} \leq z_{1-\frac{\alpha}{2}} \right) \approx 1 - \alpha$$

de unde prin pivotare obținem intervalul de încredere aproximativ

$$IC^{1-\alpha}(\theta) = [t_n - z_{1-\frac{\alpha}{2}} s_n(t_n), t_n + z_{1-\frac{\alpha}{2}} s_n(t_n)]$$

unde t_n este realizarea lui T_n . Intervalul de încredere de mai sus este aproximativ în sensul că procedura adoptată nu asigură cu exactitate un nivel $1 - \alpha$ pentru orice θ, n finit.

Pentru a exemplifica vom considera cazul unei populații Poisson: $X_1, \dots, X_n \sim \text{Pois}(\theta)$. Din *Teorema Limită Centrală* avem $(s_n(\theta) = \sqrt{\frac{\theta}{n}})$

$$\sqrt{n} \frac{\bar{X}_n - \theta}{\sqrt{\theta}} \xrightarrow{d} \mathcal{N}(0, 1)$$

de unde

$$\mathbb{P}_\theta \left(z_{\frac{\alpha}{2}} \leq \sqrt{n} \frac{\bar{X}_n - \theta}{\sqrt{\theta}} \leq z_{1-\frac{\alpha}{2}} \right) \approx 1 - \alpha.$$

Dubla inegalitate din probabilitatea de mai sus poate fi scrisă sub formă $n \frac{(\bar{X}_n - \theta)}{\theta} \leq z_{1-\frac{\alpha}{2}}^2$ ceea ce revine la a rezolva o inecuație de gradul doi. Soluția acestei inecuații conduce la intervalul de încredere aproximativ

$$\mathbb{P}_\theta \left(\bar{X}_n + \frac{z_{1-\frac{\alpha}{2}}^2}{n} - \sqrt{\frac{\bar{X}_n z_{1-\frac{\alpha}{2}}^2}{n} + \frac{z_{1-\frac{\alpha}{2}}^4}{4n^2}} \leq \theta \leq \bar{X}_n + \frac{z_{1-\frac{\alpha}{2}}^2}{n} + \sqrt{\frac{\bar{X}_n z_{1-\frac{\alpha}{2}}^2}{n} + \frac{z_{1-\frac{\alpha}{2}}^4}{4n^2}} \right) \approx 1 - \alpha.$$

Observăm că dacă în expresia de mai sus neglijăm termenii de ordin $\frac{1}{n^2}$ de sub radical și termenii de ordin $\frac{1}{n}$ din afara acestuia obținem intervalul de încredere de tip Wald:

$$\mathbb{P}_\theta \left(\bar{X}_n - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{X}_n}{n}} \leq \theta \leq \bar{X}_n + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{X}_n}{n}} \right) \approx 1 - \alpha.$$

Pentru a finaliza, mai rămâne întrebarea existenței estimatorului T_n care să verifice o expresie asimptotică precum cea de la începutul secțiunii. Am văzut că atunci când populația din care provine eșantionul îndeplinește o serie de condiții de regularitate, *estimatorul de verosimilitate maximă* $\hat{\theta}_n$ a lui θ satisfacă

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\frac{1}{n I_1(\theta)}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

unde $I_1(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f_\theta(X_1) \right)^2 \right]$ este Informația lui Fisher asociată unei observații. Cu excepția unor cazuri simple în care expresia lui $I_1(\theta)$ ne permite să pivotăm vom folosi abordarea lui Wald, înlocuind $I_1(\theta)$ cu $I_1(\hat{\theta}_n)$. Obținem astfel intervalul de încredere asimptotic pentru θ

$$IC^{1-\alpha}(\theta) = \left[\hat{\theta}_n - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{nI_1(\hat{\theta}_n)}}, \hat{\theta}_n + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{nI_1(\hat{\theta}_n)}} \right].$$

În secțiunile următoare vom exemplifica modul de construcție al unui interval de încredere în contextul unei populații normale.

Intervale de încredere pentru media unei populații normale atunci când varianța este cunoscută

Fie X_1, \dots, X_n un eșantion de talie n dintr-o populație normală $\mathcal{N}(\mu, \sigma^2)$ cu σ^2 cunoscut și μ necunoscut. Vrem să determinăm un interval de încredere de nivel $1 - \alpha$ pentru media populației normale. Am văzut că $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ ceea ce implică $\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \sim \mathcal{N}(0, 1)$.

Alegem funcția pivot $g(x_1, \dots, x_n, \mu) = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$ și pentru $\alpha \in (0, 1)$ fixat considerăm $z_{\frac{\alpha}{2}}$ și $z_{1-\frac{\alpha}{2}}$ cuantilele de ordin $\frac{\alpha}{2}$ și respectiv $1 - \frac{\alpha}{2}$ ale repartiției $\mathcal{N}(0, 1)$. Avem, ținând cont și de simetria normalei față de medie, că

$$\mathbb{P}_\mu \left(z_{\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \leq z_{1-\frac{\alpha}{2}} \right) = \mathbb{P}_\mu \left(\bar{X}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

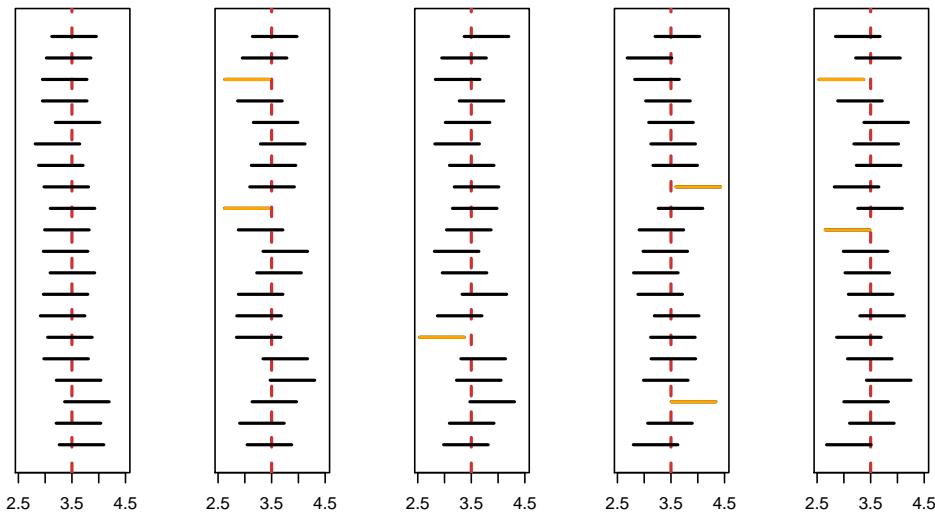
de unde concluzionăm că

$$IC^{1-\alpha}(\mu) = \left[\bar{x}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

este un interval de încredere de nivel $1 - \alpha$ pentru μ .

O ilustrare a 100 de intervalele de încredere de nivel de încredere $1 - \alpha = 0.95$ pentru media unei populații normale $\mathcal{N}(3.5, 1.5^2)$ atunci când $\sigma = 1.5$ este cunoscut este afișată mai jos:

100 intervale de încredere pentru μ (σ cunoscut)



Intervale de încredere pentru media unei populații normale atunci când varianța este necunoscută

Fie X_1, \dots, X_n un eșantion de talie n dintr-o populație normală $\mathcal{N}(\mu, \sigma^2)$ cu μ și σ^2 necunoscute. Am văzut că înlocuind σ^2 cu statistică $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ (varianța eșantionului) în $\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}$ obținem statistică $T_n = \sqrt{n} \frac{\bar{X}_n - \mu}{S_n}$ care este repartizată Student cu $n - 1$ grade de libertate, i.e.

$$T_n = \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \sim t_{n-1}.$$

Pentru funcția pivot $g(x_1, \dots, x_n, \mu) = \sqrt{n} \frac{\bar{X}_n - \mu}{S_n}$ și pentru $\alpha \in (0, 1)$ fixat, fie $t_{n-1, \frac{\alpha}{2}}$ și $t_{n-1, 1-\frac{\alpha}{2}}$ cuantilele de ordin $\frac{\alpha}{2}$ și respectiv $1 - \frac{\alpha}{2}$ ale repartiției t_{n-1} . Atunci

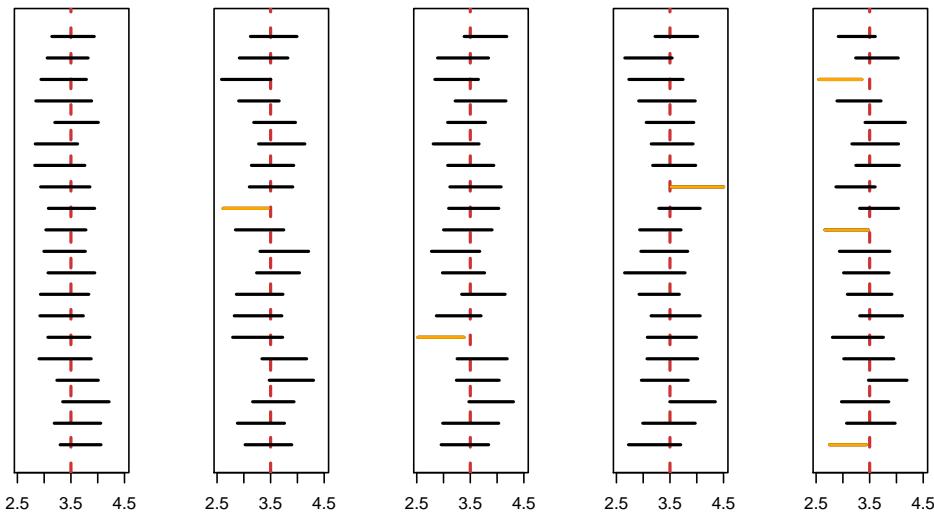
$$\mathbb{P}_\mu \left(t_{n-1, \frac{\alpha}{2}} \leq \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \leq t_{n-1, 1-\frac{\alpha}{2}} \right) = \mathbb{P}_\mu \left(\bar{X}_n - t_{n-1, 1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{n-1, 1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \right) = 1 - \alpha,$$

unde $t_{n-1, \frac{\alpha}{2}} = -t_{n-1, 1-\frac{\alpha}{2}}$ din simetria repartiție Student și găsim astfel că un interval de încredere de nivel $1 - \alpha$ pentru μ este

$$IC^{1-\alpha}(\mu) = \left[\bar{x}_n - t_{n-1, 1-\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1, 1-\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}} \right].$$

În figura de mai jos sunt prezentate 100 de intervale de încredere pentru $\mu = 3.5$ de nivel de încredere $1 - \alpha = 0.95$ atunci când σ^2 nu este cunoscut:

100 intervale de încredere pentru μ (σ necunoscut)



Intervale de încredere pentru varianța unei populații normale

Fie X_1, \dots, X_n un eşantion de talie n dintr-o populație normală $\mathcal{N}(\mu, \sigma^2)$ cu μ și σ^2 necunoscute. Până acum am dat intervale de încredere pentru media populației iar acum ne interesăm la un interval de încredere pentru varianța populației. Reamintim că varianța eşantionului S_n^2 este un estimator nedeplasat pentru σ^2 care pentru populații normale verifică $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$.

Pentru $\alpha \in (0, 1)$ fixat considerăm $\chi_{n-1, \frac{\alpha}{2}}^2$ și $\chi_{n-1, 1-\frac{\alpha}{2}}^2$ cuantilele de ordin $\frac{\alpha}{2}$ și respectiv $1 - \frac{\alpha}{2}$ ale repartiției χ_{n-1}^2 . Alegem ca funcție pivot $g(x_1, \dots, x_n, \sigma^2) = \frac{(n-1)S_n^2}{\sigma^2}$ și observăm că

$$\mathbb{P}_\mu \left(\chi_{n-1, \frac{\alpha}{2}}^2 \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi_{n-1, 1-\frac{\alpha}{2}}^2 \right) = \mathbb{P}_\mu \left(\frac{(n-1)S_n^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right) = 1 - \alpha,$$

de unde concluzionăm că un interval de încredere pentru σ^2 de nivel de încredere $1 - \alpha$ este

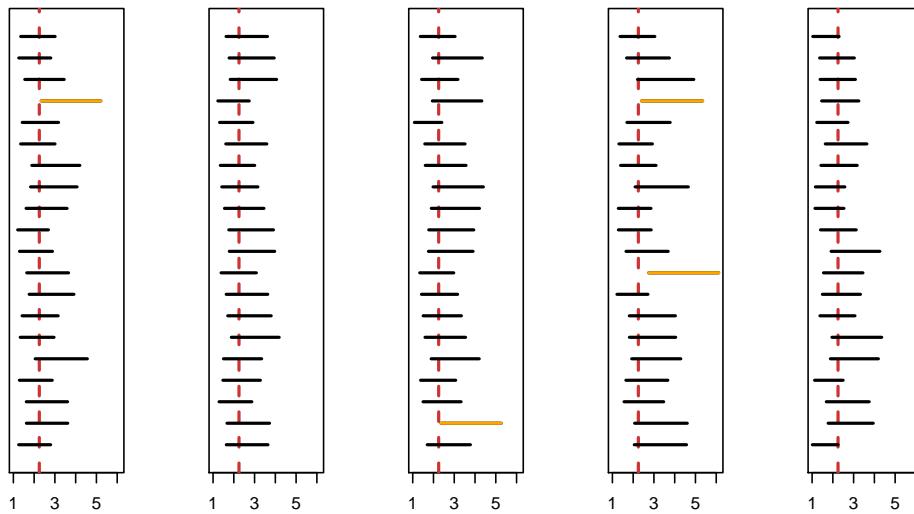
$$IC^{1-\alpha}(\sigma^2) = \left[\frac{(n-1)S_n^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}, \frac{(n-1)S_n^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right].$$

Este important de remarcat că în cazul în care media populației μ este cunoscută atunci $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \sim \chi_n^2$ ceea ce conduce la un interval de încredere de forma

$$IC^{1-\alpha}(\sigma^2) = \left[\frac{\sum_{i=1}^n (x_i - \mu)^2}{\chi_{n, 1-\frac{\alpha}{2}}^2}, \frac{\sum_{i=1}^n (x_i - \mu)^2}{\chi_{n, \frac{\alpha}{2}}^2} \right].$$

Figura de mai jos ilustrează 100 de intervale de încredere pentru σ^2 de nivel de încredere $1 - \alpha = 0.95$ atunci populația este normală și media este necunoscută.

100 intervale de încredere pentru σ^2 (μ necunoscut)



Intervale de încredere pentru diferența mediilor a două populații normale

Fie $X_1, \dots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma^2)$ și $Y_1, \dots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma^2)$ două eșantioane independente (între ele) de talie n_1 și respectiv n_2 din populații normale de medii diferite dar cu aceeași dispersie necunoscută σ^2 . Ne propunem să construim un interval de încredere de nivel $1 - \alpha$ pentru diferența mediilor $\mu_1 - \mu_2$. Ca și în cazul unei populații normale avem că $\bar{X}_{n_1} \sim \mathcal{N}\left(\mu_1, \frac{\sigma^2}{n_1}\right)$ și $\bar{Y}_{n_2} \sim \mathcal{N}\left(\mu_2, \frac{\sigma^2}{n_2}\right)$ prin urmare, ținând cont de independentă, $\bar{X}_{n_1} - \bar{Y}_{n_2} \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$. Din ultima relație deducem că

$$\frac{(\bar{X}_{n_1} - \bar{Y}_{n_2}) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0, 1).$$

Cum σ^2 este necunoscut îl vom estima plecând de la cele două eșantioane remarcând pentru început că $\frac{(n_1-1)S_1^2}{\sigma^2} \sim \chi_{n_1-1}^2$ și $\frac{(n_2-1)S_2^2}{\sigma^2} \sim \chi_{n_2-1}^2$, unde S_i^2 reprezintă varianța eșantionului $i = 1, 2$. Ținând cont de independentă dintre eșantioane avem că $S_1^2 \perp S_2^2$ prin urmare

$$\frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2} \sim \chi_{n_1+n_2-2}^2.$$

Folosind aceeași metodă ca și în cazul mediei unei populații normale cu μ și σ^2 necunoscute deducem că statistică

$$\frac{\frac{(\bar{X}_{n_1} - \bar{Y}_{n_2}) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2}}} \sim \frac{\mathcal{N}(0, 1)}{\sqrt{\frac{\chi_{n_1+n_2-2}^2}{n_1+n_2-2}}} \sim t_{n_1+n_2-2}.$$

Astfel, notând cu $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ estimatorul varianței σ^2 (pooled variance), putem re scrie

$$\frac{(\bar{X}_{n_1} - \bar{Y}_{n_2}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

ceea ce conduce la intervalul de încredere de nivel $1 - \alpha$ pentru diferența mediilor (similar cu cel din cazul mediei populației normale atunci când varianța este necunoscută)

$$IC^{1-\alpha}(\mu) = \left[(\bar{x}_{n_1} - \bar{y}_{n_2}) - t_{n_1+n_2-2, 1-\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\bar{x}_{n_1} - \bar{y}_{n_2}) + t_{n_1+n_2-2, 1-\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right].$$

Trebuie menționat că în cazul în care eșantioanele provin din populații cu varianțe diferite $\sigma_1^2 \neq \sigma_2^2$ atunci statistica $\frac{(\bar{X}_{n_1} - \bar{Y}_{n_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ nu mai este repartizată Student. În acest caz se folosește corecția lui Welch-Satterthwaite:

$$\frac{(\bar{X}_{n_1} - \bar{Y}_{n_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_\nu, \quad \text{cu} \quad \nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2-1}}.$$

Intervale de încredere pentru diferența mediilor a două populații normale atunci când datele vin în pereche

În cazul în care datele vin în perechi (de exemplu provin de la gemeni de sex diferit, de la indivizi care sunt potriviti după o serie de caracteristici sau de la aceiași indivizi înainte și după un eveniment de interes) atunci există dependență între acestea și nu mai putem aplica metoda anterioară. În acest context presupunem că $X_1, \dots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$ și $Y_1, \dots, Y_n \sim \mathcal{N}(\mu_2, \sigma_2^2)$ sunt două eșantioane cu date pereche și ne interesăm la determinarea unui interval de încredere pentru $d = \mu_1 - \mu_2$.

Fie $D_i = X_i - Y_i$ variabilele aleatoare independente cu $D_i \sim \mathcal{N}(d, s_d^2)$ unde $d = \mu_1 - \mu_2$ iar $s_d^2 = \sigma_1^2 + \sigma_2^2 - 2\text{Cov}(X_i, Y_i)$. Observăm, din normalitatea și independența variabilelor aleatoare D_i , că statistica

$$\sqrt{n} \frac{\bar{D}_n - d}{S_d} \sim t_{n-1}$$

unde $S_d^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D}_n)^2$ ceea ce conduce la intervalul de încredere de nivel de încredere $1 - \alpha$ pentru diferența mediilor

$$IC^{1-\alpha}(\mu) = \left[\bar{d}_n - t_{n-1, 1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}, \bar{d}_n + t_{n-1, 1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}} \right].$$

Intervale de încredere pentru raportul varianțelor a două populații normale

Fie $X_1, \dots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ și $Y_1, \dots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$ două eșantioane independente (între ele) de talie n_1 și respectiv n_2 din populații normale de medii și dispersii diferite. Ne propunem să construim un interval de încredere de nivel $1 - \alpha$ pentru raportul varianțelor $\frac{\sigma_1^2}{\sigma_2^2}$.

Cum $\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi_{n_1-1}^2$ și $\frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi_{n_2-1}^2$ unde S_1^2 și S_2^2 sunt varianțele primului și respectiv celui de-a doilea eșantion și ținând cont de independența dintre eșantioane, și în particular dintre S_1^2 și S_2^2 , găsim

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}}{\frac{(n_2-1)S_2^2}{\sigma_2^2}} \sim \frac{\chi_{n_1-1}^2}{\chi_{n_2-1}^2} \sim F_{n_1-1, n_2-1}.$$

Astfel, pentru $\alpha \in (0, 1)$, avem

$$\mathbb{P}_{\frac{\sigma_1^2}{\sigma_2^2}} \left(f_{n_1-1, n_2-1, \frac{\alpha}{2}} \leq \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \leq f_{n_1-1, n_2-1, 1-\frac{\alpha}{2}} \right) = 1 - \alpha,$$

unde $f_{n_1-1, n_2-1, \alpha}$ este cuantila de ordin α a repartiției Fisher-Snedecor F_{n_1-1, n_2-1} ceea ce conduce la intervalul de încredere de nivel $1 - \alpha$ pentru $\frac{\sigma_1^2}{\sigma_2^2}$

$$IC^{1-\alpha} \left(\frac{\sigma_1^2}{\sigma_2^2} \right) = \left[\frac{S_1^2}{S_2^2} f_{n_1-1, n_2-1, \frac{\alpha}{2}}, \frac{S_1^2}{S_2^2} f_{n_1-1, n_2-1, 1-\frac{\alpha}{2}} \right].$$

Intervale de încredere pentru o proporție

Sunt multe situațiile în care vrem să determinăm un interval de încredere pentru o proporție, de exemplu în cazul unui sondaj aleator dorim să determinăm un interval de încredere de nivel 95% pentru proporția p a persoanelor care verifică o anumită proprietate de interes (e.g. au votat cu partidul X). În acest context avem X_1, \dots, X_n un eșantion de talie n dintr-o populație Bernoulli $\mathcal{B}(p)$ cu p necunoscut. Din Teorema Limită Centrală știm că

$$\sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

ceea ce implică, pentru n suficient de mare (în practică ≥ 30), că

$$\mathbb{P}_\mu \left(-z_{1-\frac{\alpha}{2}} \leq \sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \leq z_{1-\frac{\alpha}{2}} \right) \approx 1 - \alpha.$$

Cantitatea $\sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}}$ se poate pivota în raport cu θ și rezolvând inecuația de ordin 2 în p

$$n \frac{(\bar{X}_n - p)^2}{p(1-p)} \leq z_{1-\frac{\alpha}{2}}^2$$

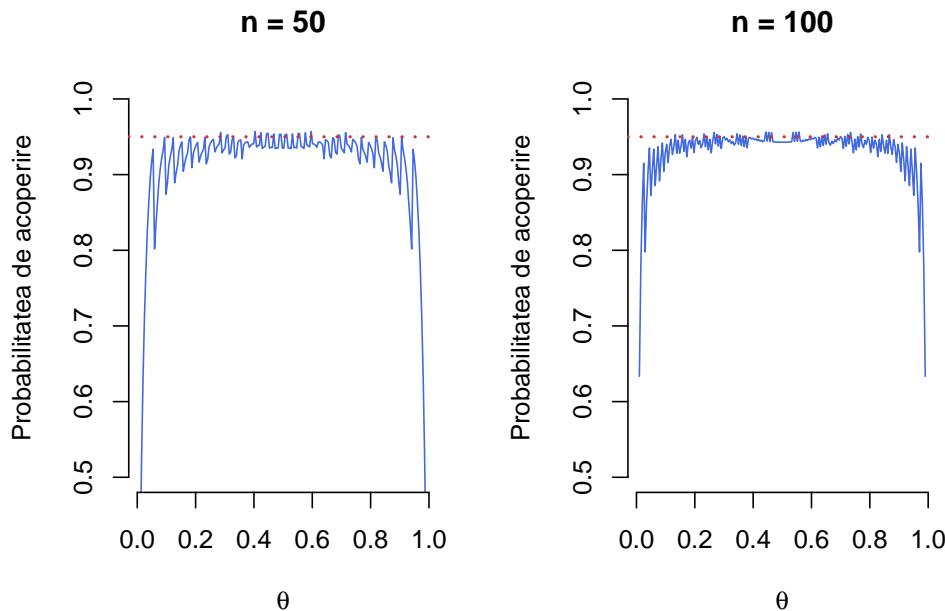
găsim un interval de încredere asimptotic de forma

$$IC^{1-\alpha}(\theta) = \left[\frac{\bar{X}_n + \frac{z_{1-\frac{\alpha}{2}}^2}{2n}}{1 + \frac{z_{1-\frac{\alpha}{2}}^2}{n}} - \frac{1}{1 + \frac{z_{1-\frac{\alpha}{2}}^2}{n}} \sqrt{\frac{z_{1-\frac{\alpha}{2}}^2}{n} \bar{X}_n (1 - \bar{X}_n) + \frac{z_{1-\frac{\alpha}{2}}^4}{4n^2}}, \frac{\bar{X}_n + \frac{z_{1-\frac{\alpha}{2}}^2}{2n}}{1 + \frac{z_{1-\frac{\alpha}{2}}^2}{n}} + \frac{1}{1 + \frac{z_{1-\frac{\alpha}{2}}^2}{n}} \sqrt{\frac{z_{1-\frac{\alpha}{2}}^2}{n} \bar{X}_n (1 - \bar{X}_n) + \frac{z_{1-\frac{\alpha}{2}}^4}{4n^2}} \right].$$

Eliminând termenii de ordin $\frac{1}{n^2}$ de sub radical și pe cei de ordin $\frac{1}{n}$ din afara radicalului găsim exact intervalul de încredere asimptotic pe care l-am fi obținut dacă aplicam metoda lui Wald:

$$IC^{1-\alpha}(\theta) = \left[\bar{X}_n - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right].$$

Probabilitatea de acoperire este:



Modele de regresie folosind limbajul R

În această secțiune vom prezenta funcțiile de bază din R folosite pentru fitarea (potrivirea) și analizarea unui model de regresie liniară. Pentru a rula modelul de regresie liniară clasic în R se folosește funcția `lm()` (*linear model*). Funcția `lm()` are două argumente esențiale: `formula` și `data`:

```
model_regresie <- lm(formula, data)
```

unde

Argument	Descriere
<code>formula</code>	O formulă de forma $y \sim x_1 + x_2 + \dots$, unde y este variabila răspuns (dependentă) iar x_1, x_2, \dots sunt variabilele explicative (covariabilele). Dacă vrem să includem toate coloanele (cu excepția lui y) ca variabile explicative putem folosi $y \sim .$
<code>data</code>	Este setul de date în format <code>data.frame</code> care conține coloanele specificate de formulă.

Următorul tabel conține o serie de corespondențe între codul R și conceptele statistice asociate modelului de regresie:

R	Concepțe statistice
<code>x1, x2, ..., xp</code>	Variabilele predictor x_1, \dots, x_p
<code>y</code>	Variabila răspuns

R	Concepțe statistice
<code>data <- data.frame(x1 = x1, ..., xp = xp, y = y)</code>	Eșantionul $(y_i, \mathbf{x}_{ij}), i = 1, \dots, n, j = 1, \dots, p$
<code>model <- lm(y ~ x, data = data)</code>	Modelul de regresie liniară
<code>model\$coefficients</code>	Coefficienții $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$
<code>model\$residuals</code>	Valorile reziduale $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$
<code>model\$fitted.values</code>	Valorile ajustate (fitate) $\hat{y}_1, \dots, \hat{y}_n$
<code>model\$df.residual</code>	Gradele de libertate $n - (p + 1)$
<code>summaryModel <- summary(model)</code>	Sumarul modelului de regresie liniară
<code>summaryModel\$sigma</code>	Estimatorul $\hat{\sigma}$
<code>summaryModel\$r.squared</code>	Coefficientul de determinare R^2
<code>summaryModel\$fstatistic</code>	Testul lui Fisher global F
<code>anova(model)</code>	Tabelul ANOVA

O componentă importantă în specificarea unui model de regresie liniară este dat de termenul `formula` în expresia funcției `lm()`. Forma de bază a unei *formule* este

$$y \sim x_1 + x_2 + \cdots + x_p$$

unde variabila răspuns y este separată de variabilele predictor prin \sim iar covariabilele sunt separate la rândul lor prin semnul $+$ și corespunde modelului clasic

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

Formula se poate modifica introducând alte simboluri pentru a permite specificarea mai multor tipuri de modele liniare (e.g. cu interacții), după cum este specificat în tabelul de mai jos:

Simbol	Mod de utilizare
\sim	Separă variabila răspuns (partea stângă) de variabilele predictor (partea dreaptă). De exemplu, un model de regresie liniară cu trei predictori x_1 , x_2 și x_3 se scrie sub forma $y \sim x_1 + x_2 + x_3$.
$+$	Separă variabilele predictor.
$:$	Se folosește pentru a specifica interacția dintre variabilele predictor. De exemplu un model de regresie liniară care consideră predicția lui y în raport cu x_1 , x_2 și interacția dintre x_1 și x_2 se scrie $y \sim x_1 + x_2 + x_1:x_2$.
$*$	O notație prescurtată pentru a descrie toate interacțiile dintre termeni, astfel codul $y \sim x_1*x_2*x_3$ se traduce prin modelul $y \sim x_1 + x_2 + x_3 + x_1:x_2 + x_1:x_3 + x_2:x_3 + x_1:x_2:x_3$.
$^$	Specifică interacțiile până la un anumit grad, de exemplu formula $y \sim (x_1 + x_2 + x_3)^2$ se traduce prin modelul $y \sim x_1 + x_2 + x_3 + x_1:x_2 + x_1:x_3 + x_2:x_3$.
$.$	Un simbol care permite o scriere prescurtată care să includă toate variabilele explicative din setul de date. De exemplu, dacă setul de date conține variabila răspuns y și variabilele predictor x_1 , x_2 și x_3 atunci codul $y \sim .$ se traduce prin $y \sim x_1 + x_2 + x_3$.
$-$	Semnul minus permite eliminarea din ecuație a unei variabile explicative. De exemplu, prin formula $y \sim (x_1 + x_2 + x_3)^2 - x_1:x_3$ se înțelege $y \sim x_1 + x_2 + x_3 + x_1:x_2 + x_2:x_3$.
-1	Acest termen elimină termenul liber (ordonata la origine - intercept) al modelului. De exemplu, în cazul modelului de regresie liniară simplă $y \sim x_1 - 1$ înțelegem că modelul forțează dreapta de regresie să treacă prin origine.

Simbol	Mod de utilizare
I()	Această funcție este folosită atunci când dorim interpretarea aritmetică a expresiei din paranteză. Pentru a exemplifica, să considerăm pentru început codul $y \sim x_1 + (x_2 + x_3)^2$ care se traduce prin $y \sim x_1 + x_2 + x_3 + x_2 \cdot x_3$. În contrast, prin formula $y \sim x_1 + I((x_2 + x_3)^2)$ se înțelege modelul $y \sim x_1 + x_4$ unde x_4 este variabila nou creată prin ridicarea la pătrat a sumei dintre x_2 și x_3 .
function	Se pot folosi și funcții uzuale în specificarea formulei, astfel modelul $\log(y) \sim x_1 + x_2 + x_3$ va folosi ca variabilă răspuns variabila $\log(y)$.

În tabelul de mai jos regăsiți o serie de funcții utile, pe lângă funcția `lm()`, necesare pentru analiza de regresie multiplă împreună cu pachetele din care fac parte.

Funcție	Pachet	Descriere
<code>summary()</code>	<code>stats</code>	Funcție care prezintă rezultate detaliate pentru modelul fitat (e.g. reziduuri, coeficienții modelului, sigma - $\hat{\sigma}$, gradele de libertate ale modelului, coeficientul de determinare, etc.)
<code>/</code>		Afișează parametrii/coeficienții modelului potrivit (fitat)
<code>summary.lm()</code>		Prezintă intervale de încredere de nivel de încredere 95% (default) pentru parametrii modelului
<code>coef()</code>		Afișează valorile ajustate \hat{y} pentru modelul considerat
<code>confint()</code>		Funcție care crează matricea de design pentru un model dat printr-o formulă sau extrage matricea de design dintr-un obiect <code>lm</code>
<code>fitted()</code>		Funcție care permite calcularea predicțiilor de noi observații pe baza unui model dat
<code>model.matrix()</code>		Afișează reziduurile unui model fitat
<code>predict()</code>		Afișează reziduurile standardizate ale unui model fitat
<code>residuals()</code>		Afișează reziduurile studentizate ale unui model fitat
<code>rstandard()</code>		Generează tabelul ANOVA pentru un model dat sau un tabel ANOVA pentru compararea a două sau mai multe modele
<code>rstudent()</code>		Afișează matricea de varianță covarianță a parametrilor modelului
<code>anova()</code>		Funcție generică care permite calcularea criteriului informațional al lui Akaike pentru unul sau mai multe modele fitate. Funcția are la bază formula $-2 * \log\text{-likelihood} + k * npar$ unde <code>npar</code> reprezintă numărul de parametrii din model, $k = 2$ pentru AIC sau $k = \log(n)$ pentru BIC - criteriu informațional Bayesian (sau SBC - Schwarz's Bayesian criterion)
<code>vcov()</code>		Funcție care permite extragerea formulei folosite în generarea unui model dat
<code>AIC()</code>		Funcție folosită pentru modificarea unui model, de obicei prin adăugarea/eliminarea unui termen.
<code>formula()</code>		Metodă a unui obiect <code>lm</code> care generează o serie de (șase) grafice de diagnostic pentru evaluarea modelului fitat. Se poate apela și prin <code>plot()</code> .
<code>update.formula()</code>		Funcție care permite selectarea unui model pe bază de AIC.
<code>plot.lm()</code>		Funcție similară cu <code>step()</code> din pachetul <code>stats</code> .
<code>stepAIC()</code>	<code>MASS</code>	Funcție care încercă potrivirea tuturor modelelor care diferă printr-un singur termen prin adăugire față de modelul considerat.
<code>addterm()</code>		Funcție similară cu <code>add1()</code> din pachetul <code>stats</code> .
<code>dropterm()</code>		Funcție care încercă potrivirea tuturor modelelor care diferă printr-un singur termen prin eliminare față de modelul considerat.
<code>leaps()</code>	<code>leaps</code>	Funcție similară cu <code>drop1()</code> din pachetul <code>stats</code> .
		Selectie de modele pe baza algoritmului <code>leaps and bounds</code> .

Funcție	Pachet	Descriere
<code>linearHypothesis()</code>	<code>car</code>	Funcție generică care permite testare ipotezelor liniare.
<code>Anova()</code>		Returnează tabelul ANOVA pentru modelul fitat.
<code>cookd()</code>		Calculează distanța lui Cook.
<code>outlierTest()</code>		Această funcție întoarce p-valorile ajustate Bonferroni pentru reziduurile studentizate.
<code>durbinWatsonTest()</code>		Testul Durbin-Watson pentru autocorelarea termenilor eroare.
<code>leveneTest()</code>		Calculează testul Levene pentru omogenitatea varianțelor între grupuri.
<code>ncvTest()</code>		Funcție care calculează un test de scor pentru testarea ipotezei de varianță constantă a erorilor versus o alternativă care presupune că varianța termenilor eroare se schimbă o dată cu modificarea valorilor ajustate sau cu o combinație liniară a predictorilor.

Referințe

- R. Dennis Cook and Sanford Weisberg. *Graphs in statistical analysis: Is the medium the message?* *The American Statistician*, 53(1):29–37, February 1999. ISSN 0003-1305. doi: 10.1080/00031305.1999.10474426. URL <https://www.tandfonline.com/doi/abs/10.1080/00031305.1999.10474426>. (Citat la pagina 37.)
- Julian Faraway. *Linear Models with R*. CRC Press, 2nd edition, 2015. ISBN 978-1-4398-8734-9. (Citat la paginile 31, 37 și 71.)
- William H. Greene. *Econometric Analysis*. Prentice Hall, 7 edition, 2011. ISBN 0131395386,9780131395381. URL <http://gen.lib.rus.ec/book/index.php?md5=A0CEC5F0E1D09DF6E99A0E52B9B60DBB>. (Citat la pagina 91.)
- H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *SIAM Review*, 23(1):53–60, 1981. doi: 10.1137/1023004. URL <https://epubs.siam.org/doi/abs/10.1137/1023004>. (Citat la pagina 78.)
- Jean Jacod and Philip Protter. *Probability essentials*. Springer, 2003. (Citat la paginile 89 și 90.)
- Mark E Johnson. *Multivariate statistical simulation: A guide to selecting and generating continuous multivariate distributions*. John Wiley & Sons, 2013. (Citat la pagina 39.)
- David J Olive. *Linear regression*. Springer, 2017. (Citat la pagina 45.)
- L. Ornea and A. Turtoi. *O introducere în geometrie*. Fundatia Theta, 2000. (Citat la paginile 64, 65 și 95.)
- Alvin C Rencher and G Bruce Schaalje. *Linear models in statistics*. John Wiley & Sons, 2008. (Citat la paginile 31, 71, 78, 89, 97, 101 și 106.)
- George Seber and Alan Lee. *Linear Regression Analysis*. Wiley, 2nd edition, 2003. ISBN 0-471-41540-5. (Citat la paginile 31 și 89.)
- Ashish Sen and Muni Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 2012. (Citat la paginile 91, 97 și 106.)
- Sanford Weisberg. *Applied Linear Regression*. Wiley, 4th edition, 2014. ISBN 978-1-118-38608-8. (Citat la pagina 31.)
- Haruo Yanai, Kei Takeuchi, and Yoshio Takane. *Projection matrices, generalized inverse matrices, and singular value decomposition*. Springer-Verlag New York, 2011. ISBN 978-1-4419-9886-6. doi: <https://doi.org/10.1007/978-1-4419-9887-3>. (Citat la pagina 65.)