

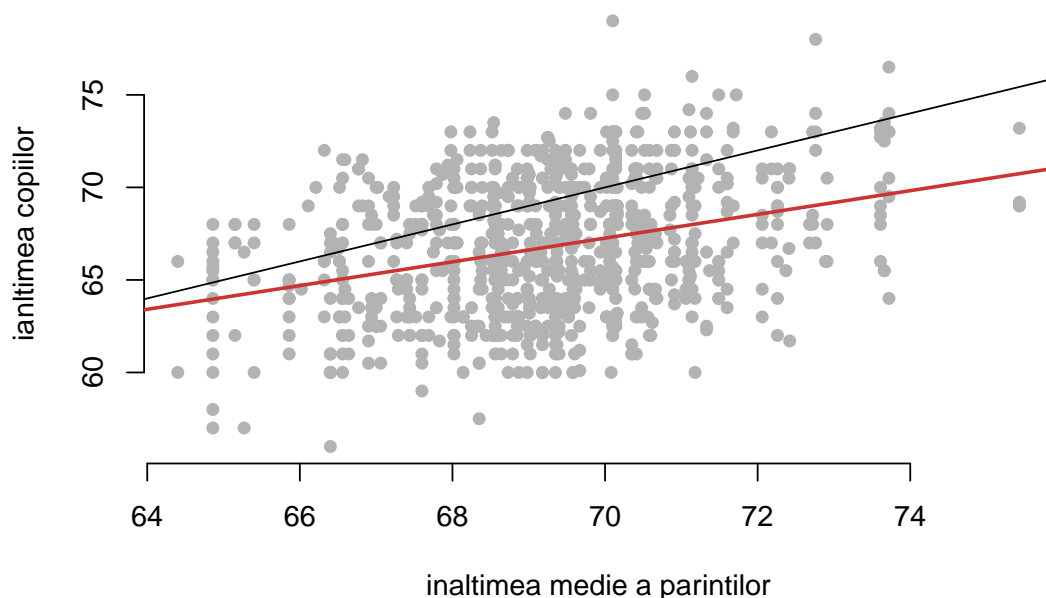
Introducere în modele de regresie

Regresie liniară simplă

Obiectivul acestor note este de a prezenta câteva aspecte ce țin de regresia liniară simplă.

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 vârstă 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* $x = \text{“înălțimea medie a părinților”}$ influențează *variabila răspuns* $y =$

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

“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

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

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 \mathcal{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 rescrie 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 parcelează și pentru fiecare parcelă inginerul atribuie o anumită cantitate de îngrășământ prestabilită: 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 vârsta populației (variabila explicativă). Astfel, unui individ ales la întâmplare din grupul țintă îi corespunde un cuplu de variabile aleatoare (X, Y) unde X este vârsta 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 valoarea lui $X = x$, $\mathbb{E}[Y|X = x] = f(x)$.

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

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

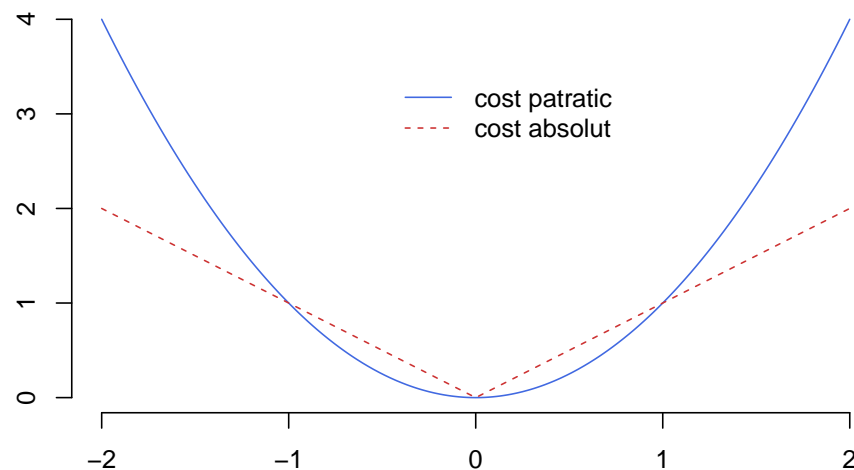
$$y = \mathbb{E}[y|x_1, \dots, x_k] + \varepsilon = f(x_1, \dots, x_k) + \varepsilon.$$

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 predictor 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 adevărata 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

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

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

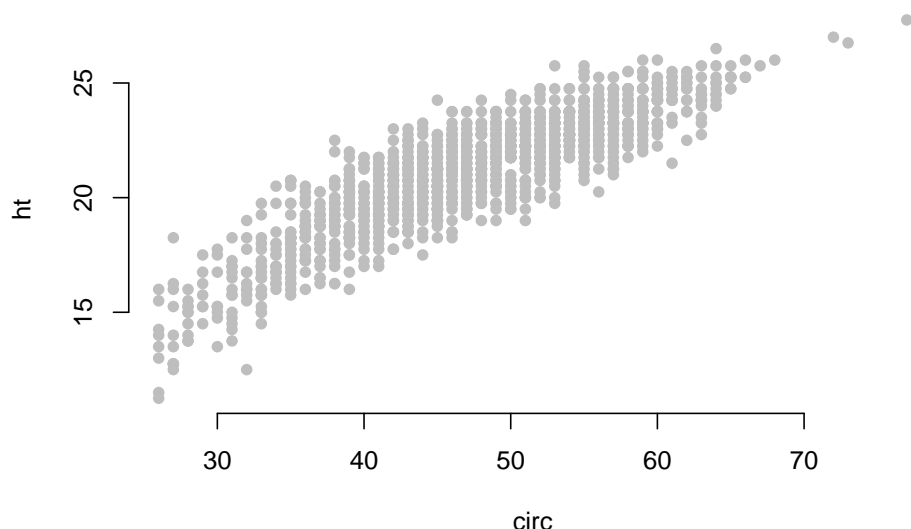
Setul de date **Eucalypt** (care poate fi descărcat de [aici](#)) face referire la înălțimea și circumferința (măsurată la 1m 30cm de sol) a 1429 arbori de eucalipt plantați într-o regiune experimentală din Franța. Cele două caracteristici sunt măsurate la vârsta de maturitate a arborilor, anume la 6 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 dat în figura de mai jos:



2.2 Prețul chiriei locuințelor în Munchen

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 Munchen, 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:

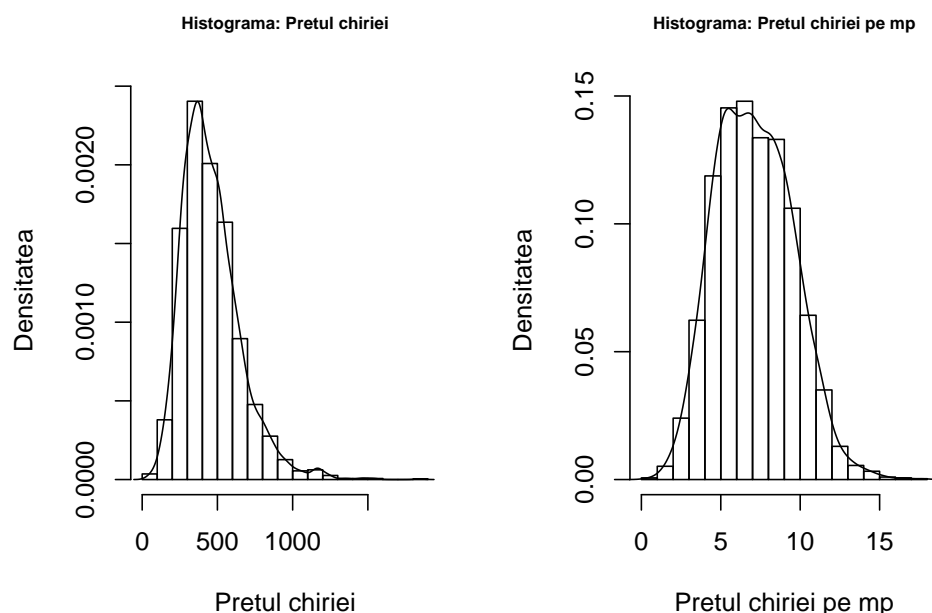
Variabila	Descriere	Media/Frecvența	Abaterea standard	Min/Max
rent	Prețul lunar net al chiriei (în Euro)	459.44	195.66	40.51/1843.38
rentsqm	Prețul lunar net al chiriei pe m^2 (în Euro)	7.11	2.44	0.41/17.72
area	Suprafața de locuit în m^2	67.37	23.72	20/160
yearc	Anul de construcție	1956	22.31	1918/1997
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 bucătăriei: 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 sumarizarea 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 Munchen, 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:



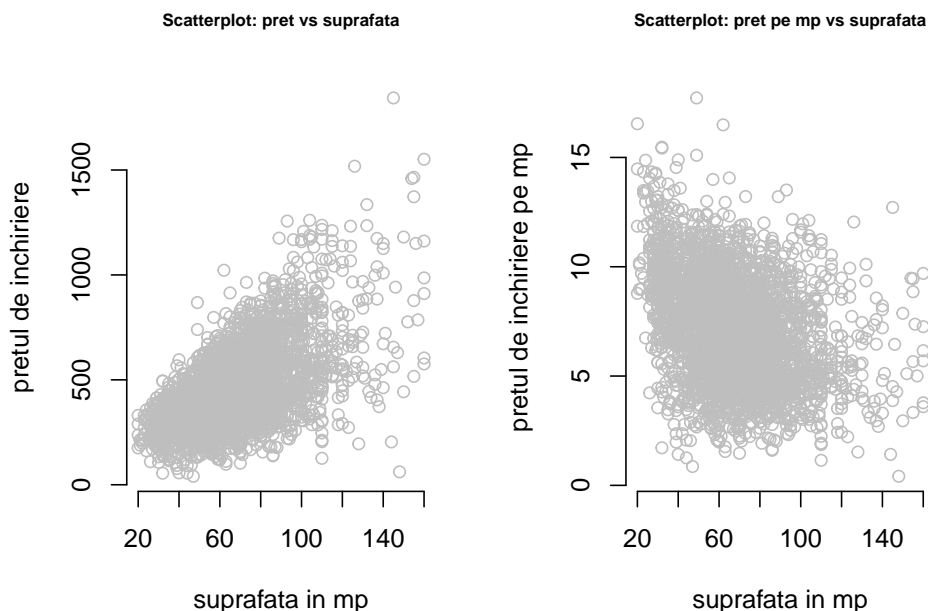
Ex. 3



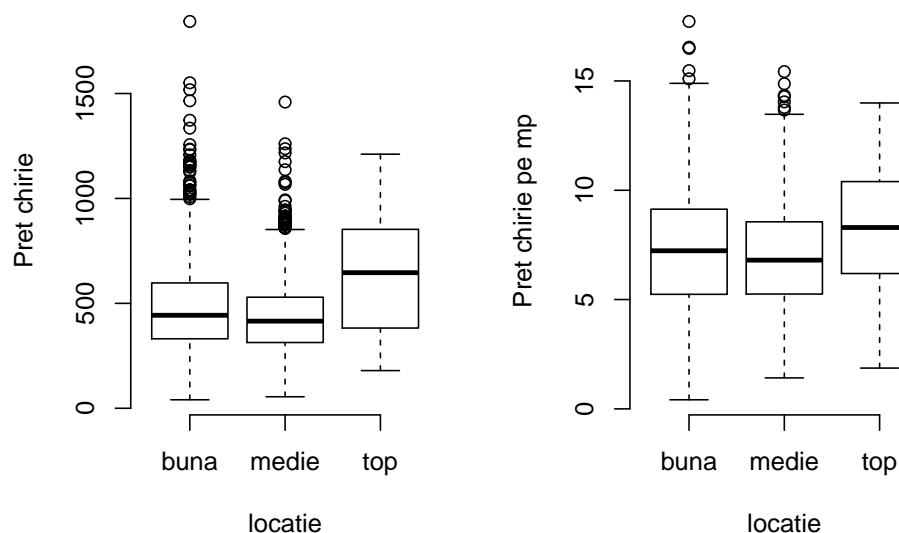
Ilustrați prin intermediul unei histograme (`hist()`) cum este repartizată suprafața de locuit (variabila `area`) și respectiv anul de construcție a imobilului (variabila `yearc`).

În cazul în care ne interesăm asupra relației dintre variabila răspuns și variabila explicativă putem folosi ca metodă grafică diagrama de împrăștiere (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 împrăștiere 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 mustăți). Astfel, în figura următoare se poate observa că valoarea mediană a prețului lunar net al chiriei crește odată cu calitatea locației apartamentului.

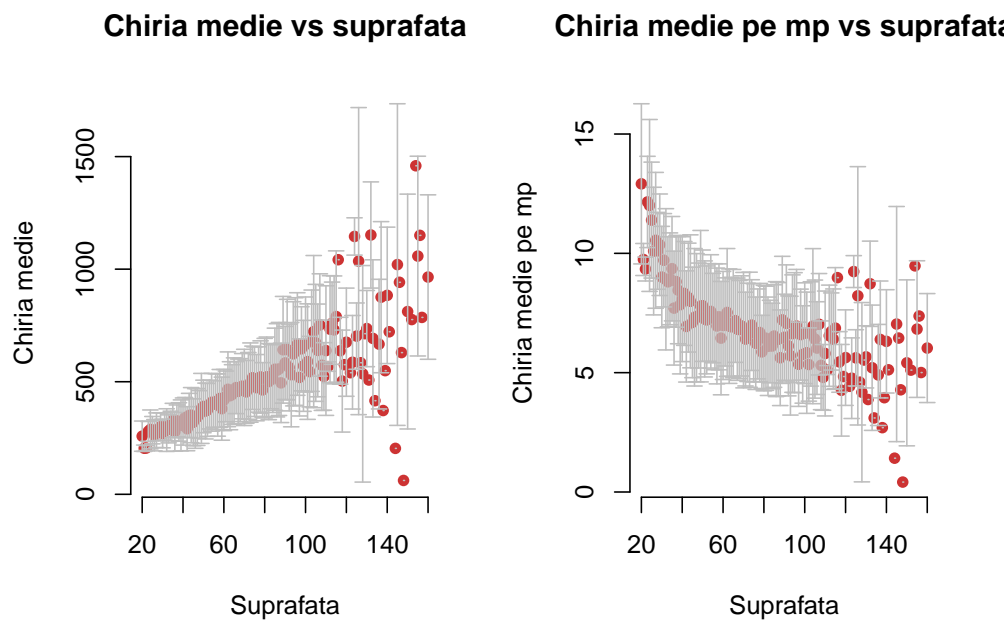


Ex. 4



Ilustrați prin intermediul unei diagrame de împrăștiere 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 (cluster) 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 variabila răspuns pentru fiecare nivel (medie, medie - abatere standard, medie + abatere standard) și să ilustrăm doar datele summarize.



Funcția 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")  
}
```

5 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)

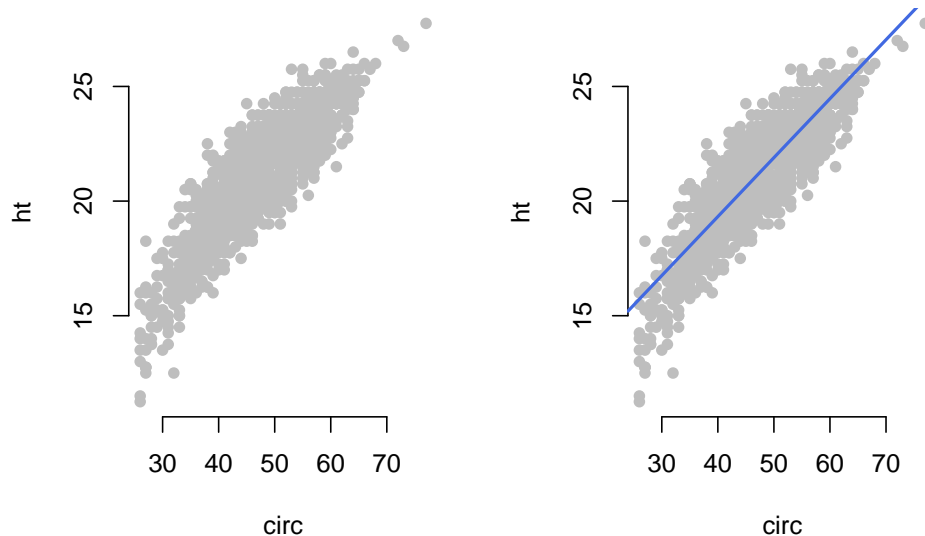
$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

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*

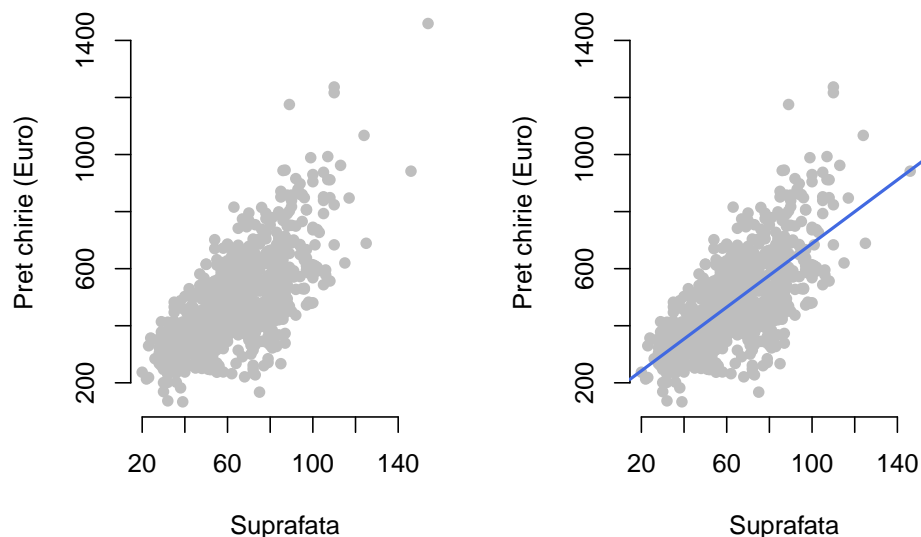
Ca prim exemplu, putem considera setul de date referitor la înălțimea și circumferința arborilor de eucalipt. Modelul de regresie liniară simplă prin care dorim să explicăm înălțimea (medie) a arborilor (variabila răspuns) în funcție de circumferința lor (variabila explicativă) este dat de

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



b) *Exemplu - Prețul chirii în Munchen*

Să considerăm acum setul de date referitor la prețul chiriilor în Munchen 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 suprafața de locuit, i.e. $\mathbb{E}[pret|suprafata] = \beta_0 + \beta_1 suprafata$.

5.1 Metoda celor mai mici pătrate

5.1.1 Calculul estimatorilor prin metoda celor mai mici pătrate

Metoda celor mai mici pătrate este o metodă deterministă de calcul a estimatorilor coeficienților drepte 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. 5.1.1



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

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

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 $\nabla S = 0$ 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

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

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.

5.1.1.1 Exemplu - Prețul chiriilor în Munchen 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

$$pret_i = 130.554 + 5.576 \cdot suprafața_i + \varepsilon_i.$$

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.

5.1.2 Proprietăți ale estimatorilor obținuți prin metoda celor mai mici pătrate

Sub ipotezele făcute asupra termenilor eroare (\mathcal{H}_1 și \mathcal{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. 5.1.2

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

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

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

În mod similar,

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

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. 5.1.3

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)

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

prin urmare

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

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

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

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

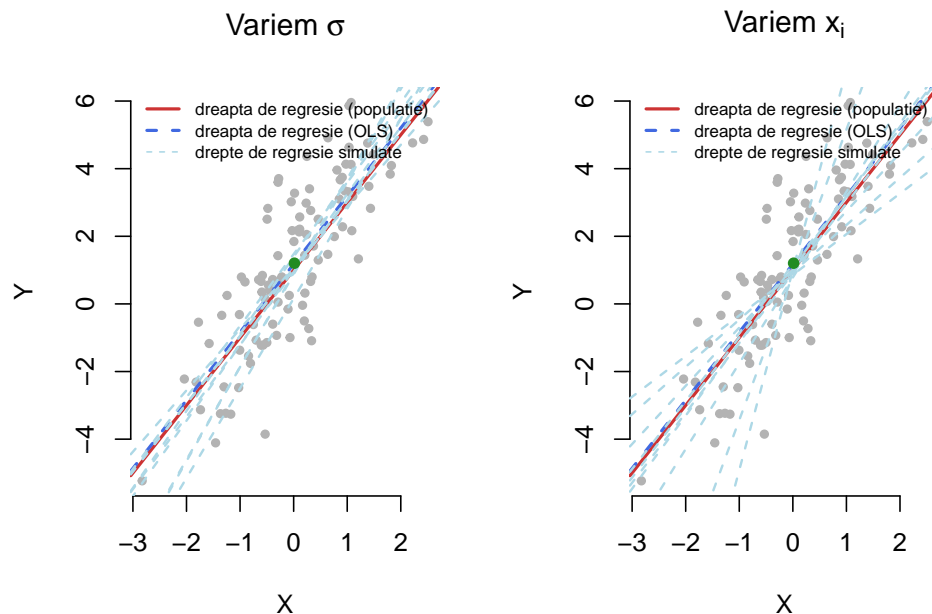
Matricea de varianță-covarianță a estimatorilor $\hat{\beta}_0$ și $\hat{\beta}_1$ devine

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

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 dispersate în jurul valorii medii \bar{x} estimarea coeficientului $\hat{\beta}_1$ este mai precisă ($Var(\hat{\beta}_1)$ este mai mică). Acest fenomen se poate observa și în figura de mai jos în care am generat 100 de valori aleatoare X și 100 de valori pentru Y după modelul

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

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



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.



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

Prop. 5.1.4

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

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

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 implică $\sum_{i=1}^n d_i = 0$ 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 ținând 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

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

5.1.3 Valori reziduale

În modelul de regresie liniară simplă am estimat prin intermediul metodei celor mai mici pătrate atât ordonata la origine a drepte 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. 5.1.5

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. 5.1.6



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

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

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

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

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

Luând media găsim că

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

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

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

5.1.3.1 Exemplu - Prețul chiriilor în Munchen Observăm că pentru setul de date care face referire la prețul chiriilor în Munchen, 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

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

5.1.4 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. 5.1.7



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

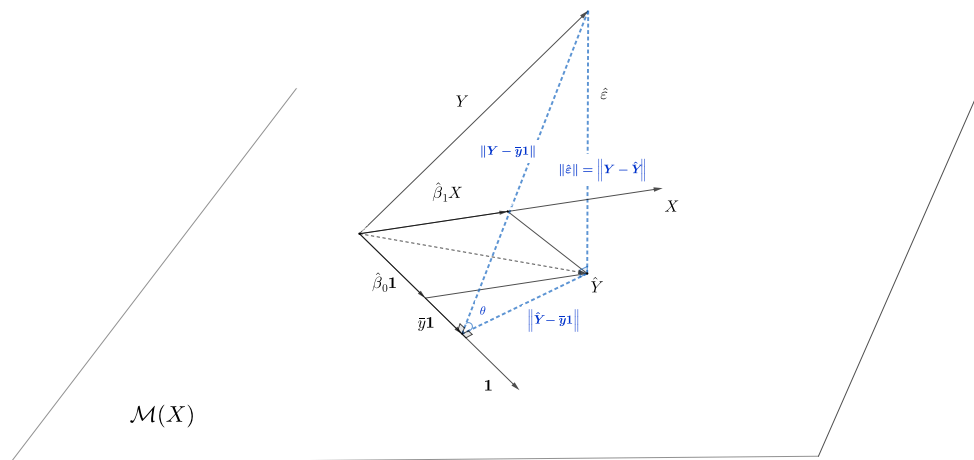
$$Var(\hat{\varepsilon}_{n+1}) = Var(y_{n+1} - \hat{y}_{n+1}) = Var(y_{n+1}) + Var(\hat{y}_{n+1}) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right].$$

5.2 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

Fie $\mathbf{1} = (1, 1, \dots, 1)^\top \in \mathbb{R}^n$ și $\mathcal{M}(X)$ subspațiul linear 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)

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).



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

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

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

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

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

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

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

De asemenea

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

și cum

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

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

5.2.0.1 Exemplu - Prețul chiriilor în Munchen Pentru a vedea dacă modelul de regresie liniară simplă este potrivit în contextul setului de date referitor la prețul chiriilor 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 chiriilor crește odată cu creșterea suprafeței de locuit. Această problemă va fi tratată într-o secțiune ulterioară care face referire la validarea ipotezelor modelului propus.

5.3 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

$$(\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \\ (\mathcal{H}_2) : \varepsilon_i \text{ independente} \end{cases}$$

Prin urmare modelul de regresie liniară simplă devine un model parametric guvernat 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 ,

$$\forall i \in \{1, \dots, n\} \quad y_i \sim \mathcal{N}(\beta_1 + \beta_2 x_i, \sigma^2),$$

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. 5.1.1) 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. 5.1.6).

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

5.3.1 Repartițiile estimatorilor și regiuni de încredere

Pentru a ușura scrierea rezultatelor din această secțiune vom folosi următoarele notații (a se vedea Prop. 5.1.3):

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

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. 5.3.1

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. 5.3.2



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 repartiția 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 repartiția 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. 5.3.3



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

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

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

5.3.1.1 Exemplu - Prețul chiriilor în Munchen Raportându-ne la setul de date referitor la prețul chiriilor în Munchen avem următoarele intervale de încredere:

- pentru $\hat{\beta}_0$

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

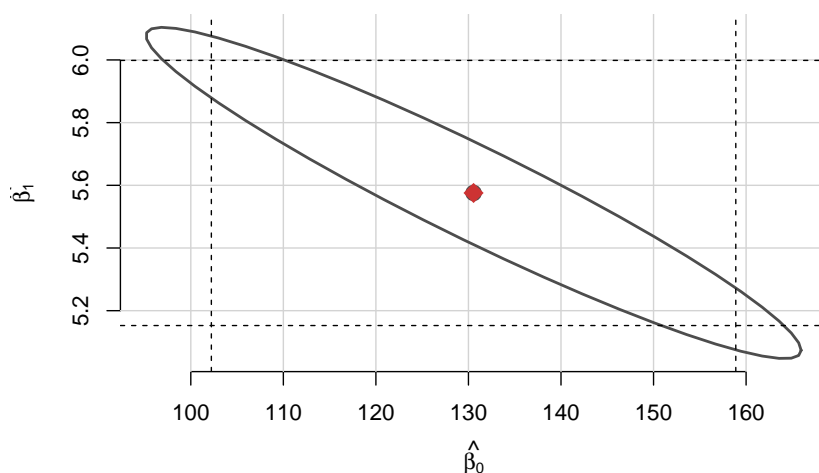
Observăm că intervalul de încredere pentru $\hat{\beta}_0$ este destul de larg (lungime 56.665), ceea ce este justificat dat fiind variabilitatea erorilor ($\hat{\sigma} = 123.455$) dar în special pentru că suprafețele sunt mai mari de zero.

- 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 suprafeței 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 parametri.

5.3.2 Predicție

Vom începe prin a da intervalul de încredere pentru răspunsul mediu care reiese din prima parte a Prop. 5.1.7:

Prop. 5.3.4



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 cuantila 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 decât 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 varianța 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. 5.3.5



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 variabile $u = x - \bar{x}$ 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 dată de inegalitatea lui Bonferroni ($\mathbb{P}(\cup_{i=1}^m A_i) \leq \sum_{i=1}^m \mathbb{P}(A_i)$) 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. 5.3.6



În contextul problemei de regresie liniară simplă cu erori repartizare gaussian 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

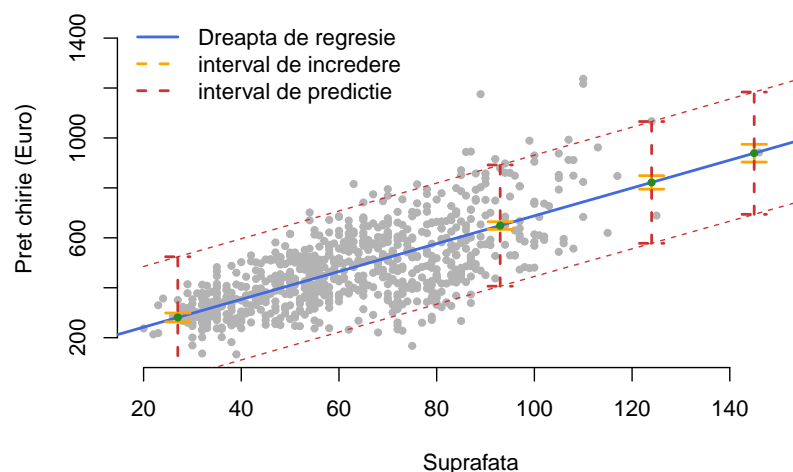
$$\mathbb{P} \left(\hat{y} - M_\alpha \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_j - \bar{x})^2}} \leq y \leq \hat{y} + M_\alpha \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_j - \bar{x})^2}}, \quad \forall x \right) \geq 1 - \alpha$$

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 $F_{2,n-2}$.

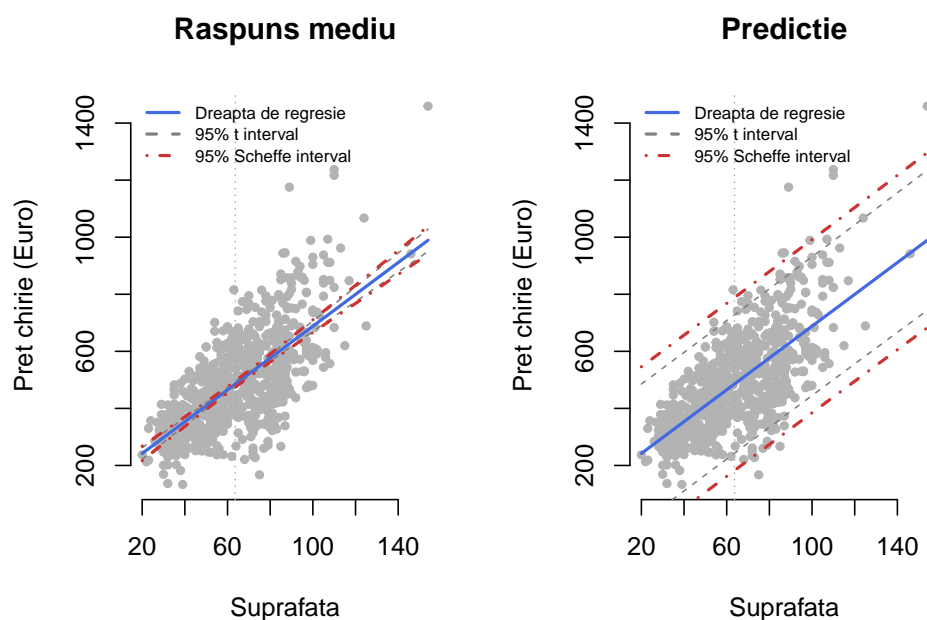
5.3.2.1 Exemplu - Prețul chiriilor în Munchen Revenind la exemplul modelului de regresie liniară simplă introdus pe setul de date **Munchen** să presupunem că ne interesăm la prețul de închiriere al apartamentelor care au suprafața de 27, 93, 124 și respectiv 145 de metri pătrați. Tabelele următoare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

Suprafata	Valori prezise	CI_Inf	CI_Sup	Suprafata	Valori prezise	Pred_Inf	Pred_Sup
27	281.0942	263.2530	298.9355	27	281.0942	38.08053	524.1079
93	649.0816	633.8394	664.3237	93	649.0816	406.24485	891.9183
124	821.9241	794.9069	848.9413	124	821.9241	578.06498	1065.7832
145	939.0110	903.4809	974.5411	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. 5.3.4 și 5.3.5 respectiv banda lui Scheffe din Prop. 5.3.6:



5.4 Aplicații numerice

Ex. 6



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

<i>Tata :</i>	65	63	67	64	68	62	70	66	68	67	69	71
<i>Fiu :</i>	68	66	68	65	69	66	68	65	71	67	68	70

Obținem următoarele rezultate numerice

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

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

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

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

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

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

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

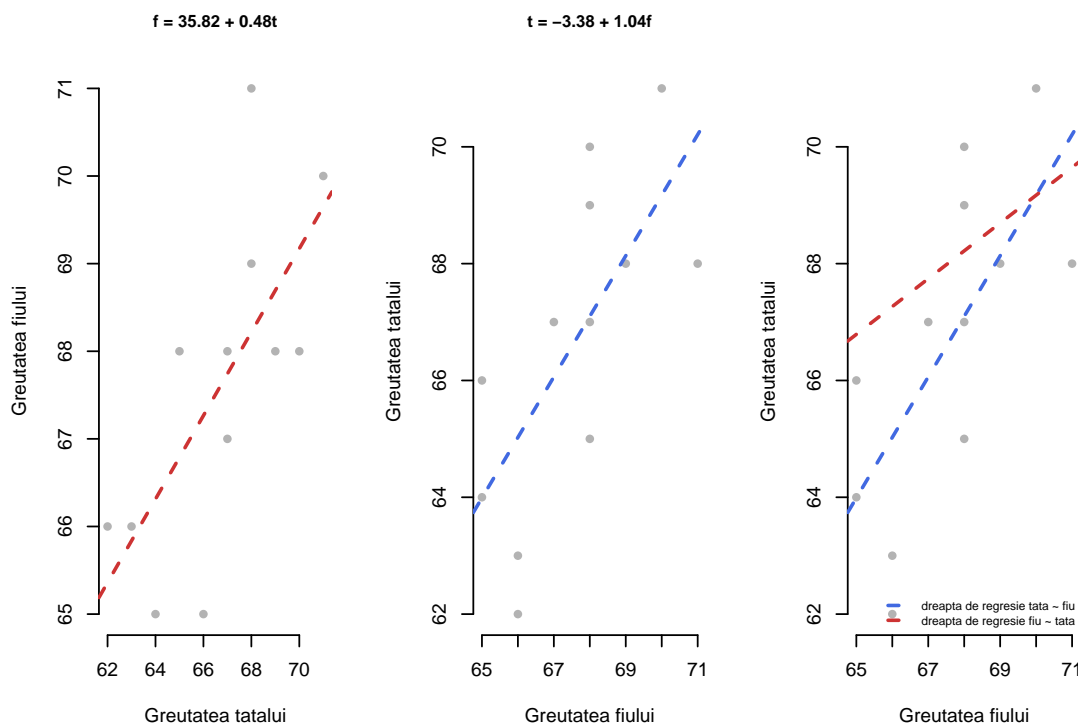
În cazul problemei, coeficienții sunt $\hat{\beta}_0 = -3.38$ și $\hat{\beta}_1 = 1.03$ iar dreapta de regresie este $t = -3.38 + 1.03f$ (a se vedea figura din mijloc).

3. Produsul pantelor celor două drepte este

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

și conform Prop. 5.2.1 și a definiției coeficientului de determinare avem

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



Ex. 7



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

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

1. Notăm cu $y = \hat{\beta}_0 + \hat{\beta}_1 x$ dreapta de regresie. Calculați coeficienții $\hat{\beta}_0$ și $\hat{\beta}_1$.
2. Dați și calculați o măsură care descrie calitatea concordanței datelor cu modelul propus.
3. 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 repartizate normal de medie 0 și varianțe egale. Vrem să testăm ipotezele $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ pentru $j = 0, 1$. De ce acest test este interesant în contextul problemei noastre?

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

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

și respectiv

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

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

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

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

3. Sub ipoteza modelului condiționat normal (erorile ε_i sunt variabile aleatoare independente 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

$$\hat{\sigma}_j^2, \text{ deducem că } \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim t_{n-2}.$$

Prin urmare, sub H_0 avem că

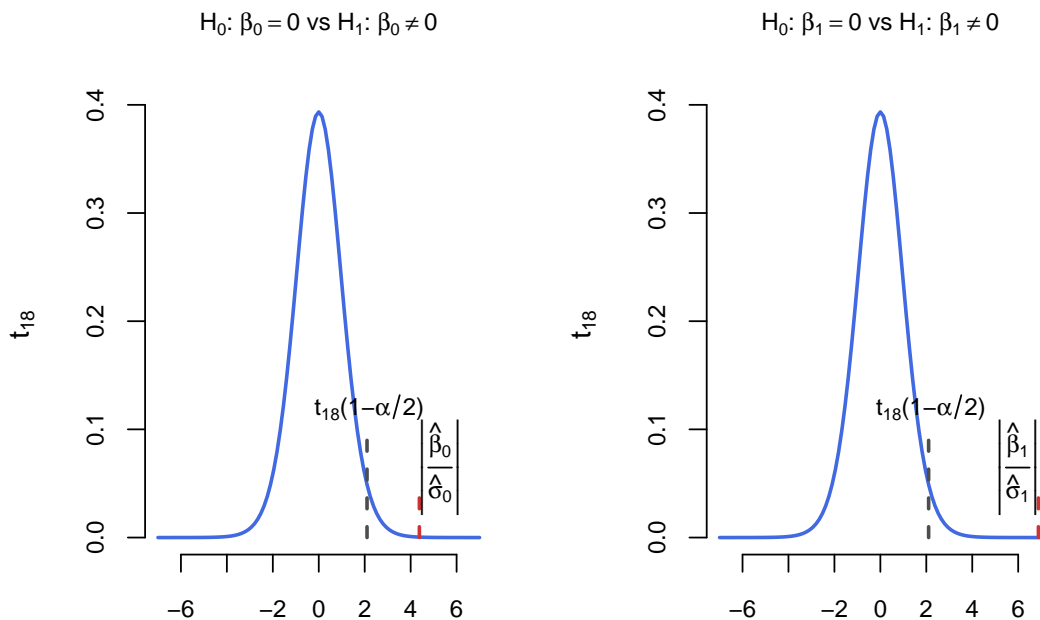
$$\frac{\hat{\beta}_0}{\hat{\sigma}_0} \sim t_{18},$$

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.



Anexe

Repartiții derivate din repartiția normală

În afară de repartiția normală, următoarele trei repartiții sunt des utilizate în inferența statistică a modelului clasic de regresie liniară: repartiția χ^2 , repartiția Student t și repartiția Fisher-Snedecor F .

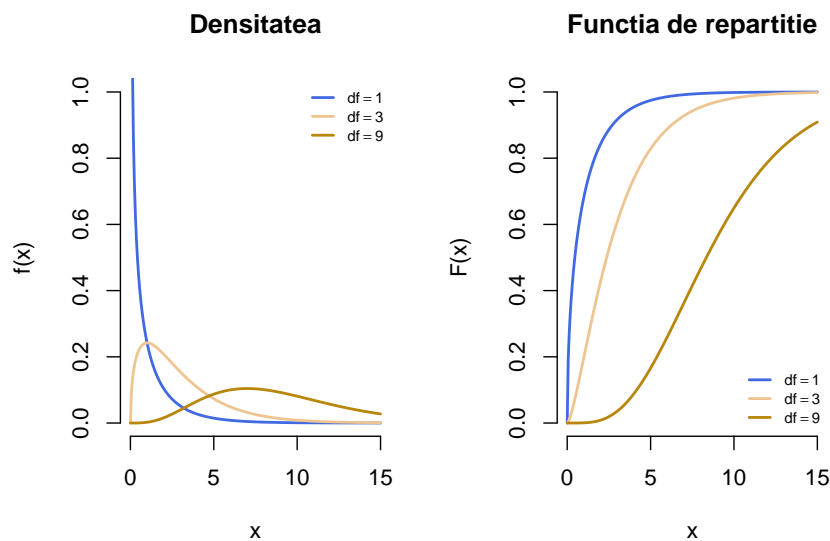
Repartiția χ^2

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

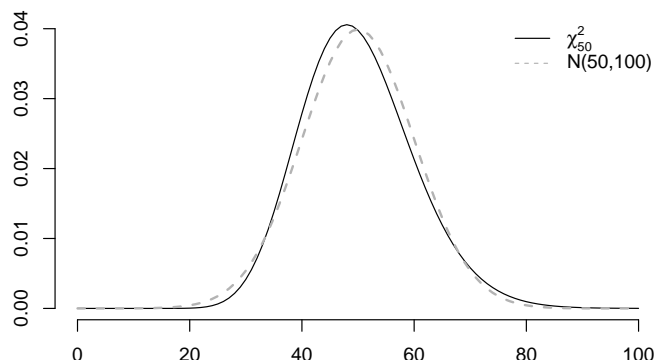
$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} \mathbf{1}_{\{x>0\}}$$

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 $\mathbb{E}[X] = n$ 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 $[n - 2\sqrt{2n}, n + 2\sqrt{2n}]$.

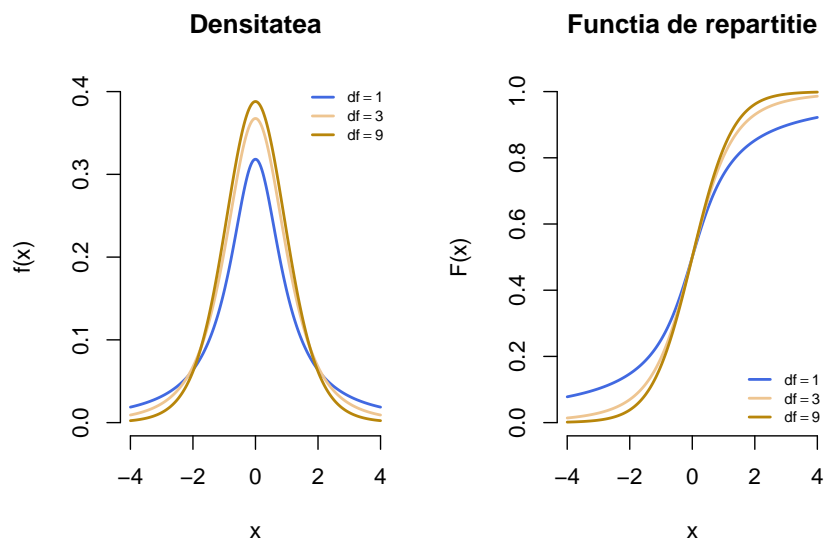


Repartiția t -Student

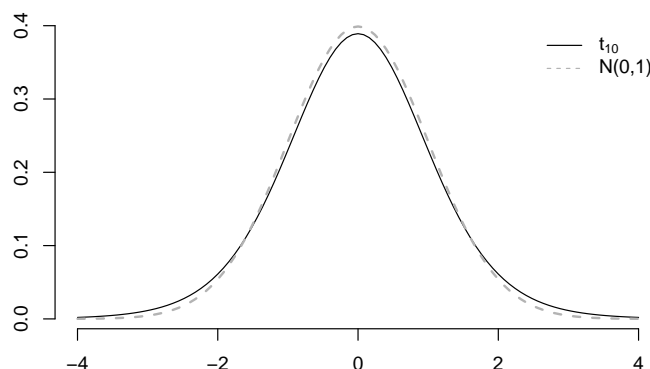
Fie U o variabilă aleatoare repartizată $\mathcal{N}(0, 1)$ și V o variabilă repartizată χ_n^2 , cu U și V independente. Repartiția variabilei aleatoare $T = \frac{U}{\sqrt{\frac{V}{n}}}$ se numește repartiția Student cu n grade de libertate și se notează cu $T \sim t_n$. Densitatea repartiției t_n este

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n} \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad x \in \mathbb{R}$$

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 $\text{Var}(T) = \frac{n}{n-2}$.



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 $n \rightarrow \infty$).

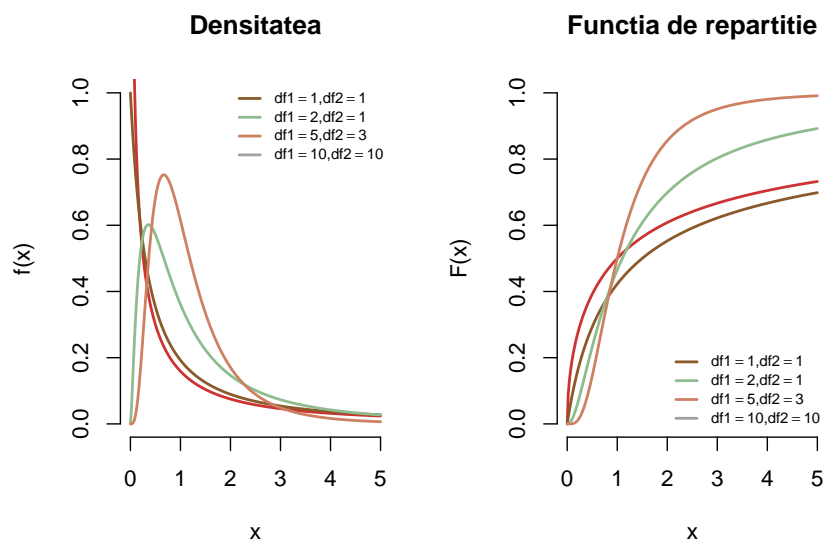


Repartiția Fisher-Snedecor

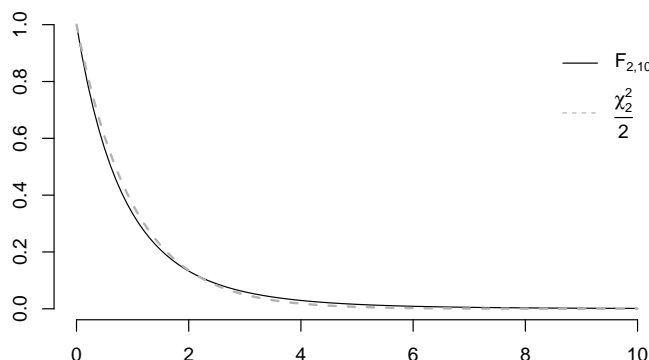
Fie U o variabilă aleatoare repartizată $\chi^2_{n_1}$ și V o variabilă aleatoare repartizată $\chi^2_{n_2}$, cu U și V independente. Repartiția variabilei aleatoare $F = \frac{U/n_1}{V/n_2}$ se numește repartiția Fisher-Snedecor cu n_1 grade de libertate la numărător și n_2 grade de libertate la numitor 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 $F \approx \frac{\chi^2_{n_1}}{n_1}$.



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. 7.0.1



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. 7.0.2



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 $\frac{\bar{X}_n - \theta}{\sqrt{1/n}} \sim \mathcal{N}(0, 1)$. Fie z_1, z_2 cuantile ale repartiției normale standard pentru care are loc relația

$$\mathbb{P}_\theta \left(z_1 \leq \frac{\bar{X}_n - \theta}{\frac{1}{\sqrt{n}}} \leq z_2 \right) = \mathbb{P}_\theta \left(\bar{X}_n - \frac{z_2}{\sqrt{n}} \leq \theta \leq \bar{X}_n - \frac{z_1}{\sqrt{n}} \right) = 1 - \alpha$$

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 sutnem în cadrul unei probleme de optimizare, mai precis avem problema (cu notațiile standard pentru funcția de repartiție 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. 7.0.3



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:

- repartiția lui $g(X_1, \dots, X_n, \theta)$ nu depinde de θ
- 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. 7.0.4



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:

- $g(x_1, \dots, x_n, \cdot)$ 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 repartiție 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. 7.0.5



Spunem că o densitate de repartiție f este unimodală dacă există x^* astfel încât $f(x)$ este crescătoare pentru $x \leq x^*$ și descrescătoare pentru $x \geq x^*$.

Prop. 7.0.6



Fie $f(x)$ o densitate de repartiție unimodală. Dacă intervalul $[a, b]$ satisface:

- 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 (în θ) $\frac{T_n - \theta}{s_n(\theta)}$ 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. 7.0.7  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 forma $n \frac{(\bar{X}_n - \theta)}{\theta} \leq z_{1-\frac{\alpha}{2}}^2$ 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 satisface

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\frac{1}{nI_1(\theta)}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

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

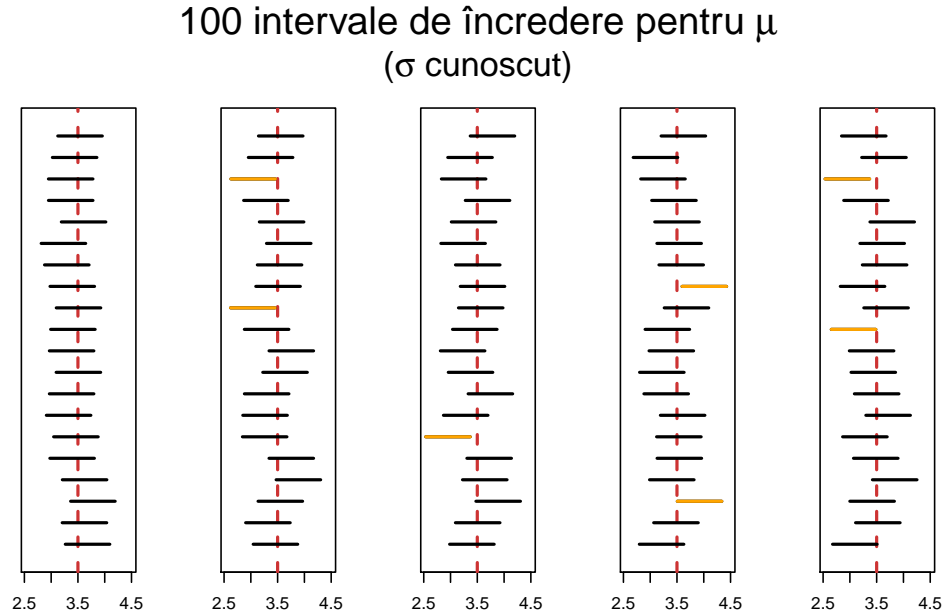
$$\mathbb{P}_\mu \left(z_{\frac{\alpha}{2}} \leq \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \leq z_{1-\frac{\alpha}{2}} \right) = \mathbb{P}_\mu \left(\bar{X}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

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:



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 statistica $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ (varianța eșantionului) în $\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}$ obținem statistica $T_n = \sqrt{n} \frac{\bar{X}_n - \mu}{S_n}$ 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

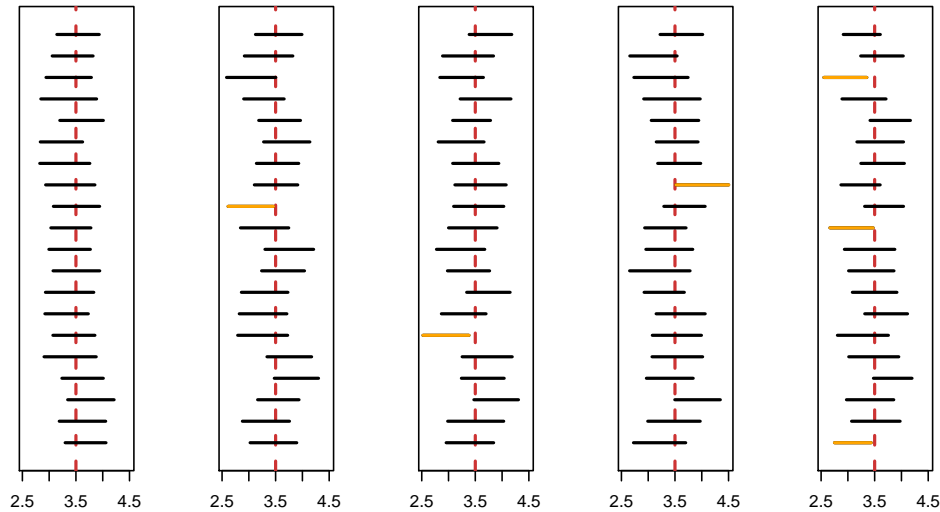
$$\mathbb{P}_\mu \left(t_{n-1, \frac{\alpha}{2}} \leq \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \leq t_{n-1, 1-\frac{\alpha}{2}} \right) = \mathbb{P}_\mu \left(\bar{X}_n - t_{n-1, 1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{n-1, 1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \right) = 1 - \alpha,$$

unde $t_{n-1, \frac{\alpha}{2}} = -t_{n-1, 1-\frac{\alpha}{2}}$ din simetria repartiției 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 eșantion 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 eșantionului 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

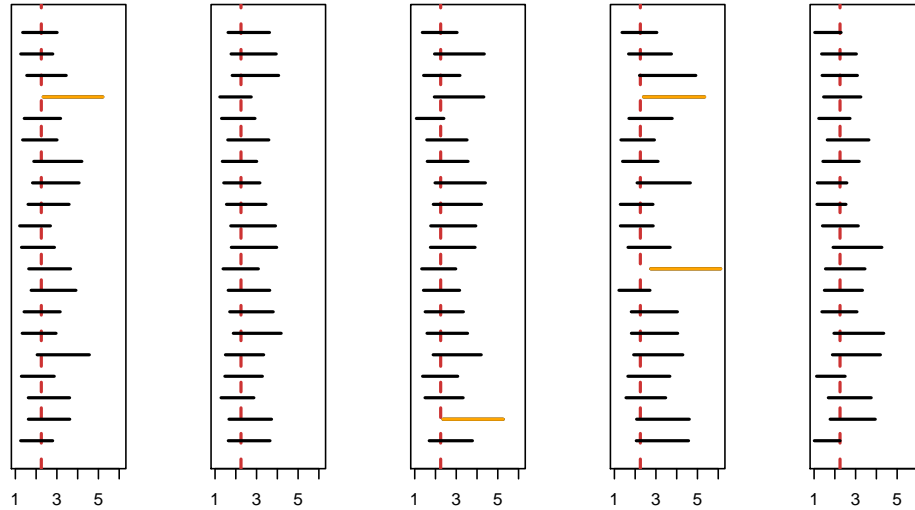
$$IC^{1-\alpha}(\sigma^2) = \left[\frac{(n-1)S_n^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}, \frac{(n-1)S_n^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right].$$

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 independență, $\bar{X}_{n_1} - \bar{Y}_{n_2} \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$. 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 independența 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 statistica

$$\frac{\frac{(\bar{X}_{n_1} - \bar{Y}_{n_2}) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2 \cdot (n_1+n_2-2)}}} \sim \frac{\mathcal{N}(0, 1)}{\sqrt{\frac{\chi_{n_1+n_2-2}^2}{n_1+n_2-2}}} \sim t_{n_1+n_2-2}.$$

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 rescrie

$$\frac{(\bar{X}_{n_1} - \bar{Y}_{n_2}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

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 potriviți 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 $s_d^2 = \sigma_1^2 + \sigma_2^2 - 2Cov(X_i, Y_i)$. 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:

