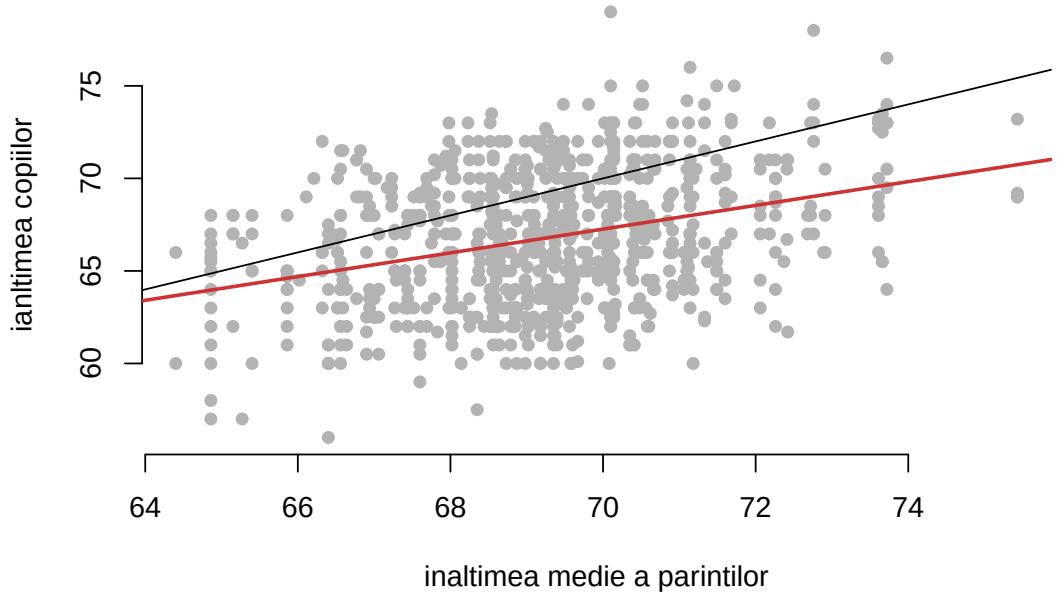
Introducere în modele de regresie

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 adolescenți[1]. 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 = "înălțimea medie a părinților" influențează *variabila răspuns* y =

[1]Setul 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 $x_1, \ldots, x_k$ 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 $f(x_1, \ldots, x_k)$ 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

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

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

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

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: $x_1, \ldots, x_n$[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 $y_1, \ldots, y_n$ sunt văzute ca realizări ale unor variabile aleatoare $Y_1, \ldots, Y_n$, 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 valoare lui $X = x$, $\mathbb{E}[Y|X = x] = f(x)$.

[2]Aici trebuie avut grijă să nu se confunde valorile $x_1, \ldots, x_n$ 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

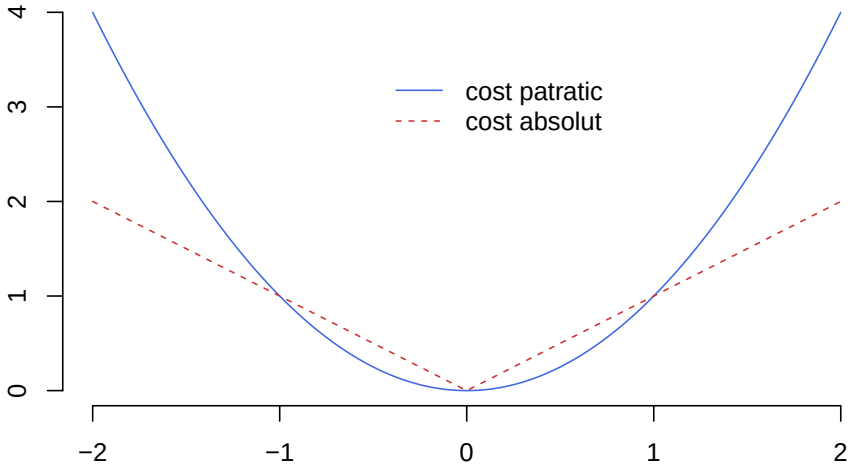
$$y = \mathbb{E}[y|x_1, \ldots, x_k] + \varepsilon = f(x_1, \ldots, x_k) + \varepsilon.$$

Modelele de regresie se pot clasifica, în funcție de forma variabilei răspuns, în modele univariate atunci când aceasta este o variabilă aleatoare și respectiv multivariate atunci când aceasta este un vector aleator iar în raport cu numărul de predictori în modele simple atunci când avem un singur predictor sau modele multiple atunci când intervin mai multe varaibile 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, $y_i, x_{i1}, \ldots, x_{ik}$, $i = 1, \ldots, n$ 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

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

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 $\mathcal{F} = \{f \mid f(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\}$, prin urmare media (condiționată) a lui y este o combinație liniară de covariabile iar variabila răspuns este continuă:

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

Atunci când înlocuim cu observațiile obținem n ecuații

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

cu parametrii necunoscuți (sau coeficienții de regresie) $\beta_0, \ldots, \beta_k$.

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

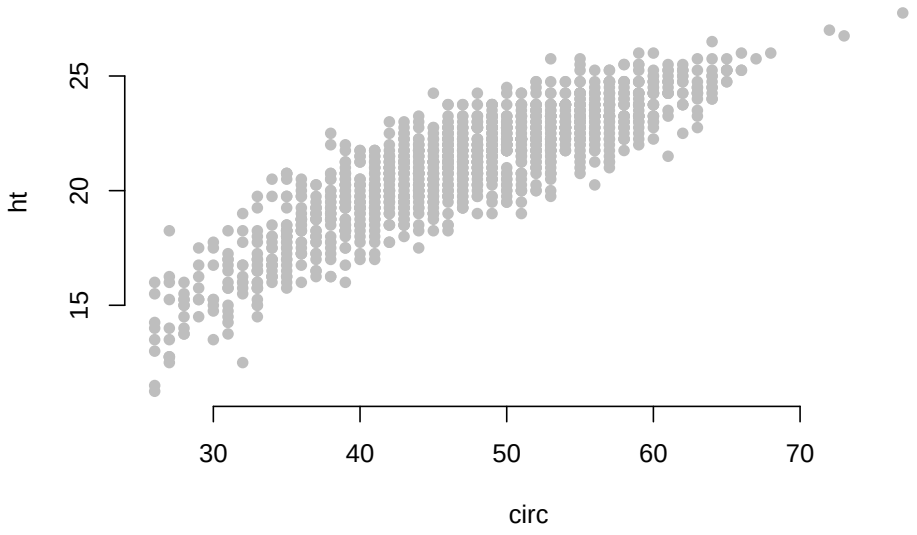
Setul de date `Eucalypt` (care poate fi descărcat de aici) face referire la înălțimea și circumferința (măsurată la 1m 30cm de sol) a 1429 arbori de eucalipt plantați într-o regiune experimentală din Franța. Cele două caracteristici sunt măsurate la vârsta de maturitate a arborilor, anume la 6 ani. O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

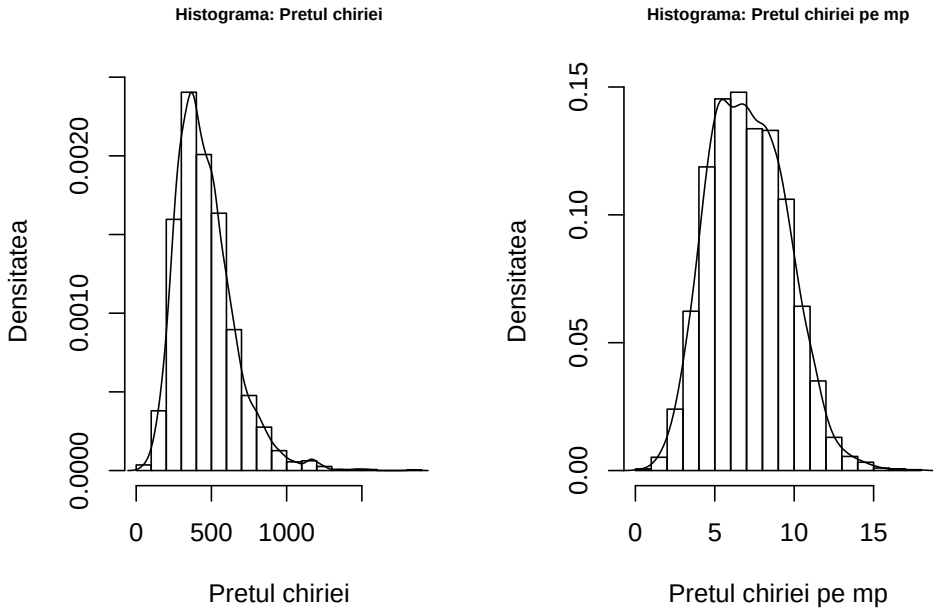
Variabila	Descriere	Media/Frecvența	Abaterea standard	Min/Max
`rent`	Prețul lunar net al chiriei (în Euro)	459.44	195.66	40.51/1843.38
`rentsqm`	Prețul lunar net al chiriei pe m^2 (în Euro)	7.11	2.44	0.41/17.72
`area`	Suprafața de locuit în m^2	67.37	23.72	20/160
`yearc`	Anul de construcție	1956	22.31	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`	alitatea bucătariei: 0 - standard, 1 - premium	95.75%, 4.25%		
`cheating`	Încălzire centralizată: 0 - fără încălzire, 1 - cu încălzire	10.42%, 89.58%		

2.3 Primii pași

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

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:

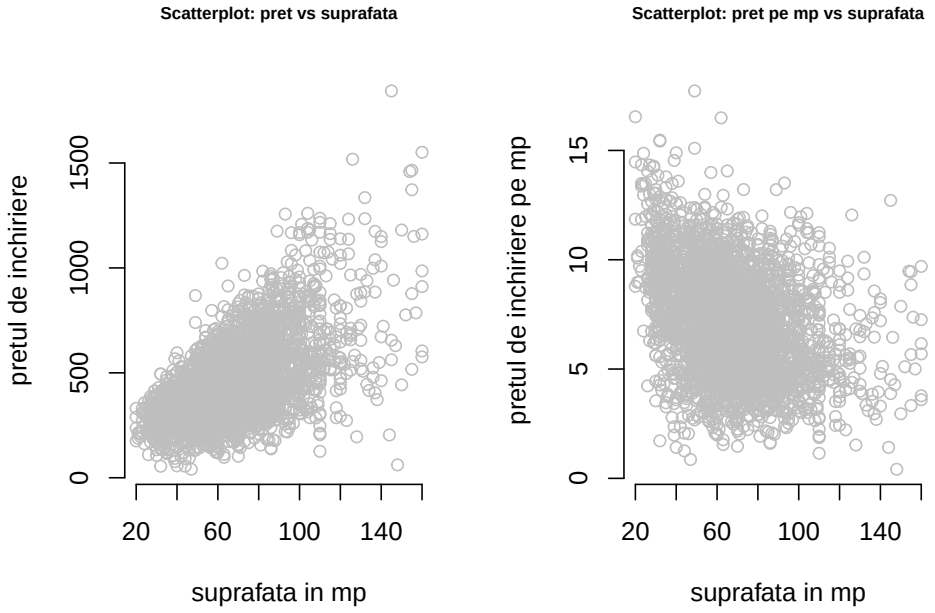


Ex. 2.1

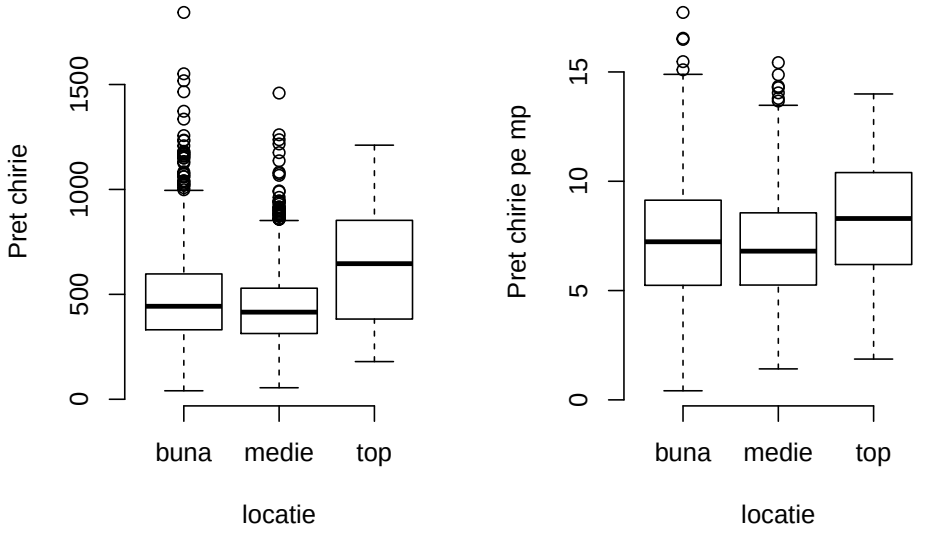
Ilustrați prin intermediul unei histograme (`hist()`) cum este repartizată suprafața de locuit (variabila `area`) și respectiv anul de construție a imobilului (variabila `yearc`).

În cazul în care ne interesăm asupra relației dintre variabila răspuns și variabila explicativă putem folsi 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.



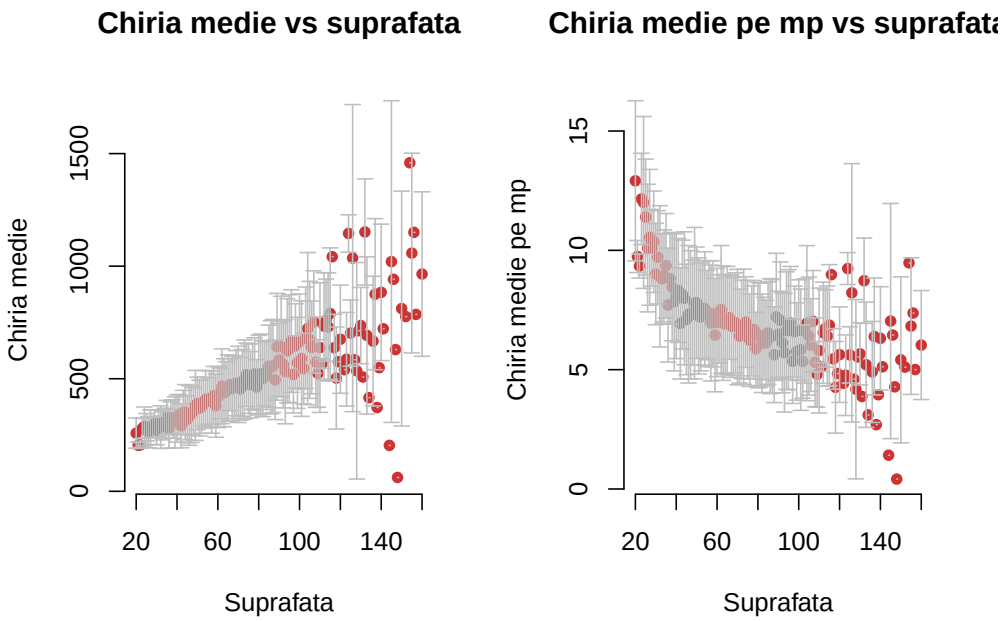
Atunci când variabila explicativă este categorială atunci este de preferat utilizarea boxplot-ului (diagramei cu mustăți). Astfel în figura următoare se poate observa valoarea mediană a prețului lunar net al chiriei crește odată cu calitatea locației apartamentului.



Ex. 2.2

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 boxplor relația dintre prețul lunar net sau prețul lunar net pe m^2 și calitatea băii sau a bucătăriei.

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



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")
}
```

3 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

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

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

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

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)

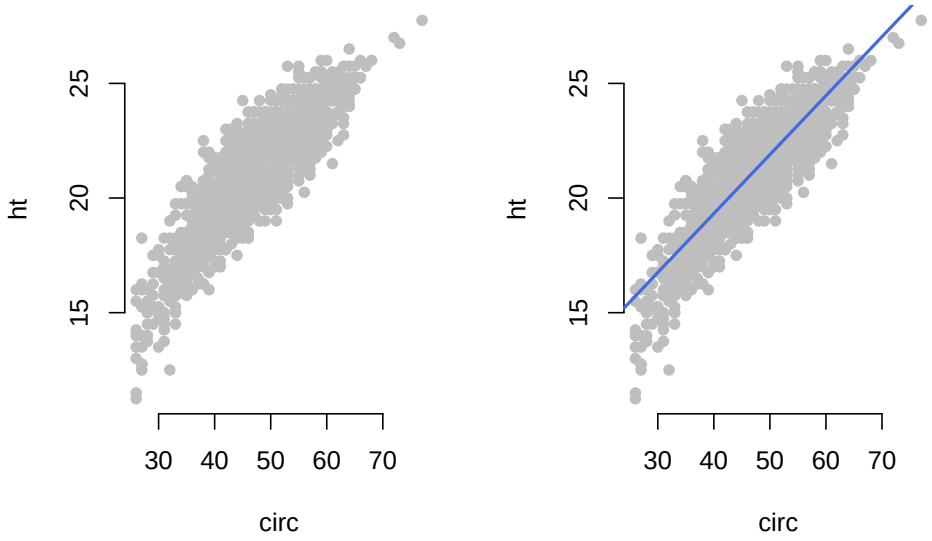
$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

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*

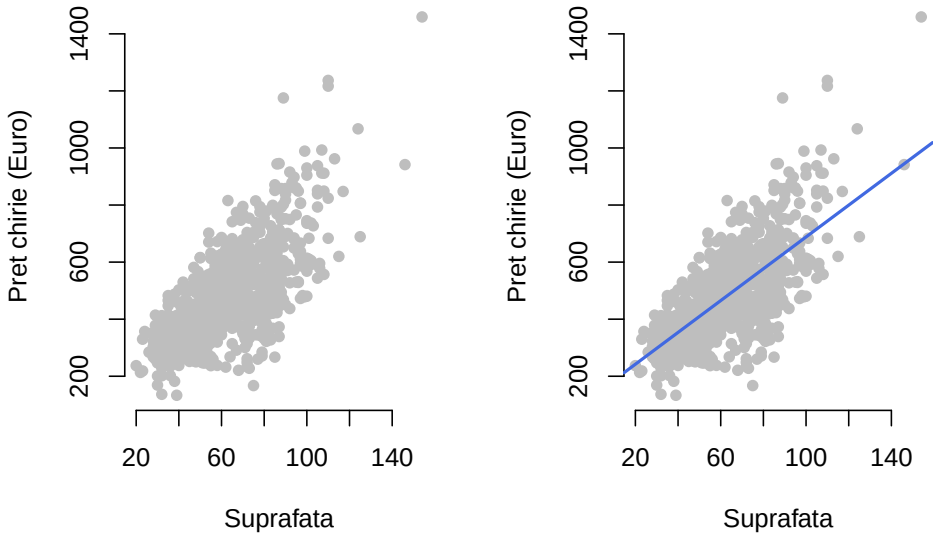
Ca prim exemplu, putem considera setul de date referitor la înălțimea și circumferința arborilor de eucalipt. Modelul de regresie liniară simplă prin care dorim să explicăm înălțimea (medie) a arborilor (variabila răspuns) în funcție de circumferința lor (variabila explicativă) este dat de

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



b) *Exemplu - Prețul chiriilor 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. $\mathbb{E}[pret|suprafata] = \beta_0 + \beta_1 suprafata$.

3.1 Metoda celor mai mici pătrate

3.1.1 Calculul estimatorilor prin metoda celor mai mici pătrate

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

Prop. 3.1

Estimatorii $\hat{\beta}_0$ și $\hat{\beta}_1$ obținuți prin metoda celor mai mici pătrate, adică valorile 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

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

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 $S(\beta_0, \beta_1)$ se scrie sub forma

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

care justifică în egală măsură soluția obținută anterior. □

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.

3.1.1.1 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

$$pret_i = 130.554 + 5.576 suprafata_i + \varepsilon_i.$$

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

3.1.2 Proprietăți ale estimatorilor obținuți prin metoda celor mai mici pătrate

Sub ipotezele făcute asupra termenilor eroare ($\mathcal{H}_1$ și $\mathcal{H}_2$), de centrare, necorelare și homoscedasticitate putem prezenta o serie de proprietăți ale estimatorilor obținuți prin metoda celor mai mici pătrate.

Prop. 3.2 Estimatorii obținuți prin metoda celor mai mici pătrate, $\hat{\beta}_0$ și $\hat{\beta}_1$, sunt estimatori nedeplasați.

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

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

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

În mod similar,

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

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

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

Prop. 3.3 Calculați matricea de varianță-covarianță a estimatorilor $\hat{\beta}_0$ și $\hat{\beta}_1$.

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

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

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

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

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

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

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

prin urmare

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

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

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

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

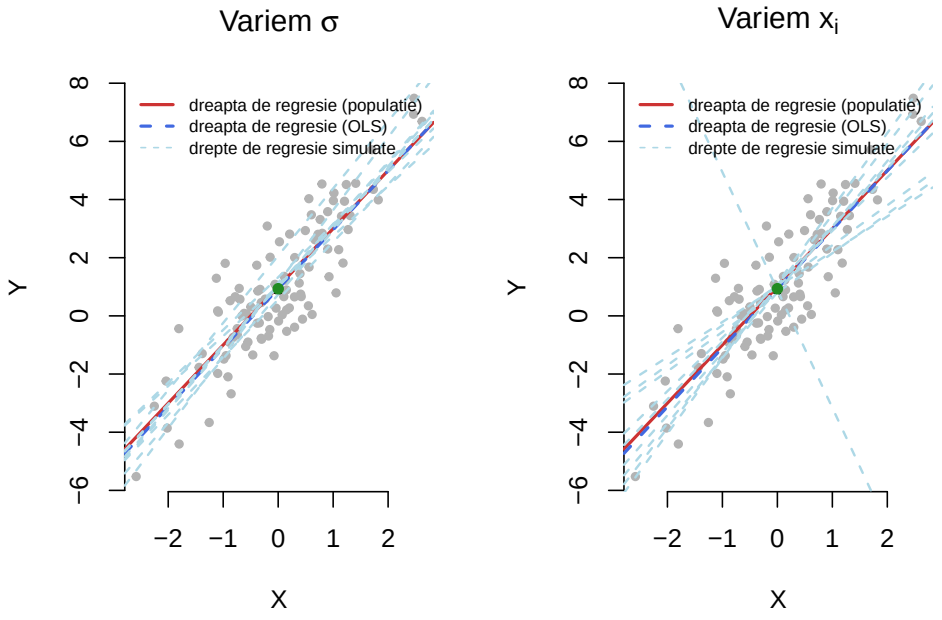
Matricea de varianță-covarianță a estimatorilor $\hat{\beta}_0$ și $\hat{\beta}_1$ devine

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

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 albastru deschis au fost generate tot cu ajutorul metodei celor mai mici pătrate atunci când variem σ^2 (în figura din stânga) și respectiv pe x_i în jurul lui $\bar{x}$ (în figura din dreapta).



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

Prop. 3.4

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

Începem prin a reaminti că un estimator este liniar în y dacă se poate scrie sub forma $\sum_{i=1}^n d_i y_i$ cu $d_1, \ldots, d_n$ 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

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

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

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

3.1.3 Valori reziduale

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

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

Prop. 3.5 În cadrul modelului de regresie liniară simplă, suma valorilor reziduale este nulă.

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

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

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

Prop. 3.6 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

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

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

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

Luând media găsim că

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

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

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

3.1.3.1 Exemplu - Prețul chiriilor în Munchen

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

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

3.1.4 Predicție

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

Prop. 3.7

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 $i = 1, \ldots, n$.

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, $i \in \{1, 2, \ldots, n\}$. Din necorelarea erorilor deducem că

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

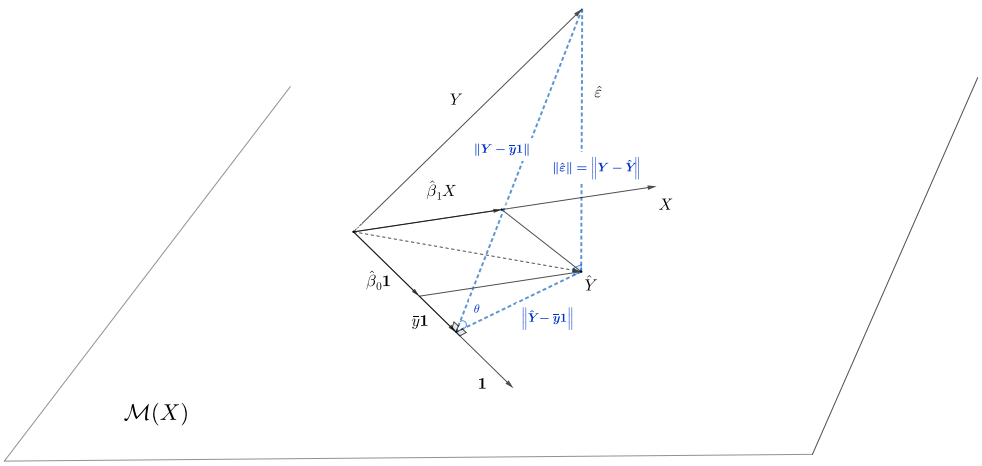
3.2 Coeficientul de determinare R^2 și coeficientul de corelație

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

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

$$\left\|Y-\hat{Y}\right\|=\sum_{i=1}^{n}[y_i-(\hat{\beta}_0+\hat{\beta}_1x_i)]^2$$

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



Observăm că, în general, vectorii $\{\mathbf{1},X\}$ nu formează o bază ortogonală în $\mathcal{M}(X)$ (cu excepția cazului în care $\langle\mathbf{1},X\rangle=n\bar{x}=0$) prin urmare $\hat{\beta}_0\mathbf{1}$ nu este proiecția ortogonală a lui Y pe $\mathbf{1}$ (aceasta este $\frac{\langle Y,\mathbf{1}\rangle}{\|\mathbf{1}\|^2}\mathbf{1}=\bar{y}\mathbf{1}$) iar $\hat{\beta}_1X$ nu este proiecția ortogonală a lui Y pe X (aceasta fiind $\frac{\langle Y,X\rangle}{\|X\|^2}X$).

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

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

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

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

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).

Din definiția coeficientului de determinare R^2 avem că

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

și conform figurii de mai sus $R^2 = \cos^2(\theta)$. Prin urmare dacă $R^2 = 1$, atunci $\theta = 0$ și $Y \in \mathcal{M}(X)$, deci $y_i = \beta_0 + \beta_1 x_i$, $i \in \{1, 2, \ldots, n\}$ (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).

Prop. 3.8

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

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

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

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

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

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

De asemenea

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

și cum

$$\begin{aligned}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})(y_i-\hat{y}_i) &= \sum_{i=1}^{n}(\hat{\beta}_0+\hat{\beta}_1x_i-\bar{y})(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)\\ &= \sum_{i=1}^{n}(\bar{y}-\hat{\beta}_1\bar{x}+\hat{\beta}_1x_i-\bar{y})[(y_i-\bar{y})-\hat{\beta}_1(x_i-\bar{x})]\\ &= \hat{\beta}_1\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})-\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2\\ &= \underbrace{\frac{S_{xy}}{S_{xx}}}_{\hat{\beta}_1}S_{xy}-\frac{S_{xy}^2}{S_{xx}^2}S_{xx}=0\end{aligned}$$

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

3.2.0.1 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.

3.3 Aplicații numerice

Ex. 3.9

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

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

Obținem următoarele rezultate numerice

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

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

1. Dreapta de regresie a greutății fiilor în funcție de greutatea taților este $f=\hat{\alpha}_0+\hat{\alpha}_1t$ unde coeficienții sunt dați de

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

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

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

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

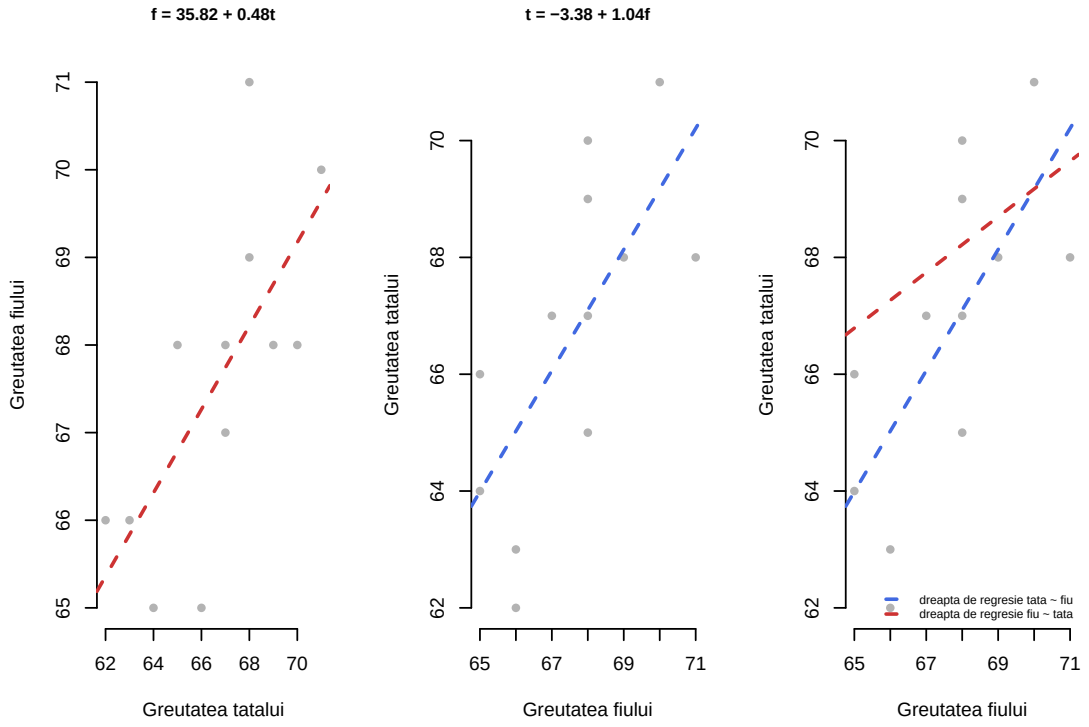
În cazul problemei, coeficienții sunt $\hat{\beta}_0 = -3.38$ și $\hat{\beta}_1 = 1.03$ iar dreapta de regresie este $t = -3.38 + 1.03f$ (a se vedea figura din mijloc).

3. Produsul pantelor celor două drepte este

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

și conform exercițiului 2 și a definiției coeficientului de determinare avem

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



Ex. 3.10

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

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

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 repartizare normal de medie 0 și variante egale. Vrem să testăm ipotezele $H_0:\ \beta_j=0$ versus $H_1:\ \beta_j\neq 0$ pentru $j=0,1$. De ce acest test este interesant în contextul problemei noastre?

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

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

și respectiv

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

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

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

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

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

Prin urmare, sub H_0 avem că

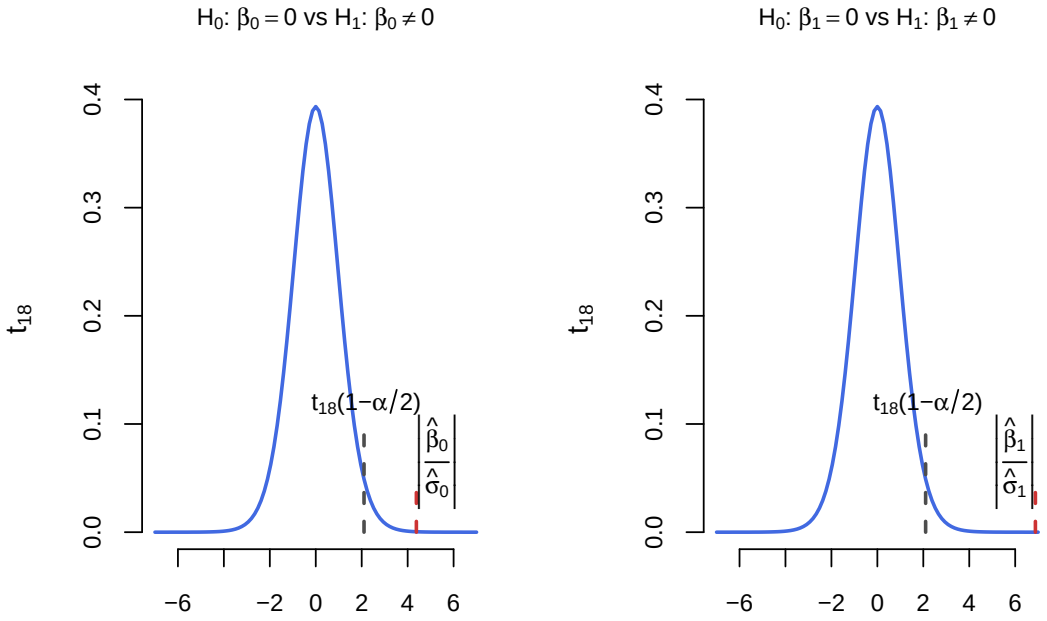
$$\frac{\hat{\beta}_0}{\hat{\sigma}_0}\sim t_{18},$$

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

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

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

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


H0: β0 = 0 vs H1: β0 ≠ 0
t18
0.0
0.1
0.2
0.3
0.4
t18(1−α/2)
|β̂0 / σ̂0|
−6
−2
0
2
4
6
H0: β1 = 0 vs H1: β1 ≠ 0
t18
0.0
0.1
0.2
0.3
0.4
t18(1−α/2)
|β̂1 / σ̂1|
−6
−2
0
2
4
6