Curs 20

Cristian Niculescu

1 Regresia liniară

1.1 Scopurile învăţării

1. Să poată folosi metoda celor mai mici pătrate pentru a potrivi o dreaptă cu date bivariate.
2. Să poată da o formulă pentru eroarea pătratică totală la potrivirea oricărui tip de curbă cu datele.
3. Să poată spune cuvintele homoscedasticitate şi heteroscedasticitate.

1.2 Introducere

Presupunem că avem colectate date bivariate $(x_i, y_i), i = 1, ..., n$. Scopul regresiei liniare este modelarea relaţiei dintre x şi y prin aflarea unei funcţii $y = f(x)$ care dă o potrivire apropiată cu datele. Presupunerile modelării pe care le vom folosi sunt că x_i nu sunt aleatoare şi că y_i este o funcţie de x_i plus un zgomot aleator. Cu aceste presupuneri, x este numită variabilă independentă sau predictor şi y este numită variabilă dependentă sau răspuns.

Exemplul 1. Costul unui timbru de clasa întâi în dolari de-a lungul timpului este dat în lista următoare:

.05 (1963)	.06 (1968)	.08 (1971)	.10 (1974)	.13 (1975)	.15 (1978)	.20 (1981)	.22 (1985)
.25 (1988)	.29 (1991)	.32 (1995)	.33 (1999)	.34 (2001)	.37 (2002)	.39 (2006)	.41 (2007)
.42 (2008)	.44 (2009)	.45 (2012)	.46 (2013)	.49 (2014)			

Folosind codul R:

```
x=c(3,8,11,14,15,18,21,25,28,31,35,39,41,42,46,47,48,49,52,53,54)
y=c(5,6,8,10,13,15,20,22,25,29,32,33,34,37,39,41,42,44,45,46,49)
lm(y~x),
```

obţinem:

```
Call:
lm(formula = y ~ x)
Coefficients:
(Intercept) x
```

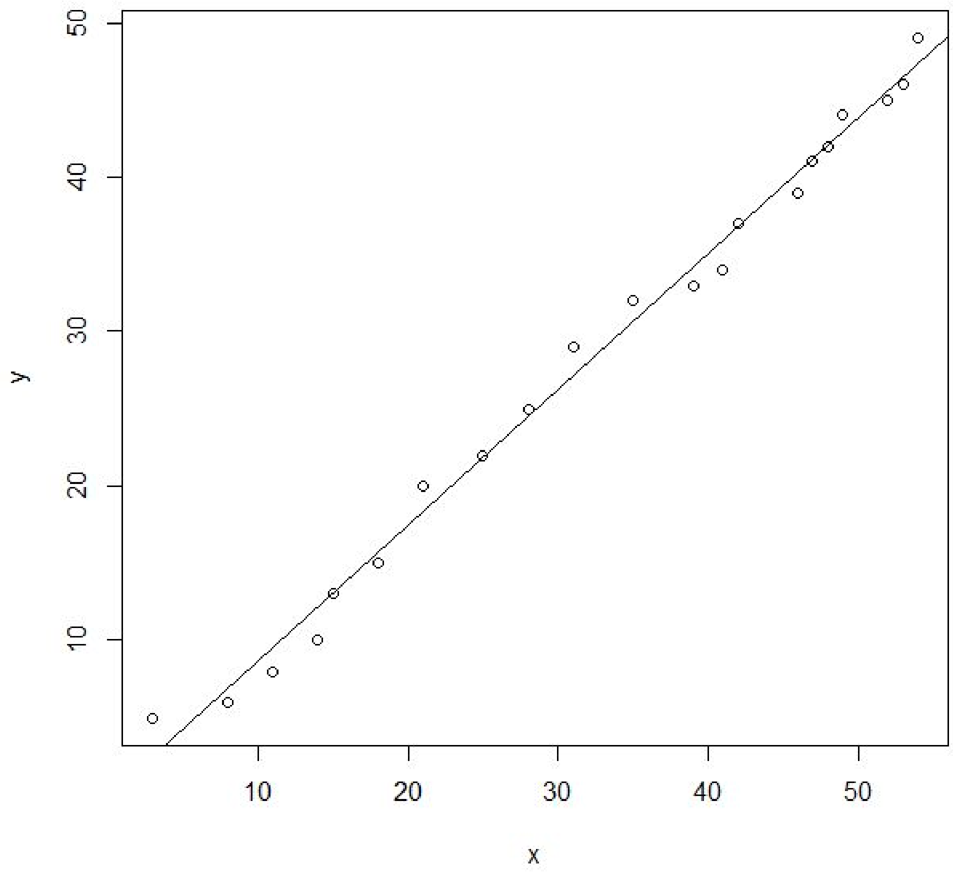
`-0.1324 0.8791.`
Am aflat că dreapta care dă "potrivirea celor mai mici pătrate" cu aceste date (dreapta de regresie) este

$$y = -0.1324 + 0.8791x,$$

unde x este numărul de ani de la 1960, iar y este în cenţi.
Folosind acest rezultat "prezicem" că în 2021 ($x = 61$), costul unui timbru va fi 53 de cenţi (deoarece $-0.1324 + 0.8791 \cdot 61 = 53.4927$).
Folosind codul `R` (în continuarea celui de mai sus):

```
a=-0.1324
b=0.8791
plot(x,y)
abline(a,b)
```

obţinem:

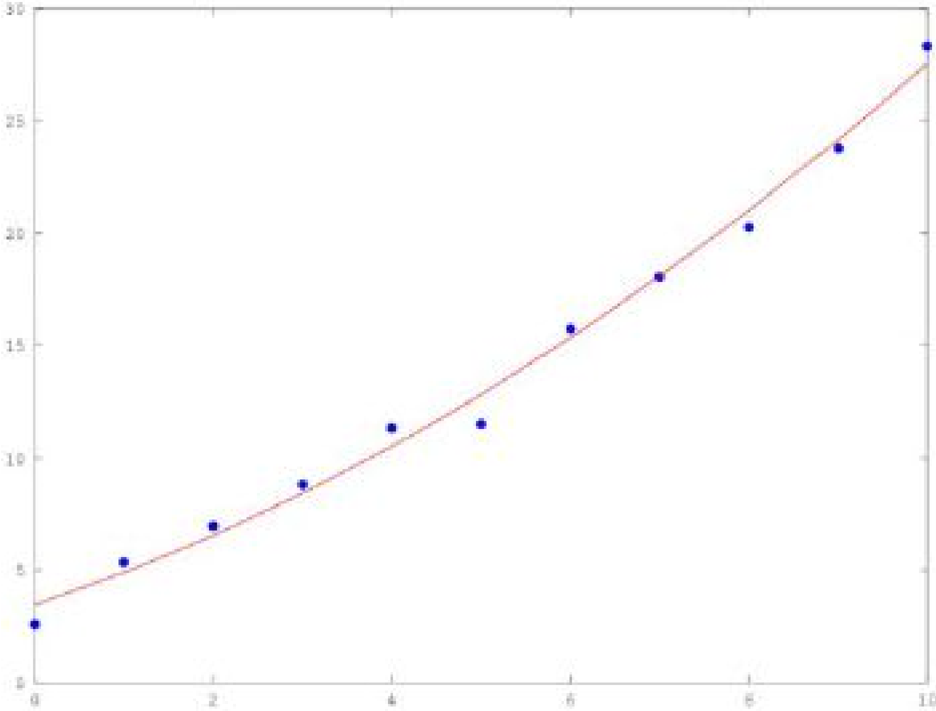


Costul timbrului (cenţi) vs. timp (ani din 1960)

Niciuna din date nu se află chiar pe dreaptă. Mai degrabă această dreaptă are "cea mai bună potrivire" în raport cu toate datele, cu o mică eroare pentru fiecare dată.

Exemplul 2. Presupunem că avem n perechi de taţi şi fii adulţi. Fie x_i şi y_i înălţimile celui de-al i-lea tată, respectiv fiu. Dreapta celor mai mici pătrate pentru aceste date poate fi folosită pentru a prezice înălţimea de adult a unui băiat tânăr din cea a tatălui lui.

Exemplul 3. Nu suntem limitaţi la drepte cu cea mai bună potrivire. $\forall d \in \mathbb{N}^*$, metoda celor mai mici pătrate poate fi folosită pentru a afla un polinom de grad d cu "cea mai bună potrivire cu datele". Iată o figură arătând potrivirea datelor cu o parabolă prin metoda celor mai mici pătrate:



Potrivirea unei parabole, $y = b_2x^2 + b_1x + b_0$ cu datele

1.3 Potrivirea unei drepte folosind cele mai mici pătrate

Presupunem că avem datele (x_i, y_i) ca mai sus. Scopul este să aflăm dreapta

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

care "se potriveşte cel mai bine" cu datele. Modelul nostru spune că fiecare y_i este prezis de x_i până la o eroare ϵ_i:

$$y_i = \beta_1 x_i + \beta_0 + \epsilon_i.$$

Deci,

$$\epsilon_i = y_i - \beta_1 x_i - \beta_0.$$

Metoda celor mai mici pătrate află valorile $\hat{\beta}_0$ şi $\hat{\beta}_1$ ale lui β_0 şi β_1 care minimizează suma pătratelor erorilor:

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

Folosind analiza matematică (detalii în adaos), aflăm

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}, \ \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}, \tag{1}$$

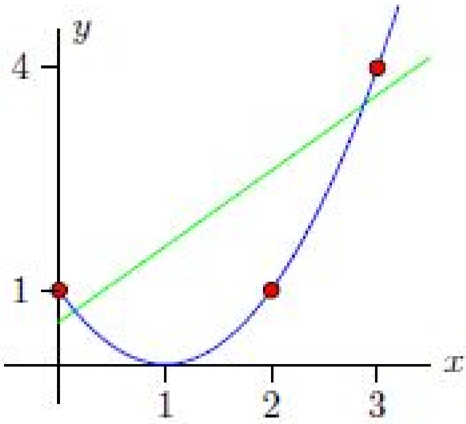
unde

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

Aici, $\overline{x}$ este media de selecţie a lui x, $\overline{y}$ este media de selecţie a lui y, s_{xx} este dispersia de selecţie a lui x şi s_{xy} este covarianţa de selecţie a lui x şi y.
Exemplul 4. Folosiţi cele mai mici pătrate pentru a potrivi cu o dreaptă următoarele date: (0,1), (2,1), (3,4).
Răspuns. În cazul nostru, $(x_1, y_1) = (0, 1)$, $(x_2, y_2) = (2, 1)$ şi $(x_3, y_3) = (3, 4)$. Deci

$$\begin{aligned}
\overline{x} &= \frac{1}{3}(x_1 + x_2 + x_3) = \frac{1}{3}(0 + 2 + 3) = \frac{5}{3},\\
\overline{y} &= \frac{1}{3}(y_1 + y_2 + y_3) = \frac{1}{3}(1 + 1 + 4) = \frac{6}{3} = 2,\\
s_{xx} &= \frac{1}{3-1}\sum_{i=1}^{3}(x_i - \overline{x})^2 = \frac{1}{2}\left[\left(0 - \frac{5}{3}\right)^2 + \left(2 - \frac{5}{3}\right)^2 + \left(3 - \frac{5}{3}\right)^2\right] = \frac{7}{3},\\
s_{xy} &= \frac{1}{2}\left[\left(0 - \frac{5}{3}\right)(1-2) + \left(2 - \frac{5}{3}\right)(1-2) + \left(3 - \frac{5}{3}\right)(4-2)\right] = 2;\\
\hat{\beta}_1 &= \frac{s_{xy}}{s_{xx}} = \frac{2}{\frac{7}{3}} = \frac{6}{7};\\
\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1\overline{x} = 2 - \frac{6}{7}\cdot\frac{5}{3} = 2 - \frac{10}{7} = \frac{4}{7}.
\end{aligned}$$

Deci dreapta de regresie a celor mai mici pătrate are ecuaţia $y = \frac{4}{7} + \frac{6}{7}x$. Aceasta este arătată ca dreapta verde din figura următoare.



Potrivirea celor mai mici pătrate a unei drepte (verde) şi a unei parabole (albastru)

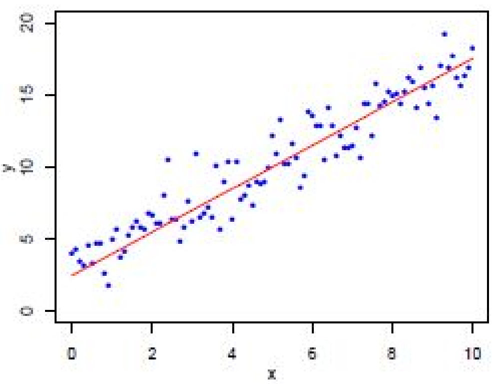
Regresie liniară simplă: Este puţin confuz, dar cuvântul "liniară" din "regresie liniară" nu se referă la potrivirea unei drepte. Totuşi, cea mai uzuală curbă pentru potrivire este o dreaptă. Potrivirea unei drepte la date bivariate este numită regresie liniară simplă.

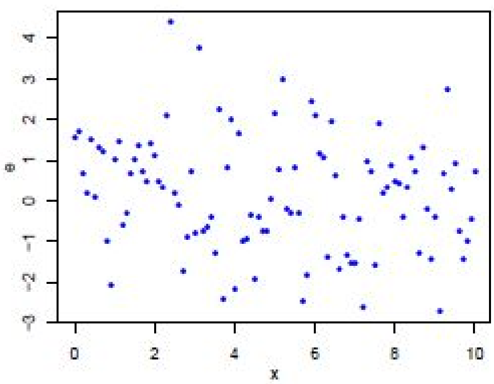
1.3.1 Reziduuri

Pentru o dreaptă, modelul este

$$y_i = \hat{\beta}_1 x_i + \hat{\beta}_0 + \epsilon_i.$$

Gândim $\hat{\beta}_1 x_i + \hat{\beta}_0$ ca prezicând sau explicând y_i. Termenul rămas ϵ_i este numit reziduul, pe care-l gândim ca pe un zgomot aleator sau o eroare de măsurare. O verificare vizuală folositoare a modelului de regresie liniară este reprezentarea reziduurilor. Datele ar trebui să fie lângă dreapta de regresie. Reziduurile ar trebui să arate cam la fel de-a lungul domeniului lui x.

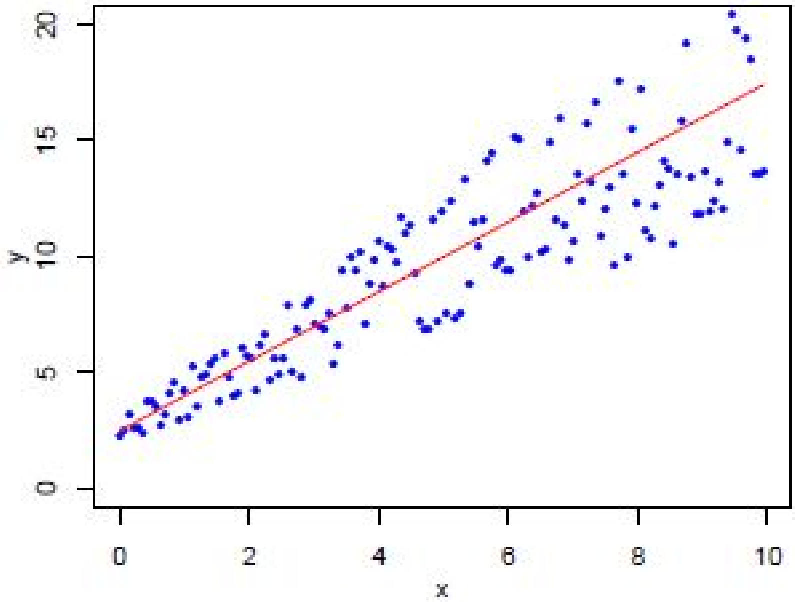




Date cu dreapta de regresie (stânga) şi reziduuri (dreapta). Observaţi homoscedasticitatea.

1.3.2 Homoscedasticitatea

O presupunere importantă a modelului de regresie liniară este că reziduurile ϵ_i au aceeaşi dispersie $\forall i$. Această presupunere este numită homoscedasticitate. Puteţi vedea aceasta în cazul ambelor figuri de mai sus. Datele sunt în banda de lăţime fixă în jurul dreptei de regresie şi la fiecare x reziduurile au cam aceeaşi împrăştiere verticală.
Mai jos este o figură arătând date heteroscedastice. Împrăştierea verticală a datelor creşte când x creşte. Înainte de a folosi cele mai mici pătrate pe aceste date ar trebui să transformăm datele pentru a fi homoscedastice.



Date heteroscedastice

1.4 Regresie liniară pentru potrivirea polinoamelor

Potrivirea unei drepte la date este numită regresie liniară simplă. Putem de asemenea folosi regresia liniară pentru a potrivi polinoame cu datele. Folosirea cuvântului "liniară" în ambele cazuri poate părea confuză. Aceasta este deoarece cuvântul "liniară" din "regresia liniară" nu se referă la potrivirea unei drepte. Mai degrabă se referă la ecuaţiile algebrice liniare pentru parametrii necunoscuţi β_i, i.e. fiecare β_i are exponentul 1.
Exemplul 5. Luaţi aceleaşi date ca în exemplul 4 şi folosiţi cele mai mici pătrate pentru a afla parabola cu cea mai bună potrivire pentru date.

Răspuns. O parabolă are formula $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Eroarea pătratică este

$$S(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{3} (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2.$$

După substituirea valorilor date pentru x_i şi y_i putem folosi analiza matematică (egalăm derivatele parţiale în raport cu $\beta_0, \beta_1, \beta_2$ cu 0, obţinând un sistem de 3 ecuaţii liniare cu 3 necunoscute) pentru a afla tripletul $(\beta_0, \beta_1, \beta_2)$ care minimizează S. Sau putem folosi codul `R`

```
x=c(0,2,3)
y=c(1,1,4)
C=cbind(1,x,x^ 2)
solve(t(C)%*%C,t(C)%*%y).
```

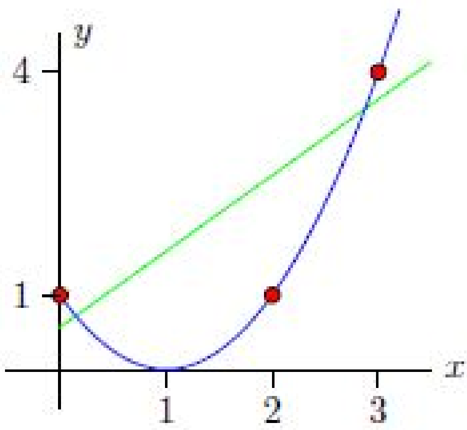
Sau codul `R`

```
x=c(0,2,3)
y=c(1,1,4)
x1=x
x2=x^ 2
lm(y~x1+x2)
```

Cu aceste date, parabola celor mai mici pătrate are ecuaţia

$$y = 1 - 2x + x^2.$$

Pentru 3 puncte, potrivirea pătratică este perfectă.



Potrivirea celor mai mici pătrate a unei drepte (verde) şi a unei parabole (albastru)

Exemplul 6. Perechile (x_i, y_i) pot da vârsta şi mărimea vocabularului a n copii. Deoarece copiii mici dobândesc cuvinte noi într-un ritm accelerat, putem ghici că un polinom de grad mai mare poate fi cea mai bună potrivire

pentru date.
Exemplul 7. (Transformarea datelor). Uneori este necesar să transformăm datele înainte de a folosi regresia liniară. De exemplu, presupunem că relaţia este exponenţială, i.e. $y = ce^{ax}$. Atunci

$$\ln(y) = ax + \ln(c).$$

Deci putem folosi regresia liniară simplă pentru a obţine un model

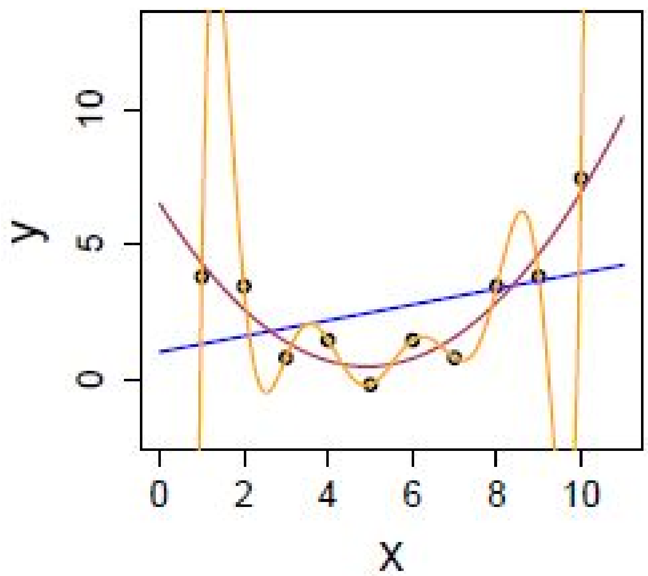
$$\ln(y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

şi apoi obţinem modelul exponenţial

$$y_i = e^{\hat{\beta}_0} e^{\hat{\beta}_1 x_i}.$$

1.4.1 Suprapotrivirea

Putem totdeauna obţine o potrivire mai bună folosind un polinom de ordin mai mare. De exemplu, date fiind 6 date bivariate (cu x_i distincte) se poate totdeauna afla un polinom de grad 5 care trece prin toate. Aceasta poate duce la suprapotrivire. Adică, potrivirea zgomotului la fel de bine ca adevărata relaţie între x şi y. Un model suprapotrivit va potrivi datele originale mai bine, dar va prezice mai puţin bine y pentru noi valori ale lui x. O povocare a modelării statistice este echilibrarea potrivirii modelului cu complexitatea modelului.
Exemplul 8. În reprezentarea de mai jos potrivim polinoame de gradul 1, 2 şi 9 la 10 date bivariate. Modelul de gradul 2 (maro) dă o potrivire semnificativ mai bună decât modelul de gradul 1 (albastru). Modelul de gradul 9 (portocaliu) dă o potrivire exactă cu datele, dar dintr-o privire am ghici că este suprapotrivit. Adică, nu ne aşteptăm că potrivească bine următoarea dată bivariată pe care o vedem. De fapt, datele au fost generate folosind un model pătratic, deci modelul de gradul 2 va tinde să facă cea mai bună potrivire cu date noi.



1.4.2 Funcţia R lm

Nu facem regresia liniară cu mâna. Regresia liniară se reduce la rezolovarea sistemelor de ecuaţii liniare, i.e. la calcul matriceal. Funcţia `R lm` poate fi folosită la potrivirea unui polinom de orice grad cu datele. (`lm` înseamnă model liniar). De fapt, `lm` poate potrivi multe tipuri de funţii, exceptând polinoamele, după cum puteţi explora folosind ajutorul lui `R` sau google.

1.5 Regresie liniară multiplă

Datele nu sunt totdeauna bivariate. Pot fi trivariate sau chiar de o dimensiune mai mare. Presupunem că avem datele de forma

$$(y_i, x_{1i}, x_{2i}, ..., x_{mi}).$$

Putem analiza aceste date într-o manieră foarte similară cu datele bivariate. Adică, putem folosi cele mai mici pătrate pentru a potrivi modelul

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m.$$

Aici fiecare x_j este o variabilă predictor şi y este variabila de răspuns. De exemplu, putem fi interesaţi de cum variază o populaţie de peşti în funcţie nivelele măsurate ale câtorva poluanţi, sau am vrea să prezicem înălţimea de adult a unui fiu pe baza înălţimilor tatălui şi a mamei.

1.6 Cele mai mici pătrate ca un model statistic

Modelul de regresie liniară pentru potrivirea unei drepte spune că valoarea y_i din perechea (x_i, y_i) este extrasă dintr-o variabilă aleatoare

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

unde termenii de "eroare" ε_i sunt variabile aleatoare independente cu media 0 şi deviaţia standard σ. Presupunerea standard este că ε_i sunt i.i.d. cu repartiţia $N(0, \sigma^2)$. În orice caz, media lui Y_i este dată de:

$$E(Y_i) = \beta_0 + \beta_1 x_i + E(\varepsilon_i) = \beta_0 + \beta_1 x_i.$$

Din această perspectivă, metoda celor mai mici pătrate alege valorile lui β_0 şi β_1 care minimizează dispersia de selecţie din jurul dreptei.
De fapt, estimarea celor mai mici pătrate $(\hat{\beta}_0, \hat{\beta}_1)$ coincide cu estimarea de verosimilitate maximă pentru parametrii (β_0, β_1); adică, dintre toţi coeficienţii posibili, $(\hat{\beta}_0, \hat{\beta}_1)$ sunt cei care fac datele observate cele mai probabile.

1.7 Regresia la medie

Motivul termenului "regresie" este că variabila de răspuns prezisă y va tinde să fie "mai aproape" de (i.e. să regreseze la) media ei decât variabila predictor x este faţă de media ei. Aici "mai aproape" este în ghilimele deoarece trebuie să controlăm scala (i.e deviaţia standard) fiecărei variabile. Modul de a controla scala este mai întâi să standardizăm fiecare variabilă.

$$u_i = \frac{x_i - \overline{x}}{\sqrt{s_{xx}}}, \ v_i = \frac{y_i - \overline{y}}{\sqrt{s_{yy}}}.$$

Standardizarea schimbă media în 0 şi dispersia în 1:

$$\overline{u} = \overline{v} = 0, \ s_{uu} = s_{vv} = 1.$$

Proprietăţile algebrice ale covarianţei arată că

$$s_{uv} = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}} = \rho,$$

coeficientul de corelaţie. Astfel, potrivirea celor mai mici pătrate pentru $v = \beta_0 + \beta_1 u$ are

$$\hat{\beta}_1 = \frac{s_{uv}}{s_{uu}} = \rho \text{ şi } \hat{\beta}_0 = \overline{v} - \hat{\beta}_1 \overline{u} = 0.$$

Deci dreapta celor mai mici pătrate este $v = \rho u$. Deoarece ρ este coeficientul de corelaţie, el este între -1 şi 1. Presupunem că este pozitiv şi mai mic ca 1 (i.e., x şi y sunt pozitiv, dar nu perfect corelate). Atunci formula $v = \rho u$ înseamnă că, dacă u este pozitiv, atunci valorea prezisă a lui v este mai mică decât u. Adică, v este mai aproape de 0 decât u. Echivalemt,

$$\frac{y - \overline{y}}{\sqrt{s_{yy}}} < \frac{x - \overline{x}}{\sqrt{s_{xx}}},$$

i.e., y regresează la $\overline{y}$. Standardizarea are grijă de scală.
Considerăm cazul extrem al corelaţiei 0 între x şi y. Atunci, indiferent de valoarea lui x, valoarea prezisă a lui y este totdeauna $\overline{y}$. Adică, y a regresat până la media lui.
Dreapta de regresie trece totdeauna prin punctul $(\overline{x}, \overline{y})$.
Exemplul 9. Regresia la medie este importantă în studiile longitudinale. Rice (*Mathematical Statistics and Data Analysis*) dă următoarul exemplu. Presupunând că li se dau copiilor un test IQ la vârsta de 4 ani şi altul la vârsta de 5 ani, ne aşteptăm ca rezultatele să fie pozitiv corelate. Analiza de mai sus spune că, în medie, acei copii care au făcut slab la primul test tind să arate îmbunătăţire (i.e. regresează la medie) în al 2-lea test. Astfel, o intervenţie inutilă poate fi interpretată greşit ca utilă deoarece pare a îmbunătăţi scorurile.
Exemplul 10. Alt exemplu cu consecinţe practice este recompensa şi pedeapsa. Imaginaţi-vă o şcoală unde performanţa înaltă la un examen este recompensată şi performanţa slabă este pedepsită. Regresia la medie ne spune că (în medie) studenţii foarte performanţi vor face puţin mai slab la următorul examen şi studenţii puţin performanţi vor face puţin mai bine. O viziune nesofisticată a datelor va face să pară că pedeapsa a îmbunătăţit performanţa şi recompensa de fapt a scăzut performanţa. Sunt reale consecinţe dacă cei cu autoritate acţionează după această idee.

1.8 Adaos

1.8.1 Demonstraţia formulei pentru potrivirea celor mai mici pătrate a unei drepte

Cea mai directă demonstraţie este cu analiza matematică. Suma erorilor pătrate este

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

Luând derivatele parţiale (şi reamintind că x_i şi y_i sunt date, deci constante)

$$\frac{\partial S}{\partial \beta_0} = \sum_{i=1}^{n} -2(y_i - \beta_1 x_i - \beta_0) = 0$$
$$\frac{\partial S}{\partial \beta_1} = \sum_{i=1}^{n} -2x_i(y_i - \beta_1 x_i - \beta_0) = 0.$$

Se obţine următorul sistem de 2 ecuaţii liniare în necunoscutele β_0 şi β_1:

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

Rezolvând sistemul obţinem formulele (1).
Pentru multe aplicaţii între discipline vezi:
`http://en.wikipedia.org/wiki/Linear_regression#Applications_of_linear_regression`.

1.8.2 Măsurarea potrivirii

Odată ce se calculează coeficienţii de regresie, este important să verificăm cât de bine modelul de regresie se potriveşte cu datele (i.e., cât de aproape cea mai potrivită dreaptă urmăreşte datele). O măsură uzuală dar brută a "bunătătăţii de potrivire" este coeficientul de determinare, notat R^2. Suma totală a pătratelor este dată de:

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Suma reziduală a pătratelor este dată de suma pătratelor reziduurilor. Când potrivim o dreaptă, aceasta este:

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

RSS este porţiunea "neexplicată" a sumei totale a pătratelor, i.e. neexplicată de ecuaţia de regresie. Diferenţa TSS−RSS este porţiunea "explicată" a sumei totale a pătratelor. Coeficientul de determianre R^2 este raportul dintre porţiunea "explicată" şi suma totală a pătratelor:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}.$$

Cu alte cuvinte, R^2 măsoară proporţia variabilităţii datelor care este contabilizată pentru modelul de regresie. O valoare aproape de 1 indică o potrivire bună, în timp ce o valoare aproape de 0 indică o potrivire slabă. În cazul regresiei liniare simple, R^2 este pur şi simplu pătratul coeficientului de corelaţie dintre valorile observate y_i şi valorile prezise $\beta_0 + \beta_1 x_i$.

Exemplul 11. În exemplul 8 de suprapotrivire, valorile lui R^2 sunt ("degree"="grad"):

degree	R^2
1	0.3968
2	0.9455
9	1.0000

Măsura bunătăţii potrivirii creşte când n (gradul) creşte. Potrivirea este mai bună, dar modelul devine de asemenea mai complex, deoarece este nevoie de mai mulţi coeficienţi pentru a descrie polinoame de grad mai mare.