VALOARE MEDIE CONDITIONATA

MODELE DE REGRESIE; ESTIMAREA PARAMETRILOR REGRESIEI LINIARE

Problema:
Pentru perechea de variabile aleatoare $(X, Y) =$ (efect, cauza), cum evidentiem dependenta lor (cantitativ si calitativ)?
Exemplu: $(X, Y) =$ (valoarea tensiunii arteriale sistolice, nivelul colesterolului)

COEFICIENT DE CORELATIE

Fie (X, Y) pentru care exista momentele de ordinul $2.$ Reamintim definitiile covariantei si a coeficientului de corelatie:

$$cov\,(X, Y) = M\left((X - M\,(X))\,(Y - M\,(Y))\right) = M\,(XY) - M\,(X)\,M\,(Y)$$

$$\rho = \frac{cov\,(X, Y)}{\sqrt{D^2\,(X)\,D^2\,(Y)}}$$

Proprietate: $|\rho| \leq 1$ (rezulta din inegalitatea Schwartz)

- $\rho = 1,$ corelatie pozitiva maxima
- $\rho = -1,$ corelatie negativa maxima
- $\rho = 0,$ necorelare

Repartitii asociate:

$$P \circ (X, Y)^{-1} = \begin{cases} \sum\limits_{x \in A} \sum\limits_{y \in B} p\,(x, y) \cdot \delta_{(x,y)}, & \text{rep. discreta} \\ \text{sau} & \\ f\,(x, y) \cdot l^2, & \text{rep. continua} \end{cases}$$

$$P \circ X^{-1}\,(C_1) = \begin{cases} P \circ (X, Y)^{-1}\,(C_1 \times B)\,, & \text{rep. discreta} \\ \text{sau} & \\ P \circ (X, Y)^{-1}\,(C_1 \times R)\,, & \text{rep. continua} \end{cases}$$

$$P\circ Y^{-1}(C_2)=\begin{cases} P\circ(X,Y)^{-1}(A\times C_2), & \text{rep. discreta}\\ \text{sau} & \\ P\circ(X,Y)^{-1}(R\times C_2), & \text{rep. continua}\end{cases}$$

In cazul repartitiilor discrete,

$$\begin{aligned} p_X(x) &= \sum_{y\in B} p(x,y),\ x\in A\\ p_Y(y) &= \sum_{x\in A} p(x,y),\ y\in B\end{aligned}$$

$$X,Y \ \text{ independente } \ \Leftrightarrow \ p(x,y)=p_X(x)\cdot p_Y(y) \quad \forall x\in A, y\in B$$

In cazul repartitiilor continue,

$$\begin{aligned} f_X(x) &= \int_R f(x,y)\,dy,\ x\in R\\ f_Y(y) &= \int_R f(x,y)\,dx,\ y\in R\end{aligned}$$

$$X,Y \ \text{ independente } \ \Leftrightarrow \ f(x,y)=f_X(x)\cdot f_Y(y) \quad \forall x,y\in R$$

Proprietate:
X,Y independente $\Longrightarrow$ X,Y necorelate

Coeficientul de corelatie apare ca o masura cantitativa a dependentei dintre X si Y.
Introducem si un model stocastic al acestei dependente (al relatiei "cauza - efect")

VALOARE MEDIE CONDITIONATA

Lema

Fie $(\Omega,\mathcal{K},P)$, $\mathcal{F}\subset\mathcal{K}$, $\mathcal{F}$ corp borelian si fie $h:\Omega\longrightarrow\overline{R}$ o variabila aleatoare nenegativa sau integrabila, $\mathcal{F}-$masurabila. Atunci

$$\int_\Omega h\ dP_{|\mathcal{F}}=\int_\Omega h\ dP$$

Demonstratie:
Notam aplicatia identitate cu $i:(\Omega,\mathcal{K})\longrightarrow(\Omega,\mathcal{F})$. Rezulta ca i este masurabila si $P\circ i^{-1}=P_{|\mathcal{F}}$

$$\int_\Omega h\ dP_{|\mathcal{F}}=\int_\Omega h\ dP\circ i^{-1}=\int_\Omega h\circ i\ dP=\int_\Omega h\ dP$$

Teorema (existenta si unicitate)

Fie $(\Omega, \mathcal{K}, P)$, $\mathcal{F} \subset \mathcal{K}$, $\mathcal{F}$ corp borelian.

a) Daca X este o variabila aleatoare nenegativa, atunci exista o variabila aleatoare nenegativa $M(X \mid \mathcal{F})$ astfel incat

$$i) \ M(X \mid \mathcal{F}) \ \text{ este } \ \mathcal{F}\text{-masurabila}$$

$$ii) \ \int_A M(X \mid \mathcal{F})\, dP = \int_A X dP \ \ \forall A \in \mathcal{F}$$

In particular, daca X este integrabila rezulta ca $M(X \mid \mathcal{F})$ este integrabila.

$M(X \mid \mathcal{F})$ este unica $(P - a.s.)$ variabila aleatoare cu proprietatile $i)$ si $ii)$.

b) Daca X este o variabila aleatoare integrabila, atunci exista si este unica $(P - a.s.)$ o variabila aleatoare integrabila $M(X \mid \mathcal{F})$, cu proprietatile $i)$ si $ii)$.

Demonstratie:

a) :

- Demonstram intai unicitatea: Daca exista g_1, g_2 variabile aleatoare cu proprietatile $i)$ si $ii)$, rezulta

$$\int_A g_1 dP = \int_A g_2 dP \ \ \forall A \in \mathcal{F}$$

Dar g_1, g_2 sunt $\mathcal{F}-$masurabile. Rezulta $g_1 = g_2 \ P - a.s.$

- Fie X variabila aleatoare nenegativa si fie

$$\begin{aligned} \mu &: \mathcal{F} \longrightarrow \overline{R_+} \\ \mu(A) &= \int_A X dP \end{aligned}$$

μ este o masura $\sigma-$finita, absolut continua in raport cu $P_{|\mathcal{F}}$. Rezulta din teorema Radon - Nicodym ca exista o unica aplicatie

$$g : \Omega \longrightarrow \overline{R_+}$$

$\mathcal{F}-$masurabila, asa incat

$$\mu(A) = \int_A g dP_{|\mathcal{F}} \quad \forall A \in \mathcal{F}$$

Aplicam Lema:

$$\int_A g dP_{|\mathcal{F}} = \int_\Omega I_A \cdot g dP_{|\mathcal{F}} = \int_\Omega I_A \cdot g dP = \int_A g dP$$

Deci

$$\int_A X dP = \int_A g dP \quad \forall A \in \mathcal{F}$$

Vom nota aceasta unica aplicatie cu $g = M(X \mid \mathcal{F})$ si o vom numi "media lui X conditionata de $\mathcal{F}$".

b) :

Fie X variabila aleatoare integrabila. Atunci

$$X = X^+ - X^-,$$

cu X^+ si X^- pozitive, integrabile, $X^+ = \max\{X, 0\}$, $X^- = \max\{-X, 0\}$.

Din a), $(\exists)(!)\, M(X^+ \mid \mathcal{F}), M(X^- \mid \mathcal{F})$ variabile aleatoare nenegative, integrabile, cu proprietatile $i)$ si $ii)$. Luam

$$M(X \mid \mathcal{F}) = M\left(X^+ \mid \mathcal{F}\right) - M\left(X^- \mid \mathcal{F}\right),$$

care satisface prorpietatile din enuntul teoremei.

■

CAZURI PARTICULARE

- $A \in K,\ X = 1_A$. Atunci notam

$$M(1_A \mid \mathcal{F}) = P(A \mid \mathcal{F})$$

- Y variabila aleatoare, $\mathcal{F} = \mathcal{B}(Y) = Y^{-1}(\mathcal{B})$. Atunci notam

$$M(X \mid \mathcal{B}(Y)) = M(X \mid Y)$$

- $A \in K,\ X = 1_A$ si $\mathcal{F} = \mathcal{B}(Y)$. Atunci notam

$$M(1_A \mid \mathcal{B}(Y)) = P(A \mid Y)$$

VERSIUNE A MEDIEI CONDITIONATE

Fie X si Y variabile aleatoare, cu X nenegativa sau integrbila.

Se numeste versiune a mediei conditionate $M(X \mid Y)$ functia masurabila

$$M(X \mid Y = y) : R \longrightarrow R$$

cu proprietatea

$$M(X \mid Y = y) \circ Y = M(X \mid Y) \quad P - a.s.$$

Propozitie

Fie X si Y variabile aleatoare, cu X nenegativa sau integrabila. Functia masurabila $\varphi : R \longrightarrow R$ este versiune a mediei conditionate $M(X \mid Y)$ daca si numai daca

$$\int_B \varphi(y)\, dP \circ Y^{-1}(y) = \int_{Y^{-1}(B)} X dP, \quad \forall B \in \mathcal{B}$$

Demonstratie:

$$\begin{aligned} \varphi \circ Y &= M(X \mid Y) \quad P - a.s. \quad \Leftrightarrow \\ \int_A \varphi \circ Y dP &= \int_A M(X \mid Y)\, dP, \quad \forall A \in \mathcal{B}(Y) \end{aligned}$$

Dar $\mathcal{B}(Y) = Y^{-1}(\mathcal{B})$. Deci, pentru orice $B \in \mathcal{B}$

$$\int_B \varphi(y)\, dP \circ Y^{-1}(y) = \int_{Y^{-1}(B)} \varphi \circ Y dP = \int_{Y^{-1}(B)} M(X \mid Y)\, dP = \int_{Y^{-1}(B)} X dP$$

■

MODALITATI DE CALCUL PENTRU $M(X \mid Y = y)$

(a) Cazul repartitiilor discrete

Presupunem

$$\begin{aligned} P \circ Y^{-1} &= \sum_{k \in I} P(Y = a_k) \cdot \delta_{\{a_k\}} \\ P(Y = a_k) &> 0 \ \forall k, \ \sum_{k \in I} P(Y = a_k) = 1 \end{aligned}$$

cu I cel mult numarabila. Aratam ca

$$M(X \mid Y = a_k) = \frac{1}{P(Y = a_k)} \int\limits_{\{Y = a_k\}} X dP.$$

Notam cu φ o functie $\mathcal{B}$–masurabila, asa incat

$$\varphi(a_k) = \frac{1}{P(Y = a_k)} \int\limits_{\{Y = a_k\}} X dP, \quad k \in I$$

Notam suportul lui $P \circ Y^{-1}$ cu $A = \{a_k, k \in I\}$. Fie $B \in \mathcal{B}$. Avem

$$\begin{aligned}
\int\limits_B \varphi(y)\, dP \circ Y^{-1}(y) &= \int\limits_{B \cap A} \varphi(y)\, dP \circ Y^{-1}(y) = \sum_{a_k \in B \cap A} \varphi(a_k) \cdot P(Y = a_k) = \\
&= \sum_{a_k \in B \cap A} \int\limits_{\{Y = a_k\}} X dP = \int\limits_{Y^{-1}(B)} X dP
\end{aligned}$$

Aplicand propozitia anterioara, obtinem c.t.d.

Daca presupunem chiar mai mult, si anume ca (X, Y) este un vector aleator cu repartitie discreta

$$\begin{aligned}
P \circ (X, Y)^{-1} &= \sum_{x \in A'} \sum_{y \in A} p(x, y) \cdot \delta_{\{(x,y)\}} \\
A' &= \{a'_k,\ k \in I\} \\
A &= \{a_k,\ k \in I\}
\end{aligned}$$

atunci

$$M(X \mid Y = a_k) = \sum_{k \in I} a'_k \cdot \frac{P(X = a'_k, Y = a_k)}{P(Y = a_k)} = \sum_{k \in I} a'_k \cdot P(X = a'_k \mid Y = a_k)$$

(b) Cazul repartitiilor continue

Presupunem ca (X, Y) are densitatea de repartitie $f(x, y)$. Notam

$$f_Y(y) = \int\limits_R f(x, y)\, dx$$

Aratam ca

$$M(X \mid Y = y) = \int\limits_R x \cdot \frac{f(x, y)}{f_Y(y)} dx$$

Observam ca definitia este corecta pentru y cu $f_Y(y) > 0$. In punctele in care $f_y(y) = 0$ se ia $M(X \mid Y = y)$ egala cu o constanta arbitrara.

Notam functia masurabila

$$\varphi(y) = \int_R x \cdot \frac{f(x,y)}{f_Y(y)} dx$$

Fie $B \in \mathcal{B}$

$$\begin{aligned} \int_B \varphi(y)\, dP \circ Y^{-1}(y) &= \int_B \left(\int_R x \cdot \frac{f(x,y)}{f_Y(y)} dx \right) f_Y(y)\, dy = \\ &= \int_{R \times B} x \cdot f(x,y)\, dxdy = \int_{R \times R} x \cdot 1_B(y) \cdot f(x,y)\, dxdy = \\ &= \int_\Omega (1_B \circ Y) \cdot X dP = \int_{Y^{-1}(B)} X dP \end{aligned}$$

Aplicand propozitia anterioara, obtinem c.t.c.

■

Notatie (densitatea de repartitie conditionata a lui X)

$$f(x \mid y) = \frac{f(x,y)}{f_Y(y)}$$

$$M(X \mid Y = y) = \int_R x \cdot f(x \mid y)\, dx$$

Definitie

Fie vectorul aleator (X, Y) cu componente integrabile. Se numeste regresia lui X in Y functia

$$y \longrightarrow M(X \mid Y = y)$$

Regresia este liniara daca

$$M(X \mid Y = y) = a + by$$

Dreapta de regresie este data de ecuatia

$$x = a + by$$

REGRESIA LINIARA PENTRU MODELUL NORMAL BIDIMENSIONAL

Fie urmatorii parametri:

$$\boldsymbol{\mu} = (\mu_x, \mu_y)' \in R^2$$

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix},$$

Σ matrice simetrica, pozitiv definita.

Vectorul aleator $(X,Y)'$ are o repartitie normala bidimensionala $N(2;\boldsymbol{\mu},\boldsymbol{\Sigma})$ daca are densitatea de repartitie

$$f(x.y) = \frac{1}{2\pi\sqrt{\sigma_x^2\sigma_y^2(1-\rho^2)}}\cdot$$

$$\cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}$$

Proprietatea 1

Repartitiile marginale ale lui $N(2;\boldsymbol{\mu},\boldsymbol{\Sigma})$ sunt

$$P\circ X^{-1} = N(\mu_x,\sigma_x^2), \quad P\circ Y^{-1} = N(\mu_y,\sigma_y^2)$$

Demonstratie:
Adunand si scazand $\rho^2\left(\frac{y-\mu_y}{\sigma_y}\right)^2$ la exponent obtinem

$$f(x.y) = \frac{1}{\sqrt{2\pi\sigma_y^2}\sqrt{2\pi\sigma_x^2(1-\rho^2)}}\cdot$$

$$\cdot \exp\left\{\frac{1}{2\sigma_x^2(1-\rho^2)}\left[x-\left(\mu_x+\rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)\right)\right]^2 - \frac{1}{2\sigma_y^2}(y-\mu_y)^2\right\}$$

Repartitia marginala a lui Y este

$$f_Y(y) = \int_R f(x,y)\,dx = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left\{-\frac{1}{2\sigma_y^2}(y-\mu_y)^2\right\}$$

Analog se obtine si repartitia marginala a lui X.

Proprietatea 2

Repartitia lui X conditionata de Y este normala,

$$N\left(\mu_x+\rho\frac{\sigma_x}{\sigma_y}\left(y-\mu_y\right);\sigma_x^2\left(1-\rho^2\right)\right)$$

Proprietatea rezulta imediat, calculand

$$f\left(x\mid y\right)=\frac{f\left(x,y\right)}{f_Y\left(y\right)}$$

Corolar

$$\begin{aligned}M\left(X\mid Y=y\right) &= \mu_x+\rho\frac{\sigma_x}{\sigma_y}\left(y-\mu_y\right)\\ D^2\left(X\mid Y=y\right) &= \sigma_x^2\left(1-\rho^2\right)\end{aligned}$$

Rezulta ca, pentru modelul normal bidimensional, regresia lui X in Y este liniara, iar ecuatia dreptei de regresie este

$$x=\left(\mu_x-\rho\frac{\sigma_x}{\sigma_y}\mu_y\right)+\rho\frac{\sigma_x}{\sigma_y}\cdot y$$

ESTIMAREA PARAMETRILOR DREPTEI DE REGRESIE

(a) Fara specificarea repartitiei lui (X,Y)

Fie vectorul aleator $(X,Y)'$ pentru care facem ipoteza

$$M\left(X\mid Y=y\right)=a+by$$

astfel incat ecuatia dreptei de regresie este $x=a+by$.

Fie observatiile $\left(X_i,Y_i\right)',=1,...,n,$ care sunt vectori aleatori independenti, identic repartizati ca si $(X,Y)'$ si fie $\left(x_i,y_i\right)'$ $i=1,...,n$ datele statistice corespunzatoare.

$$M\left(X_i\mid Y_1=y_1,...,Y_i=y_i,...,Y_n=y_n\right)=M\left(X_i\mid Y_i=y_i\right)=a+by_i$$

Lucrand cu repartitia conditionata, apare modelul liniar $n-$dimensional

$$X_i=\left(a+by_i\right)+Z_i,\quad i=1,...,n$$

unde $Z_1, ..., Z_n$ sunt variabile aleatoare indep, de medie zero. Aplicam metoda celor mai mici patrate:

$$SS(a,b) = \sum_{i=1}^{n} (x_i - a - by_i)^2$$

Sistemul de ecuatii normale $\frac{\partial SS}{\partial a} = \frac{\partial SS}{\partial b} = 0$ se scrie sub forma

$$\begin{cases} na + b\sum\limits_{i=1}^{n} y_i = \sum\limits_{i=1}^{n} x_i \\ a\sum\limits_{i=1}^{n} y_i + b\sum\limits_{i=1}^{n} y_i^2 = \sum\limits_{i=1}^{n} x_i y_i \end{cases}$$

Determinantul matricii sistemului liniar este egal cu zero doar in cazul degenerat (cand toti $y_i = \overline{y}, \forall i$), caz care apare cu probabilitatea zero:

$$\Delta = \begin{vmatrix} n & \sum\limits_{i=1}^{n} y_i \\ \sum\limits_{i=1}^{n} y_i & \sum\limits_{i=1}^{n} y_i^2 \end{vmatrix} = n\sum_{i=1}^{n} y_i^2 - (n\overline{y})^2 = n\sum_{i=1}^{n} (y_i - \overline{y})^2 > 0$$

Notatie:

$$\begin{aligned} s_x^2 &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 \\ s_y^2 &= \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2 \\ s_{xy} &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \\ r &= \frac{s_{xy}}{s_x s_y} \end{aligned}$$

Solutia unica a sistemului de ecuatii normale este

$$\begin{aligned} \widehat{b} &= \frac{s_{xy}}{s_y^2} = r\frac{s_x}{s_y} \\ \widehat{a} &= \overline{x} - \widehat{b} \cdot \overline{y} \end{aligned}$$

Obtinem dreapta de regresie de selectie

$$x - \overline{x} = r\frac{s_x}{s_y}(y - \overline{y})$$

Estimatorii obtinuti prin metoda celor mai mici patrate,

$$\begin{aligned}
\widehat{b}\left(X_1,...,X_n\right) &= \frac{1}{\sum\limits_{i=1}^{n}\left(y_i-\overline{y}\right)^2}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(y_i-\overline{y}\right)=\frac{1}{\sum\limits_{i=1}^{n}\left(y_i-\overline{y}\right)^2}\sum_{i=1}^{n}X_i\left(y_i-\overline{y}\right)\\
\widehat{a}\left(X_1,...,X_n\right) &= \overline{X}-\widehat{b}\left(X_1,...,X_n\right)\cdot\overline{y}
\end{aligned}$$

sunt nedeplasati (medierea conditionata):

$$\begin{aligned}
M\left(\widehat{b}\mid Y_1=y_1,...,Y_n=y_n\right) &= b\\
M\left(\widehat{a}\mid Y_1=y_1,...,Y_n=y_n\right) &= a
\end{aligned}$$

Putem calcula valoarea minima a sumei abaterilor patratice,

$$SS_{\min}=\sum_{i=1}^{n}\left(x_i-\widehat{a}-\widehat{b}y_i\right)^2\overset{notat}{=}SS_{resid}$$

(b) Cu specificarea repartitiei normale a lui (X,Y)

Fie vectorul aleator $(X,Y)'$ pentru care facem ipoteza ca urmaza o repartitie normala bidimensionala $N\left(2;\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$. Utilizand proprietatile modelului, avem

$$D^2\left(X_i\mid Y_1=y_1,...,Y_n=y_n\right)=\sigma_x^2\left(1-\rho^2\right),\quad i=1,...,n$$

Proprietatea 3.

Variabila aleatoare

$$SS_{resid}=\sum_{i=1}^{n}\left(X_i-\widehat{a}-\widehat{b}y_i\right)^2$$

are proprietatea

$$\frac{1}{\sigma_x^2\left(1-\rho^2\right)}\cdot SS_{resid}\quad\sim\quad\chi^2\left(n-2\right)$$

Rezulta din Proprietatea 8 de la "Estimarea parametrilor" (metoda celor mai mici patrate).

In continuare facem o analiza a surselor de variabilitate ale datelor, utilizand modelul regresiei liniare (ANOVA pentru dreapta de regresie)

In acest moment dispunem de urmatoarele valori:

- $y_i,\ i = 1, .., n,$ valorile observate ale covariatei (ale variabilei "cauza")
- $x_i,\ i = 1, ..., n,$ valorile observate ale variablei raspuns ("efect")
- $\widehat{x_i} = \widehat{a} + \widehat{b} \cdot y_i,\ i = 1, ..., n,$ predictorii dati de modelul regresiei liniare (fitted values)
- $x_i - \widehat{x_i},\ i = 1, ..., n,$ reziduuri

Introducem urmatoarele "sume de abateri patratice" (sum of squares):

$$SS_{resid} = \sum_{i=1}^{n} (x_i - \widehat{x_i})^2 = \sum_{i=1}^{n} \left(x_i - \widehat{a} - \widehat{b} y_i\right)^2$$

$$SS_{regresie} = \sum_{i=1}^{n} (\widehat{x_i} - \overline{x})^2$$

$$SS_{total} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

(vom utiliza aceste notatii atat pentru valorile numerice calculate ale $SS-$urilor, cat si pentru variabilele aleatoare corespunzatoare)

Proprietatea 4 (ecuatia ANOVA)

$$SS_{total} = SS_{regresie} + SS_{resid}$$

Demonstratie:

$$\begin{aligned} SS_{total} &= \sum_{i=1}^{n} (x_i - \widehat{x_i} + \widehat{x_i} - \overline{x})^2 = \\ &= SS_{resid} + SS_{regresie} + 2 \sum_{i=1}^{n} (x_i - \widehat{x_i}) (\widehat{x_i} - \overline{x}) \end{aligned}$$

$$\sum_{i=1}^{n}(x_i-\widehat{x}_i)(\widehat{x}_i-\overline{x})=\sum_{i=1}^{n}\left(x_i-\widehat{a}-\widehat{b}y_i\right)\left(\widehat{a}+\widehat{b}y_i-\overline{x}\right)=$$

$$=\sum_{i=1}^{n}\left(x_i-\overline{x}+\widehat{b}\overline{y}-\widehat{b}y_i\right)\left(\overline{x}-\widehat{b}\overline{y}+\widehat{b}y_i-\overline{x}\right)=$$

$$\begin{aligned} &= -\widehat{b}\sum_{i=1}^{n}\left[(x_i-\overline{x})-\widehat{b}\,(y_i-\overline{y})\right](y_i-\overline{y})= \\ &= -\widehat{b}\left\{ns_{xy}-\frac{s_{xy}}{s_y^2}\cdot ns_y^2\right\}=0\end{aligned}$$

■

Cunoastem repartitia variabilei aleatoare $\frac{1}{\sigma_x^2(1-\rho^2)}\cdot SS_{resid}$ (proprietatea 3).

Ne propunem sa stabilim repartitiile variabilelor aleatoare

$$\frac{1}{\sigma_x^2\left(1-\rho^2\right)}\cdot SS_{regresie} \text{ si } \frac{1}{\sigma_x^2\left(1-\rho^2\right)}\cdot SS_{total},$$

in situatia in care am avea

$$b=0$$

AUXILIAR: TEOREMA LUI COCHRAN

Propozitie (rezultat algebric, pentru variabile scalare)

Fie vectorul $\mathbf{y} = (y_1, ..., y_N)' \in R^N$. Presupunem ca suma de patrate

$$\sum_{i=1}^{N} y_i^2$$

se descompune in suma a m forme patratice

$$q_j = \sum_{\alpha,\beta=1}^{N} a_{\alpha\beta}^j \cdot y_\alpha y_\beta, \quad j = 1, ...m,$$

$$\sum_{i=1}^{N} y_i^2 = \sum_{j=1}^{m} q_j,$$

unde, pentru orice $j = 1, ..., m$,

$$A_j = \left\| a_{\alpha\beta}^j \right\|_{\alpha,\beta=1,...,N}$$

este matrice simetrica, de rang r_j.

O conditie necesara si suficienta ca sa existe o transformare ortogonala

$$\mathbf{z} = B\mathbf{y}$$

asa incat

$$q_j = \sum_{k=r_1+...+r_{j-1}+1}^{r_1+...+r_j} z_k^2, \quad j = 1, ...m$$

este ca

$$r_1 + ... + r_m = N$$

Demonstratie:

" $\Longrightarrow$ "

Presupunem ca exista transformarea $z = By,\ B'B = \mathbf{I}$, cu proprietatea din enunt. Transformarea

$$(y_1, ..., y_N) \longrightarrow (z_1, ..., z_{r_1+...+r_m})$$

trebuie sa fie nesingulara. Rezulta

$$r_1 + ... + r_m \leq N$$

Scriem matriceal relatia de descompunere din ipoteza

$$\mathbf{y}'\mathbf{y} = \sum_{j=1}^{m} \mathbf{y}' A_j \mathbf{y}$$

Rezulta

$$\sum_{j=1}^{m} A_j = \mathbf{I}$$

$$rang\left(\sum_{j=1}^{m} A_j\right) = N$$

Dar

$$rang\left(\sum_{j=1}^{m} A_j\right) \leq \sum_{j=1}^{m} rang\left(A_j\right) = \sum_{j=1}^{m} r_j$$

Deci

$$N \leq r_1 + ... + r_m$$

" $\Longleftarrow$ "

Vom construi matricea B intr-o forma partitionata,

$$B = \begin{pmatrix} B_1 \\ ..... \\ \vdots \\ \vdots \\ ..... \\ B_m \end{pmatrix}$$

– Pentru $i = 1$:

A_1 este $N \times N-$dimensionala, simetrica, de rang r_1. Rezulta ca exista o matrice nesingulara D_0 asa incat

$$D_0 A_1 D_0' = \begin{bmatrix} \mathbf{I}_q & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_{r_1 - q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

unde q este numarul de valori proprii pozitive ale lui A_1 si $(r_1 - q)$ este numarul de valori proprii negative ale lui A_1.

Notam

$$\begin{aligned} D' &= D_0^{-1} \\ D &= \|d_{\alpha\beta}\| \end{aligned}$$

si avem

$$A_1 = D' \begin{bmatrix} \mathbf{I}_q & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_{r_1-q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} D$$

Retinem

$$b^{(1)}_{\alpha\beta} = d_{\alpha\beta}, \quad \alpha = 1, ..., r_1;\ \beta = 1, ..., N$$

$$B_1 = \left\| b^{(1)}_{\alpha\beta} \right\|_{\alpha=1,...,r_1;\ \beta=1,...,N}$$

Consideram transformarea liniara definita de aceasta matrice,

$$\begin{aligned} z_\alpha &= \sum_{\beta=1}^{N} b^{(1)}_{\alpha\beta} y_\beta, \quad \alpha = 1, ..., r_1 \\ \mathbf{z}^{(1)} &= (z_1, ..., z_{r_1})' = B_1 \mathbf{y} \end{aligned}$$

Atunci

$$\begin{aligned} q_1 &= \mathbf{y}' A_1 \mathbf{y} = \mathbf{y}' D' \begin{bmatrix} \mathbf{I}_q & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_{r_1-q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} D\mathbf{y} = \\ &= z_1^2 + ... + z_q^2 - z_{q+1}^2 - ... - z_{r_1}^2 \end{aligned}$$

$$q_1 = \sum_{\alpha=1}^{r_1} c_\alpha z_\alpha^2, \quad c_\alpha \in \{-1, 1\}.$$

– Pentru i arbitrar:

In mod analog obtinem

$$z_\alpha = \sum_{\beta=1}^{N} b^{(i)}_{\alpha\beta} y_\beta,\ \alpha = r_1 + ... + r_{i-1} + 1, ..., r_1 + ... + r_i$$

$$B_i = \left\| b^{(i)}_{\alpha\beta} \right\|_{\substack{\alpha=r_1+...+r_{i-1}+1,...,r_1+...+r_i; \\ \beta=1,...,N}}$$

$$q_i = \sum_{\alpha=r_1+...+r_{i-1}+1}^{r_1+...+r_i} c_\alpha z_\alpha^2, \quad c_\alpha \in \{-1, 1\}.$$

– Atunci

$$\sum_{i=1}^{m} q_i = \sum_{\alpha=1}^{N} c_\alpha z_\alpha^2, \quad c_\alpha \in \{-1, 1\}.$$

Dar

$$\sum_{i=1}^{m} q_i = \mathbf{y}'\mathbf{y} > 0 \ \ \forall \mathbf{y} \neq \mathbf{0}$$

Deci $\sum\limits_{\alpha=1}^{N} c_\alpha z_\alpha^2$ este pozitiv definita si deci $c_\alpha = 1 \ \forall \alpha = 1, ..., N$. Am obtinut

$$q_i = \sum_{\alpha=r_1+...+r_{i-1}+1}^{r_1+...+r_i} z_\alpha^2, \ \ i = 1, ..., m$$

Formam matricea $B = \|b_{\alpha\beta}\|$, de dimensiune $N \times N$, partitionata in componentele B_i. Avem

$$z_\alpha = \sum_{\beta=1}^{N} b_{\alpha\beta} \cdot y_\beta, \ \ \alpha = 1, ..., N$$

$$\sum_{\alpha=1}^{N} y_\alpha^2 = \sum_{\alpha=1}^{N} z_\alpha^2$$

Ultima relatie este echivalenta cu

$$\mathbf{y}'\mathbf{y} = (B\mathbf{y})'(B\mathbf{y}) = \mathbf{y}'B'B\mathbf{y},$$

deci $B'B = \mathbf{I}$, adica transformarea este ortogonala.
■

TEOREMA LUI COCHRAN

Fie $Y_1, ..., Y_N$ variabile aleatoare independendente, identic repartizate $N(0,1)$. Notam $\mathbf{Y} = (Y_1, ..., Y_N)'$. Presupunem ca $\mathbf{Y}'\mathbf{Y}$ se descompune in suma a m forme patratice

$$Q_i = \mathbf{Y}'A_i\mathbf{Y}, i = 1, ..., m,$$

cu $A_i = \left\|a_{\alpha\beta}^{(i)}\right\|_{\alpha,\beta=1,...,N}$ matrici simetrice, de rang r_i, $i = 1, ..., m$, asa incat

$$\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{m} Q_i.$$

O conditie necesara si suficienta ca variabilele aleatoare Q_i sa fie repartizate $\chi^2(r_i)$, $i = 1, ..., m$ si Q_i sa fie independenta de Q_j pentru orice $i \neq j$ este ca

$$r_1 + ... + r_m = N$$

Demonstratie

" $\Longrightarrow$ "

Aceasta implicatie rezulta cu aceleasi argumente ca cele utilizate in demonstrarea implicatiei similare din rezultatul algebric.

" $\Longleftarrow$ "

Folosind rezultatul algebric rezulta ca exista o transformare $\mathbf{Z} = B\mathbf{Y}$, $B = \|b_{\alpha\beta}\|$, asa incat

$$Q_i = \sum_{\alpha=r_1+...+r_{i-1}+1}^{r_1+...+r_i} Z_\alpha^2, \quad i = 1, ..., m$$

$$Z_\alpha = \sum_{\beta=1}^{N} b_{\alpha\beta} \cdot Y_\beta, \quad \alpha = 1, ..., N$$

Din proprietatile combinatiilor liniare de variabile independente, repartizate normal rezulta ca Z_α este repartizata $N(0,1)$ pentru orice $\alpha = 1, ..., N$ si $Z_1, ..., Z_N$ sunt independente. Atunci, din avem $Q_i \sim \chi^2(r_i)$, $i = 1, ..., m$ si, din asociativitatea independentei, Q_i este independenta de Q_j pentru orice $i \neq j$.

■

Corolar 1

Fie $Y_1, ..., Y_k$ variabile aleatoare independente, identic repartizate $N(0,1)$. Notam $\mathbf{Y} = (Y_1, ..., Y_k)'$. O conditie necesara si suficienta ca $\mathbf{Y}'A\mathbf{Y}$ sa fie repartizata χ^2 este ca $A^2 = A$, caz in care numarul de grade de libertate este egal cu $rang(A)$.

Corolar 2.

Fie $Y_1, ..., Y_k$ variabile aleatoare independente, identic repartizate $N(0,1)$. Notam $\mathbf{Y} = (Y_1, ..., Y_k)'$. Presupunem ca $\mathbf{Y}'\mathbf{Y} = Q_1 + Q_2$, unde

$$Q_1 = \mathbf{Y}'A\mathbf{Y} \sim \chi^2(r)$$

Atunci $Q_2 \sim \chi^2(k-r)$.

Corolar 3.

Fie $Y_1, ..., Y_k$ variabile aleatoare independente, identic repartizate $N(0,1)$. Notam $\mathbf{Y} = (Y_1, ..., Y_k)'$. Fie Q, Q_1, Q_2 forme patratice in $\mathbf{Y}$ asa incat $Q = Q_1 + Q_2,\ Q \sim \chi^2(a),\ Q_1 \sim \chi^2(b)$. Atunci $Q_2 \sim \chi^2(a-b)$.

Corolar 4.

Fie $Y_1, ..., Y_k$ variabile aleatoare independente, identic repartizate $N(0,1)$. Notam $\mathbf{Y} = (Y_1, ..., Y_k)'$. Fie $\mathbf{Y}'A_1\mathbf{Y} \sim \boldsymbol{\chi}^2(a)$ si $\mathbf{Y}'A_2\mathbf{Y} \sim \boldsymbol{\chi}^2(b)$. O conditie necesara si suficienta ca cele doua forme patratice sa fie independente este ca $A_1A_2 = \mathbf{0}$.

==

Revenim la ANOVA pentru dreapta de regresie:

Proprietatea 5.

Daca $b = 0$, atunci

$$\frac{1}{\sigma_x^2(1-\rho^2)} \cdot SS_{regresie} \sim \chi^2(1)$$

$$\frac{1}{\sigma_x^2(1-\rho^2)} \cdot SS_{total} \sim \chi^2(n-1)$$

iar variabilele $\frac{1}{\sigma_x^2(1-\rho^2)} \cdot SS_{regresie}$ si $\frac{1}{\sigma_x^2(1-\rho^2)} \cdot SS_{resid}$ sunt independente (in raport cu repartitia conditionata).

Demonstratie:

Daca $b = 0$, atunci repartitia conditionata a lui X_i este $N\left(a, \sigma_x^2\left(1-\rho^2\right)\right), \forall i$.

(i) Ne ocupam intai de $SS_{regresie}$

$$\begin{aligned} SS_{regresie} &= \sum_{i=1}^{n}\left(\widehat{X_i} - \overline{X}\right)^2 = \sum_{i=1}^{n}\left(\widehat{a} + \widehat{b}y_i - \overline{X}\right)^2 = \sum_{i=1}^{n}\left(\overline{X} - \widehat{b}\overline{y} + \widehat{b}y_i - \overline{X}\right)^2 = \\ &= \left(\widehat{b}\right)^2 \sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2}\left(\sum_{i=1}^{n}(y_i - \overline{y})X_i\right)^2, \end{aligned}$$

$$SS_{regresie} = \frac{1}{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2}(X_1, ..., X_n) \cdot B \cdot \begin{pmatrix} X_1 \\ \vdots \\ \vdots \\ X_n \end{pmatrix}$$

unde

$$B = \left\| (y_i - \overline{y})(y_j - \overline{y}) \right\|_{i,j=1,...,n} \overset{notat}{=} \|b_{ij}\|$$

Presupunem ca nu suntem in cazul degenerat si observam ca pentru $1 \le i < j \le n$ avem

$$\frac{y_j - \overline{y}}{y_i - \overline{y}} \cdot \begin{pmatrix} b_{1i} \\ \vdots \\ \vdots \\ b_{ni} \end{pmatrix} - \begin{pmatrix} b_{1j} \\ \vdots \\ \vdots \\ b_{nj} \end{pmatrix} = \mathbf{0}$$

Deci $rang(B) = n - (n-1) = 1$. Prin calcul direct se verifica

$$\left(\frac{1}{ns_y^2} B \right)^2 = \frac{1}{ns_y^2} B$$

Cum

$$\frac{1}{\sigma_x^2 (1-\rho^2)} SS_{regresie} = \left(\frac{1}{\sqrt{\sigma_x^2 (1-\rho^2)}} X \right)' \cdot \frac{1}{ns_y^2} B \cdot \left(\frac{1}{\sqrt{\sigma_x^2 (1-\rho^2)}} X \right)$$

putem aplica Corolarul 1 si obtinem faptul ca

$$\frac{1}{\sigma_x^2 (1-\rho^2)} SS_{regresie} \sim \chi^2 (1).$$

(ii) Continuam cu variabila aleatoare SS_{total}:

$$SS_{total} = \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$$

Putem scrie

$$SS_{total} = \sum_{i=1}^{n} \left(X_i - \overline{X} \right) X_i = \frac{1}{n^2} (X_1, ..., X_n) \cdot A \cdot \begin{pmatrix} X_1 \\ \vdots \\ \vdots \\ X_n \end{pmatrix}$$

unde $A = \|a_{ij}\|_{i,j=1,...,n}$, $a_{ii} = n(n-1)$, $a_{ij} = -n$ pentru $i \neq j$.

Aplicam succesiv transformarile elementare pe coloane ($C_i \longrightarrow C_i - C_{i+1}$, $i = 1, ..., n-1$) si obtinem

$$\frac{1}{n^2} \cdot A = \begin{pmatrix} 0 & 0 & ..... & 0 & -1/n \\ -1 & 1 & ..... & 0 & -1/n \\ 0 & -1 & ..... & 0 & -1/n \\ ..... & ..... & ..... & ..... & ..... \\ 0 & 0 & ..... & -1 & -1/n \\ 0 & 0 & ..... & 1 & 1-1/n \end{pmatrix}$$

Notam $\widetilde{C}_1, ..., \widetilde{C}_n$ coloanele acestei matrice si observam ca

$$\frac{1}{n}\widetilde{C}_1 + \frac{2}{n}\widetilde{C}_2 + ... + \frac{n-1}{n}\widetilde{C}_{n-1} + \widetilde{C}_n = \mathbf{0}$$

iar $\widetilde{C}_1, ..., \widetilde{C}_n$ sunt vectori liniar independenti. Deci $rang\left(\frac{1}{n^2}A\right) = n-1.$

Rezulta ca

$$\frac{1}{\sigma_x^2\left(1-\rho^2\right)} SS_{total} \ \sim \ \chi^2\left(n-1\right).$$

(iii) Prin calcul direct se verifica relatia

$$\left(\frac{1}{n^2}A - \frac{1}{ns_y^2}B\right) \cdot \frac{1}{ns_y^2}B = \mathbf{0}$$

Cum avem si

$$\frac{1}{\sigma_x^2\left(1-\rho^2\right)} SS_{resid} = \frac{1}{\sigma_x^2\left(1-\rho^2\right)}\left(SS_{total} - SS_{regresie}\right),$$

$$\frac{1}{\sigma_x^2\left(1-\rho^2\right)} SS_{regresie} = \frac{1}{\sigma_x^2\left(1-\rho^2\right)} \cdot \frac{1}{s_y^2}\left(X_1, ..., X_n\right) \cdot B \cdot \begin{pmatrix} X_1 \\ : \\ : \\ X_n \end{pmatrix} \sim \chi^2\left(1\right),$$

$$\frac{1}{\sigma_x^2\left(1-\rho^2\right)} SS_{resid} = \frac{1}{\sigma_x^2\left(1-\rho^2\right)}\left(X_1, ..., X_n\right) \cdot \left(\frac{1}{n^2}A - \frac{1}{ns_y^2}B\right) \cdot \begin{pmatrix} X_1 \\ : \\ : \\ X_n \end{pmatrix} \sim \chi^2\left(n-2\right),$$

putem aplica Corolar 4 si obtinem independenta variabilelor $\frac{1}{\sigma_x^2(1-\rho^2)} SS_{regresie}$ si $\frac{1}{\sigma_x^2(1-\rho^2)} SS_{resid.}$

■

TABELUL ANOVA PENTRU DREAPTA DE REGRESIE

Sursa de variabilitate	SS	Grade de libertate	$\overline{SS}$ (mean SS)
abaterile predictorilor de la $\overline{x}$	$SS_{regresie}$	1	$\overline{SS_{regresie}} = SS_{regresie}$
reziduuri aleatoare	SS_{resid}	$n-2$	$\overline{SS_{resid}} = \frac{1}{n-2} SS_{resid}$
abaterile observatiilor de la $\overline{x}$	SS_{total}	$n-1$	

FUNCTII IN R

$> cauza \longleftarrow c\,(y_1, ..., y_n)$
$> efect \longleftarrow c\,(x_1, ..., x_n)$
$> model \longleftarrow lm\,(efect \sim cauza)$

Functia lm returneaza

- coefficients $\left(\widehat{a}, \widehat{b}\right)$
- summary: statistica descriptiva pentru reziduuri

$$\{x_i - \widehat{x_i}, i = 1, ..., n\}$$

$> anova($model$)$

Functia $anova$ returneaza tabelul ANOVA si teste pentru ipoteza $\{b = 0\}$ despre care discutam in ultima parte a cursului.

APLICATIE

longley {datasets}   R Documentation
Longley's Economic Regression Data

Description
A macroeconomic data set which provides a well-known example for a highly collinear regression.

Usage
longley

Format
A data frame with 7 economical variables, observed yearly from 1947 to 1962 (n=16).
GNP.deflator: GNP implicit price deflator (1954=100)
GNP: Gross National Product.
Unemployed: number of unemployed.
Armed.Forces: number of people in the armed forces.
Population: 'noninstitutionalized' population >= 14 years of age.
Year: the year (time).
Employed: number of people employed.

The regression lm(Employed ~.) is known to be highly collinear.
Alegem ca variabila raspuns "Employed", cu covariata "Population"

```
> X <- longley[, "Employed"]
> Y <- longley[,"Population"]
> model1<-lm(X~Y2)
> model1
Call:
lm(formula = X ~Y)
Coefficients:
(Intercept)...........Y
 8.3807 .........0.4849
```

> summary(model1)
Call:
lm(formula = X ~Y2)
Residuals:

Min........ ....1Q.......... Median....... 3Q ....Max
-1.4362 ...-0.9740 ....0.2021...... 0.5531 ......1.9048

Coefficients:

....................Estimate .....Std. Error...... t value.......Pr(>|t|)
(Intercept) ...8.3807 ....4.4224 ....1.895 ....0.079 .
Y................ 0.4849 ....0.0376 ....12.896 .....3.69e-09

Residual standard error: 1.013 on 14 degrees of freedom
Multiple R-Squared: 0.9224, Adjusted R-squared: 0.9168
F-statistic: 166.3 on 1 and 14 DF,
p-value: 3.693e-09

p-value < 0.05, deci modelul regresiei liniare este corect

> anova(model1)
Analysis of Variance Table

Response: X
...................Df...... Sum Sq........Mean Sq ....F value........Pr(>F)

Y........ ....1....... 170.643 ......170.643 ....166.30 ......3.693e-09
Residuals ...14 ......14.366 ....1.026