REGRESIA LINIARĂ MULTIPLĂ

1. Regresia prin origine
2. Prezentarea modelului liniar multiplu
3. Estimarea parametrilor modelului liniar multiplu
4. Testarea parametrilor modelului liniar multiplu

Regresia prin origine (I)

- Situaţii în care am putea construi un model de regresie prin origine:
  - În urma testării parametrilor modelului, parametrul β_0 are o valoare nesemnificativă statistic, iar parametrul β_1 este semnificativ
  - Există suport teoretic care să impună estimarea unui model care trece prin origine – lipsa influenţei variabilei independente conduce la o medie zero pentru variabila dependentă (analiza de cost, legătura dintre lungimea şi greutatea frunzelor unui copac).
- În cazul modelului de regresie $Y = \beta_1 X + \varepsilon$ aplicarea metodei celor mai mici pătrate se simplifică.
- Problema de minim care trebuie rezolvată este

$$S = \sum_i (y_i - \hat{\beta}_1 x_i)^2 = min.$$

Condiția de extrem conduce la ecuația:

$$2\sum_i (y_i - \hat{\beta}_1 x_i)(-x_i) = 0 \quad \text{sau} \quad \hat{\beta}_1 \sum_i x_i^2 = \sum_i y_i x_i \quad \text{sau} \quad \boxed{\hat{\beta}_1 = \frac{\sum_i y_i x_i}{\sum_i x_i^2}}.$$

Estimatorul $\hat{\beta}_1$ este nedeplasat

Avem n-1 grade de libertate

Probleme ale utilizării în practică:

- Suma erorilor nu mai este zero;
- R^2 poate fi negativ sau poate avea o valoare foarte mare, prin urmare interpretarea acestuia nu mai are sens. Se utilizează o variantă a lui R^2, şi anume:

$$r^2 = \frac{(\sum_i y_i x_i)^2}{\sum_i x_i^2 \sum_i y_i^2}, \text{ unde } 0<r^2<1.$$

Aceste probleme dispar dacă modelul de regresie liniară are variabilele standardizate. În acest caz, panta dreptei de regresie are aceeaşi valoare cu coeficientul de corelație Pearson.

Modelul liniar multiplu

Forma generală a modelului liniar multiplu este dată prin relația:

$$Y = M\left(Y / X_1, \ldots X_i, \ldots, X_p\right) + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

unde:

- Y - variabila dependentă;
- $X_1, X_2, \ldots, X_i, \ldots, X_p$ - variabile independente (predictori);
- **ε** - variabilă reziduu de modelare (variabila aleatoare);
- β_i - parametrii modelului de regresie
- k - numărul de parametri din model, **k=p+1**.

Exemplu: Pentru un eşantion de 50 de mărci de cereale, se poate studia legătura dintre ratingul acordat de consumatori unei mărci de cereale şi factorii de influență (nr. de calorii, de grame de grăsimi, de zahăr, de fibre, etc.)

Cei ***k*** parametri ai modelului liniar multiplu au următoarea semnificație:

β_0 – valoarea medie a variabilei dependente Y, în condițiile în care influența variabilelor independente ar fi nulă;

$\beta_i = \frac{\partial Y}{\partial X_i}, i = \overline{1, p}$ - variația absolută a variabilei

dependente la o variaţie absolută cu o unitate a variabilei independente X_i, în condițiile în care influența celorlalte variabile independente este menținută constantă. Arată influența parțială a fiecărei variabile independente asupra variabilei dependente.

Ipotezele modelului clasic de regresie

- variabilele independente sunt ***nestochastice***
- **normalitatea erorilor :** $\varepsilon_i \sim N(0,\sigma^2)$
- **homoscedasticitate**: $V(\varepsilon_i) = M(\varepsilon_i^2) = \sigma^2$
- **necorelarea erorilor**: $cov(\varepsilon_i, \varepsilon_j) = 0$
- **lipsa corelației dintre variabilele independente şi variabila eroare**
- **lipsa coliniarității** sau a unei legături liniare între variabilele independente

Estimarea parametrilor modelului multiplu liniar

Se consideră modelul de regresie liniară multiplă cu două variabile independente:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

La nivelul unui eşantion, modelul devine:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\varepsilon}_i \quad \text{sau} \quad y_i = \hat{y}_i + \hat{\varepsilon}_i$$

Rezultă

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}$$

Estimarea parametrilor modelului prin metoda celor mai mici pătrate presupune respectarea condiției:

$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = minim$, adică $\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2 = \min$

Pentru satisfacerea condiției MCMMP trebuie ca derivatele parțiale de ordin I în raport cu coeficienții modelului să se anuleze. Astfel se va obține un sistem de 2+1=3 ecuații cu 3 necunoscute.

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} = \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{1i} x_{2i} = \sum_{i=1}^{n} y_i x_{1i} \\ \hat{\beta}_0 \sum_{i=1}^{n} x_{2i} + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} x_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i}^2 = \sum_{i=1}^{n} y_i x_{2i} \end{cases}$$