Régression multiple

Soit la variable Y_i à expliquer par plusieurs variables quantitatives X_j:

Y : variable à expliquer, quantitative

$X_1, X_2, \dots X_K$: variables explicatives, quantitatives

le Modèle s'écrit comme suit:

$$Y_{ij} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$$

ε_i : terme aléatoire les erreurs

les Hypothèses à mettre en évidence pour obtenir ce modèle:

$$\varepsilon_i \rightsquigarrow N(0, \sigma_i)$$

$\varepsilon_1, \varepsilon_2 \dots \varepsilon_n$: ne sont pas corrélés entre elles.

on a: i : les individus $\quad j = 1 \dots k$ les variables

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \dots + \beta_k X_{1k} + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \dots + \beta_k X_{2k} + \varepsilon_2$$

$$\vdots$$

$$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \dots + \beta_k X_{nk} + \varepsilon_n$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ | \\ y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ | \\ \beta_k \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ | & x_{21} & x_{22} & & x_{2k} \\ 1 & x_{n1} & x_{n2} & & x_{nk} \end{pmatrix}$$

$(n,1)$ $\quad (k+1,1)$ $\quad (k+1, n)$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ | \\ \varepsilon_n \end{pmatrix}$$

$(n,1)$

le modèle s'écrit sur cette forme :

$$Y = \beta X + \varepsilon$$

on cherche l'ajustement qui minimise en β :

$$Q(\beta) = \sum \varepsilon_i^2 = \varepsilon' \varepsilon$$

ε $(n,1)$ donc pour $\varepsilon\varepsilon$ existe il faut $\varepsilon'\varepsilon$ $(1,n)(n,1)$ =

$$\varepsilon = Y - \beta X$$

$$\varepsilon' \varepsilon = (Y - \beta X)' (Y - \beta X)$$

$$Q(\beta) = Y'Y - Y'X\beta - \beta' X' Y + \beta' X' X \beta$$

on a $Y'X\beta = \beta' X' Y$

$$Q(\beta) = Y'Y - 2 Y'X\beta + \beta' X' X \beta$$

$$\frac{\partial Q(\beta)}{\partial \beta} = 0$$

$$\frac{\partial Y'Y}{\partial \beta} - \frac{\partial 2Y'X\beta}{\partial \beta} + \frac{\partial \beta' X'X\beta}{\partial \beta} = 0$$

$$0 - 2XY' + 2X'X\beta = 0$$

$$2X'X\beta = 2XY'$$

$$\boxed{\hat{\beta} = (X'X)^{-1} X'Y}$$

la commande R : $lm(Y \sim X_1 + X_2 + X_3 + \dots X_k)$

la fonction summary ( ) : permet de produire les sorties suivantes :

- les coefficients estimés (associés à chaque variable explicative) leur écart-type (Std Error), la valeur de la statistique de Student (t value) et la p-value (Pr(>|t|) : (la probabilité que le coefficient soit significativement différent de zéro) associées à chaque coefficient.
- R^2 ajusté (Adjusted R-squared) : Qualité d'ajustement du modèle
- la statistique de Fisher (testant la significativité globales des variables), son degré de liberté et la p-value associé.

la fonction coef( ) : permet d'extraire les coefficients estimés

la fonction fitted( ) : permet d'extraire les valeurs prédites (estimées) $\hat{Y}_i$

la fonction resid( ) : permet d'extraire les résidus.

valeur prédite - valeur réelle = $\hat{Y}_i - Y_i$

Plusieurs hypothèses sous-jacentes au modèle de regression linéaire portent sur les résidus de la régression :

* indépendance
* homoscédasticité (même variance)
* normalité (shapiro.test(residus))