LE MODÈLE LINÉAIRE

La régression linéaire simple

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

on observe : $Y_i = \alpha + \beta X_i + e_i$

on estime par : $\hat{Y} = a + bX$

$\hat{Y}_i = a + bX_i$

$\sum e_i^2$ minimum

$\sum e_i = 0$

$\sum e_i \cdot X_i = 0$

$\bar{X} = \frac{1}{m} \sum X_i$

$\text{Var } X = \frac{1}{m} \sum (X_i - \bar{X})^2 = E(X^2) - E(X)^2$

Estimation des coefficients

$$b = \frac{\sum_{i=1}^{m} X_i Y_i - m \bar{X} \bar{Y}}{\sum_{i=1}^{m} X_i^2 - m \bar{X}^2} \quad \text{et} \quad a = \bar{Y} - b\bar{X}$$

$$b = \frac{\sum_{i=1}^{m} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{m} (X_i - \bar{X})^2}$$

$$b = \frac{\text{cov}(X,Y)}{V(X)} = r \frac{s_Y}{s_X}$$

avec $s_Y^2 = \frac{1}{m-1} \sum (Y_i - \bar{Y})^2$

$s_X^2 = \frac{1}{m-1} \sum (X_i - \bar{X})^2$

$\text{cov}(X,Y) = \frac{1}{m} \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{m} \sum X_i Y_i - \bar{X}\bar{Y}$

Écarts-types des estimateurs

on note $s^2 = \frac{1}{m-2} \sum (Y_i - \hat{Y}_i)^2 = \frac{m-1}{m-2} (s_Y^2 - b^2 s_X^2)$

(estimateur sans biais de σ^2) $= (1 - r^2) s_Y^2$

écart-type de b : $SE_b = \frac{s}{\sqrt{\sum_{i=1}^{m} (X_i - \bar{X})^2}} = \sqrt{\frac{s_Y^2/s_X^2 - b^2}{m-2}}$

écart-type de a : $SE_a = s \sqrt{\frac{1}{m} + \frac{\bar{X}^2}{\sum_{i=1}^{m} (X_i - \bar{X})^2}}$

Intervalles de confiance de a et b

au niveau de confiance au seuil de 5%, pour tester $\beta = 0$, on regarde si $0 \in IC_\beta$

IC de α : $a \pm t_{0,025;m-2} \cdot SE_a$

IC de β : $b \pm t_{0,025;m-2} \cdot SE_b$

loi de Student à $m-2$ ddl.

Prédiction de Y pour X donné

IC de μ_0 la moyenne de Y en X_0 :

$$a + bX_0 \pm t_{0,025} \cdot s \cdot \sqrt{\frac{1}{m} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

IC pour une simple observation Y_0

$$a + bX_0 \pm t_{0,025} \cdot s \cdot \sqrt{1 + \frac{1}{m} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

Analyse de la variance

appliquée à la régression

on définit les sommes des carrés des écarts

Totale : $SCE_T = \sum_{i=1}^{m} (Y_i - \bar{Y})^2$

Expliquée (= régression) : $SCE_E = \sum_{i=1}^{m} (\hat{Y}_i - \bar{Y})^2$

Résiduelle (= erreur = non expliquée) : $SCE_R = \sum_{i=1}^{m} (Y_i - \hat{Y}_i)^2 = (m-2)s^2$

on démontre : $SCE_T = SCE_E + SCE_R$

Source de variation	Somme des carrés	Degrés de liberté	Moyenne des carrés	Fisher
Expliquée	SCE_E	1	$Q_E = \frac{SCE_E}{1}$	$F = \frac{Q_E}{Q_R}$
Résiduelle	SCE_R	$m-2$	$Q_R = \frac{SCE_R}{m-2}$ $Q_R = s^2$	
Totale	SCE_T	$m-1$		

Sous l'hypothèse $\beta = 0$, F suit une Fisher $(1; m-2)$.

Coefficient de corrélation

$R \in [-1; 1]$, R a le même signe que b

$R^2 = \frac{SCE_E}{SCE_T} = b^2 \cdot \frac{s_X^2}{s_Y^2}$ $\quad \rho = \beta \cdot \frac{\sigma_X}{\sigma_Y}$

on définit $R_{théo} = \rho = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}$

$t_c = \frac{R\sqrt{m-2}}{\sqrt{1-R^2}}$ (on a $F = t_c^2$)

Sous l'hypothèse $\beta = 0$, t_c suit une loi de Student à $(m-2)$ ddl.

Analyse de la variance à 1 facteur

Un facteur A a a modalités : $i : 1 \to a$;

Pour chaque modalité on a n_i observations. $\sum n_i = m$

Source de variation	Somme des carrés des écarts	ddl	Carré moyen	F
Facteur A	$SCE_A = \sum_{i=1}^{a} n_i (\bar{X}_i - \bar{\bar{X}})^2$	$a-1$	$Q_A = \frac{SCE_A}{a-1}$	$F = \frac{Q_A}{Q_R}$
Résiduelle	$SCE_R = \sum_{i=1}^{a} \sum_{t=1}^{n_i} (X_{it} - \bar{X}_i)^2$	$\sum_{i=1}^{a} (n_i - 1) = m - a$	$Q_R = \frac{SCE_R}{m-a}$	
Totale	$SCE_T = \sum_{i=1}^{a} \sum_{t=1}^{n_i} (X_{it} - \bar{\bar{X}})^2$	$m-1$		

Sous $H_0 (\mu_1 = \mu_2 = \dots = \mu_a)$, F suit une loi de Fisher à $(a-1; m-a)$ ddl

On a $SCE_T = SCE_A + SCE_R$