République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER BISKRA

Faculté des Sciences Exactes et Sciences de la Nature et de la Vie
Département de Mathématiques

Première Année Master

Notes de Cours

Analyse de Données

Chapitre 1 : Régression linéaire simple (Séance 3)

Auteur des notes :

Dr. Sana BENAMEUR

Année universitaire : 2021-2022

Théorème 1.5 (Lois des estimateurs avec variance résiduelle estimée)**.**

Si σ_ε^2 est inconnue, dans ce cas σ_ε^2 est estimé par S^2, nous avons :

i) $\dfrac{\hat{b}_0 - b_0}{\hat{\sigma}_{\hat{b}_0}} \sim \mathcal{T}_{n-2}$, *où* $\mathcal{T}_{n-2}$ *la loi de Student à* $(n-2)$ *degrés de liberté (dll).*

ii) $\dfrac{\hat{b}_1 - b_1}{\hat{\sigma}_{\hat{b}_1}} \sim \mathcal{T}_{n-2}$.

ii) $\dfrac{1}{2}\begin{pmatrix} \hat{b}_0 - b_0 \\ \hat{b}_1 - b_1 \end{pmatrix}^t \hat{\Gamma}^{-1}\left(\hat{b}_0, \hat{b}_1\right)\begin{pmatrix} \hat{b}_0 - b_0 \\ \hat{b}_1 - b_1 \end{pmatrix} \sim \mathcal{F}(2, n-2)$, *où* $\mathcal{F}(2, n-2)$ *est la loi de Fisher à* 2 *et* $(n-2)$ *ddl.*

Remarque 1.3. *Ce dernier théorème nous permet de tester la signification des paramètres, la validation du modèle et de trouver des intervalles et des régions de confiance pour les paramètres du modèle.*

1.2.4 Intervalles de Confiance

Soit $\alpha \in [0, 1]$, cherchons une intervalle de confiance (IC) des paramètres b_0 et b_1 au niveau de confiance $(1-\alpha)$.

- Utilisant le théorème 1.5,

$$P\left(\frac{\left|\hat{b}_j - b_j\right|}{\hat{\sigma}_{\hat{b}_j}} < t\right) = 1 - \alpha, \ j = 0, 1,$$

on obtient les IC des paramètres b_0 et b_1

$$b_j = \hat{b}_j \pm t\hat{\sigma}_{\hat{b}_j},$$

où t est le fractile d'ordre $(1-\alpha/2)$ de la loi de Student à $(n-2)$ dll.

- De même, puisque $\dfrac{(n-2)}{\sigma^2}\hat{\sigma}_\varepsilon^2 \sim \mathcal{X}_{n-2}^2$, où $\mathcal{X}_{n-2}^2$ la loi du Khi2 à $(n-2)$ ddl. Alors

$$P\left(t_2 < \frac{(n-2)}{\sigma_\varepsilon^2}\hat{\sigma}_\varepsilon^2 < t_1\right) = 1 - \alpha,$$

On obtient l'IC de σ_ε^2 :

$$\sigma_\varepsilon^2 \in \left]\frac{(n-2)\,\hat{\sigma}_\varepsilon^2}{t_1}, \frac{(n-2)\,\hat{\sigma}_\varepsilon^2}{t_2}\right[,$$

où t_1 (respt. t_2) est le fractile d'ordre $(1-\alpha/2)$ (respt. $\alpha/2$) de la loi de $\mathcal{X}_{n-2}^2$.

- Finalement, la région de confiance de $\begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \end{pmatrix}$ est

$$\frac{1}{2S^2}\left[n\left(\hat{b}_0-b_0\right)^2+2n\bar{x}\left(\hat{b}_0-b_0\right)\left(\hat{b}_1-b_1\right)+\left(\hat{b}_1-b_1\right)^2\sum_{i=1}^{n}x_i^2\right]<f$$

f est le fractile d'ordre $(1-\alpha)$ d'une loi de Fisher $\mathcal{F}(2,n-2)$.

Remarque 1.4. *Un intervalle de confiance pour la moyenne*

$$E(y_i)=b_0+b_1x_i,$$

est

$$E(y_i)=\hat{y}_i\pm t_{1-\frac{\alpha}{2}}S\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}.$$

1.3 Tests sur les paramètres

Proposition 1.1 (Test de Student)**.**

Définissons les hypothèses du test de Student (test bilatéral de significativité)

$$\begin{cases} H_0: b_j=0, \\ H_1: b_j\neq 0. \end{cases} \quad j=0,1$$

Au niveau de confiance $(1-\alpha)\,\%$*, avec* $\alpha\in\,]0,1[$*, on dit que le paramètre* b_j *est significativement nulle si.*

$$T=\frac{\left|\hat{b}_j-b_j\right|}{\hat{\sigma}_{\hat{b}_j}}=\frac{\left|\hat{b}_j\right|}{\hat{\sigma}_{\hat{b}_j}}\leq t,$$

où T *la statistique du test et* t *est le fractile d'ordre* $(1-\alpha/2)$ *de la loi de Student à* $(n-2)$ *ddl (lue à partir la table de Student). Dans ce cas nous acceptons l'hypothèse* H_0*.*

Proposition 1.2 (Test de Fisher)**.**

Soit les hypothèses du test :

$$\begin{cases} H_0: b_0=0 \text{ et } b_1=0, \\ H_1: b_0\neq 0 \text{ et } b_1\neq 0. \end{cases}$$

On accepte l'hypothèse H_0 si

$$\frac{1}{2S^2}\left\{n\left(\hat{b}_0-b_0\right)^2+2n\overline{x}\left(\hat{b}_0-b_0\right)\left(\hat{b}_1-b_1\right)+\left(\hat{b}_1-b_1\right)^2\sum_{i=1}^{n}x_i^2\right\}\leq f$$

f est le fractile d'ordre $1-\alpha$ de la loi de Fisher $\mathcal{F}(2,n-2)$ lue à partir de la table statistique de Fisher. On dit dans ce cas que le modèle est non valide (n'est pas significatif globalement)

1.4 Qualité d'Ajustement

Pour juger la qualité d'ajustement du modèle nous utilisons l'équation de l'analyse de la variance, c-à-d. cherchons tout d'abord à décomposer la variance des y_i autour de leur moyenne en une somme de deux autre variances.

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{SCT}=\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SCR}+\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}_{SCE},$$

où

SCT : somme des carrés totale (ou variation totale des y_i),

SCR : somme des carrés résiduelle (ou variation des résidus $\hat{\varepsilon}_i$ dite aussi variation résiduelle),

SCE : somme des carrés expliqués (variation expliquée).

La qualité d'ajustement peut être déterminée par le coefficient de détermination qui exprime le rapport entre la variation expliquée et la variation totale.

Définition 1.1 (Coefficient de détermination)**.**
On appelle coefficient de détermination, noté D, la quantité

$$D=R^2=\frac{S_{xy}^2}{S_xS_y}=\frac{SCE}{SCT}\leq 1. \tag{1.1}$$

Plus la variation résiduelle est proche de 0, c-à-d la variance expliquée est proche de la variance totale, plus R^2 proche de 1, plus l'ajustement est meilleure.

1.5 Analyse de la variance

1.5.1 Tableau d'analyse de la variance

Source de Variation	ddl	Somme des carrés SC	Moyenne des carrés MC	F
Expliquée	1	SCE	$MCE = SCE/1$	$F = \frac{MCE}{MCR}$
Résiduelle	$n-2$	SCR	$MCR = SCR/(n-2)$	
Totale	$n-1$	SCT		

1.5.2 Test de Fisher

Nous acceptons l'hypothèse de signification globale du modèle si

$$T^2 = F = \frac{SCE/1}{SCR/(n-2)} > f_{(1-\alpha)}(1, n-2),$$

avec : $T = \dfrac{\left|\hat{b}_1\right|}{\hat{\sigma}_{\hat{b}_1}}$,

et $f_{(1-\alpha)}(1, n-2)$ est le fractile d'ordre $1-\alpha$ de la loi de Fisher $\mathcal{F}(1, n-2)$ lue à partir de la table statistique de Fisher.

1.6 Prévision

La prévision est l'un des buts de la régression, c-à-d de prévoir une valeur de la variable à expliquer Y en présence d'une nouvelle valeur de la variable explicative X. Supposons que la nouvelle valeur x_p de X est connue, peut-on prévoir la valeur y_p ?. Pour cela, on postule que la validité du modèle est encore vraie pour la $p^{ème}$ observation, c-à-d.

$$y_p = b_0 + b_1 x_p + \varepsilon_p. \tag{1.2}$$

On estime y_p par un estimateur sans biais, donné par

$$\hat{y}_p = \hat{b}_0 + \hat{b}_1 x_p.$$

De plus

$$\hat{y}_p - y_p = \left(\hat{b}_0 - b_0\right) + \left(\hat{b}_1 - b_1\right) x_p - \varepsilon_p,$$

cette quantité est la somme des variables aléatoires normales, elle suit une loi normale centrée de variance donnée dans la proposition suivante.

Proposition 1.3 (Variance de l'erreur de la prévision)**.**

La variance de l'erreur de la prévision satisfait :

$$\sigma_e^2 = Var(\hat{y}_p - y_p) = \sigma_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right).$$

Démonstration. (Exercice 3) □

Pour un nombre $\alpha \in]0,1[$, l'intervalle de prévision de y_p au niveau de confiance $(1-\alpha)\,\%$ est donnée par :

$$]\hat{y}_p - t\hat{\sigma}_e, \hat{y}_p + t\hat{\sigma}_e[\,,$$

où t est le fractile d'ordre $(1-\alpha/2)$ de la loi de Student $\mathcal{T}(n-2)$.