L2 MASS UE 42 – Correction du contrôle du 4/05/2

Ex 1 : 1) Soient $(x_i, y_i)_{i\leq m}$ des données. Développons l'expression indiquée :

$$\sum_{i=1}^{m}\left[(x_i-\bar{x})(y_i-\bar{y})\right] = \sum_{i=1}^{m} x_i y_i - \sum_{i=1}^{m}\bar{x}y_i - \sum_{i=1}^{m} x_i\bar{y} + \sum_{i=1}^{m}\bar{x}\bar{y}$$

$$= \sum_{i=1}^{m} x_i y_i - \bar{x}\sum_{i=1}^{m} y_i - \bar{y}\sum_{i=1}^{m} x_i + m\bar{x}\bar{y}$$

$$= \sum_{i=1}^{m} x_i y_i - \bar{x}\cdot m\bar{y} - \bar{y}\cdot m\bar{x} + m\bar{x}\bar{y}$$

$$= \sum_{i=1}^{m} x_i y_i - m\bar{x}\bar{y}.$$

lorsque, pour tout i, $y_i = x_i$, on obtient ainsi :

$$\sum_{i=1}^{m}(x_i-\bar{x})^2 = \sum_{i=1}^{m} x_i^2 - m\bar{x}^2 = m\,\mathrm{var}(x_i)$$

2) On observe des triplets $(x_{1,j}, x_{2,j}, y_j)$. Ces triplets sont des réalisations des variables $(X_{j,1}, X_{j,2}, Y_j)$ où les X_1 et X_2 sont déterministes et où Y_j est de la forme

$$Y_j = \beta_0 + \beta_1 X_{1,j} + \beta_2 X_{2,j} + \varepsilon_j$$

avec $\beta_0, \beta_1, \beta_2$ des constantes (déterministes, les paramètres du modèle) et (ε_j) une suite de variables aléatoires indépendantes et de loi $\mathcal{N}(0,\sigma^2)$ avec $\sigma > 0$ (un paramètre également).

On cherche à minimiser la quantité T définie pour tout $b = (b_0, b_1, b_2) \in \mathbb{R}^3$, par

$$T(b) = \sum_{j=1}^{m}\left(Y_j - b_0 - b_1 X_{1,j} - b_2 X_{2,j}\right)^2$$

On peut vérifier que, matriciellement, le modèle s'écrit : $Y = X\beta + \varepsilon$ avec $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{pmatrix}$ et $X = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & X_{m,1} & X_{m,2} \end{pmatrix}$

(2) On a $T(b) = \|Y - Xb\|^2$

$= {}^t(Y-Xb)\cdot(Y-Xb)$

le minimum est obtenu pour $\hat{b} = \begin{pmatrix} \hat{b_0} \\ \hat{b_1} \\ \hat{b_2} \end{pmatrix} = ({}^tXX)^{-1}\,{}^tXy$.

L'estimateur $\hat{y_i}$ de y_i est alors

$\hat{y_i} = \hat{b_0} + \hat{b_1} x_{i1} + \hat{b_2} x_{i2}$

3) Appliquons la méthode de la régression linéaire multiple au cas $p=1$:

Le modèle s'écrit $Y = X\beta + \varepsilon$ avec

$X = \begin{pmatrix} 1 & x_1 \\ 1 & \vdots \\ 1 & x_n \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ et $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ un vecteur de vaid $\mathcal{N}(0,\sigma^2)$.

On obtient comme estimateur $\hat{\beta}$ de β :

$\hat{\beta} = ({}^tXX)^{-1}\,{}^tXY$

Or ${}^tXX = \begin{pmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{pmatrix}$ donc $({}^tXX)^{-1} = \begin{pmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{pmatrix} \cdot \frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2}$

et ${}^tXY = \begin{pmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{pmatrix}$

d'où $\hat{\beta} = \frac{1}{n\Sigma x_i^2 - n^2\bar{x}^2} \begin{pmatrix} \Sigma x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \begin{pmatrix} n\bar{y} \\ \Sigma x_i y_i \end{pmatrix}$

$= \frac{1}{n(\Sigma x_i^2 - n\bar{x}^2)} \begin{pmatrix} n\bar{y}\Sigma x_i^2 - n\bar{x}\Sigma x_i y_i \\ n\Sigma x_i y_i - n^2\bar{x}\bar{y} \end{pmatrix}$

On a donc $\hat{b_1} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2}$: c'est l'estimateur attendu ($\hat{a}$) de la pente de la droite de régression

et $\hat{b_0} = \frac{\bar{y}\Sigma x_i^2 - \bar{x}\Sigma x_i y_i}{\Sigma x_i^2 - n\bar{x}^2}$.

$= \frac{\bar{y}(\Sigma x_i^2 - n\bar{x}^2) + n\bar{x}\bar{y} - \bar{x}\Sigma x_i y_i}{\Sigma x_i^2 - n\bar{x}^2}$

$= \bar{y} - \bar{x} \cdot \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2} = \bar{y} - \hat{b_1}\bar{x}$

(3) On retrouve donc là aussi la valeur attendue (b) pour l'ordonnée à l'origine dans le modèle de la régression linéaire simple.

Exercice 2 :

1) Les données permettent de calculer (avec (x_i) : surfaces, (y_i) : loyers)

- $\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y} = 7275,4$
- $\sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = 813,6$

On en déduit la pente estimée : $\hat{a} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

et l'ordonnée à l'origine estimée : $\hat{b} = \bar{y} - \hat{a}\bar{x} = 152,7$

On en déduit les valeurs estimées des y_i ainsi que les résidus $e_i = y_i - \hat{y}_i$ avec $\hat{y}_i = \hat{a}x_i + \hat{b}$.

Obs. i	1	2	3	4	5	6	7	8	9	10
surface x_i	17	18	19	22	27	31	32	33	36	47
loyer y_i	390	305	310	320	396	427	370	430	480	620
estim. $\hat{y}_i$	304,68	331,02	322,56	349,38	394,08	429,84	438,78	447,72	474,54	572,88
résidu e_i	85,32	-8,62	-12,56	-29,38	1,92	-2,84	-68,78	-17,72	5,46	47,12

La somme des résidus doit être nulle (aux erreurs d'arrondis près : ici on trouve -0,08).

2) Déterminons les sommes des carrés :

$SM = \sum (\hat{y}_i - \bar{y})^2 = \sum \hat{y}_i^2 - n\bar{y}^2 = 65090,81$

$ST = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = 80739,8$

On en déduit $SR = ST - SM = 15648,99$.

On peut alors dresser le tableau d'analyse de la variance :

(4)

Source	ddl	Somme des carrés	Carré moyen
Modèle	$p=1$	$SM = 65090,81$	$SM/1 = 65090,81$
Résidus	$n-p-1=8$	$SR = 15648,49$	$SR/8 = 1956,06$.
Total	$n-1=9$	$ST = 80739,6$	

3) Le coefficient de détermination de la régression est $R^2 = \frac{SM}{ST} = 0,8$

4) Le loyer estimé pour une surface de 30 m^2 est donné (dans ce modèle) par $\hat{y}_{30} = \hat{a}.30 + \hat{b}$
$= 420,9$.

L'intervalle de confiance est alors

$$\left[\hat{y}_{30} - t_8\,\hat{\sigma}\sqrt{1+\frac{1}{10}+\frac{(30-\bar{x})^2}{\sum_{i=1}^{10}(x_i-\bar{x})^2}}\;,\; \hat{y}_{30} + t_8\,\hat{\sigma}\sqrt{1+\frac{1}{10}+\frac{(30-\bar{x})^2}{\sum(x_i-\bar{x})^2}}\right]$$

t_8 est lu dans la table de la loi de Student à 8 degrés de liberté : $P(|T_8| \le t_8) = 0,95$, ou $P(T_8 \le t_8) = 0.975$, pour l'intervalle de niveau 95% : $t_8 = 2,31$.

$\hat{\sigma}$ est estimé dans la question suivante

5) L'estimateur sans biais de la variance de l'erreur est : $\hat{\sigma}^2 = \frac{1}{10-2}\sum(\hat{y}_i - \bar{y})^2$

$= \frac{1}{8}\left(\sum_{i=1}^{10}\hat{y}_i^2 - 10\bar{y}^2\right) = \frac{SR}{8} = 1960.$

⑤ Exercice 3.

1) En notant $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_6 \end{pmatrix}$, $X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \\ 1 & x_{61} & x_{62} \end{pmatrix}$,

le modèle s'écrit $Y = X\beta + \varepsilon$ avec $\beta = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$ et $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_6 \end{pmatrix}$. Le vecteur ε est le vecteur des erreurs : les ε_i sont supposés de loi normale centrée, de variance σ^2 (inconnue), et indépendants.
La matrice X est déterministe.
Le vecteur β est le vecteur des paramètres de la régression. Il est lui aussi déterministe.
Remarquons que l'on a pour tout $i \leq 6$:

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \varepsilon_i.$$

L'objet de la regression linéaire est d'estimer b_0, b_1, b_2 ainsi que la variance σ^2 des ε_i, à partir des observations $(x_{i1}, x_{i2}, Y_i)_{i \leq 6}$.

2)
```
x1 = c(3, 2, -4, 0, 1, -2)
x2 = c(1, -1, 0, -8, 5, 3)
y = c(-4, -2, 5, 7, -6, -4)
reg = lm (y ~ x1 + x2)
```
reg $ residuals : pour avoir les résidus
reg $ fitted.values : les $\hat{y}_i$ estimés par le modèle
summary (reg) ...

3) D'après le cours, on a : $\hat{\beta} = ({}^tX . X)^{-1} {}^tXY$

On a ${}^tXX = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 34 & 0 \\ 0 & 0 & 100 \end{pmatrix}$ - la matrice étant diagonale, elle s'inverse aisément : $({}^tX.X)^{-1} = \begin{pmatrix} 1/6 & 0 & 0 \\ 0 & 1/34 & 0 \\ 0 & 0 & 1/100 \end{pmatrix}$

On a ${}^tXY = -\begin{pmatrix} 4 \\ 34 \\ 100 \end{pmatrix}$ d'où $\hat{\beta} = \begin{pmatrix} -2/3 \\ -1 \\ -1 \end{pmatrix}$

4) Les estimations obtenues sont donc :

$$\hat{y}_i = -\frac{2}{3} - x_{i1} - x_{2i}$$

On obtient $\hat{Y} = {}^t\left(-\frac{14}{3}, \frac{-5}{3}, \frac{10}{3}, \frac{22}{3}, \frac{-20}{3}, \frac{-5}{3}\right)$

et le vecteur des résidus est égal à :

$$\hat{\varepsilon} = Y - \hat{Y} = \left(\frac{2}{3}, \frac{-1}{3}, \frac{5}{3}, \frac{-1}{3}, \frac{2}{3}, \frac{-7}{3}\right)$$

La variance est estimée (sans biais) par :

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \frac{1}{3}\left(\frac{4}{9} + \frac{1}{9} + \frac{25}{9} + \frac{1}{9} + \frac{4}{9} + \frac{49}{9}\right)$$

$$= \frac{84}{27} = \frac{28}{9} = 3,11$$

Source	ddl	Somme des carrés des écarts	Carré moyen	Fisher
Régression	2	$SM = \sum(\hat{y}_i - \bar{y})^2 = 134$	$\frac{SM}{2} = 67$	$\frac{SM/2}{SR/3} = 21,54$
Résidus	3	$SR = \sum(y_i - \hat{y}_i)^2 = 9,33$	$\frac{SR}{3} = 3,11$	
Total	5	$ST = 143,33$		

6) Pour tester la significativité du modèle, on calcule $\frac{SM/2}{SR/3}$, et on le compare au fractile d'ordre 0,95 de la loi de Fisher (2,3), de. En effet sous l'hypothèse H_0 = " $b_1 = b_2 = 0$ ", $\frac{SM/2}{SR/3}$ suit une loi de Fisher (2,3)

La valeur observée est égale ici à 21,54, et $F_{2;3;0,95} = 9,55$.

$9,55 > 21,54$ ce qui conduit à rejeter l'hypothèse H_0 de non significativité. Au risque 5%, on peut donc supposer que l'un (au moins) des coefficients b_1 et b_2 est non nul.

(7) 7) L'étape suivante consiste alors à tester si X_1 est dans le modèle sachant que X_2 y est : c'est à dire que l'on teste l'hypothèse

H'_0 : $b_1 = 0 \mid b_2 \neq 0$.

contre H'_1 : $b_1 \neq 0 \mid b_2 \neq 0$.

Nous testerons ensuite de façon similaire si X_2 est dans le modèle sachant que X_1 y est : H''_0 : $b_2 = 0 \mid b_1 \neq 0$

H''_1 : $b_2 \neq 0 \mid b_1 \neq 0$

Testons H'_0 : Le modèle réduit est le même que si l'on effectuait la régression linéaire (simple) de Y en X_2. Dans ce modèle, la somme des résidus est donnée : $SR_{x_2} = 43,33$.

Donc la somme des carrés des écarts dûs à ce modèle est $SM_{x_2} = ST - 43,33$

$= 143,33 - 43,33 = 100$.

Le test s'effectue en comparant $\frac{SM - SM_{x_2}}{SR/(6-2-1)}$ au fractile d'ordre 0,95 de la loi de Fisher (1,3), où on a noté SM et SR les sommes des carrés des écarts dûs au modèle complet (resp: la somme des carrés des résidus dans le modèle complet). On obtient ici : $\frac{SM - SM_{x_2}}{SR/3} = 10,92$, qui est supérieur à $f_{1,3,0.95} = 10,13$. On rejette donc H'_0 et on accepte H'_1, au risque 0,05.

De façon similaire : $\frac{SM - SM_{x_1}}{SR/(6-2-1)} = \frac{SM - (ST - SR_{x_1})}{SR/(6-2-1)} = 52,14$.

on rejette là aussi H''_0.