

Solution des exercices

Solution de l'Exercice 1 (*Régression linéaire simple*)

- À partir de la présentation graphique (voir figure 1), on constate que le nuage des points est distribué sous une forme linéaire, à priori le modèle proposé est adéquat pour l'explication de Y en fonction de x .

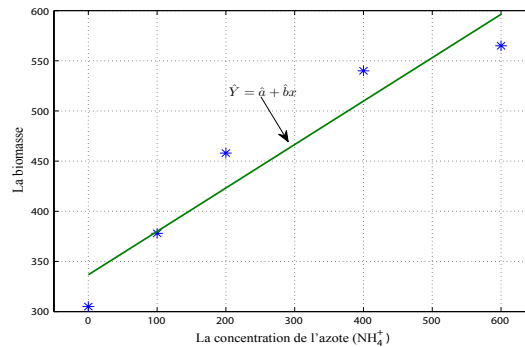


Figure 1: Présentation graphique du nuage des points (X_i, Y_i)

- On a d'une part :

$$\hat{b} = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X} \bar{Y}}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{X}^2} \quad \text{et} \quad \hat{a} = \bar{Y} - \hat{b} \bar{X}. \quad (1)$$

et d'autre part :

$$\begin{aligned} \bar{X} &= \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{5} 1300 = 260, & \bar{Y} &= \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{5} 2246 = 449.2, \\ Cov(x, y) &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X} \bar{Y} = \frac{1}{5} (684400) - (260) (449.2) = 20088, \\ Var(x) &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{X}^2 = \frac{1}{5} (570000) - (260)^2 = 46400, \end{aligned}$$

ainsi,

$$\hat{b} = 0.4329, \quad \text{et} \quad \hat{a} = 336.6460,$$

de ce fait, la droite de régression de la biomasse (Y) en fonction de la concentration (x) est :

$$\hat{Y} = 0.4329 x + 336.6460.$$

- On a d'une part,

$$r = r(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y}, \quad (2)$$

et d'autre part : $Cov(x, y) = 20088$, $\text{Écart-type}(x) = \sqrt{46400} = 215.4066$ et
 $\text{Écart-type}(Y) = \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{Y}^2} = \sqrt{\frac{1}{5}(1056498) - (449.2)^2} = \sqrt{9518.36} = \mathbf{97.5621}$, alors

$$\rho = \rho(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} = 0.9559 = 95.59\%, \quad (3)$$

Le fait que la valeur de $\rho \approx 1$, on déduit qu'il y a une forte liaison linéaire entre x et Y .

4. Afin de valider le modèle nous aurons besoin des $\hat{Y}_i = 0.4329 x_i + 336.6460$ dont leurs valeurs sont rangées dans le tableau suivant :

Concentration (μmol)	0	100	200	400	600
Biomasse (mg)	305	378	458	540	565
\hat{y}_i (mg)	336.646	379.936	423.226	509.806	596.386
$e_i = y_i - \hat{y}_i$	-31.646	-1.936	34.774	30.194	-31.386

On a d'une part

$$f_c = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2 / 1}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n-2)} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2 / 1}{\sum_{i=1}^n e_i^2 / (n-2)} = \frac{43477.3591 / 1}{4111.2071 / (5-2)} = 31.7260,$$

et d'autre par

$$f_\alpha = f(1, n-2, 1-\alpha) = f(1, 3, 0.95) = 10.1.$$

On constate que $f_c > f_\alpha$, alors le modèle est valide (pertinent), c'est-à-dire on admet qu'on peut expliquer la Biomasse de la plante en fonction de la concentration de l'azote par la droite

$$\hat{Y} = 0.4329 x + 336.6460.$$

5. On a : $\hat{Y} = 0.4329 x + 336.6460$ alors la Biomasse qu'on peut prévoir à une concentration $500 \mu mol$ est $\hat{Y} = 0.4329 * 500 + 336.6460 = 553.0960 mg$.

Solution de l'Exercice 2 (Régression linéaire simple)

1. Par définition le coefficient de corrélation linéaire est donné par :

$$r = \frac{Cov(x, y)}{\sigma_x \sigma_y}.$$

on a : $\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{19690}{10} = 1969$, $\bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{20.3}{10} = 2.03$,

$$\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{X}^2} = \sqrt{\frac{1}{10} 42925500 - 1969^2} = 679.5333,$$

$$\sigma_y = \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{Y}^2} = \sqrt{\frac{1}{10} 162.4100 - 2.03^2} = 3.6697,$$

$$Cov(x, y) = \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) - \bar{X} \bar{Y} = \left(\frac{1}{10} 17671 \right) - 1969 \times 2.03 = -2.22997,$$

alors le coefficient de corrélation est :

$$r = -0.9936$$

- Calculer les estimations des paramètres a , b et σ^2 pour la régression linéaire de Y sur X .
- Le modèle linéaire de Y sur X est donné par :

$$Y = aX + b + \epsilon,$$

en utilisant la méthode des moindres carrés les estimateurs de a et b sont définis comme suite :

$$\hat{a} = \frac{Cov(x,y)}{Var(x)} = -0.0054. \quad \text{et} \quad \hat{b} = \bar{Y} - \hat{a}\bar{X} = 12.5953 \quad (4)$$

c'est-à-dire, la droite de régression est :

$$\hat{Y} = -0.0054X + 12.5953.$$

on a,

$$\hat{\sigma}_c^2 = var(\epsilon) = var(y - \hat{a} - \hat{b}x) = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i)^2, \quad (5)$$

donc

$$\hat{\sigma}_c^2 = \frac{1}{10-2} \sum_{i=1}^{10} (y_i - \hat{a} - \hat{b}x_i)^2 = 0.1931.$$

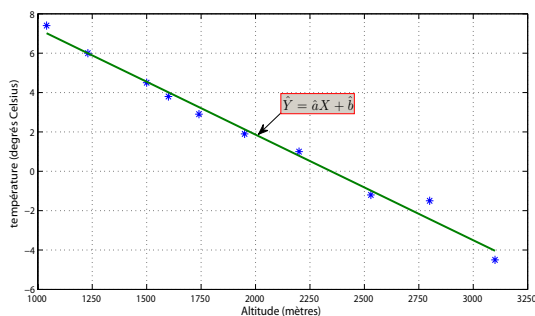


Figure 2: Nuage des points observés et la droite de régression

- Les températures moyennes correspondantes à aux altitudes 1100 m et 2300 m.

(a) 1100m est : $y = -0.0054 * 1100 + 12.5953 = 6.6929$.

(b) 2300m est : $y = -0.0054 * 2300 + 12.5953 = 0.2539$.

Solution de l'Exercice 3 On note X est le poids, Y est le Prix et le modèle de régression est $Y = aX + b$.

- A partir des données on a :

variable	Moyenne	Variance	Carrée Moyenne
X	138.1667	1426.4722	20516.50
Y	166.6667	3222.2222	31000
$X * Y$	24643.33		

d'où : $Cov(X, Y) = 1615.54$, $\rho = 0.754$, $\hat{a} = 1.133$, $\hat{b} = 10.186$ et $\hat{Y} = 1.133X + 10.186$.

2. Si on augmente le poids du Sandwich S_6 à 180 g, alors son nouveau prix sera :
 $\hat{Y} = 1.133(180) + 10.186 = 214.126DA$.

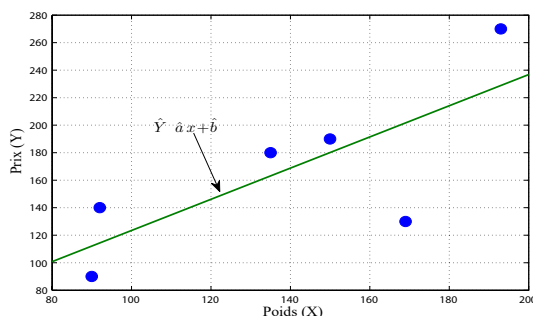


Figure 3: Nuage de variation du Prix des Sandwichs en fonction de leurs Poids.

3. La table d'analyse de la variance, du modèle, est donnée comme suite :

Source	SC	ddl	MC	f_c	Fisher (f_α)
Régression	10978.215	1	10978.215	5.256
Résidu	8355.118	4	2088.780		
Total	19333.333	5			

Table 1: Table d'ANOVA du modèle

A partir de ces résultats, on constate que pour un risque de 5%, le modèle linéaire n'est pas adéquat pour la description de la relation entre les variables Poids et Prix. Mais on peut conclure que ce modèle est adéquat pour un risque de 10%.

Solution de l'Exercice 4 (*Régression linéaire simple et transformation des variables*)

Pour faire une régression linéaire, on effectue un changement de variable en posant $X = \ln(D)$ et $Y = \ln(H)$. Après le calcul des valeurs des variable X et Y on aura les résultats suivants :

X	-1.61	-1.20	-0.97	-0.51	-0.42
Y	2.22	2.27	2.38	2.60	2.65
\hat{Y}	2.1690	2.3255	2.4133	2.5889	2.6232
ϵ	0.0510	-0.0555	-0.0333	0.0111	0.0268

1. Le calcul du coefficient de corrélation linéaire entre X et Y nécessite les quantités suivantes : $\bar{X} =$

$$\frac{1}{n} \sum_{i=1}^n x_i = -0.9420, \bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i = 2.4240,$$

$$\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{X}^2} = \sqrt{0.1945} = 0.4410,$$

$$\sigma_y = \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{Y}^2} = \sqrt{0.0299} = 0.1728,$$

$$Cov(x, y) = \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) - \bar{X} \bar{Y} = 0.0742,$$

ainsi on aura le coefficient de corrélation :

$$r = 0.9737.$$

2. On a,

$$\hat{a} = \frac{Cov(x, y)}{Var(x)} = 0.3817 \text{ et } \hat{b} = \bar{Y} - \hat{a}\bar{X} = 2.7836.$$

alors,

$$Y = 0.38172X + 2.78358, \quad (6)$$

3. Le test de validation du modèle se base sur la statistique :

$$F = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2 / 1}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - 2)} \rightsquigarrow f_{(1, n-2)},$$

or on a,

$$\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2 / 1 = 0.1417$$

et

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - 2) = 0.0025$$

alors la réalisation, f , de la statistique F est égale à : 55.7266.

A partir de la table de Fisher pour un seuil de risque $\alpha = 5\%$, on obtient $f_{(1, n-2, 1-\alpha)} = f_{(1, 3, 0.95)} = 10.1$, on constate que la valeurs de la réalisation de la statistique F est supérieur à la valeurs tabulée de fisher, cela signifier que le modèle est valide c'est-à-dire le modèle linéaire définie dans (6) est adéquat pour l'explication de la variable Y en fonction de la variable X .

4. Donner la hauteur prévue d'un arbre de diamètre 0.7. on a,

$$\hat{Y} = 0.38172X + 2.78358 \Rightarrow \ln(\hat{H}) = 0.38172 \ln(D) + 2.78358 \Rightarrow \hat{H} = e^{0.38172 \ln(D) + 2.78358}.$$

Alors, pour un diamètre $D=0.7$, on prévoit une hauteur $H = e^{0.38172 \ln(0.7) + 2.78358} = 14.1177$.

Solution de l'Exercice 5 (Régression linéaire simple et changement des variables)

							Somme
X $\mu g/\mu l$	0	20	40	60	80	100	300
Y	0	0.205	0.331	0.515	0.584	0.671	2.3060
X^2	0	400	1600	3600	6400	10000	22000
Y^2	0	0.0420	0.1096	0.2652	0.3411	0.4502	1.2081
$X * Y$	0	4.10	13.24	30.90	46.72	67.10	162.06

a) Afin de modéliser ces données, nous avons proposé le modèle linéaire suivant :

$$Y = a_1 x + b_1.$$

1. Calcul des estimateurs des paramètres a_1 et b_1 . On a :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{6} 300 = 50.$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{6} 3002.3060 = 0.3843.$$

$$Cov(x, y) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X} \bar{Y} = \frac{1}{6} (162.06) - (50) (0.3843) = 7.7950$$

$$Var(x) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{X}^2 = \frac{1}{6} (22000) - (50)^2 = 1166.6667$$

alors,

$$\hat{a}_1 = \frac{Cov(x, y)}{Var(x)} = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X} \bar{Y}}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{X}^2} = 0.0067.$$

$$\hat{b}_1 = \bar{Y} - \hat{a}_1 \bar{X} = 0.0503,$$

de ce fait la droite de régression de l'absorbance (Y) en fonction de la concentration (x) est donnée par :

$$\hat{Y} = 0.0067 x + 0.0503.$$

2. Quelle absorbance prévoyez-vous à une concentration 50 $\mu\text{g}/\mu\text{l}$?

$$\hat{Y} = 0.0067 (50) + 0.0503 = 0.3853.$$

3. Quelle absorbance prévoyez-vous à une concentration 40 $\mu\text{g}/\mu\text{l}$? Que peut-on conclure?

$$\hat{Y} = 0.0067 (40) + 0.0503 = 0.3183.$$

On constate que la valeur de régression est très proche de la vraie valeur (0.331), donc à priori le modèle retenu est adéquate pour la représentation des données du tableau.

4. Calcul du coefficient de corrélation linéaire.

$$r = r(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} = 0.9851,$$

avec $\sigma_y = \sqrt{var(Y)} = \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{Y}^2} = \sqrt{0.0536} = 0.2316$; La valeur du coefficient de corrélation est très proche de 1, i.e. X et Y sont fortement linéairement liés donc le modèle est efficace ce qui confirme les résultats de la question 3).

5. Pour un seuil de risque $\alpha = 5\%$, le modèle proposé est-il pertinent?

Pour répondre à cette question on utilise le test de validation du modèle (Fisher). On d'une part

$$f_c = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2 / 1}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - 2)} = \frac{0.3124 / 1}{0.0095 / (6 - 2)} = 131.5368,$$

et d'autre par

$$f_\alpha = f(1, n - 2, 1 - \alpha) = f(1, 4, 0.95) = 7.71.$$

On constate que $f_c > f_\alpha$, alors on accepte le modèle proposé, c'est-à-dire le modèle est valide (pertinent)

- b) Vue les doutes qu'on a sur le modèle précédent, nous avons proposé le modèle suivant :

$$Z = e^Y = a_2 x + b_2.$$

1. Complétez le tableau suivant :

							Somme
X $\mu\text{g}/\mu\text{l}$	0	20	40	60	80	100	300
Z	1.0000	1.2275	1.3924	1.6736	1.7932	1.9562	9.0429
Z ²	1.0000	1.5068	1.9387	2.8011	3.2156	3.8267	14.2888
X * Z	0	24.5505	55.6944	100.4183	143.4558	195.6193	519.7382

2. Calculer les estimations des paramètres a_2 et b_2 pour la régression linéaire de Z sur X .

$$\begin{aligned}\bar{X} &= 50. \\ \bar{Z} &= \frac{1}{n} \sum_{i=1}^n z_i = \frac{1}{6}(9.0429) = 1.5072. \\ Cov(x, z) &= \frac{1}{n} \sum_{i=1}^n x_i z_i - \bar{X} \bar{Z} = \frac{1}{6}(162.06) - (50)(1.5072) = 7.7950 \\ Var(x) &= 1166.6667\end{aligned}$$

alors,

$$\hat{a}_2 = \frac{Cov(x, z)}{Var(x)} = 0.0097.$$

$$\hat{b}_2 = \bar{Z} - \hat{a}_2 \bar{X} = 1.0243,$$

de ce fait la droite de régression de (Z) en fonction de (x) est donnée par :

$$\hat{Z} = 0.0097 x + 1.0243.$$

3. Quelle absorbance prévoyez-vous à une concentration $40 \mu g/\mu l$. Que peut-on conclure par rapport au premier modèle?

On $Z = 0.0097 (40) + 1.0243 = 1.4123$ donc l'absorbance $y = \log(z) = \log(1.4123) = 0.3452$.

On constate que ce modèle nous fournit une valeur proche à la vraie valeur mais c'est le premier modèle qui nous fournit une valeur plus proche de la vraie valeur il se peut que c'est le premier modèle qui est meilleur.

4. Calculer le coefficient de corrélation linéaire de ce nouveau modèle.

$$r = r(x, y) = \frac{Cov(x, z)}{\sigma_x \sigma_z} = 0.9946,$$

5. On constate que le coefficient de corrélation est plus grand pour le deuxième modèle donc le meilleur modèle est le deuxième.