Série 3 STT 5100

Exercice 1.

On examine l'évolution d'une variable réponse y_i en fonction de deux variables explicatives x_i et z_i. Soit $X = (\mathbb{1}\ x\ z)$ la matrice $n \times 3$ du plan d'expérience.

1. Nous avons obtenu les résultats suivants :

$$X'X = \begin{pmatrix} 25 & 0 & 0 \\ ? & 9.3 & 5.4 \\ ? & ? & 12.7 \end{pmatrix} \qquad (X'X)^{-1} = \begin{pmatrix} 0.04 & 0 & 0 \\ 0 & 0.1428 & -0.0607 \\ 0 & -0.0607 & 0.1046 \end{pmatrix}.$$

   (a) Donner les valeurs manquantes.

   (b) Que vaut n ?

   (c) Calculer le coefficient de corrélation linéaire empirique entre x et z.

2. La régression linéaire de Y sur $(\mathbb{1}, x, z)$ donne

$$Y = -1.6\mathbb{1} + 0.61x + 0.46z + \hat{\varepsilon}, \qquad SCR = \|\hat{\varepsilon}\|^2 = 0.3.$$

   (a) Déterminez la moyenne empirique $\bar{y}$.

   (b) Calculer la somme des carrés expliquée (SCE), la somme des carrés totale (SCT), le coefficient de détermination et le coefficient de détermination ajusté.

Exercice 2.

On considère un modèle de régression de la forme :

$$y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \varepsilon_i, \qquad 1 \leq i \leq n.$$

Les $x_{i,j}$ sont supposées non aléatoires. Les erreurs ε_i du modèle sont supposées aléatoires indépendantes gaussiennes centrées de même variance σ^2. On pose comme d'habitude :

$$X = \begin{bmatrix} 1 & x_{1,2} & x_{1,3} & x_{1,4} \\ \vdots & \vdots & \vdots & \\ 1 & x_{n,2} & x_{n,3} & x_{n,4} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}.$$

Un calcul préliminaire a donné

$$X'X = \begin{bmatrix} 50 & 0 & 0 & 0 \\ 0 & 20 & 15 & 4 \\ 0 & 15 & 30 & 10 \\ 0 & 4 & 10 & 40 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 100 \\ 50 \\ 40 \\ 80 \end{bmatrix}, \quad Y'Y = 640.$$

On admettra que

$$\begin{bmatrix} 20 & 15 & 4 \\ 15 & 30 & 10 \\ 4 & 10 & 40 \end{bmatrix}^{-1} = \frac{1}{13720} \begin{bmatrix} 1100 & -560 & 30 \\ -560 & 784 & -140 \\ 30 & -140 & 375 \end{bmatrix}.$$

1. Calculer $\hat{\beta}$, estimateur des moindres carrés de β, la somme des carrés des résidus $\sum_{i=1}^{50} \hat{\varepsilon}_i^2$, et donner l'estimateur de σ^2.
2. Donner un intervalle de confiance pour β_2, au niveau 95%. Faire de même pour σ^2 (on donne $c_1 = 29$ et $c_2 = 66$ pour les quantiles d'ordre 2,5% et 97,5% d'un chi-deux à 46 ddl).

3. Tester la "validité globale" du modèle ($\beta_2 = \beta_3 = \beta_4 = 0$) au niveau 5% (on donne $f^3_{46}(0,95) = 2.80$ pour le quantile d'ordre 95% d'une Fisher à (3,46) ddl).
4. On suppose $x_{51,2} = 1$, $x_{51,3} = -1$ et $x_{51,4} = 0,5$. Donner un intervalle de prévision à 95% pour y_{51}.

Exercice 3.

On souhaite expliquer la hauteur y (en mètres) d'un arbre en fonction de sa circonférence x (en centimètres) à 1m30 du sol et de la racine carrée de celle-ci. On a relevé $n = 1429$ couples (x_i, y_i), le nuage de points étant représenté figure 3.3. On considère donc le modèle de régression suivant :

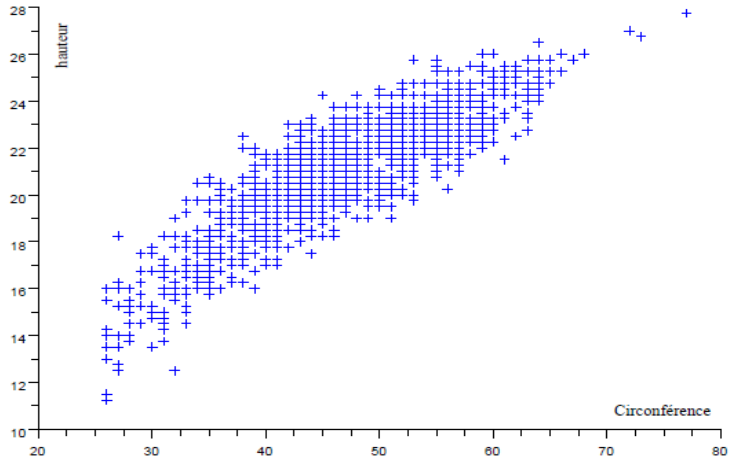

FIGURE 3.3 – Nuage de points pour les eucalyptus.

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 \sqrt{x_i} + \varepsilon_i, \qquad 1 \leq i \leq n.$$

Les ε_i sont des variables aléatoires indépendantes, de loi normale centrée admettant la même variance σ^2. En posant :

$$X = \begin{bmatrix} 1 & x_1 & \sqrt{x_1} \\ \vdots & \vdots & \vdots \\ 1 & x_n & \sqrt{x_n} \end{bmatrix} \quad \text{et} \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$

on a observé :

$$X'X = \begin{bmatrix} ? & ? & 9792 \\ ? & 3306000 & ? \\ ? & 471200 & 67660 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 30310 \\ 1462000 \\ 209700 \end{bmatrix}, \quad Y'Y = 651900.$$

1. Déterminer les '?' dans la matrice $X'X$.
2. Que vaut la circonférence moyenne empirique $\bar{x}$?
3. Le calcul donne (en arrondissant !)

$$(X'X)^{-1} = \begin{bmatrix} 4.646 & 0.101 & -1.379 \\ 0.101 & 0.002 & -0.030 \\ -1.379 & -0.030 & 0.411 \end{bmatrix} \quad \text{et} \quad (X'X)^{-1}X'Y = \begin{bmatrix} -16.8 \\ -0.30 \\ 7.62 \end{bmatrix}.$$

   Que valent les estimateurs $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ par la méthode des moindres carrés ? Grâce au calcul de quelques points, représenter la courbe obtenue sur la figure 3.3.
4. Calculer l'estimateur de σ^2 pour les moindres carrés.
5. Calculer pour β_3 un intervalle de confiance à 95%.
6. Tester l'hypothèse $\beta_2 = 0$ au niveau de risque 10%.
7. Que vaut la hauteur moyenne empirique $\bar{y}$? En déduire le coefficient de détermination ajusté R^2_a.
8. Construire un intervalle de prévision à 95% de y_{n+1} connaissant $x_{n+1} = 49$.
9. Construire un intervalle de prévision à 95% de y_{n+1} connaissant $x_{n+1} = 25$.
10. Des deux intervalles précédents, lequel est le plus grand ? Pouvait-on s'y attendre ?

Exercice 4.

Mr Derek Whiteside de la *UK Building Research Station* a collecté la consommation hebdomadaire de gaz et la température moyenne externe de sa maison au sud-est de l'Angleterre pendant une saison. Une régression pour expliquer la consommation de gaz en fonction de la température est réalisée avec le logiciel R. Les résultats numériques sont les suivants.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.97802 -0.11082  0.02672  0.25294  0.63803

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.72385    0.12974       ?  < 2e-16 ***
Temp        -0.27793          ?  -11.04 1.05e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3548 on 28 degrees of freedom
Multiple R-Squared: 0.8131,     Adjusted R-squared: 0.8064
F-statistic: 121.8 on 1 and 28 DF,  p-value: 1.046e-11
```

1. Donner le modèle et les hypothèses de la régression.
2. Compléter le tableau.
3. Soit Z une variable aléatoire de loi de Student de degré de liberté 28. Quelle est la probabilité que $|Z|$ soit supérieure à 11.04 ?
4. Préciser les éléments du test correspondant à la ligne "`Temp`" du tableau (H_0, H_1, la statistique de test, sa loi sous H_0, la règle de décision).
5. Interpréter le nombre "`Multiple R-Squared: 0.8131`" du tableau.
6. Donner une estimation de la variance du terme d'erreur dans le modèle de régression simple.
7. Expliquer et interpréter la dernière ligne du tableau :
   "`F-statistic: 121.8 on 1 and 28 DF, p-value: 1.046e-11`".
   Voyez-vous une autre façon d'obtenir cette p-value ?
8. Pensez-vous que la température extérieure a un effet sur la consommation de gaz ? Justifiez votre réponse.