Modéle de régression linéaire - Feuille 5

Modèle Gaussien

Exercice 1 (Analyse de sorties logiciel)
Nous voulons expliquer la concentration de l'ozone sur Rennes en fonction des variables `T9`, `T12`, `Ne9`, `Ne12` et `Vx`. Les sorties données par R sont (à une vache près) :

Coefficients :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	62	10	$\boxed{1}$	0
T9	-4	$\boxed{2}$	-5	0
T12	5	0.75	$\boxed{3}$	0
Ne9	-1.5	1	$\boxed{4}$	0.13
Ne12	-0.5	0.5	$\boxed{5}$	0.32
Vx	0.8	0.15	5.3	0

```
--
Multiple R-Squared: 0.6233, Adjusted R-squared: 0.6081
Residual standard error: 16 on 124 degrees of freedom
F-statistic: [6] on [7] and [8] DF, p-value: 0
```

1. Compléter approximativement la sortie ci-dessus.
2. Rappeler la statistique de test et tester la nullité des paramètres séparément au seuil de 5 %.
3. Rappeler la statistique de test et tester la nullité simultanée des paramètres autres que la constante au seuil de 5 %.
4. Les variables `Ne9` et `Ne12` ne semblent pas influentes et nous souhaitons tester la nullité simultanée de $\beta_{\texttt{Ne9}}$ et $\beta_{\texttt{Ne12}}$. Proposer un test et l'effectuer à partir des rèsultats numèriques suivants :

Coefficients :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	66	11	6	0
T9	-5	1	-5	0
T12	6	0.75	8	0
Vx	1	0.2	5	0

```
--
Multiple R-Squared: 0.5312, Adjusted R-squared: 0.52
Residual standard error: 16.5 on 126 degrees of freedom
```

Exercice 2

1. Nous considérons le modèle de régression linéaire $Y = X\theta + \varepsilon$ avec : $Y \in \mathbb{N}^n$, $X \in \mathcal{M}_{n,p}(\mathbb{N})$ de rang p, $\theta \in \mathbb{N}^p$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.
   (a) Quelle est l'interprétation géométrique de $\widehat{Y} = X\widehat{\theta}$?

   (b) Donner $\widehat{\theta}$ et calculer son esprérance et sa matrice de covariance de $\widehat{\theta}$.

   (c) Montrer que $\widehat{\theta}$ et $\widehat{\sigma^2}$ sont indépendants et que $(n-p)\widehat{\sigma^2}/\sigma^2 \sim \chi^2(n-p)$.

2. Considérons un modéle avec 4 variables explicatives (la première étant la constante). Nous avons observé :

$$X'X = \begin{bmatrix} 100 & 20 & 0 & 0 \\ 20 & 20 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad X'Y = \begin{bmatrix} -60 \\ 20 \\ 10 \\ 1 \end{bmatrix}, \quad Y'Y = 159.$$

   (a) Estimer θ, σ^2.

   (b) Donner un estimateur de la variance de $\widehat{\theta}$.

   (c) Donner un intervalle de confiance pour β_2 au niveau 95%.

Exercice 3 (Un modèle à 3 variables explicatives)
On considère un modèle de régression de la forme :

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

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 $f_{46}^3(0,95) = 2.80$ 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 4 On considère le modèle de régression $y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i, \quad 1 \le i \le n$, que l'on écrit sous la forme $Y = X\beta + \varepsilon$. Les $x_{i,j}$ sont des variables exogénes du modèle, les ε_i sont des variables aléatoires indépendantes, de loi normale centrée admettant la même variance σ^2. On a observé :

$$X^TX = \begin{bmatrix} 30 & 20 & 0 \\ 20 & 20 & 0 \\ 0 & 0 & 10 \end{bmatrix}, \quad X^TY = \begin{bmatrix} 15 \\ 20 \\ 10 \end{bmatrix}, \quad Y^TY = 59.5.$$

1. Déterminer n, la moyenne des $x_{i,3}$, le coefficient de corrélation des $x_{i,2}$ et des $x_{i,3}$.
2. Estimer $\beta_1, \beta_2, \beta_3, \sigma^2$ par la méthode des moindres carrés ordinaires.
3. Calculer pour β_2 un intervalle de confiance à 95% et tester l'hypothése $\beta_3 = 0.8$ au niveau 10%.
4. Tester $\beta_2 + \beta_3 = 3$ contre $\beta_2 + \beta_3 \neq 3$, au niveau 5%.
5. Calculer $\bar{y}$ et déduire le coefficient de détermination ajusté R_a^2.
6. Construire un intervalle de prévision à 95% de y_{n+1} si $x_{n+1,2} = 3$ et $x_{n+1,3} = 0.5$.