TP n° 4 : Régression linéaire multiple 1

Exercice 1. On souhaite expliquer une variable quantitative Y à partir de 10 variables quantitatives $X1, \ldots, X10$. Pour ce faire, on considère le modèle de *rlm* :

$$Y = \beta_0 + \beta_1 X1 + \ldots + \beta_{10} X10 + \epsilon,$$

où $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Les paramètres $\beta_0, \beta_1, \ldots, \beta_{10}$ et σ sont des réels inconnus. On les estime alors avec n observations de $(Y, X1, \ldots, X10)$ par la méthode des *mco*. Le tableau de ce modèle de *rlm* renvoyé par la commande `summary` est donné ci-dessous :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	•	•	4.14	0.0005	***
X1	-0.0033	0.0011	•	0.0061	**
X2	-0.0427	0.0149	-2.87	0.0092	**
X3	0.0497	0.0678	0.73	0.4722	
X4	-0.5389	•	•	0.1709	
X5	0.1362	0.0707	1.93	0.0676	.
X6	-0.4224	•	-0.39	0.7004	
X7	0.0459	0.6801	0.07	0.9468	
X8	•	0.1520	-0.25	0.8041	
X9	-0.3626	0.5593	-0.65	•	
X10	-0.5980	0.4966	-1.20	0.2419	

```
Residual standard error:  0.5537 on 21 degrees of freedom
Multiple R-squared:  0.6903, Adjusted R-squared:  •
F-statistic:  • on 10 and 21 DF, p-value:  •
```

Les points • représente des informations volontairement effacées.

1. Quelle est la valeur de n ?
2. Donner l'*emco* ponctuel de β_8.
3. Est-ce que la régression est hautement significative en $X2$?
4. Donner un intervalle de confiance pour β_5 au niveau 95%.
5. Peut-on affirmer que $\beta_{10} \neq -0.51$ au risque 5% ?
6. Donner la valeur du R^2 ajusté.
7. Donner une estimation ponctuelle de σ^2.
8. Donner ete_6.
9. Donner f_{obs} et la p-valeur du test global de Fisher. Quelle est l'hypothèse nulle associée à ce test statistique ?

Exercice 2. On souhaite expliquer le temps en minutes (variable Y) pour approvisionner un réseau de distributeurs de boissons à partir du nombre de caisses de bouteilles placées (variable $X1$) et de la distance parcourue en mètres (variable $X2$). Le jeu de données est disponible ici :

`https://chesneau.users.lmno.cnrs.fr/boisson.txt`

On considère le modèle de *rlm* :

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \epsilon,$$

avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Les paramètres $\beta_0, \beta_1, \beta_2$ et σ sont des réels inconnus.

1. Mettre les données sous la forme d'une data frame `w` en attachant les noms $X1$, $X2$ et Y aux colonnes correspondantes.

2. Représenter les nuages de points associés à $(X1, Y)$ et $(X2, Y)$. Est-ce que ceux-ci laissent à penser la pertinence du modèle de *rlm* pour ce problème ?

3. Reproduire et comprendre l'enjeu des commandes suivantes :

```
library(scatterplot3d)
s3d = scatterplot3d(X1, X2, Y, type = "h", highlight.3d = TRUE, angle = 65,
scale.y = 0.7, pch = 16)
```

4. Reproduire et comprendre l'enjeu des commandes suivantes :

```
X = cbind(1, X1, X2)
X
n = nrow(X)
p = ncol(X) - 1
b = solve(t(X) %*% X) %*% t(X) %*% Y
b
s2 = sum ((Y - X %*% b)^2) / (n - (p + 1))
s2
v = solve(t(X) %*% X) * s2
sqrt(diag(v))
tobs = b / sqrt(diag(v))
pvaleurs = 2 * (1 - pt(abs(tobs), n - (p + 1)))
pvaleurs
R2 = 1 - sum((X %*% b - Y)^2) / sum((mean(Y) - Y)^2)
R2
R2aj = 1 - ((n - 1)/(n - (p + 1))) * (1 - R2)
R2aj
fobs = ((n - (p + 1)) / p) * (R2 / (1 - R2))
fobs
pvaleurfisher = pf(fobs, p, n - (p + 1), lower.tail = F)
pvaleurfisher
```

5. Retrouver les résultats de la question précédente avec les commandes `lm` et `summary`.

6. Donner la valeur prédite de Y lorsque $(X1, X2) = (8, 281)$.

7. Représenter le graphique des résidus. Est-ce que les hypothèses standards semblent être satisfaites ?