Exercice 1.3 (Poids des pères et des fils)
L'étude statistique ci-dessous porte sur les poids respectifs des pères et de leur fil aîné.

Père	65	63	67	64	68	62	70	66	68	67	69	71
Fils	68	66	68	65	69	66	68	65	71	67	68	70

Voici les résultats numériques que nous avons obtenus :

$$\sum_{i=1}^{12} p_i = 800 \qquad \sum_{i=1}^{12} p_i^2 = 53418 \qquad \sum_{i=1}^{12} p_i f_i = 54107 \qquad \sum_{i=1}^{12} f_i = 811 \qquad \sum_{i=1}^{12} f_i^2 = 54849.$$

1. Calculez la droite des moindres carrés du poids des fils en fonction du poids des pères.
2. Calculez la droite des moindres carrés du poids des pères en fonction du poids des fils.
3. Montrer que le produit des pentes des deux droites est égal au carré du coefficient de corrélation empirique entre les p_i et les f_i (ou encore au coefficient de détermination).

Exercice 1.4 (Hauteur d'un arbre)
Nous souhaitons exprimer la hauteur y (en pieds) d'un arbre d'une essence donnée en fonction de son diamètre x (en pouces) à 1m30 du sol. Pour ce faire, nous avons mesuré 20 couples (diamètre,hauteur) et effectué les calculs suivants : $\bar{x} = 4.53$, $\bar{y} = 8.65$ et

$$\frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})^2 = 10.97 \qquad \frac{1}{20}\sum_{i=1}^{20}(y_i - \bar{y})^2 = 2.24 \qquad \frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})(y_i - \bar{y}) = 3.77$$

1. On note $y = \hat{\beta}_0 + \hat{\beta}_1 x$ la droite de régression. Calculer $\hat{\beta}_0$ et $\hat{\beta}_1$.
2. Donner et commenter une mesure de la qualité de l'ajustement des données au modèle. Exprimer cette mesure en fonction des statistiques élémentaires. Commenter le résultat.
3. On donne les estimations de l'écart-type de $\hat{\beta}_0$, $\hat{\sigma}_0 = 1.62$, et de $\hat{\beta}_1$, $\hat{\sigma}_1 = 0.05$. On suppose les perturbations ε_i gaussiennes, centrées, de même variance et indépendantes. Tester $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$ pour $j = 0, 1$. Pourquoi ce test est-il intéressant dans notre contexte ? Que pensez-vous du résultat ?

Exercice 1.5 (Droite de régression et points atypiques)
Douze personnes sont inscrites à une formation. Au début de la formation, ces stagiaires subissent une épreuve A notée sur 20. A la fin de la formation, elles subissent une épreuve B de niveau identique. Les résultats sont donnés dans le tableau suivant :

Epreuve A	3	4	6	7	9	10	9	11	12	13	15	4
Epreuve B	8	9	10	13	15	14	13	16	13	19	6	19

1. Représenter le nuage de points. Déterminer la droite de régression. Calculer le coefficient de détermination. Commenter.
2. Deux stagiaires semblent se distinguer des autres. Les supprimer et déterminer la droite de régression sur les dix points restants. Calculer le coefficient de détermination. Commenter.

Exercice 1.2 (R^2 et corrélation empirique)
Rappeler la formule définissant le coefficient de détermination R^2 et la développer pour montrer qu'il est égal au carré du coefficient de corrélation empirique entre x et y, noté $\rho_{x,y}$, c'est-à-dire qu'on a :

$$R^2 = \rho_{x,y}^2 = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \right)^2$$

Exercice 1.7 (Forrest Gump for ever)
On appelle "fréquence seuil" d'un sportif amateur sa fréquence cardiaque obtenue après trois quarts d'heure d'un effort soutenu de course à pied. Celle-ci est mesurée à l'aide d'un cardio-fréquencemètre. On cherche à savoir si l'âge d'un sportif a une influence sur sa fréquence seuil. On dispose pour cela de 20 valeurs du couple (x_i, y_i), où x_i est l'âge et y_i la fréquence seuil du sportif. On a obtenu $(\bar{x}, \bar{y}) = (35,6; 170,2)$ et :

$$\sum_{i=1}^n (x_i - \bar{x})^2 = 1991 \qquad \sum_{i=1}^n (y_i - \bar{y})^2 = 189,2 \qquad \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = -195,4$$

1. Calculer la droite des moindres carrés pour le modèle $y = \beta_1 + \beta_2 x + \varepsilon$.
2. Calculer le coefficient de détermination R^2. Commenter la qualité de l'ajustement des données au modèle.
3. Avec ces estimateurs, la somme des carrés des résidus vaut $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = 170$. Si on suppose les perturbations ε_i gaussiennes, centrées, indépendantes et de même variance σ^2, en déduire un estimateur non biaisé $\hat{\sigma}^2$ de σ^2.
4. Donner un estimateur $\hat{\sigma}_2^2$ de la variance de $\hat{\beta}_2$.
5. Tester l'hypothèse $H_0 : \beta_2 = 0$ contre $H_1 : \beta_2 \neq 0$ pour un risque de 5%. Conclure sur la question de l'influence de l'âge sur la fréquence seuil.

Exercice 1.11 (Forces de frottement et vitesse)
Au 17$^{\text{ème}}$ siècle, Huygens s'est intéressé aux forces de résistance d'un objet en mouvement dans un fluide (eau, air, etc.). Il a d'abord émis l'hypothèse selon laquelle les forces de frottement étaient proportionnelles à la vitesse de l'objet, puis, après expérimentation, selon laquelle elles étaient proportionnelles au carré de la vitesse. On réalise une expérience dans laquelle on fait varier la vitesse x d'un objet et on mesure les forces de frottement y. Ensuite, on teste la relation existant entre ces forces de frottement et la vitesse.

1. Quel(s) modèle(s) testeriez-vous ?
2. Comment feriez-vous pour déterminer le modèle adapté ?

Exercice 3.9 (Consommation de gaz)
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.

SOLUTIONS

Exercice 1.3 (Poids des pères et des fils)

1. La droite des moindres carrés du poids des fils en fonction du poids des pères s'écrit (cf. figure 1.10 à gauche) : $f = \hat{\alpha}_1 + \hat{\alpha}_2 p = 35.8 + 0.48p$.
2. La droite des moindres carrés du poids des fils en fonction du poids des pères s'écrit (cf. figure 1.10 à droite) : $p = \hat{\beta}_1 + \hat{\beta}_2 f = -3.38 + 1.03f$.
3. Le produit des pentes des deux droites est

$$\hat{\alpha}_2 \hat{\beta}_2 = \frac{\left(\sum (f_i - \bar{f})(p_i - \bar{p})\right)^2}{\left(\sum (f_i - \bar{f})^2\right)\left(\sum_{i=1}^{n} (p_i - \bar{p})^2\right)} = R^2,$$

où R^2 est le coefficient de détermination, carré du coefficient de corrélation linéaire.

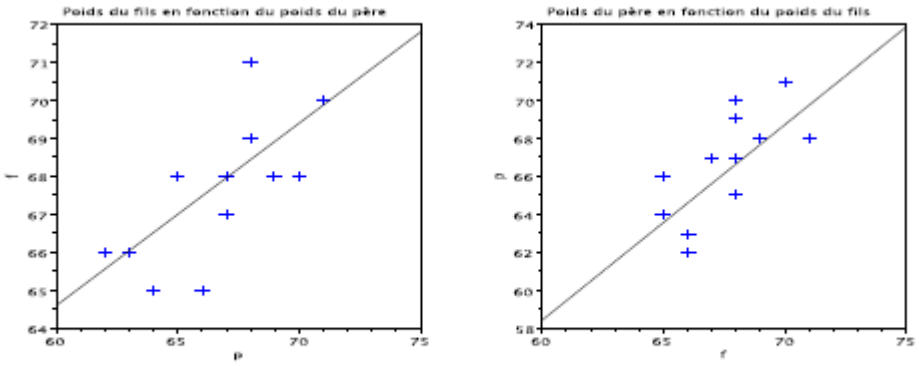

FIGURE 1.10 – Nuages de points et droites de régression pour les poids des pères et des fils.

Exercice 1.4 (Hauteur d'un arbre)
Nous souhaitons exprimer la hauteur y (en pieds) d'un arbre d'une essence donnée en fonction de son diamètre x (en pouces) à 1m30 du sol. Pour ce faire, nous avons mesuré 20 couples (diamètre,hauteur) et effectué les calculs suivants : $\bar{x} = 4.53$, $\bar{y} = 8.65$ et

$$\frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})^2 = 10.97 \qquad \frac{1}{20}\sum_{i=1}^{20}(y_i - \bar{y})^2 = 2.24 \qquad \frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})(y_i - \bar{y}) = 3.77$$

1. Les estimateurs de la droite des moindres carrés $y = \hat{\beta}_0 + \hat{\beta}_1 x$ sont respectivement :

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \approx 0.344$$

et

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \approx 7.09$$

2. Une mesure de la qualité de l'ajustement des données au modèle est donnée par le coefficient de détermination R^2, dont on a vu qu'il correspond au carré du coefficient de corrélation linéaire empirique :

$$R^2 = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \right)^2 \approx 0.58.$$

Le modèle de régression linéaire simple explique donc un peu plus de la moitié de la variance présente dans les données.

3. Sous H_0, on sait que

$$\frac{\hat{\beta}_0}{\hat{\sigma}_0} \sim \mathcal{T}_{18},$$

loi de Student à 18 degrés de liberté. Pour un niveau de confiance de 95%, on compare donc la valeur absolue obtenue dans notre cas particulier, à savoir $|\hat{\beta}_0/\hat{\sigma}_0| \approx 4.38$ au quantile $t_{18}(0.975) \approx 2.1$. On en déduit qu'on rejette l'hypothèse selon laquelle β_0 serait nul. De même pour le test d'hypothèse sur β_1, ce qui donne la statistique de test :

$$\left| \frac{\hat{\beta}_1}{\hat{\sigma}_1} \right| \approx 6.88 > 2.1$$

donc on rejette également l'hypothèse selon laquelle β_1 serait nul.

A priori, un arbre de diamètre nul a une hauteur égale à zéro, donc on aurait pu s'attendre à ce que le coefficient β_0 soit nul. Ceci est en contradiction avec le résultat du test d'hypothèse ci-dessus, mais il n'y a rien d'étonnant à ça : le modèle de régression proposé est pertinent dans l'intervalle considéré, c'est-à-dire pour des arbres de hauteur moyenne 8.65 pieds, avec un écart-type égal à $\sqrt{2.24} \approx 1.5$, non pour des arbres tout petits.

Exercice 1.5 (Droite de régression et points aberrants)
Douze personnes sont inscrites à une formation. Au début de la formation, ces stagiaires subissent une épreuve A notée sur 20. A la fin de la formation, elles subissent une épreuve B de niveau identique. Les résultats sont donnés dans le tableau suivant :

Epreuve A	3	4	6	7	9	10	9	11	12	13	15	4
Epreuve B	8	9	10	13	15	14	13	16	13	19	6	19

1. Pour l'explication de la note B à partir de la note A, la droite de régression (cf. figure 1.11 à gauche) est donnée par $y = \hat{\beta}_1 + \hat{\beta}_2 x$, où :

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \approx 0.11$$

et $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \approx 12.0$ Le coefficient de détermination vaut :

$$R^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)} \approx 0,01$$

Le modèle de régression linéaire expliquerait donc 1% de la variance des données, ce qui est très faible.

2. Si on supprime les deux derniers stagiaires, on obtient cette fois (cf. figure 1.11 à droite) $y = \hat{\alpha}_1 + \hat{\alpha}_2 x = 5.47 + 0.90x$ et $R^2 \approx 0.81$. Sans ces deux stagiaires, le modèle de régression linéaire expliquerait donc 81% de la variance des données, ce qui le rend tout à fait pertinent. Les deux derniers stagiaires correspondent à ce qu'on appelle des points aberrants.

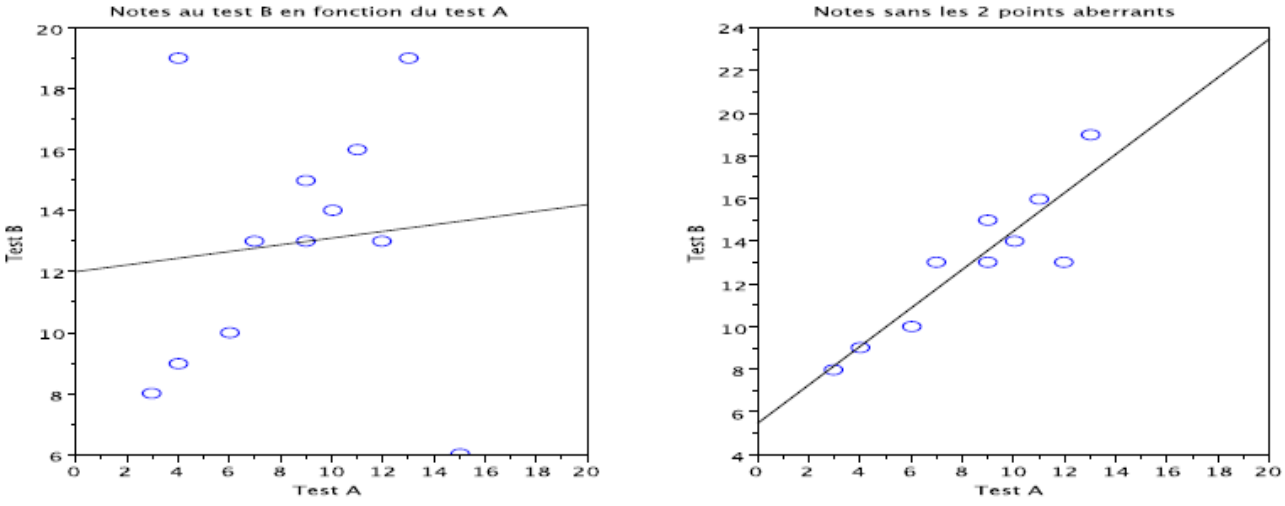

FIGURE 1.11 – Droites de régression et points aberrants.

Exercice 1.2 (R^2 et corrélation empirique)
Le coefficient R^2 s'écrit

$$\begin{aligned} R^2 &= \frac{\|\hat{Y} - \bar{y}\mathbb{1}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2} = \frac{\sum_{i=1}^n \left(\hat{\beta}_1 + \hat{\beta}_2 x_i - \bar{y}\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n \left(\bar{y} - \hat{\beta}_2\bar{x} + \hat{\beta}_2 x_i - \bar{y}\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\hat{\beta}_2^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right]^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2 \sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} = \rho_{x,y}^2, \end{aligned}$$

et la messe est dite.

Exercice 1.7 (Forrest Gump for ever)

1. La méthode des moindres carrés ordinaires donne pour estimateur de β_2 :

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \approx -0,098.$$

   Et pour estimateur de β_1 :

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x} \approx 173.7.$$

2. Le coefficient de détermination R^2 est égal au carré du coefficient de corrélation linéaire entre les variables x et y, ce qui donne :

$$R^2 = \frac{\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)\left(\sum_{i=1}^n (y_i - \bar{y})^2\right)} \approx 0,101.$$

   On en conclut que 10% de la variance des fréquences seuils y_i est expliquée par l'âge. Ce modèle de régression linéaire simple ne semble donc pas efficace.

3. Un estimateur non biaisé $\hat{\sigma}^2$ de σ^2 est tout simplement :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{18} \approx 9.44.$$

4. Un estimateur $\hat{\sigma}_2^2$ de la variance de $\hat{\beta}_2$ est alors donné par :

$$\hat{\sigma}_2^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \approx 0,0047.$$

5. On sait que l'estimateur centré et normalisé de β_2 suit une loi de Student à $(n-2) = 18$ degrés de liberté :

$$\frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_2} \sim \mathcal{T}_{18},$$

donc sous l'hypothèse $H_0 : \beta_2 = 0$, ceci se simplifie en $\frac{\hat{\beta}_2}{\hat{\sigma}_2} \sim \mathcal{T}_{18}$, et cette statistique de test donne ici :

$$t = T(\omega) \approx \frac{-0,098}{\sqrt{0,0047}} \approx -1.43 > -2.101 = t_{18}(0.025).$$

Ainsi on accepte l'hypothèse H_0 selon laquelle la pente de la droite de régression est nulle. Ceci signifie qu'au vu des données dont nous disposons, on serait tenté de considérer que l'âge n'a pas d'influence sur la fréquence seuil. Vu la valeur du coefficient de détermination, il faut toutefois tenir compte du fait que le modèle n'explique pas grand-chose...

Exercice 1.11 (Forces de frottement et vitesse)

1. Le premier modèle, supposant que les forces de frottement sont proportionnelles à la vitesse de l'objet, s'écrit : pour tout $i \in \{1, \dots, n\}$

$$f_i = \alpha v_i + \varepsilon_i,$$

où n est le nombre d'observations et les ε_i représentent les erreurs du modèle, typiquement supposées centrées, décorrélées et de même variance σ^2.

Le second modèle, supposant que les forces de frottement sont proportionnelles au carré de la vitesse de l'objet, s'écrit : pour tout $i \in \{1, \dots, n\}$

$$f_i = \beta v_i^2 + \eta_i,$$

où n est le nombre d'observations et les η_i représentent les erreurs du modèle, typiquement supposées centrées, décorrélées et de même variance s^2.

2. Pour déterminer le modèle adapté, une méthode élémentaire consiste à comparer les pourcentages de variation des données $(f_i)_{1 \leq i \leq n}$ expliqués par chacun des modèles. Ceci se fait en calculant les coefficients de détermination respectifs R_1^2 et R_2^2 pour chaque modèle. On optera pour celui qui a le R^2 le plus grand.

Exercice 3.9 (Consommation de gaz)

1. Le modèle considéré ici est : pour tout $i \in \{1, \dots, 30\}$

$$C_i = \beta_1 + \beta_2 T_i + \varepsilon_i,$$

avec les erreurs ε_i gaussiennes centrées, indépendantes et de même variance σ^2.

2. cf. ci-dessus.

3. Soit Z une variable aléatoire de loi de Student de degré de liberté 28. D'après le tableau, la probabilité que $|Z|$ soit supérieure à 11.04 est de l'ordre de 1.05×10^{-11}.

4. Pour la ligne "`Temp`" du tableau, l'hypothèse H_0 correspond à $\beta_2 = 0$ contre $H_1 : \beta_2 \neq 0$. Sous H_0, $\hat{\beta}_2/\hat{\sigma}_{\hat{\beta}_2}$ suit une loi de Student à 28 degrés de liberté. On décide de rejeter H_0 si la statistique de test $|T(\omega)| = |\hat{\beta}_2/\hat{\sigma}_{\hat{\beta}_2}|$ correspond à une p-value très faible (typiquement inférieure à 5%). En l'occurence, la règle de décision ci-dessus est calculée à partir des valeurs obtenues $\hat{\beta}_2 = -0.27793$, $\hat{\sigma}_{\hat{\beta}_2} = 0.0252$, $|T(\omega)| = |\hat{\beta}_2/\hat{\sigma}_{\hat{\beta}_2}| = 11.04$ et la p-value correspondante pour une loi de Student à 28 degrés de liberté est : $\mathbb{P}(|T| > 11.04) = 1.05 \times 10^{-11}$.

5. Le nombre `Multiple R-Squared: 0.8131` correspond au coefficient de détermination R^2 du modèle. Il signifie qu'environ 81% de la variation des données de consommation est expliquée par ce modèle de régression linéaire simple.

6. Un estimateur de la variance σ^2 du terme d'erreur est donné par le carré du terme `Residual standard error` du tableau, à savoir $\hat{\sigma}^2 = 0.3548^2 \approx 0.126$.

7. La dernière ligne du tableau correspond au test de Fisher de validité globale du modèle. Avec les notations du cours, on sait que sous l'hypothèse $H_0 : \beta_2 = 0$, nous avons

$$F = \frac{n-p}{q} \times \frac{SCR_0 - SCR}{SCR} = \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\hat{\sigma}^2} \sim \mathcal{F}^1_{28},$$

loi de Fisher à 1 et 28 degrés de liberté. La statistique de test donne ici $F(\omega) = 121.8$, ce qui correspond à une p-value de 1.046×10^{-11}. Nous rejetons donc l'hypothèse selon laquelle β_2 serait nul. Remarquons que ce test correspond au test de Student effectué dans la ligne "`Temp`" du tableau.

8. Au vu des résultats du test de Student (ou de l'équivalent Test de Fisher de la dernière ligne), il est clair que la température a un impact sur la consommation de gaz. Ceci est tout à fait naturel, puisque plus il fait froid, plus on chauffe.