6. Modèles linéaire et linéaire gaussien : intervalles de confiance, modèles emboités, lecture des résultats

Objectifs : Savoir lire les résultats d'une régression linéaire à la main, tester des modèles emboités, traiter le cas d'une variable quantitative. Travailler autour de l'erreur de prédiction. L'exercice 6.1 est à faire pendant le TD, le reste est à chercher de votre côté.

Exercice 6.1 (Durée de vie d'une pièce industrielle)**.** Une entreprise de fabrication automobile cherche à expliquer et prédire la durée de vie d'une pièce de sa fabrication, en fonction de variables d'intérêt. Cette pièce, en alliage d'aluminium, est située dans le moteur des véhicules produits par l'entreprise. C'est un parrallélépipède rectangle dont la base est de surface constante, mais d'épaisseur (ou hauteur) variable. Elle est de densité variable selon l'alliage d'aluminium utilisé. On dispose d'un jeu de données de $n = 71$ observations dont les variables sont les suivantes :

- `duree_vie` : durée de vie de la pièce, définie comme le nombre de km parcourus au compteur du véhicule avant de devoir changer la pièce ;
- `epaisseur` : épaisseur de la pièce, en cm ;
- `poids` : poids de la pièce, en hg (1 hg $= 10^2$ g) ;
- `finition` : en fin de fabrication de la pièce en aluminium, une finition lui est appliquée. On distingue trois types de finitions, notés A, B et C.

Dans cet exercice, tous les modèles linéaires rencontrés sont supposés gaussiens. On donne ci-dessous un aperçu du début du jeu de données ainsi qu'un résumé de celui-ci.

duree_vie <dbl>	epaisseur <dbl>	poids <dbl>	finition <chr>
120913.10	6.265404	7.116102	C
139490.76	4.198606	3.115564	B
108487.94	5.286604	2.315019	A
98329.29	2.989635	1.562986	A
115342.99	5.635362	3.820599	C
97965.86	2.804801	1.294772	C

Figure 2 – *Aperçu du jeu de données*

```
   duree_vie        epaisseur         poids          finition
 Min.   : 76750   Min.   :1.889   Min.   :0.3094   Length:71
 1st Qu.:102958   1st Qu.:3.183   1st Qu.:0.4542   Class :character
 Median :118248   Median :3.997   Median :0.7420   Mode  :character
 Mean   :118682   Mean   :3.935   Mean   :0.9292
 3rd Qu.:131886   3rd Qu.:4.726   3rd Qu.:1.1978
 Max.   :161501   Max.   :6.444   Max.   :3.1897
```

Figure 3 – *Résumé du jeu de données*

1. On commence par une régression linéaire classique dont le résultat est présenté sur la figure 4. Attention, certains éléments sont effacés.

   Au vu de ces résultats, quelles sont les éléments expliquant significativement la durée de vie de la pièce ? Dans quel sens sont-ils corrélés à la variable réponse ? On appuiera la réponse sur des arguments quantititatifs.

2. Soucieux d'étudier les données plus en détail, on mène une analyse bivariée des variables quantitatives et on obtient les corrélations empiriques présentées sur la figure 5(a).

   2.(a). Que dire quant à la redondance des variables explicatives ? Donner une explication de cette éventuelle redondance. On souhaite remplacer la variable `poids` par

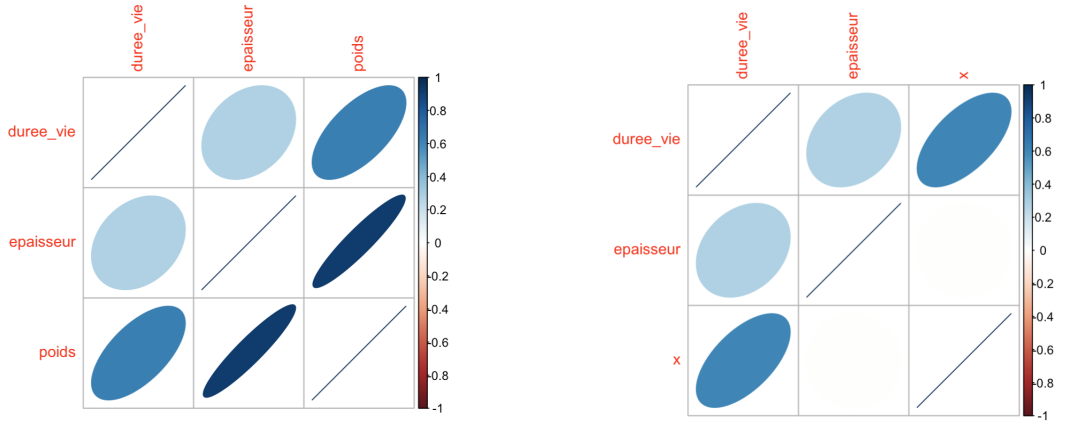
```
Call:
lm(formula = duree_vie ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-26797.8  -6220.6   -128.4   5958.2  28778.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 100976.0     5344.9  18.892  < 2e-16
epaisseur      162.6     1299.7   0.125    0.901
poids         3266.2      522.2   6.254 3.35e-08
finitionB    20831.8     3812.8   5.464 7.64e-07
finitionC    -3387.5     3494.3  -0.969    0.336
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10980 on [redacted] degrees of freedom
Multiple R-squared:  0.7004,    Adjusted R-squared:  0.6822
F-statistic: 38.57 on [redacted] and [redacted] DF,  p-value: < 2.2e-16
```

Figure 4 – *Résultat de la première régression*



(a) – *Corrélations empiriques*

(b) – *Nouvelles corrélations empiriques*

Figure 5 – *Corrélations entre les variables*

une variable notée `x`. Proposer une expression pertinente de la variable `x` pour supprimer la redondance dans les données.

2.(b). Nous appliquons à notre jeu de données la modification proposée en question 2.(a). : la nouvelle variable `x` remplace désormais la variable `poids`. Commenter la figure 5(b).

3. Sur la figure 6, on présente le résultat de la régression linéaire pour ce jeu de données modifié. Compléter les trois dernières lignes du résultat affiché dans la figure 6 en expliquant votre raisonnement. *On pourra utiliser les résultats de la figure 7.*
4. Ce deuxième modèle (avec `x`) est-il plus pertinent que le premier (avec `poids`) pour expliquer les données ? Donner deux indicateurs quantitatifs pour justifier votre réponse.
5. La figure 6 donne des extraits de code. Retrouver les éléments cachés à la ligne `epaisseur` de la figure 5. On ne demande pas de retrouver les étoiles. *On pourra là encore utiliser les résultats de la figure 7.*
6. On s'intéresse à la variable `finition`. Pour $f \in \{A, B, C\}$, on note (f) le groupe des pièces telles que `finition` $= f$.

   6.(a). Les groupes (A) et (B) sont-ils statistiquement équivalents ?

   6.(b). Les groupes (A) et (C) sont-ils statistiquement équivalents ?

```
Residuals:
     Min       1Q   Median       3Q      Max
-25561.9  -5573.7   -713.6   4851.0  28125.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    86464       5572  15.517  < 2e-16
epaisseur       3508                    0.00387
finitionB      20366       3704   5.498 6.69e-07
finitionC      -2867       3392  -0.845  0.40114
x              14192       2098   6.765 4.23e-09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10640 on      degrees of freedom
Multiple R-squared:  0.7182,    Adjusted R-squared:
F-statistic: 42.06 on      and      DF,  p-value: < 2.2e-16
```

Figure 6 – *Résultats de la régression pour le jeu de données modifié*

```r
X = as.matrix(cbind(rep(1,n),epaisseur,finition=='B',finition=='C',x))
colnames(X) <- NULL # on enlève le nom des colonnes
t(X)%*%X
```

```
         [,1]      [,2]     [,3]      [,4]      [,5]
[1,]  71.00000  279.38630 23.00000  34.00000  65.97137
[2,] 279.38630 1184.43593 96.23539 131.34674 259.54139
[3,]  23.00000   96.23539 23.00000   0.00000  26.96826
[4,]  34.00000  131.34674  0.00000  34.00000  26.55636
[5,]  65.97137  259.54139 26.96826  26.55636  89.22035
```

```r
solve(t(X)%*%X)
```

```
            [,1]         [,2]        [,3]        [,4]         [,5]
[1,]  0.27406085 -0.045651410 -0.03859205 -0.068088863 -0.037914974
[2,] -0.04565141  0.012117616 -0.00612221 -0.001874548  0.000914027
[3,] -0.03859205 -0.006122210  0.12111580  0.071188752 -0.011453203
[4,] -0.06808886 -0.001874548  0.07118875  0.101582443  0.004045512
[5,] -0.03791497  0.000914027 -0.01145320  0.004045512  0.038842193
```

Figure 7 – *Extraits de code : on lit X^TX dans le premier prompt, et $(X^TX)^{-1}$ dans le second*

6.(c). Les groupes (B) et (C) sont-ils statistiquement équivalents ? Justifier soigneusement votre réponse.

7. Combien y a-t-il de pièces de finition A dans l'échantillon ?
8. L'ingénieur auquel vous présentez les résultats vous déclare : "au vu de cette étude statistique, le fait d'appliquer la finition B augmente la durée de vie de la pièce. Suggérez-vous de n'utiliser plus que la finition B dans notre processus de fabrication ? "

   Que répondre à cette question au vu de l'étude réalisée ? Argumentez. Quelle nouvelle étude recommanderiez-vous éventuellement afin de vous donner plus d'éléments pour répondre à cette question ?

Exercice 6.2 (Minimisation de l'erreur de prédiction)**.** Dans cet exercice, on s'intéresse à l'erreur de prédiction dans le modèle linéaire. On cherche à apprendre le modèle suivant pour prédire le réel Y_i en fonction d'une variable explicative réelle z_i :

$$\forall 1 \leq i \leq n, \quad Y_i = \beta_0 + \beta_1 z_i + \varepsilon_i \,.$$

On notera $\overline{y}$ et $\overline{z}$ les moyennes empiriques de $z = (z_1, \dots, z_n)^T$ et de $Y = (Y_1, \dots, Y_n)^T$ (attention, elles dépendent de n). On suppose que le modèle est identifiable et on note $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ l'estimateur des moindres carrés de β. On note σ^2 la variance (constante) du bruit.

Sous le même modèle, on observe une nouvelle valeur z_{n+1} de la variable explicative et on cherche à prédire la variable réponse Y_{n+1} avec l'estimateur

$$\hat{y}_{n+1} := \hat{\beta}_0 + \hat{\beta}_1 z_{n+1} = x_{n+1}^T \hat{\beta}, \quad \text{où } x_{n+1} := \begin{bmatrix} 1 \\ z_{n+1} \end{bmatrix} .$$

L'erreur de prédiction est définie par :

$$\mathrm{err}(z_{n+1}) := \mathbb{E}\left[(Y_{n+1} - \hat{y}_{n+1})^2\right] .$$

Notons que dans cette espérance, l'aléa vient du bruit ε_{n+1} dans Y_{n+1} ainsi que de l'aléa dans le $\hat{\beta}$.

1. Montrer que si S est un vecteur aléatoire de $\mathbb{R}^m$ de matrice de covariance (finie) C, alors pour tout vecteur $u \in \mathbb{R}^m$, $\mathrm{Var}(u^T S) = u^T C u$.
2. Que vaut $\mathbb{E}\left[Y_{n+1} - \hat{y}_{n+1}\right]$?
3. Ecrire la matrice X du plan d'expérience dans ce modèle (on mettra l'intercept dans la première colonne), et montrer que

$$(X^T X)^{-1} = \frac{1}{n v(z)} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n z_i^2 & -\overline{z} \\ -\overline{z} & 1 \end{bmatrix},$$

   avec $v(z) := \frac{1}{n}\sum_{i=1}^n z_i^2 - (\overline{z})^2$ la variance empirique de z.
4. En utilisant les résultats des questions 1, 2 et 3, montrer que

$$\mathrm{err}(z_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(z_{n+1} - \overline{z})^2}{\sum_{i=1}^n (z_i - \overline{z})^2}\right) .$$

5. Pour quelle(s) valeur(s) de z_{n+1} l'erreur de prédiction est-elle minimale ? Interpréter ce résultat.
6. Quelle est la limite de l'erreur de prédiction minimale lorsque $n \to \infty$? Interpréter cette valeur limite.

Exercice 6.3 (Equivalence Student/Fisher pour la nullité d'un coefficient)**.** Dans cet exercice souhaite montrer l'équivalence entre les tests de Student et de Fisher pour la nullité d'un paramètre. On considère donc le modèle linéaire Gaussien identifiable $Y = X\theta + \varepsilon$ pour lequel on veut tester la nullité du dernier coefficient θ_p.

1. Donner la statistique T du test de Student pour le test de nullité du dernier coefficient θ_p.
2. Donner la statistique F du test de Fisher pour les modèles emboîtés correspondants. On fera apparaître $\hat{\sigma}^2$ au dénominateur.
3. Soit T_d une variable suivant une loi de Student à d degrés de liberté. Rappeler sa définition et en déduire la loi suivie par la variable $F_d = T_d^2$.
4. On note la matrice du plan d'expérience sous forme de bloc $X = [X_0 \,|\, X_p]$, où $X_0 = [X_1 \,|\, \dots \,|X_{p-1}]$ est la matrice de taille $n \times (p-1)$ des $(p-1)$ premières colonnes de X, et X_p est sa dernière colonne. Ecrire la matrice $X^T X$ sous forme de 4 blocs.
5. On donne (ou on rappelle) le lemme suivant :

Lemme (Inversion par blocs). Soit A une matrice inversible s'écrivant par blocs $A = \begin{bmatrix} T & U \\ V & W \end{bmatrix}$ avec T inversible. Alors $Q = W - VT^{-1}U$ est inversible et l'inverse de A s'écrit :

$$A^{-1} = \begin{bmatrix} T^{-1} + T^{-1}UQ^{-1}VT^{-1} & -T^{-1}UQ^{-1} \\ -Q^{-1}VT^{-1} & Q^{-1} \end{bmatrix}.$$

Grâce à la formule d'inversion matricielle par blocs, en déduire que

$$[(X^T X)^{-1}]_{i,i} = (X_p^T (I_n - \Pi_0) X_p)^{-1},$$

où Π_0 est la matrice $n \times n$ de projection orthogonale sur l'espace engendré par les colonnes de X_0.

6. En reprenant les notations du cours $\hat{Y}$ et $\hat{Y}_0$, vérifier que $\hat{Y}_0 = \Pi_0 \hat{Y}$ puis montrer que

$$\hat{Y} - \hat{Y}_0 = (\hat{\theta}_{MC})_p (I_n - \Pi_0) X_p \,.$$

7. En déduire que dans ce cas $F = T^2$. Conclure.