

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

Objectifs : Savoir lire les résultats d'une régression linéaire à la main, tester des modèles emboîtés, 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 parallélépipède 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 ($1 \text{ hg} = 10^2 \text{ g}$) ;
- **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 quantitatifs. *Solution. c'est une question très gentille ; les facteurs expliquant significativement la durée de vie de la pièce sont a priori sont poids, et le fait ou non qu'elle ait reçu la finition B. Ce sont les seuls coefficients pour lesquels la probabilité d'erreur en rejetant H_0 (coefficient nul) dans test de Student associé est très faible ($\leq 10^{-6}$). Ces deux facteurs sont positivement corrélés à la variable réponse. Les autres ne semblent pas significatifs.*

```
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  degrees of freedom
Multiple R-squared:  0.7004,    Adjusted R-squared:  0.6822
F-statistic: 38.57 on  and  DF,  p-value: < 2.2e-16
```

Figure 4 – Résultat de la première régression

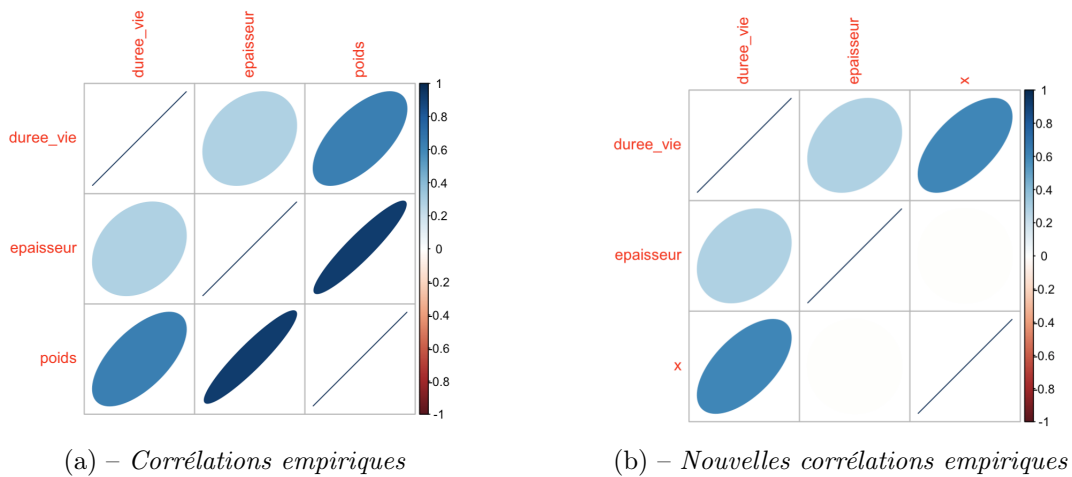


Figure 5 – Corrélations entre les variables

2. Soucieux d'étudier les données plus en détail, on mène une analyse bvariée 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 une variable notée **x**. Proposer une expression pertinente de la variable **x** pour supprimer la redondance dans les données. *Solution. Bien sûr, le poids et l'épaisseur sont intrinsèquement corrélés : la pièce étant de surface constante, il y a relation linéaire directe entre le poids et l'épaisseur. Au vu de l'énoncé, les alliages étant de densité différentes, c'est la densité qui est intéressante. On crée donc cette variable $x = \frac{\text{poids}}{\text{épaisseur}}$ qui supprimera la redondance observée ci-dessus.*

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). *Solution. On a en effet supprimé la redondance dans les variables: x n'est plus corrélée à épaisseur.*

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. *Solution.*

```

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         [redacted]  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 [redacted] degrees of freedom
Multiple R-squared:  0.7182,    Adjusted R-squared: [redacted]
F-statistic: 42.06 on [redacted] and [redacted] 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 $X^T X$ dans le premier prompt, et $(X^T X)^{-1}$ dans le second

Grâce à $X^T X$ (où à la figure 2 si on observe bien), on lit $n = 71$. De plus, ici $p = 1 + 1 + 1 + 2 = 5$ en comptant l'intercept et la variable qualitative. On en déduit que c'est un $n - p = 66$ qui manque pour l'estimée de σ^2 , puis $p - 1 = 4$ et $n - p = 66$ à la dernière ligne. Le R^2 ajusté vaut d'après le cours : $R_a^2 = 1 - \frac{n-1}{n-p}(1 - R^2)$. Application numérique : $R_a^2 = 1 - \frac{70}{66}(1 - 0.7182) \sim 0.7011$.

- 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. Solution. Oui ; le R^2 et le R^2 ajusté sont meilleurs, et la F statistique est plus haute à mêmes degrés de libertés : on rejette encore plus la nullité du modèle.
- 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. Solution. Ce sont l'estimée de l'écart-type du coefficient correspondant à **epaisseur**, ainsi que la statistique de Student correspondante. L'estimateur sans biais de l'écart-type du coefficient θ_i est donné par $\hat{\sigma} \sqrt{[(X^T X)^{-1}]_{i,i}}$, l'application numérique donne $10640 \times \sqrt{0.012117616} \sim 1171$ et la t -value est juste le

quotient de l'estimée par son écart-type, autrement $t = \frac{3508}{1171} \sim 2.9957$.

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 ? *Solution. On lit dans la figure 5 les résultats du test de nullité du paramètre θ_B avec A comme référence. La p-valeur est très très faible : les groupes ne sont pas du tout équivalents statistiquement.*

6.(b). Les groupes (A) et (C) sont-ils statistiquement équivalents ? *Solution. Même chose ici, mais la p-valeur est de 0.4 environ : on ne rejette pas l'hypothèse nulle, les groupes sont donc équivalents statistiquement.*

6.(c). Les groupes (B) et (C) sont-ils statistiquement équivalents ? Justifier soigneusement votre réponse. *Solution. En toute rigueur, il faudrait réaliser le test avec B ou C comme référence. Cependant, on voit que l'estimée de θ_B est très positive, celle de θ_C est négative, et comme (A) et (B) sont déjà distinguables, (B) et (C) le sont aussi.*

7. Combien y a-t-il de pièces de finition A dans l'échantillon ? *Solution. Il faut lire dans $X^T X$ que $n_B = 23, n_C = 34$. On en déduit $n_A = 71 - 23 - 34 = 14$.*

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 ?

Solution. Nous touchons ici au phénomène de causalité (déjà rencontré en cours). Ce que l'on a établi, c'est que le fait d'appliquer la finition B expliquait (au sens de la corrélation) une durée de vie supérieure. Mais cette corrélation n'est pas nécessairement causalité. Par exemple il se peut que les pièces n'aient pas été sélectionnées à l'aveugle (premier souci) ou, pire encore, que dans le processus de fabrication la finition B soit appliquée seulement à des pièces déjà plus résistantes. Il faudrait faire une étude où l'on fabrique des pièces et où le choix de la finition est fait aléatoirement (et indépendamment de tout le reste). A ce moment là, on pourra mettre en lumière un effet causal. C'est ce qu'on appelle un test randomisé.

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 \bar{y} et \bar{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 :

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

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 $u \in \mathbb{R}^m$, $\text{Var}(u^T S) = u^T C u$. *Solution. On peut développer pour voir que*

$$\begin{aligned}\text{Var}(u^T S) &= \text{Var}\left(\sum_{i=1}^m u_i S_i\right) \\ &= \sum_{1 \leq i, j \leq m} \text{Cov}(u_i S_i, u_j S_j) \\ &= \sum_{1 \leq i, j \leq m} u_i u_j C_{i,j} = u^T C u.\end{aligned}$$

2. Que vaut $\mathbb{E}[Y_{n+1} - \hat{y}_{n+1}]$? *Solution. Elle est égale à $\mathbb{E}\left[x_{n+1}^T(\beta - \hat{\beta}) + \varepsilon_{n+1}\right] = x_{n+1}^T 0 + 0 = 0$ par nullité du biais de $\hat{\beta}$ et d'après les hypothèses sur les ε_i .*
3. Ecrire la matrice X du plan d'expérience dans ce modèle (on mettra l'intercept dans la première colonne), et montrer que

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

avec $v(z) := \frac{1}{n} \sum_{i=1}^n z_i^2 - (\bar{z})^2$ la variance empirique de z . *Solution. Dans notre cas on a $X = [\mathbf{1}_n \mid z] \in \mathbb{R}^{n \times 2}$, et $X^T X = n \begin{bmatrix} 1 & \bar{z} \\ \bar{z} & \frac{1}{n} \sum_{i=1}^n z_i^2 \end{bmatrix}$ de déterminant $nv(z)$. Il vient donc $(X^T X)^{-1} = \frac{1}{nv(z)} \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n z_i^2 & -\bar{z} \\ -\bar{z} & 1 \end{bmatrix}$.*

4. En utilisant les résultats des questions 1, 2 et 3, montrer que

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

Solution. Comme la variable dont on prend le moment d'ordre deux est centré d'après la question 2, on calcule en fait une variance.

$$\begin{aligned}\text{err}(z_{n+1}) &= \text{Var}(\varepsilon_{n+1} + x_{n+1}^T(\beta - \hat{\beta})) \\ &= \sigma^2 + x_{n+1}^T (\sigma^2 (X^T X)^{-1}) x_{n+1}.\end{aligned}$$

D'après la question 3 on a $(X^T X)^{-1} = \frac{1}{nv(z)} \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n z_i^2 & -\bar{z} \\ -\bar{z} & 1 \end{bmatrix}$ avec $v(z)$ variance empirique de z . Cela donne

$$\begin{aligned}\text{err}(z_{n+1}) &= \sigma^2 + x_{n+1}^T (\sigma^2 (X^T X)^{-1}) x_{n+1} \\ &= \sigma^2 + \frac{\sigma^2}{nv(z)} \left(\frac{1}{n} \sum_{i=1}^n z_i^2 - 2\bar{z}z_{n+1} + z_{n+1}^2 \right) \\ &= \sigma^2 + \frac{\sigma^2}{nv(z)} (v(z) + \bar{z}^2 - 2\bar{z}z_{n+1} + z_{n+1}^2) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(z_{n+1} - \bar{z})^2}{nv(z)} \right).\end{aligned}$$

5. Pour quelle(s) valeur(s) de z_{n+1} l'erreur de prédiction est-elle minimale ? Interpréter ce résultat. *Solution.* L'erreur est minimale en $z_{n+1} = \bar{z}$ et vaut $\sigma^2 + \sigma^2/n$. On est toujours meilleur pour prédire lorsqu'on observe le barycentre de ce qu'on a déjà observé. En effet, intuitivement, c'est le point le plus typique.
6. Quelle est la limite de l'erreur de prédiction minimale lorsque $n \rightarrow \infty$? Interpréter cette valeur limite. *Solution.* Cette erreur tend vers σ^2 lorsque n est grand, c'est l'erreur minimale de prédiction qu'on puisse faire, car la variable Y_{n+1} est bruitée avec un bruit indépendant et de variance σ^2 : on ne peut pas faire mieux.

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 . *Solution.* c'est $T = \frac{(\hat{\theta}_{MC})_p}{\hat{\sigma} \sqrt{(X^T X)^{-1}_{i,i}}}$ avec pour rappel $\hat{\sigma} = \sqrt{\frac{SCR(\hat{\theta}_{MC})}{n-p}}$.
2. Donner la statistique F du test de Fisher pour les modèles emboîtés correspondants. On fera apparaître $\hat{\sigma}^2$ au dénominateur. *Solution.* Ici le modèle de plus faible dimension est ω , l'espace engendré par les $p-1$ premières colonnes de X . La statistique de Fisher s'écrit donc

$$F = \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\hat{\sigma}^2},$$

avec \hat{Y}_0 le projecté orthogonal de Y sur ω .

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 $F_d = T_d^2$. *Solution.* On rappelle qu'en loi T_d est égale à $\frac{Z}{\sqrt{K/d}}$ où $Z \sim \mathcal{N}(0,1)$ et $K \sim \chi^2(d)$ sont indépendantes. En passant au carré il apparaît que $F = \frac{Z^2/1}{K/d} \sim \mathcal{F}(1, d)$.
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. *Solution.* il vient

$$X^T X = \begin{bmatrix} X_0^T X_0 & X_0^T X_p \\ X_p^T X_0 & X_p^T X_p \end{bmatrix}$$

on pourra vérifier la cohérence des dimensions.

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 . *Solution.* On calcule Q dans le lemme en haut. $X_0^T X_0$ est bien inversible par hypothèse d'indentifiabilité (les $p-1$ premières colonnes de X sont encore libres). Q s'écrit ici $X_p^T X_p - X_p^T X_0 (X_0^T X_0)^{-1} X_0^T X_p$ et on reconnaît la matrice de projection Π_0

(cf cours sur l'estimateur des moindres carrés). $Q = X_p^T(I_n - \Pi_0)X_p$ qui est un réel non nul d'après le Lemme, et donc inversible. On a le résultat en identifiant l'entrée en bas à droite.

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

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

Solution. Comme ω est un sous-espace de Ω , on a $\Pi_\Omega \Pi_0 = \Pi_0$. En effet, $\Pi_0(x) \in \omega \subset \Omega = \text{Im}(\Pi_\Omega)$, et un projecteur est constant sur son image. On obtient $\hat{Y} - \hat{Y}_0 = (I_n - \Pi_0)\hat{Y}$. Mais $\hat{Y} = X\theta_{MC} = (\theta_{MC})_1 X_1 + \dots + (\theta_{MC})_p X_p$. Et pour tout $1 \leq j \leq p-1$, par définition $(I_n - \Pi_0)X_j = 0$. Cela donne bien $\hat{Y} - \hat{Y}_0 = (\hat{\theta}_{MC})_p(I_n - \Pi_0)X_p$.

7. En déduire que dans ce cas $F = T^2$. Conclure. Solution. En utilisant les questions précédentes on a que

$$\begin{aligned} F &= \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\hat{\sigma}^2} = \frac{(\hat{\theta}_{MC})_p^2 \|(I_n - \Pi_0)X_p\|^2}{\hat{\sigma}^2} = \frac{(\hat{\theta}_{MC})_p^2 X_p^T (I_n - \Pi_0)(I_n - \Pi_0)X_p}{\hat{\sigma}^2} \\ &= \frac{(\hat{\theta}_{MC})_p^2 X_p^T (I_n - \Pi_0)X_p}{\hat{\sigma}^2} = \frac{(\hat{\theta}_{MC})_p^2}{\hat{\sigma}^2 [(X^T X)^{-1}]_{i,i}} = T^2 \end{aligned}$$

Conclusion, ici tester si $|T| > t_{1-\alpha/2}^{n-p}$ (test de Student) revient exactement à tester si $F > f_{1-\alpha}^{1, n-p}$ (test de Fisher).