Examen Final

Durée 2h30. Les documents, la calculatrice, les téléphones portables, tablettes, ordinateurs ne sont pas autorisés. Les exercices sont indépendants. La qualité de la rédaction sera prise en compte.

Exercice 1. Interprétation géométrique en régression linéaire simple

1. On considère un modèle de régression linéaire simple sans constante :

$$\mathbf{y} = \beta\mathbf{x} + \boldsymbol{\epsilon}$$

où $\mathbf{y}$ est un vecteur aléatoire à valeurs dans $\mathbb{R}^2$, $\mathbf{x} = (2,1)^T$, $\beta \in \mathbb{R}$ et $\boldsymbol{\epsilon}$ est un vecteur aléatoire vérifiant les conditions classiques d'un modèle de régression linéaire.

   (a) Rappelez la définition de l'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β, et donnez une interprétation géométrique du vecteur des valeurs ajustées $\hat{\mathbf{y}} = \mathbf{x}\hat{\beta}$.

   L'estimateur des MCO de β est défini par $\hat{\beta} = \text{argmin}_{b\in\mathbb{R}}\|\mathbf{y} - \mathbf{x}b\|^2$. Si $\mathcal{M}$ désigne l'espace vectoriel engendré par le vecteur $\mathbf{x}$, le vecteur $\mathbf{x}\hat{\beta}$ est le vecteur de $\mathcal{M}$ dont la distance à $\mathbf{y}$ est minimale, c'est-à-dire le projeté orthogonal de $\mathbf{y}$ sur $\mathcal{M}$.

   (b) La valeur observée de $\mathbf{y}$ est $\mathbf{y}^{obs} = (3,4)^T$. Faites un dessin. À l'aide de la question précédente exclusivement, déterminer graphiquement sans faire de calcul la valeur de $\hat{\beta}$.

   On constate que $\hat{\mathbf{y}} = \mathbf{x}\hat{\beta} = 2\mathbf{x}$. On en déduit $\hat{\beta} = 2$.

2. On considère maintenant le modèle de régression linéaire simple avec constante

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

où $\mathbf{y}$ est un vecteur aléatoire à valeurs dans $\mathbb{R}^2$, $\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$, $\boldsymbol{\beta} \in \mathbb{R}^2$ et $\boldsymbol{\epsilon}$ est un vecteur aléatoire vérifiant les conditions standards d'un modèle de régression linéaire.
La valeur observée de $\mathbf{y}$ est toujours $\mathbf{y}^{obs} = (3,4)^T$.
Déterminez géométriquement sans faire de calcul la valeur de $\hat{\boldsymbol{\beta}}$ dans ce cas.

   Les deux vecteurs colonnes de $\mathbf{X}$ sont indépendants, l'espace engendré $\mathcal{M}$ est donc $\mathbb{R}^2$ tout entier, et $\hat{\mathbf{y}} = \mathbf{y}$.
   On remarque de plus que $\hat{\mathbf{y}} = \mathbf{y} = 5\mathbf{X}_1 - \mathbf{X}_2$. On en déduit $\hat{\beta}_1 = 5$ et $\hat{\beta}_2 = -1$.

3. Retrouvez les valeurs de $\hat{\boldsymbol{\beta}}$ de la question précédente en utilisant les formules classiques modèle de régression linéaire simple avec intercepte.

   En appliquant les formules:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^2 x_i y_i - 2\bar{x}\bar{y}}{\sum_{i=1}^2 x_i^2 - 2\bar{x}^2} = \frac{2\times 3 + 4\times 1 - 2\times 1.5\times 3.5}{4 + 1 - 2\times 1.5^2} = \frac{10 - 10.5}{5 - 4.5} = -1$$
$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x} = 3.5 - (-1)\times 1.5 = 5$$

   On retrouve bien les valeurs précédentes.

Exercice 2. Moindres carrés pondérés

On suppose le modèle suivant :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

avec :

- $\mathbf{X}$ la matrice $(n \times p)$ du plan d'expérience,
- $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$ un vecteur de $\mathbb{R}^p$,
- $\mathbf{y}$ le vecteur $(n \times 1)$ des observations y_i,
- $\boldsymbol{\epsilon}$ le vecteur $(n \times 1)$ des erreurs.

Comme dans la régression linéaire classique, on suppose que les erreurs ϵ_i sont centrées ($\mathbb{E}[\epsilon_i] = 0$ pour tout $1 \leq i \leq n$) et de covariances nulles ($\mathbb{C}[\epsilon_i, \epsilon_j] = 0$ pour tout $1 \leq i \neq j \leq n$). En revanche, on suppose que la variance de chaque erreur est *variable*, pondérée par un "poids" $w_i > 0$, tel que: $\mathbb{V}[\epsilon_i] = \frac{\sigma^2}{w_i}$ pour tout $1 \leq i \leq n$. Dans ce modèle, les valeurs des poids w_i sont supposées connues, mais les paramètres $\boldsymbol{\beta}$ et σ^2 sont inconnus.

1. On considère le modèle transformé:

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\epsilon}^* \tag{2}$$

où :
- $\mathbf{X}^*$ est la matrice $(n \times p)$ tel que $x^*_{ij} = \sqrt{w_i} \times x_{ij}$ pour tout $1 \leq i, j \leq n$,
- $\mathbf{y}^*$ le vecteur $(n \times 1)$ tel que $y^*_i = \sqrt{w_i} \times y_i$ pour tout $1 \leq i \leq n$,
- $\boldsymbol{\epsilon}^*$ le vecteur $(n \times 1)$ tel que $\epsilon^*_i = \sqrt{w_i} \times \epsilon_i$ pour tout $1 \leq i \leq n$.

On défini de plus $\mathbf{W} = \text{Diag}(w_1, \ldots, w_n)$ la matrice diagonale telle que $W_{ii} = w_i$ et $W_{ij} = 0$ et $\mathbf{W}^{1/2} = \text{Diag}(\sqrt{w_1}, \ldots, \sqrt{w_n})$ la matrice diagonale telle que $W^{1/2}_{ii} = \sqrt{w_i}$ et $W^{1/2}_{ij} = 0$ pour tout $1 \leq i \neq j \leq n$.

(a) Écrire $\mathbf{X}^*$ (respectivement $\mathbf{y}^*$, $\boldsymbol{\epsilon}^*$), en fonction de $\mathbf{X}$ (respectivement $\mathbf{y}$, $\boldsymbol{\epsilon}$) et $\mathbf{W}$.

$$\mathbf{X}^* = \mathbf{W}^{1/2}\mathbf{X} \qquad \mathbf{y}^* = \mathbf{W}^{1/2}\mathbf{y} \qquad \boldsymbol{\epsilon}^* = \mathbf{W}^{1/2}\boldsymbol{\epsilon}^*$$

(b) Déterminer le vecteur moyenne et la matrice de covariance du vecteur aléatoire $\boldsymbol{\epsilon}$.

$$\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}_n \qquad \mathbb{V}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{W}^{-1}$$

(c) Déterminer le vecteur moyenne et la matrice de covariance du vecteur aléatoire $\boldsymbol{\epsilon}^*$.

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\epsilon}^*] &= \mathbf{W}^{1/2}\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}_n \\
\mathbb{V}[\boldsymbol{\epsilon}^*] &= \mathbb{V}[\mathbf{W}^{1/2}\boldsymbol{\epsilon}] = \mathbf{W}^{1/2}\mathbb{V}[\boldsymbol{\epsilon}][\mathbf{W}^{1/2}]^T = \mathbf{W}^{1/2}\sigma^2\mathbf{W}^{-1}\mathbf{W}^{1/2} \\
&= \mathbf{W}^{1/2}[\mathbf{W}^{1/2}]^{-1}\sigma^2[\mathbf{W}^{1/2}]^{-1}\mathbf{W}^{1/2} = \sigma^2\mathbf{I}_n
\end{aligned}$$

(d) Sous quelle(s) condition(s) le modèle de l'équation (2) suit-il les hypothèses classiques du modèle linéaire vu en cours ?

On a bien:
- $\mathbb{E}[\boldsymbol{\epsilon}^*] = \mathbf{0}_n$
- $\mathbb{V}[\boldsymbol{\epsilon}^*] = \sigma^2\mathbf{I}_n$.

Pour que le modèle suive bien les hypothèses, il faut de plus supposer $rg(\mathbf{X}^*) = p$.

(e) En supposant $\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)$ inversible, et en utilisant le modèle de l'équation (2), déterminer l'estimateur des moindres carrés $\hat{\boldsymbol{\beta}}^*$ de $\boldsymbol{\beta}$ en fonction de $\mathbf{X}$, $\mathbf{W}$ et $\mathbf{y}$. Préciser son biais et sa matrice de covariance.

Le modèle de l'équation (2) est le modèle classique vu en cours. On obtient donc:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^* &= ([\mathbf{X}^*]^T\mathbf{X}^*)^{-1}[\mathbf{X}^*]^T\mathbf{y}^* \\ &= (\mathbf{X}^T[\mathbf{W}^{1/2}]^T\mathbf{W}^{1/2}\mathbf{X})^{-1}\mathbf{X}^T[\mathbf{W}^{1/2}]^T\mathbf{W}^{1/2}\mathbf{y} \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}\end{aligned}$$

et:

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}}^*] &= \boldsymbol{\beta} \\ \mathbb{V}[\hat{\boldsymbol{\beta}}^*] &= \sigma^2([\mathbf{X}^*]^T\mathbf{X}^*)^{-1} = \sigma^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\end{aligned}$$

(f) Montrer que l'on peut écrire:

$$\hat{\boldsymbol{\beta}}^* = \underset{\mathbf{b}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \times \left(y_i - \sum_{j=1}^{p} x_{ij}b_j\right)^2.$$

Justifier le terme de "moindres carrés pondérés".

Par définition:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^* &= \underset{\mathbf{b}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{\sum_{i=1}^{n}\left(y_i^* - \sum_{j=1}^{p} b_j x_{ij}^*\right)^2\right\} \\ &= \underset{\mathbf{b}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n}\left(\sqrt{w_i}y_i - \sum_{j=1}^{p} b_j\sqrt{w_i}x_{ij}\right)^2 \\ &= \underset{\mathbf{b}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} w_i\left(y_i - \sum_{j=1}^{p} b_j x_{ij}\right)^2\end{aligned}$$

Chaque terme i de la somme des carrés est pondéré par un poids w_i, d'où le terme de "moindres carrés pondérés".

(g) Proposer un estimateur sans biais $\hat{\sigma}_*^2$ de σ^2.

Le modèle de l'équation (2) est le modèle classique vu en cours. Un estimateur sans biais de σ^2 est donc donné par:

$$\hat{\sigma}_*^2 = \frac{1}{n-p}\sum_{i=1}^{n} w_i\left(y_i - \sum_{j=1}^{p}\hat{\beta}_j^* x_{ij}\right)^2$$

2. En revenant au modèle initial (1), on fait l'hypothèse supplémentaire que les erreurs ϵ_i sont gaussiennes, c'est-à-dire $\epsilon_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{w_i}\right)$ pour tout $1 \leq i \leq n$. Les erreurs sont supposées indépendantes deux à deux.

(a) Donner la loi des vecteurs aléatoire $\boldsymbol{\epsilon}$ et $\mathbf{y}$.

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{W}^{-1}) \qquad \mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{W}^{-1})$$

(b) Donner la log-vraisemblance $\ell(\boldsymbol{\beta}, \sigma^2 | \mathbf{y})$ du modèle.
On rappelle que la densité d'une variable aléatoire gaussienne $Z \sim \mathcal{N}(m, s^2)$ est donnée par: $f : z \mapsto \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{1}{2}\frac{(z-m)^2}{s^2}}$.

Comme $\mathbf{W}$ est diagonale, les y_i sont indépendants de loi $y_i \sim \mathcal{N}(\mathbf{x}^i\boldsymbol{\beta}, \sigma^2/w_i)$, et:

$$\begin{aligned}
\ell\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}\right) &= \sum_{i=1}^{n} \ell(\boldsymbol{\beta}, \sigma^2 | y_i) \\
&= \sum_{i=1}^{n} -\frac{1}{2}\log(2\pi\sigma^2/w_i) - \frac{1}{2(\sigma^2/w_i)}(y_i - \mathbf{x}^i\boldsymbol{\beta})^2 \\
&= -\frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2}\sum_{i=1}^{n}\log(w_i) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} w_i(y_i - \mathbf{x}^i\boldsymbol{\beta})^2
\end{aligned}$$

(c) En déduire la relation entre l'estimateur du maximum de vraisemblance $\hat{\boldsymbol{\beta}}_{mv}$ et l'estimateur des moindres carrés pondérés $\hat{\boldsymbol{\beta}}^*$.

On retrouve la somme des carrés pondérés dans la log-vraisemblance, si bien que $\hat{\boldsymbol{\beta}}_{mv} = \hat{\boldsymbol{\beta}}^*$.

(d) Dérivez l'expression de l'estimateur du maximum de vraisemblance $\hat{\sigma}^2_{mv}$ de σ^2. Quelle est sa relation avec l'estimateur $\hat{\sigma}^2_*$ de la section précédente ?

En dérivant la log-vraissemblance par rapport à σ^2 comme vu en cours, on trouve:

$$\hat{\sigma}^2_{mv} = \frac{1}{n}\sum_{i=1}^{n} w_i \left(y_i - \sum_{j=1}^{p} \hat{\beta}^*_j x_{ij}\right)^2$$

d'où:

$$\hat{\sigma}^2_{mv} = \frac{n}{n-p}\hat{\sigma}^2_*.$$

(e) Donnez les lois de $\hat{\boldsymbol{\beta}}^*$ et de $\hat{\sigma}^2_*$.

Les estimateurs $\hat{\beta}^*$ et de $\hat{\sigma}^2_*$ sont obtenus à partir du modèle de l'équation (2), qui vérifie les hypothèse classiques, avec des résidus iid gaussiens. On obtient donc directement:

$$\hat{\beta}^* \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}) \qquad \frac{(n-p)\hat{\sigma}^2_*}{\sigma^2} \sim \chi^2_{n-p}$$

3. Supposons maintenant le modèle classique de régression linéaire

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

avec des erreurs centrées et de matrice de covariance $\mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$. Néanmoins, on n'observe pas comme d'habitude les $\mathbf{x}^i$ et y_i, mais des moyennes par classe. Spécifiquement, les n données sont réparties en L classes $C_1, \dots, C_L$ d'effectifs respectifs connus $n_1, \dots, n_L$ (avec $\sum_{\ell=1}^{L} n_\ell = n$), et on a seulement accès aux moyennes par classe, à savoir pour tout $1 \leq \ell \leq L$:

$$\bar{y}_\ell = \frac{1}{n_\ell} \sum_{i \in C_\ell} y_i \qquad \& \qquad \bar{x}_{\ell j} = \frac{1}{n_\ell} \sum_{i \in C_\ell} x_{ij}$$

(a) En notant $\bar{\epsilon}_\ell = \frac{1}{n_\ell} \sum_{i \in C_\ell} \epsilon_i$, vérifier que le modèle liant les observations $\bar{y}_\ell$ aux régresseurs $\bar{x}_{\ell j}$ peut se mettre sous la forme

$$\bar{\mathbf{y}} = \bar{\mathbf{X}} \boldsymbol{\beta} + \bar{\boldsymbol{\epsilon}}. \tag{3}$$

On note:
- $\bar{\mathbf{y}} = [\bar{y}_1, \dots, \bar{y}_L]^T$ le vecteur de taille L des observations moyennes,
- $\bar{\mathbf{x}}_j = [\bar{x}_{1j}, \dots, \bar{x}_{\ell j}]^T$ le vecteur colonne de taille L,
- $\bar{\mathbf{x}} = [\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_p]$ la matrice de taille $L \times p$ des prédicteurs,
- $\bar{\boldsymbol{\epsilon}} = [\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_L]^T$ le vecteur de taille L des erreurs moyennes.

On a alors bien:

$$\bar{\mathbf{y}} = \bar{\mathbf{X}} \boldsymbol{\beta} + \bar{\boldsymbol{\epsilon}}.$$

(b) Donner la moyenne et la matrice de covariance de $\bar{\boldsymbol{\epsilon}}$.

Pour tout $\ell \neq m$, comme les ϵ_i sont de covariance nulle:

$$\mathbb{E}[\bar{\epsilon}_\ell] = \frac{1}{n_\ell} \sum_{i \in C_\ell} \mathbb{E}[\epsilon_i] = 0$$

$$\mathbb{C}\left[\bar{\epsilon}_\ell, \bar{\epsilon}_m\right] = \mathbb{C}\left[\frac{1}{n_\ell} \sum_{i \in C_\ell} \epsilon_i, \frac{1}{n_m} \sum_{j \in C_m} \epsilon_j\right] = 0$$

$$\mathbb{V}[\bar{\epsilon}_\ell] = \frac{1}{n_\ell^2} \sum_{i \in C_\ell} \mathbb{V}[\epsilon_i] = \frac{\sigma^2}{n_\ell}$$

En notant $\mathbf{W}$ la matrice diagonale de termes $W_{\ell\ell} = n_\ell$, on obtient:

$$\mathbb{E}[\bar{\boldsymbol{\epsilon}}] = \mathbf{0}_L \qquad \mathbb{V}[\bar{\boldsymbol{\epsilon}}] = \sigma^2 \mathbf{W}^{-1}$$

(c) En déduire que le modèle (3) peut s'écrire comme une régression pondérée. Interprétez les poids donnés à chaque observation dans ce modèle.

On se retrouve bien dans le cas précédent, avec un modèle à L points, chaque points étant pondérée par n_ℓ le nombre d'observations associés au point ℓ. Plus un point a d'observation, plus sont poids sera grand dans la régression pondérée, c'est-à-dire plus il comptera dans les estimateurs des moindres carrés.

(d) Déduire des questions précédentes des estimateurs de $\boldsymbol{\beta}$ et σ^2.

On peut utiliser les estimateurs non biaisés suivants:

$$\hat{\boldsymbol{\beta}}^* = (\bar{\mathbf{X}}^T \mathbf{W} \bar{\mathbf{X}})^{-1} \bar{\mathbf{X}}^T \mathbf{W} \bar{\mathbf{y}} \qquad \hat{\sigma}^2_* = \frac{1}{L-p} \sum_{\ell=1}^{L} n_\ell \left(\bar{y}_\ell - \sum_{j=1}^{p} \hat{\beta}^*_j \bar{x}_{\ell j} \right)^2$$

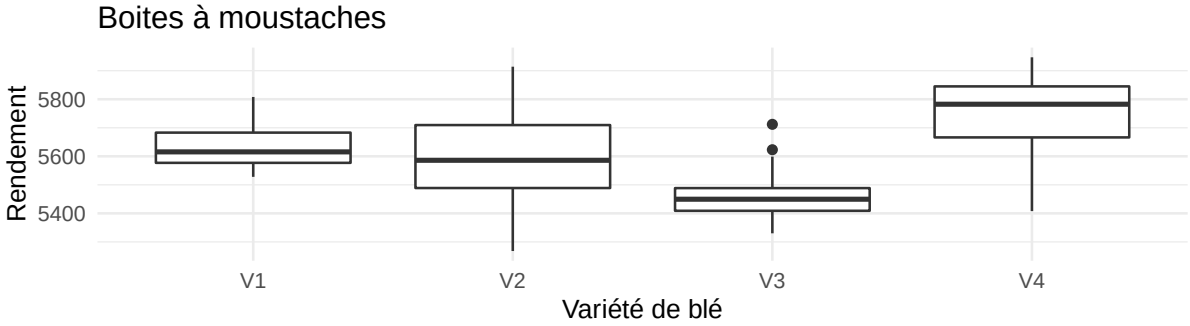
Exercice 3. ANOVA

Le fichier `ble.txt` contient les rendements de blé pour 80 parcelles en fonction de la variété de blé (variable `variete` avec les modalités `V1`, `V2`, `V3` ou `V4`) et de la présence ou non de traitement phytosanitaire (variable `phyto` avec les modalités `avec` ou `sans`).

```
ble <- read.table("ble.txt" ,header = TRUE, sep = ";", dec = ".")
head(ble)

##   parcelle variete phyto  rdt
## 1        1      V1  Avec 5652
## 2        2      V1  Avec 5583
## 3        3      V1  Avec 5612
## 4        4      V1  Avec 5735
## 5        5      V1  Avec 5704
## 6        6      V1  Avec 5544
```

1. On veut étudier ici l'influence de la variété de blé sur le rendement. On peut visualiser l'influence de la variété en affichant ces boîtes à moustaches :



(a) On fait la régression linéaire suivante:

```
lm_variete <- lm(rdt ~ variete, data = ble)
```

Écrire le modèle linéaire associé à cette commande.

Soit R_{ik} la variable aléatoire représentant le rendement de la parcelle i ($1 \leq i \leq n = 80$) plantée avec la variété k ($1 \leq k \leq K = 4$). Le modèle s'écrit:

$$R_{ik} = \mu + \beta_k + \epsilon_i$$

avec les ϵ_i iid gaussiens de variance inconnue σ^2, $\beta_1 = 0$, et β_k la différence de moyenne entre le groupe 1 de référence et le groupe $k > 1$.

(b) Écrivez la statistique du test de Student ainsi que les hypothèses associées à la nullité des différents coefficients du modèle. Concluez en utilisant la sortie `R` suivante:

```r
summary(lm_variete)
```

```
##
## Call:
## lm(formula = rdt ~ variete, data = ble)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -344.20  -69.30   -6.60   89.15  329.90
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5633.80      26.30 214.211  < 2e-16 ***
## varieteV2     -49.70      37.19  -1.336  0.18546
## varieteV3    -169.20      37.19  -4.549    2e-05 ***
## varieteV4     118.40      37.19   3.183  0.00211 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 117.6 on 76 degrees of freedom
## Multiple R-squared:  0.4476,Adjusted R-squared:  0.4258
## F-statistic: 20.53 on 3 and 76 DF,  p-value: 7.674e-10
```

Le test de Student permet de tester la nullité de chaque coefficient du modèle, *i.e.* pour chaque k, $\mathcal{H}_0 : \beta_k = 0$ ou $\mathcal{H}_0 : \mu = 0$. Il utilise la statistique $\frac{\hat{\beta}_k}{\hat{\sigma}_k}$, $\frac{\hat{\mu}}{\hat{\sigma}}$. On rejette significativement $\mathcal{H}_0$ pour μ, β_3 et β_4.

(c) En utilisant la commande précédente uniquement, pouvez-vous répondre aux questions suivantes :

- Les variétés `V1` et `V2` ont elles un rendement significativement différent ?

  Non, la p-valeur du test $\beta_2 = 0$ est 0.18.

- Les variétés `V1` et `V3` ont elles un rendement significativement différent ?

  Oui, la p-valeur du test $\beta_3 = 0$ est 2.10^{-5}.

- Les variétés `V3` et `V4` ont elles un rendement significativement différent ?

  Les boîtes à moustaches suggère qu'il y a bien une différence significative entre β_3 et β_4. Cependant aucun test de la sortie précédente ne permet de le justifier.

Si vous n'avez pas assez d'élément pour répondre à une question, indiquez pourquoi vous ne pouvez pas y répondre. Pour chaque questions, interprétez le résultat au vu des boîtes à moustaches ci-dessus.

(d) On procède à une analyse de la variance:

```r
anova(lm(rdt ~ 1, data = ble), lm_variete)
```

```
## Analysis of Variance Table
##
## Model 1: rdt ~ 1
## Model 2: rdt ~ variete
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
```

```
## 1     79 1903232
## 2     76 1051387  3    851845 20.525 7.674e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

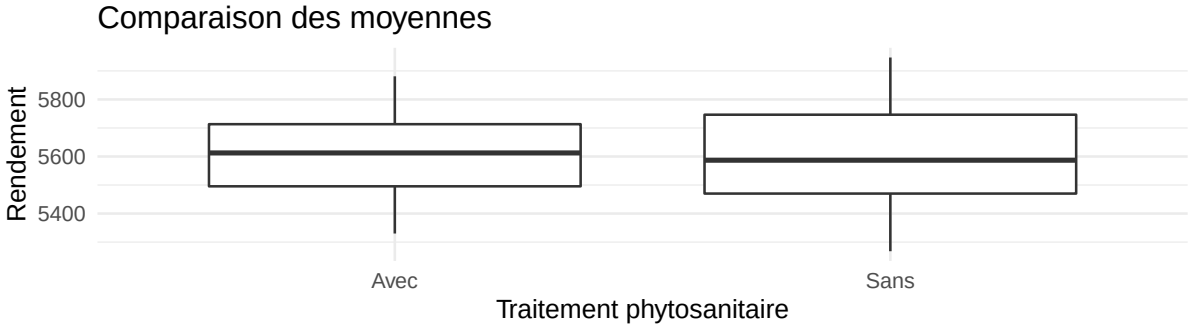
Quelles hypothèses (donner $\mathcal{H}_0$ et $\mathcal{H}_1$) a-t-on testées? Donnez la statistique de test. Quel modèle faut-il conserver? La variété de blé influe-t-elle sur le rendement? Aurait-on pu obtenir ce résultat directement à partir de la sortie de la fonction `summary` ? Si oui, indiquez précisément comment.

$\mathcal{H}_0 : \beta_2 = \beta_3 = \beta_4 = 0$, $\mathcal{H}_1 : \exists 2 \leq k \leq 4 | \ \beta_k \neq 0$. La statistique de test est:

$$F = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2/(4-1)}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-4)},$$

où $\hat{\mathbf{y}}_0 = \bar{y}\mathbf{1}$ est la valeur ajustée avec le modèle ne contenant que l'intercepte.
On rejette l'hypothèse nulle suivant laquelle la variété n'a pas d'influence sur le rendement (p-valeur très faible).
Ce test correspond exactement a test de Fisher de la nullité des coefficients hors de l'intecepte d'une régression linéaire. On retrouve donc ce même test dans la sortie de `summary`, ligne `F-statistic`.

2. Réalisons maintenant l'Analyse de la Variance sur le pesticide utilisé :



(a) On fait la régression linéaire suivante:

```
lm_phyto <- lm(rdt~ -1 + phyto, data  =ble)
```

Écrire le modèle linéaire associé à cette commande.

Soit $R_{i\ell}$ la variable aléatoire représentant le rendement de la parcelle i ($1 \leq i \leq n = 80$) traitée suivant la procédure ℓ ($1 \leq \ell \leq L = 2$). Le modèle s'écrit:

$$R_{i\ell} = \alpha_\ell + \epsilon_i$$

avec les ϵ_i iid gaussiens de variance inconnue σ^2, et α_ℓ la moyenne du groupe ℓ.

(b) En utilisant le résultat de la fonction `summary` ci-dessous uniquement, pouvez-vous répondre à la question suivante: "Les parcelles traitées et non traitées ont elles un rendement significativement différent ?" Si vous n'avez pas assez d'élément pour y répondre, indiquez pourquoi. Interprétez le résultat au vu des boîtes à moustaches ci-dessus.

On ne peut pas répondre à cette question, car les coefficients donnent les myennes de chaque groupe, et non les différences entre les deux groupes. Les tests sur les coefficients ne permettent donc pas de conclure à une différence ou non. D'après les boîtes à moustache, il semble qu'il n'y ai pas beaucoup de différence.

```r
summary(lm_phyto)
```

```
##
## Call:
## lm(formula = rdt ~ -1 + phyto, data = ble)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -337.12 -127.95   -4.18  106.02  341.88
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## phytoAvec  5612.23      24.69   227.3   <2e-16 ***
## phytoSans  5605.12      24.69   227.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 156.2 on 78 degrees of freedom
## Multiple R-squared:  0.9992,Adjusted R-squared:  0.9992
## F-statistic: 5.16e+04 on 2 and 78 DF,  p-value: < 2.2e-16
```

(c) On procède à une analyse de la variance:

```r
anova(lm(rdt ~ 1, data = ble), lm_phyto)
```

```
## Analysis of Variance Table
##
## Model 1: rdt ~ 1
## Model 2: rdt ~ -1 + phyto
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     79 1903232
## 2     78 1902223  1    1008.2 0.0413 0.8394
```

Quelles hypothèses (donner $\mathcal{H}_0$ et $\mathcal{H}_1$) a-t-on testées? Quel modèle faut-il conserver? La présence de traitement phytosanitaire a-t-il un effet sur le rendement?

On trouve ici une p-valeur de 0.8, ce qui est très au-dessus de 5%. On ne rejette donc pas l'hypothèse $H_0 : \alpha_1 = \alpha_2$ selon laquelle les rendements sont égaux.

3. Jusqu'ici, nous avons étudié les 2 facteurs (`variete` et `phyto`) séparément. Cependant, la variété et le traitement phytosanitaire peuvent avoir des interactions qui influent sur le rendement.

(a) On fait la régression linéaire suivante:

```r
lm_variete_phyto <- lm(rdt ~ variete*phyto, data = ble)
lm_variete_phyto
```

```
##
## Call:
## lm(formula = rdt ~ variete * phyto, data = ble)
##
## Coefficients:
##         (Intercept)           varieteV2           varieteV3
##              5628.1               -34.4               -167.6
```

```
##           varieteV4            phytoSans  varieteV2:phytoSans
##               138.5                 11.4                -30.6
## varieteV3:phytoSans  varieteV4:phytoSans
##                -3.2                -40.2
```

Écrire le modèle linéaire associé à cette commande.

Soit $R_{ik\ell}$ la variable aléatoire représentant le rendement de la parcelle i ($1 \leq i \leq n = 80$) plantée avec la variété k ($1 \leq k \leq K = 4$), et traitée suivant la procédure ℓ ($1 \leq \ell \leq L = 2$).

$$R_{i,k,l} = \mu + \alpha_k + \beta_\ell + \gamma_{k\ell} + \epsilon_i$$

Avec, pour tout $1 \leq k \leq K$ et $1 \leq \ell \leq L$:

$$\alpha_1 = 0, \qquad \beta_1 = 0, \qquad \gamma_{1\ell} = 0, \qquad \gamma_{k,1} = 0.$$

(b) On procède à une analyse de la variance à deux facteurs:

```
anova(lm_variete_phyto)
```

```
## Analysis of Variance Table
##
## Response: rdt
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## variete        3  851845  283948 19.5749 2.205e-09 ***
## phyto          1    1008    1008  0.0695    0.7928
## variete:phyto  3    5968    1989  0.1371    0.9375
## Residuals     72 1044411   14506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En utilisant la commande précédente, pouvez-vous répondre aux questions suivantes:

- L'ajout du facteur `phyto` à un modèle contenant déjà le facteur `variete` améliore-t-il significativement le modèle ?

  Non. La p-valeur des interactions (79.28%) est très largement supérieure à 5%.

- L'ajout du facteur `variete` à un modèle contenant déjà le facteur `phyto` améliore-t-il significativement le modèle ?

  On ne peut pas conclure, car il s'agit d'une table de type I.

- L'ajout de l'interaction entre les deux facteurs à un modèle contenant déjà les facteurs `variete` et `phyto` améliore-t-il significativement le modèle ?

  Non. La p-valeur des interactions (93.75%) est très largement supérieure à 5% ; on en déduit donc que les interactions n'ont pas d'impact sur le rendement.

Si vous n'avez pas assez d'élément pour répondre à une question, indiquez pourquoi vous ne pouvez pas y répondre.

(c) Aurait-on obtenu des résultats différents aux questions précédentes en exécutant la commande suivante ?

```
anova(lm(rdt ~ phyto*variete, data = ble))
```

Justifiez.

Non, car la table de type I n'est pas symmetrique. On aurait pu répondre à la question 2 ci-dessus.