Examen Final – Session 2

Durée 2h. 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. Régression Multiple.

On souhaite expliquer la hauteur y (en mètres) d'un arbre en fonction de sa circonférence x (en centimètres) à 1m30 du sol et de la racine carrée de celle-ci. On a relevé $n = 1429$ couples $(x_i; y_i)$. On considère donc le modèle de régression suivant :

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 \sqrt{x_i} + \epsilon_i \qquad 1 \leq i \leq n.$$

Les ϵ_i sont des variables aléatoires indépendantes, de loi normale centrée admettant la même variance σ^2. En posant :

$$\mathbf{x} = \begin{pmatrix} 1 & x_1 & \sqrt{x_1} \\ \vdots & \vdots & \vdots \\ 1 & x_n & \sqrt{x_n} \end{pmatrix} \quad \text{et} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

on a observé (approximativement):

$$\mathbf{x}^T\mathbf{x} = \begin{pmatrix} ? & ? & 9800 \\ ? & 3.3 \times 10^6 & ? \\ ? & 4.7 \times 10^5 & 6.8 \times 10^4 \end{pmatrix} \quad , \quad \mathbf{x}^T\mathbf{y} = \begin{pmatrix} 3 \times 10^4 \\ 1.5 \times 10^6 \\ 2.1 \times 10^5 \end{pmatrix} \quad , \quad \mathbf{y}^T\mathbf{y} = 6.5 \times 10^5.$$

1. Déterminer (sans faire de calcul, mais en explicitant votre raisonnement) les valeurs manquantes ("?") dans la matrice $\mathbf{x}^T\mathbf{x}$.

$$\mathbf{x}^T\mathbf{x} = \begin{pmatrix} 1429 & 6.8 \times 10^4 & 9800 \\ 6.8 \times 10^4 & 3.3 \times 10^6 & 4.7 \times 10^5 \\ 9800 & 4.7 \times 10^5 & 6.8 \times 10^4. \end{pmatrix}$$

2. Que valent la circonférence moyenne empirique $\bar{\mathbf{x}}$ et la hauteur moyenne empirique $\bar{\mathbf{y}}$? (Donnez les formules, et faites le calcul "à la louche" de manière approchée.)

$$\bar{\mathbf{x}} = 47.3477957 \qquad \bar{\mathbf{y}} = 21.2123863.$$

On effectue les commandes suivantes dans `R`. Pour toutes les questions suivantes, vous répondrez en utilisant les résultats des commandes ci-dessous uniquement.

```
eucalyptus <- read.csv("eucalyptus.csv")
fit <- lm(Height ~ Circumference + sqrt(Circumference), data = eucalyptus)
coef(fit)

##         (Intercept)       Circumference sqrt(Circumference)
##         -24.3520033          -0.4829455           9.9868881

sum(residuals(fit)^2) / (nobs(fit) - 0:3)
```

```
## [1] 1.288073 1.288975 1.289878 1.290783

confint(fit)

##                             2.5 %     97.5 %
## (Intercept)          -29.4805676 -19.223439
## Circumference         -0.5965919  -0.369299
## sqrt(Circumference)    8.4561795  11.517597

newdata <- data.frame(Circumference = c(49, 20))
predict(fit, newdata, interval = "conf", level = 0.95)

##        fit      lwr      upr
## 1 21.89189 21.82025 21.96352
## 2 10.65181 10.07558 11.22804

predict(fit, newdata, interval = "pred", level = 0.95)

##        fit       lwr      upr
## 1 21.89189 19.662077 24.12169
## 2 10.65181  8.349862 12.95376
```

3. Donnez les expressions et les valeurs des estimateurs des moindres carrés $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}^2$.

$$\hat{\boldsymbol{\beta}} = (-24.3520033, -0.4829455, 9.9868881)^T \qquad \hat{\sigma}^2 = 1.2907827.$$

4. Donnez les expressions et les valeurs des estimateurs du maximum de vraisemblance $\hat{\boldsymbol{\beta}}^{mv}$ et $\hat{\sigma}^2_{mv}$.

$$\hat{\boldsymbol{\beta}}^{mv} = \hat{\boldsymbol{\beta}} = (-24.3520033, -0.4829455, 9.9868881)^T \qquad \hat{\sigma}^2_{mv} = 1.2880729.$$

5. Pouvez-vous donner un intervalle de confiance à 95% de $\hat{\beta}_1$, $\hat{\beta}_2$ et $\hat{\beta}_3$ et $\hat{\sigma}^2$? Pouvez-vous donner un intervalle de confiance à 97.5% de ces paramètres ? À 99% ?

   L'intervalle de confiance à 95% de $\hat{\beta}_1$, $\hat{\beta}_2$ et $\hat{\beta}_3$ est donné par la commande `confint(fit)`. Les autres ne peuvent pas être calculés à partir des commandes fournie.

6. Pouvez-vous tester hypothèse $\beta_2 = 0$ contre $\beta_2 \neq 0$ au niveau de risque 10% ? Au niveau de risque 5% ? Au niveau de risque 2.5% ?

   L'intervalle de confiance à 95% de $\hat{\beta}_2$ ne contient pas 0, on peut donc rejeter l'hypothèse nulle $\beta_2 = 0$ au niveau de risque 5%. Si l'on est prêt à prendre un risque supérieur de 10%, on pourra toujours rejeter. Par contre, pour un risque inférieur de 2.5%, on ne peut pas conclure à partir des commandes ci-dessus.

7. On observe un nouvel arbre d'une circonférence de $x_{n+1} = 49\ cm$. Pouvez-vous prédire sa hauteur y_{n+1} ? Au niveau de confiance à 95%, pouvez-vous donner un intervalle pour votre prédiction ?

   Le résultat est donné par la commande `predict(fit, newdata, interval = "pred", level = 0.95)`.

8. Mêmes questions pour un nouvel arbre d'une circonférence de $x_{n+2} = 20\ cm$. Pour lequel de ces deux arbres l'incertitude est-elle la plus grande ? Pouvait-on s'y attendre ?

   Idem. L'intervalle est plus grand, car 25 est plus éloingé du centre de gravité du point que 49 ($\bar{\mathbf{x}} = 47.3477957$).

Exercice 2. Tests de Student et de Fisher

On considère le modèle de régression linéaire classique à n variables et p prédicteurs:

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

On souhaite montrer l'équivalence entre les tests de Student et de Fisher pour tester la nullité du dernier coefficient:

$$\mathcal{H}_0 : \beta_p = 0 \qquad \text{contre} \qquad \mathcal{H}_1 : \beta_p \neq 0.$$

1. Soient $U \sim \mathcal{N}(0,1)$ et $V \sim \chi^2_k$ deux variables aléatoires indépendantes (avec k un entier strictement positif). Quelle est la loi de $T = \frac{U}{\sqrt{V/k}}$? Quelle est la loi de $F = T^2$?

   Par définition, T suit une loi de Student $\mathcal{T}_k$.
   De plus, $W = U^2 \sim \chi^2_1$ par définition. Donc $F = T^2 = \frac{W/1}{V/k}$ suit une loi de Fisher $\mathcal{F}^1_k$.

2. Rappelez les hypothèses classiques du modèle linéaire gaussien. Quelles sont les dimensions de $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ et $\boldsymbol{\epsilon}$? On se place sous ces hypothèses dans toute la suite.

   Voir le cours.

3. Rappelez la statistique T_p du test de Student pour la nulité du coefficient β_p, et sa loi sous l'hypothèse $\mathcal{H}_0$.

   D'après le cours:

   $$T_p = \frac{\hat{\beta}_p}{\sqrt{\hat{\sigma}^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{pp}}},$$

   avec $\hat{\sigma}^2 = \frac{1}{n-p}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$. Sous $\mathcal{H}_0$, T_p suit une loi de Student à $n-p$ degrés de libertés.

4. On décompose $\mathbf{X}$ en blocs:

   $$\mathbf{X} = [\mathbf{X}_0\ \mathbf{X}_p] \quad \text{avec} \quad \mathbf{X}_0 = [\mathbf{X}_1\ \cdots\ \mathbf{X}_{p-1}],$$

   où $\mathbf{X}_0$ est la matrice de taille $n \times (p-1)$ des $(p-1)$ premières colonnes de $\mathbf{X}$.
   Écrivez les deux modèles emboîtés qui correspondent au test de la nullité du coefficient β_p. Donnez la statistique F_p du test de Fisher correspondant, et sa loi sous $\mathcal{H}_0$.

   On note

   $$\text{Modèle 1} : \mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}'$$
   $$\text{Modèle 2} : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

   La statistique de test s'écrit:

   $$F_p = \frac{\|\hat{\mathbf{y}} - \mathbf{X}_0\hat{\boldsymbol{\beta}}_0\|^2/(p-(p-1))}{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(n-p)} = \frac{\|\hat{\mathbf{y}} - \mathbf{X}_0\hat{\boldsymbol{\beta}}_0\|^2}{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(n-p)}$$

   Sous $\mathcal{H}_0$, F_p suit une loi de Fisher à 1, $n-p$ degrés de libertés.

5. En utilisant la décomposition $\mathbf{X} = [\mathbf{X}_0\ \mathbf{X}_p]$, donnez la matrice $\mathbf{X}^T\mathbf{X}$ sous forme de 4 blocs.

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

6. On admet le lemme d'inversion matricielle par blocs suivant:
Soit $\mathbf{M}$ *une matrice par blocs,* $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$, *avec* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ *de dimensions respectives* $q \times q$, $q \times r$, $r \times q$, *et* $r \times r$. *On suppose* $\mathbf{M}$ *et* $\mathbf{A}$ *inversibles. Alors, on peut écrire* $\mathbf{M}^{-1}$ *sous la forme:*

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{G} & \mathbf{H} \end{pmatrix}, \quad \textit{avec} \quad \mathbf{H}^{-1} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}.$$

Montrez la relation suivante :

$$\frac{1}{[(\mathbf{X}^T\mathbf{X})^{-1}]_{pp}} = \mathbf{X}_p^T\mathbf{X}_p - \mathbf{X}_p^T\mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T\mathbf{X}_p$$

D'après le lemme d'inversion matricielle:

$$[[(\mathbf{X}^T\mathbf{X})^{-1}]_{pp}]^{-1} = \mathbf{X}_p^T\mathbf{X}_p - (\mathbf{X}_p^T\mathbf{X}_0)(\mathbf{X}_0^T\mathbf{X}_0)^{-1}(\mathbf{X}_0^T\mathbf{X}_p).$$

7. On note $\mathbf{P}_0$ la matrice de projection orthogonale sur l'espace $\mathcal{M}_0$ engendré par les $p-1$ colonnes de $\mathbf{X}_0$, et $\mathbf{P}$ la matrice de projection orthogonale sur $\mathcal{M}$ engendré par les p colonnes de $\mathbf{X}$.
Donnez les expressions de $\mathbf{P}_0$ et $\mathbf{P}$ en fonction de, respectivement, $\mathbf{X}_0$ et $\mathbf{X}$, puis montrez la relation suivante :

$$\frac{1}{[(\mathbf{X}^T\mathbf{X})^{-1}]_{pp}} = \mathbf{X}_p^T\left(\mathbf{I}_n - \mathbf{P}_0\right)\mathbf{X}_p.$$

On a $\mathbf{P}_0 = \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T$, d'où:

$$\frac{1}{[(\mathbf{X}^T\mathbf{X})^{-1}]_{pp}} = \mathbf{X}_p^T\mathbf{X}_p - \mathbf{X}_p^T\mathbf{P}_0\mathbf{X}_p = \mathbf{X}_p^T(\mathbf{I}_n - \mathbf{P}_0)\mathbf{X}_p.$$

8. On décompose $\hat{\boldsymbol{\beta}}$ en deux blocs: $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\beta}_p \end{pmatrix}$. Montrez : $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}_0\hat{\boldsymbol{\beta}}_0 + \mathbf{X}_p\hat{\beta}_p$.

$$\mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{X}_0\ \mathbf{X}_p)\begin{pmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\beta}_p \end{pmatrix} = \mathbf{X}_0\hat{\boldsymbol{\beta}}_0 + \mathbf{X}_p\hat{\beta}_p.$$

9. On note $\hat{\mathbf{y}}$ et $\hat{\mathbf{y}}_0$ les projetés orthogonaux de $\mathbf{y}$ sur $\mathcal{M}$ et $\mathcal{M}_0$. Justifiez l'égalité:

$$\hat{\mathbf{y}}_0 = \mathbf{P}_0\hat{\mathbf{y}}.$$

En déduire:

$$\hat{\mathbf{y}} - \hat{\mathbf{y}}_0 = (\mathbf{I}_n - \mathbf{P}_0)\mathbf{X}_p\hat{\beta}_p$$

Du fait que les espaces vectoriels sont emboîtés:

$$\hat{\mathbf{y}}_0 = \mathbf{P}_0\mathbf{y} = \mathbf{P}_0\mathbf{P}\mathbf{y} = \mathbf{P}_0\hat{\mathbf{y}}.$$

D'où:

$$\begin{aligned}\hat{\mathbf{y}} - \hat{\mathbf{y}}_0 &= \hat{\mathbf{y}} - \mathbf{P}_0\hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{P}_0)\hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{P}_0)\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= (\mathbf{I}_n - \mathbf{P}_0)(\mathbf{X}_0\hat{\boldsymbol{\beta}}_0 + \mathbf{X}_p\hat{\beta}_p) = (\mathbf{I}_n - \mathbf{P}_0)\mathbf{X}_p\hat{\beta}_p.\end{aligned}$$

10. Montrez que $T_p^2 = F_p$. En déduire l'équivalence des deux tests.

$$\begin{aligned}T_p^2 &= \frac{\hat{\beta}_p^2}{\hat{\sigma}^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{pp}} = \frac{\mathbf{X}_p^T(\mathbf{I}_n - \mathbf{P}_0)\mathbf{X}_p\hat{\beta}_p^2}{\hat{\sigma}^2} \\ &= \frac{\|(\mathbf{I}_n - \mathbf{P}_0)\mathbf{X}_p\hat{\beta}_p\|^2}{\hat{\sigma}^2} = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2}{\hat{\sigma}^2} \\ &= F_p\end{aligned}$$

Les deux tests sont donc bien équivalents.

Exercice 3. Regression Ridge

On considère le modèle de régression $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ où $\mathbf{Y}$ est un vecteur de $\mathbb{R}^n$, $\mathbf{X}$ est une matrice de taille $n \times p$, $\boldsymbol{\beta}$ un vecteur de $\mathbb{R}^p$ et $\boldsymbol{\epsilon}$ un vecteur de $\mathbb{R}^n$ de variables aléatoires supposées iid, centrées et de variance σ^2.

1. On suppose que $p > n$. Que dire de l'estimateur des moindres carrés dans ce cas là ?

   Si $p > n$, alors la matrice $\mathbf{X}$ ne peut pas être de plein rang ($\text{rg}(\mathbf{X}) < p$), et $\mathbf{X}^T\mathbf{X}$ n'est pas inversible. L'estimateur des moindres carrés n'est donc pas défini.

2. On appelle estimateur *ridge* de paramètre λ ($\lambda > 0$) l'estimateur de $\boldsymbol{\beta}$ suivant:

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}} \left\{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right\}.$$

   Exprimez $\hat{\boldsymbol{\beta}}_\lambda$ en fonction de $\mathbf{Y}$, $\mathbf{X}$ et λ. Est-ce un estimateur linéaire en $\mathbf{Y}$? Que dire de cet estimateur lorsque $p < n$?

   En annulant la différentielle en β de l'expression, on obtient:

$$2\mathbf{X}^T\mathbf{X}\hat{\beta}_\lambda - 2\mathbf{X}^T\mathbf{Y} + 2\lambda\hat{\beta}_\lambda = 0 \iff \hat{\beta}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^T\mathbf{Y}$$

   avec $(\mathbf{X}^T\mathbf{X} + \lambda I_p)$ toujours inversible lorsque $\lambda > 0$, même lorsque $\mathbf{X}^T\mathbf{X}$ n'est pas de plein rang, et en particulier si $p > n$. C'est bien un estimateur linéaire en $\mathbf{Y}$.

3. Calculez l'espérance et la variance de l'estimateur ridge. Est-il sans biais ?

$$\mathbb{E}[\hat{\beta}_\lambda] = (\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^T\mathbb{E}[\mathbf{Y}] = (\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^T\mathbf{X}\beta \neq \beta.$$

$$\mathbb{V}[\hat{\beta}_\lambda] = (\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^T\mathbb{V}[\mathbf{Y}]\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}$$

4. On suppose ici que $p = 1$, c'est-à-dire que le modèle de régression est simple avec une seule variable (non nulle) et sans constante. Montrer qu'il existe une valeur de λ pour laquelle le risque de l'estimateur ridge est strictement inférieur au risque de l'estimateur des moindres carrés. Est-ce en contradiction avec le théorème de Gauss-Markov ?

   Lorsque $p = 1$, on a $\mathbf{X}^T\mathbf{X} = \sum_{i=1}^n x_i^2$, donc $\mathbf{X}^T\mathbf{X} = \alpha > 0$ est un réel strictement positif. On obtient donc $\mathbb{V}[\hat{\boldsymbol{\beta}}_\lambda] = \frac{\alpha}{(\alpha+\lambda)}$ et $\mathbb{V}[\hat{\boldsymbol{\beta}}] = \frac{1}{\alpha}$ pour l'estimateur des moindres carrés. Si l'on prend par exemple $\lambda = \alpha$, on trouve: $\mathbb{V}[\hat{\boldsymbol{\beta}}_\lambda] = \frac{\alpha}{4\alpha^2} = \frac{1}{4}\frac{1}{\alpha} < \frac{1}{\alpha} = \mathbb{V}[\hat{\boldsymbol{\beta}}]$.
   L'estimateur ridge est linéaire, mais n'est pas sans biais, il ne vérifie donc pas les conditions d'application du théorème de Gauss-Markov, et le fait que sa variance soit plus faible que le BLUE n'est pas contradictoire.