Examen Intra MAT7381

(2 heures)

Mars 2020

Exercice 1. *On a simulé les données suivantes : $x_i = i$ pour $i = 1, \cdots, 50$, les ε_i sont tirés suivant des lois $\mathcal{N}(0,1)$, et $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ avec $\beta_0 = 2$ et $\beta_1 = 1$. On obtient la sortie donnée ci-dessous. Que vaut le biais de $\widehat{\beta}_1$, estimé par moindres carrés (associé à la variable x) ?*

```
> x = 1:50
> epsilon = rnorm(50)
> y = 2+x+epsilon
> summary(lm(y~x))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.250      0.281    7.99 2.26e-10 ***
x              0.980      0.014    69.9  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Exercice 2. *On travaille ici avec le modèle linéaire homoscédastique $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ satisfaisant les hypothèses $\mathcal{H}_1 - \mathcal{H}_2$ du cours, et soit $\hat{\boldsymbol{\beta}}$ l'estimateur des MCO:*

1. *Si les covariables sont orthogonales, rappelez ce que vaut $\mathrm{Var}(\hat{\beta}_j)$ pour $j = 1, \ldots, p$.*

2. *Supposons $p = 2$, montrer que $\mathrm{Var}(\hat{\beta}_1) \geq \dfrac{\sigma^2}{\mathbf{x}_1^\top \mathbf{x}_1}$ (indication: calculer le 1er terme diagonal de $(\mathbf{X}^\top \mathbf{X})^{-1}$).*

Exercice 3. *Considérons* $\mathbf{X}$ *une matrice* $n \times p$ *de covariances, et notons* $\mathbf{x}_i^\top$ *la* i*ème ligne de la matrice* $\mathbf{X}$*, et* $\mathbf{X}_{(i)}$ *la matrice* $(n-1) \times p$ *la matrice* $\mathbf{X}$ *privée de la ligne* i*. De manière similaire,* $\mathbf{y}$ *est un vecteur de taille* n*,* y_i *désigne la* i*ème observation, et* $\mathbf{y}_{(i)}$ *le vecteur de taille* $n-1$ *obtenu à partir de* $\mathbf{y}$ *en enlevant la* i*ème observation. On notera* H_{ii} *le* i*ème terme sur la diagonale de la matrice* H *de projection (tel que les prévision obtenues par moindres carrés s'écrivent* $\widehat{\mathbf{y}} = H\mathbf{y}$*). On appelle* M *le modèle construit à partir des* n *observations* $\{\mathbf{y}, \mathbf{X}\}$ *et* M_i *le modèle construit à partir des* $n-1$ *observations* $\{\mathbf{y}_{(i)}, \mathbf{X}_{(i)}\}$

1. *Montrer que* $\mathbf{X}^\top\mathbf{X} = \mathbf{X}_{(i)}^\top\mathbf{X}_{(i)} + \mathbf{x}_i\mathbf{x}_i^\top$ *pour tout* $i = 1, \ldots, n$
2. *Montrer que* $\mathbf{X}_{(i)}^\top\mathbf{y}_{(i)} = \mathbf{X}^\top\mathbf{y} + \mathbf{x}_i y_i$ *pour tout* $i = 1, \ldots, n$
3. *Montrez que*
$$\left(\mathbf{X}_{(i)}^\top\mathbf{X}_{(i)}\right)^{-1} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1} + \frac{1}{1-H_{ii}}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x}_i\mathbf{x}_i^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$$
4. *Montrez que la prévision de l'observation* $\mathbf{x}_i$ *à l'aide du modèle* M_i *est*
$$\widehat{y}_i^{(i)} = \frac{1}{1-H_{ii}}\widehat{y}_i - \frac{H_{ii}}{1-H_{ii}}y_i$$
où $\widehat{y}_i$ *est la prévision obtenue par* M*.*
5. *En notant* $\widehat{\varepsilon}_i$ *les résidus estimés du modèle* M*, les résidus studentisés sont*
$$\widehat{t}_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{1-H_i i}}$$
Montrez qu'on peut les écrire
$$\widehat{t}_i = \frac{\widehat{y}_i - \widehat{y}_i^{(i)}}{\widehat{\sigma}\sqrt{1+\mathbf{x}_i^\top(\mathbf{X}_{(i)}^\top\mathbf{X}_{(i)})^{-1}\mathbf{x}_i}}$$
6. *Si* $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbb{I})$*, quelle est la loi des* $\widehat{t}_i$

Indice*: on admettra que si* $\mathbf{M}$ *est une matrice symétrique inversible* $p \times p$*, et si* $\mathbf{u}$ *et* $\mathbf{v}$ *sont deux vecteurs de dimension* p*,*
$$\left(\mathbf{M} + \mathbf{u}\mathbf{v}^\top\right)^{-1} = \mathbf{M}^{-1} - \frac{\mathbf{M}^{-1}\mathbf{u}\mathbf{v}^\top\mathbf{M}^{-1}}{1+\mathbf{u}^\top\mathbf{M}\mathbf{v}}$$

Exercice 4. *Soit* $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ *un modèle de régression linéaire homoscédastique gaussien. On suppose que la matrice de design* $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2)$ *est de taille* $(n, 2)$*. Le vecteur des paramètres* $\boldsymbol{\beta}$ *est donc de dimension* 2*. On notera* σ^2 *le paramètre de variance du bruit. Enfin, on supposera en outre que* $\mathbf{x}_1$*,* $\mathbf{x}_2$ *sont centrés et que, pour* ρ *est un paramètre réel,*

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

1. *Quelle est la condition sur* n *et* ρ *pour que la matrice de design* $\mathbf{X}$ *soit de rang plein et que* $\mathbf{X}^\top\mathbf{X}$ *soit définie positive? Cette condtion sera supposée par la suite.*

2. *Soit* $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)$ *l'estimateur obtenu par moindres carrés ordinaires de ce modèle. Montrez que pour* $j = 1, 2$

$$\hat{\beta}_j = \frac{1}{1-\rho^2}\mathbf{z}_j^\top \mathbf{Y}, \qquad \text{avec } \mathbf{z}_1 = \mathbf{x}_1 - \rho\mathbf{x}_2 \text{ et } \mathbf{z}_2 = \mathbf{x}_2 - \rho\mathbf{x}_1.$$

3. *Calculez* $\mathrm{Var}(\hat{\beta}_j)$ *en fonction de* σ^2 *et* ρ*, pour* $j = 1, 2$*.*

4. *On rappelle que le critère du facteur d'augmentation de la variance du* j*ème régresseur* VIF_j *vaut* $\left((\mathbf{X}^\top\mathbf{X})^{-1}\right)_{jj}$*. Pour quelle condition sur* ρ*,* VIF_j *est-il supérieur à 4? supérieur à 10?*

5. *Soit* $\hat{\sigma}^2$ *l'estimateur de la variance du bruit (=* $RSS/(n-2)$*). Définir les statistiques des tests de significativité des paramètres* β_1 *et* β_2 *en fonction de* ρ*. Que se passe-t-il lorsque* $|\rho| \to 1$*? Commentez.*

Exercice 5. *On dispose de* n *observations* $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$*, et on estime un modèle linéaire par moindres carrés. Soit* $\mathbf{X}$ *la matrix* $n \times p$ *associée. On suppose que* $\mathcal{H}_1 - \mathcal{H}_2$ *sont vérifiées. On obtient une nouvelle observation* (y_{n+1}, bx_{n+1})*.*

1. *Montrez que l'erreur de prédiction* $e_{n+1} = Y_{n+1} - \hat{Y}_{n+1}$ *vérifie*

$$e_{n+1} = \varepsilon_{n+1} - \mathbf{x}_{n+1}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}, \quad \mathrm{Var}(e_{n+1}) = \sigma^2\left(1 + \mathbf{x}_{n+1}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{n+1}\right).$$

2. *Montrez également que dans le cas* $p = 2$*, i.e.* $\mathbf{x} = (1, x)$

$$\mathrm{Var}(e_{n+1}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum(x_i - \overline{x})^2}\right).$$