TD DE MODÈLES LINÉAIRES - FEUILLE 7
Validation de modèle

Exercice 1 (QCM - révision des chapitres précédents)

1. Nous pouvons justifier les MC quand $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ via l'application du maximum de vraisemblance :
   (a) oui,
   (b) non,
   (c) aucun rapport entre les deux méthodes.
2. Y a-t-il une différence entre les estimateurs $\hat{\beta}$ des MC et $\widetilde{\beta}$ du maximum de vraisemblance ?
   (a) oui,
   (b) non,
   (c) pas toujours, cela dépend de la loi des erreurs.
3. Y a-t-il une différence entre les estimateurs $\hat{\sigma}^2$ des MC et $\widetilde{\sigma}^2$ du maximum de vraisemblance quand $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$?
   (a) oui,
   (b) non,
   (c) pas toujours, cela dépend de la loi des erreurs.
4. Le rectangle formé par les intervalles de confiance de niveau α individuels de β_1 et β_2 correspond à la région de confiance simultanée de niveau α de la paire (β_1, β_2).
   (a) oui,
   (b) non,
   (c) cela dépend des données.
5. Nous avons n observations et p variables explicatives, nous supposons que ε suit une loi normale, nous voulons tester $\mathcal{H}_0 : \beta_2 = \beta_3 = \beta_4 = 0$. Quelle va être la loi de la statistique de test ?
   (a) $\mathcal{F}_{p-3,n-p}$,
   (b) $\mathcal{F}_{3,n-p}$,
   (c) une autre loi.

Exercice 2 (QCM)

1. Lors d'une régression multiple, la somme des résidus vaut zéro :
   (a) toujours,
   (b) jamais,
   (c) cela dépend des variables explicatives utilisées.
2. Les résidus studentisés sont-ils
   (a) homoscédastiques,
   (b) hétéroscédastiques,
   (c) on ne sait pas.

3. Un point levier peut-il être aberrant ?
   (a) toujours,
   (b) jamais,
   (c) parfois.
4. Un point aberrant peut-il être levier ?
   (a) toujours,
   (b) jamais,
   (c) parfois.
5. La distance de Cook est-elle basée sur un produit scalaire ?
   (a) oui,
   (b) non,
   (c) cela dépend des données.

Exercice 3 (Comparaison de modèles)
On effectue une régression de y sur deux variables explicatives x et z à partir d'un échantillon de n individus, c'est-à-dire que $X = [\mathbb{1}, \mathbf{x}, \mathbf{z}]$, où $\mathbb{1}$ est le vecteur de taille n composé de 1. On a obtenu le résultat suivant :

$$X'X = \begin{bmatrix} 5 & 3 & 0 \\ 3 & 3 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

1. Que vaut n ?
2. Que vaut le coefficient de corrélation linéaire empirique entre $\mathbf{x}$ et $\mathbf{z}$?
3. La régression par moindres carrés ordinaires a donné le résultat suivant

$$\hat{y}_i = -1 + 3x_i + 4z_i + \hat{\varepsilon}_i$$

   et la somme des carrés résiduelle vaut $\|\hat{\varepsilon}\|^2 = 3$.
   (a) Exprimer $X'Y$ en fonction de $(X'X)$ et $\hat{\beta}$, et calculer $X'Y$. En déduire $\bar{y}$.
   (b) Calculer $\|\hat{Y}\|^2$. En déduire $\|Y\|^2$.
   (c) Calculer la somme des carrés totale $\|Y - \bar{y}\mathbb{1}\|^2$, le coefficient de détermination R^2 et le coefficient de détermination ajusté.
4. On s'intéresse maintenant au modèle privé du régresseur z, c'est-à-dire $Y = X_0\beta_0 + \varepsilon_0$, où $X_0 = [\mathbb{1}, \mathbf{x}]$.
   (a) Déterminer $X_0'X_0$ et $X_0'Y$. En déduire $\hat{\beta}_0$.
   (b) Calculer $\|\hat{Y}_0\|^2$.
   (c) Justifier l'égalité $\|\hat{Y}_0\|^2 + \|\hat{\varepsilon}_0\|^2 = \|\hat{Y}\|^2 + \|\hat{\varepsilon}\|^2$. En déduire $\|\hat{\varepsilon}_0\|^2$, le coefficient de détermination R_0^2 et le coefficient de détermination ajusté.
5. On veut maintenant comparer les deux modèles précédents.
   (a) Effectuer un test de Fisher entre ces deux modèles grâce aux coefficients de détermination. Qu'en concluez-vous au niveau de risque 5% ?
   (b) Proposer un autre moyen d'arriver au même résultat.