

On dispose des données suivantes concernant la fréquentation des bus sur 25 villes françaises (voir fichier EXCEL). On veut utiliser ces données dans le cadre d'une étude sociologique. Les variables suivantes sont considérées:

NB est le nombre de passagers en millions,

CAR est le prix du carburant en euros,

REV est le revenu moyen par habitant en milliers d'euros.

On désire expliquer la variable "NB" linéairement par les variables "CAR" et "REV".

- 1) Représenter les différentes variables en utilisant des histogrammes. Etablir quelques statistiques. D'après ces résultats que peut-on remarquer pour la variable CAR ? Que peut-on craindre si on effectue une régression incluant une constante? Que pouvez vous remarquer en ce qui concerne la variable NB ?
- 2) On considère la régression suivante:

$$M1 : NB = \underline{a}_0^{(1)} + \underline{a}_1^{(1)}CAR + \underline{a}_2^{(1)}REV + E_1, \quad (1)$$

où E_1 est le vecteur contenant les erreurs. Estimer le modèle et étudiez la normalité des résidus (procédure "univariate" avec l'option "qq-plot"). Ecrire l'expression de la statistique de test de nullité d'un paramètre.

- a) Cas où les erreurs sont gaussiennes: Sous l'hypothèse que E_1 suit une loi normale $N(0, \sigma_E^2)$, rappeler la démarche permettant d'établir que cette statistique suit une loi t_{n-3} . Que décidez vous concernant l'hypothèse $\underline{a}_1^{(1)} = 0$, d'après le résultat du test de Student fourni par SAS?

- b) Cas où les erreurs ne sont pas gaussiennes: En se basant sur les résultats donnés par la procédure "univariate", on décide que l'hypothèse de normalité des erreurs n'est pas vérifiée. On construit dans la suite de la question 2 un test où l'on suppose simplement que les erreurs sont indépendantes entre elles et indépendantes de la matrice X sans être nécessairement gaussiennes. Nous supposons que les variables REV et CAR sont iid et que $E(X'X) < \infty$, et nous maintenons l'hypothèse que les erreurs sont iid et telles que $E(E_1) = 0$ d'écart-type σ_E . Montrer que

$$\hat{\underline{a}} - \underline{a} = \{n^{-1}(X'X)\}^{-1} \{n^{-1}X'E_1\}.$$

où $\underline{a} = (\underline{a}_0^{(1)}, \underline{a}_1^{(1)}, \underline{a}_2^{(1)})'$, $\hat{\underline{a}}$ l'estimateur des moindres carrés correspondant à $M1$ et n le nombre d'observations.

- c) En remarquant que l'élément k, j de la matrice $\{n^{-1}(X'X)\}$ s'écrit

$$n^{-1} \sum_{i=1}^n X_{ik}X_{ij}$$

et la k ème composante du vecteur $\{n^{-\frac{1}{2}}X'E_1\}$ s'écrit

$$n^{-\frac{1}{2}} \sum_{i=1}^n X_{ik}E_{1i},$$

spécifiez la limite de $\{n^{-1}(X'X)\}$ et $\{n^{-\frac{1}{2}}X'E_1\}$. Donner la limite de $n^{\frac{1}{2}}\{\hat{\underline{a}} - \underline{a}\}$ quand n tend vers l'infini.

- d) Sous l'hypothèse nulle, en déduire le comportement à la limite de

$$\left[\frac{n^{\frac{1}{2}}\hat{\underline{a}}_1^{(1)}}{\hat{\sigma}_E \sqrt{n^{-1}(X'X)_{jj}^{-1}}} \right]^2,$$

où $\hat{\sigma}_E$ est l'estimateur usuel de l'écart-type des erreurs et $(X'X)_{jj}^{-1}$ le j ème élément de la diagonale de la matrice $(X'X)^{-1}$. Construire le test permettant de tester la nullité du paramètre $\underline{a}_1^{(1)}$.

Commenter l'ensemble des résultats de la question 2.

- 3) Dans cette question on étudie les conséquences de la colinéarité entre deux variables explicatives sur la précision des estimateurs et des prévisions.
- Pour une ville A , on définit x_A le vecteur ayant pour composantes le revenu, le prix du carburant de cette ville et une composante égale à un pour la constante du modèle: $x_A = (1, REV_A, CAR_A)^\top$. Ainsi on écrit $\widehat{NB}_A = x_A^\top \hat{a}$, la prévision du nombre de passagers pour cette ville. Exprimer $\widehat{NB}_A - NB_A$ et donner la loi de cette quantité sous l'hypothèse $E_1 \sim \mathcal{N}(0, \sigma_E)$.
 - En utilisant l'estimateur de la variance usuel de E_1 , en déduire un intervalle de confiance pour la vraie valeur NB_A .
 - Dans le cas où il y a colinéarité, les valeurs de $(X'X)^{-1}$ sont explosives. Quelle est la conséquence sur la précision des estimateurs de la prévision? Peut-on négliger un problème de colinéarité?
- 4) On considère la variable "CAR" exprimée en centimes dans le modèle (1). Effectuer la transformation de variable et la regression. Que constatez-vous?
- 5) On décide finalement de considérer un modèle linéaire incluant la variable REV seulement:

$$M2 : NB = \underline{a}_0^{(2)} + \underline{a}_1^{(2)} REV + E_2.$$

Commenter les résultats. En éliminant les éventuelles valeurs aberrantes, peut-on affirmer que la variable REV est pertinente pour expliquer la variable NB ?

- 6) Le sociologue avec qui vous collaborez décide de ne pas supprimer les valeurs aberrantes des données pour des raisons sous jacentes à l'étude. Votre collègue veut faire une prévision du nombre de passagers pour la ville Tédém où le revenu moyen annuel est de 16500 euros. Trouver cette prévision $\widehat{NB}_{tedem} = E(NB_{tedem} | REV = 16500)$ et son intervalle de confiance. Trouver également l'intervalle de confiance pour le nombre de passagers NB_{tedem} .

Exercice d'approfondissement du cours.

Exercice 2: On procède a des essais très couteux de fiabilité d'un produit. Pour cela on considère les effets de deux facteurs A,B et leur interaction notée AB sur le comportement du produit. On envisage de considérer un modèle linéaire avec pour variable dépendante Y les résultats de l'expérience, et comme régresseurs les effets des facteurs, leurs interactions plus une constante qui sont consignés dans la matrice X .

- 1) Les expériences coûtant très cher, vos collègues ingénieurs n'ayant pas fait TDMM désirent effectuer moins de quatre expériences. Peut-on estimer les paramètres du modèle? Peut-on analyser les résultats de la régression dans ce cas?
- 2) Prenant en compte vos remarques vos collègues consentent à effectuer 4 expériences. Dans la matrice X du modèle linéaire ci-dessous la valeur 1 correspond à la présence de l'effet ou de l'interaction, 0 correspond à son absence. La première colonne correspond à la constante du modèle. Lancer une procédure "REG" que constatez vous?

$$X = \begin{pmatrix} Cste & A & B & AB \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

En utilisant la formule donnant les résidus, expliquer ce résultat.

- 3) Si vos collègues veulent faire une analyse complète des résultats d'expérience du point de vue statistique, que pouvez vous leur conseiller?