Correction CC, L3 économétrie (2017-2018)

Exercice 1 : Régression simple (9.5 points)

Partie 1

Question 1

Les données dont nous disposons sont des données en coupe transversales car nous observons des ventes de biens immobiliers à une date donnée.

Question 2

On note N le nombre d'observations de la base de données. La variance du prix est donnée par la formule suivante :

$$Var(P) = \frac{1}{N} * \sum_{n=1}^{21599} P_i^2 - \bar{P^2} \tag{1}$$

Dans un premier temps, on calcule $\bar{P}$:

$$\bar{P} = \frac{1}{N} * \sum_{i=1}^{21599} P_i$$

$$\bar{P} = \frac{1}{21599} * 11602.91 \approx 0.53$$

En appliquant la formule de la variance, on a :

$$Var(P) = \frac{1}{21599} * 8837.45 - 0.53^2 \approx 0.12$$

$$Var(P) = 0.12$$

De même pour la surface, on a :

$$Var(S) = \frac{1}{N} \sum_{i=1}^{21599} S_i^2 - \bar{S^2}$$

Dans un premier temps, on calcule $\bar{S}$:

$$\bar{S} = \frac{1}{N} * \sum_{i=1}^{21599} S_i$$

$$\bar{S} = \frac{1}{21599} * 4164734.93$$

$$\bar{S} \approx 192.82 \text{ et } \bar{S^2} \approx 37179.8336$$

On obtient donc pour variance de S :

$$Var(S) = \frac{1}{21599} * 954363573.1 - 37179.8336 = 7005.71077$$

$$Var(S) = 7005.71077$$

Le coefficient de corrélation entre le prix et la surface est donnée par la formule suivante :

$$r(P,S) = \frac{COV(P,S)}{\sigma_p * \sigma_s}$$

La covariance entre le prix et la surface est donnée par la formule suivante :

$$COV(P,S) = \frac{1}{T}\sum_{i=1}^{21599} P_i S_i - \bar{P}\bar{S}$$
$$COV(P,S) = \frac{1}{21599} * 2672744.28 - 0.53 * 192.82$$
$$COV(P,S) \approx 20.1612$$

On a donc pour le coefficient de corrélation, le résultat suivant :

$$r(P,S) = \frac{20.1612}{\sqrt{Var(P)} * \sqrt{Var(S)}}$$
$$r(P,S) = \frac{20.1612}{\sqrt{0.12} * \sqrt{7005.71}}$$
$$r(P,S) \approx 0.69$$

Question 3
Le coefficient de corrélation a un signe positif et il est assez élevé. Il semblerait donc exister un lien linéairement positif entre le prix et la surface des immobiliers dans le comté de King. Oui ce signe était prévisible au vu du nuage de points qui paraît indiquer une relation croissante entre les 2 variables.

Partie 2
Question 4
P_i est la variable endogène . u_i constitue le terme d'erreur du modèle. α est l'ordonnée à l'origine de la droite de régression tandis que β est la pente de la droite de régression. Les paramètres à estimer sont α et β.

Question 5
L'application des MCO consiste à chercher la droite de régression qui minimise l'écart entre les données et la droite de régression. Autrement dit, la méthode des MCO consite à chercher la valeur des paramètres tel que la somme des carrés des résidus soit minimale.

Partie 3
Question 6
L'équation de la droite en fonction des paramètres estimés est la suivante :

$$\widehat{P_i} = \widehat{\alpha} + \widehat{\beta} S_i$$

Question 7
Nous remplaçons $\widehat{\alpha}$ et $\widehat{\beta}$ par leurs valeurs dans (2), i.e $\widehat{\alpha} = -0.0177$ et $\widehat{\beta} = 0.0029$. On a donc l'estimation suivante :

$$\widehat{P_i} = -0.0177 + 0.0029 S_i$$

Question 8
Dans le cas présent, il n'est pas possible d'interpréter l'ordonnée à l'origine pour plusieurs raisons. Tout d'abord, l'ordonnée à l'origine correspond à une surface nulle ce qui n'a pas de sens au vu de notre régression. Par ailleurs, d'un point de vue économique, un prix nul pour un bien immobilier n'a pas de sens.

Question 9
Dans le cas présent, on peut interpréter économiquement la pente de la droite. Le $\widehat{\beta}$ indique l'effet d'une augmentation de la surface d'une unité sur le prix des biens immobiliers dans le comté de King. En utilisant le résultat de la question 7, l'effet marginal de la surface sur le prix des biens immobiliers est donné comme suit :

$\frac{\partial \widehat{P_i}}{\partial \widehat{\beta}} = 0.0029$ Ainsi, une augmentation de 1m2 de la surface d'un bien immobilier se traduit par une hausse du prix de 0.0029 millions de dollars US, soit une augmentation de 2900 dollars.

Question 10
On utilise l'équation estimée de la question 7 en fixant la surface à 200m2, on a donc : $S_i = 200$

$$\widehat{P_i} = \widehat{\alpha} + \widehat{\beta} * S_i$$

$$\widehat{P_i} = -0.0177 + 0.0029 * 200 \approx 0.56$$

Ainsi, pour une surface de 200m2 le prix estimé est égal à 0.56 millions de dollars soit 560.000 dollars US.

Exercice 2 : régression multiple

Question 1
Au vu des informations données, nous choisisons la variable binaire et l'indicateur de la qualité de la construction. Nous retenons ces 2 variables pour différentes raisons. Premièrement, ces 2 variables sont les plus corrélés avec le prix (resp. 0.249 et 0.677, contre 0.054 pour l'année de construction). Deuxièment, les boîtes à moustache peuvent également être utilisés pour choisir les variables. Les boîtes à moustache montrent clairement une différence dans la médiane des prix pour les maisons situés en bord de mer et celles qui ne le sont pas. Ainsi, au vu de la boîte à moustache , on note une différence significative du prix médian dépendante de la localisation du bien. De même pour la qualité des biens, nous constatons qu'une augmentation de la qualité des biens se traduit par une hausse du prix médian de ceux-ci, renforçant la pertinence supposée de cette variable. Par ailleurs, le choix de l'année de construction en tant que variable supplémentaire ne semble pas très pertinent au vu du nuage de points. En effet, il ne semble pas indiquer l'existence d'une relation entre l'année de construction et le prix.

Question 2
La matrice X est la matrice qui comprend les variables explicatives et la constante.

Nous avons 3 variables explicatives et 1 constante dans le modèle, soit 4 variables au total. Les dimensions de la matrice X sont donc les suivantes :
X a 21599 lignes et 4 colonnes (1 colonnes par variables explicatives et 1 colonne pour la constante).
β est de dimension (4,1) et U_i est de dimension(21599,1).

Question 3
Il s'agit de mener un test de Student. Les test doivent **impérativement** comprendre les éléments suivants :le jeu d'hypothèses, la statistique de test et la règle de décision. Le jeu d'hypothèses pour le test de significativité de $\widehat{\beta_1}$ est donc le suivant :
Jeu d'hypothèses :

$\begin{cases} Ho : \widehat{\beta_1} = 0 \text{ => la surface n'est pas significative} \\ H1 : \widehat{\beta_1} \neq 0 \text{ => la surface est significative} \end{cases}$

La statistique de test est la suivante :

$$t_{\widehat{\beta_1}} = \frac{\widehat{\beta_1} - \beta_1}{\widehat{\sigma}_{\widehat{\beta_1}}}$$

Sous Ho, on a :

$$t_{\widehat{\beta_1}} = \frac{\widehat{\beta_1}}{\widehat{\sigma}_{\widehat{\beta_1}}} \sim t(T-k-1)$$

Règle de décision :

Si $|t_{\widehat{\beta_1}|}$ > t(T-k-1) => On rejete l'hypothèse nulle.
Si $|t_{\widehat{\beta_1}|}$ < t(T-k-1) => On ne pas rejete l'hypothèse nulle.

$t_{\widehat{\beta_1}} = \dfrac{0.0017}{0.0000287} \approx 59.23$
En lisant la table de Student, on a t(21599-3-1)=1.96.
On a donc $| t_{\widehat{\beta_1}|} > 1.96$. Nous rejetons donc Ho, la surface est donc significative à 5%.

Question 4
Le coefficient de détermination fournit une mesure du pouvoir explicatif d'un modèle de régression. Selon l'équation d'analyse de la variance, nous avons :

Somme des Carrés Totaux=Somme des Carrés Expliquées + Somme des Carrés Résiduelles

SCT=SCE+SCR

Le Coefficient de détermination est donnée par la formule suivante :

$$R^2 = \frac{SCE}{SCT} = \frac{Var(\widehat{P_i})}{Var(P_i)}$$

Le R^2 représente donc la variance expliquée par le modèle. Ainsi, 56.8% de la variance du prix de l'immobilier dans le comté de King est expliquée par le modèle considéré. Le modèle est donc d'assez bonne qualité.

Question 5
Il s'agit de faire un test de Fisher afin de tester la significativité globale du modèle considérée. C'est un modèle avec 3 variables explicatives et une constante. Nous considérons donc le modèle suivant :

$$P_i = \alpha + \beta_1 S_i + \beta_2 M_i + \beta_3 Q_i + u_i$$

Le jeu d'hypothèses pour le test de Fisher est le suivant :

$\begin{cases} Ho : \alpha = \beta_1 = \beta_2 = \beta_3 = 0 \text{ => le modèle est globalement non significatif} \\ H1 : \alpha \neq 0 \text{ ou } \beta_1 \neq 0 \text{ ou } \beta_2 \neq 0 \text{ , } \beta_3 \neq 0 \text{ => le modèle est globalement significatif} \end{cases}$
La statistique de test pour le test de Fisher est la suivante :

$$\text{F= } \frac{R^2/k}{1 - R^2/T - k - 1}$$

Sous Ho, on a :
$\text{F= } \frac{R^2/k}{1 - R^2/T - k - 1} \sim F(k; T - k - 1)$
où k est le nombre de variables explicatives du modèle. La règle de décision pour le test de Fisher est la suivante :
Si F>$F^{0,05}$ (k,T-k-1) => On rejette Ho et le modèle est globalement significatif.
Si F<$F^{0,05}$ (k ;T-k-1) => On rejette Ho et le modèle n'est pas globalement significatif.

La statistique de test est donnée par :

$$\text{F= } \frac{0.568/3}{1 - 0.568/21599 - 3 - 1} \approx 9464.47$$

Nous avons donc $F = 9464.67 > F^{0.05}(3, 21599 - 3 - 1) = 2.60$. Nous rejetons donc l'hypothèse nulle et le modèle est globalement significatif.

Question 6
Comme la variable binaire est significative, cela signifie qu'il y a une différence de prix significative en moyenne entre les maisons situées en bord de mer et celles qui ne le sont pas. Lorsqu'une maison est située en bord de mer, son prix est en moyenne 0.76 millions de dollars plus élevée qu'une maison qui n'est pas située en bord de mer, soit 760 000 dollars de plus (en moyenne).

Exercice 3
1. réponse 1
2. réponse 1
3. réponse 2