Contrôle continu : économétrie L3

29 novembre 2017

Année universitaire 2017-2018.
Cours : Valérie Mignon.
TD : Florian Morvillier et Benjamin Egron.
Durée : 2 heures.
Calculatrice autorisée.

Exercice 1 : régression simple (9.5 pts)

Vous venez d'être embauché comme économètre au sein d'une société immobilière opérant sur le comté de King (état de Washington et comprenant entre autre la ville de Seattle). Tous les jours, de nouveaux biens immobiliers sont mis en vente par des agents économiques, ces derniers font appel à votre société afin de réaliser la vente. La première étape consiste à fournir un prix aux biens immobiliers qui ne soit pas déconnecté du prix du marché afin que la vente ait une chance d'avoir lieu. Bien sûr une fois ce prix initial donné, ce dernier peut-être revu à la hausse ou la baisse selon les caractéristiques spécifiques du bien immobilier.

Votre rôle en tant qu'économètre est justement de fournir ce prix initial, pour cela vous disposez de la base de données suivante :

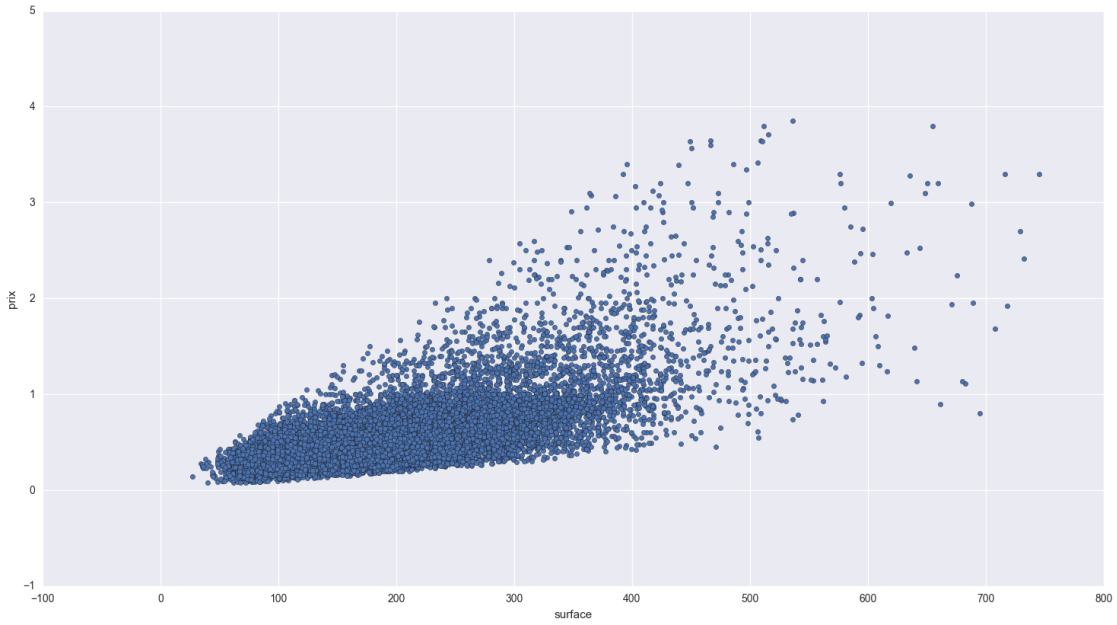
Id	prix	surface
1	0.223	109.626
2	0.588	238.760
3	0.180	71.530
...	...	...
21598	0.400	148.645
21599	0.325	94.761

La base de données est donc constituée de 21599 observations correspondant à autant de ventes de biens immobiliers ayant eu lieu dans le comté de King. Pour chaque observation i nous disposons de deux variables : le prix de vente exprimé en millions de dollars (noté P_i par la suite) et la surface du bien immobilier exprimée en m^2 (notée S_i par la suite). La colonne de gauche représente simplement l'identifiant de la vente.

De plus nous vous donnons quelques informations supplémentaires :

$$\sum_{i=1}^{21599} P_i = 11602.91 \quad \sum_{i=1}^{21599} S_i = 4164734.93$$

$$\sum_{i=1}^{21599} P_i^2 = 8837.45 \quad \sum_{i=1}^{21599} S_i^2 = 954363573.31 \quad \sum_{i=1}^{21599} P_i S_i = 2672744.28$$



Partie 1

Question 1 (0.5 pt) Quelle est la structure des données dont vous disposez : série temporelle, coupe transversale ou bien données de panel ? Justifiez votre réponse.

Question 2 (2 pt) Calculer la variance de chacune des variables ainsi que le coefficient de corrélation entre ces deux variables. Explicitez les formules que vous utilisez.

Question 3 (1 pt) Interpréter le coefficient de corrélation. Son signe était-il prévisible à partir des informations fournies ? Expliquez pourquoi.

Partie 2

Vous souhaitez maintenant prédire le prix de vente des biens immobiliers. Pour cela vous estimez, à partir de la base de données, le modèle de régression linéaire suivant :

$$P_i = \alpha + \beta S_i + u_i$$

Question 4 (1 pt) Précisez les différents termes intervenant dans cette équation, en particulier, quels sont les paramètres à estimer ?

Question 5 (1 pt) Vous décidez d'estimer les paramètres du modèle par la méthode des moindres carrés ordinaires, expliquez en quelques lignes le principe de cette méthode.

Partie 3

Une fois le modèle estimé, nous pouvons représenter graphiquement les valeurs prédites par le modèle sous la forme d'une droite reportée sur le graphique suivant :

Question 6 (0.5 pt) Ecrivez l'équation de la droite en fonction des paramètres estimés.

Question 7 (0.5 pt) La valeur de la pente de la droite est de 0.0029 et la valeur de son ordonnée à l'origine est de -0.0177. Reprenez l'équation de la droite (question 6) et incorporez ces deux informations.

Question 8 (1 pt) Dans le cas présent, peut-on interpréter économiquement l'ordonnée à l'origine de la droite ? Si non pourquoi ? Si oui, interprétez.

Question 9 (1 pt) Dans le cas présent, peut-on interpréter économiquement la pente de la droite ? Si non pourquoi ? Si oui, interprétez.

Question 10 (1 pt) On vous demande d'estimer le prix d'une maison de 200 mètres carrés, quel prix lui accordez-vous ?

Exercice 2 : régression multiple (9 pts)

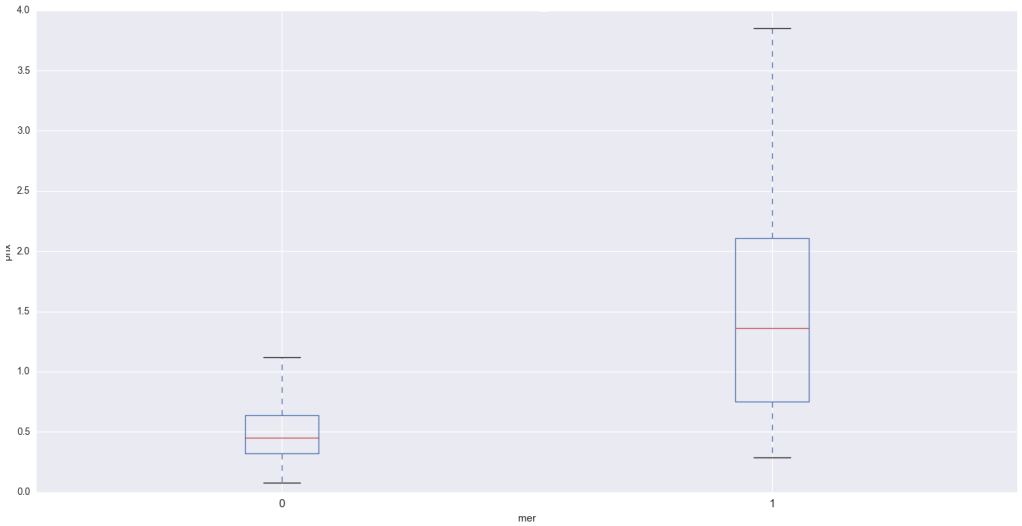
Vous disposez à présent d'une nouvelle base de données. Comme précédemment, celle-ci contient pour chaque observation i le prix de vente exprimé en millions de dollars (P_i) et la surface exprimée en m^2 (S_i). A cela s'ajoutent :

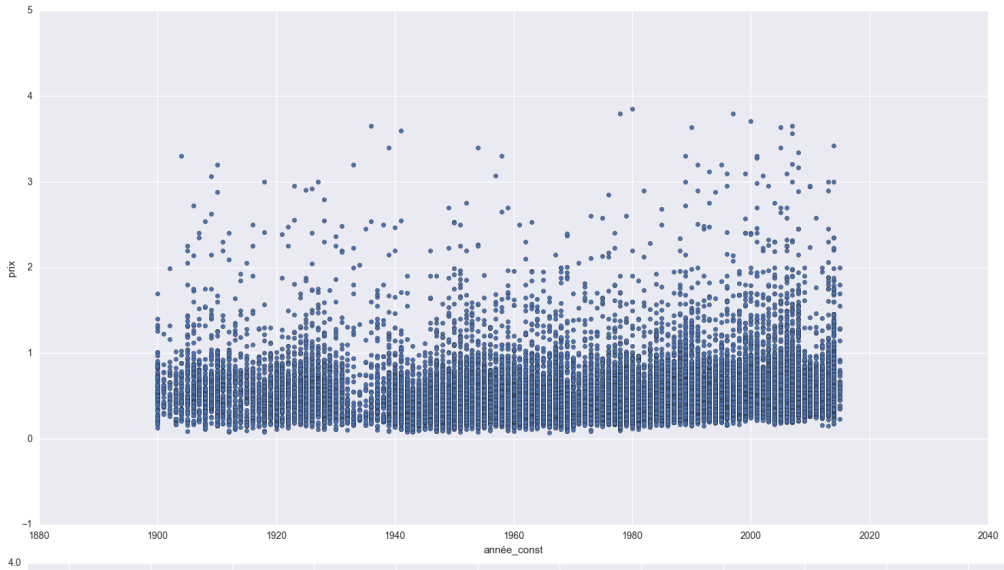
1. *Une variable binaire prenant la valeur 1 si la maison se trouve en bord de mer et la valeur 0 sinon. Cette variable est notée M_i par la suite.*
2. *L'année de construction du bien immobilier allant de 1900 à 2015. Cette variable est notée C_i par la suite.*
3. *La qualité de la construction, prenant des valeurs entières de 1 (construction de très mauvaise qualité) jusqu'à 13 (construction de très bonne qualité). Cette variable est notée Q_i par la suite.*

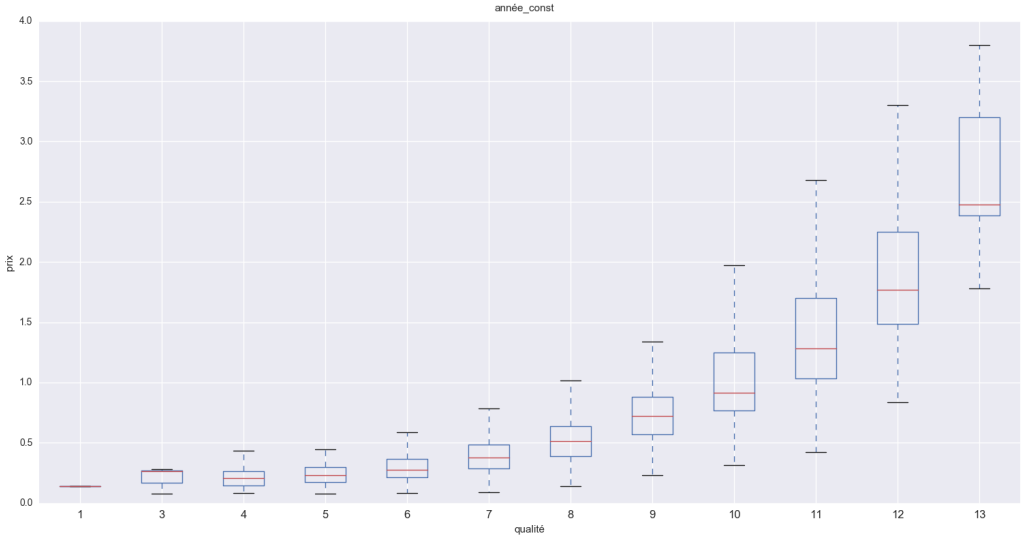
La base de données prend donc à présent la forme suivante :

Id	prix	surface	mer	annee const	qualité
1	0.223	109.626	0	1955	7
2	0.588	238.760	0	1951	6
3	0.180	71.530	1	1987	8
...	...	...	...	...	...
21598	0.400	148.645	1	2004	8
21599	0.325	94.761	0	2008	7

Nous vous fournissons plusieurs coefficients de corrélation : $Corr(P, M) = 0.249$, $Corr(P, C) = 0.054$ et $Corr(P, Q) = 0.677$, ainsi que quelques graphiques :







Question 1 (1 pt) On vous demande d'intégrer deux variables (parmi les trois proposées) supplémentaires au modèle de régression linéaire précédemment estimé, lesquelles choisissez-vous au vu des informations que nous vous donnons [1] ? Justifiez votre réponse en quelques lignes.

Nous avons donc maintenant un modèle de régression multiple : une variable expliquée qui est le prix, trois variables explicatives (la surface plus les deux variables que vous avez choisies) et une constante. Ce modèle s'écrit de la manière suivante :

$$P = X\beta + u$$

1. En fin d'énoncé un bref rappel est fait sur les boîtes à moustache

où $P = (P_1, P_2, ..., P_{21599})'$ est le vecteur colonne des prix de dimension 21599, X est la matrice incluant les variables explicatives et β est le vecteur des paramètres.

Question 2 (1 pt) Quelles sont les dimensions de la matrice X, du vecteur β et du vecteur u (n'oubliez pas que le modèle inclut une constante) ?

Question 3 (2 pts) Le paramètre associé à la variable surface est noté β_1. Après estimation du modèle nous obtenons $\hat{\beta}_1 = 0.0017$ et $\hat{\sigma}_{\hat{\beta}_1} = 0.0000287$. La valeur du paramètre β_1 est-elle significativement différente de zéro (vous prendrez un risque de 5%) ? N'oubliez pas de présenter le test statistique que vous utilisez.

Question 4 (1 pt) Expliquez ce qu'est le coefficient de détermination (R^2) et son interprétation en utilisant l'équation d'analyse de la variance. Dans le cas présent, le coefficient de détermination s'élève à 0.568, interprétez ce résultat.

Question 5 (2 pts) Le modèle de régression linéaire est-il globalement significatif (vous prendrez un risque de 5%) ? N'oubliez pas de présenter le test statistique que vous utilisez.

Question 6 (2 pts) Supposons que nous ayons incorporé la variable M_i au modèle, la valeur du paramètre associé est de 0.76 et est significativement différente de zéro. Comment interprétez-vous économiquement ce résultat ?

Exercice 3 : questions de cours (3 pts)

Une mauvaise réponse fait perdre 0.5 points au sein de cet exercice, vous ne perdez donc pas de points sur les exercices précédents.

Question 1 (1 pt) Que se passe-t-il si l'hypothèse de nullité de l'espérance du terme d'erreur n'est pas vérifiée ?

1. L'estimateur MCO est biaisé.
2. Il n'est pas possible d'obtenir l'estimateur MCO.
3. Aucune des deux réponses précédentes.

Question 2 (1 pt) On suppose que l'hypothèse d'homoscédasticité du terme d'erreur n'est pas vérifiée. Quel est l'effet de la violation de cette hypothèse ?

1. Les t-stat (statistiques de Student) ne sont plus fiables.
2. Les termes en dehors de la diagonale de la matrice de variance-covariance du terme d'erreur sont non nuls.
3. L'estimateur MCO est biaisé.

Question 3 (1 pt) On suppose que l'hypothèse d'absence d'autocorrélation du terme d'erreur n'est pas vérifiée. Quel est l'effet de la violation de cette hypothèse ?

1. La taille de l'intervalle de prévision augmente.
2. Les termes en dehors de la diagonale de la matrice de variance-covariance du terme d'erreur sont non nuls.
3. La matrice des variables explicatives n'est plus certaine.

Rappel : boîte à moustaches

Les boîtes à moustaches permettent de résumer la distribution d'une variable en quelques valeurs : le segment indique la médiane, la boîte indique le premier et le quatrième quartiles enfin les extrémités indiquent le premier et le neuvième déciles.