Synthèse - Projet économétrie

Baptiste Viola - Lucas Monteiro

Introduction

On a à disposition une base de données portant sur des biens immobiliers dans la région de Seattle, dans l'état de Washington aux États-Unis. Cette base comporte 21 163 observations et décrit 20 caractéristiques pour chaque bien. L'objectif de notre étude est de déterminer le modèle le plus intéressant pour fournir une bonne prédiction du prix.

Démarche

Dans un premier temps, nous avons suivi notre intuition et nous avons supprimé les variables qui ne semblaient pas être les plus pertinentes concernant le sujet de notre analyse, notamment le numéro d'identification, la date de vente et le code postal. De plus, nous avons transformé les variables "prix" ainsi que les variables liées aux surfaces du bien car elles étaient très élevées et cela nous permettait d'exprimer les élasticités afin de comprendre d'où pouvaient venir les variations du prix.

Dans un second temps, nous avons réalisé différentes statistiques descriptives afin de pouvoir faire apparaître les premiers liens entre le prix et les différentes variables étudiées. On a également recodé quelques variables pour pouvoir mener une analyse plus simple et plus intéressante. Par exemple, nous avons étudié le prix selon différentes latitudes et nous avons fini par créer une nouvelle variable catégorielle nommée "Nord" qui vaut 0 si le bien est au Sud et qui vaut 1 si le bien est au Nord.

Dans un troisième temps, nous sommes passés aux tests de régression linéaire dans le but de dégager des variables qui apportent une information intéressante quant aux prix pour pouvoir élaborer des modèles simples et pertinents.

Modèles retenus

Suite à nos différents tests, nous avons pu, d'un côté, éliminer des variables qui expliquaient peu les variations du prix, et d'un autre côté, nous avons pu retenir celles nous permettant de mieux prédire le prix.

Nous avons d'abord élaboré un premier modèle simple, ne contenant que trois variables explicatives (*lnliv15*, *nord* et *waterfront*) car elles étaient celles qui nous apportaient le plus d'information concernant les variations du prix. Ensuite nous avons affiné un peu plus ce premier modèle en lui ajoutant les variables *bathrooms* et *renov*. Enfin nous avons établi un modèle final, reprenant les variables des deux premiers modèles ainsi qu'en ajoutant *young*, *basement* et *condition*. On a procédé aux calculs des VIF et on en a conclu qu'il n'y avait pas de forte colinéarité entre les variables explicatives.
Le modèle final que l'on cherche à estimer est donc :

$$lnprice_i = \beta_0 + \beta_1 lnliv15_i + \beta_2 nord_i + \beta_3 bathrooms_i + \beta_4 waterfront_i + \beta_5 renov_i + \beta_6 condition_i + \beta_7 basement_i + \beta_8 young_i + \epsilon_i$$

On obtient donc les estimations suivantes (avec les écart-types de chacune) avec la méthode des moindres carrés ordinaires :

$$\beta_0 = 7.088\ [0.059] \qquad \beta_1 = 0.662\ [0.008] \qquad \beta_2 = 0.421\ [0.004] \qquad \beta_3 = 0.184\ [0.004] \qquad \beta_4 = 0.740\ [0.025]$$

$$\beta_5 = 0.218\ [0.011] \qquad \beta_6 = 0.091\ [0.004] \qquad \beta_7 = 0.086\ [0.005] \qquad \beta_8 = 0.044\ [0.006]$$

Diagnostic

Tous les coefficients estimés sont significatifs au seuil de 1%, de plus ils sont tous positifs et par conséquent, toutes les variables ont tendance à faire augmenter les prix. On va expliquer les coefficients estimés ci-dessus :

- β_1 : Ce coefficient signifie que, *ceteris paribus*, une augmentation de 1% de la surface habitable entraîne en moyenne une augmentation du prix de 0.662% , selon notre modèle.
- β_2 : Ce coefficient signifie que, *ceteris paribus*, un bien situé au Nord est en moyenne 42.1% plus cher qu'un bien situé au Sud, selon notre modèle et notre découpage Nord/Sud.

— β_3 : Ce coefficient signifie que, *ceteris paribus*, une salle de bain supplémentaire entraîne en moyenne une augmentation du prix de 18.4%, selon notre modèle.

— β_4 : Ce coefficient signifie que, *ceteris paribus*, un bien situé au bord d'un point d'eau est en moyenne 74.0% plus cher qu'un bien qui n'a pas de vue sur une point d'eau, selon notre modèle.

— β_5 : Ce coefficient signifie que, *ceteris paribus*, un bien rénové est en moyenne 21.8% plus cher qu'un bien non rénové, selon notre modèle.

— β_6 : Ce coefficient signifie que, *ceteris paribus*, un niveau de condition supplémentaire entraîne en moyenne une augmentation du prix de 9.1%, selon notre modèle.

— β_7 : Ce coefficient signifie que, *ceteris paribus*, un bien qui possède un sous-sol est en moyenne plus cher de 8.6% qu'un bien qui n'en possède pas, selon notre modèle.

— β_8 : Ce coefficient signifie que, *ceteris paribus*, un bien construit avant 1995 est en moyenne moins cher de 4.4% qu'un bien construit après cette date, selon notre modèle.

Le R^2 et le R^2 ajusté sont égaux et valent 0.637, c'est-à-dire que 63.7% de la variance du logarithme du prix est expliquée par notre modèle. Lorsque le prix augmente de 1%, le modèle est capable d'en expliquer 0.637%.

La statistique de Fisher est significative au seuil de 1%, signifiant que le modèle est globalement significatif.

On a ensuite analysé les résidus de notre modèle.

Premièrement, on a testé si l'hypothèse de linéarité sous la méthode des moindres carrés ordinaires était respectée et suite au test de Rainbow, on en a conclu que cette hypothèse était valide. Deuxièmement, nous avons examiné la normalité des résidus, grâce au test de Jarque-Bera. Suite à ce test, nous avons conclu que les résidus ne suivaient pas une distribution normale. Nous avons ensuite analysé si le modèle présentait de l'hétéroscédasticité ou non. Nous avons conclu, grâce au test de Breush-Pagan, que notre modèle présentait effectivement de l'hétéroscédasticité. Finalement, nous avons testé s'il y avait des résidus trop importants dans notre modèle grâce aux distances de Cook et nous en avons conclu qu'il n'y en avait pas.

Nous avons réalisé notre modèle suivant la méthode des moindres carrés ordinaires, cependant cette méthode requiert que l'hypothèse d'homoscédasticité des résidus soit respectée, or ce n'est pas le cas. Ceci entraîne une mauvaise estimation de la matrice de variance-covariance et par conséquent des écart-types des coefficients de notre modèle. Pour pallier le biais de cette matrice, nous utilisons la méthode des écart-types robustes et l'estimateur de White.

Il en résulte que nos variables restent significatives et que les écart-types n'ont que sensiblement changé.

Conclusion

Nous avons sélectionné les variables les plus pertinentes compte tenu l'objet de notre analyse. Ces variables reprennent différentes caractéristiques qui pourraient s'avérer être utiles à une agence immobilière, notamment pour chercher à comprendre les variations de prix ou encore les prix en eux-mêmes.

On a cherché à garder une variable pour chaque catégorie. Par exemple, on disposait de plusieurs variables portant sur la surface du bien et nous avons décider de ne garder que sur la surface habitable actualisée. Même chose pour la localisation avec "Nord", pour la vue avec la variable "waterfront", pour les pièces de la maison avec "bathrooms".

Limites

Cependant notre analyse se concentre sur la région de Seattle entre 2014 et 2015, ce qui empêche toute généralisation à d'autres régions car les coefficients estimés sont basés sur l'étude des biens à Seattle.

Par ailleurs, on ignore la signification de la variable "grade", plusieurs recherches nous suggèrent qu'il s'agirait d'un critère de qualité et d'esthétique du bien mais on ignore la réelle signification. Concernant les autres variables, bien qu'on ait un léger descriptif, on ne sait pas vraiment sur quoi sont basées certaines variables, par exemple "view", "sqft_lot15 ou "condition", dont ces dernières manquent de précision.

De plus, les valeurs que prennent certaines variables dépendent du découpage que l'on a effectué. En effet un découpage différent modifierait les valeurs et donc pourrait modifier les résultats de notre modèle, c'est le cas de "Nord" et "young".

Il faut également prendre en compte le fait qu'on ne dispose pas de valeurs faibles dans nos échelles logarithmiques pour le prix et la surface habitable et que par conséquent, on ne peut estimer la relation entre le prix et la variables pour ces valeurs faibles. Notre modèle est donc limité par l'intervalle de valeurs étudiées et essayer de prédire un prix avec une surface qui serait en dehors de cet intervalle est une démarche peu rigoureuse. De plus, si on pose l'hypothèse qu'une surface nulle entraîne un prix nul ($\beta_0 = 0$), il se pose alors la question de la relation entre l'origine et notre nuage de points.