SAE Régression Linéaire

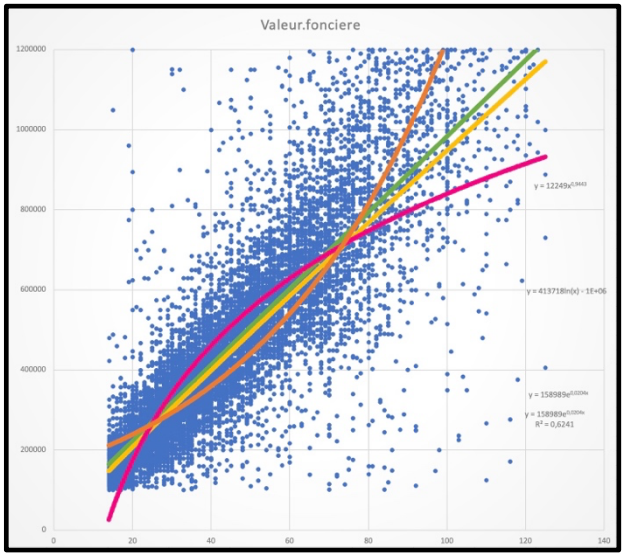
1) Introduction

La capitale française regorge de logements dont les caractéristiques comme la surface peuvent faire varier le prix. Est-il possible de prédire la valeurs des logements en fonction de leur spécificités ?

L'objectif de notre travail est d'estimer la valeur de biens immobiliers de la ville de Paris au premier semestre 2023. Pour cela, nous disposons de deux fichiers CSV nommé Train.csv et Test.csv. Le fichier Train.csv contient les valeurs foncières ainsi que d'autres variables comme la surface habitable ou encore l'arrondissement. Le fichier Test.csv ne contient pas les valeurs foncières et nous devons les prédire à partir d'autres variables à l'aide d'un modèle statistique de régression. Nous devons réaliser cette étude statistique à l'aide du langage de programmation statistique R et ensuite importer nos résultats dans un fichier CSV.

2) Démarche pour élaborer notre modèle

Premièrement, nous avons analysé le fichier Train à l'aide d'Excel pour prendre connaissance du jeu de données. Nous avons trié le fichier en supprimant les valeurs aberrantes qui pouvaient corrompre l'élaboration de notre modèle, afin de le rendre le plus proche possible des valeurs du fichier Test. Nous avons commencé par supprimer tous les logements possédant un nombre de pièces égal à 0. Nous avons conservé les logements dont le prix se situe entre 100 000 et 2 000 000€. Ensuite, nous avons gardé les logements dont la surface est comprise entre 14 et 125 m^2. Enfin, nous avons supprimé les logements où la surface du terrain est supérieure à 600. Cette démarche nous a permis de passer initialement de 12 797 logements à 10 949.



Dans un second temps, nous avons déterminé quel était le modèle le plus pertinent pour effectuer l'estimation des prix. Nous avons dès le début choisi de baser notre modèle sur la surface des logements car d'un point de vue statistique, c'est la variable quantitative discrète qui peut prendre le plus grand nombre de différentes valeurs dans notre jeu de données. De plus, d'un point de vue logique, la surface est le premier indice de prix dans le monde de l'immobilier. Nous avons appliqué les formules fournies par le graphique ci-dessus à la surface habitable et nous avons déduis que la méthode la plus efficace était d'utiliser le modèle puissance et de créer par la suite un coefficient en fonction de l'arrondissement. Ce coefficient se calcule à partir des moyennes des valeurs foncières du fichier Train.csv, en prenant le 17^{e} arrondissement pour base du coefficient, car c'est l'arrondissement médian en valeur moyenne.

3) Exposition du modèle retenu

Nous avons donc commencé les calculs sur R. Tout d'abord, nous avons repris le fichier Train.csv pour le trier selon les critères que l'on a déjà définit sur Excel.

Ensuite, nous avons appliqué la fonction logarithme (log sur R) sur les colonnes des Surface.reelle.bati et Valeur.fonciere.

Grâce à cela, nous avons pu trouver le coefficient de corrélation (fonction cor sur R). Ensuite nous avons calculé la covariance ainsi que la variance.

Coef de corrélation = 0,874
Covariance = 0,286
Variance = 0,280
Suite à cela, nous avons calculé la pente et l'ordonnée à l'origine de la droite de régression linéaire.

Pente = 1,20
Ordonné à l'origine = 9,159

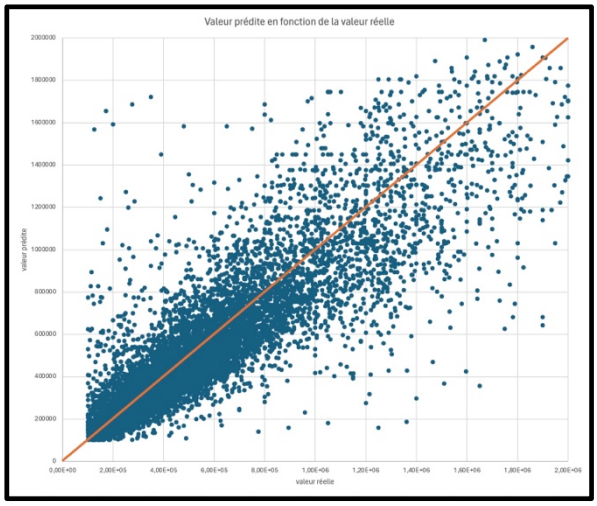
Grâce à la pente et à l'ordonnée, nous avons pu calculer la prédiction des valeurs au logarithme avec la fonction du modèle puissance, qui est y = ax + b .

Ce qui donne : Pente * ln(surface du logement) + ordonnée

Ensuite, nous avons intégré le coefficient lié à l'arrondissement dans lequel se trouve le logement. Premièrement, nous avons déterminé la moyenne des valeurs des logements pour chaque arrondissement, en créant une table contenant l'arrondissement et sa moyenne correspondante. Suite à cela, nous avons joint les deux tables en rattachant la bonne moyenne à chaque logement par rapport à l'arrondissement, puis nous avons rajouté une autre colonne contenant la moyenne du 17e arrondissement pour chaque logement afin de faire une troisième colonne où nous créons le coefficient en divisant la moyenne des valeurs de l'arrondissement du logement par la moyenne du 17e.

Nous pouvons ainsi déterminer le prix final estimé pour chaque logement en mettant à l'exponentielle la prédiction des valeurs au logarithme, puis en multipliant celle-ci par le coefficient de l'arrondissement.

Arrondissement	Moyenne	Coefficient
PARIS 01	627617,0056	1,168441585
PARIS 02	590617,3633	1,099558938
PARIS 03	659801,8874	1,228360538
PARIS 04	674676,676	1,256053098
PARIS 05	648334,0224	1,207010686
PARIS 06	846328,8242	1,575619819
PARIS 07	839881,1575	1,563616126
PARIS 08	734945,65	1,36825652
PARIS 09	612833,132	1,140918282
PARIS 10	477600,5084	0,889154197
PARIS 11	465943,8755	0,867452914
PARIS 12	453846,1219	0,844930391
PARIS 13	427322,5083	0,795551083
PARIS 14	469891,0092	0,874801337
PARIS 15	511966,5342	0,953133811
PARIS 16	748543,3353	1,393571483
PARIS 17	537140,2506	1
PARIS 18	393923,565	0,733371898
PARIS 19	401911,4948	0,748243116
PARIS 20	391306,014	0,728498774



Notre modèle permet d'avoir un coefficient de corrélation d'environ 0,83. Nous pouvons donc importer le fichier Test et appliquer notre modèle en fonction des valeurs de chaque bien. On utilise l'équation y = Pente * ln(surface du logement) + ordonnée, puis on multiplie les valeurs obtenue par le coefficient correspondant à l'arrondissement. Nous obtenons par la suite le fichier final nommé prediction.csv qui contient nos prédictions pour les valeurs foncières pour les logements du fichier test.csv .

4) Conclusion

Pour conclure, nous avons ainsi réalisé un modèle pour prédire les valeurs de biens immobiliers à Paris en fonction de leurs surfaces et de leurs arrondissements, nous sommes conscients du fait qu'il possède certaines limites et que nous aurions pu prendre en compte d'autres variables comme le nombre de pièces ou la surface du terrain habitable.

Cependant le fichier Test.csv ne contient que 3 maisons ce qui ne change pas réellement la globalité des valeurs. Pour le nombre de pièces, celui-ci est fortement corrélé à la surface, nous n'avons donc pas pris en compte cette variable. Nous sommes alors satisfaits de notre travail puisque notre modèle semble fidèle à la réalité. Le coefficient de corrélation entre la valeur foncière estimée et la valeur réelle égal à 0,83 nous confirme que notre modèle s'approche de la réalité, sans pour autant entrer dans l'over-fiting.

Le temps de travail en cours nous a été suffisant et le travail chez soi a été réduit. Ce projet nous a également permis de mettre en application le cours et de comprendre l'enjeu et l'intérêt de celui-ci.