Modélisation statistique - TD3
Régression linéaire simple avec R

1 Exercice : Prix de l'immobilier

Dans le fichier `Apartment_prices.csv` vous trouverez le prix de vente relevé pour 50 appartements vendus dans la même année, dans une même ville, ainsi que le nombre de mètres carrés de chaque appartement.

1. Importer dans **R**, le fichier de données et créez un *data.frame* que vous nommerez `house`.
2. Dans un premier temps, vous observerez le lien entre les deux variables `price` et `Squaremeter`.
   (a) A l'aide d'un nuage de points (fonction `plot`)
   (b) En calculant le coefficient de corrélation linéaire (fonction `cor`)
   (c) Que concluez-vous ?
   (d) Rappelez quel est le lien entre la pente de la droite de régression définie par $E(Y|X) = \alpha + X\beta$ et le coefficient de correlation entre les variables X et Y.
   (e) Vous paraît-il judicieux de modéliser le prix de l'appartement par sa surface ?
3. Ecrire le modèle de régression linéaire simple qui relie le prix d'un appartement à sa surface.
4. Le modèle est écrit sur les 50 individus de l'échantillon, mais quelle est la population ?
5. Estimer le modèle de régression avec l'instruction suivante

   ```
   res=lm(Price~Squaremeter,data=house)
   ```

   La sortie de la fonction `lm` est un objet de la classe `lm`, pour lequel un certains nombres de fonction ont été implémentées. On utilisera par exemple `summary`, `coefficients`,`anova`, `plot`, `predict`, `residuals`. Pour obtenir de l'aide sur ces fonctions, on écrira par exemple `?plot.lm`. Cela permet de la distinguer de la fonction généraliste `plot`. Vous pouvez observer la structure de l'objet `lm` avec l'instruction `str(lm)`.
6. Nous souhaitons tester l'utilité d'un tel modèle de régression simple par rapport à un modèle plus simple, où toutes les variables auraient la même moyenne.
   (a) Quel test devons nous réaliser ?
   (b) Ce test s'appuie sur des sommes carrés, donner la définition des trois sommes de carrés, leur degré de liberté associé et leur relation.
   (c) Quelle est la stastistique de test et sa loi sous $\mathcal{H}_0$?
   (d) Avec l'instruction `anova(res)`. Donner la valeur des trois sommes de carré, des degrés de liberté, de la statistique F, et sa p-valeur.
   (e) Quelle est la conclusion du test ?

7. Nous souhaitons maintenant estimer la variance des résidus.
   (a) L'avez-vous déjà repérée ?
   (b) Donner son estimateur et son estimation.
   (c) Vous pouvez également la trouver avec la fonction `summary(res)`, sous le terme *Residual standard error*. Il s'agit ici de l'estimation de l'écart-type, vérifier que la valeur donnée ici est bien égale à la racine carrée de l'estimation de la variance donnée dans la table d'analyse de la variance.
8. Nous souhaitons maintenant estimer les coefficients.
   (a) Quelle est la méthode retenue pour l'estimation des deux coefficients α et β ?
   (b) Donner leur estimation (formule).
   (c) A l'aide de la fonction `summary` donner leur estimation (valeur), n'oubliez pas les unités.
   (d) Comment interprétez vous chaque coefficient ?
   (e) Rappelez la variance des estimateurs A et B.
   (f) A la suite de chaque estimation, une statistique de test et une p-valeur est donnée. Dire de quel test il s'agit (hypothèses nulle et alternative, statistique de test et sa loi sous $\mathcal{H}_0$)
   (g) Avec la fonction `lines` ajouter la droite de régression au nuage de points. Celle-ci passe t-elle par le centre de gravité du nuage ?
9. Nous souhaitons prédire le prix d'un nouvel appartement dont la surface est de 80 m^2.
   (a) Cela est-il réalisable avec notre modèle et nos données ?
   (b) Donner une estimation de $E(Y|X = 80)$, ainsi que son intervalle de confiance. Vous pourrez utiliser la fonction `predict`.
   (c) Donner l'intervalle de prévidiction pour la prédiction
   $$Y_{80}^P = A + B80 + E^P \text{ avec } E^P \sim \mathcal{N}(0, S^2)$$
   (d) Calculer l'intervalle de confiance et de prédiction pour toutes les surfaces allant de 0 à 150 m^2 avec un pas de 1 m^2. Représentez les sur une figure, que constatez-vous ?
10. Nous allons maintenant vérifier les hypothèses du modèle.
   (a) A l'aide de la fonction `plot(res)` afficher le graphique des résidus vs valeurs prédites. Les hypothèses de linéarité et homoscédasticité vous paraissent-elles vérifiée ?
   (b) A l'aide du graphique "quantile-quantile" vérifiez l'hypothèse de normalité.
   (c) Y a t-il des résidus qui vous paraissent aberrants ?
   (d) Si vous avez répondu "oui" à la question précédante, supprimer ces résidus du tableau de données et reprenez l'estimation du modèle. Observez les changements.