UE Statistiques
TP 1 : Modèle linéaire

Enseignant : François Bachoc

Language: R.

Ceux qui n'ont jamais utilisé R peuvent regarder des manuels d'introduction en ligne, par exemple : `https://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf`

1 Jeu de données : qualité de l'air

Charger le jeu de données airquality avec les commandes

- `library("datasets")`
- `data(airquality)`
- `table=airquality`

(A l'issue de ces commandes, `table` est une variable de type `data.frame` qui contient toutes les quantités utiles à ce TP)

Le jeu de données correspond à une matrice de taille $N \times P$. La ligne i de cette matrice est un vecteur fournissant les valeurs de certaines variables pour le jour de mesure numéro i. Ainsi, chaque colonne de la matrice correspond à toutes les mesures d'une variable. L'une des variables est la quantité d'ozone dans l'aire, et on souhaite prédire cette variable en fonction des autres variables, qu'on appelle variables explicatives. Une description du jeu de données est disponible sur `https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/airquality.html`

1) Donner les valeurs de N et P. Lister les variables explicatives.

2) A l'aide de la commande `plot(table)`, décrivez visuellement et qualitativement l'influence de chacune des variables explicatives sur la quantité d'ozone dans l'air.

2 Gérer les données manquantes

Certaines valeurs dans le jeu de données sont manquantes (`NA` pour not available). Pour chacune des variables explicatives radiation solaire, vent et température, on veut savoir si elles ont, en moyennes, des valeurs différentes selon que la quantité d'ozone est observée.

1) Chacune des trois variables explicatives ci-dessus correspond à un vecteur $(v_i)_{i=1,\dots,N}$. La quantité d'ozone correspond à un vecteur $(O_i)_{i=1,\dots,N}$. Pour le vecteur extrait $(v_i)_{i=1,\dots,N;O_i=NA}$ (les valeurs des variables explicatives pour les jours ou on n'a pas pu observer la quantité d'ozone), quelle est la proportion de valeurs non-observées? Quelle est la moyenne des valeurs observées?

2) Répéter la question 1), en considérant cette fois le vecteur extrait $(v_i)_{i=1,\dots,N;O_i\neq NA}$ (les valeurs des variables explicatives pour les jours ou on n'a pu observer la quantité d'ozone).

3) Au vu des résultats des questions 2) et 3), diriez-vous que les trois variables explicatives prennent, en moyenne, des valeurs significativement différentes selon que la quantité d'ozone est observée ou non?

3 Supprimer les lignes avec des données manquantes

1) A partir de la variable `table` contenant toutes les données, supprimer les colonnes correspondant aux variables autres que la quantité d'ozone, la radiation solaire, le vent et la température. Puis, supprimer toutes les lignes pour lesquels au moins l'une des quatre variables ci-dessus n'es pas observée. Vous obtenez ainsi une nouveau jeu de donnée, correspondant à une matrice de taille $n \times 4$.

2) Donner la valeur de n.

4 Mettre en œuvre le modèle linéaire "à la main"

Dans cette partie, il est demandé pour répondre aux questions de n'utiliser aucune fonction toute faite de R concernant le modèle linéaire. Il est demandé de programmer chaque élément d'analyse du modèle linéaire (estimation, intervalles de confiance, tests...), en utilisant les outils de calcul matriciel de R.

1) Créer la matrice X de taille $n \times 4$, dont la première colonne est composée uniquement de 1 et dont les 3 autres colonnes sont les vecteurs des valeurs prises par les 3 variables explicatives radiation solaire, vent et température. Créer un vecteur Y de taille $n \times 1$, contenant les valeurs prises par la variable d'intérêt quantité d'ozone. Assurez vous que la ligne i de X et l'élément i de Y contiennent des valeurs qui étaient situées à la même ligne dans le jeu de données `table` initial. On suppose alors que X et Y proviennent d'un modèle linéaire $Y = X\beta^* + \epsilon$.

2) Donner la valeur de $\hat{\beta}$. L'interprétation des signes des composants de $\hat{\beta}$ est-elle en accord avec l'interprétation visuelle de la partie 1?

3) Donner la valeur de $\hat{\sigma}$.

4) Calculer le vecteur $\hat{Y} = X\hat{\beta}$. Calculer la matrice de C de taille $n \times n$ telle que $\sigma^2 C = cov(Y - \hat{Y})$. Montrer que le vecteur $((Y - \hat{Y})_i/(\sigma\sqrt{C_{i,i}}))_{i=1,...,n}$ est composé de n variables gaussiennes standards. Tracer les valeurs du vecteur $((Y-\hat{Y})_i/(\hat{\sigma}\sqrt{C_{i,i}}))_{i=1,...,n}$. Discuter par quels aspects ce tracé ressemble à n réalisations de variables gaussiennes standards, et par quels aspects ce n'est pas le cas.

5) Proposer des estimateurs des écarts-types des 4 variables aléatoires $\hat{\beta}_i$, $i = 1, ..., 4$. Donner les valeurs de ces estimateurs.

6) (Plus difficile) A l'aide de la commande `plot` appliquée à un objet de type `data.frame`, tracer les valeurs de $\hat{Y}$ en fonction des trois variables explicatives. Ces tracés devraient être visuellement "plus linéaires" que les tracés de la partie 1. Expliquer pourquoi.

7) Pour chacune des trois variables explicatives radiation solaire, vent et température, on teste si elles ont une influence sur Y. Donner les valeurs des trois statistiques de test, pour les trois tests de Student des hypothèses $\beta_2 = 0$, $\beta_3 = 0$ et $\beta_4 = 0$. On note t_2, t_3 et t_4 ces trois valeurs.

8) Au niveau de confiance 95%, rejetez-vous les trois hypothèses de la question 7)?

9) Calculer, pour i allant de 2 à 4 les probabilités $P(|S| \geq t_i)$, ou S suit une loi de Student à $n - p$ degrés de liberté. Expliquer pourquoi ces trois probabilités correspondant à des P-values, similaires à celles vues pour le test de la somme des rangs de Wilcoxon.

10) Donner la valeur de la statistique de test pour le test de Fisher de l'hypothèse $\beta_2 = \beta_3 = \beta_4 = 0$. On note f cette valeur.

11) Calculer la probabilité $P(F \geq f)$, ou F suit une loi de Fisher à $3, n - p$ degrés de liberté. Expliquer pourquoi cela correspond aussi à la notion de P-value.

5 Utilisation de la fonction modèle linéaire de R

En notant par exemple `table_bis` l'objet de type `data.frame` obtenu à l'issu de la partie 3, utiliser la fonction modèle linéaire de R avec la commande `summary(lm(data=table_bis,Ozone~.))`. [Cela correspond à un modèle linéaire dans lequel la variable d'intérêt est la quantité d'ozone, et les variables explicatives sont la variable constante égale à 1 (inclusion d'un intercept dans le modèle) et toutes les variables du jeu de données `table_bis` autre que la quantité d'ozone.]

1) Faire le lien entre chacune des informations fournies par la commande précédente `summary(lm(data=table_bis,Ozone~.))` (à l'exception de `MultipleR-squared` et `AdjustedR-squared`) et les questions de la partie 4.