Fiche 3 - Régression linéaire

Tran Viet Chi, chi.tran@math.univ-lille1.fr, bureau 316 (bâtiment M3).

Les données sont disponibles sur la page http://math.univ-lille1.fr/~tran/enseignements.html

Exercice 1 (Ozone de l'air)
La table `ozone.dta` contient les variables suivantes, pour une série de journées (qui sont ici nos individus) :
- l'identifiant de la journée,
- le maximum d'ozone (variable `maxO3`)
- l'heure à laquelle le maximum d'ozone a été obtenu (`heure`),
- les températures à 6h, 9h, 12h, 15h, 18h (resp. `T6` à `T18`)
- la nébulosité à 6h, 9h, 12h, 15h, 18h (resp. `Ne6` à `Ne18`)
- la projection du vent sur l'axe est-ouest à 12h (`Vx`),
- le maximum d'ozone de la veille (`maxO3v`).

Le but est de modéliser la valeur des pics d'ozone en fonction de grandeurs physiques facilement mesurables (température, heure, nébulosité, vent) afin d'avoir des approximations de la qualité de l'air faciles et rapides à obtenir.

Rque : Ce jeu de données ne correspond pas à la même période que celui utilisé au TD2.

Partie A Explication du pic d'ozone par la température à midi

Dans cette première partie, nous souhaitons étudier les liens entre la valeur du pic d'ozone `maxO3` et la température à midi `T12`.

1. Importer les données avec la commande :
`donnees<-read.table(chemin,header = TRUE)`
où *chemin* est le chemin d'accès du fichier `ozone.dta`, par exemple `H:/TISD/ozone.dta`.

2. Analyser les variables `maxO3` et `T12` indépendamment (moyenne, écart-type, boxplot, histogramme avec densité superposée ...). Reconnaît-on l'allure de lois usuelles ?

3. Dessiner `maxO3` en fonction de "T12" avec la commande `plot`. Qu'en pensez-vous ?

4. Effectuer la régression de `maxO3` en fonction de `T12` avec la commande
`resmco<-lm(donnees$maxO3 ~ donnees$T12)`.

5. Extraire les coefficients de la régression à l'aide de la commande `coef`. Vérifier que l'on retrouve les mêmes valeurs avec les formules du cours.

6. Extraire de `resmco` la droite de régression en utilisant la commande `fitted` et la superposer au nuage de points obtenu à la question **3**. Recommencer en utilisant la commande `abline`, et recommencer en utilisant la série `T12` et les coefficients de la régression.

7. Demander une analyse de la régression avec la commande `summary.lm(resmco)`. Commenter.

8. Faire une analyse de la variance avec la commande `anova.lm(resmco)`.

9. Dessiner l'intervalle de confiance de la droite de régression.

10. Extraire les résidus de la régression avec la commande `residuals`. Vérifier que l'on obtient la même chose "à la main" en utilisant les séries `maxO3`, `T12` et les coefficients de la régression.

11. Tracer la densité estimée des résidus, leur évolution en fonction du temps, puis dessiner les résidus en fonction de `T12`. Enfin, calculer la moyenne des résidus et la covariance entre ces résidus et `T12`.

Partie B Explication du pic d'ozone par une régression linéaire multiple

1. Dessiner `maxO3` en fonction des différentes variables. Quelles sont celles qui sont *a priori* intéressantes ?

2. Effectuer la régression de `maxO3` en fonction de toutes les variables et utiliser la commande `summary.lm` pour obtenir les détails de la régression.

3. Effectuer "à la main" une procédure "backward" pour sélectionner les variables : on estime le modèle, on retire la variable la moins significative et on recommence jusqu'à ce que toutes les variables soient significatives.

Exercice 2 (Simulations)
1. Simuler deux vecteurs de 100 variables $\mathcal{U}[0,1]$ indépendantes, x_1 et x_2. Définir $\beta_1 = 0,5$, $\beta_2 = -4$ et $\beta_3 = 3,8$.

2. Créer une fonction qui :
- simule un vecteur ε de longueur 100 suivant une loi normale de moyenne 0 et de variance $\sigma^2 = 0,2$.
- calculer ensuite le vecteur $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 + \varepsilon$.
- retourne les estimations MCO de β_1, β_2, β_3 et σ^2.

3. Appeler 1000 fois la fonction précédente et dessiner une approximation de la distribution des estimateurs β_1, β_2, β_3 et σ^2.