Fiche 5 - TISD - Master Pro
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`

1 Régression linéaire avec R

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.

2 Régression linéaire avec SAS

Exercice 3 (Ozone (suite))
Cet exercice est la suite de l'exercice 1. On utilise la table `ozone2.sas7bdat`.

1. Etudier à l'aide de la `PROC CORR` la corrélation linéaire entre les variables de cette table.

2. Estimer la régression de `maxO3` en fonction de toutes les variables en utilisant la `PROC REG`.

3. Utiliser l'option `stepwise` de la `PROC REG` pour sélectionner les variables intéressantes :

```
proc reg data=malib.ozone;
model maxO3 = .../selection=stepwise;
run;
```

4. Réaliser les sélection `forward`, `backward` et `stepwise`. Commenter et comparer-les.

5. Reprendre la régression de `maxO3` sur la seule température `T12` :

```
proc reg data=malib.ozone;
model maxO3=T12;
plot maxO3*T12='*';
run;
```

où la ligne `plot` trace la droite de régression.

Exercice 4 (Entreprises : Rendements d'échelle et choix d'une fonction de Cobb-Douglas)
Nous disposons d'un échantillon de 1658 entreprises (table `tdeco.sas7bdat`, extrait de fichier de la Comptabilité Nationale de 1979) pour lesquelles sont donnés :

- l'effectif L (variable EFFEC)
- la valeur ajoutée brute aux coûts des facteurs VA, en francs constants de 1970 (variable VABCF)
- le capital K, en francs constants (IBD)
- la branche d'appartenance, avec une nomenclature de niveau 40 (N40B).

Partie A : Rendements d'échelle (**SAS**)

Modéliser la valeur ajoutée par une fonction de production Cobb-Douglas revient à la supposer de la forme :

$$VA = \lambda L^b K^c, \tag{1}$$

où λ, b et c sont des paramètres réels à estimer.

1. (théorique) En déduire que le modèle linéaire associé est :

$$\log(VA) = \log(\lambda) + b\log(L) + c\log(K). \tag{2}$$

Quelles sont les interprétations des coefficients $\log(\lambda)$, b et c ?

2. Estimer ce modèle par MCO.
2.1. Donner les valeurs de $\widehat{\lambda}$, $\widehat{b}$ et $\widehat{c}$.
2.2. Quel est le coefficient de détermination ?
2.3. Donner les intervalles de confiance pour les paramètres a, b et c.
2.4. Tester l'hypothèse H_0 : $c = 0$ contre l'hypothèse alternative H_a : $c \neq 0$.
2.5. Comment tester l'hypothèse de constance des rendements d'échelle ? Effectuer ce test.
2.6. Effectuer le test d'égalité du coefficient b à 1.

3. Effectuer les régressions par branche pour :

- $T07$: Minerais et métaux ferreux, première transformation de l'acier,
- $T15B$: Bien d'équipement ménager,
- $T18$: Industrie textile et de l'habillement.

3.1. Commenter la valeur des coefficients de détermination obtenus.
3.2. Quels sont les coefficients significatifs ?
3.3. Les trois branches sont-elles régies par le même modèle ?
3.4. Comment pouvez-vous avoir une indication sur le respect de l'hypothèse "les variances des perturbations sont les mêmes dans les trois branches" ?

Partie B : Fonction de production Cobb-Douglas et CES (**SAS**)

La fonction de production CES s'écrit :

$$VA = B\left(\delta K^{-\rho} + (1-\delta)L^{-\rho}\right)^{-\mu/\rho}. \tag{3}$$

1. (théorique) Montrer que lorsque $\rho \to 0$, on retrouve une fonction de production du type Cobb-Douglas.
Le problème qui nous intéresse dans la suite est alors de tester :
H_0 La fonction de production est une fonction de Cobb-Douglas ($\rho = 0$)
H_a La fonction de production est de type CES ($\rho \neq 0$).

2. (théorique) Montrer que le modèle linéaire associé à (3) est :

$$\log\left(\frac{VA}{L}\right) = \log(B) + (\mu-1)\log(L) - \frac{\mu}{\rho}\log\left(\delta\left(\frac{K}{L}\right)^{-\rho} + 1 - \delta\right). \tag{4}$$

Montrer par un développement limité au voisinage de $\rho = 0$ l'approximation de Kmenta :

$$\log\left(\frac{VA}{L}\right) = \log(B) + (\mu-1)\log(L) + \mu\delta\log\left(\frac{K}{L}\right) - \frac{\mu\rho\delta(1-\delta)}{2}\left(\log\left(\frac{K}{L}\right)\right)^2 + (\rho^2). \tag{5}$$

3. Nous nous intéressons donc au modèle linéaire suivant :

$$\log\left(\frac{VA}{L}\right) = a + b\,\log(L) + c\,\log\left(\frac{K}{L}\right) - d\left(\log\left(\frac{K}{L}\right)\right)^2. \tag{6}$$

3.1. Quel test est équivalent à H_0 ?
3.2. Effectuer la régression par les MCO et conclure.

Partie C Détection et correction d'hétéroscédasticité

Nous reprenons le modèle (2).

1. Une manière de tester l'hétéroscédasticité consiste à régresser les résidus élevés au carré du modèle (2) sur les variables explicatives, puis à tester la significativité de tous les coefficients autres que la constante. Si la statistique de Fisher globale rejette l'hypothèse nulle (tous les coefficients autres que la constante sont nuls), une hétéroscédasticité éventuelle est détectée. Effectuer un test d'hétéroscédasticité suivant cette approche.

2. Même si les paramètres estimés par MCO restent convergents en présence d'hétéroscédasticité, ils sont alors non-efficaces. Pour obtenir une estimation efficace, il faut utiliser la méthode des Moindres Carrés Généralisés. Une manière de faire consiste à utiliser les valeurs ajustées W de la régression des résidus au carré sur les régresseurs pour transformer les données et estimer par MCO les coefficients (b_0, b_1, b_2) dans le modèle :

$$\frac{\log(VA)}{\sqrt{\mathrm{W}}} = \frac{b_0}{\sqrt{\mathrm{W}}} + b_1 \frac{\log(L)}{\sqrt{\mathrm{W}}} + b_2 \frac{\log(K)}{\sqrt{\mathrm{W}}} \tag{7}$$

(attention, vérifier que toutes les valeurs ajustées W sont positives) Estimer par MCG le modèle puis comparer les résultats avec l'estimation MCO.

Partie D Synthèse

Rédiger en quelques phrases (pas plus d'une demi-page !) une petite synthèse destinée à des non-spécialistes (par exemple un article du journal de la fac).