Devoir 2
STT 2400 Eté 2024

Instructions

- Date limite de remise : **4 août 2024 à 23h 59**.
- Le Devoir est sur **100**. A partir du 5 août à 00h 00, une pénalité de **25 points** sera déduite. Chaque jour de retard supplémentaire entraînera également une pénalité de **25 points**. Ainsi, le **8 août à 00h 00**, la note sera de **0** pour ceux qui n'auront pas rendu leur devoir.
- Ce devoir est à faire en **groupe de deux**. **Vous devez garder les mêmes groupes que pour le devoir 1**. Aucune communication ou collaboration n'est permise entre les étudiants de la classe ou avec une tierce partie.
- Le dépôt du devoir doit être fait au format PDF et déposer sur Studium. **Vous devez le faire avec les logiciels Word ou Latex excepté l'exercice 1 que vous pouvez faire de façon manuscrite, mais il faudra le numériser en s'assurant que tout est clair.**
- Les programmes informatiques doivent être écrits dans R ou SAS et **déposer sur Studium**.
- Pour les questions nécessitant R ou SAS, vous devez ajouter les résultats des sorties R ou SAS dans le pdf avec des commentaires si nécéssaire.

Exercice 1 : Test d'égalité de moyenne entre deux groupes (25 points)

Vous avez des données provenant de deux groupes distincts. La variable dépendante quantitative est $y = (y_1, y_2, \ldots, y_{n_1}, y_{n_1+1}, \ldots, y_{n_1+n_2})$ et la variable indicatrice de groupe est $x = (\underbrace{0, \ldots, 0}_{n_1 \text{ observations}}, \underbrace{1, \ldots, 1}_{n_2 \text{ observations}})$, où $x_i = 0$ pour les observations du groupe 1 et $x_i = 1$ pour les observations du groupe 2. On suppose que $y_i \sim \mathcal{N}(\mu_1, \sigma^2)$, pour $i = 1, \ldots, n_1$ et $y_i \sim \mathcal{N}(\mu_2, \sigma^2)$, $i = n_1 + 1, \ldots, n_1 + n_2$. La statistique pour tester l'égalité des moyennes ($H_0 : \mu_1 = \mu_2$ entre les deux groupes est

$$t = \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

et cette dernière suit une loi de student $T_{n_1+n_2-2}$. On note

$$n = n_1 + n_2, \quad \bar{y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i, \quad \bar{y}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n} y_i, \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2, \quad s_2^2 = \frac{1}{n_2 - 1} \sum_{i=n_1+1}^{n} (y_i - \bar{y}_2)^2.$$

On considère maintenant le modèle de régression linéaire $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ où $var(\epsilon_i) = \sigma^2$, $cov(\epsilon_i, \epsilon_j) = 0$ $(i \neq j)$, $i, j = 1, \ldots, n_1 + n_2$.

a) Montrez que :

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \frac{n_1 n_2}{n}$$

b) Montrez que l'estimateur des MCO de β_1 s'écrit $\hat{\beta}_1 = \bar{y}_2 - \bar{y}_1$.

c) Montrez que l'estimateur des MCO de β_0 s'écrit $\hat{\beta}_0 = \bar{y}_1$.

Indice : Pour les questions b) et c), vous pouvez utiliser le fait que : $\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1+n_2}$.

d) Montrez que l'estimateur MCO de σ^2 est égale à s_p^2.

e) Utilisez ces résultats pour montrer l'égalité entre la statistique de test t_1 pour tester H_0 : $\beta_1 = 0$ dans le modèle de régression et la statistique t énoncer en début d'exercice pour tester l'égalité de moyenne.

Exercice 2 (30 points)

Les données `mtcars` sont tirées de la revue "Motor Trend US" de 1974. Elles portent sur la consommation d'essence (mpg) et 10 autres variables sur le design et la performance de 32 automobiles des modèles 1973-74.

— mpg Miles/(US) gallon
— cyl Number of cylinders
— disp Displacement (cu.in.)
— hp Gross horsepower
— drat Rear axle ratio
— wt Weight (1000 lbs)
— qsec 1/4 mile time
— vs Engine (0 = V-shaped, 1 = straight)
— am Transmission (0 = automatic, 1 = manual)
— gear Number of forward gears
— carb Number of carburetors

1. Déterminer les 4 meilleurs modèles parmi les 45 modèles à 2 variables explicatives (à part l'ordonnée) selon le critère R^2. Est-ce que la variable `wt` semble jouer un rôle important ?
2. Le modèle à deux variables explicatives `wt` et `qsec` est un des 4 modèles en (a). Discuter au moyen d'un graphique de la régression partielle pour la variable `wt` de l'utilité d'ajouter un terme quadratique en `wt` à ce modèle.
3. Le modèle à deux variables explicatives `wt` et `qsec` est un des 4 modèles en (a). Discuter au moyen d'un graphique du résidu partiel pour la variable `wt` de l'utilité d'ajouter un terme quadratique en `wt` à ce modèle.
4. Le modèle à deux variables explicatives `wt` et `qsec` est un des 4 modèles en (a). Vérifier s'il y a des valeurs aberrantes dans ce modèle.
5. Le modèle à deux variables explicatives `wt` et `qsec` est un des 4 modèles en (a). A l'aide d'un graphique approprié, vérifier la présence d'hétéroscédasticité.
6. Le modèle à deux variables explicatives `wt` et `qsec` est un des 4 modèles en (a). Vérifier la normalité des résidus. Y a-t-il des points leviers, influents ?

Exercice 3 (30 points)

Une compagnie de conseil en gestion a obtenu les salaires et d'autres informations sur 100 dirigeants de différentes compagnies. Leur but était de prévoir le salaire à partir d'autrtes variables et de déterminer les variables explicatives utiles. Les données se trouvent dans le fichier `execsal2.txt` et portent sur les variables suivantes :

— Row number (ignorer)
— Log of annual salary
— experience (years)
— education (years)
— gender (1=male, 0=female)
— number of employees supervised
— corporate assets (millions of dollars)
— board member (1=yes, 0=no)
— age (years)
— company profits (past 12 months, millions of dollars)
— has international responsibility (1=yes, 0=no)
— company's total sales (past 12 months, millions of dollars)

Le conseil a utilisé le logartithme du salaire car sa relation avec les autres variables lors d'études passées similaires semble plus linéaire. Une conséquence intéressante de ce logarithme est qu'une augmentation d'une unité d'une des variables explicatives est associée avec une ceratine augmentaion en pourcentage du salaire annuel, ce qui est souvent sensé. Donner tous les résultats demandés accompagnés du code `SAS ou R`.

1. Lire les données en `SAS ou R` et attribuer aux variables les noms se trouvant dans la première ligne.
2. Faire une régression pour prévoir "Log of annual salary" à partir de toutes les autres variables, excluant la variable "Row number".
3. Quelle variable est la moins significative ? Identifier la variable et refaire la régression sans cette variable.
4. Continuer à enlever une à une la variable la moins significative jusqu'à ce que toutes les variables restantes soient très significatives.
5. Quelles variables explicatives se trouvent dans le modèle final ? Donner leur nom complet comme dans la liste ci-dessus.
6. Observer les coefficients de régression. Sont-ils positifs ou négatifs ? Est-ce que cela fait du sens dans notre contexte ?

Exercice 4 (15 points)

Pour une marque particulière d'automobile, une expérience avait pour but d'étudier l'usure de pneu y. Deux facteurs d'intérêt sont la température en degrés Fahrenheit x_1 et le nombre de miles parcourus x_2 sur une chaussée mouillée. On effectue 4 tests pour chaque combinaison de (x_1, x_2). Les données se trouve dans le fichier `usure.txt`.

1. Lire les données en `SAS ou R`.
2. Effectuer le test de validité au moyen des erreurs pures pour le modèle I
$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$
3. Refaire le test en (b) pour le modèle II
$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_2^2 + \epsilon.$$