TP noté

Romain Lesauvage

16/11/2020

Ce TP noté est à faire individuellement. Il est à rendre par mail (lesauvageromain@gmail.com) avant le 20 novembre 23h59.

Pour ceux qui maîtrisent RMarkDown, c'est l'idéal. Sinon, vous pouvez me rendre d'un côté votre code R, et de l'autre vos réponses aux questions dans un LaTeX, un word, une photo lisible d'une feuille manuscrite etc.

Bon courage !

Exercice

On considère un jeu de données venant d'une étude de Stamey et al. (1989). Dans cette étude ils étudient la corrélation entre le niveau d'un antigène spécifique de la prostate et un certain nombre de variables cliniques, chez des hommes ayant subi prostatectomie complète. Les variables sont le log du volume de la tumeur cancéreuse (*lcavol*), le log du poids de la prostate (*lweight*), l'âge *age*, le log du pourcentage d'hypertrophie bénine de la prostate (*lbph*), l'invasion séminale de (*svi*), log de la pénétration capsulaire (*lcp*), le score de gleason (*gleason*), et le pourcentage de score de Gleason valant 4 et 5 (*pgg*45). Les variables *svi* et *gleason* étant qualitatives, on les enlève de la modélisation.

Dans ce qui suit on considère

$$Y = \text{level of prostate-specific antigen,}$$

et

$$X = (\text{lcavol}, \text{lweight}, \text{age}, \text{lbph}, \text{lcp}, \text{pgg45})^T.$$

1. Charger les données "prostate" de R
2. Transformer les variables *svi* et *gleason* en variables qualitatives et les laisser de coté pour l'instant.
3. Commencer par le modèles de régression linéaire simple expliquant *lpsa* en fonction de *lcavol*.

   Pour ce modèle

   (a) Donner la variable expliquée et son type ainsi que les variables explicatives et leurs types.

   (b) En déduire le type de modèle à mettre en oeuvre.

   (c) Donner le modèle mis en oeuvre ainsi que toutes les quantités qui apparaissent.

   (d) Donner la dimension de ce modèle.

   (e) Donner l'expression des estimateurs des moindres carrés des paramètres.

   (f) Donner l'équation de la droite des moindres carrés obtenue à partir des observations.

   (g) En déduire une prédiction de la variables *lpsa* pour valeur de *lcavol* égale à 1.35.

   (h) Quels sont les tests mis en oeuvre? On précisera les hypothèses testées, les p-values et les conclusions de ces tests.

(i) On s'intéresse au test de nullité de la pente. Ecrire les deux hypothèses testées, la statistique de test, sa loi sous l'hypothèse nulle, ainsi que la forme de sa zone de rejet.

(j) Réécrire ce test sous la forme d'une comparaison de modèles en décrivant les modèles testés, la statistique de test associée et la conclusion du test.

(k) Tracer le plot des résidus en fonction des valeurs prédites et le QQ-plot pour comparer les résidus à une loi normale. Que concluez-vous ?

4. Considérons maintenant le modèle complet.

   (a) Décrire le modèle mis en oeuvre en prenant soin de définir toutes les quantités qui apparaissent.

   (b) Donner la dimension de ce modèle.

   (c) Donner l'expression matricielle du modèle.

   (d) On note V le sous-espace vectoriel engendré par les colonnes de la matrice $\mathbf{X}$ apparaissant dans l'écriture matricielle du modèle. Que vaut l'estimateur de $\mathbb{E}(Y)$? Qulles équations doit-il satisfaire?

   (e) En déduire l'expression matricielle de l'estimateur du vecteur des paramètres.

   (f) Décrire les tests mis en oeuvre par la commande R et donner les conclusions des tests.

5. On va maintenant faire des comparaisons de modèles emboités. On va par exemple comparer le modèle *lpsa* en fonction de $lcavol, pgg45$, *lpsa* en fonction de *lcavol* et *lweight* ou *lpsa* en fonction de $lcavol, lweight$ et *age*. Que concluez-vous ?