Régression linéaire

Enfin les élélements de correction sont disponibles sur le site http://josephsalmon.eu/index.php?page=teaching_13_14&lang=fr sous forme de fichiers R.

- Rappels sur le modèle linéaire -

Nous considérons le modèle statistique suivant :

$$\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{1}$$

où

- $\mathbf{y} = (y_i)_{1\leq i\leq n} \in \mathbb{R}^n$ est un vecteur colonne $n \times 1$,
- $X = (X_{i,j})_{1\leq i\leq n, 0\leq j\leq p}$ est une matrice $n \times (p+1)$ de **rang plein** telle que $X_{i,0} = 1$ pour tout $1 \leq i \leq n$,
- $\boldsymbol{\theta} = (\theta_i)_{0\leq i\leq p} \in \mathbb{R}^{p+1}$ est un vecteur colonne $(p+1) \times 1$,
- $\boldsymbol{\varepsilon} = (\varepsilon_i)_{1\leq i\leq n} \in \mathbb{R}^n$ un vecteur colonne $n \times 1$ aléatoire.

On suppose de plus que le vecteur $\boldsymbol{\varepsilon}$ suit la distribution :

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \ .$$

On rappelle les notations suivantes :

- $\hat{\boldsymbol{\theta}} = (X^\top X)^{-1} X^\top \mathbf{y}$ l'estimateur par moindres carrés de $\boldsymbol{\theta}$ quand la matrice $(X^\top X)$ est inversible.
- $\hat{\mathbf{y}} = X\hat{\boldsymbol{\theta}}$, la prédiction sur les valeurs observées
- $\bar{x}_n = n^{-1}\sum_{i=1}^n x_i$ et $\bar{y}_n = n^{-1}\sum_{i=1}^n y_i$, les moyennes empiriques
- $\mathrm{var}_n(\mathbf{x})$ et $\mathrm{var}_n(\mathbf{y})$, les variances empiriques
- Le vecteur $r = \mathbf{y} - \hat{\mathbf{y}}$ est appelé vecteur des résidus.
- $\mathbf{1}_n = (1, \ldots, 1)^\top$ est le vecteur "tout à un" de taille $n \times 1$
- On note $\mathrm{RSS} = ||\mathbf{y} - \widehat{\mathbf{y}}||^2$ (*Residual Sum of Squares* en anglais).

- Régression linéaire simple -

On utilisera le langage R et par exemple le logiciel RStudio pour faire ce TP. Le mot "régression" a été introduit par Sir Francis Galton (cousin de C. Darwin) alors qu'il étudiait la taille des individus au sein d'une descendance.

1. Récupérer les données du fichier http://www.math.uah.edu/stat/data/Galton.txt. La seconde colonne contient la taille du parent "moyen", c'est-à-dire $\frac{1}{2}$ (taille(pere) + 1.08taille(mere)). La première colonne contient la taille d'un de leurs enfants (à l'âge adulte). On note x_i la taille du parent moyen pour la famille i et y_i la taille de l'enfant. On écrit $y_i = \theta_1 x_i + \theta_0 + \varepsilon_i$ et on modélise les variables ε_i comme gaussienne centrées, indépendantes de même variance σ^2 inconnue.
2. Tracer le nuage de points $(x_i.y_i)$ pour $1 \leq i \leq n$ où n est le nombre de familles figurant dans les données. Utiliser la fonction `plot`.
3. Estimer θ_0, θ_1, par $\hat{\theta}_0$, $\hat{\theta}_1$ en utilisant la fonction `lm` puis en vérifiant les formules vues en cours pour le cas unidimensionnel.
4. Calculer et visualiser les valeurs prédites $\hat{y}_i = \hat{\theta}_1 x_i + \hat{\theta}_0$ et y_i sur un même graphique.
5. Visualiser l'histogramme des résidus $r_i = y_i - \hat{y}_i$. L'hypothèse de normalité est-elle crédible ? On pourra aussi s'appuyer sur la fonction `qqnorm`.
6. Régresser $\mathbf{x}$ sur $\mathbf{y}$ et comparer les coefficients $\hat{\alpha}_0$ et $\hat{\alpha}_1$ obtenus par rapport aux $\hat{\theta}_0$ et $\hat{\theta}_1$ du modèle original. Vérifier numériquement (et éventuellement en exercice formellement) que :

$$\begin{aligned}\hat{\alpha}_0 &= \bar{x}_n + \frac{\bar{y}_n}{\bar{x}_n}\frac{\mathrm{var}_n(\mathbf{x})}{\mathrm{var}_n(\mathbf{y})}(\hat{\theta}_0 - \bar{y}_n)\\ \hat{\alpha}_1 &= \frac{\mathrm{var}_n(\mathbf{x})}{\mathrm{var}_n(\mathbf{y})}\hat{\theta}_1\end{aligned}$$

- Régression linéaire multiple -

On travaille maintenant sur le fichier `auto-mpg.data` et on cherche à régresser la consommation des voitures sur leurs caractéristiques : nombre de cylindres, cylindrés (*engine displacement* en anglais), puissance, poids, accélération, année, pays d'origine et le nom de la voiture. On utilise le modèle (1), où $\mathbf{y}$ est le vecteur contenant les consommations des voitures (plus précisément la distance parcourue en miles par Gallon, ou mpg) ; les colonnes de X sont les régresseurs quantitatifs [1].

7. Importer la base de données avec la commande `read.table`.
8. Calculer $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{y}}$ sur une sous partie de la base : garder les 9 premières lignes et les 8 premières colonnes. Que constatez-vous ?
9. Calculer $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{y}}$ cette fois sur l'intégralité des données.
10. Calculer le carré de la norme du vecteur des résidus $\text{RSS} = \|r\|^2$, puis la moyenne des ces écarts $MSE = SSE/(n-p-1)$ (*Mean Square Errors* en anglais). Vérifier numériquement que :

$$\|\mathbf{y} - \bar{y}_n \mathbf{1}_n\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \bar{y}_n \mathbf{1}_n\|^2.$$

    On proposera une méthode matricielle pure, et une méthode utilisant la fonction `lm` de R.
11. Supposons que l'on vous fournisse les caractéristiques suivantes d'un nouveau véhicule :

cylinders	displacement	horsepower	weight	acceleration	year	origin
6	225	100	3233	15.4	76	1

    Prédire sa consommation [2].
12. Calculer de nouveau $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{y}}$ mais cette fois sur sur les données centrées réduites.

Liens pour aller plus loin :

⋆⋆⋆ http://cbio.ensmp.fr/~jvert/svn/tutorials/practical/linearregression/linearregression.R

⋆⋆ http://perso.univ-rennes1.fr/bernard.delyon/regression.pdf

1. sauf la variable du nom, et la variable "origine". Pour cette dernière, si on veut l'intégrer il faut introduire 3 nouvelles variables explicatives binaires (une pour chaque origine).
2. A titre d'information, la consommation effectivement mesurée sur cet exemple était de 22 mpg.