Etude de cas : Régression linéaire

Rappels

Lois utiles

La loi de Student T_r est la loi suivie par la variable aléatoire

$$\frac{\sqrt{r}U}{\sqrt{\sum_{i=1}^{r} V_i^2}} ,$$

où les U et V_i sont des $\mathcal{N}(0,1)$ indépendantes. La loi de Fisher $F_{q,r}$ est la loi suivie par la variable aléatoire

$$\frac{r\sum_{i=1}^{q} U_i^2}{q\sum_{i=1}^{r} V_i^2} ,$$

où les U_i et V_i sont des $\mathcal{N}(0,1)$ indépendantes.

Modèle linéaire

Nous considérons le modèle statistique suivant :

$$Y = \Phi\theta + \varepsilon, \tag{1}$$

où

- $Y = (Y_i)_{1\le i\le n} \in \mathbb{R}^n$ est un vecteur colonne $n \times 1$,
- $\Phi = (\Phi_{i,j})_{1\le i\le n, 0\le j\le p}$ est une matrice $n \times (p+1)$ de rang plein telle que $\Phi_{i,0} = 1$ pour tout $1 \le i \le n$,
- $\theta = (\theta_i)_{0\le i\le p} \in \mathbb{R}^{p+1}$ est un vecteur colonne $(p+1) \times 1$,
- $\varepsilon = (\varepsilon_i)_{1\le i\le n} \in \mathbb{R}^n$ un vecteur colonne $n \times 1$ aléatoire.

On suppose de plus que le vecteur ε suit la distribution :

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

On rappelle les notations suivantes :

- $\hat{\theta} = (\Phi^T\Phi)^{-1}\Phi^T Y$ l'estimateur par moindres carrés de θ.
- $\hat{Y} = \Phi\hat{\theta}$
- $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$.
- Le vecteur $e = Y - \hat{Y}$ est appelé vecteur des résidus.
- $\mathbf{1} = (1, \dots, 1)^T$ est le vecteur "tout à un" de taille $n \times 1$
- $\mathrm{SST} = ||Y - \overline{Y}\mathbf{1}||^2$, $\mathrm{SSR} = ||\widehat{Y} - \overline{Y}\mathbf{1}||^2$ (parfois aussi appelé SSM), $\mathrm{SSE} = ||Y - \widehat{Y}||^2$.
- L'estimateur de la variance est $\hat{\sigma}^2 = \mathrm{SSE}/(n-p-1)$.

1 Préliminaires

1. Rappelez quelle est la loi suivie par

$$\frac{\widehat{\theta}_j - \theta_j}{\sqrt{\hat{\sigma}^2 \left(\Phi^T \Phi\right)^{-1}_{jj}}}$$

pour $0 \leq j \leq p$. En déduire l'expression d'un intervalle de confiance pour θ_j au niveau $1-\alpha$ à l'aide des fonctions quantiles de la loi de Student à m degré de liberté, notée qt_m.

2. Rappelez la loi suivie par

$$\frac{\text{SSE}}{\sigma^2} .$$

En déduire l'expression d'un intervalle de confiance pour σ^2 au niveau $1-\alpha$ à l'aide des fonctions quantiles des lois du χ^2 à m degrés de liberté, notée $q\chi^2_m$.

3. Soit $\phi_0 \in \mathbb{R}^{p+1}$ un nouveau vecteur de variables explicatives et $Y_0 \in \mathbb{R}$ une variable telle que

$$Y_0 = \phi_0^T \theta + \varepsilon_0,$$

où $\varepsilon_0 \sim \mathcal{N}(0, \sigma^2)$, indépendant de ε et de même variance σ^2. On définit $\hat{Y}_0 = \phi_0^T \hat{\theta}$. Rappelez quelle est la loi suivie par :

$$\frac{\hat{Y}_0 - Y_0}{\sqrt{(1 + \phi_0^T (\Phi^T \Phi)^{-1} \phi_0)\, \hat{\sigma}^2}} , \tag{2}$$

et en déduire un intervalle de confiance pour Y_0 au niveau $1-\alpha$.

2 Régression linéaire simple

On utilisera la boîte à outils `stixbox` qui regroupe des fonctions matlab ayant trait aux statistiques et est distribuée sous la license GPL. On peut librement la télecharger à l'adresse http ://www.maths.lth.se/matstat/stixbox/. Les fonctions `rt.m`, `dt.m`, `pt.m` et `qt.m` permettent d'obtenir un échantillon suivant une loi de Student, la densité, la fonction de répartition et la fonction quantile d'une loi de Student. Pour la loi de Fisher, il s'agit des fonctions `rf.m`, `df.m`, `pf.m` et `qf.m`. Pour la loi du χ^2, il s'agit des fonctions `rchisq.m`, `dchisq.m`, `pchisq.m` et `qchisq.m` [1].

Le mot "régression" a été introduit par Sir Francis Galton alors qu'il étudiait la taille des individus au sein d'une descendance.

1. Si le nombre de degrés de liberté est trop important `qchisq.m` renvoie `NaN`, on utilisera alors l'approximation de `qchisq(a,m)` par `qnorm(a,0,1)*sqrt(2*m)+m`

4. Récupérer les données du fichier `galton.dat`. La première colonne contient la taille du parent "moyen", c'est–à–dire $\frac{1}{2}$ (taille(pere) + 1.08taille(mere)). La seconde colonne contient la taille d'un de leur enfant (à l'age 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.
5. 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.
6. Estimer θ_0, θ_1 et σ^2.
7. Déterminer les intervalles de confiance correspondants.
8. Calculer et visualiser les valeurs prédites $\hat{Y}_i = \hat{\theta}_1 x_i + \hat{\theta}_0$ et Y_i sur un même graphique.
9. Visualiser l'histogramme des résidus $e_i = Y_i - \hat{Y}_i$. L'hypothèse de normalité est-elle crédible ? On pourra aussi s'appuyer sur la fonction `qqnorm.m` de `Stixbox`.
10. Régresser x sur y et comparer les coefficients obtenus.

On souhaite élaborer une procédure de test permettant de discriminer les deux hypothèses suivantes :

$$\begin{aligned} H_0 : &\quad \theta_1 = 0 \\ H_1 : &\quad \theta_1 \neq 0 \ . \end{aligned}$$

11. Rappeler quelle est la loi suivie par la statistique

$$\mathcal{S} := \frac{\mathrm{SSR}}{\mathrm{SSE}/(n-2)}$$

sous l'hypothèse H_0. Tracer la fonction de répartition complémentaire $\bar{F}$ de cette loi.
12. Calculer le rapport $\mathcal{S}$ pour le jeu de données et évaluer la p-valeur $\bar{F}(\mathcal{S})$.
Que concluez-vous ?

3 Régression linéaire multiple

On travaille maintenant sur le fichier `cars.dat` et on cherche à régresser la consommation des voitures sur leurs caractéristiques : cylindrée, cylindrée, puissance, poids, accélération, année, pays d'origine. On utilise le modèle (1), où Y est le vecteur contenant les consommations des voitures, et où les colonnes de Φ sont les régresseurs quantitatifs [2].

13. Calculer $\hat{\theta}$, $\hat{Y}$, $\hat{\sigma}^2$.
14. Visualiser les intervalles de confiance à 95% de θ_j pour $0 \leq j \leq p$.
Y a-t-il des intervalles de confiance qui contiennent la valeur zéro ? Intuitivement, que peut-on en déduire quant à l'influence des régresseurs correspondants ?

2. tous sauf la dernière variable "origine". Pour cette dernière, si on veut l'intégrer il faut introduire 3 nouvelles variables explicatives binaires (une pour chaque origine).

15. 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[3]. Donner l'intervalle de confiance à 95% associé à cette prédiction.

4 Sélection de variables

On souhaite mettre en œuvre une procédure permettant de conserver uniquement les variables explicatives jugées utiles. En indexant les variables explicatives de 1 à p, cela revient à sélectionner un certain ensemble $A = \{i_1, \cdots, i_k\} \subset \{1, \cdots, p\}$.
Pour tout sous-ensemble $A = \{i_1, \cdots, i_k\}$, on note $\hat{\theta}_A$ l'estimateur des moindres carrés obtenu en ne conservant que les variables explicatives indexées $i_1, \cdots, i_k$, c'est à dire :

$$\hat{\theta}_A = (\Phi_A^T \Phi_A)^{-1} \Phi_A^T Y \ ,$$

où Φ_A est la matrice obtenue en ne conservant de Φ que les régresseurs d'indices $i_1, \cdots, i_k$.
On note $\hat{Y}_A = \Phi_A \hat{\theta}_A$ le prédicteur associé au sous-modèle A. Enfin, on note :

$$\mathrm{SSE}_A = \|Y - \hat{Y}_A\|^2$$

la norme au carré du vecteur des résidus.

16. Expliquer pourquoi il serait maladroit de déterminer A par minimisation de SSE_A :

$$\min_{A \in 2^{\{1,\cdots,p\}}} \mathrm{SSE}_A \ .$$

Séléction de variables par stratégie descendante

Afin de choisir parmi les régresseurs, on commence par formuler un simple test binaire :

$$\begin{aligned} H_0 : &\quad \text{Seuls les coefficients } \theta_0, \theta_{i_1}, \ldots, \theta_{i_k} \text{ sont non-nuls.} \\ H_1 : &\quad \text{Tous les coefficients } \theta_0, \theta_1, \ldots, \theta_p \text{ sont non-nuls.} \end{aligned}$$

On propose de rejeter H_0 pour de fortes valeurs de la statistique

$$\mathcal{S}_A = \frac{(\mathrm{SSE}_A - \mathrm{SSE})/(p-k)}{\mathrm{SSE}/(n-p-1)} \ .$$

17. Montrer que sous H_0, $\mathcal{S}_A$ suit une loi de Fisher de paramètres $(p-k, n-p-1)$.
18. Tester sur le jeu de données la procédure de sélection descendante de la Table 1.

3. A titre d'information, la consommation effectivement mesurée sur cet exemple était de 22 mpg.

1. Initialisation : Considérer le grand modèle avec tous les régresseurs.
2. Calculer la statistique de Fisher correspondant au retrait de chaque variable.
3. Si toutes les p-valeurs sont inférieures au niveau α, stopper la procédure. Sinon, enlever du modèle le régresseur de plus forte p-valeur.
4. Retourner à l'étape 2.

TABLE 1 – Stratégie de sélection descendante

Sélection de variables par validation croisée

On note $Y^{(-i)} = (Y_1 \cdots Y_{i-1}, Y_{i+1} \cdots Y_n)^T$ le vecteur Y duquel on a retiré la $i$ème composante. On note $\Phi_A^{(-i)}$ la matrice Φ_A de laquelle on a éliminé la $i$ème ligne, et on considère l'estimée

$$\hat{\theta}_A^{(-i)} = ({\Phi_A^{(-i)}}^T \Phi_A^{(-i)})^{-1} {\Phi_A^{(-i)}}^T Y^{(-i)} \ .$$

On définit le critère CV (Cross Validation) par :

$$\mathrm{CV}(A) = \sum_{i=1}^{n} (Y_i - [\Phi_A \hat{\theta}_A^{(-i)}]_i)^2 \ ,$$

où $[x]_i$ représente la $i$ème composante du vecteur x.

19. Quelle interprétation peut-on donner à $\mathrm{CV}(A)$?
20. Pour le jeu de données `cars.dat`, trouver le jeu de variables explicatives minimisant $\mathrm{CV}(A)$.
21. Plutôt que d'utiliser un critère de test statistique on peut utiliser un critère de type CV dans l'algorithme descendant : on compare le CV du modèle complet avec les modèles où on enlève une variable, et on garde le plus petit CV ; on recommence tant qu'un modèle petit "gagne". Expliquer l'intérêt d'une telle méthode.

5 Détection de données aberrantes

On rappelle que si le modèle (1) est valide, le vecteur e suit la loi $\mathcal{N}(0, \sigma^2(I - H))$ où H est le projecteur orthogonal $H = \Phi(\Phi^T\Phi)^{-1}\Phi^T$. En particulier,

$$e_i \sim \mathcal{N}(0, \sigma^2(1 - h_i)) \tag{3}$$

où h_i est le $i$ème coefficient diagonal de H. Si la $i$ème observation est en adéquation avec le modèle, le résidu normalisé

$$\frac{e_i}{\sigma\sqrt{1 - h_i}}$$

doit donc se trouver dans l'intervalle ± 1.96 avec probabilité 0,95. En pratique, on n'a malheureusement pas accès à la statistique (3) car la variance σ^2 est inconnue. Une première idée consiste à remplacer (3) par $t_i = e_i/\hat{\sigma}\sqrt{1 - h_i}$. Mais comme le numérateur et le

dénominateur ne sont pas indépendants, la distribution de t_i n'a pas une forme facile à manipuler. On considèrera donc le *résidu studentisé* :

$$t_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}\,,$$

où $\hat{\sigma}_{(i)}$ est l'estimateur de la variance obtenue en éliminant la i-ème observation. Nous pouvons montrer que :

$$\hat{\sigma}^2_{(i)} = \frac{(n-p)\hat{\sigma}^2}{n-p-1} - \frac{e_i^2}{(n-p-1)(1-h_i)}\,. \tag{4}$$

Il s'avère que t_i^* est distribuée suivant une loi de Student à $(n-p-1)$ degrés de liberté.

22. En évaluant l'intervalle de confiance sur t_i^* au niveau $1-\alpha$, détecter la présence d'éventuelles données pour lequel le modèle (1) ne paraît pas adéquat. Représenter les valeurs des résidus e_i en fonction de $\hat{Y}_i$ en représentant les outliers d'une couleur différente.