TD1 : Modèle de régression linéaire

23/01/2025

Le but de ce premier TD et de présenter le modèle classique de régression linéaire et de sa mise en oeuvre à l'aide du logiciel `R`. Nous allons illustrer ces techniques pour la base de données `Carseats` du package `ISLR`.

On cherche à expliquer/prédire une variable statistique Y à l'aide de p variables statistiques numériques $X_1, \ldots, X_p$. Lorsque la variable Y est numérique "continue", on parle de problème de régression. Y s'appelle la variable à expliquer (ou encore, variable réponse, dépendante, endogène...), et $X_1, \ldots, X_p$ les variables explicatives (variables de contôle, indépendantes, exogènes, régresseurs, prédicteurs...). Notons le vecteur aléatoire (ligne)

$$\mathbf{X} := (X_1, \ldots, X_p) \in \mathbb{R}^p$$

des p variables explicatives $X_1, \ldots, X_p$. Le problème de régression consiste à trouver une fontion $f : \mathbb{R}^p \to \mathbb{R}$ telle que $Y \approx f(\mathbf{X})$:

$$Y = f(X_1, \ldots, X_p) + \varepsilon,$$

où ε est une variable aléatoire, non observable, représentant le terme d'erreur (le bruit).

Pour la fonction de perte quadratique, la fonction optimale

$$f^* := \arg\inf_f \ \mathbb{E}\left((Y - f(\mathbf{X})^2)\right)$$

est l'espérance conditionnelle de Y sachant $\mathbf{X} = \mathbf{x}$, i.e., $\forall \mathbf{x} \in \mathbb{R}^p,\ f^*(\mathbf{x}) = \mathbb{E}\left(Y \,|\, \mathbf{X} = \mathbf{x}\right)$.

Si la fonction de régression optimale $f^*(\cdot)$ est paramétrique et est linéaire, i.e., de la forme

$$f(\mathbf{x}) =: f(\mathbf{x}, \mathbf{w}) := w_0 + w_1\, x_1 + \cdots + w_p\, x_p,$$

on obtient le modèle de régression linéaire multiple (RLM) réécrit sous la forme

$$Y = w_0 + w_1\, X_1 + \cdots + w_p\, X_p + \varepsilon,$$

avec $\mathbb{E}(\varepsilon \,|\, \mathbf{X} = \mathbf{x}) = 0$. Le problème est alors d'estimer les paramètres

$$\mathbf{w} := (w_0, w_1, \ldots, w_p)^\top \in \mathbb{R}^{1+p}$$

à partir de n observations,

$$(\mathbf{X}_1, Y_1) \in \mathbb{R}^{p+1},\ \ldots,\ (\mathbf{X}_n, Y_n) \in \mathbb{R}^{p+1},$$

du vecteur aléatoire $(\mathbf{X}, Y) \in \mathbb{R}^{p+1}$. Nous allons utiliser les notations matricielles suivantes

$$\mathbf{Y} := \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n, \quad \mathbb{X} := \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{bmatrix}, \quad \mathbf{w} := \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix} \in \mathbb{R}^{1+p}, \quad \boldsymbol{\varepsilon} := \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in \mathbb{R}^n,$$

qui représentent, respectivement, le vecteur des valeurs observées de la variable réponse Y, la matrice de design (matrice de dimension $n \times (1+p)$), le vecteur des paramètres du modèle de RLM, et enfin le vecteur des termes d'erreur du modèle, $\varepsilon_1, \ldots, \varepsilon_n$, non observés. Le modèle de RLM, associé aux données, s'écrit donc sous la forme matricielle suivante

$$\mathbf{Y} = \mathbb{X}\,\mathbf{w} + \boldsymbol{\varepsilon}.$$

L'estimateur MC

L'estimateur des moindres carrés $\widehat{\mathbf{w}}$, du vecteur $\mathbf{w}$, est défini par

$$\begin{aligned}\widehat{\mathbf{w}} &:= \underset{(w_0,w_1,\ldots,w_p)\in\mathbb{R}^{1+p}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}(Y_i - w_0 - w_1\,X_{i,1} - \cdots - w_p\,X_{i,p})^2 \\ &=: \underset{\mathbf{w}\in\mathbb{R}^{1+p}}{\arg\min}\ \frac{1}{n}\left\|\mathbf{Y}-\mathbb{X}\,\mathbf{w}\right\|^2. \\ &= (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\ \mathbf{Y}. \end{aligned} \tag{1}$$

La dernière égalité a lieu si la matrice $\mathbb{X}^\top\mathbb{X}$ est inversible, ce qui est équivalent à supposer que les colonnes de $\mathbb{X}$ sont linéairement indépendantes (donc le nombre d'observations n doit être supérieur à $1+p$, p étant le nombre de variables explicatives). Cette hypothèse garantit l'existence et l'unicité de l'estimateur MC précédent.

Lien avec l'estimation par maximum de vraisemblance (MV)

On suppose que les termes d'erreur,

$$\varepsilon_1,\ \ldots,\ \varepsilon_n,$$

sont i.i.d. de même loi normale $\mathcal{N}(0,\sigma^2)$, $\mathbb{E}(\varepsilon\,|\,\mathbf{X}=\mathbf{x})=0$, et $\mathrm{Var}(\varepsilon\,|\,\mathbf{X}=\mathbf{x})=\sigma^2$ ne dépendant pas de $\mathbf{x}$ (hypothèse d'homoscédasticité). Par conséquent, la loi de Y, conditionnellement à $\mathbf{X}=\mathbf{x}$, est une loi normale d'espérance $w_0+w_1x_1+\ldots+w_px_p$, et de variance σ^2 ne dépendant pas de $\mathbf{x}$. La log-vraisemblance (conditionnelle) s'écrit donc sous la forme

$$\mathcal{L}_n(\mathbf{w},\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\|\mathbf{Y}-\mathbb{X}\,\mathbf{w}\|^2.$$

On peut voir dans ce cas que l'estimateur du maximum de vraisemblance (MV), de $\mathbf{w}$, coïncide avec l'estimateur des MC : Notons $(\widetilde{\mathbf{w}},\widetilde{\sigma}^2)$ l'EMV de $(\mathbf{w},\sigma^2)$, i.e.,

$$(\widetilde{\mathbf{w}},\widetilde{\sigma}^2) := \underset{\mathbf{w}\in\mathbb{R}^{1+p},\,\sigma^2\in\mathbb{R}_+^*}{\arg\max}\ \mathcal{L}_n(\mathbf{w},\sigma^2).$$

Il est clair que l'EMV $\widetilde{\mathbf{w}} = \underset{\mathbf{w}}{\arg\min}\ \|\mathbf{Y}-\mathbb{X}\,\mathbf{w}\|^2 =: \widehat{\mathbf{w}} =: EMC$. L'EMV $\widetilde{\sigma}^2$, de σ^2, vaut

$$\widetilde{\sigma}^2 = \frac{1}{n}\left\|\mathbf{Y}-\mathbb{X}\,\widehat{\mathbf{w}}\right\|^2.$$

La log-vraisemblance du modèle (évaluée à l'EMV) est donnée par

$$\mathcal{L}_n(\widetilde{\mathbf{w}},\widetilde{\sigma}^2) = -\frac{n}{2}\log\frac{\|\widehat{\boldsymbol{\varepsilon}}\|^2}{n} - \frac{n}{2}(1+\log(2\pi)), \tag{2}$$

où

$$\widehat{\boldsymbol{\varepsilon}} := (\widehat{\varepsilon}_1,\ldots,\widehat{\varepsilon}_n)^\top := \mathbf{Y}-\mathbb{X}\,\widehat{\mathbf{w}} =: \mathbf{Y}-\widehat{\mathbf{Y}}$$

représentant le vecteur des résidus : les écarts entre les valeurs observées $\mathbf{Y} := (Y_1,\ldots,Y_n)^\top$, de la variable réponse Y, et les valeurs ajustées

$$\widehat{\mathbf{Y}} := \mathbb{X}\,\widehat{\mathbf{w}} =: (\widehat{Y}_1,\ldots,\widehat{Y}_n)^\top.$$

Propriétés des estimateurs MC

Si les conditions suivantes sont vérifiées

(i) Les erreurs, $\varepsilon_i, i=1,\ldots,n$, sont non corrélées;

(ii) $\mathbb{E}(\varepsilon_i \,|\, \mathbb{X}) = 0$, $\mathrm{Var}(\varepsilon_i \,|\, \mathbb{X}) = \sigma^2$, $i = 1, \ldots, n$, ne dépendant pas de $\mathbb{X}$ (hypothèse d'homoscédasticité);

alors on a les propriétés suivantes

(1) $\widehat{\mathbf{w}}$ est un estimateur sans biais de $\mathbf{w}$;

(2) La matrice de variance-covariance de $\widehat{\mathbf{w}}$, conditionnellement à $\mathbb{X}$, est donnée par

$$\mathrm{Var}(\widehat{\mathbf{w}} \,|\, \mathbb{X}) = \sigma^2 \, (\mathbb{X}^\top \, \mathbb{X})^{-1}.$$

Lois des estimateurs MC

Dans cette section, on suppose que les termes d'erreur, $\varepsilon_1, \ldots, \varepsilon_n$, sont i.i.d. (et indépendentes de $\mathbb{X}$) de même loi normale $\mathcal{N}(0, \sigma^2)$, de variance σ^2 ne dépendant pas de $\mathbf{x}$ (hypothèse d'homoscédasticité). Soit

$$\widehat{\boldsymbol{\varepsilon}} := \mathbf{Y} - \mathbb{X}\,\widehat{\mathbf{w}} =: \mathbf{Y} - \widehat{\mathbf{Y}}$$

le vecteur des résidus, et $\widehat{\sigma}^2$ l'estimateur, de σ^2, défini par

$$\widehat{\sigma}^2 := \frac{1}{n-p-1} \, \|\widehat{\boldsymbol{\varepsilon}}\|^2 .$$

Notons que les résulats de cette section restent valables (asymptotiquement, i.e., pour n suffisamment grand) si les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ sont i.i.d. centrées de variance σ^2 (ne dépendant pas de $\mathbf{x}$, homoscédasticité), même si la normalité n'est pas vérifiée. Les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ ne sont pas observables. Elles sont alors approchées par les résidus, pour tester par exemple les hypothèses d'homoscédasticité ou de non-corrélation des erreurs.

Proposition

On a

(1) $\widehat{\mathbf{w}}$ est un vecteur gaussien d'espérance $\mathbf{w}$ et de matrice de variance-covariance $\sigma^2 \left(\mathbb{X}^\top \mathbb{X}\right)^{-1}$;

(2) La statistique $(n-p-1)\,\frac{\widehat{\sigma}^2}{\sigma^2}$ suit la loi du $\chi^2_{(n-p-1)}$ (la loi du χ^2 à $n-p-1$ degrès de liberté);

(3) $\widehat{\mathbf{w}}$ et $\widehat{\sigma}^2$ sont indépendants.

Intervalles de confiance et tests

On note $\widehat{\sigma}_j^2 := \widehat{\sigma}^2 \left[\left(\mathbb{X}^\top \mathbb{X}\right)^{-1}\right]_{j,j}$, pour $j = 0, 1, \ldots, p$. On a

$$\forall j = 0, \ldots, p, \;\; \text{la statistique } \frac{\widehat{w}_j - w_j}{\widehat{\sigma}_j} \;\; \text{suit la loi } \mathcal{T}_{(n-p-1)},$$

(la loi de Student à $n-p-1$ degrès de liberté), ce qui permet de construire des intervalles de confiance, au niveau $1-\alpha$, pour les paramètres w_j, et de réaliser des tests d'hypothèses du type $\mathcal{H}_0 : w_j = 0$ contre $\mathcal{H}_1 : w_j \neq 0$.

Intervalle de confiance, au niveau $1-\alpha$, pour w_j

$$\left[\widehat{w}_j - t_{(n-p-1)}(1-\alpha/2)\,\widehat{\sigma}_j, \;\; \widehat{w}_j + t_{(n-p-1)}(1-\alpha/2)\,\widehat{\sigma}_j\right],$$

où $t_{(n-p-1)}(1-\alpha/2)$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à (n-p-1) degrés de liberté.

Test de Student

Considérons le problème de test de l'hypothèse nulle $\mathcal{H}_0 : w_j = 0$ contre l'alternative $\mathcal{H}_1 : w_j \neq 0$.

Notons $t := \frac{\widehat{w}_j}{\widehat{\sigma}_j}$ (la statistique de Student). La P-value du test (de Student) est donnée par

$$\text{P-value} = \mathbb{P}\left(|T| > |t_{obs}|\right),$$

où T est une variable aléatoire suivant la loi $\mathcal{T}_{(n-p-1)}$.

Prévision et intervalles de confiance

On dispose d'une nouvelle observation $\mathbb{X}_{n+1} := (1, X_{n+1,1}, \ldots, X_{n+1,p})$. On prédit la valeur Y_{n+1} correspondante par

$$\widehat{Y}_{n+1} := \mathbb{X}_{n+1}\, \widehat{\mathbf{w}}.$$

Intervalle d'estimation, au niveau $1-\alpha$, pour Y_{n+1} :

$$\left[\widehat{Y}_{n+1} - t_{(n-p-1)}(1-\alpha/2)\,\widehat{\sigma}\,\sqrt{\mathbb{X}_{n+1}\,\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}_{n+1}^\top},\ \widehat{Y}_{n+1} + t_{(n-p-1)}(1-\alpha/2)\,\widehat{\sigma}\,\sqrt{\mathbb{X}_{n+1}\,\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}_{n+1}^\top}\right];$$

Intervalle de prévision, au niveau $1-\alpha$, pour Y_{n+1} :

$$\left[\widehat{Y}_{n+1} - t_{(n-p-1)}(1-\alpha/2)\,\widehat{\sigma}\,\sqrt{1+\mathbb{X}_{n+1}\,\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}_{n+1}^\top},\ \widehat{Y}_{n+1} + t_{(n-p-1)}(1-\alpha/2)\,\widehat{\sigma}\,\sqrt{1+\mathbb{X}_{n+1}\,\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}_{n+1}^\top}\right].$$

Équation de l'analyse de variance

On a d'après le théorème de Pythagore

$$\begin{array}{rcl} \|\mathbf{Y} - \overline{\mathbf{Y}}\mathbf{1}\|^2 & = & \|\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\mathbf{1}\|^2 + \|\widehat{\boldsymbol{\varepsilon}}\|^2 \\ SST & = & SSR + SSE. \end{array}$$

(somme totale des carrés = somme des carrés de la régression + somme des carrés résiduels)

(total sum of squares = regression sum of squares + sum of squared errors)

Coefficient de détermination R^2

Le coefficient de détermination R^2 est défini par

$$R^2 := \frac{\|\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\mathbf{1}\|^2}{\|\mathbf{Y} - \overline{\mathbf{Y}}\mathbf{1}\|^2} =: \frac{SSR}{SST}.$$

Il vérifie les propriétés suivantes

(i) $0 \leq R^2 \leq 1$;

(ii) Si $R^2 = 1$, la variabilité de la variable réponse est entièrement expliquée par le modèle;

(iii) Si $R^2 = 0$, toute la variabilité se trouve dans le bruit (le terme d'erreur).

Test du modèle global

Le modèle de RLM s'écrit

$$Y_i = w_0 + w_1\, X_{i,1} + \cdots + w_p\, X_{i,p} + \varepsilon_i,\ \ i = 1, \ldots, n,$$

où les termes d'erreurs $\varepsilon_i, i = 1, \ldots, n$, sont supposés ici i.i.d. de même loi $\mathcal{N}(0, \sigma^2)$.

On veut tester

$$\mathcal{H}_0 : w_1 = \cdots = w_p = 0 \ \text{ contre } \ \mathcal{H}_1 : \exists j \in \{1, \ldots, p\} \text{ t.q. } w_j \neq 0.$$

Sous $\mathcal{H}_0$, la statistique de Fisher

$$F := \frac{R^2}{1 - R^2} \frac{n - p - 1}{p} \ \text{ suit la loi } \ \mathcal{F}_{(p, n-p-1)} \text{ (la loi de Fisher à } p \text{ et } n - p - 1 \text{ degrés de liberté).}$$

On rejette $\mathcal{H}_0$ si $F_{obs} > F_{(p,n-p-1)}(1 - \alpha)$, où $F_{(p,n-p-1)}(1 - \alpha)$ est le quantile d'ordre $(1 - \alpha)$ de la loi $\mathcal{F}_{(p,n-p-1)}$.

La P-value est donc donnée par

$$\text{P-value} = \mathbb{P}(Z > F_{obs}),$$

où Z est une variable aléatoire suivant la loi $\mathcal{F}_{(p,n-p-1)}$.

Tests entre modèles emboîtés

On veut tester le modèle réduit $\mathcal{M}_0$

$$Y_i = w_0 + w_{q+1}\, X_{i,q+1} + \cdots + w_p\, X_{i,p} + \varepsilon_i,$$

avec $1 \leq q < p$, à l'intérieur du modèle plus large $\mathcal{M}$

$$Y_i = w_0 + w_1\, X_{i,1} + \cdots + w_p\, X_{i,p} + \varepsilon_i.$$

Cela revient à tester (à l'intérieur du modèle $\mathcal{M}$) l'hypothèse nulle

$$\mathcal{H}_0 : w_1 = 0, \cdots, w_q = 0 \ \text{ contre } \ \mathcal{H}_1 : \exists j \in \{1, \ldots, q\} \text{ t.q. } w_j \neq 0.$$

Pour réaliser le test précédent, on utilise la statistique de Fisher suivante

$$F := \frac{\|\widehat{\mathbf{Y}}_0 - \widehat{\mathbf{Y}}\|^2 / q}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 / (n - p - 1)}$$

qui suit la loi $\mathcal{F}_{(q,n-p-1)}$, sous $\mathcal{H}_0$. ($\widehat{\mathbf{Y}}_0$ désigne le vecteur des valeurs ajustées du modèle réduit).

On rejette $\mathcal{H}_0$ si $F_{obs} > F_{(q,n-p-1)}(1 - \alpha)$.

La P-value est donc définie par

$$\text{P-value} := \mathbb{P}\left(Z > F_{obs}\right),$$

où Z est une variable aléatoire suivant la loi $\mathcal{F}_{(q,n-p-1)}$.

Ce test peut se faire sous R à l'aide de la fonction `anova()` qu'on applique à des modèles linéaires calculés avec la fonction `lm()`.

Exercice

(1) Construire le modèle de RLM de la variable `Sales` en fonction de toutes les variables de la base de données `Carseats` : utiliser la fonction `lm()`;

(2) Commenter les résultats;

(3) Vérifier graphiquement : (i) la non-corrélation des erreurs (utiliser la fonction `acf()`); (ii) la relation linéaire entre la variable réponse et les variables explicatives; (iii) l'hypothèse d'homoscédasticité des erreurs (appliquer la fonction `plot()` au modèle calculé par la fonction `lm()`);

(4) Tester l'hypothèse de non corrélation des erreurs (Utiliser le test de Durbin-Watson);

(5) Tester l'hypothèse d'homoscédasticité des erreurs: utiliser le test de Breusch-Pagan (fonction `bptest()` du package `lmtest`);

(6) Vérifier graphiquement l'hypothèse de normalité du terme d'erreur : utiliser la représentation Q-Q plot, et comparer l'histogramme des erreurs et la densité gaussienne;

(7) Tester l'hypothèse de normalité du terme d'erreur : utiliser le test de Shapiro-Wilk;

(8) Donner les résulats du test de Student de l'hypothèse nulle de non-significativité de chacune des variables explicatives. Ordonner les variables explicatives de la plus significative à la moins significative;

(9) Donner les résulats du test de Fisher de l'hypothèse nulle de non-significativité de chacune des variables explicatives. Ordonner les variables explicatives de la plus significative à la moins significative;

(10) Comparer les résultats des deux questions précédentes, et commenter;

(11) Parmi les deux tests précédents, lequel choisiriez-vous? Justifiez.