Analyse des données prostate: les moindres carrés ordinaires

julien.chiquet@genopole.cnrs.fr

Module MPR – option modélisation, 12 novembre 2009

1 Le modèle linéaire

Le modèle linéaire est défini par

$$y = \beta_0 + \sum_{i=1}^{p} X_i \beta_i + \varepsilon,$$

où y est la réponse à expliquer par le vecteur de variables d'entrées $X = (X_1, \ldots, X_p)$. Le scalaire β_0 (appelé biais ou *intercept*) et les $(\beta_i)_{i=1,\ldots,p}$ sont les paramètres à estimer. On suppose ici que les résidus ε sont gaussiens, c'est-à-dire que $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

On dispose de n données relatives aux variables y et X. Notons $\mathbf{y}$ le vecteur de taille n contenant les n observations de y, et $\mathbf{X}$ la matrice de taille $n \times (p+1)$ dont la $(i+1)^{\text{e}}$ colonne contient les n observations relatives à la variable X_i ; la première colonne de $\mathbf{X}$ est remplie de 1, ce qui permet de gérer l'intercept. Ainsi, le modèle linéaire s'écrit pour ce n–échantillon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\intercal$ est le vecteur des paramètres à estimer.

2 L'estimateur des moindres carrés

L'estimateur des moindres carrés ordinaires est défini comme le vecteur minimisant la somme des carrés résiduels :

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \mathrm{RSS}(\boldsymbol{\beta}), \qquad \mathrm{RSS}(\boldsymbol{\beta}) = \frac{1}{2} \left\| \mathbf{X}\boldsymbol{\beta} - \mathbf{y} \right\|_2^2 . \tag{1}$$

Le théorème suivant donne la solution de ce problème et les grandeurs caractéristiques standards associées à un estimateur (biais, variance).

Théorème. *Lorsque* $\mathbf{X}^\intercal\mathbf{X}$ *est inversible*[1]*, la solution* $\widehat{\boldsymbol{\beta}}$ *des équations normales,*

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}, \tag{2}$$

[1] c'est-à-dire, lorsque $n > p+1$: moralement, cela signifie qu'on a suffisamment de données pour estimer nos $p+1$ paramètres.

est unique et est appelée estimateur de Gauss-Markov. *C'est un estimateur sans biais de $\boldsymbol{\beta}$ et de matrice de variance-covariance*

$$\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}. \tag{3}$$

Démonstration. La fonction RSS est convexe : son minimum est atteint lorsque son gradient $\nabla_{\boldsymbol{\beta}}\mathrm{RSS}(\boldsymbol{\beta})$ s'annule. Ainsi

$$\nabla_{\boldsymbol{\beta}}\mathrm{RSS}(\boldsymbol{\beta}) = \mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \Leftrightarrow \mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\intercal\mathbf{y},$$

d'où la solution (2) pour $\mathbf{X}^\intercal\mathbf{X}$ inversible.

Comme[2] $\mathbb{E}(\mathbf{y}) = \mathbb{E}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta}$, on a

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}}\right) = \mathbb{E}\left((\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}\right) = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbb{E}\left(\mathbf{y}\right) = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

et $\widehat{\boldsymbol{\beta}}$ est donc sans biais. Concernant sa matrice de variance-covariance, on a[3]

$$\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \mathrm{Var}\left((\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}\right) = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathrm{Var}(\mathbf{y})\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}.$$

Or, $\mathrm{Var}(\mathbf{y}) = \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$, avec $\mathbf{I}_n$ la matrice identité de taille $n \times n$; d'où l'expression (3). □

2.1 Risque quadratique, erreur de prédiction

Définition (Risque quadratique)**.** *On appelle risque quadratique (*Mean Squared Error*) de l'estimateur $\widehat{\boldsymbol{\theta}}$ du paramètre $\boldsymbol{\theta}$ la grandeur définie par :*

$$\mathrm{MSE}(\widehat{\boldsymbol{\theta}}) = \mathbb{E}[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2] = \mathrm{biais}^2(\widehat{\boldsymbol{\theta}}) + \mathrm{Var}(\widehat{\boldsymbol{\theta}}). \tag{4}$$

Évidemment, le risque quadratique de $\widehat{\boldsymbol{\beta}}$ se ramène simplement à sa variance puisqu'il est sans biais.

Définition (Erreur de prédiction)**.** *On appelle erreur de prédiction du modèle de régression*

$$y = f(X) + \varepsilon$$

la grandeur définie par[4]

$$\mathrm{EPE}(f) = \mathbb{E}[(y - f(X))^2]. \tag{5}$$

Dans le cas du modèle linéaire, l'estimateur de Gauss-Markov prévoit pour $\mathbf{y}$ le vecteur de valeurs $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$. Ainsi, nous avons

$$\mathrm{EPE}(X\widehat{\boldsymbol{\beta}}) = \mathbb{E}[(y - X\widehat{\boldsymbol{\beta}})^2] = \mathbb{E}(y - X\widehat{\boldsymbol{\beta}})^2 + \mathrm{Var}(y - X\widehat{\boldsymbol{\beta}}) = \sigma^2 + \mathrm{MSE}(X\widehat{\boldsymbol{\beta}}).$$

[2] Rappel : $\mathbb{E}(AX) = A\mathbb{E}(X)$ pour A une matrice déterministe et X un vecteur aléatoire
[3] Rappel :$\mathrm{Var}(AX) = A\mathrm{Var}(X)A^\intercal$ pour A une matrice déterministe et X un vecteur aléatoire
[4] EPE pour *Expected Prediction Error*

Ensemble de test, ensemble d'apprentissage. Pour évaluer les performance d'un modèle, il est classique lorsque l'on dispose de suffisamment de données de les séparer en un ensemble d'*apprentissage*, qui permet de calculer les paramètres du modèle (ici $\widehat{\boldsymbol{\beta}}$), et un ensemble de *test*, qui permet d'évaluer la qualité du modèle. Sans faire cette procédure, on peut montrer que l'on sous-estime l'erreur de prédiction.

Estimation du risque quadratique. Le risque quadratique dépend de la variance de $\widehat{\boldsymbol{\beta}}$: il faudrait connaître la valeur σ^2 pour le calculer. Si l'on note $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, les valeurs prédites par le modèle pour $\mathbf{y}$, on peut estimer le paramètre σ^2 par la variance empirique corrigée des résidus empiriques $\mathbf{y} - \widehat{\mathbf{y}}$, c'est-à-dire

$$\widehat{\sigma}^2 = \frac{1}{n-p-1} \|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2 . \tag{6}$$

Ce calcul se fait sur les données de l'ensemble d'apprentissage, puisque ce sont celles qui ont permis de déterminer $\widehat{\boldsymbol{\beta}}$.

Estimation de l'erreur de prédiction. L'erreur de prédiction s'estime par exemple par

$$\widehat{\mathrm{EPE}} = \frac{1}{n^{\mathrm{test}}} \sum_{i=1}^{n^{\mathrm{test}}} \left(\widehat{\mathbf{y}}_{\mathrm{test}} - \mathbf{X}_{\mathrm{test}} \widehat{\boldsymbol{\beta}} \right)^2 . \tag{7}$$

Nous verrons plus tard d'autres manière d'estimer cette erreur, en particulier par validation croisée.

2.2 Le Z–score : une première méthode de sélection de variables

Sous l'hypothèse de normalité des résidus ε, le vecteur $\widehat{\boldsymbol{\beta}}$ suit évidemment une loi normale multivariée donc on connaît espérance et variance. Ainsi,

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(0, \sigma^2 \mathbf{X}^\intercal \mathbf{X}).$$

De même,

$$(n-p-1)\widehat{\sigma}^2 \sim \sigma^2 \chi^2 (n-p-1),$$

puisque $\widehat{\sigma}^2$ est définie comme la somme du carré des résidus (qui sont gaussiens). On peut donc construire le test d'hypothèse suivant sur la nullité des composants de $\widehat{\boldsymbol{\beta}}$:

$$\begin{cases} H_0 : & \widehat{\beta}_i = 0, \\ H_1 : & \widehat{\beta}_i \neq 0. \end{cases}$$

Le Z–score, défini par $z_i = \widehat{\beta}_i / \widehat{\sigma} \sqrt{\mathrm{diag}(\mathbf{X}^\intercal \mathbf{X})}$, suit une loi de Student à $n-p-1$ degrés de liberté sous H_0, puisque défini comme le rapport d'une loi normale et d'un χ^2.

Ainsi, on rejette H_0 au niveau de confiance α lorsque $|z_i| > t_{1-\alpha/2;n-p-1}$, le fractile d'ordre d'une loi de Student à $n-p-1$ degrés de liberté.

3 Commande R commentée

3.1 Préparation des données

On charge les données prostate, qui contiennent trois variables : un vecteur `x`, un vecteur `y` et un vecteur de booléens `set` indiquant les données de l'ensemble d'apprentissage.

```
> load("prostate.rda")
```

La manipulations des matrices (objet `matrix`) est plus simple que celle des tableaux de données (objet `data.frame`). Transformons donc `x` en matrice à l'aide de la fonction `as.matrix` :

```
> x <- as.matrix(x)
```

Pour retrouver les résultats du livre [ESLII], page 50, nous devons commencer par centrer et réduire les données[5] :

```
> x <- scale(x)
```

Ensuite, nous construisons les données test (vecteurs `x.test` et `y.test`) et les donnés d'apprentissage (vecteurs `x` et `y`).

```
> test <- !set
> n <- sum(set)
> n.test <- sum(test)
> x.test <- x[test, ]
> y.test <- y[test]
> x <- x[set, ]
> y <- y[set]
```

Concaténons un vecteur $\mathbf{1}_n$ aux matrices $\mathbf{X}$ et $\mathbf{X}_{\text{test}}$ afin afin d'inclure l'intercept dans la procédure d'estimation des paramètres :

```
> p <- ncol(x)
> x <- cbind(rep(1, n), x)
> x.test <- cbind(rep(1, n.test), x.test)
```

3.2 Estimation

Calculons l'estimateur du vecteur des paramètres (notez l'utilisation de `solve` : pour inverser une matrice A, on résout $AX = I$) :

```
> beta.ols <- solve(t(x) %*% x, t(x) %*% y)
```

[5] Théoriquement, on devrait faire l'opération de centrage/réduction *après* avoir construit les ensembles de test et d'apprentissage – tenons-nous en à l'approche du livre, afin de retrouver exactement les mêmes valeurs

Vient ensuite l'estimation de σ^2, de la variance de $\widehat{\boldsymbol{\beta}}$ et de l'erreur de prédiction.

```
> sigma.hat <- sqrt(1/(n - p - 1) * sum((y - x %*% beta.ols)^2))
> var.beta.ols <- solve(t(x) %*% x) * sigma.hat^2
> std.error <- sqrt(diag(var.beta.ols))
> epe <- 1/n.test * sum((y.test - x.test %*% beta.ols)^2)
```

On teste au niveau $\alpha = 5\%$ la nullité des $\widehat{\beta}_i$ et on stocke dans un vecteur `rejet` les résultats (1 en cas de rejet de H_0).

```
> Z.score <- beta.ols/std.error
> rejet <- rep(FALSE, p + 1)
> alpha <- 0.05
> rejet[which(abs(Z.score) > qt(1 - alpha/2, df = n + p + 1))] <- TRUE
```

3.3 Résultats

On stocke tout les résultats dans une matrice `results` en y nommant les colonnes de manières appropriée.

```
> results <- cbind(beta.ols, std.error, Z.score, rejet)
> colnames(results) <- c("beta.hat", "std.error", "Z.score", "H1 à 5%")
```

Voici pour appeler l'affichage des résultats :

```
> cat("\n Modèle linéaire sur données prostate\n\n")
> print(results)
> cat("\nValeur estimée de l'erreur de prédiction:", epe, "\n")
```

Ce qui donne à l'écran :

```
 Modèle linéaire sur données prostate

           beta.hat  std.error    Z.score H1 à 5%
          2.46493292 0.08931498 27.5982031       1
lcavol    0.67952814 0.12662903  5.3662905       1
lweight   0.26305307 0.09562821  2.7507894       1
age      -0.14146483 0.10134245 -1.3959090       0
lbph      0.21014656 0.10221904  2.0558456       1
svi       0.30520060 0.12360027  2.4692552       1
lcp      -0.28849277 0.15452934 -1.8669126       0
gleason -0.02130504 0.14524723 -0.1466812       0
pgg45     0.26695576 0.15361357  1.7378397       0

Valeur estimée de l'erreur de prédiction: 0.521274
```

Références

[ESLII] Hastie, T., Tibshirani, R. and Friedman, J., *The Elements of Statistical Learning*, second edition, Springer, 2008.