MAP433 Statistique

PC 5: Régression

25 septembre 2015

1 Modèle de régression multiple

On considère le modèle de regression multiple

$$\mathbf{Y} = \beta_0 \mathbf{e} + \mathbf{X}\beta + \sigma\xi, \quad \text{où } \mathbb{E}[\xi] = 0,\ \mathbb{E}[\xi\xi^T] = I_n,\ \mathbf{e} = (1, 1, \dots, 1)^T$$

avec $\mathbf{X}$ une matrice $n \times k$ de rang k et $\mathbf{Y}$, ξ des vecteurs de $\mathbb{R}^n$. Les paramètres $\beta_0 \in \mathbb{R}$ et $\beta \in \mathbb{R}^k$ sont inconnus. On note $\hat{\beta}_0$ et $\hat{\beta}$ les estimateurs des moindres carrés de β_0 et β.

1. On note $\widehat{\mathbf{Y}} = \hat{\beta}_0 \mathbf{e} + \mathbf{X}\hat{\beta}$ $\quad \overline{Y} = n^{-1}\mathbf{e}^T\mathbf{Y}$ et $\overline{\widehat{\mathbf{Y}}} = n^{-1}\mathbf{e}^T\widehat{\mathbf{Y}}$. Montrer que $\overline{\widehat{\mathbf{Y}}} = \overline{Y}$. En déduire que $\overline{Y} = \hat{\beta}_0 + \left(n^{-1}\mathbf{e}^T\mathbf{X}\right)\hat{\beta}$.
2. Montrer l'équation d'analyse de la variance :
$$\|\mathbf{Y} - \overline{Y}\mathbf{e}\|^2 = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 + \|\widehat{\mathbf{Y}} - \overline{Y}\mathbf{e}\|^2.$$
En déduire que le *coefficient de détermination*
$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \qquad \text{où } \mathbf{Y} = \begin{bmatrix} Y_1 \\ \cdots \\ Y_n \end{bmatrix} \qquad \widehat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \cdots \\ \hat{Y}_n \end{bmatrix}$$
est toujours inférieur à 1.
3. Supposons que $\mathbf{Z} = [\mathbf{e}, \mathbf{X}]$ est de rang $k+1$. Calculez en fonction de $\mathbf{Z}$ la matrice de covariance de $(\hat{\beta}_0, \hat{\beta})$. Comment accède-t-on à $\mathrm{Var}(\hat{\beta}_j)$, pour $j = 0, \dots, k$?
4. Proposer un estimateur sans biais de σ^2 puis de la matrice de covariance de $(\hat{\beta}_0, \hat{\beta})$.
5. On suppose dorénavant que $\beta_0 = 0$ et donc $\mathbf{Y} = \mathbf{X}\beta + \sigma\xi$ avec $\mathbb{E}[\xi] = 0$ et $\mathbb{E}[\xi\xi^T] = I_n$. L'estimateur des moindres carrés $\tilde{\beta}$ dans ce modèle est-il égal à $\hat{\beta}$?
6. A-t-on la relation $\overline{\widehat{\mathbf{Y}}} = \overline{Y}$? Que dire du R^2 dans ce modèle ?

2 Le modèle ANOVA

On dispose d'observations de variables aléatoires

$$Y_{ij} = m_i + \xi_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, l,$$

où $\mathbf{m} = (m_1, \dots, m_k)^T \in \mathbb{R}^k$ et les ξ_{ij} sont des variables aléatoires i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

1. Montrer qu'il s'agit d'un modèle de régression linéaire avec la matrice $\mathbf{X}$ que l'on précisera. Que vaut $B = \mathbf{X}^T\mathbf{X}$?
2. Montrer que la condition $m_1 = m_2 = \cdots = m_k$ s'écrit sous la forme $\mathbf{G}\mathbf{m} = 0$ avec une matrice $\mathbf{G}$ que l'on précisera.
3. On estime $\mathbf{m}$ par l'estimateur des moindre carrés $\widehat{\mathbf{m}}$. Quelle est la covariance de $\widehat{\mathbf{m}}$?
4. Proposer un estimateur de $\mathbf{G}\mathbf{m}$. Quel est son biais ? sa covariance ?
5. Proposer un estimateur $\hat{\sigma}^2$ de σ^2. Quelle est sa distribution ?

3 Théorème de Gauss-Markov

On considère le modèle de régression $\underset{(n,1)}{\mathbf{Y}} = \underset{(n,k)}{\mathbf{X}}\underset{(k,1)}{\beta} + \sigma \underset{(n,1)}{\xi}$. On suppose que $\mathbf{X}$ est une matrice déterministe, $\mathbb{E}[\xi] = 0$, $\mathbb{E}[\xi\xi^T] = I_n$, $\mathrm{Rang}(\mathbf{X}) = k$. On note $\hat{\beta}$ l'estimateur des MC de β.

1. Montrer que $\hat{\beta}$ est sans biais et expliciter sa matrice de covariance.
2. Soit $\tilde{\beta}$ un estimateur de β linéaire en $\mathbf{Y}$, i.e., $\tilde{\beta} = \mathbf{L}\mathbf{Y}$ pour une matrice $\mathbf{L} \in \mathbb{R}^{k\times n}$ déterministe. Donner une condition nécessaire et suffisante sur $\mathbf{L}$ pour que $\tilde{\beta}$ soit sans biais. On supposera maintenant cette hypothèse vérifiée.
3. Calculer la matrice de covariance de $\tilde{\beta}$. En posant $\Delta = \mathbf{L} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ montrer que $\Delta\mathbf{X} = 0$ et $\mathrm{cov}(\tilde{\beta}) = \mathrm{cov}(\hat{\beta}) + \sigma^2\Delta\Delta^T$. En déduire que
$$\mathbb{E}[(\tilde{\beta}-\beta)(\tilde{\beta}-\beta)^T] \geq \mathbb{E}[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T] \quad \text{(inégalité au sens matriciel).}$$
4. En passant aux risques quadratiques $\mathbb{E}\big[\|\tilde{\beta}-\beta\|^2\big]$ et $\mathbb{E}\big[\|\hat{\beta}-\beta\|^2\big]$, en déduire que l'estimateur des MC est optimal dans la classe de tous les estimateurs linéaires sans biais.

4 Régression Ridge

On considère le modèle de régression $\underset{(n,1)}{\mathbf{Y}} = \underset{(n,k)}{\mathbf{X}}\underset{(k,1)}{\beta} + \sigma \underset{(n,1)}{\xi}$. On suppose que $\mathbf{X}$ est une matrice déterministe, $\mathbb{E}[\xi] = 0$, $\mathbb{E}[\xi\xi^T] = I_n$.

1. On suppose que $k > n$. Que dire de l'estimation par moindres carrés ?
2. On appelle estimateur `Ridge regression` de paramètre de régularisation $\lambda > 0$ l'estimateur
$$\hat{\beta}_\lambda = \arg\min_{\beta\in\mathbb{R}^k} \left\{\|\mathbf{Y}-\mathbf{X}\beta\|^2 + \lambda\|\beta\|^2\right\}.$$
Exprimez $\hat{\beta}_\lambda$ en fonction de $\mathbf{X}$, $\mathbf{Y}$ et λ. Cet estimateur est-il défini pour $k > n$?
3. Calculez la moyenne et la matrice de covariance de l'estimateur Ridge. Est-il sans biais ?
4. On suppose maintenant que $k = 1$, ce qui correspond au modèle de régression simple. Montrer qu'il existe une valeur de λ telle que, pour certaines valeurs de β, le risque $\mathbb{E}\big[(\hat{\beta}_\lambda-\beta)^2\big]$ de l'estimateur Ridge de paramètre λ est inférieur au risque $\mathbb{E}\big[(\hat{\beta}_0-\beta)^2\big]$ de l'estimateur des MC.