ÉCOLE POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

MAP433 Statistique

PC 5 et 6: Régression

25 septembre et 2 octobre 2015

1 Modèle de régression multiple

On considère le modèle de regression multiple

$$y = \theta_0 e + X\theta + \xi, \quad \text{où} \;\; \mathbb{E}[\xi] = 0, \; \mathbb{E}[\xi\xi^T] = \sigma^2 I_n, \; e = (1, 1, \ldots, 1)^T$$

avec X une matrice $n \times k$ de rang k et y, ξ des vecteurs de $\mathbb{R}^n$. Les paramètres $\theta_0 \in \mathbb{R}$ et $\theta \in \mathbb{R}^k$ sont inconnus. On note $\hat{\theta}_0$ et $\hat{\theta}$ les estimateurs des moindres carrés de θ_0 et θ.

1. On note $\hat{y} = \hat{\theta}_0 e + X\hat{\theta}$. Montrer que $\bar{\hat{y}} = \bar{y}$, où $\bar{y}$ (resp. $\bar{\hat{y}}$) est la moyenne des y_i (resp. des $\hat{y}_i$). En déduire que $\bar{y} = \hat{\theta}_0 + \bar{X}\hat{\theta}$ où $\bar{X} = \frac{1}{n} e^T X = \left[\ldots, \bar{X}_{\cdot j}, \ldots\right]$.
2. Montrer l'équation d'analyse de la variance :

   $$\|y - \bar{y}e\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - \bar{y}e\|^2.$$

   En déduire que le *coefficient de détermination*

   $$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

   est toujours inférieur à 1.
3. Supposons que $Z = [e, X]$ est de rang $k+1$. Calculez en fonction de Z la matrice de covariance de $(\hat{\theta}_0, \hat{\theta})$. Comment accède-t-on à $\mathrm{Var}(\hat{\theta}_j)$, pour $j = 0, \ldots, p$?
4. Proposer un estimateur sans biais de σ^2 puis de la matrice de covariance de $(\hat{\theta}_0, \hat{\theta})$.
5. On suppose dorénavant que $\theta_0 = 0$ et donc

   $$y = X\theta + \xi, \quad \mathbb{E}[\xi] = 0, \; \mathbb{E}[\xi\xi^T] = \sigma^2 I_n.$$

   L'estimateur des moindres carrés $\tilde{\theta}$ dans ce modèle est-il égal à $\hat{\theta}$?
6. A-t-on la relation $\bar{\hat{y}} = \bar{y}$? Que dire du R^2 dans ce modèle ?

2 Le modèle ANOVA

On dispose d'observations de variables aléatoires

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

où $(m_1, \ldots, m_k) \in \mathbb{R}^k$ 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 $Gm = 0$ avec une matrice G que l'on précisera.
3. On estime m par l'estimateur des moindre carrés $\hat{m}$. Quelle est la covariance de $\hat{m}$?
4. Proposer un estimateur de Gm. 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

$$\underset{(n,1)}{Y} = \underset{(n,k)(k,1)}{X}\,\underset{}{\theta} + \underset{(n,1)}{\xi}\,.$$

On suppose que X est une matrice déterministe, $\mathbb{E}[\xi] = 0$, $\mathbb{E}[\xi\xi^T] = \sigma^2 I_n$, $\mathrm{Rang}(X) = k$. On note $\hat{\theta}$ l'estimateur des MC de θ.

1. Montrer que $\hat{\theta}$ est sans biais et expliciter sa matrice de covariance.
2. Soit $\tilde{\theta}$ un estimateur de θ linéaire en Y, i.e., $\tilde{\theta} = LY$ pour une matrice $L \in \mathbb{R}^{k\times n}$ déterministe. Donner une condition nécessaire et suffisante sur L pour que $\tilde{\theta}$ soit sans biais. On supposera maintenant cette hypothèse vérifiée.
3. Calculer la matrice de covariance de $\tilde{\theta}$. En posant $\Delta = L - (X^TX)^{-1}X^T$ montrer que $\Delta X = 0$ et $\mathrm{cov}(\tilde{\theta}) = \mathrm{cov}(\hat{\theta}) + \sigma^2\Delta\Delta^T$. En déduire que
$$\mathbb{E}[(\tilde{\theta}-\theta)(\tilde{\theta}-\theta)^T] \geq \mathbb{E}[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T] \quad \text{(inégalité au sens matriciel).}$$
4. En passant au risques quadratiques $\mathbb{E}\big[\|\tilde{\theta}-\theta\|^2\big]$ et $\mathbb{E}\big[\|\hat{\theta}-\theta\|^2\big]$, 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

$$\underset{(n,1)}{Y} = \underset{(n,k)(k,1)}{X}\,\underset{}{\theta} + \underset{(n,1)}{\xi}\,.$$

On suppose que X est une matrice déterministe, $\mathbb{E}[\xi] = 0$, $\mathbb{E}[\xi\xi^T] = \sigma^2 I_n$.

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
$$\hat{\theta}_\lambda = \arg\min_{\theta\in\mathbb{R}^k} \left\{\|Y - X\theta\|^2 + \lambda\|\theta\|^2\right\}.$$
Exprimez $\hat{\theta}_\lambda$ en fonction de X, 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 θ, le risque $\mathbb{E}\big[(\hat{\theta}_\lambda - \theta)^2\big]$ de l'estimateur Ridge de paramètre λ est inférieur au risque $\mathbb{E}\big[(\hat{\theta}_0 - \theta)^2\big]$ de l'estimateur des MC.

5 Analyse de données atmosphériques

Nous allons analyser des relevés atmosphériques effectués par l'association"Air Breizh". Ces relevés se présentent sous la forme d'un tableau dont chaque ligne donne les mesures de l'ozone du jour (O3), de la température à 12h (T12) et 15h (T15), d'un indice de nébulosité à 12h (Ne12), des relevés de vents à 12h (N12, S12, E12, W12), d'un indice du vent moyen (Vx) et de la concentration en ozone de la veille (03v). Notre objectif sera de trouver parmi les facteurs précédents ceux qui sont influents sur la quantité d'ozone (O3) présente dans la basse atmosphère.

Les analyses seront réalisées avec `R` : c'est un logiciel gratuit et très largement utilisé par les statisticiens car la plupart des méthodes statistiques (anciennes et nouvelles) ont été implémentées dans ce langage. Il est téléchargeable sur : `http://cran.r-project.org/`. Pour apprendre à s'en servir : `http://cran.r-project.org/doc/manuals/R-intro.pdf`. Vous pouvez aussi consulter l'ouvrage *Régression avec R* de Cornillon & Matzner-Lober. La syntaxe est proche de scilab et matlab.

Pour commencer, téléchargez les données sur la page `http://www.cmap.polytechnique.fr/~giraud/MAP433/ozone.Rdata`. Lancez `R`, puis chargez les données dans `R` avec la commande `load("ozone.Rdata")`. Nous effectuerons une régression linéaire à l'aide de la fonction `lm`. Par exemple

```
reg = lm(O3~T12+Vx, data=ozone)
```

réalise la régression de O3 par rapport aux variables T12 et Vx. Tapez `?lm` pour avoir une description de cette fonction. Si `reg` est le résultat d'une régression de $Y \in \mathbb{R}^n$ contre $X \in \mathbb{R}^{n\times k}$, l'instruction `summary(reg)` retourne un tableau de valeurs dont la première colonne donne l'estimateur

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^k} \|Y - X\theta\|^2$$

et l'avant dernière colonne donne les t-values

$$\hat{t}_j = \frac{\widehat{\theta}_j}{\sqrt{\hat{\sigma}^2[(X^TX)^{-1}]_{jj}}} \quad \text{où} \quad \widehat{\sigma}^2 = \frac{1}{n-k}\|Y - X\widehat{\theta}\|^2.$$

La dernière colonne donne les p-values $\hat{p}_j = \mathcal{T}_{n-k}(\hat{t}_j)$ où $\mathcal{T}_{n-k}(t) = \mathbb{P}(|T_{n-k}| > |t|)$ avec T_{n-k} une variable de Student à $n-k$ degrés de liberté.

A) Inspection des résidus

1. Calculer avec la fonction `lm` la régression de O3 par rapport aux autres variables. Identifier Y et X dans ce cas. Que vaut n ? Que vaut k ?
2. On note $\hat{\xi} = Y - \hat{Y}$ où $\hat{Y} = X\widehat{\theta}$. Tracer l'histogramme des $\{\hat{\xi}_i : i = 1, \dots, n\}$ à l'aide de la fonction `hist`.
3. L'histogramme suggère que les résidus pourraient suivre une loi Gaussienne. On va inspecter cette hypothèse en regardant les quantiles de la loi empirique. On note $x_q(Q) = \min\{x : Q(]-\infty, x]) \geq q\}$ le quantile d'ordre q d'une loi Q. Tracer le QQplot

   ```
   qqnorm(lm(O3~.,data=ozone)$residuals)
   ```

Que représente ce graphique ?

4. Pour savoir si la variance dépend du signal tracer les points $\left\{(\hat{Y}_i, |\hat{\xi}_i|) : i = 1, \ldots, n\right\}$ à l'aide de la fonction `plot`. Effectuer la régression des $|\hat{\xi}_i|$ en fonction des $\hat{Y}_i$.

B) Choix des variables

1. Quelles variables j ont une p-value $\hat{p}_j$ inférieure à 5% ?
2. Calculer la régression de O3 par rapport à Ne12+O3v et inspecter les résidus comme précédemment.
3. Calculer la régression de O3 par rapport à Ne12+O3v+T15+Vx. Que constatez-vous au niveau des p-values $\hat{p}_j$?

C) Régressions partielles

Dorénavant on ne travaille qu'avec les variables Ne12, O3v, T15 et Vx. On veut inspecter les questions suivantes :

— le modèle linéaire par rapport à la variable j est-il raisonnable ?
— quelle est l'influence de la variable j ?

1. Montrer que si le modèle $Y = \sum_k \theta_k X_k + \xi$ est vrai, alors :
   lm($Y \sim -X_j$)\$residuals = $\theta_j \times$ lm($X_j \sim -X_j$)\$residuals + lm($\xi \sim -X_j$)\$residuals
   où lm($Z \sim -X_j$) représente la régression de Z par rapport à toutes les variables $X_1, \ldots, X_k$ sauf X_j.
2. Si les $\xi_1, \ldots, \xi_n$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$, quelle est la loi du vecteur lm($\xi \sim -X_j$)\$residuals ?
3. Calculer la régression de lm($Y \sim -X_j$)\$residuals par lm($X_j \sim -X_j$)\$residuals pour j =Ne12. Le modèle linéaire semble-t-il raisonnable pour cette variable ?
4. Même question avec la variable j =T15.