Régression linéaire (suite)

On pourra reprendre la première partie de ce TP (*cf.* le fichier `TP_linear_regression.pdf`) sachant que des élélements de correction sont disponibles sur le site http://josephsalmon.eu/index.php?page=teaching_13_14&lang=fr sous forme de fichiers R. Au cours de cette séance, on mettra l'accent sur les questions d'implémentation et les questions théoriques pourront être mises de côté en première lecture.

- Régression linéaire simple -

Reprendre les données du fichier http://www.math.uah.edu/stat/data/Galton.txt.

1. Redonner les estimations de θ_0, θ_1, par $\hat{\theta}_0$, $\hat{\theta}_1$ et proposer un estimateur $\hat{\sigma}^2$ de σ^2.

- Régression linéaire multiple -

On travaille de nouveau sur le fichier `auto-mpg.data` et l'on reprend la régression de la consommation des voitures sur les variables considérées au TP précèdent.

2. Calculer $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{y}}$ et $\hat{\sigma}^2$ sur l'intégralité des données.
3. Afficher un histogramme des résidus, et faire apparaître la densité de loi gaussienne convenablement normalisée.
4. Calculer de nouveau $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{y}}$ et $\hat{\sigma}^2$ mais cette fois sur sur les données centrées réduites.

- Intervalles de confiance (IC) -

On s'intéresse ici à construire un intervalle de confiance de la moyenne dans un modèle gaussien. Ainsi, on observe un n-échantillon, distribué suivant la loi gaussienne $\mathcal{N}(\mu, \sigma^2)$. Notre but est de déterminer un IC de niveau 95% pour μ et d'étudier ses propriétés. On sait que μ est bien estimé par la moyenne empirique $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$.

On suppose dans cette partie que σ^2 est connu.

1. Montrer que la variable aléatoire $v_n = \frac{(\bar{x}_n - \mu)\sqrt{n}}{\sigma}$ suit une loi gaussienne centrée réduite.

On cherche à avoir une confirmation empirique de ce résultat. Pour cela, on génère N échantillons de taille n et de loi $\mathcal{N}(\mu, \sigma^2)$, pour des valeurs de μ et de σ^2 données. Pour chacun de ces N échantillons, on calcule la valeur de v_n. Cela nous donne une série numérique $v_1^N, \ldots, v_n^N$.

```
CheckGauss=function(N,n,mu,sigma)
{
X=matrix(rnorm(N*n,mean=mu,sd=sigma),N,n)
v=sqrt(n)*(apply(X,1,mean)-mu)/sigma
return(v)
}
```

2. Si cette série est vraiment distribuée suivant une loi normale centrée réduite, alors son histogramme doit être proche de la densité de la loi $\mathcal{N}(0,1)$. Exécuter les commandes suivantes pour le verifier :

```
N=5000; n=30;
v=CheckGauss(N,n,2,3)
par(bg="cornsilk",lwd=2,col="darkblue")
hist(v,breaks=30,freq=F,col="cyan")
curve(dnorm(x),add=T,lwd=2)
```

3. Déterminer les quantiles d'ordre 2.5% et 97.5% de la série numérique $v_1^N, \ldots, v_n^N$ (on lira l'aide en ligne de la commande `quantile`).
4. Faire varier les valeurs de μ et de σ^2. Cela influence t-il le résultat ?
5. Augmenter N jusqu'à 40.000 (si les capacités de la machine le permettent) et refaire l'expérience. Quelles sont les valeurs obtenus pour les deux quantiles ?
6. Soient a et b les valeurs obtenues dans la question précédente. Si Z est une v.a. distribuée suivant la loi $\mathcal{N}(0,1)$, quelle est la probabilité $\mathbb{P}(Z \in [a,b])$?
7. Vérifier par calcul que $v_n \in [a,b]$ équivaut à

$$\mu \in [\bar{x}_n - \frac{b\sigma}{\sqrt{n}}, \bar{x}_n + \frac{a\sigma}{\sqrt{n}}] \tag{1}$$

   puis comparer ce résultat avec le quantile théorique de la densité d'une gaussienne centrée réduite (on lira l'aide en ligne de la commande `qnorm`).
8. Quel est le pourcentage des échantillons (parmi les N échantillons générés) pour lesquels la relation (1) est satisfaite ?

On suppose dans cette partie que σ^2 est inconnu.

On utilise alors les rappels suivants. Tout d'abord, σ^2 est bien estimé par la variance empirique sans biais $s_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$. De plus, la théorie assure que la variable aléatoire $t_n = \frac{(\bar{x}_n - \mu)\sqrt{n}}{s_n}$ suit la loi de Student à $n-1$ degrés de liberté (*cf.* [1, Théorème 5.2]). De nouveau, on cherche à avoir une confirmation empirique de ce résultat. Pour cela, on génère N échantillons de taille n de loi $\mathcal{N}(\mu, \sigma^2)$, pour des valeurs de μ et de σ^2 données. Pour chacun de ces N échantillons, on calcule la valeur de t_n. Cela nous donne une série numérique $t_1^N, \ldots, t_n^N$.

```
CheckStudent=function(N,n,mu,sigma)
{
X=matrix(rnorm(N*n,mean=mu,sd=sigma),N,n)
t=sqrt(n)*(apply(X,1,mean)-mu)/apply(X,1,sd)
return(t)
}
```

9. Si cette série est vraiment distribuée suivant la loi de Student t_{n-1}, alors son histogramme doit être proche de la densité de la loi t_{n-1}. Pour vérifier cela, on exécute les commandes :

```
N=5000; n=30;
t=CheckStudent(N,n,2,3)
par(bg="cornsilk",lwd=2,col="darkblue")
hist(t,breaks=30,freq=F,col="cyan")
curve(dt(x,n-1),add=T,lwd=2)
```

10. Déterminer les quantiles d'ordre 2.5% et 97.5% de la série numérique $t_1^N, \ldots, t_n^N$.
11. Faire varier les valeurs de μ et de σ^2. Cela influence t-il le résultat ?
12. Augmenter N jusqu'à 40.000 (si les capacités de la machine le permettent) et refaire l'expérience. Quelles sont les valeurs obtenus pour les deux quantiles ?
13. Soient α et β les valeurs obtenues dans la question précédente. Si T est une v.a. distribuée suivant la loi t_{n-1}, quelle est la probabilité $\mathbb{P}(T \in [\alpha, \beta])$?
14. Vérifier par calcul que $T_n \in [\alpha, \beta]$ équivaut à

$$\mu \in [\bar{x}_n - \frac{\beta s_n}{\sqrt{n}}, \bar{x}_n + \frac{\alpha s_n}{\sqrt{n}}] \tag{2}$$

15. Quel est le pourcentage des échantillons (parmi les N échantillons générés) pour lesquels la relation (2) est satisfaite ?

Bonus

Revenons dans le cadre du modèle de régression linéaire multiple.

16. Rappelez quelle est la loi suivie par
$$\frac{\widehat{\theta}_j - \theta_j}{\sqrt{\hat{\sigma}^2 \, (X^\top X)^{-1}_{jj}}}$$
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 $p-n-1$ degré de liberté (on suspose la matrice X de plein rang).

- Classification binaire par moindre carrés -

On considère dans cette partie deux populations gaussiennes dans $\mathbb{R}^p$ ayant la **même** structure de covariance. On observe ensuite des points générer par un mélange de ces deux populations.

Les lois conditionnelles de X sachant $Y=+1$ (respectivement $Y=-1$) sont des gaussiennes multivariées $\mathcal{N}_p(\mu_+, \Sigma)$ (respectivement $\mathcal{N}_p(\mu_-, \Sigma)$). On notera leur densités respectives f_+ et f_-. Les vecteurs μ_+ et μ_- sont dans $\mathbb{R}^p$ et la matrice Σ est (symétrique) de taille $p \times p$. On note également $\pi_+ = \mathbb{P}\{Y=+1\}$. On tire donc avec probabilité π_+ une étiquette $Y=+1$ ou $Y=-1$, qui indique si X est tiré selon la loi f_+ ou f_-. On rappelle que la densité p-dimensionnelle de la loi $\mathcal{N}_p(\mu, \Sigma)$ est donnée par :

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(\Sigma)}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right\} \, .$$

et que la matrice de covariance d'un vecteur aléatoire X est définie par $\Sigma = \mathbb{E}\big[(X - \mathbb{E}X)(X-\mathbb{E}X)^\top\big]$.

1. Générer un jeu de données simulées selon le modèle de mélange précèdent, pour lequel on observe autant de variables dans la première classe que dans la seconde. On prendra comme valeurs numériques : $\pi_+ = 0.5, p = 2, n = 500, \mu_- = (-1,-1), \mu_+ = (1,1), \Sigma = 3\,\mathrm{Id}_p$,
2. On suppose que l'échantillon considéré contient n observations notées $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ et que $\sum_{i=1}^n \mathbb{1}\{y_i = +1\} = m$. En utilisant par exemple la méthode des moments (*i.e.,* on remplace les espérances par leurs contreparties empiriques), proposer des estimateurs $\widehat{\pi}_+$, $\widehat{\mu}_+$, $\widehat{\mu}_-$ et $\widehat{\Sigma}$ des paramètres.

On se propose d'étudier un classifieur (*i.e.,* une fonction qui à tout point associe létiquette prédite) obtenu par minimisation du critère des moindres carrés :

$$\mathbb{R}^{p+1} \ni \hat{\boldsymbol{\theta}} = \underset{\theta_0 \in \mathbb{R}, \theta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^n (y_i - \theta_0 - \theta^\top \mathbf{x}_i)^2,$$

et l'on obtient alors un prédicteur de la forme suivante : $\mathrm{sign}\left(\theta_0 + \theta^\top \mathbf{x}\right)$, où la fonction sign est définie par

$$\mathrm{sign}(x) = \begin{cases} 1 & \text{si } x > 0, \\ -1 & \text{si } x < 0, \\ 0 & \text{sinon.} \end{cases}$$

3. Construire numériquement (avec par exemple la fonction `lm`) le classifieur ci-dessus, et l'appliquer aux jeu de données simulées.
4. Écrire la condition d'annulation du gradient et mettre en évidence le fait que la solution $\hat{\boldsymbol{\theta}}$ doive satisfaire une équation de la forme :
$$(\alpha\widehat{\Sigma} + \gamma\widehat{\Sigma}_B)\hat{\boldsymbol{\theta}} = n(\widehat{\mu}_+ - \widehat{\mu}_-)$$
où $\widehat{\Sigma}_B = (\widehat{\mu}_+ - \widehat{\mu}_-)(\widehat{\mu}_+ - \widehat{\mu}_-)^\top$ et α, γ sont des facteurs dépendant de n et m à préciser.
5. Montrer alors que $\widehat{\Sigma}_B\hat{\boldsymbol{\theta}}$ est porté par la direction $(\widehat{\mu}_+ - \widehat{\mu}_-)$. En déduire que $\hat{\boldsymbol{\theta}}$ est proportionnel à $\widehat{\Sigma}^{-1}(\widehat{\mu}_+ - \widehat{\mu}_-)$.
6. Vérifier cette dernière propriété numériquement.

Références

[1] M. Lejeune. *Statistiques, la théorie et ses applications*. Springer, 2010. 2