MAP433 Statistique

PC8 - Tests et regression

1 Test dans un modèle ANOVA

1.1 Formalisation et construction du test

Soit K un entier supérieur ou égal à 2.
Pour tout $k = 1, 2, \dots K$, on dispose de n_k observations, que l'on suppose être des réalisations de n_k variables aléatoires $X_{k,1}$, $X_{k,2}$, …, X_{k,n_k} indépendantes et de loi $\mathcal{N}(m_k, \sigma^2)$.
La variance σ^2 est commune à tous les groupes et est supposée inconnue.
On chercher à tester l'hypothèse nulle H_0: $m_1 = \cdots = m_K$.
On posera dans la suite $\mathbf{X} = (X_{ki}, 1 \leq k \leq K, 1 \leq i \leq n_k)$.

1. Ecrire la vraisemblance des observations $\mathbf{X}$.
2. Calculer l'estimateur de maximum de vraisemblance des paramètres du modèle sous H_0.
3. Calculer l'estimateur de maximum de vraisemblance des paramètres du modèle.
4. Montrer que le test de rapport de vraisemblance généralisé conduit à utiliser une statistique de test dont la loi suit une loi de Fisher dont on précisera les degrés de liberté.
5. Quelle est la forme de la région de rejet de H_0 pour ce test?

1.2 Un exemple

On s'intéresse à l'impact de différents régimes alimentaires sur le poids de chiens de même race. On dispose d'un groupe témoin (ou contrôle) de 10 animaux qui reçoivent une alimentation classique et de deux groupes tests à qui l'on fait suivre deux régimes différents. Les résultats (en kg) par groupe sont résumés dans le tableau suivant:

Groupe	effectif	$\sum_i x_{ki}$	$\bar{x}_k$	$\sum_i x_{ki}^2$	$\sum_i (x_{ki} - \bar{x}_k)^2$
Contrôle	10	50.32	5.032	256.2702	3.05996
Régime 1	10	46.61	4.661	222.9185	5.66929
Régime 2	10	55.26	5.526	307.1296	1.76284

On vérifie alors que $\bar{x} = 5.073$ et que $\sum_k \sum_i (x_{ki} - \bar{x}_k)^2 = 14.2584$. Peut-on conclure que ces différents régimes alimentaires n'ont pas tous le même effet sur le poids des animaux?

2 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 $\{(\hat{Y}_i, |\hat{\xi}_i|) : i = 1, \ldots, n\}$ 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.