M1- Mathématiques Appliquées - ENSTA / Paris-Saclay
Modélisation statistique (MAP-STA201), 2020-2021
christine.keribin@universite-paris-saclay.fr
Chargé.e.s de TD: Olivier Coudray, Henri Mermoz Kouye, Zacharie Naulet, Camille Palmier

Projet à rendre pour le 4 novembre 18h

1 La transformation de Box-Cox

Une transformation non linéaire des variables peut permettre de proposer une modélisation plus adaptée au jeu de données étudié. Par exemple, la transformation de Box-Cox étudie les transformations en puissance de la variable à expliquer, ce qui peut être intéressant pour stabiliser la variance lorsque l'hypothèse d'homoscédasticité (variance constante) n'est pas vérifiée, ou "gaussianiser" les données lorsque leur distribution est fortement dissymétrique.

Le principe de la méthode est le suivant: y étant la variable à expliquer, on définit une famille de transformations paramétriques (h_λ) de y en $h_\lambda(y)$ et on considère le modèle d'observation (x_i, Y_i):

$$h_\lambda(Y_i) = Z_i = \mathrm{x}_i\theta + \varepsilon_i, \;\; \varepsilon_i \sim_{iid} \mathcal{N}(0, \sigma^2) \tag{1}$$

où x_i est le vecteur ligne des conditions d'expérience et ε_i le bruit iid gaussien. L'objectif est de déterminer λ_{opt} permettant un meilleur ajustement en régression de y que celui obtenu en régression linéaire (pour λ_{lin} tel que $h_{\lambda_{lin}}$ est l'identité).

Box et Cox[1] ont proposé la famille de transformations $(\tilde{h}_\lambda)$ suivante, pour $\lambda \in \mathbb{R}$:

$$\forall y > 0, \;\; \tilde{h}_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases}$$

tandis que Bickel et Doksum[2] définissent, pour $\lambda > 0$

$$\forall y, \;\; h_\lambda(y) = \frac{sgn(y)|y|^\lambda - 1}{\lambda}$$

où $sgn(y)$ est le signe de y (1 si $y > 0$, -1 si $y < 0$, 0 si $y = 0$).

1. Expliquer pourquoi la transformation $\tilde{h}_\lambda$ de Box et Cox n'est pas en théorie compatible avec (1) sauf pour $\lambda = 0$, mais peut être utilisée de façon pratique quand toutes les observations du jeu de données sont positives.

   Dans la suite de l'étude théorique, on considèrera h_λ.

2. Montrer que la vraisemblance des paramètres $(\lambda, \theta, \sigma^2)$ à l'observation de $Y = (Y_1, \ldots, Y_n)$ s'écrit

$$L(\lambda, \theta, \sigma^2; Y) = \frac{J(\lambda; Y)}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(h_\lambda(Y) - X\theta)'(h_\lambda(Y) - X\theta)\right)$$

   avec X la matrice du plan d'expérience, $h_\lambda(Y)$ le vecteur de composantes $h_\lambda(Y_i)$ et $J(\lambda; Y)$ un terme à déterminer.

[1] *An Analysis of Transformations*, Journal of the Royal Statistical Society. Series B (Methodological) , 1964, Vol. 26, No. 2 (1964), pp. 211-252

[2] *An Analysis of Transformations Revisited*, JASA, June 1981, Vol. 76, Number 371

3. A λ fixé, déterminer l'estimateur du maximum de vraisemblance (emv) $\widehat{\theta}(\lambda)$ et $\widehat{\sigma}^2(\lambda)$ en fonction de X et λ.

   Montrer que

   $$L_{max}(\lambda) := \log L(\lambda, \widehat{\theta}(\lambda), \widehat{\sigma}^2(\lambda)) = -\frac{n}{2}\log \widehat{\sigma}^2(\lambda) + (\lambda - 1)\sum_i \log |Y_i| + a(n)$$

   où $a(n)$ est une constante ne dépendant que de n. Calculer l'équation à vérifier par $\widehat{\lambda}$. Quelle méthode proposer pour calculer $\widehat{\lambda}$ (on ne demande pas sa mise en œuvre)? L'emv est-il gaussien à distance finie?

   Dans la suite, on considèrera que le plan d'expérience X est tel que $(X'X)/n$ tend, quand n tend vers l'infini, vers une matrice définie positive, et on admettra que dans ce cas, l'emv est asymptotiquement gaussien, de matrice de variance de la loi limite Σ dont l'expression n'est pas demandée.

4. Quel estimateur proposer pour la variance de l'emv $\widehat{\lambda}$ de λ?

   Construire un intervalle de confiance asymptotique de λ de niveau $1-\alpha$, puis le test de Wald de $(H_0) : \lambda = \lambda_0$ contre $(H_1) : \lambda \neq \lambda_0$.

5. Rappeler la loi asymptotique suivie par la statistique $LRV = 2(L_{max}(\widehat{\lambda}) - L_{max}(\lambda))$ lorsque les données ont été générées avec le modèle (1) et h_λ, puis construire le test du rapport de vraisemblance $(H_0) : \lambda = \lambda_0$ contre $(H_1) : \lambda \neq \lambda_0$

2 Test de la méthode sur des données simulées

On souhaite mettre en œuvre la méthodologie précédente sur des données simulées suivant le modèle

$$h_{0.3}(Y_i) = Z_i = a + bx_i + \varepsilon_i, \quad \varepsilon_i \sim_{iid} \mathcal{N}(0, \sigma^2) \tag{2}$$

où $a = 5$, $b = 1$, $\sigma^2 = 2$, $i = 1, \ldots, n$. Les x_i sont issues d'une loi gaussienne centrée réduite.

1. Afin de rendre vos simulations reproductibles, initialiser le germe du générateur aléatoire (`set.seed(999)`), puis générer les x_i (`rnorm(n)`) que vous conserverez ainsi tout au long de la section.

   La condition de convergence indiquée dans la section 1 est-elle vérifiée?

   Générer un $n = 50$-échantillon de loi gaussienne $\mathcal{N}(0, \sigma^2 = 2)$, en déduire l'échantillon $Z = (Z_1, \ldots, Z_n)$, puis l'échantillon $Y = (Y_1, \ldots, Y_n)$

   Estimer la régression linéaire simple de z fonction de x et celle de y fonction x. Étudier dans chaque cas les résidus studentisés (normalité, graphe fonction des valeurs observées) et commenter.

2. Créer la matrice `X` du plan d'expérience (ne pas y oublier l'intercept), puis interpréter le code suivant:

```
Q = diag(1,n) - X%*%solve(t(X)%*%X)%*%t(X)
Lmle = function(Z){
  n = length(Z)
  sig2 = ( t(Z)%*%Q%*%Z )/n
  -n/2*log(sig2)
}
```

Coder la fonction `lmin(lambda, Y)` qui calcule $-L_{max}(\lambda)$.

Tracer $-L_{max}(\lambda)$ pour $\lambda \in [0; 2]$, en déduire une valeur de $\widehat{\lambda}$ lue sur le graphique.

Remarque: on peut vectoriser la fonction `lmin`: `Vlmin = Vectorize(lmin,"lambda")` pour éviter une boucle.

3. Utiliser la fonction `nlm` pour optimiser $-L_{max}(\lambda)$ en complétant le code suivant:

```
resopt = nlm(lmin,Y=Y,p=??,hessian=TRUE) # ?? à préciser
```

En déduire $\widehat{\lambda}$ et l'estimation de sa variance.

4. Déterminer un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour λ.

Tester par Wald (toutes les alternatives sont bilatères):

(H_0) : les données Y ne nécessitent pas de transformation;

(H_0) : la transformation à appliquer aux observations est en racine carrée.

(H_0) : $\lambda = 0.3$

(H_0) : la transformation à appliquer aux observations est $\tilde{h}_0$ (on supposera qu'il est possible d'utiliser les résultats de la section 1 dans ce cas, sans chercher de preuve théorique).

Commenter les résultats.

5. Tester les mêmes hypothèses avec le test de rapport de vraisemblance.

6. Retrouver les résultats précédents avec la fonction `powerTransform` du package `car`.

7. Simuler le niveau du test du rapport de vraisemblance de $(H_0) : \lambda = 0.3$ contre $(H_1) : \lambda \neq 0.3$. Pour ce faire, générer $B = 100$ n-échantillons Y sous (H_0), et observer le nombre moyen de rejets du test.

Indiquer ce qu'il faut changer dans le processus précédent pour simuler la puissance du même test en l'alternative $\lambda = \lambda_1$, puis estimer cette puissance en $\lambda_1 = 0.1$.

Tracer la courbe de puissance simulée pour λ_1 compris entre 0.05 et 1.5.

[*Bonus*] Reprendre la question 7 pour le test de $(H_0) : \lambda = 0$, et représenter la courbe de puissance entre 0.1 et 0.3. Commenter.

3 Cas pratique

Le jeu de données `NbCycleRupture.csv` donne le nombre de cycles à rupture (`y`) d'un fil peigné en fonction de trois variables `x1`, `x2` et `x3` définies à partir des mesures de Box et Draper (1987) par

$$\begin{aligned} \text{Longeur (en mm)} &: x_1 = \frac{\text{longueur} - 300}{50} \\ \text{Amplitude du cycle de chargement (en mm)} &: x_2 = \text{amplitude} - 9 \\ \text{chargement (en gr)} &: x_3 = \frac{\text{chargement} - 45}{5} \end{aligned}$$

Lire les données, vérifier que le data.frame obtenu comporte 27 observations.

1. Définir le modèle de régression avec effets additifs (M1), puis l'ajuster. Expliquer ce qui change et ce qui ne change pas dans les résultats lorsqu'on applique ou non la transformation des variables indiquée.

   Commenter les valeurs du R^2, la significativité globale de la régression, des variables. Les hypothèses sur le modèle sont-elles vérifiées?

2. Ajouter les termes `xi*xj` d'ordre 2, puis analyser ce nouveau modèle (M2) et en commenter les résultats.

   Construire le test de (M1) contre (M2). L'ajout des variables d'interaction est-il à préconiser?

   Note: il est possible de définir les termes d'ordre 2 sans les calculer formellement. Par exemple, pour deux variables: `Y~x1 + x2 + I(x1^2) + I(x2^2) + I(x1*x2)`

3. Proposer le choix d'une transformation pour stabiliser la variance du modèle (M1). Réestimer le modèle additif (M1bis), puis le modèle avec interaction (M2bis) suivant cette transformation.

4. Quel modèle proposez-vous?

Quelques consignes Ce projet donne lieu à un compte-rendu *rédigé* à effectuer en binôme et à remettre sur ecampus. **Un seul des deux membres du binôme met en ligne** sur son compte e-campus. Le rendu pour l'autre sera donc vide, mais une note bien attribuée.

- sous forme d'un `pdf` et d'un fichier `.R` contenant les commandes. Les fichiers seront nommés avec votre `NOM`, soit `NOM1-NOM2.pdf` et `NOM1-NOM2.R`.
- Définir en particulier un titre informatif, une introduction pour préciser la problématique étudiée et le plan du travail, une conclusion.
- Inclure dans le CR toutes les valeurs intervenant dans les applications numériques, mais pas le code qui sera reporté dans le fichier .R.
- Il est important de commenter les résultats obtenus, même si ce n'est pas formellement demandé.