Université de Strasbourg
M1 - Magistère
Statistique - projet

Ségolen Geffray
geffray@math.unistra.fr
Année 2020/2021

TP2: régression linéaire

Notons β le vecteur des paramètres de régression à estimer dans le cadre d'un modèle de régression linéaire et $\widehat{\beta}$ son estimateur.

1. Proposer des simulations de Monte-Carlo permettant d'illustrer la convergence asymptotique de $\widehat{\beta}$ vers β dans le cadre de l'ajustement d'un modèle de régression linéaire gaussien lorsque le modèle proposé s'adapte bien aux données.

2. Proposer des simulations de Monte-Carlo permettant de comparer le comportement des différents types de résidus introduits en cours dans le cadre de l'ajustement d'un modèle de régression linéaire gaussien.

3. Proposer des simulations de Monte-Carlo permettant d'illustrer la robustesse du modèle dans le cadre d'un modèle de régression linéaire gaussien.

Vous veillerez notamment à

- faire varier n, la taille de l'échantillon,
- étudier l'impact du nombre de prédicteurs,
- étudier l'impact de l'inclusion de prédicteurs superflus,
- étudier l'impact d'une éventuelle colinéarité entre prédicteurs.

Pensez également à quantifier votre analyse au moyen de critères objectifs tels que biais empirique, écart-type estimé, probabilité de couverture dans le cas d'intervalles de confiance, erreurs de type I et II dans le cas de tests.

Quelques fonctions utiles et quelques recommandations

Pour ajuster un modèle de régression linéaire gaussien, utiliser

```
lm(formula,data)
```

Supposons que `Y`, `X1`, `X2` et `X3` sont des variables quantitatives (donc du type `numeric`), `A` est une variable qualitative (donc du type `factor`). Voici quelques formules possibles pour écrire un modèle linéaire.

`Y∼X1`	régression linéaire simple avec intercept implicite
`Y∼1+X1`	régression linéaire simple (identique au précédent) avec intercept explicite
`Y∼-1+X1`	régression linéaire simple sans intercept
`Y∼0+X1`	régression linéaire simple sans intercept (identique au précédent)
`Y∼X1-1`	régression linéaire simple sans intercept (identique au précédent)
`log(Y)∼X1+X2`	régression linéaire multiple sur `log(Y)` (avec intercept implicite)
`Y∼X1*X2`	régression linéaire multiple avec interaction d'ordre 2
`Y∼X1*X2*X3-X1:X2:X3`	régression linéaire multiple avec interaction d'ordre 2
`Y∼(X1+X2+X3)^2`	régression linéaire multiple avec interaction d'ordre 2 (identique au précédent)
`Y∼A`	analyse de la variance à un critère de classification
`Y∼A+X1`	analyse de la covariance
`Y∼A2%in%A1`	régression linéaire avec 2 covariables, `A2` étant emboitée dans `A1`

str (utils)	donne la structure d'un jeu de données
head (utils)	permet de voir les premières lignes d'un jeu de données
rmvnorm (mvtnorm)	simule des vecteurs gaussiens
scatter.smooth (stats)	fournit des tracés exploratoires
scatterplotMatrix (car)	fournit des tracés exploratoires
ggPairs (GGally)	idem avec estimation des coefficients de corrélation linéaire
pairs (graphics)	trace les covariables 2 par 2
boxcox (MASS)	transformation de Box-Cox dans le cas d'un LNM
bcPower (car)	transformation de Box-Cox, Yeo-Johnson ou puissance
lm (stats)	ajuste un modèle de régression linéaire gaussien standard
summary (base)	renvoie les résultats de l'ajustement du modèle
model.matrix (stats)	renvoie la matrice expérimentale
logLik (stats)	renvoie la log-vraisemblance du modèle
confint (stats)	détermine un IC pour chaque coefficient d'un LNM
shapiro.test (stats)	effectue un test de Shapiro-Wilk
ks.test (stats)	effectue un test de Kolmogorov-Smirnov
qqnorm (stats)	effectue un QQ-plot
qqline (stats)	ajoute à un QQplot une droite qui passe par les 1^{ers} et $3^{\text{èmes}}$ quartiles
influence.measures (stats)	effectue un diagnostic d'individus influents
hatvalues (stats) =hat (stats)	renvoie les leviers
cooks.distance (stats) =cookd (car)	calcule la distance de Cook
dffits (stats)	calcule les *dffits*
dfbetas (stats)	calcule les *dfbetas*
coeftest (lmtest)	effectue de Student sur chacun des coefficients
anova (stats)	comparaison de deux modèles emboités avec un F-test
waldtest (lmtest)	effectue un test de Wald pour modèles emboités
bptest (lmtest)	test d'hétéroscédasticité de Breusch-Pagan
ncvTest (car)	test d'hétéroscédasticité
leveneTest (car)	test d'homogénéité des variances entre différents groupes
linearHypothesis (car)	effectue des tests linéaires d'hypothèses
residuals (stats)	renvoie les résidus de base
rstandard (stats)	renvoie les résidus standardisés d'un LNM
rstudent (stats)	renvoie les résidus studentisés d'un LNM