Travaux dirigés - Régression linéaire simple

Julien Chiquet et Guillem Rigaill

1er octobre 2015

Objectifs de la séance

- maîtriser la régression linéaire simple
- interpréter les sorties de `R` associées

Remarques

- les TD doivent être faits en binômes
- vous devez utilisez l'interface `R studio`
- vous préparerez un rapport succint sur l'exercice portant sur la régression linéaire à l'aide de `R studio` (fichier `R markdown`) que vous enverrez à vos chargés de TD respectifs, en intégrant votre code et commentant les sorties associées. Une aide à la syntaxe markdown se trouve à cette page.
- dans la mesure du possible, vous préférerez l'utilisation des outils du package `ggplot2` pour vos représentations graphiques à celles intégréées à `R`.

Pour ceux qui n'ont jamais fait de `R`

Rendez-vous à la page http://tryr.codeschool.com/. Compléter les premiers niveaux du challenge directement en ligne.

Quelques révisions de `R`

1. **Manipulation de vecteur**. On rappelle que

$$e^x = \sum_{k \geq 0} \frac{x^k}{k!}.$$

   Créer dans un vecteur `exp2` les 20 premiers termes de cette suite. Supprimer toutes les valeurs inférieures à 10^{-8}. En déduire une approximation de e^2 et comparer avec la valeur `exp(2)`.

2. **Simuler des données**. Simuler avec la fonction `rnorm` (l'aide est accessible avec la commande `?rnorm`) un vecteur X de taille 100 issu d'une loi normale de moyenne 2 et variance 1. Simuler un autre vecteur Y de taille 100 obtenu en multipliant X par 9.8 et en rajoutant un bruit Gausssien d'écart-type 1/10.

3. **Lire et écrire un fichier de données**. Mettre le vecteur X et Y dans un `data.frame`. Sauvegardez le `data.frame` avec la commande `write.table`. Le relire avec la commande `read.table`. Comparer le tableau obtenu après relecture avec le tableau initial.

4. **Lire et écrire un fichier de données au format RData**. Mettre le vecteur X et Y dans un `data.frame`. Sauvegardez le `data.frame` avec la commande `save`. Le relire avec la commande `load`. Comparer le tableau obtenu après relecture avec le tableau initial.

5. **Nuage de points**. Tracer le nuage de point de Y en fonction de X, d'abord avec la commande `plot` puis à l'aide de la librairie `ggplot2`.

6. **Histogramme**. Tracer l'histogramme de X.

7. **Boucle For**. Une variable aléatoire suit une loi du χ^2. On ne connait pas son nombre de degré de liberté. On souhaite estimer ce degré à partir de la moyenne empirique de n realisations de cette variable aléatoire. Utiliser `R` pour évaluer la qualité de cette estimation pour $n = 3$ et $n = 100$. Vous utiliserez une boucle `for`.

8. **Boucle for** (plus rapide, plus élégant, plus dans l'esprit). Même question en utilisant la fonction sapply.

Régression linéaire simple: Brochet et DDT

Le DDT (dichlorodiphényltrichloroéthane) est un insecticide relativement puissant. Il est toxique et n'est pas dégradé de manière naturelle. Il s'accumule dans certains tissus tels que le foie et les tissus adipeux. On étudie ici l'accumulation du DDT chez les brochets.

1. Préliminaires

a) Importer les données 'Brochet.txt'.
b) Calculer la moyenne, la médiane, la variance de l'age des brochets et du taux de DDT.
c) Tracer l'histogramme de l'âge et du taux de DDT. Augmenter le nombre de barres. Que constate-t-on ?
d) Tracer, sous la forme d'un nuage de point, le graphe du taux de DTT en fonction de l'âge des brochets.
e) Tracer, sous la forme de box-plot, le graphe du taux de DTT en fonction de l'âge des brochets. Que constate-t-on ?
f) Calculer la variance du Taux de DDT par classe d'âge.

2. Un premier modèle

a) Écrire un modèle de regression linéaire permettant d'expliquer le taux de DDT en fonction de l'âge.
b) Utiliser `R` pour estimer les paramètres de ce modèle. Vous appliquerez tout d'abord les formules du cours, puis vous utiliserez la fonction `lm`. Donner l'ordonnée à l'origine, la pente et la variance résiduelle.
c) Tester les paramètres du modèles. Faites une analyse de la variance. Calculer `à la main` la valeur de la statistique de Fisher ainsi que la valeur du coefficient d'ajustement.
d) Tracer la droite de regression. Ajouter les intervalles de confiance de prédictions. Vous les calculerez d'abord à l'aide des formules du cours puis en vous aidant de la commande `predict`.
e) Faites un graphes des résidus pour évaluer la pertinence de votre modèle et effectuer les diagnostics d'usage. Vous pourrez également utiliser la commande `plot` de `R` appliqué à l'objet issu de la fonction `lm`.

3. Modèle quadratique

a) Écrire un nouveau modèle de regression linéaire permettant d'expliquer le taux de DDT en fonction de l'âge au carré.
b) Utiliser `R` pour estimer les paramètres de ce modèle.
c) Tester les paramètres du modèles. Faites une analyse de la variance.
d) Tracer la droite de regression. On pourra utiliser la fonction `geom_smooth` de `ggplot`
e) Effectuer le diagnostic du modèle

4. Modèle transformation logarithmique

a) Écrire un nouveau modèle de regression linéaire permettant d'expliquer le log du taux de DDT en fonction de l'âge.

b) Tracer, sous la forme de box-plot, le graphe du log du taux de DTT en fonction de l'âge des brochets. Que constate-t-on ?
c) Calculer la variance du Taux de DDT par classe d'âge.
d) Utiliser `R` pour estimer les paramètres de ce modèle.
e) Tester les paramètres du modèles. Faites une analyse de la variance.
f) Tracer la droite de regression. On pourra utiliser la fonction `geom_smooth` de `ggplot`.
g) Effectuer le diagnostic du modèle

5. Vers la régression multiple

a) Écrire un modèle de regression linéaire permettant d'expliquer le log du taux de DDT en fonction de l'âge et l'âge au carré.
b) Utiliser R pour estimer les paramètres de ce modèle.
c) Tester les paramètres du modèles. Faites une analyse de la variance pour comparer les 3 modèles M0, M1, M2 (intercept, + âge, + le carré de l'âge).
d) Tracer la courbe de regression.
e) Validation des hypothèses. Utiliser R pour évaluer la pertinence du modèle. Qu'en pensez-vous ?

Pour ceux qui ont été vite: maximisation numérique de la vraisemblance

Considérons un phénomène modélisé par une loi normale $\mathcal{N}(\mu, \sigma^2)$.

1. Calculer analytiquement $\log L(x_1, \ldots, x_n; \mu, \sigma^2)$.
2. Déterminer les estimateurs du maximum de vraisemblance en dérivant successivement $\log L$ par rapport à μ et σ^2.
3. Générer un jeu de données gaussien de taille $n = 100$, de moyenne $\mu = \pi/2$ et d'écart type $\sigma = \sqrt{2}$. Calculer les valeurs prises par les estimateurs du maximum de vraisemblance de μ et σ^2 obtenus dans les questions précédentes.
4. Écrire une fonction `loglikelihood` qui prend en argument `x,mu,sigma` et renvoie la valeur de la fonction de log-vraisemblance pour $(x_1, \ldots, x_n), \mu$ et σ donnés.
5. À l'aide de la fonction `optimize`, déterminer numériquement les valeurs de μ et σ maximisant la fonction `loglikelihood`.
6. Dans une même fenêtre graphique, représenter
   - l'histogramme des données,
   - la fonction $\log L$ pour σ fixée à sa vraie valeur, en faisant varier μ sur $[\pi/2 - \varepsilon, \pi/2 + \varepsilon]$; situer également les valeurs estimées analytiquement et par `optimize` à l'aide de `abline`.
   - la fonction $\log L$ pour μ fixée, en faisant varier σ sur $[\sqrt{2} - \varepsilon, \sqrt{2} + \varepsilon]$; de même, situer les valeur estimées analytiquement et par `optimize`. Représenter également la variance empirique corrigée.
7. Créer une matrice `logL` contenant les valeurs de la log-vraisemblance en faisant varier à la fois μ et σ. Créer une liste `data=list(x,y,z)` et utiliser les fonction `persp`, `contour`, `image` pour représenter la vraisemblance en 3D et 2D.