Polycopié de travaux dirigés

ENSIIE

Régression avancée et R

chrisotphe.ambroise@genopole.cnrs.fr
Julien.chiquet@genopole.cnrs.fr

Semestre de printemps 2012
Université d'Évry Val d'Essonne

Table des matières

Travaux dirigés 1

Régression linéaire multiple

Exercice 1.1. Le jeu de données `gavote` décrit le vote présidentiel aux états-unis en 2000, dans l'état de Géorgie. Chacun des 159 "canton" est décrit par les variables suivantes :

- `equip` Le système physique de vote
  - LEVER : machine à levier
  - OS-CC : Scan optique comptage centralisé ("central count"),
  - OS-PC : Scan optique comptage local ("precinct count")
  - PAPER : vote par bulletin papier
  - PUNCH : vote par poinçon
- `econ` le statut économique du "canton" (`middle`, `poor`, `rich`).
- `perAA`, le pourcentage d'afro-américains
- `rural` indicateur de la ruralité du canton (urban, rural)
- `atlanta` indicateur de l'appartenance ou non à Atlanta
- `gore` nombre de votes pour Gore
- `bush` nombre de votes pour Bush
- `other` number of votes for other candidates
- `votes` nombre de votes validés
- `ballots` nombre de bulletins

1. Charger le jeu de données `gavote` et faites un résumé numérique des données.
2. Créer la variable `undercount` qui est la proportion de bulletins de vote considérés comme nuls. Représenter la distribution de cette nouvelle variable. Créer la variable `pergore` (pourcentage de votants pour `Gore`) et tracer le diagramme de dispersion croisant `pergore` avec `perAA`.
3. Tracer la droite de régresssion.
4. Représenter la distribution de la variable `equip`
5. Régresser `undercount` sur `perAA`. Interprétez ce modèle `model1`.
6. Calculer la somme des résidus au carré, le R^2 et le R^2 ajusté.
7. Régresser `undercount` sur `rural`. Interprétez ce modèle `model2`.
8. Centrer les variables `pergore` et `perAA`, et ajuster un modèle (`model3`) de régression linéaire qui explique `undercount` en fonction de `cperAA`, `cpergore`, `rural` et `equip`. Commentez.
9. Supprimer toutes les variables non significatives et comparer le modèle obtenu avec le précédent. Commentez.
10. En utilisant la procédure `step`, simplifier le modèle `model3`. Commentez.
11. Faites le diagnostique de la régression finale.

Travaux dirigés 2

Régression logistique

Exercice 2.1. Considérons le jeu de données `esoph` sur le cancer de l' œsophage en Ile-et-Vilaine L'objectif est de mener une étude cas-témoins [1] et de comprendre quels sont les facteurs influant sur le déclenchement de la maladie.

1. En regardant la structure du tableau de données expliquer la nature de chaque variable.
2. Faites un résumé numérique du tableau.
3. Créer un tableau de contingence à 2 lignes, 4 colonnes dont chaque case contient le nombre de cancer et les contrôles en fonction de leur consommation de tabac. Dessinez un diagramme mosaïque [2] du tableau.
4. Estimer un modèle `model0` de régression logistique en utilisant la seule variable `tobgp` (consommation de tabac).
5. Interpréter les coefficients en terme d'odd ratios (rapport de chance).
6. En recodant la première modalité comme non fumeur et fumeur pour les autres, estimer un modèle `model1` et calculer l'effet de la cigarette.
7. Estimer un modèle `model2` de régression logistique en utilisant la seule variable `tobgp` (consommation d'alcool).
8. Estimer un modèle `model3` de régression logistique en utilisant toutes les variables et interactions.
9. Utiliser la fonction `unclass` pour transformer toutes vos variables qualitatives en variables quantitatives. Estimer un dernier modèle `model4` avec ces dernières variables
10. Considérons le model `model4` : quel est la probabilité q'un homme de 25 ans qui ne boit pas ni ne fume développe un cancer de l'œsophage.
11. Faites une analyse graphique des résidus

1. Un groupe de personnes atteintes d'une maladie (cas) est comparé à un groupe de sujets qui n'ont pas la maladie étudiée (témoins). Le but est la recherche d'un ou des facteurs d'exposition antérieurs à la maladie susceptibles de pouvoir l'expliquer. Ce type d'étude sert donc à tester une hypothèse spécifique avec une association d'un facteur de risque.

2. Ce diagramme est composé d'autant de rectangles (et/ou carrés) qu'il y a cellules dans le tableau de contingence de départ. Pour la représentation de chaque mosaïque la largeur de la bande sera proportionnelle aux fréquences marginales. La hauteur de chaque mosaïque sera proportionnelle au rapport de l'effectif de la cellule sur le total de la colonne.

Travaux dirigés 3

Régression non paramètrique et modèle additif

Exercice 3.1. Considèrons le jeu de données `uswages` issu d'une étude de 1988. Regardons les variables nombre d'années d'éducation (`educ`) et salaire (`wage`).

1. Calculez un premier estimateur du salaire en fonction du nombre d'années d'études :
2. Calculez un second estimateur du salaire en fonction du nombre d'année en utilisant un estimateur à noyau (via `ksmooth`).
3. Comparer l'erreur quadratique moyenne sur les données d'apprentissage des deux estimations précédentes et commentez.
4. Calculez l'erreur quadratique moyenne en validation croisée. Commentez.

Exercice 3.2. Considérons les données `kyphosis`, qui décrivent les résultat d'une chirurgie corrective sur 81 enfants atteints d'une déformation de la colonne vertébrale. Ajuster un modèle additif et pour prédire la variable réponse `Kyphosis`, qui vaut un lorsqu'une déformation est présente après l'opération et zéro sinon. Commentez.

Travaux dirigés 4

Projet

Le jeu de données `dvisists` provient d'un étude du ministère de la santé Australien réalisée en 1977-1978. Pour accéder aux données, il vous suffit d'installer le module `faraway` et charger ensuite les données :

```
library(faraway)
data(dvisits)
```

1. Expliquer en quelques lignes la nature d'une régression log-linéaire ainsi que son intérêt.
2. Construire un modèle de regression de Poisson (log-linéaire) avec pour variable réponse le dvisists et comme variables explicatives `sex`, `age`, `agesq`, `income`, `levyplus`, `freepor`, `freerepa`, `illness`, `actdays`, `hscore`, `chcond1`, `chcond2`.

   Est-ce que ce modèle est un modèle raisonnable (utiliser la déviance pour répondre à cette question).
3. Afficher les résidus versus les données estimées. Comment expliquez vous les lignes sur le graphique affiché ?
4. Utiliser une procédure d'élimination descendante avec un seuil à 5% pour réduire la taille de votre modèle autant que possible.
5. Quel type de personne serait d'après votre modèle la plus susceptible de consulter le médecin ?
6. Pour la dernière personne du jeu de données prédire la probabilité de visite chez le docteur 0, 1, ou 2 fois.
7. Utiliser un modèle gaussien pour résoudre le problème. Décrire les différences.
8. Proposer et tester des modèles / méhodes de régression aternatives (non paramètrique par exemple) pour améliorer votre modèle.
9. Donner une estimation de l'erreur de prédiction de vos modèles.