Practical 2 - Multiple Linear Regression

J. Chiquet, G. Rigaill and A.A Charantonis

september 2016

Session's Objectives

- Master multiple linear regression
- Interpret the associated outputs in `R`

Notes

- The practicals must be made in pairs
- You will make a report (`R markdown` file) and send it to julien.chiquet@gmail.com.
- For any technical questions, do not hesitate to ask the supervisors.
- For theoretical questions please use as much as possible the slides available at [http://julien.cremeriefamily.info/teachings_ensai.html]
- Reports are to be sent by mail at the end of the session. They will be graded and they count for the final grade. Use comments to show that you understand what you do and that you understand the linear model and its integration in `R`.

2000 US Presidential vote in Georgia

The dataset `gavote` described the presidential vote in the United States in 2000, in the state of Georgia. Each of the 159 cantons is described by the following variables:

- `Atlanta`, indicates whether the canton is in Atlanta or not
- `Ballots`, number of ballots
- `Bush`, votes for Bush
- `Econ`, town economic status (`middle`, `poor`,`rich`).
- `Equip`, physical voting system
  - STAND: lever machine
  - OS-CC: optical scan - centralized counting,
  - OS-PC: optical scan - local counting
  - PAPER: paper ballot
  - PUNCH: punched card
- `Gore`, votes for Gore
- `Other`, votes for candidates other than Bush and Gore
- `PerAA`, the percentage of African Americans
- `Rural`, indicator of rural town (urban‘`,`rural‘)
- `Votes`, number of valid votes

1. Preliminaries

- Load the `gavote` data set at [http://julien.cremeriefamily.info/reglin/gavote.txt]
- Create the variable `undercount` (proportion of invalid ballots) and add it to the table. We'll look at the prediction of this variable using others. It is therefore the "response variable". Delete the variables `votes` and `ballots` (and explain why).
- Create variables `pergore`, `perbush` and `perother` (percentage of ballots for Gore, Bush and other candidates). Add them to the table. Delete the `gore` , `bush` and `other` variables of the table.

2. Descriptive Analysis

Make a descriptive analysis. This analysis will help you to choose a particular linear models and simplify your interpretation of the results.

- There is a wide choice of tools in `R`: summary, histogram, boxplot, barplot, scatter plot, distribution function, correlation matrix, hierarchical clustering, etc. Be imaginative! The idea is not to integrate all possible graphs. Integrate only those that you find usefull and for which you have something to say. *

3. Some simple linear models

Votes among African Americans

- Draw the scatterplot between `pergore` and `perAA`. Fit a simple regression model and draw the regression line between `pergore` and `perAA`.
- Regress `undercount` with `pergore`, then with `perAA` and finally with `pergore` and `perAA`. Compare these models and interpret the results.

Votes according to the standard of living and one-way ANOVA

To study the effect of the `econ` factor on the variable `undercount`, we propose the following ANOVA model:

$$Y_i = \mu + \mathbf{1}_{\{i \in k\}} \mu_k + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where y_i describes the variable `undercount` associated with the i th individual, μ the intercept and μ_k an additive term associated with the group k (that is to say to the modality k of the variable `econ`).

- Observe the levels of the categorical variables `econ`. Recode this variable by classifying these terms in an easily readable order.
- Represent the distribution of the variable `econ`.
- Show that the above model can be written as a linear model of the form

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

- Study the effect of `econ` factor.
- Pay attention to the interpretation of the regression coefficients: the ANOVA model is over parameterized. Constraints on the values of μ_k are needed. In `R`,the first level of the factor is set to zero. It is used as a reference for the values of the coefficients associated with other levels. *

Votes according ...

Illustrate a relevant behavior of your choice in the dataset by fitting a simple linear model different from those considered so far.

4. Construction of multiple linear models

Construction of a model using tests

- Fit a linear regression model that explains `undercount` based on the quantitative variables `perbush`, `pergore`, `perother` and `perAA`. How do you explain the results? Propose an alternative model by removing a variable (and justify). Remove this variable from the `gavote` data table.
- Fit a model that explain `undercount` based on all the (remaining) explanatory variables. Build a model `Model.test` that integrates only the significant variables.
- Study the residue of `model.test`.

Construction by stepwise regression

"Stepwise" regression is a strategy that builds a model step by step starting from the null model and adding / removing a new explanatory variable based on a criterion (the R^2 for example). We will see in class how to define a model selection criterion aiming for a compromise between fit to the data and the number of parameters. In this context, the AIC and BIC are the most conventionally used criteria.

- Using the procedure `step`, propose a model using the forward / backward procedureusing the AIC and BIC criteria. We will call these model `model.BIC` and `model.AIC`.
- Study the residues of these models.

Exhaustive search

When the number of predictors remains reasonable (<30), it is possible to test all linear models and choose the best in the sense of a criterion.

- Use the package `regsubsets` and its procedure `leaps`, to identify the best model with respect to the adjusted R^2. Comment.
- Study the residue of `model.R2`.

5. Study and refine of the final model

- Choose a model between `model.test`,`model.AIC`, `model.BIC` and `model.R2`. Make a complete diagnosis (residue, test hypothesis related to the linear model, Cook's distance, levers). You can use the functions `plot.lm`, `cook.distance`, `rstudent` ...
- Try to improve the fit to the data by excluding outliers, transforming the answer or some predictors. You can also merge the levels of categorical variable if they have similar effects.
- Interpret the final model that you have.