Linear Models
Winter 2015 – 2016

MATH EPFL
V.M. Panaretos

PRACTICAL 2 - November 25, 2015 (due 19:00, 17/12/2015)

Purpose

The purpose of this practical is to familiarize you with ANOVA, model selection procedures, checking for multicollinearities, and their implementation in `R`.

Data

We continue with the automobile mileage dataset from Practical 1. In Practical 1, we fitted and investigated the following model:

$$\frac{100}{\text{City MPG}} = \beta_0 + \beta_1 \text{Weight} + \beta_2 \frac{\text{Horsepower}}{\text{Weight}} + \varepsilon, \tag{1}$$

with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. For the data analysis of this practical, you should keep all the 82 observations of the original dataset.

Questions to be addressed

In the first practical, the questions to be addressed were posed in a way that guided you to structure the report correctly. For this practical, only the questions will be posed. When writing your report, you should, as before, follow closely the structure and guidelines given in Practical 1. You should also briefly explain in the methodology section of your report what are the algorithms you are going to use and how they work.

1. Conduct a full analysis of variance for model (1), give the ANOVA table (`?aov, ?anova`), and interpret the results at 5% significance level. Are the results the same if you change the order of inclusion of the terms? Comment on this.
2. We are interested in seeing whether we should build other more complicated models. In particular, we would like to investigate models in which we regress the response $\frac{100}{\text{City MPG}}$ on variables with label in the set $\{11, 12, \dots, 25, 26\}$ (see Table 1).
    a) Fit the full model with all the predictors $\{11, 12, \dots, 25, 26\}$, and look at the individual t-tests. What do you notice? Compute the variance inflation factors (`library(car); ?vif`) of the model. What is the problem here?
    b) Use backward deletion and forward selection to build models, using AIC as a model selection criterion (`?step`). Are the results the same? Do the results change if you use BIC as the model selection criterion (option `k = log(n)` in `step`)?
    c) Look at the models you found in b): do they make sense? Also, did performing model selection remove the collinearity problems? (You do not need to include all the tables here; just point out the problems you notice.)
    d) Now redo backward and forward selection with AIC as well as with BIC, but after removing the predictors $\{17, 19, 20, 21, 22\}$. Justify the removal of these variables using scatter plots (`?pairs`), correlations (`?cor`) and your insight into the problem.

e) After the model selection in point d), you end up with two distinct models. Examine these models — do they seem reasonable to you? Do we still have problems with multicollinearity? Choose the more parsimonious model as your final model and perform regression diagnostics for it (standardized residuals vs. fitted values, QQ plot, Cook's distances).

Some useful advice

- Use the covariate `Horsepower/Weight` only in part 1 of the practical. In part 2, there is no need to consider transformations of the covariates (but do remember to transform the response!).
- The `Sum Sq` values returned by `aov` and `anova` are the *changes* in the residual sum of squares when variables are added. To access the plain RSS values, you can use `anova(m1,m2,m3)`, where `m1`, `m2` and `m3` are nested `lm` objects in the order of the addition of variables.
- You can use the shortcut `lm(y∼., data = my.data)` to regress `my.data$y` on all the other variables in the data frame `my.data`.
- When performing forward selection with `step`, it is convenient to set the `scope` argument using `formula(my.data)`, where `my.data` is a data frame containing the response as the 1st column and the covariates as the subsequent columns.

Specifications for the report

1. This report must be done in groups of 2.
2. You can write the report in French or in English.
3. The report must follow the structure and the specifications given in Practical 1.
4. You must return the code you have written in the appendix of the report (see Practical 1).
5. We strongly suggest that you write the report in LaTeX.
6. All figures and tables should be numbered, have a caption and be elegant. Reference should be made to each figure/table from within the text.
7. Mention any references you have used and provide a detailed bibliography. References should be made to scientific articles or books, *not* to a course or a website.
8. Pasting plain computer output is not acceptable.

Finally, the most important remarks:

The report should be maximum 14 pages long (excluding the appendix), in 12pt font. You should print it out and return the *printed* version to the box next to office MA B1 493 before *Thursday December 17, 19:00.*

Reports that do not match these requirements will not be considered.

Table 1: Variables in car dataset

Column	Description
1	Manufacturer
2	Model
3	Type: Small, Sporty, Compact, Midsize, Large
4	Minimum Price (in $1,000) - Price for the base version
5	Midrange Price (in $1,000) - Average of Min and Max prices
6	Maximum Price (in $1,000) - Price for the fully loaded version
7	City MPG (miles per gallon as rated by EPA)
8	Highway MPG
9	Air Bags standard [0 = none, 1= driver only, 2= driver and passenger]
10	Drive train type [0 = rear wheel drive, 1= front wheel drive, 2 = all wheel drive]
11	Number of cylinders
12	Engine size (liters)
13	Horsepower (max)
14	RPM (revolutions per minute at maximum horsepower)
15	Engine revolutions per mile (in highest gear)
16	Manual transmission available [0 = No, 1= yes]
17	Fuel tank capacity (gallons)
18	Passenger capacity (persons)
19	Length (inches)
20	Wheelbase (inches)
21	Width (inches)
22	U-turn space (feet)
23	Rear seat room (inches)
24	Luggage capacity (cu. ft.)
25	Weight (pounds)
26	Domestic? [0= non-U.S. manufacturer, 1= U.S. Manufacturer]