**Linear Models** MATH EPFL
Winter 2015 – 2016 V.M. Panaretos

---

**PRACTICAL 1 - October 26, 2015** (due 10:15 am, 16/11/2015)

---

## Purpose

The purpose of this first practical is two-fold:

1. Gaining familiarity with the various practical aspects involved in fitting a linear model in `R`. These include fitting the model, assessing the appropriateness of the model, making statistical inferences, and critically commenting on the results.
2. Making a first attempt at writing a report resembling the format(s) employed by the scientific community.

## Data

This practical considers an automobile mileage dataset. You will find the dataset on the course website `http://smat.epfl.ch/courses/regression.php`. The data come from the April 1993 issue of U.S. consumer reports. We have data on 82 cars. For each car we have 26 variables. Table 1 gives a full description of each of these variables. The data are available on the website as an `R` data frame (`cars.RData`, to load this in `R` use `load("cars.RData")`).

We are interested in which factors have an effect on automobile gas mileage (i.e. the efficiency in the use of gasoline). To this aim, we employ regression methodology, and in particular the Gaussian linear model. For this practical, we single out three variables from the data set: weight, horsepower and city MPG. The basic model we will be investigating is:

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

The response variable $y := \frac{100}{\text{City MPG}}$ is indicative of the fuel efficiency of the car, its interpretation being how many gallons the car consumes to cover 100 miles within a city (i.e. not on a freeway). Since horsepower often depends on weight, which is already included in the equation, we use the variable $\frac{\text{Horsepower}}{\text{Weight}}$ as a measure of the "pure power" of the car. In particular, the latter variable measures "how many horses" the car has per pound.

## Writing your report

When writing a report, you should not answer questions directly (i.e., in the form 1, 2, 3,...) as when solving exercises. Instead make sure your report covers the material discussed in each point, but structure your report as a scientific paper.

Additionally, your should not directly copy-paste to your report any `R` output. You should fit the model using `R` but explain and interpret the results using elegant tables, figures and plain text.

**The course website contains some documents that may be helpful in writing your report.**

Table 1: Names of the variables in the 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] |

Since this is the first practical, we are going to walk you through the process of writing a scientific report. For the second practical, you will be expected to do this on your own.

Your report should start with a summary of your findings. In addition, the report should include the following sections:

**I.** Introduction

**II.** The data

**III.** Methodology

**IV.** Analysis

**V.** Discussion and conclusions

**A** Appendix with your `R` code

## The content of each section in detail

Here is a description of what each section of the report should consist of:

**Summary** This section should consist of one or two paragraphs describing concisely what you did and what you have found. Write this section last!

**I. Introduction** Without using mathematics, describe what you are going to do in the report. For this practical, the introduction should be only one short paragraph and should address the following points:

1. What questions are you trying to answer in the report?
2. What techniques (and why) are you going to use to answer the questions?

**II. The data** Write a brief description of the data including interpretation of the variables used in the analysis based on the information given on the first page of this document. Additionally, do an exploratory analysis of the data. In particular, include boxplots of the response and each covariate, as well as scatter plots of the response versus each of the explanatory variables. Based on these plots, comment on whether a linear model might be appropriate for these data.

**III. Methodology** The methodology section should describe the statistical models and methods you intend to use in the analysis section. Introduce the Gaussian linear model and explain that you will employ such a model in order to answer the questions set forth in the introduction. Propose a Gaussian linear model of the form (1), and explain how conducting inference on the unknown parameters of such a model will help answer the questions set forth in the introduction. Explain what are the estimators that you will be using. Mention the basic distributional properties of your estimators, and how you can construct confidence intervals using these properties. Mention also that you are going to use the `lm` function in `R` (use the `citation()` function to get a citation for `R`). Try to be concise but informative!

**IV. Analysis** Fit the models described in **III.**

1. Provide estimates for the parameters and their standard errors. Comment on the values of the parameter estimates and interpret these within the framework of the problem. Comment on the value of $R^2$.
2. Construct 95% confidence intervals for each parameter, and use these to determine whether one can reject the null hypothesis that the parameter is zero at 5% significance level (individually for each parameter).
3. You have made some assumptions when fitting your model. Do these assumptions appear to be valid? Some plots that may help you are plots of the standardized residuals against the fitted values and against the explanatory variables, the normal QQ plot, and the plot of Cook's distances. Comment on each plot.

4. When you plot the Cook's distances, you will notice that two influential observations stand out from the rest. Based on the other diagnostic plots, the leverages (`hatvalues` in `R`) and your exploratory data analysis, conclude whether these points appear to be outliers or leverage points.
5. Refit the model without the two influential observations (argument `subset` in `lm`) and repeat the previous analysis (including the model diagnostics) for the cleaned data. Comment on these results and compare them with the original fit. Comment also on the leverages you obtain. Comment whether removing the influential observations is a sensible course of action and what dangers it may entail.

**V. Discussion and conclusions**

1. Present the conclusions that may be drawn from the analysis (in the context set forth in the introduction).
2. Comment on any interesting features that arose from the analysis.
3. Mention whether the model has any apparent drawbacks.

**A. Appendix** You should include the `R` code used in solving this practical in the appendix of the report. The code should look clean, should be properly indented and should contain a sufficient number of plain text comments. You may find the environment `verbatim` or the package `listings` useful for typesetting your code.

## Specifications for the report

1. **This report must be written in groups of 2.**

2. You can write the report in French or in English.

3. We strongly suggest that you write the report in LaTeX. The file "practicalTemplate.zip" on the course webpage might be helpful. If you are not familiar with LaTeX, here is a tutorial: `tobi.oetiker.ch/lshort/lshort.pdf`, and some free software that include a text editor are Texmaker (Mac), Miktex (PC) and Kile (Linux).

4. All figures and tables should be numbered and have a caption briefly explaining what is in the figure/table. Reference should be made to each figure/table from within the text. The captions of tables are placed above the table, while the figure captions go below the figure.

5. Please **read you report carefully** before handing it in and use a spell checker to find any spelling mistakes. The report should be elegant and self-consistent. An excessive number of spelling mistakes, typographic issues, inconsistencies and other careless mistakes will affect your grade negatively.

6. Mention any references you have used and provide a detailed bibliography. References should be made to scientific articles or books and *not* to the lecture notes of the course. We strongly suggest using BibTeX for building the bibliography.

7. Pasting plain computer output is not acceptable.

## Some useful advice

Here are some tips that may help you writing your code in `R`:

- You should do Exercise 4 of Problem set 1 before solving this practical.
- Do not write your code directly to the `R` command line. Create instead `R` scripts and source them. You may also write your own `R` functions if you consider that to be useful.
- We highly recommend performing the data analysis in RStudio (available for Windows, OS X and Linux at `www.rstudio.com`)
- Some commands that you may find useful (use `help()` or `?` to find the documentation): `lm`, `confint`, `model.matrix`, `fitted`, `rstandard`, `qqnorm`, `qqline`, `cooks.distance`, `hatvalues`, `abline`, `identify`, `text`
- Use the `par` command to create figures containing multiple plots. For example, to create a 2 × 3 matrix of plots, write:

```
opar <- par(mfrow = c(2,3))
# Plot commands
par(opar)
```

- Pay attention to the quality of your plots. Export them from `R` using the `pdf` command; you can also play with its arguments `width` and `size`. Here is an example:

```
pdf(file="plot.pdf", width = 7, height = 6)
# Plot command(s)
dev.off()
```

And here is some advice that you may find useful when writing the report:

- You should under no circumstances directly copy-paste or translate material from some other source. Express your ideas using your own words!
- Include a table with the variables in the dataset, such as Table 1. You can find a tex version of it (in English) on the course website.
- Put the estimates and the confidence intervals in a table.
- Use a special font for variables/predictors — it helps the reader. For instance, you could use “`\texttt{}`” in LaTeX, which gives: “. . . in conclusion, we notice that the variable `age` has a significant impact on . . . ”
- **Re-read your work!** Finish a couple of days before the deadline, do other things for 1-2 days, then read again your report with a fresh mind: you’ll spot many small mistakes, and it will really improve the quality of your work!

## Finally, the most important remarks:

The report should be maximum 14 pages long (excluding the appendix), in 12pt font. It should be printed and deposited in the box in front of office MA B1 493 before *Monday November 16, 10:15 am.* Alternatively, the report can be handed in at the beginning of the exercise session on that day.

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