# SB1: 1st Assessed Practical

Week 5, MT 2019

- This practical sheet contains two sections. **Only the exercise in Section 2 is assessed** – it contributes 8.5% to your raw SB1 total mark.
- Write your answer to the exercise in Section 2 as a report.
- The deadline for submission, which is officially published in the Part B synopses, is:

  **12 noon Monday week 8, Michaelmas Term 2019,**
  **at the Statistics Department reception, 24–29 St Giles'.**

- On the cover page of your report:

  **please use your candidate number, not your name (nor your student number)**.

Your report should be clearly written. There are no marks awarded for presentation but there are marks awarded for clarity. When you make a statistical test do not just report the p-value but also report your conclusion using plain language. The same holds for interpreting models; reporting the coefficients and their standard errors is required but try to link your results to the research question as well. You should use captions for your tables and figures and include your commented R-code, preferably in an appendix.

**HINTS**

- You must hand-in a paper copy of your report by the deadline. Note that hand-in is at the Department of Statistics.
- Your report must be accompanied by a completed declaration of authorship form (on paper) saying that the report you are submitting is your own work. The declaration form will be available on the course material page:
  `http://www.stats.ox.ac.uk/current_students/bammath/course_material`
- There are some tips about writing reports, an example practical report and accompanying R code on my SB1 practicals page:
  `http://www.stats.ox.ac.uk/~laws/SB1/SB1_practicals.html`
- Your report should probably be between 5 and 10 pages including any figures/tables. The best way to include your R code is as an appendix (not as part of the report). If you need, an extra page or two for R code is ok, but you should aim for the main report to be no more than 10 pages – part of writing clearly is writing concisely.
- Your report doesn't need to be in LaTeX, but you can of course use LaTeX if you wish.
- There is plenty of time until the deadline, nearly 3 weeks. The report should only take a fraction of this time. But please don't leave writing your report to the last minute.

# 1 Examples for practice, NOT ASSESSED

### (a) A quadratic term

If you want e.g. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, then you should use `y ~ x + I(x^2)`. It is important to use `I()` here.

```
## one of the introductory examples
plot(dist ~ speed, data = cars)

cars0.lm <- lm(dist ~ speed, data = cars)
cars1.lm <- lm(dist ~ speed + I(speed^2), data = cars)
cars2.lm <- lm(dist ~ speed + speed^2, data = cars)

summary(cars0.lm)
summary(cars1.lm)
summary(cars2.lm)
## cars2.lm is the same as cars0.lm and is probably not what was intended
```

### (b) Box-Cox transformation

Suppose a normal linear model applies not to $y$, but to some power of $y$, say to $y^\lambda$. We can use the Box-Cox method to find the best value of $\lambda$. Where possible we might hope for an interpretable value of $\lambda$. Faraway (2015): "If explaining the model is important, you should round $\lambda$ to the nearest interpretable value."

As $\lambda$ varies in the range $(-2, 2)$ we get the inverse transformation ($\lambda = -1$), square and cube roots ($\lambda = \frac{1}{2}, \frac{1}{3}$), the original scale ($\lambda = 1$), as well as the squared case ($\lambda = 2$). We want a sensible $\lambda = 0$ case as well, so the method actually works with the transformation to $y^{(\lambda)}$ where

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0. \end{cases}$$

Note this is consistent because $\lim_{\lambda \to 0} \left( \frac{y^\lambda - 1}{\lambda} \right) = \log y$.

We assume all $y_i$ values satisfy $y_i > 0$ (if not we could add a small constant to all $y_i$s).

We can treat $\lambda$ as a parameter and find the MLE: see Davison (2003, p389–390), or Faraway (2015, p134–137) for details.

### (c) Interactions

```
# an example from lectures
data(whiteside, package = "MASS")
gas2.lm <- lm(Gas ~ Temp * Insul, data = whiteside)
```

The term `Temp * Insul` in `gas2.lm` is shorthand for `1 + Temp + Insul + Temp:Insul`

In particular, the term `Temp:Insul` is the interaction between `Temp` and `Insul`

If we had three variables, say `a`, `b` and `c`, then for a model involving the main effects of `a`, `b` and `c`, plus all of the two-way interactions `a:b`, `a:c` and `b:c`, we can use the shorthand

```
y ~ (a + b + c)^2
```

which is the same as

```
y ~ 1 + a + b + c + a:b + a:c + b:c
```

(and for a model that also includes the three-way interaction `a:b:c` as well, use `y ~ a * b * c`).

**(d) Electrodes**

The `electrode` data give skin resistance (a positive quantity) in ohms measured using five different electrodes on sixteen subjects. The experiment was designed to search for differences in response across electrodes. The electrodes were applied to randomly chosen areas of skin on the upper right arm of each subject.

In the accompanying R-file `RnotesForElectrodes.R` we perform exploratory data analysis and give a brief summary of the data. We model the relation between `Resistance` and the available explanatory variables using a normal linear model for the response. We have to transform the response to make it normal (using the Box-Cox method outlined above). We carry out outlier analysis and model selection.

`http://www.stats.ox.ac.uk/~laws/SB1/RnotesForElectrodes.R`

## 2 ASSESSED EXERCISE

The data in the file `restr.csv` are concerned with price of restaurant meals in a city.

Each row of the dataset corresponds to one restaurant. The response variable `price` is the price per person of a meal in the restaurant. The available explanatory variables are:

- `food`: customer rating of the quality of the food in the restaurant (out of 30)
- `decor`: customer rating of the quality of the decor of the restaurant (out of 30)
- `service`: customer rating of the quality of service of the restaurant (out of 30)
- `guide`: a categorical variable (with levels `A`, `B` or `C`) indicating whether the restaurant is given an A, B or C rating in a certain guidebook
- `location`: an indicator of whether the restaurant is in the north or the south of the city (`0` = north, `1` = south).

*Write a report on the exercise below.*

**Exercise:**

You are asked to investigate how `price` depends on the available explanatory variables. The aim is to: (i) perform some exploratory analysis; (ii) obtain a model that explains how `price` depends on the other variables; (iii) interpret the model you obtain.

1. Perform an exploratory analysis of the data and summarise your findings – give a brief summary of the problem and data. You may wish to consider some numerical summaries as well as some exploratory plots.
2. Model the relation between `price` and the available explanatory variables using a normal linear model for the response. (Stick to normal linear models, with fixed effects.)
3. It is of interest to investigate:
   - which of the predictor variables has the largest estimated effect on price
   - whether the effects of any of `food`, `decor` or `service` differ according to the `location` of the restaurant.
4. Interpret the model you obtain, comment on your findings.

```
# to load the data from the web
restr <- read.csv("http://www.stats.ox.ac.uk/~laws/SB1/data/restr.csv")
# and look at the first few rows
head(restr)

# maybe better is to save a copy of restr.csv and load it:
#
# if you have saved a copy of restr.csv you could use:
# restr <- read.csv(file.choose())
```

```
# and then choose the restr.csv file
#
# or you could use something like:
# restr <- read.csv("../data/restr.csv")
# where you need to specify where to find the restr.csv file
```