Workshop 2

1. Consider some data for deformation (in mm), y_i, of 3 different types of alloy, under different loads (in kg), x_i. When there is no load, there is no deformation, and the deformation is expected to vary linearly with load, in exactly the same way for all three alloys. However, as the load increases the three alloys deviate from this ideal linear behaviour in slightly different ways, with the relationship becoming slightly curved (possibly suggesting quadratic terms). The loads are known very precisely, so errors in x_i's can by ignored, whereas the deformations, y_i, are subject to larger measurement errors, that do need to be taken into account. Define a linear model suitable for describing these data, assuming that the same 6 loads are applied to each alloy, and write it out in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

2. Suppose that $\mathbf{y}$ is a response vector and $\mathbf{x}$, $\mathbf{v}$ and $\mathbf{z}$ are vectors of predictor variables. What is wrong with testing

$$\mathrm{H}_0 : y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \epsilon_i, \quad \epsilon_i \text{ i.i.d. } N(0, \sigma^2),$$

against

$$\mathrm{H}_1 : y_i = \beta_1 + \beta_2 v_i + \beta_3 z_i + \beta_4 z_i^2 + \epsilon_i, \quad \epsilon_i \text{ i.i.d. } N(0, \sigma^2),$$

using the F-ratio results from section 4.3.2 of the notes?

3. A statistician has fitted two alternative models to response data y_i. The first is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{1}$$

and the second is

$$y_i = \beta_0 + \beta_1 x_i + \gamma_j + \epsilon_i \quad \text{if } y_i \text{ from group } j. \tag{2}$$

In `R` the factor variable containing the group labels is `trt`. The statistician wants to test the null hypothesis that model (1) is correct against the alternative that model (2) is correct. To do this, both models are fitted in R, and a fragment of the summary for each is shown below.

```
summary(b0)
lm(formula = y ~ x)
...
Residual standard error: 0.3009 on 98 degrees of freedom

summary(b1)
lm(formula = y ~ x + trt)
...
Residual standard error: 0.3031 on 95 degrees of freedom
```

In `R` this test could be conducted via `anova(b0,b1)`, but instead perform the test using just the information given.

4. A 'quant' working for a prestigious international financial firm proposes the following model for predicting the half hour return rate, r_i, on a particular short term trade, based on simple market index and volatility data:

$$r_i = \beta_0 + \beta_1 \cos(2t_i\pi/48) + \beta_2 \sin(2t_i\pi/48) + \beta_3 \mathtt{DJ}_{i-1} + \beta_4 \mathtt{DAX}_{i-1} + \beta_5 \mathtt{FT}_{i-1} + \beta_6(\mathtt{DJ}_{i-1} - \mathtt{DJ}_{i-2})^2 + \beta_7(\mathtt{DAX}_{i-1} - \mathtt{DAX}_{i-2})^2 + \beta_8(\mathtt{FT}_{i-1} - \mathtt{FT}_{i-2})^2 + \epsilon_i \quad (3)$$

where the ϵ_i are i.i.d. $N(0, \sigma^2)$, r_i is the return in the i^{th} half hour period, t_i is the time measured in half hours from some arbitrary start point, $\mathtt{DJ}_i$, $\mathtt{DAX}_i$ and $\mathtt{FT}_i$ are the state of the Dow Jones, DAX and FT100 indices in the i^{th} half hour period. The $\sin$ and $\cos$ terms are there to model daily fluctuations. The model is to be estimated by least squares using data for the fortnight leading up to the trading half hour of interest. There is money to be made if r_i can be predicted along with accurate assessment of the prediction uncertainty. In test runs the prediction accuracy seems reasonable.

What is likely to be wrong with this model, and what residual plot would you use to show the problem up?

5. A distillery sets up and sponsors a hill race dubbed the 'Whisky challenge', for promotional purposes. To generate extra interest from elite fell runners in the first year, it is proposed to offer a prize for every runner who completes the course in less than a set time, T_0. The organisers need to set T_0 high enough to generate a big field of participants, but low enough that they do not bankrupt the distillery. To this end they approach you to come up with a predicted winning time for the race. To help you do this, the `hills` data frame in R package `MASS`, provides winning times for 35 Scottish hill races. To load the data and examine it type `library(MASS);hills` in R. Find and estimate a suitable linear model for predicting winning `times` (in minutes) in terms of race distance `dist` (in miles) and the total height climbed `climb` (in feet). The 'Whisky Challenge' is to be a 7 mile race, with 2400 feet of ascent. Produce a short report for the race organisers. *Note*: It can be instructive to look at 'Naismith's rule' on Wikipedia.

6. This question is about model checking, transformation and using factor variables in `R`. You will use the `InsectSprays` data supplied with R. The data frame contains `counts` of pest insects on plots of an agricultural experiment which had each been treated with one of 6 insecticide formulations (the factor variable `spray` indicates which). Interest is in whether formulation makes a difference to the final count, and if so, which formulation is best.

   (a) Produce a plot of counts against spray type. How would you interpret the plot?

   (b) Fit a linear model to the data, with `count` as the response and `spray` as the predictor and check the residuals against the fitted values (or the square root of the absolute value of the standardised residuals against the fitted values). Is the constant variance assumption of the linear model met?

   (c) Is the normality assumption tenable for this linear model?

   (d) Count data are often more naturally modelled as following a Poisson distribution, rather than a normal. Using the Normal as an approximation to Poisson (or a transformation of Poisson) is usually OK, provided that we can somehow fix the problem that Poisson data do not have constant variance. Transformation can often do this (at least, approximately). It can then be shown that we should be able to approximately meet the constant variance assumption by using square root of `count` as the response, rather than `count`. Please try this. Does it improve matters?

(e) Now look at a `summary` of your model. What is the interpretation of the `(Intercept)` parameter here? What about the other parameters?

(f) Sometimes it is necessary to change the 'reference' level of a factor variable. For example, suppose that you wanted to compare all the formulations to formulation `C`. As shown in the notes, you can do this by using the `relevel` function:

```
InsectSprays$spray <- relevel(InsectSprays$spray, "C")
```

and refitting the model. Please try it and look at the summary output again: how has it changed? What would you conclude about the data now?