STAT3015/7030 Generalised Linear Modelling

Tutorial 1

1. By the time you attend this tutorial, we will already have discussed a number of examples using *R* in lectures. For all the examples that have been covered in lectures, the *R* command files used and any associated data files will be available on Wattle.

   A good approach to regular study in this course will be to download these examples (and any associated data files) and review them, using *R*, on your own computer or on one of the ANU InfoCommons computers. In the first half of this tutorial, do this for one of the recent examples covered in lectures and ask your tutor for help when you encounter problems or cannot understand the details of the example. If the first half of the current tutorial is the first time you have attempted to do this, you will probably only have time to get started on one of the examples, and you will have to review the other examples in your own time, but you can still ask questions in some later tutorial, via the Questions and Answers forum on Wattle, by e-mail, during one of the advertised consultation times or by making an appointment to see me.

   Each time you review one of the examples, ask yourself the following questions:

   (a) Can you get the *R* codes to execute correctly and show you the output that was discussed in lectures and/or as discussed in the "brick" of lecture notes? If not, what modifications do you need to make? Note one of the hardest parts of mastering any statistical computing package is importing and managing data. There is a good manual on "R Data Import/Export" on the *R* Project website (https://cran.r-project.org/manuals.html) and there are also a few tips about the modifications you may need to make in the section on "Additional Revision Examples using *R*" in the Revision Topic on Wattle. Once you have got the code working, execute the code in small blocks or line-by-line; then examine the output and read the comments in the *R* command file.

   (b) Each example covers an analysis of some dataset and typically fits a series of models to the data. What is the underlying research question that we are trying to address with this analysis of the data?

   (c) Typically the analysis will arrive at a model that is in some way the "best" model for the data, given the research question we are trying to address. Review the model development process in the analysis and identify this "best" model and check if

it is a good fit to the data. In particular check whether all of the underlying assumptions have been satisfied or if there are still some questions about any of the underlying assumptions or about the status of particular observations. Any unresolved questions will qualify the usefulness (i.e., impose some caveat on) of the model to address the research question. Also check if the model is simple and parsimonious (see if there any insignificant terms). Why have these terms been left in the "best" model? Are they necessary in order to use the model to address the research question?

(d) Finally, have a think about how you might use the model to address the research question and how you might present the results to the "clients" (the people who brought the data and the research question to you, as the statistical consultant).

2. This question was adapted from Exercise 4 of Chapter 3 of Gelman and Hill (2007), Page 50. Now have a go at modifying the R code in one of the revision examples to perform a new multiple regression analysis.

   The file `child.iq.dta` (available on Wattle) contains data on the cognitive test scores of 3 and 4 year old children and data on the characteristics of the mothers of the children (these data were discussed as an example earlier in the same chapter in Gelman and Hill (2007)). You have access to children's test scores at age 3, mother's education, and the mother's age at the time she gave birth for a sample of 400 children. The data are in a `Stata` file which you can read into R by saving the file in your working directory and then typing the following:

```
library(foreign)
child.iq <- read.dta("child.iq.dta")
```

   (a) Fit a regression of the child test scores (`ppvt`) on the mother's age (`momage`). Display the data and fitted model on a graph, check the assumptions, and interpret the slope coefficient. When do you recommend mother should give birth? What are you assuming in making these recommendations?

   (b) The other variable you have access to is `educ_cat`, which is a categorical variable for mother's education with four categories (1 = "primary school", 2 = "high school", 3 = "bachelor", 4 = "post graduate"). We will be considering how to correctly treat such variables as explanatory variables in lectures over the next couple of weeks, but at this stage it will not hurt to have a preview of the R commands and output. Fit the regression of `ppvt` on `factor(educ_cat)` and produce residuals plots, the ANOVA table and a summary of the coefficients for this regression model. Now also include mother's age in the model and again produce the same output for this modified model. Do you think mother's education is having an effect on the children's test scores, adjusting for the effects of mother's age?

(c) Now create an indicator variable reflecting whether or not the mother has completed high school. Add this indicator variable to the regression model from part (a) and Page 2 also consider interactions between the high school completion and mother's age at birth. Also, create a plot that shows the separate regression lines for the two different levels of high school completion.

(d) Finally, fit a regression of child test scores on mother's age and high school completion for the first 200 children and use this model to predict the test scores for the next 200. Graphically display comparisons of the predicted and actual scores for the final 200 children and comment on your results. One way to do this would be to use the `predict()` function, but that would require you to remember a bit more of the *R* code used in the prerequisite course (STAT2008/STAT6038 Regression Modelling) than was covered in the revision examples, so do not worry if you cannot do it at this stage of this course, as we will revisit the `predict()` function later in this course.

References

Gelman, A. and Hill, J. (2007), Data analysis using regression and multilevel/hierarchical models, Cambridge University Press New York, NY, USA.