G12SMM Statistical Models and Methods

Linear Models, Assessed Coursework 1 — 2016/2017

Please hand in your work by 3.00pm on Wednesday 15 March to a student services centre. Please remember to complete and attach a coursework submission sheet to your work. Your solutions should contain all relevant plots and `R` output needed to justify your answers/arguments, together with appropriate discussion, but please do not include pages of irrelevant plots/output which you do not discuss. The easiest way to include `R` output is to use R Markdown to produce your solutions, but you do not have to do so. You do not need to include your `R` code, though you can include it if you wish. If you are using R Markdown, and do not wish to include your `R` code, then you can suppress the `R` code using the `echo = FALSE` argument, i.e. enclose the code
in an {`r, echo=FALSE`} environment in the Markdown file.

There will be a Moodle forum specifically for answering queries about the coursework, so you may post questions and I will answer them there so that everyone receives the same assistance. Please be careful to not inadvertently give away parts of your answer if you do post a question. Note that as this is assessed work, I can only answer queries relating to clarification, and I will only answer queries via the forum so that everyone can see my responses.

Unauthorised late submission will be penalised by 5% of the full mark per day. Work submitted more than one week late will receive zero marks. You are reminded to familiarise yourself with the guidelines concerning plagiarism in assessed coursework (see the student handbook), and note that this applies equally to computer code as it does to written work.

You can answer the questions one-by-one as you would normally answer an exercise sheet, and in particular there is no need to produce a formal "report". However, you should explain carefully your conclusions about the practical questions being asked, and how they follow from the models you fit, and include all relevant output to support your conclusions.

The Data

You are a medical statistician who has been tasked with investigating associations between the birthweight of children and various potential explanatory variables. Data are available regarding the birthweight of 427 children, together with various other measurements. The data are contained in the file `BirthData.txt` on Moodle. The variables are:

`age`	Age of mother.
`gest`	Gestation period.
`sex`	Sex of child.
`smokes`	Whether the mother smoked during pregnancy, with levels 'No', 'Light' and 'Heavy'.
`weight`	Pre-pregnancy weight of mother.
`rate`	Rate of growth of child in the first trimester.
`bwt`	Birthweight of child.

You can read the data into R (after saving the file in your working directory) using

```
Births = read.table("BirthData.txt",header = TRUE)
```

The variables 'smokes' and 'sex' should be treated as factors, the rest as numerical variables. After reading in the data, you should first check that R is treating each variable as intended, and change this behaviour if necessary.

Interest lies in determining the variables associated with birthweight, which can then be investigated further to understand any possible causal relationships.

The Tasks

(a) Produce a pairwise matrix scatterplot of the continuous variables. Comment on the relationship between birthweight and the other variables, and any other notable relationships between the variables. For the two factor variables, produce suitable plots for exploring the relationship between birthweight and the different levels of the factor. **[7]**

(b) Fit an appropriate one-way ANOVA model to test for an association between birthweight and sex of the child, and give the estimated parameters. Use the command `model.matrix` to check the coding of the factor used by R, and hence provide an interpretation of your parameter estimates. **[6]**

(c) Consider fitting a one-way ANOVA model of birthweight against smoking level. Explain why the following form of model is not appropriate:

$$\mathbb{E}[\texttt{bwt}] = \begin{cases} \mu + a & \text{if } \texttt{smokes} = \text{"No"}, \\ \mu + b & \text{if } \texttt{smokes} = \text{"Light"}, \\ \mu + c & \text{if } \texttt{smokes} = \text{"Heavy"} \end{cases}$$

Hint: Consider the form of the design matrix Z corresponding to this model, and determine its rank. **[5]**

(d) Now fit an appropriate one-way ANOVA model to test for an association between smoking level and birthweight. What conclusions do you draw? **[5]**

(e) Not surprisingly, gestation period is strongly associated with birthweight. Is there evidence of an association between sex and birthweight, after controlling for gestation period? **[5]**

(f) For the model you fitted in (e), use R to calculate an unbiased estimate of σ^2 (the variance of the errors in the model) and the variance-covariance matrix of the estimated parameters. What is the deviance of the model fit? **[5]**

(g) Using the appropriate model, find 95% confidence and prediction intervals of the birthweight of a female child whose mother is 30 years old and the gestation period is 275 days. **[5]**

(h) Finally, you now have free reign to model the dependence of birthweight on any or all of the other variables. Investigate possible models, and choose the model(s) which you feel are most appropriate, justifying your choice with any relevant quantities and plots. Interpret the parameter estimates from your chosen model(s), in order to summarise the variables associated with birthweight and the nature of the associations (i.e. is the variable associated with increase/decrease in birthweight?).

HINT: It is sufficient to only consider models with additive terms and not interactions - i.e. models of the form $y = a + b_1x_1 + b_2x_2$. Interactions occur when a covariate is defined as the product of individual covariates - e.g. the model $y = a + b_1x_1 + b_2x_2 + b_3x_1x_2$, which includes an interaction between x_1 and x_2 as well as the individual (additive) terms. (Note that such models are still linear though, since they are linear in the *parameters*.) You do not need to consider interactions here, so the number of possible models is quite small. One tactic is to first fit the (additive) model with all terms included, and consider which terms could be removed (one at a time), either by looking at the variable with largest (non-significant) p-value, or looking for significant reductions in deviance between nested models. **[12]**