STATS321-19B – Assignment 2 Due: 11:55pm, Monday 2nd September 2019
(50 marks total)

Please prepare answers to these questions in a Word document (or similar word-processed document) and submit the document to the appropriate link on Moodle.

Question 1 (25 marks)

Consider the data in the dataframe `phbirths` available from the `faraway` package in `R`.

This dataset concerns the birth weights of 1115 babies in the US state of Philadelphia in 1990. The variables recorded are:

`grams`	A baby's birth weight, in grams (the response variable).
`black`	Whether the birth mother was black (boolean).
`educ`	The number of years the birth mother had received education.
`smoke`	Whether the birth mother smoked during pregnancy (boolean).
`gestate`	A baby's gestational age at birth, in weeks.

Use linear modelling and any techniques you have been taught in this course (or other techniques, if you want!) to model the birth weight of babies as a function of the other predictor variables in this dataset. Describe what you are doing at each stage, even if you are applying automated techniques like stepwise regression. Include plots and regression output where appropriate.

The aim of this exercise is not necessarily to come up with the "right" model for the data (hopefully some of you will settle on the same model as I have, but quite possibly some of you will find better models!). Rather, the aim is to demonstrate an understanding of the model-fitting process.

Techniques you might consider include: interactions, polynomial terms for predictor variables, transforming the response or predictor variables, centring predictor variables, stepwise regression, excluding influential observations, F-tests for model selection, diagnostic plots etc. However, you do not have to use all of these ideas to get full marks.

Provide an interpretation of your results. Remember, as statisticians, we often like to express our uncertainty in point-estimates. Some context specific questions you might like to consider:

- Is there a difference in the average weight of children born to black and non-black mothers?
- Does smoking during pregnancy affect a baby's birth weight?
- Is it sensible to use this model to describe the rate at which a baby puts on weight over a pregnancy?
- Are there any extremely premature babies (the typical gestational period for humans is

40 weeks) who influence the fit of this model, and should they have this influence?

- Are the effects of any variable different for different levels of another variable e.g. does the longer a mother experiences in education have the same relationship with birth weight for black and non-black mothers?

You might like to illustrate some of your findings with tables and/or graphs.

Question 2 **(25 marks)**

Consider Finney's 1947 vasoconstriction data in the file `finney2.txt` available on Moodle. It has been found that a single deep breath can cause the blood vessels in the skin of the fingertips to constrict. Although this phenomenon could be detected, it was not practicable to measure the degree of constriction (in 1947). Therefore the response (`Response`) is binary. The data was collected to examine how the incidence of vasoconstriction was affected by the rate of inhalation (`Rate`, in litres per second) and the volume of air inhaled (`Volume`, in litres) in the breath.

Note: the data were collected as multiple measurements from three separate individuals. Technically, we should include the individuals as a predictor in our model, because there could be characteristic differences between individuals. The results might also not be quite as generalisable as we might have supposed, based on the "sample" size. However, the appropriate modelling technique to deal with this situation is beyond the scope of this course however, and most examples based on this data ignore this complexity.

Show your commands and output.

(a) Produce a scatterplot of the data with `Volume` on the x-axis and `Rate` on the y-axis, using different symbols for the different responses.

(b) Suggest why there is a lack of observations in the top-right quadrant of the plot.

(c) Fit binomial regression models to the data using the `Volume` and `Rate` variables as predictors, and the following link functions:

   i) the logit link

   ii) the probit link

   iii) the complementary log-log link

(d) Which of the three models seems to fit the data best? Can you say that this model is significantly better than the other models?

(e) Repeat part (c), for all three link functions, but this time using the log of `Volume` and the log of `Rate` as your predictor variables.

(f) Which of these three models seems to fit the data the best? Does the best model from this set seem to be superior to the best model using the untransformed predictor variables?

For the rest of this question, consider just the *two* models you have created with the **logit link function**.

(g) Use the `predict()` function to predict the probabilities of vasoconstriction under *both* logit models for the following cases:

i) Someone who takes a breath of 1 litre at a rate of 0.9 litres per second

ii) Someone who takes a breath of 3.1 litres at a rate of 1.2 litres per second

iii) Someone who takes a breath of 0.4 litres at a rate of 2.6 litres per second

Hint: probabilities have to be values between 0 and 1

(h) What are the p-values for the goodness-of-fit statistics for these two logit models?

(i) Indicate the combinations of `Volume` and `Rate` that lead to an estimated probability of vasoconstriction of 0.1 for both logit models on the plot from part (a). Also plot the three points used for the predictions in (g).
Hint: since we are fixing the value of the response we are looking for at 0.1 but we have two unknowns (Volume and Rate) choose a set of values for one of these unknowns then, for each value in that set work out from the estimated linear equation for the link function the corresponding values of the other unknown that produce the required response value of 0.1.