RESEARCH SCHOOL OF FINANCE, ACTUARIAL STUDIES AND STATISTICS
REGRESSION MODELLING
(STAT2008/STAT2014/STAT4038/STAT6014/STAT6038)
Assignment 2 for Semester 2, 2020

INSTRUCTIONS:

- This assignment is worth 20% of your overall marks for this course.
- The data file is on Wattle.
- Please name your submission "Name Uid".
- Please submit your assignment on Wattle. When uploading to Wattle you must submit the following, combined into a single pdf document:
  1. Your assignment/report.
  2. An appendix including the R codes you used. Failure to upload the R code will result in a penalty.
- Please do not include the cover sheet as it caused errors in Turnitin in Assignment 1. Instead, use the "Assignment 2 cover sheet" link on Wattle.
- Assignments should be typed. Your assignment may include some carefully edited computer output (e.g. graphs, tables) showing the results of your data analysis and a discussion of these results, as well as some carefully selected code. Please be selective about what you present and only include as many pages and as much computer output as necessary to justify your solution. Clearly label each part of your report with the part of the question that it refers to.
- Unless otherwise advised, use a significance level of 5%.
- Marks may be deducted if these instructions are not strictly adhered to, and marks will certainly be deducted if the total report is of an unreasonable length, i.e. more than 10 pages including graphs and tables. You may include an appendix that is in addition to the above page limits; however the appendix will not be assessed. It will only be used if there is some question about what you have actually done.
- Late submissions will attract a penalty of 5% of your mark for each day of delay. No assignments will be accepted 10 days beyond the due date.
- Extensions will usually only be granted on medical or compassionate grounds on production of appropriate evidence, but must have my permission by no later than 24 hours before the submission date. If you are granted an extension and submit your assignment after the extended deadline then the late submission penalty will still apply.

Question 1 [97 Marks]

We have cross-section data originating from the May 1985 Current Population Survey by the US Census Bureau. These data consist of a random sample of 534 workers, with information on wages and other characteristics of the workers. We are interested in the relationship between the wage of a person and these characteristics. Please use the data set called "Wage.csv" on Wattle. The variables considered are:

`wage` Wage (in dollars per hour).

`education` Number of years of education.

`experience` Number of years of potential work experience (age - education - 6).

`age` Age in years.

`gender` Factor indicating gender.

`occupation` Factor with levels "worker" (tradesperson or assembly line worker), "technical" (technical or professional worker), "services" (service worker), "office" (office and clerical worker), "sales" (sales worker), "management" (management and administration).

`married` Factor. Is the individual married?

Throughout this assignment, you should use the log-transformed `wage` as the response. You do not need to center the variables.

You could use the following code to import the data:

```
read.csv("Wage.csv", stringsAsFactors = T)
```

(a) [5 marks] Fit a multiple linear regression (MLR) model with `log(wage)` as the response variable and all other numeric variables as predictors. Test whether this model is significant.

(b) [12 marks] What are the estimated coefficients of the (MLR) model in part (a) and the standard errors associated with these coefficients? Interpret the values of each of the estimated coefficients with regards to model specification. Construct 95% Bonferroni joint confidence intervals for all the slope parameters.

(c) [15 marks] Comment on the t-test results in the summary output. Do they contradict with the test result in part (a)? Why or why not? Conduct a diagnostic check for this particular problem with the fitted model both qualitatively and quantitatively. What should be done to solve this problem?

(d) [8 marks] You decide to include in the model only `education` and `experience` as potential predictors. However, you are not sure what kind of marginal relationship

is between `experience` and the response `log(wage)`, given that `education` is already included in the model. Generate an appropriate plot to visually check this relationship and comment on the plot. Then conduct a test whether a second-order term is needed for `experience` given `education` is in the model.

(e) [9 marks] How does marital status affect the wage? Conduct a test of whether married people earn more salary than unmarried people by fitting a simple linear regression model. Then provide a 95% confidence interval on the slope coefficient and interpret this interval.

(f) [8 marks] Construct an appropriate model to test whether married people earn more salary than unmarried people given that `education`, `experience` are controlled. (Include, or not, the second-order of `experience` based on your result in part (d).) Compare the test result with part (e) and comment on the reason of difference if any.

(g) [20 marks] Using the model in (f), produce a plot of externally studentized residuals against fitted values, a normal QQ plot, a leverage plot, a Cook's distance plot and a number of DFBETAs plots for all the slope coefficients in your model. Comment on the model assumptions and unusual points. What are the characteristics of the workers identified as unusual data points?

(h) [5 marks] Now consider the model with only first-order `experience` and `occupation` as the covariates. Based on this model, you would like to know whether there is any difference in the wages of different occupations with the same length of experience. Conduct a test.

(i) [15 marks] With the model in part (h), consider adding the interaction term between `experience` and `occupation`. Generate a scatter plot of `log(wage)` against `experience` and use different colors for different `occupation` levels. Add fitted lines for each `occupation` level in a different color. Comment on the plot whether there is visible interaction. Then test whether the interaction is significant.