STAT4006 Midterm Project

This report/project forms 40% of your total assessment. It is split into two equally-weighted parts. The first part asks you to apply what you learned about contingency tables (the first half of the course); the second ask you to use what you learned about generalized linear models (the second half of the course).

Visit this link: http://lib.stat.cmu.edu/datasets/CPS_85_Wages and download the data. The data is from a 1985 survey of 534 US citizens, with 11 variables recorded for each. The purpose of the study was to investigate what affects a person's wage. For more details, read the summary and the each variable's description.

We label the eleven variables as follows:

EDUCATION → 1
SOUTH → 2
SEX → 3
EXPERIENCE → 4
UNION → 5
WAGE → Y

AGE → 6
RACE → 7
OCCUPATION → 8
SECTOR → 9
MARR → 0

Part I

Enter your student ID number into the red box provided in the spreadsheet "Midterm Project – Variables". The three numbers to the right of the red box are the codes for three variables. Use the data to construct a three-way table of these variables*. Use what you learned in Chapters 1-6 (inclusive) to analyse and comment on this data.

Part II

Treat variable Y (WAGE) as a categorical* response. Ignoring the 3 variables you used in Part I, you have 7 variables left. **Choose four of the remaining seven.** Use what you learned in Chapters 7-10 (inclusive) to model Y with these four variables, analyse and comment on this data.

*Interval variables (like WAGE, AGE, etc.) can be converted into categorical data by grouping observations by intervals. For example in Example 5.4.3 of Chapter 5, the observations of "Alcohol Consumption" would originally have been interval in nature: $\{2.5, 1, 0.5, 7, \dots\}$. By grouping by intervals $0, (0,1), [1,2.5), [2.5,5.5), [5.5, \infty)$ we turn it into an (ordinal) categorical variable. We leave the choice of intervals up to you – best to play safe and make them "reasonable".

Example

Ronald F.'s student ID is 1155???886. The spreadsheet tells him to create a three-way table from variables $7, 8$ and 6 for Part I. RACE and OCCUPATION are categorical, but AGE is interval, so he groups it by intervals $[18,35], [36,50], [51,65], [65, \infty)$. For Part II, he has seven possible variables to choose from: $\{1,2,3,4,5,9,0\}$. The four he chooses are $\{1,3,4,9\}$: EDUCATION, SEX, EXPERIENCE, SECTOR. Response Y (WAGE) is interval, so he groups it by interval.

Deadline

Please submit your projects to link provided on the course Blackboard assignment page by 2359 on Wednesday 23rd December. You may submit as early as you like. Please follow the submission policies as described on the assignment page.