WASEDA UNIVERSITY SCHOOL OF POLITICAL SCIENCE AND ECONOMICS

Quantitative Analysis [2024 Autumn]

Instructor: Rob Fahey <robfahey@aoni.waseda.jp>

WEEK 10: Homework Assignments

For this assignment, you'll need to start with a blank file in R, and figure out from there what code you'll need to solve the following problems. As ever, submit your R file through Moodle to complete the assignment.

Figuring out which of the techniques we have used thus far is the core point of this assignment - it's a test of how well you have understood regression techniques, not a test of your programming abilities, and you won't be graded harshly for programming mistakes.

This assignment is worth twice as much as previous homework assignments, and has an extended deadline of January 5th.

ASSIGNMENT 1: Vocational School Graduate Income

The file **grad_income.csv** contains a large number of observations of the income of vocational school graduates in a certain country. The data includes the gender and age of each person, along with their annual income.

Your task is to analyse this data and find the most appropriate linear model to explain the differences in income.

Some questions you should consider as you're looking for the best model…
- Is the relationship between age and income linear, or might it be described better by a curvilinear model?
- Is the relationship between gender and income independent of the relationship with age, or is there an interaction between these terms?

1) **Find the most appropriate and effective model for this data. Show your workings in your R code (i.e. what other models you tried, and any tests you did to compare models).**

2) **Write a few lines justifying your choice of model.**

3) **Create a graph which illustrates the relationship between age, gender, and income.**

ASSIGNMENT 2: Titanic Survivors

The file **titanic.csv** contains the passenger data from the Titanic, including their survival or non-survival of the ship's sinking. We saw in class that paying more for a ticket increased the chances of survival, as did being female.

Your task is to look at a few additional factors and see how they may have impacted the chances of surviving the Titanic. Bear in mind that many of these columns will need to be turned into factors before they can be effectively used in a regression. Also take careful note of the data in your outcome variable, and make sure you're using the right regression model.

1) **The column `pclass` tells us which class the passenger was in - the highest class is first, with third-class being the cheapest accommodation on the ship. We know `fare` is correlated with survival; we want to know if this is just because the more expensive first-class accommodation was better positioned for survival.**

   Test whether `pclass` is also correlated with survival, and see if `fare` is still significant even once `pclass` is included.

2) **"Women and children first" was the famous policy for boarding lifeboats in shipwrecks at this time. We found that women were much more likely to survive than men, but what about children?**

   To test this, you will need create a new factor variable that separates out women, men, and children (we probably don't care about the gender of children, so we can just treat them as a third separate category).

   To construct this new column, I suggest that you look for information about the ifelse() function (which we discussed a little in class) and the case_when() function. You'll need to use a combination of information from the `age` column and the `sex` column.

   Once it's created, test whether being a woman or a child makes a major difference to survival.

3) **Finally, combine these factors and find the most effective model to explain the odds of surviving the Titanic, based on whatever you think is the most effective combination of ticket price, passenger class, gender, and age. You don't have to include all these factors if you don't think they're all necessary.**

4) **Write a few lines to explain and justify your choice of this model.**