Massey University
College of Sciences

161.122: Statistics

Assessment C, 2020

GENERAL INFORMATION: This assignment is assessed. It is expected that you do all work on the R markdown document supplied below, and that you hand that in. This way we can see your R code inline with output and explanations. Work must be submitted via the dropbox on stream.

EXPECTATIONS: All graphs should have clear axis labelling and legends if needed. The answer alone will not give full marks – you should include explanation (e.g. R code) and/or descriptions of plots alongside your answers. Be sure to give explanations from a statistical perspective (i.e. with reference to the data given) where possible. Marks are available for good presentation, so think carefully about how you present your answers.

DEADLINE: Your work for this assignment must be submitted by **11:55pm on October 21st 2020** on stream.

PLAGIARISM: The work that you submit must be your sole effort (i.e. not copied from anyone else). If you are found guilty of plagiarism you will be penalized.

General Guideline for Assessment C

Start by downloading the 'AssessmentC.Rmd' (click the blue one!) file from Stream, and loading it into RStudio. This is already pre-formatted, and contains some R code blocks in which you can do analyses. The first code block labelled 'setup' is loading the packages needed and getting all the data setup correctly.

You'll need to make sure you've installed the `tidyverse`, `tsibble`, `lubridate`, and `visreg` packages in RStudio. You should be prompted to do this when you load up the 'AssessmentC.Rmd' file.

Make sure you can `Knit` the 'AssessmentC.Rmd' into HTML without error before you start.

EXERCISE C1: Predicting Sony PlayStation 4 (PS4) Sales

The data file '`ConsoleSales.csv`' contains the monthly global sales of Sony PlayStation 4 from Nov. 2013 to Aug. 2020. The **monthly PS4 sales** has been loaded and tidied as a `tsibble` in 'AssessmentC.Rmd'.

1. Conduct an exploratory analysis on the time series. Describe the feature of this time series and explain the possible reasons why the monthly PS4 sales shows this kind of pattern.

2. Fit a regression model with only a linear trend to the monthly global sales of Sony PlayStation 4 between Nov. 2013 and Dec. 2019. Visualise your fitted model and comment on the R summary.

3. Suggest a better regression model for the time series in C1.2. Fit your suggested model, visualise your fitted model and comment on the R summary. [Hint: you shall consider transformation, modifying trends, adding seasonal effects, etc.]

4. Produce model diagnostic plots for your selected model. Are the assumptions of the linear model for time series met? Explain your reasonings.

5. Use your selected models to forecast the PS4 sales for the first six months of the year 2020. Compute the mean square error of your model. Comment on the prediction performance of your model. [PS4 sales for the first six months in 2020 has been loaded in the R code chunk.]

EXERCISE C2: Exploring US Population Survey

The file '`NZIncomes11.csv`' contains a synthetic dataset from the 2011 New Zealand Income Survey. The dataset contains weekly incomes in 2011 (`income`) for a sample of 13291 kiwis with their ethnicities (`ethnicity`: `European`, `Maori`, `Pacific`, `Asian`), genders (`sex`: `male`, `female`), age groups (`agegrp`: 15 to 65 years by 5), highest qualifications (`qualification`: 1, 2, 3, 4 for no qualification, school, vocational, Bachelor or higher, respectively), and weekly work hours (`hours`). This data set has been has been loaded and tidied as a `tibble` in 'AssessmentC.Rmd'.

Of interest is whether there is a difference in between different ethnicities, and whether this is dependent on ages or other factors.

1. Produce plots of the weekly incomes (`dollars`) versus the variables `ethnicity`, `sex`, `agegrp`, `qualification` and `hours`, ensuring you clearly label axes. Comment on the plots, particularly with respect to whether there is (graphical) evidence for an association with the weekly wages.

2. Fit a multivariable linear model for log weekly incomes against the variables `ethnicity`, `sex`, `agegrp`, `qualification` and `hours`. Peruse the R summary and interpret the practical meaning of each significant coefficient at the level 0.05.

3. Produce model diagnostic plots. Are the assumptions of the linear model met? Explain your reasonings.

4. Produce a prediction for the weekly income of a 25 years old Asian man with a Bachelor degree working 40 hours per week along with a 95% prediction interval. Interpret this interval in words that a client may understand.

5. Refit a multivariable linear model for log weekly incomes by adding any interaction term(s) in which you are interested. Peruse the R summary and discuss your findings. [Hint: It is pretty OK if you have no findings!]