RESEARCH SCHOOL OF FINANCE, ACTUARIAL STUDIES AND STATISTICS
REGRESSION MODELLING
(STAT2008/STAT2014/STAT4038/STAT6014/STAT6038)
Assignment 1 for Semester 2, 2020

INSTRUCTIONS:

- This assignment is worth 15% 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. The assignment cover sheet.
    2. Your assignment/report.
    3. An appendix including the R codes you used. Failure to upload the R code will result in a penalty.
- 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 24hours 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 [92 Marks]

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Please use the data set called "CDI.csv" on Wattle. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. The information generally pertains to the years 1990 and 1992. The information on the 17 variables are on the last page:

(a) [10 marks] We are interested in the relationship between `Percent below poverty level` and `Per capita income`. However, there are missing values (represented as "NA"s) in this dataset. First tidy the data by removing the observations with missing values and use the remaining data. What is the sample size? Fit a simple linear regression model with `Per capita income` as X and `Percent below poverty level` as Y. What are the estimated coefficients of the fitted model and the standard errors associated with these coefficients? Interpret the values of these estimated coefficients.

(b) [7 marks] Generate the ANOVA table and test whether the model in part (a) is significant.

(c) [15 marks] Conduct diagnostic checks on the fitted model in part (a). Show the appropriate plots and comment on all the assumptions and unusual observations.

(d) [5 marks] Generate a scatter plot of Y against X and identify on the plot the outliers using their `County` names.

(e) [8 marks] Experiment with applying natural log transformations and square root transformations to one or both of the predictor and the response variable. Select a best model with the help of scatter plots and sample correlation values. Write out the form of the selected model.

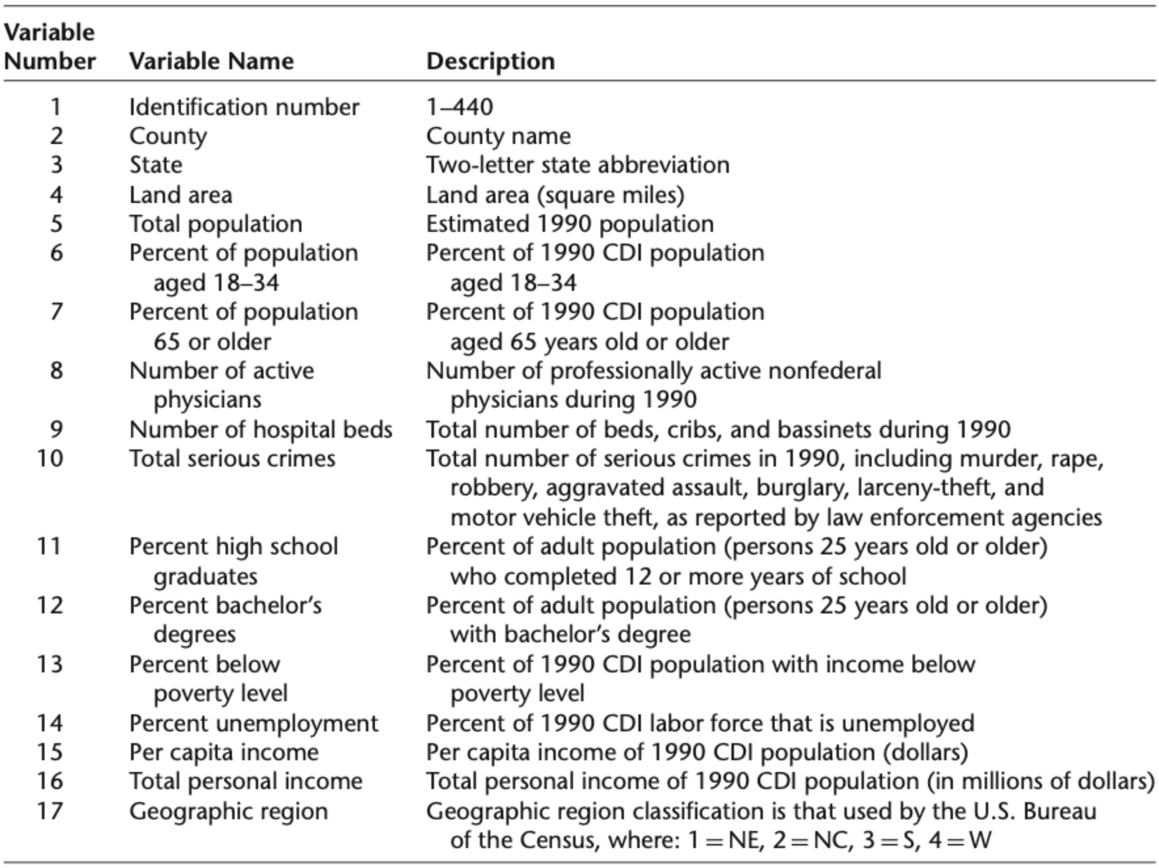
(f) [15 marks] With your chosen transformations, fit a simple linear regression (SLR) model. Conduct diagnostic checks on the selected model in part (d). Show the appropriate plots and comment on all the assumptions and unusual observations.

(g) [5 marks] With the selected model in part (e), test whether the slope is smaller than -1.

(h) [6 marks] With the selected model in part (e), what is the estimated error variance? What is the coefficient of determination value and how to interpret this value?

(i) [15 marks] Write the estimated model in terms of original untransformed variables (both X and Y). Based on the mathematical expression, what happens to $\widehat{Y}$ when the value of X is multiplied by a factor of k? Generate a plot of X and Y on the original scale, along with the fitted model on the original scale.

(j) [6 marks] For a county whose `Per capita income` is 25,000, construct a 90% interval estimate for the mean `Percent below poverty level` of this county. Interpret this interval.



Variable Number	Variable Name	Description
1	Identification number	1–440
2	County	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18–34	Percent of 1990 CDI population aged 18–34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 years old or older
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI labor force that is unemployed
15	Per capita income	Per capita income of 1990 CDI population (dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	Geographic region classification is that used by the U.S. Bureau of the Census, where: 1 = NE, 2 = NC, 3 = S, 4 = W