DATA 303/473 Assignment 1

Due 1159pm Thursday 17 March 2022

Instructions

- Prepare your assignment using Rmarkdown
- Submit your solutions in two files: an Rmarkdown file named e.g. `a1.Rmd` and the PDF file named `a1.pdf` that results from knitting the Rmd file. The Rmarkdown file (`tut1sol.Rmd`) used to create the Tutorial 1 Solutions is provided in the Tutorial Solutions section of Blackboard as an example for you to follow should you wish.
- The YAML header of your Rmarkdown file must contain your name and ID number in the author field, and should have the output format set to `pdf_document`. For example:

```
---
title: "DATA 303/473 Assignment 1"
author: "Nokuthaba Sibanda, 301111111"
date: "Due: 17 March 2022"
output: pdf_document
---
```

- While you are developing your code you may find it easiest to have the output set to `html_document`, but change it to `pdf_document` when you submit.
- In your submission, embed any executable R code in code chunks, and make sure both the R code and the output is displayed correctly when you knit the document.
- If there are any R code errors, then the Rmarkdown file will not knit, and no output will be created at all. So if you can't get your code to work, but want to show your attempted code, then put `error=TRUE` in the header of the R code chunk that is failing.

~~~
```{r, error=TRUE}
your imperfect R code
```
~~~

- **You will receive an email confirming your submission. Check the email to be sure it shows both the Rmd and PDF files have been submitted.**
- Title each question answer with its question numbers as `Q1.`, `Q2`,... instead of `1.`,`2.`,....
- Where you are asked to perform a hypothesis test, state the hypotheses being tested and give the test statistic, p-value and conclusion.

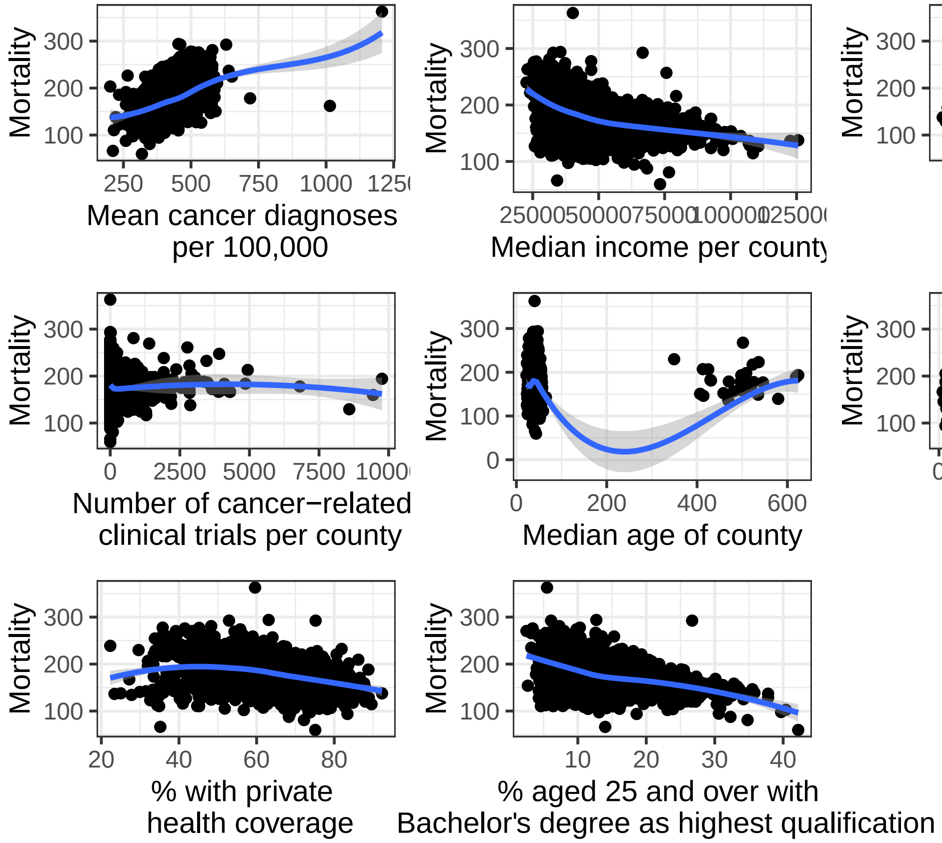
Assignment 1 Questions

Q1. (28 marks) Data on US cancer mortality rates for over 3000 counties are available in the dataset `cancer_reg.csv` available on Blackboard. The data were obtained from the Data World website (https://data.world/nrippner/ols-regression-challenge). Read the data set into R and use it to answer the questions that follow. We'll use the subset of variables listed below:

- `incidencerate`: Mean per capita (100,000) cancer diagnoses[1]
- `medincome`: Median annual income (dollars) per county ([2]
- `povertypercent`: Percent of county population in poverty[2]
- `studypercap`: Per capita number of cancer-related clinical trials per county[1]
- `medianage`: Median age (in years) of county residents[2]
- `pctunemployed16_over`: Percent of county residents aged 16 and over that are unemployed[2]
- `pctprivatecoverage`: Percent of county residents with private health coverage[2]
- `pctbachdeg25_over`: Percent of county residents aged 25 and over with bachelor's degree as highest education attained[2]
- `target_deathrate`: Response variable. Mean per capita (100,000) cancer mortalities[1]

[1] Years 2010-2016 [2] 2013 Census Estimates

a. **(6 marks)** Create a new dataset called `cancer2` that contains only the subset of variables listed above. Based on a summary of the variables in the dataset and the plots below, identify any variable or variables that have obviously incorrect values. For the variables you identify, write and implement code to filter out the incorrect values. Give the number of observations left in the dataset.



b. **(4 marks)** Some data cleaning is done on `cancer2` and a new dataset `cancer3.csv` (available on Blackboard) is created. Construct a scatterplot matrix of all variables in the new dataset. List any key points of note from the scatterplot matrix, including any considerations you might make during a regression analysis.

c. **(3 marks)** Fit a linear model to the data in `cancer3`, including all predictors with no transformations or interactions. Present a summary of the model in a table. Give an estimate of σ^2, the error variance.

d. **(2 marks)** Suppose two counties differ by 1 per 100,000 in mean cancer diagnoses with all else being equal. Based on the model fitted in part (c), what is the difference in expected cancer mortality for these two counties?

e. **(2 marks)** Does it make practical sense to interpret the intercept for the model in part (c)? Justify your answer.

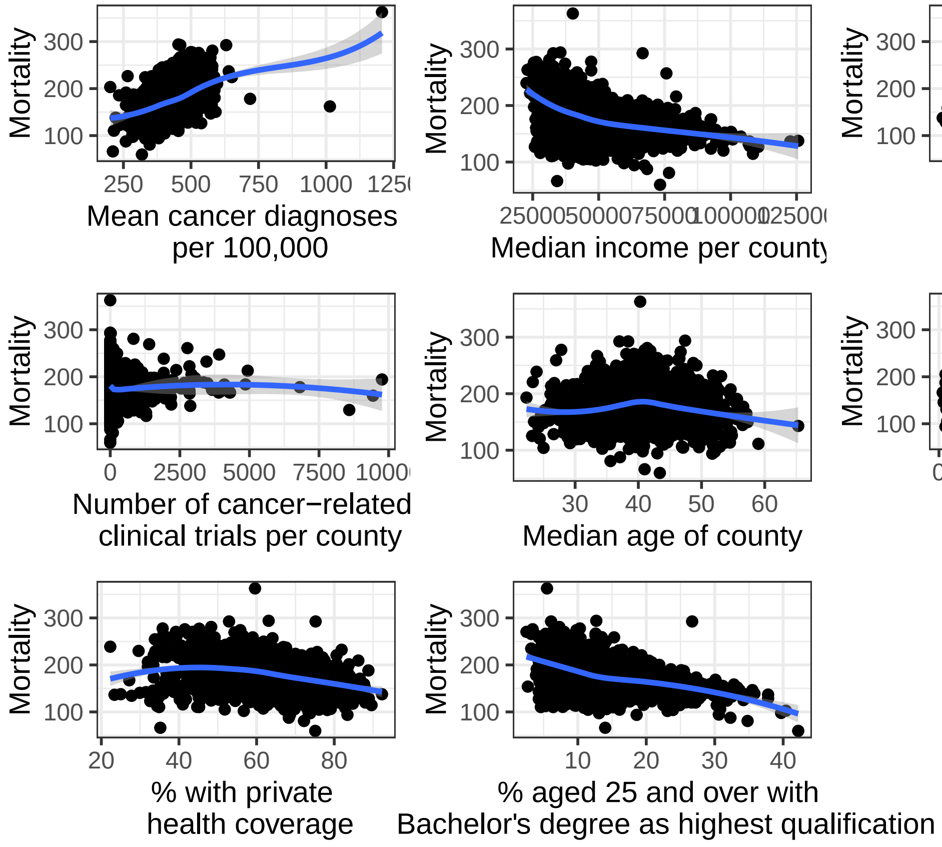
f. **(3 marks)** The model fitted in part (c) is to be used to predict cancer mortality for a county with the predictor values below. Obtain 95% confidence and prediction intervals for such a county. Explain briefly why the prediction interval is wider than the confidence interval.

- `incidencerate`: 452
- `medincome`: 23000
- `povertypercent`: 16
- `studypercap`: 150
- `medianage`: 40
- `pctunemployed16_over`: 8
- `pctprivatecoverage`: 70
- `pctbachdeg25_over`: 50

g. **(3 marks)** Assuming all regression assumptions hold, are the intervals you obtained in part (f) likely to be valid? Explain your answer briefly.

h. **(3 marks)** Based on a global usefulness test, is it worth going on to further analyse and interpret a model of `target_deathrate` against each of the predictors? Carry out the test, give the conclusion and justify your answer.

i. **(2 marks)** The plots below are constructed from the cleaned dataset `cancer3`. Which predictors, if any, would you consider applying log or polynomial transformations to? Explain your answer briefly.

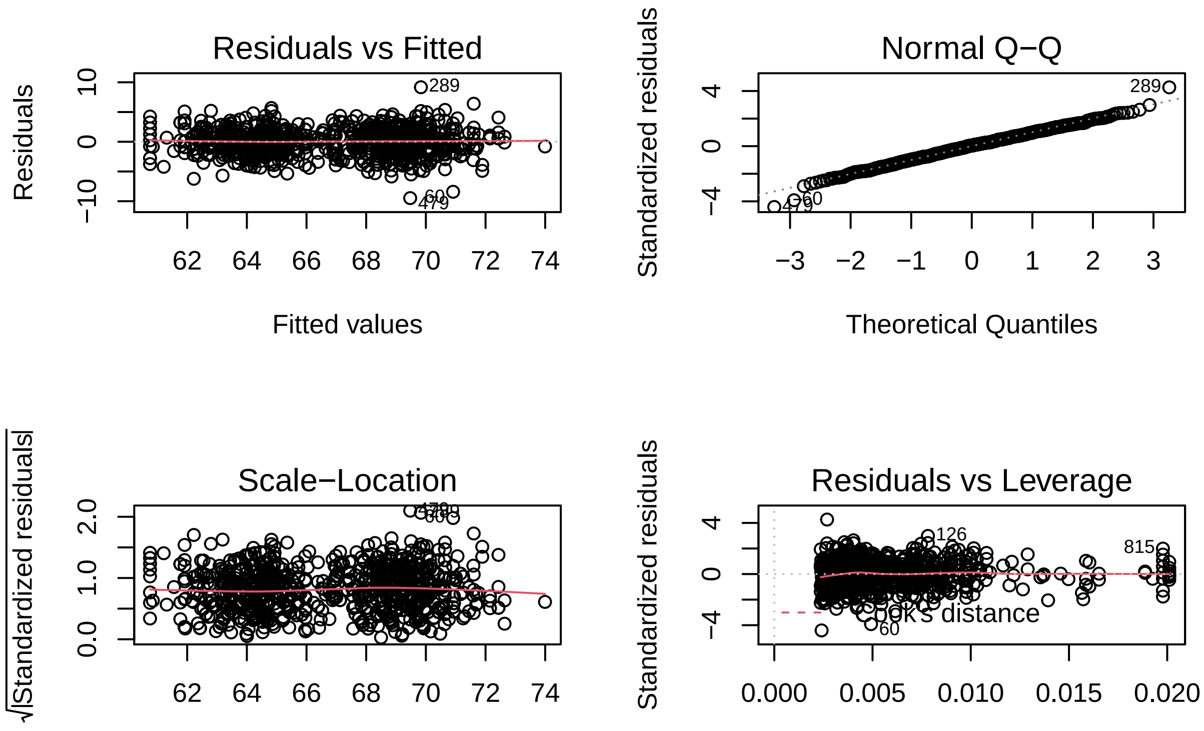


Q2. (12 marks) Francis Galton's 1866 dataset (cleaned) lists individual observations on height for 899 children. Galton coined the term "regression" following his study of how children's heights related to heights of their parents. The data are available in the file `galton.csv` and contain the following variables:

- `familyID`: Family ID
- `father`: Height of father
- `mother`: Height of mother
- `gender`: gender of child
- `height`: Height of child
- `kids`: Number of childre in family
- `midparent`: Mid-parent height calculated as ('father + 1.08*mother)/2
- `adltchld`: `height` if `gender`=M, otherwise 1.08*`height` if `gender`= F

All heights are measured in inches.

a. **(3 marks)** Read the data into R and fit a linear model for `height` with the variables `father`, `mother`, `gender`, `kids` and `midparent` as predictors. Provide a summary of the fitted model. You will notice that estimates for `midparent` are listed as `NA`. Why might this be the case and what regression problem does this point to?
b. **(2 marks)** What action might you take to resolve the problem identified in part (a)?
c. **(2 marks)** Based on the model fitted in part (a) give an interpretation of the coefficient for `genderM`.
d. **(2 marks)** Determine the number of families in the dataset.
e. **(3 marks)** The problem in part (a) is resolved and a new linear model is fitted.No observations are excluded. The plots below are obtained to investigate regression assumptions for this new model. Based on your answer in part (d) and the plots below, do the data meet all the regression assumptions? Explain your answer briefly.



Assignment total: 40 marks