The University of British Columbia
Irving K. Barber Faculty of Science
DATA 101

Assignment 5

Please submit your assignment as an R script file named with your last name, student number, assignment number and with the suffix R. For example, if Joe Smith, student number 87654321 hands in Assignment 4, he would name the file `Smith87654321A2.R`.

Within your answer file, include answers with your R code preceded by the `#` sign. For example, to answer the 5th question on an assignment which is "Perform the calculation $2 + 2$", you would type

```
# Question 5
2 + 2 #coding
# 4 (your answer here )
```

Due Date: Dec 3, 2020

1. We have made a function called `hornerpoly()` in question 3 lab 8. You can use this function to write another function that can evaluate $\sin(x)$ by using the formula below:

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

or

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}.$$

Of course, you cannot evaluate the infinite sum, so you will need to evaluate only the first m terms of the sum, say, with $m = 9$ or $m = 10$. Remember that $n!$ can be calculated with the function in R `factorial()`. Call your function `sintaylor()`. (5 points)

2. Population figures for all Canadian provinces and territories for 2017 through 2019 can be found in the 13th column of the file *population.csv.* The first two columns of that file give the year and the region. Read the data into R and create a new data frame object called `CanPop` that contains only those three columns, with labels `Year`, `Province` and `Pop`. In addition, create a *lattice* dotplot that has three panels showing the populations for each region, one panel for each of the years. (4 points)

3. Numbers of homicides for all Canadian provinces and territories for 2017 through 2019 can be found in the 11th column of the file *murders.csv.* The first two columns of that file give the year and the region.

    (a) Read the data into R and create a new data frame object called `CanMurders` that contains only those three columns (i.e. 1, 2 and 11), with labels `Year`, `Province` and `Murders`. (2 points)

    (b) Create *lattice* dotplots showing the numbers of murders in each region, by year and a single dotplot showing the numbers of murders each year. (You will need to convert the `Year` column to character data using the `as.character()` function beforehand.) (3 points)

(c) Create a new column in the data frame `CanMurders`, called `MurdersPerCapita` which contains the number of murders divided by the provincial or territorial population multiplied by 100000. This is the number of murders per 100000 population and is a more useful basis for regional comparisons. (The population for the provinces and territories is in the data frame discussed in the previous question.) Then create dotplots showing murders per capita in each region by year and a dotplot showing numbers of murders in each year. (2 points)

(d) Using the data frame `CanMurders`, create a new data frame called `SumMurders` that contains three columns and three rows. The first column name is `Year` and will contain the year data. The second column name is `Murders` that shows the total number of murders for all of Canada in that year. And the third column names MurdersPerCapita that shows the total number of murders per capita in that year for all of Canada. (You may use the data frame `CanPop` to help construct this column.) And creat the dot plot showing the total numbers of murders and the numbers of murders per capita for all nation each year.(10 points)

(e) Create a bar plot of 2017 provincial and territorial homicide counts and compare it with the corresponding bar plot of 2017 provincial and territorial per capita homicide rates. Include labels for the bars using the `names.arg` argument containing the entries of the `Province` column and using `cex.name = .35` and `las=2` for readability. (4 points)

(f) Repeat the previous exercise for 2018 and 2019. (4 points)

(g) Create a 3 column matrix called `CanHomicides` consisting of the per capita murder rates for the provinces and territories for the three years, using the `cbind` function. Name the columns of the matrix `2017`, `2018` and `2019`.

Also, create side-by-side bar plots of the per capita homicide rates for the provinces and territories, one for each of the three years. Use the `rep` function and the `Province` column of `CanMurders17` in the `names.arg` argument to print bar labels below the horizontal axis. (5 points)

4. Consider the data in *radon.R*.

(a) Fit a regression model relating measurement to diameter. Compute the PRESS statistic using the `PRESS` function in the *MPV* package. (3 points)

(b) Fit a regression model relating measurement to both diameter and time. The time is the order for recording measurment for each diameter such as
`time<- rep(1:4, 6)` Compute the PRESS statistic. Which of your two models is better? (4 points)

(c) Predict the first measurement for a diameter of 0.9 (i.e. assume `time` is 1). Compare with what you would predict for the fourth measurement for a diameter of 0.9 (i.e. assume `time` is 4). What are the 95% prediction intervals in each case? (2 points)

5. The mouse tumor data `mousetumours.R` were collected at University Hospital in London, Ontario. They concern measurements of tumours growing in three different mice at different times following injection of a known carcinogen. Use `xyplot()` to construct a graph of volume against time, for each mouse. Overlay a smooth curve in each panel. (4 points)

6. The `table.b4` data frame in the *MPV* package has 24 observations on property valuation. There are 10 columns in the data frame. Use `help()` function to check what the variables are. Try out the models below and identify the one that you think is best. (Using `PRESS()`)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \varepsilon \quad (1)$$
$$y = \beta_0 + \beta_1 x_1 + \varepsilon \quad (2)$$
$$y = \beta_0 + \beta_1 x_1 + \beta_8 x_8 + \varepsilon \quad (3)$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad (4)$$
$$y = \beta_0 + \beta_1 x_1 + \beta_5 x_5 + \varepsilon \quad (5)$$
$$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_8 x_8 + \varepsilon \quad (6)$$
$$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \varepsilon \quad (7)$$
$$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_9 x_9 + \varepsilon \quad (8)$$

If we don't have any information on taxes, which aspects of the house are most useful for us to look at in order to predict house price? Write down the predictive model. (5 points)

Data References

1. Statistics Canada. Table 35-10-0068-01 Number, rate and percentage changes in rates of homicide victims DOI: `https://doi.org/10.25318/3510006801-eng`

2. Statistics Canada. Table 17-10-0060-01 Estimates of population as of July 1st, by marital status or legal marital status, age and sex DOI: `https://doi.org/10.25318/1710006001-eng`