# Lab 1 - Basic RStudio and SLR

---

## Lab Goals

In this lab we will:

1. Review some of the basic functionality of R and RStudio including:
    a. the purpose and format of "code chunks" in .Rmd documents
    b. basic math in R
    c. how to correctly denote math equations and symbols within your R Markdown document
2. Download a data file into the workspace for this R Markdown document and do some basic checks on the downloaded file
3. Run a linear model in this R Markdown document using the `lm()` function and review how to interpret the results

## Exploring R and RStudio

### Code Chunks

1. All R code to be run within an R Markdown document must be included in a *code chunk*. You can add a code chunk to this document by clicking on *Chunks->Insert Chunk* in the toolbar above for this R markdown document. Include a code chunk here.
2. Remember that at its most basic level, R can be used as a calculator. Also, keep in mind that R code included in a code chunk can be highlighted and run in the console below by selecting "Run" from the toolbar for the R Markdown document. Run the following calculations in the console below.

```
2+3
```

```
## [1] 5
```

```
2-3
```

```
## [1] -1
```

```
2*3
```

```
## [1] 6
```

```
2/3
```

```
## [1] 0.6666667
```

```
2^3
```

```
## [1] 8
```

3. Any code included in a code chunk will be run within the associated R Markdown document once the document has been knitted. Knit this R Markdown document into pdf or Word document and look at the output from the code chunk above. It should match the output from running this code in the console below.

4. It is important to understand that the console below and this R Markdown document have **separate** workspaces. You cannot access information or variables from either workspace that has not been previously defined in that workspace.

## Proper Syntax for the Equation Editor

Often in the text of a R Markdown document, we would like to include mathematical symbols or equations. **It is important that all mathematical symbols are surrounded by $ signs for the equation editor to interpret it correctly.** While this may seem cumbersome, it enables us to write some nice mathematical formulas in our knitted documents. Here are some examples. Knit this document to see the effect of denoting each line below as "math code."

1. $5 > 4$
2. $x \leq 2$
3. $\pi$
4. $x^3$
5. $Y_{ij}$

**Remember this IS NOT R code. It cannot be used in a code chunk or in the console below. This is strictly for mathematical expressions in the text of the R Markdown document.**

## Loading Data

We will explore some of the capabilities of R and RStudio using the Old Faithful data. Perform the following steps to load this data into the workspace for this document:

1. For this lab, we will run a linear model using eruption data from the geyser Old Faithful. This data is found in the file *Faithful.csv* in the Lab 1 folder on blackboard. Download this file into the folder for Lab 1 on your computer.

2. We will first load the Old Faithful data into the working directory for the console below. Start by making the following selections from the menu for Rstudio above, *Tools->Import Data Set->From Text File...*. A window will pop up in which you can navigate to where you downloaded *Faithful.csv* on your computer. Open this file.

3. You should now see a preview of the Old Faithful data. Whenever you load data into R, it is important to check that it loaded correctly. Does the preview of the data look OK? If so, load the data into the workspace of the console.

4. Now we will include the necessary code in a code chunk here to load the Old Faithful data into the workspace of this R Markdown document. This is easily done by copying the code from the console below that R used to load the data into the workspace for the console. Include this code in a code chunk here.

*Do not include the line below using the `View()` function. This function will cause an error when you try to knit this .Rmd document.*

```
Faithful <- read.csv("~/Personal/6020/labs/Faithful.csv")
```

5. Knit this document into a Word or pdf document.

## Data Checks

The data is now loaded into this R Markdown document, but did it load correctly? Complete the following steps to do some basic checks.

1. Insert a code chunk here and use the following R functions with `Faithful` as the argument: `head()`, `tail()`, `names()`. Run this code in the console. What information can be determined by running these fuctions in R?

```
head(Faithful)
```

```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```

```
tail(Faithful)
```

```
##     eruptions waiting
## 267     4.750      75
## 268     4.117      81
## 269     2.150      46
## 270     4.417      90
## 271     1.817      46
## 272     4.467      74
```

```
names(Faithful)
```

```
## [1] "eruptions" "waiting"
```

They return, respectively, the first 6 rows of the dataset, the last 6, and the column names.

2. Each row in this data set consists of two observations associated with an eruption of Old Faithful. The variable `eruptions` is the elapsed time of the eruption in minutes. The variable `waiting` is the time from the last eruption to the current eruption in minutes. Suppose we would like a summary of `eruptions`. Type `summary(eruptions)` in a code chunk below and try to knit your document. Did your document knit? If not, try to determine why by looking at the error message in the R Markdown tab below. How can we fix this error (there are a couple of options)? Fix the error and re-knit your document. 'summary(eruptions)' returns 'Error in summary(eruptions) : object 'eruptions' not found' eruptions is a column in the dataset Faithful, not a dataset itself. I can get around this by running 'summary(Faithful$eruptions)' to summarize just the column eruptions, or I could find a way to make eruptions its own dataset.

```
summary(Faithful$eruptions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.600   2.163   4.000   3.488   4.454   5.100
```

## Simple Linear Regression

Under certain conditions, a simple linear regression model is appropriate to model the relationship between a response, Y, and a predictor, X. For a bivariate random sample, $(X_1, Y_1), ..., (X_n, Y_n)$, the model assumes the following relationship between $X_i$ and $Y_i$

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

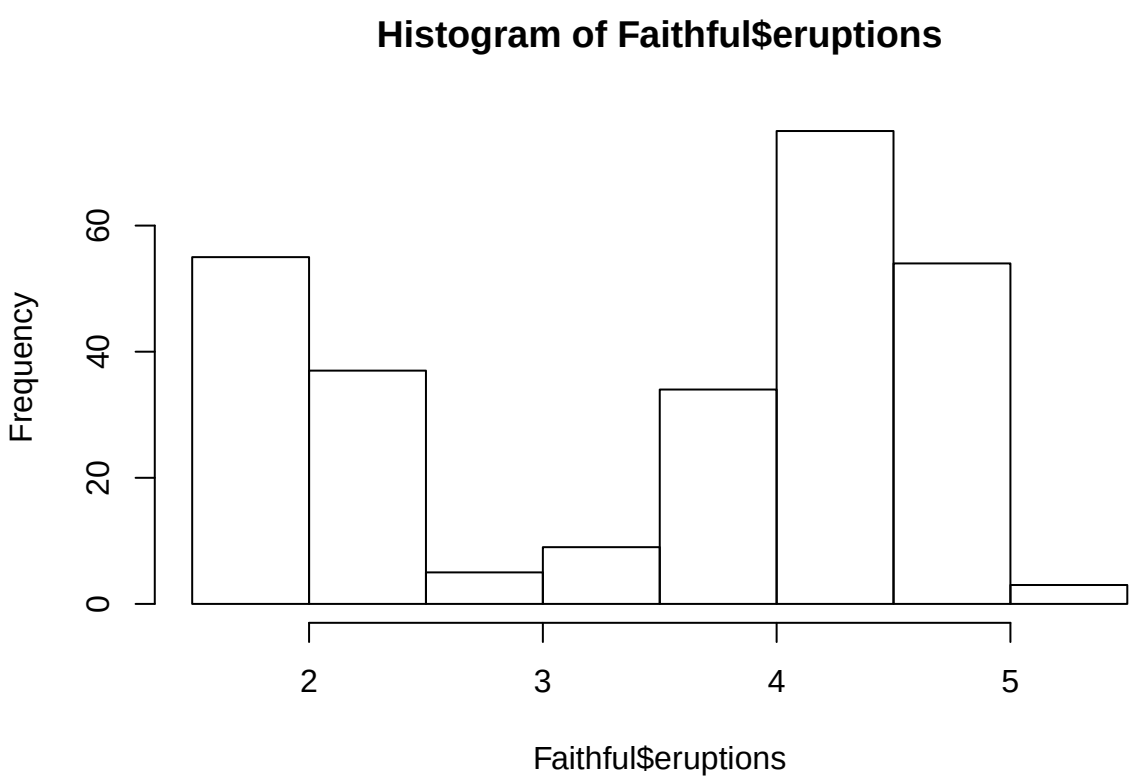
where

1. $\beta_0$ is the intercept of the regression line
2. $\beta_1$ corresponds to the amount the expected value of $Y_i$ increases per unit increase in $X_i$ (the slope of the regression line)
3. $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$
4. $\epsilon_i$ is the error term for observation $Y_i$
5. $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ for all $i = 1, \ldots, n$ and are mutually independent

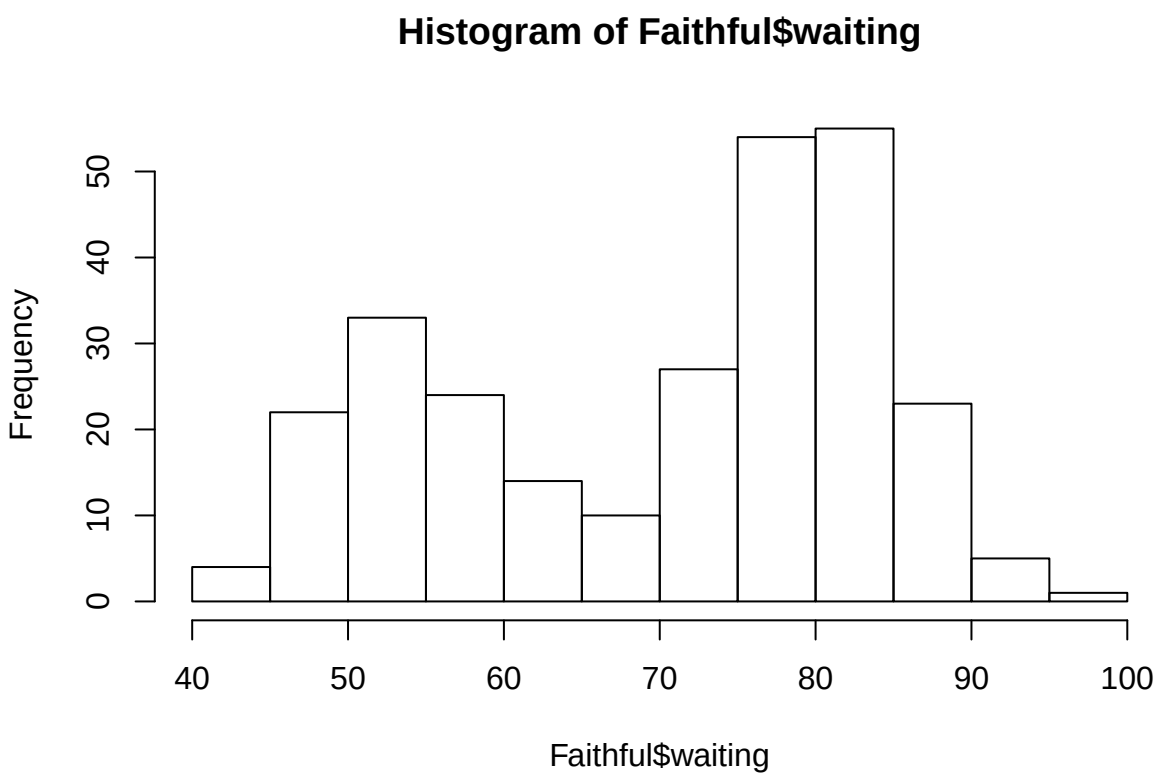
Here we will run a simple linear regression model in R with `eruptions` as the response and `waiting` as the predictor. Complete the following steps.

1. It is always a good idea to look at your data. In a code chunk here include a histogram of both variables and a scatterplot of `eruptions` by `waiting`. Use the R functions `hist()` and `plot()`.

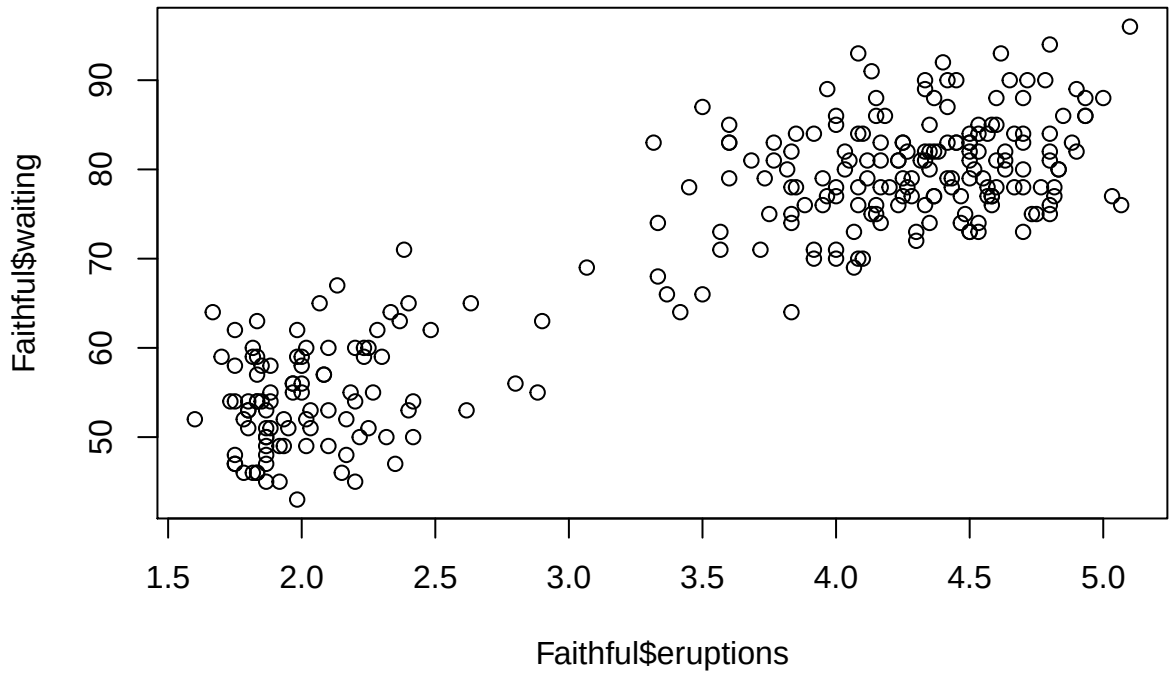
```
hist(Faithful$eruptions)
```



```
hist(Faithful$waiting)
```




```
plot(Faithful$eruptions,Faithful$waiting)
```




a. What is notable about the sample distributions of `eruptions` and `waiting`?
they are bimodal

b. Does there appear to be a linear relationship between the time between eruptions and the length of t
yes, although the scatterplot shows a more complex pattern

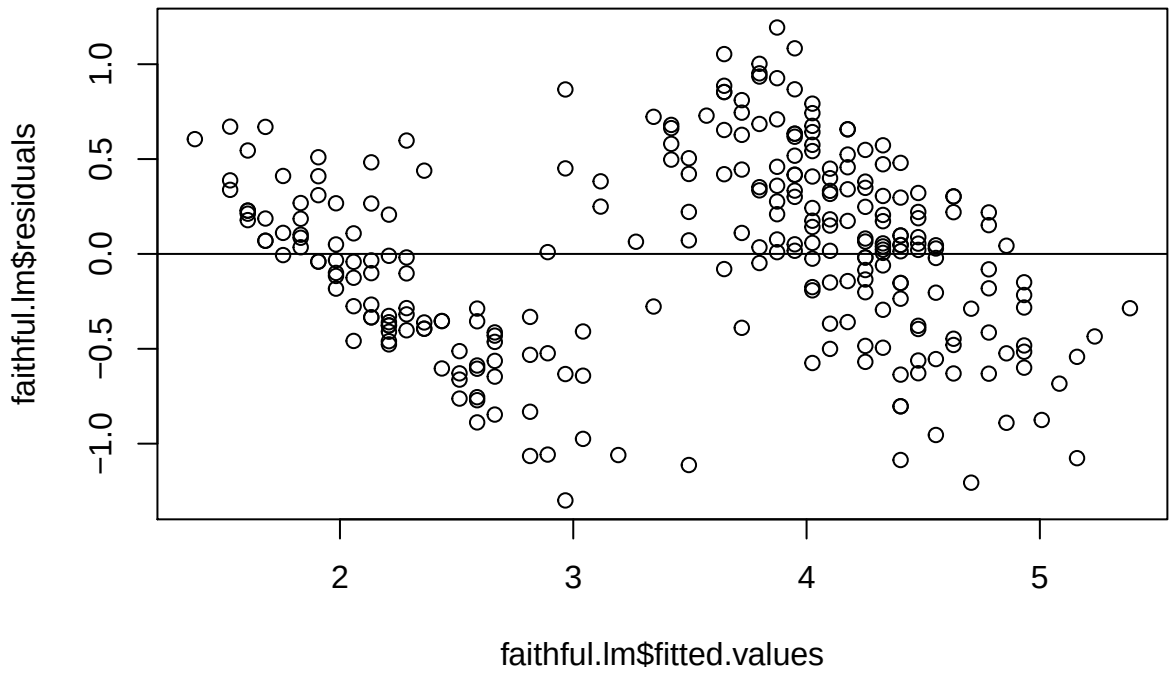
2. Lets check the assumptions for a SLR model.

   a. Are the observations independent? no

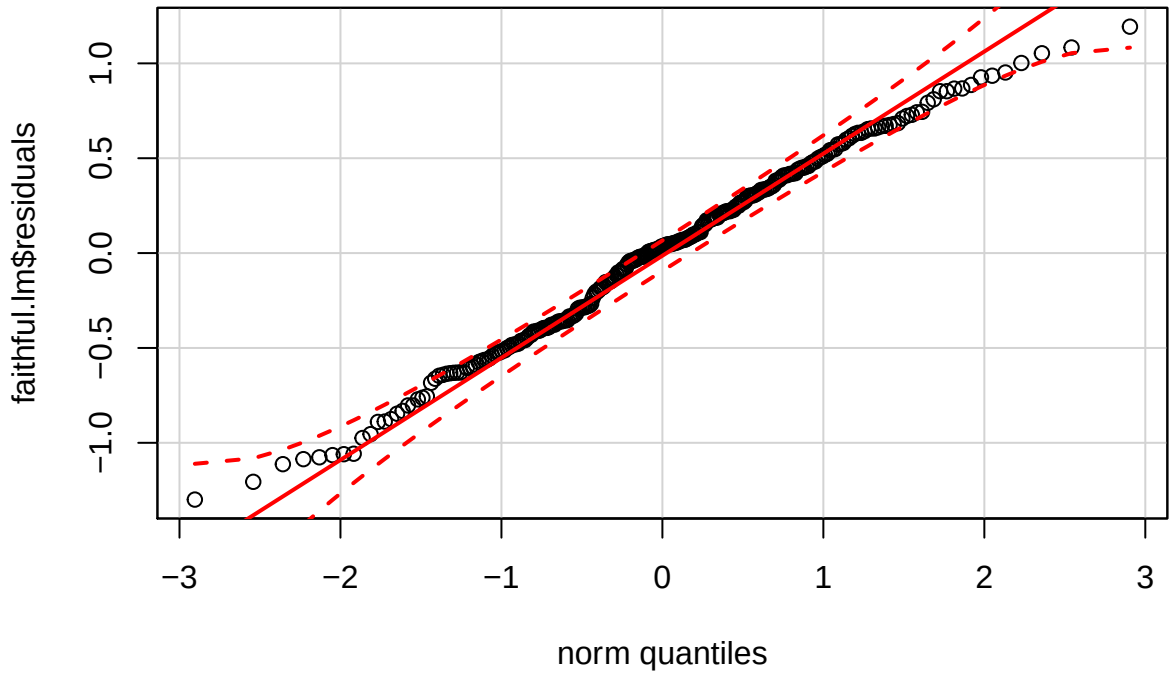
   b. The following code will (Reminder: set `eval=TRUE`):

      1. Run the linear model in R
      2. Create a plot of the residuals by the fitted values
      3. Create a Q-Q plot of the residuals. For this plot, you will need the `car` package. If the `car` package is not already installed, select *Tools>Install Packages...* from the menu for RStudio. Type `car` for the package in the window that pops up. Then click on "Install." It may take a minute or so for this package to install in R. You will know this package is not installed if you get an error related to this package when you try to run the code in the code chunk below.

```
faithful.lm <- lm(Faithful$eruptions ~ Faithful$waiting)  # lm named faithful.lm is run
plot(faithful.lm$fitted.values, faithful.lm$residuals)  # plots residuals by the fitted values
abline(0, 0)  # plots line at y=0
library(car)  # accesses the 'car' library that contains qqPlot()
```




```
qqPlot(faithful.lm$residuals)  # Q-Q plot of residuals
```




c. Knit your document to see the output from the R code from (b). Is the homogeneity of variance assumption met? Is the normality assumption met? no and no

3. For the purpose of reviewing the R output from running a linear model, we will temporarily ignore the violation of assumptions for the regression model. Use the `summary()` function in a code chunk below to obtain parameter estimates and other information about this model. You can look at this information in the console without knitting your document by first running the linear model in the console and then looking at a summary of that model.

```
summary(faithful.lm)
```

```
##
## Call:
## lm(formula = Faithful$eruptions ~ Faithful$waiting)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29917 -0.37689  0.03508  0.34909  1.19329
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.874016   0.160143  -11.70   <2e-16 ***
## Faithful$waiting  0.075628   0.002219   34.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 270 degrees of freedom
## Multiple R-squared:  0.8115, Adjusted R-squared:  0.8108
## F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

4. What is the estimated relationship between expected length of eruption and the time since the last eruption? eruption time = 0.0756 times waiting time -1.87

5. Interpret the $R^2$ value of this model. 0.81. 81 percent of the variation in length of eruption can be explained by variation in length of waiting.

6. Is there a significant linear relationship between expected `eruptions` and `waiting`?

yes, highly significant to more than 0.001

7. What is the value of $\hat{\beta}_1$? Interpret $\hat{\beta}_1$ in the context of this analysis.

   what does beta one represent again? the y intercept? the estimate is -1.87 , meaning that when the model uses a waiting time of zero it predicts a negative eruption duration. obviously the model is not useful at such an extreme