

## Simple Linear Regression and Correlation

In this assignment you will complete a variety of tasks related to correlation and simple linear regression.

**Good habits:** I strongly recommend creating a new RStudio project for every assignment and for each lecture as you follow-along. Using a good directory structure will make it much easier for you to find your work later. For this assignment, you might create a directory for all Module 2 work. Within this directory, create a folder called “Correlation and Simple Linear Regression Assignment” (or similar). Create an R Project (with an appropriate name) in this folder.

**Deliverable:** All of your work for this assignment should be done in an R Markdown document. Knit your document into a Word file and submit the Word file as the deliverable for this assignment.

**Libraries:** For this assignment you will need the following libraries: tidyverse, GGally

**Task 1:** Read-in the airquality data set (a default R dataset) as a data frame called “air”. To do this use the code below:

```
air = airquality
```

Details concerning this dataset can be found here: [http://rpubs.com/Nitika/linearRegression\\_Airquality](http://rpubs.com/Nitika/linearRegression_Airquality).

- Describe this dataset.
- How many variables and observations are there?
- Is there any missing data?
- Which variable is likely to be the response (Y) variable?

**Task 2:** In Task 1 you would have discovered that there is missing data in two of the variables: “Ozone” and “Solar.R”. We have three approaches that we can typically select from to deal with missing data:

- Delete the rows with missing data
- Delete the columns with missing data
- Impute (i.e., estimate or guess) values to replace the missing values.

Here we’ll choose deletion of any rows with missing data. Run the tidyverse-style code below to implement this approach (Note that “!is.na” means “not NA”):

```
air2 = air %>% filter(!is.na(Ozone)) %>% filter(!is.na(Solar.R))
```

How many rows and columns remain in this new (air2) data frame?

**Task 3:** Use the code below to develop a visualization of and to calculate correlation for the combinations of variables in this dataset. Then use the “ggcorr” to develop a correlation matrix for the variables. Hint: Use “label = TRUE” in the ggcorr function to show the correlation values.

- Which variable is most strongly correlated with the “Ozone” variable?
- Which variable is least strongly correlated with the “Ozone” variable?

**Task 4:** Plot “Temp” (x axis) versus “Ozone” (y axis) using the “ggplot” function. Choose an appropriate chart type. Describe the relationship between “Temp” and “Ozone”.

**Task 5:** Create a linear regression model (called in model1) using “Temp” to predict “Ozone”. **a.** Discuss the quality of this model (mention the R square value and significance of the predictor variable). **b.** Use the code “confint(model1)” to generate 95% confidence intervals for the coefficients. In what range does the slope coefficient likely fall?

**Task 6:** Re-do Task 4 to include the regression line. Hint: Add “geom\_smooth(method=“lm“, se = FALSE)”.

**Task 7:** Develop a prediction for “Ozone” when “Temp” is 80. Include the prediction interval for this prediction.

**Task 8:** Plot “Day” (x axis) versus “Ozone” (y axis) using the “ggplot” function. Choose an appropriate chart type. Describe the relationship between “Day” and “Ozone”.

**Task 9:** Create a linear regression model (called in model2) using “Day” to predict “Ozone”.

- a. Discuss the quality of this model (mention the R square value and significance of the predictor variable).
- b. Use the code “confint(model2)” to generate 95% confidence intervals for the coefficients. In what range does the slope coefficient likely fall?

**Task 10:** Re-do Task 8 to include the regression line. Hint: Add “geom\_smooth(method=“lm“, se = FALSE)”.