# Homework 1

*Your Name*

*February 7, 2018 at 11:59pm*

**Homework Policies:**

*You are encouraged to discuss problem sets with your fellow students (and with the Course Instructor of course), but you must write your own final answers, in your own words. Solutions prepared "in committee" or by copying someone else's paper are not acceptable. This violates the Brown standards of plagiarism, and you will not have the benefit of having thought about and worked the problem when you take the examinations.*

*All answers must be in complete sentences and all graphs must be properly labeled.*

***For the PDF Version of this assignment: PDF***

***For the R Markdown Version of this assignment: RMarkdown***

**Turning the Homework in:**

*Please turn the homework in through canvas. You may use a pdf, html or word doc file to turn the assignment in.*

PHP 1511 Assignment Link

PHP 2511 Assignment Link

## The Data

This homework will use the following data:

- `hw1a` - https://raw.githubusercontent.com/php-1511-2511/php-1511-2511.github.io/master/Data/hw1a.csv
- `hw1b` - https://raw.githubusercontent.com/php-1511-2511/php-1511-2511.github.io/master/Data/hw1b.csv

## Part 1

1. The data set hw1a is simulated from a famous example that illustrates how variables in multiple regression can be used to predict the response variable, here `Y`, jointly even though they do not necessarily predict `Y` very well individually. There are two predictor variables in the data set, `X1` and `X2`.

a. Open the data set and begin by looking at all possible two-way scatterplots. Comment on the relationships that you observe.
b. Next, examine the simple linear regressions of each predictor to explain `Y`. Comment on whether the predictors seem to relate to `Y`. What percent of the variability in Y does each predictor explain by itself?
c. Now use `lm()` to build a multiple regression model using both predictor variables `X1` and `X2`. Comment on the fit and the statistical significance of each predictor variable. What percent of the variability in Y is explained by the model now that both predictors are included? Give an explanation for what you think is happening with both predictors in the model.
d. Comment on the change in the estimated coefficients from the simple linear regression models compared to those from the multiple regression model. Are the changes qualitative (direction), quantitative (magnitude) or both?

e. Run the following code to create a 3d scatterplot. Notice that multiple linear regression is now a plane and not just a line. Why do you think `X1` and `X2` predict `Y` so well together when they do not alone?

```
library(plotly)
hw1a <- read.csv("https://raw.githubusercontent.com/php-1511-2511/php-1511-2511.github.io/master/Data/hw

p <- plot_ly(hw1a, x = ~x1, y = ~x2, z = ~y) %>%
  add_markers() %>%
  layout(scene = list(xaxis = list(title = 'x1'),
                      yaxis = list(title = 'x2'),
                      zaxis = list(title = 'y')))

p
```

---

## Part 2

### The Data

Data set `hw1b` contains air pollution data from 41 U.S. cities. Our goal is to try to build a multiple regression model to predict SO2 concentration using the other variables.

| Variable Name | Description |
|---|---|
| `so2` | SO2 air concentration in micrograms per cubic meter. |
| `temp` | Average Annual temperature in degrees F. |
| `empl20` | The number of manufacturing companies with 20 or more workers. |
| `pop` | The population in thousands. |
| `wind` | The average annual wind speeds in miles per hour. |
| `precipin` | The average annual precipitation in inches. |
| `precipdays` | The average number of days with precipitation per year. |

---

2. Load data set HW1b and answer the following questions. *Display all useful code and output inline. Do not just display all that R gives you but display parts that show why you chose the model you did.*

a. Begin by examining univariate summaries of the 7 variables. Do any of the points seem to have extreme values? Comment on whether cities with extreme values also have extremes on one or more other variables.

b. Start your model building by looking at simple linear regressions for each of the 6 predictor variables. Display and Examine relevant plots. Summarize the simple linear regression results using the broom package. Note that you can combine tidy statements:

```
tidy1 <- tidy(model1)
tidy2 <- tidy(model2)
rbind(tidy1, tidy2)
```

---

c. Build a multiple regression model by sequentially adding variables that you feel are important from the simple linear regressions.

d. State your final multiple regression model, interpret the parameter estimates and the R2. Comment on any differences in coefficients between the simple linear regression models and the multiple regression model.

e. Complete the following with the final model you built in part d.
   i. What does the adjusted $R^2$ tell you about your model fit?
   ii. Perform a hypothesis test on the slope estimates for each variable in the model.