# Homework 4: Harder Simple Linear Regression

Mario Giacomazzo

September 13, 2023

## Instructions:

The purpose of this homework assignment is to look more into assessing our conditions, looking for outliers/influential points, and performing transformations. Make sure you read each question carefully. In each question, I will give you a task to do, and I will tell you what I want you to output. You can write as much code as you want in each code chunk, but make sure you complete the task and only print the output I asked you to print. Don't sort the data unless you are told to sort the data. You should remove the "#" sign in each code chunk before writing your code. Also, if you see the comment "#DO NOT CHANGE", then I don't want you to make any modifications to that code. **You should knit your RMD file to a PDF after you answer every question.**

After you are done, knit the RMarkdown file to PDF and submit the PDF to Gradescope under HW4.

## Questions

### Q1 (2 Points)

For this homework, I am going to simulate data. The code below will use random sampling from the uniform distribution to sample values for $X$ in the interval from 0 to 10. We then fix our y-intercept and slope parameters to generate values for $Y$ that are dependent on $X$. We add an error that is randomly sampled from a gamma distribution to ensure that there is not a perfect linear relationship between $X$ and $Y$. The gamma distribution is a theoretical distribution for a variable that is nonnegative. Therefore, all of our errors would tend to be positive, unless I multiply by a negative number. I use the sample function to modify the positive error to be negative 10 percent of the time.

I want you to create a data frame in R called **SIM1** which is based on the two vectors: **X** and **Y**. I want the names of the variables in **SIM1** to be "X" and "Y" so we know which variable is the response and which variable is the predictor.

I would recommend learning about the function named `data.frame()`. The website https://statisticsglobe.com/data-frame-function-r would be helpful reading.

After you create **SIM1**, I want you to display the first 10 rows. This should be the only output. The `set.seed(110)` code hopefully will ensure that every student in the class has the same sample when the entire code chunk is run. When you modify this code chunk, you need to run the entire code chunk to

```
set.seed(110) #DO NOT CHANGE

X = runif(50,0,10) #DO NOT CHANGE
beta0 = -2 #DO NOT CHANGE
beta1 = 5 #DO NOT CHANGE

Y = beta0 +
    beta1*X +
```

```
    rgamma(50,shape=1*X,scale=25)*sample(c(-1,1),size=50,replace=T,prob=c(0.1,0.9)) #DO NOT CHANGE

#
```

### Q2 (2 Points)

Fit the linear regression model for the relationship of $Y$ versus $X$ and create an appropriate scatter plot with the fitted regression line added to the plot. The only output I want to see is your graphic.

```
#
```

### Q3 (4 Points)

I want you to create two variables named "stdX" and "stdY" that are standardized versions of the variables named "X" and "Y", respectively. To standardize each variable, you need to subtract the sample mean (centering) and then, divide by the sample standard deviation (scaling). You can look into using the `scale()` function in R to do this.

After you create these variables, I want you to make the same type of plot you did in the previous question. I want to see only a scatter plot with a regression line in the output. You are using *stdY* vs *stdX*. Then, I want you to reflect on what the differences are between your plot in Q2 and your plot in Q3. How did standardizing the data impact our scatterplot and regression line? Put your answer in the designated area below the code chunk.

```
#
```

**Response in Complete Sentences:** REPLACE EVERYTHING IN ALL CAPS HERE WITH YOUR ANSWER IN COMPLETE SENTENCES

### Q4 (2 Points)

Now I want you to calculate the `log()` for every value of the variable $Y$ in the data. You can run the `log()` function on a full vector of numbers. For example, try running the code `log(c(1,2,3,4))` and notice you get the vector `c(log(1),log(2),log(3),log(4))`. In your output, I want to only see a vector of the log of each of the 50 values for $Y$ in the simulated dataset.

If you did this correctly, you will see "NaN" for some values. In the appropriate space below, explain why you got "NaN" for some calculations.

```
#
```

**Response in Complete Sentences:** REPLACE EVERYTHING IN ALL CAPS HERE WITH YOUR ANSWER IN COMPLETE SENTENCES

### Q5 (7 Points)

Using the linear regression of $Y$ on $X$ (like Q2), I want you to save the fitted values and residuals as new variables named **regfit** and **regres** in **SIM1**. Then, I want you to create the following visuals based on this simple linear regression. I only want to see these four visuals in your output.

1. Histogram of the residuals. I want you to use the `hist()` function, but change the number of breaks from the default to 12 so there are more bins.
2. Create a scatter plot of the residuals versus the predictor variable. Make a horizontal reference line at 0.
3. Create a scatter plot of fitted values versus the actual values, then add a 45 degree reference line with a y-intercept of 0 and a slope of 1. You can do this using the `abline()` function. This is not something I taught, but I believe it is a useful plot that can be used when we get into models with more than 1 predictor variable.
4. Create a normal quantile plot in R of the residuals. Remember to add the reference line.

After creating these visuals and inspecting them, I want you to identify three potential problems from looking at these four plots. I want you to write three sentences. Each sentence should be about a different potential problem. You should reference the figure (for example, "in figure 3 . . . " or "in the normal quantile plot"), you should reference the potential problem, and explain why that figure indicates that their may be a potential problem in our model.

```
#
```

**Response in Complete Sentences:** REPLACE EVERYTHING IN ALL CAPS HERE WITH YOUR ANSWER IN COMPLETE SENTENCES

### Q6 (3 Points)

We see some problems in the simple linear regression model. We want to see if a log transformation on $Y$ will help; however, we cannot just calculate `log(SIM1$Y)` due to problems seen in Q4. What I want you to do is to add a constant *C=200* to the variable $Y$ before taking the log.

In **SIM1**, I want you to create a new variable called "logY" which is the transformed version of your variable named "Y". After you do this, I want you to fit a linear regression of $logY$ on $X$ so that you can create the a graphic that is similar to what you were asked to do in Q2 and Q3. I want the scatter plot to use the knew variable named "logY" and not the original variable $Y$.

```
#
```

### Q7 (4 Points)

Now, I want you to save the fitted values of the previous model and the residuals of the previous model into your dataset **SIM1**. I want the fitted values to be called "lregfit" and I want the residuals to be called "lregres".

After this, I want you to create all four figures from the Q5 for these fitted values and residuals. Think about whether or not this transformation helped in modifying the data so we have less problems with the outliers and the conditions of a simple linear regression.

Think about the fact that "lregfit" is a prediction for the log transformed $Y$ and not the original variable $Y$. We will address this later.

```
#
```

### Q8 (3 Points)

Now, I want you to calculate and print out the sum of squared errors (SSE) for the simple linear regression model of $Y$ on $X$.

Then, I want you to calculate the SSE for the simple linear regression model of our version of $logY$ on $X$. However, we cannot compare the two numbers fairly since in one model we are predicting $Y$ and in the other model we are predicting *log(Y+200)*. You need to adjust or "untransform" your predictions so that you are still predicting $Y$, and then you need to compute the residuals manually by comparing these predictions of $Y$ to the actual $Y$. This will allow you to compute SSE so that we can find out if there is evidence that use of the log transformation led to a smaller (better) SSE.

In your output, I only want to see the two numbers: SSE from the simple linear regression of $Y$ on $X$ and SSE from simple linear regression of $logY$ on $X$ that is calculated after reversing the transformation.

```
#
```