Homework 4

Due date: Friday, October 7

Complete the following exercises and submit your assignment via gradescope (linked on the course webpage).

Problems to start after class Sept 30

Q1

Chapter 2 exercise E.13. In parts (b)-(d), only report the best transformation you found.

In your answers to parts (b)-(d), please include your fitted model equation, a scatterplot with the transformed model superimposed on the **original scale** of the variables, and residual plots justifying that you improved the fit.

Note: Look at the handout from class to help you specify the appropriate formula for your `gf_lm()` layer in the scatterplot.

Q2

For the models you selected in parts (b)-(d) of Q1 go back and provide an interpretation of the slope in the context of the problem. Be sure to read the "Interpretation of transformed models" section at the end of the activity from Friday before doing this.

Q3

Consider the following data on penguin heart rate as a function of duration of dive (in minutes).

```
penguins <- read.csv("http://aloy.rbind.io/data/penguins.csv")
```

(a) Plot heart rate against duration. What problems do you see in fitting the simple linear regression model?

(b) Plot the standardized residuals against fitted values, what problem do you see? Are they they same problems you saw previously?

(c) Fit the simple linear regression model of `log(heart.rate)` against `duration`. Report the fitted regression equation.

(d) Interpret the estimated slope in the context of the problem.

(e) Calculate a 90% confidence interval for the slope and interpret it in the context of the problem.

(f) Does the model seem more appropriate? That is, do the assumptions/conditions for the model appear to be met? Explain.

(g) Create a scatterplot with the transformed model superimposed on the **original scale** of the variables.

Problems to start after class Oct 3

Q4

Researchers have a data set consisting of the isotopic composition structural bone carbonate and the isotopic composition of the coexisting calcite cements in 18 bone samples from a specimen of the dinosaur *Tyrannosaraus rex*.

```
trex <- read.csv("https://aloy.rbind.io/data/trexbones.csv")
```

(a) Create a scatterplot of `Calcite` (y) against `Carbonate` (x) and describe the relationship.

(b) Model 1: Using all 18 data cases, fit the regression of Calcite on Carbonate. Report the estimated slope, the standard error for the slope, the and R^2 value. Add the fitted regression line to the scatterplot.

(c) Do you think any points are influential in model 1? Clearly justify your answer.

(d) Model 2: now fit a model removing the first observation (i.e., the observation with the smallest carbonate value). Again, report the estimated slope, the standard error for the slope, the and R^2 value. Add the fitted regression line to the scatterplot using a different `color` and `linetype` to distinguish it from model 1. (There should now be two lines on your scatterplot.)

> **Note**
>
> To add this regression line, you can pass a different data set into the `gf_lm()` command. For example,
>
> ```
> # fill in the blank with the row number(s) to delete
> gf_point(Calcite ~ Carbonate, data = trex) |>
>   gf_lm() |>
>   gf_lm(Calcite ~ Carbonate, data = dplyr::slice(trex, -c(___)),
>         linetype = 2, color = "orange")
> ```

(e) Do you think any points are influential in model 2? Clearly justify your answer.

(f) Model 3: finally, fit a model removing the first and second observations. Again, report the estimated slope, the standard error for the slope, the and R^2 value. Add the fitted regression line to the scatterplot using a different `color` and `linetype` to distinguish it from models 1 and 2. (There should now be three lines on your scatterplot.)

(g) Do you think any points are influential in model 3? Clearly justify your answer.

(h) Compare the R^2 values, slope estimates, and standard errors for the three models. Why is there such a big difference?

(i) Why might pairs (or groups) of influential observations not be found with the usual case influence statistics?

Problems to start after class Oct 5

Q5

The `RailTrails` data set contains information about 104 homes sold in Northampton, Mass., in 2007. In this problem, you will use the following variables:

Variables	Description
`Adj2007`	Estimated 2007 price (in thousands of 2014 dollars)
`Distance`	Distance (in miles) to the nearest entry point to the rail trail network
`SquareFeet`	Square footage of interior finished space (in thousands of square feet)

Suppose that researchers are interesting in determining whether the selling price of a home in Northampton is impacted by its proximity to a bike trail, and if so how is it impacted. The researchers also want to control for the size of a home.

(a) Fit the simple linear model predicting `Adj2007` using `Distance`. Report the fitted regression equation and the R^2 value.

(b) Fit the multiple linear regression model predicting `Adj2007` using both `Distance` and `SquareFeet`. Report the fitted regression equation and the R^2 value. How has the addition of `SquareFeet` changed the model?

(c) Using the MLR model, interpret the estimated slope for `Distance` in context.

(d) Using the MLR model, calculate a 95% confidence interval for the `Distance` coefficient and interpret it in context.

(e) Using the MLR model, predict the price of a particular home that is 1500 ft2 and 0.5 miles from a bike trail.

(f) Create the appropriate interval (either confidence or prediction) for your prediction in part (e). You can use the same R code as with simple linear regression, but now your `newdata` data frame need to have two variables:

```
data.frame(Distance = 0.5, SquareFeet = 1.5)
```

(g) Assess the assumptions necessary for your analysis to be valid.

Q6

Read the description of the `Politics` data set given in Chapter 3 exercise E.12 and answer the following questions.

```
politics <- read.csv("https://aloy.rbind.io/kuiper_data/Politics.csv")
```

(a) Complete part (a) in Chapter 3 exercise E.12. You can use the following code.

(b) You will fit a multiple linear regression model using the response variable you just created in part (a) and the quantitative variables as explanatory variables (`Unemployed`, `Bachelor.s`, `Retirement`, `MedianIncome`, `Christian`, `NumberPeoplePerFamily`, `HealthInsurance`, `Voted`). Before you do this, formulate hypotheses about how each explanatory variable will be related to voting patterns.

(c) Fit the multiple linear regression model using the response variable you just created in part (a) and the quantitative variables as explanatory variables.

(d) Complete part (d) in Chapter 3 exercise E.12. When you state your conclusion about each hypothesis, be sure to use some inferential procedure to support your conclusion.