STAT 632, HW 6
Due: Tuesday, April 7

Reading: Chapter 3, pp. 82–90 and 99–102, from *An Introduction to Statistical Learning*. Chapter 5, pp. 140–146, and Chapter 6, pp. 195–203, from *A Modern Approach to Regression*.

Exercise 1. For this exercise use the `bdims` data set from the `openintro` package. Since `ggplot()` requires that a categorical variable be coded as a factor type in R, run the following code:

```
library(openintro)
bdims$sex2 <- factor(bdims$sex, levels=c(0,1), labels=c("F", "M"))
```

(a) Use `ggplot2` to make a scatter plot with `hgt` on the x-axis, and `wgt` on the y-axis. Color the points according to the gender variable `sex2`. Use `geom_smooth()` to add the least squares lines for each gender to the scatter plot.

(b) Use `lm()` to fit a linear regression model with `wgt` as the response variable, and `hgt` and `sex` as the predictors. Use `summary()` to print the results.

(c) Write down the regression equation for the model fit in part (b). The model describes two parallel lines; what are the equations for these two lines?

(d) Use `lm()` to fit a linear regression model with `wgt` as the response variable, and `hgt`, `sex`, and the interaction between `hgt` and `sex` as predictors. Use `summary()` to print the results. Is the interaction term significant?

(e) Write out the regression equation for the model fit in part (d). The model describes two unrelated regression lines (with different slopes and intercepts); what are the equations for these two lines?

Exercise 2. For this exercise use the HDI data set discussed in the previous assignment:

```
hdi <- read.csv("https://ericwfox.github.io/data/hdi2018.csv")
```

Fit the full model with `hdi_2018` as the response, and the other four variables as predictors (`median_age`, `pctpop65`, `pct_internet`, and `pct_labour`).

(a) Compute the correlation matrix between the predictors. Are there strong correlations between some of the predictors?

(b) Compute the variance inflation factors (VIFs) for the predictors. Do the VIFs indicate that multicollinearity is an issue? [Load the `faraway` package to use `vif()` function]

(c) Provide an interpretation of the VIF for the `median_age` predictor.

Exercise 3. For this exercise use the `ozone` data from the `faraway` package. Fit a model with `O3` as the response, and `temp`, `humidity` and `ibh` as predictors. Use the Box-Cox method to determine the best transformation on the response. Use residual versus fitted plots to evaluate the fit of the model with and without the response transformation.

Additional practice on categorical predictors with more than 2 levels (not to be collected)

Practice Problem. For this exercise use the `Carseats` data set from the `ISLR` package.

(a) Use `ggplot2` to make scatter plots with with `Price` on the x-axis, and `Sales` on the y-axis; use `facet_wrap()` to create 3 panels for each level (Bad, Good, Medium) of the categorical predictor `ShelveLoc` (quality of shelving location). Add a regression line to each panel with `geom_smooth()`.

(b) Fit a linear regression model with `Sales` as the response, and `Price` and `ShelveLoc` as predictors. In the regression model, what is the baseline level for the categorical predictor `ShelveLoc`?

(c) Fit a linear regression model with `Sales` as the response, and `Price`, `ShelveLoc`, and the interaction between `Price` and `Shelveloc` as predictors.

(d) Use the partial F-test to compare the models fit parts (b) and (c). What is the conclusion of the test?