BACS - HW (Week 11)

Let's go back and take another look at our analysis of the cars dataset. Recall our variables:

1. **mpg:** **miles-per-gallon (dependent variable)**
2. cylinders: cylinders in engine
3. displacement: size of engine
4. horsepower: power of engine
5. weight: weight of car
6. acceleration: acceleration ability of car (seconds to achieve 0-60mph)
7. model_year: year model was released
8. origin: place car was designed (1: USA, 2: Europe, 3: Japan)

Did you notice the following from doing a full regression model of mpg on all independent variables?

- *Only weight, year, and origin had significant effects*
- *Non-significant factors cylinders, displacement & horsepower were highly correlated with weight*
- *Displacement has the opposite effect in the regression from its visualized effect!*
- *Several factors, like horsepower, seem to have a nonlinear (exponential) relationship with mpg*

Question 1) Let's deal with **nonlinearity** first. Create a new dataset that log-transforms several variables from our original dataset (called `cars` in this case):

```
cars_log <- with(cars, data.frame(log(mpg), log(cylinders), log(displacement),
log(horsepower), log(weight), log(acceleration), model_year, origin))
```

Note: unless you specify column names, each log transformed column `<col>` is named: `<col>.log`.

a. Run a new regression on the `cars_log` dataset, with `mpg.log.` dependent on all other variables
   i. Which log-transformed factors have a significant effect on `log.mpg.` at 10% significance?
   ii. Do some new factors now have effects on mpg, and why might this be?
   iii. Which factors still have insignificant or opposite (from correlation) effects on mpg? Why might this be?

b. Let's take a closer look at weight, because it seems to be a major explanation of mpg
   i. Create a regression (call it `regr_wt`) of `mpg` over `weight` from the original `cars` dataset
   ii. Create a regression (call it `regr_wt_log`) of `log.mpg.` on `log.weight.` from `cars_log`
   iii. Visualize the residuals of both regression models (raw and log-transformed):
      1. density plots of residuals
      2. scatterplot of `log.weight.` vs. residuals
   iv. Which regression produces better distributed residuals for the assumptions of regression?
   v. How would you interpret the slope of `log.weight.` vs `log.mpg.` in simple words?

c. Let's examine the 95% confidence interval of the *slope* of `log.weight.` vs. `log.mpg.`
   i. Create a *bootstrapped* confidence interval
   ii. Verify your results with a confidence interval using traditional statistics (i.e., estimate of coefficient and its standard error from `lm()` results)

Question 2) Let's tackle **multicollinearity** next. Consider the regression model:

```
regr_log <- lm(log.mpg. ~ log.cylinders. + log.displacement. + log.horsepower. +
                          log.weight. + log.acceleration. + model_year +
                          factor(origin), data=cars_log)
```

a. Using regression and R^2, compute the VIF of `log.weight.` using the approach shown in class
b. Let's try a procedure called *Stepwise VIF Selection* to remove highly collinear predictors.
   Start by Installing the '`car`' package in RStudio -- it has a function called `vif()`
   (note: CAR package stands for Companion to Applied Regression -- it isn't about cars!)
   i. Use `vif(regr_log)` to compute VIF of the all the independent variables
   ii. Eliminate from your model the single independent variable with the *largest* VIF score that is also greater than 5
   iii. Repeat steps (i) and (ii) until no more independent variables have VIF scores above 5
   iv. Report the final regression model and its summary statistics
c. Using stepwise VIF selection, have we lost any variables that were previously significant?
   If so, how much did we hurt our explanation by dropping those variables? (hint: look at model fit)
d. From only the *formula* for VIF, try deducing/deriving the following:
   i. If an independent variable has no correlation with other independent variables, what would its VIF score be?
   ii. Given a regression with only two independent variables (X1 and X2), how correlated would X1 and X2 have to be, to get VIF scores of 5 or higher? To get VIF scores of 10 or higher?

Question 3) Might the relationship of weight on mpg be different for cars from different origins?
Let's try visualizing this. First, plot all the weights, using different colors and symbols for the three origins:

```
origin_colors = c("blue", "darkgreen", "red")
with(cars.log, plot(log.weight., log.mpg., pch=origin, col=origin_colors[origin]))
```

a. Let's add <u>three separate regression lines</u> on the scatterplot, one for each of the origins:
   Here's one for the US to get you started:

```
cars_us <- subset(cars_log, origin==1)
wt_regr_us <- lm(log.mpg. ~ log.weight., data=cars_us)
abline(wt_regr_us, col=origin_colors[1], lwd=2)
```

b. *[not graded]* Do cars from different origins appear to have different weight vs. mpg relationships?

We will investigate these relationships more in class!