Selecting an appropriate model in multiple regression is a very difficult process. This question will give you some practice at doing this by starting with a simple linear regression model, and trying to add to it "useful" explanatory variables. These are called "added variables" and

A researcher develops a multiple regression model to predict the highway miles per gallon of a vehicle based on the city miles per gallon of the vehicle. They run the following codes to do this.

```
> library(mosaic)
> ?mpg
> View(mpg)
> plot(hwy ~ cty, data = mpg)
> mpg.lm <- lm(hwy ~ cty, data=mpg)
> summary(mpg.lm)
> par(mfrow=c(1,2)); plot(mpg.lm, which=1:2)
```

Note that the city miles per gallon is a significant predictor of the highway miles per gallon, p-value < 2e-16.

The researcher is wondering if their model would be improved by including a second explanatory variable. They are considering whether they should add (A) the number of cylinders of the vehicle, (B) the the drive type of the vehicle, or (C) the displacement of the vehicle to the model.

Create added variable plots for each of these three variables using the code

```
> par(mfrow=c(1,3))
> plot(mpg.lm$residuals ~ cyl , data=mpg)
> plot(mpg.lm$residuals ~ as.factor(drv), data=mpg)
> plot(mpg.lm$residuals ~ displ, data=mpg)
```

Select all variables that appear to show pattterns in the residuals. These patterns imply that there is extra information to add to the mpg.lm regression model. In other words, we should add these variable(s) to the current regression.

☐ Number of cylinders

☐ Displacement of the engine

☑ Drive type of the vehicle

To see how a t Test helps to decide this same result, perform three two-variable regressions. Note that in each regression the original explanatory variable of **city** is included with one of the three extra variables under consideration included each time. For example the first regression would use

```
mpg.lm <- lm(hwy ~ cty + cyl, data=mpg)
summary(mpg.lm)
```

Your turn, now run three such regresions in R where you change out **cyl** on two of the regressions for **drv** and then **disp**, respectively.

What is the p-value for the number of cylinders variable when added to the original mpg.lm regression?

p-value = 0.4664

This shows that there is insufficient evidence to conclude that the coefficient for cylinders is different from zero. In other words, cylinders doesn't add anything to the original regression model.

What is the t Test p-value for the displacement of the engine when added to the original mpg.lm regression?

p-value = 0.8167

This shows that there is insufficient evidence to conclude that the coefficient for displacement of the engine is different from zero. In other words, displacement of the engine doesn't add anything to the original regression model.

What is the t Test p-value for the drive type of the vehicle when added to the original mpg.lm regression?

Note that since drive type is a categorical variable with 4-wheel, front-wheel, and rear-wheel drive categories, there are two p-values.

p-value for front-wheel drive types = 1.42e-14 (copy and paste the answer)

p-value for rear-wheel drive types = 9.702e-10 (copy and paste the answer)

This shows that there is sufficient evidence to conclude that the coefficients for front-wheel and rear-wheel drive types is different from zero. In other words, the various drive types add information to the original regression model. They should be included in the model.

Optional: To verify this result, see if you can use mPlot(mpg) to visually recreate a near depiction this regression. Email your code to your instructor to see how you did. Doing this will help you on your Analysis for this week and is recommended even though it is optional.