# G4058 Assignment 3

Due: Friday, March 25, 2016 by 6PM

## 1 Visualization

Load the `countyComplete` dataset from the openintro R package

```
suppressPackageStartupMessages(library(openintro))
data(countyComplete)
```

You can get some information on the variables by executing

```
str(countyComplete)
help(countyComplete)
```

although you should not include those two lines in your writeup.

Create a population growth variable by executing

```
countyComplete$growth <- with(countyComplete, (pop2010 - pop2000) / pop2000 * 100)
```

Now divide the `countyComplete` `data.frame` into a training `data.frame` and a testing `data.frame` by executing

```
include <- sample(nrow(countyComplete), size = 2000, replace = FALSE)
training <- countyComplete[include, ]
testing <- countyComplete[-include, ]
```

Using the training data only create a scatterplot that shows an important relationship between a county's population growth on the vertical axis and some predictor on the horizontal axis. You should use different colors and or plotting symbols to reflect other variables, possibly after utilizing the `cut` function to discretize a continuous variable into a few bins.

## 2 Prediction with Linear Models

Use the `lm` function to estimate a good linear model for population growth in the training `data.frame`, using whatever predictors and interactions thereof you believe are relevant. Note that if you include a great many variables, there may be only a small number of observations that have no missingness.

Then use the `predict` function with `newdata = testing` to generate $\widehat{y}_i$ for each observation in the `testing data.frame`. Subtract $\widehat{\mathbf{y}}$ from $\mathbf{y}$ in the `testing data.frame` to obtain the prediction errors. What is the mean squared error of your model in the `testing data.frame`?

## 3 Prediction with Alternative Linear Models

Use the `lars` function in the lars package to estimate your model via the lasso criterion using the training data.

Use the `pcr` function in the pls package to estimate your model via principal components regression using the training data.

In both cases, use the `predict` function as in problem 2 to obtain the mean squared error in the testing data for the "best" value of the tuning parameter. Which of the three models has the smallest mean squared error?