Problem Set 1: Learning and Regression

Statistical and Machine Learning
Supervised learning deals with data that is well-labeled with an expected target in mind. Unsupervised learning, on the other hand, deals with fairly "messy," unstructured data; you are assuming there is some latent structure to the data and you are attempting to make sense of it or provide some order to its chaos.

Supervised learning takes the existing labeled data and helps to predict future data. Are target, therefore, is predictions about future trends of our data. X's are used to help predict values of Y's. Examples of supervised learning include clasification, where you want to maximize your classification accuracy (putting data into different "zones" or categories), and regression, which seeks to predict or explain the variance between our line of best fit and out actual data with the goal of minimizing loss between the two. The "learning," therefore, is the ability to predict future values not included in our original dataset.

Unsupervised learning is useful for unlabeled data, where the model learns to discover information and patterns to organize the data. Therefore, we aren not predicting future trends, but rather the model is learning to discover how to organize our current existing dataset. Our target, therefore, is order rather than prediction. While we have our initial data, our models are meant to generate new variables for the data to better organize it. Examples of unsupervised learning include clustering/grouping, where you are assuming an underlying latent structure exists and you are trying to recover it, and dimension reduction, where you identify a common variance explained across some common features of the data.

Linear Regression
a. Predict miles per gallon (mpg) as a function of cylinders (cyl). What is the output and parameter values for your model?

```r
data(mtcars)
model1<- lm(mpg~cyl, data=mtcars)
summary(model1)
```

```
##
## Call:
## lm(formula = mpg ~ cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9814 -2.1185  0.2217  1.0717  7.5186
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.8846     2.0738   18.27  < 2e-16 ***
## cyl          -2.8758     0.3224   -8.92 6.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.206 on 30 degrees of freedom
## Multiple R-squared:  0.7262, Adjusted R-squared:  0.7171
## F-statistic: 79.56 on 1 and 30 DF,  p-value: 6.113e-10
```

Slope: For every 1 unit increase in cylinders, there is a -2.8758 decrease in the mpg of a vehicle.
Intercept: When there are no cylinders in a vehicle, the vehicle gets 37.8846 mpg.

b. Write the statistical form of the simple model in the previous question (i.e., what is the population regression function?).

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon$$

where $Y_i = mpg$
β_0 =intercept; when cyl=0
β_1 =slope; for every 1 increase in cylinders there is a -2.8758 decrease in mpg
x_i =value of cyl; number of cylinders
ϵ =random error

$$mpg = 37.8846 - 2.8758(cyl) + \epsilon$$

c. Add vehicle weight (wt) to the specification. Report the results and talk about differences in coefficient size, effects, etc.

```r
model2<- lm(mpg~cyl+wt, data=mtcars)
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2893 -1.5512 -0.4684  1.5743  6.1004
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.6863     1.7150  23.141  < 2e-16 ***
## cyl          -1.5078     0.4147  -3.636 0.001064 **
## wt           -3.1910     0.7569  -4.216 0.000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.568 on 29 degrees of freedom
## Multiple R-squared:  0.8302, Adjusted R-squared:  0.8185
## F-statistic: 70.91 on 2 and 29 DF,  p-value: 6.809e-12
```

When weight is accounted for in the model, the overall intercept of the data (when cyl and weight are both 0) increases from 37.89 to 39.69. Interestingly, the effect the number of cylinders has on mpg decreases, as evident from the decrease in the magnitude of its slope from model 1 (-2.88) to model 2 (-1.51) and its decrease in pvalue from model 1 (6.11e-10) to model 2 (0.001064). Indeed, the number of cylinders decline from a extremely significant pvalue in model 1 to a moderately significant pvalue in model 2, while weight has an extremely significant pvalue in model 2. Therefore, weight has more influence over mpg than the number of cylinders does, having more explanatory value in our model.

d. Interact weight and cylinders and report the results. What is the same or different? What are we theoretically asserting by including a multiplicative interaction term in the function?

```r
model3<- lm(mpg~cyl+wt+cyl*wt, data=mtcars)
summary(model3)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + cyl * wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2288 -1.3495 -0.5042  1.4647  5.2344
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.3068     6.1275   8.863 1.29e-09 ***
## cyl          -3.8032     1.0050  -3.784 0.000747 ***
## wt           -8.6556     2.3201  -3.731 0.000861 ***
## cyl:wt        0.8084     0.3273   2.470 0.019882 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.368 on 28 degrees of freedom
## Multiple R-squared:  0.8606, Adjusted R-squared:  0.8457
## F-statistic: 57.62 on 3 and 28 DF,  p-value: 4.231e-12
```

By including a multiplicative interaction term in the function, we are asserting that the number of cylinders and the weight of a vehicle are not necessarily independent of one another, they are linked in some way. For example, the bigger a vehicle is, the more it will weigh, but it will also need more cylinders.

By including an interaction term in the model, the intercept has increased from 37.88 mpg to 54.3068 mpg when the number of cylinders and the weight are both 0. Additionally, both the number of cylinders and weight of the vehicle are deemed extremely effective for predicting mpg, both yielding extremelly significant pvalues. On that note, both slopes become larger in magnitude. For every 1 increase in cylinders, there's a -3.8032 decrease in mpg, compared to a -2.8758 decrease in model 1 and a -1.5078 decrease in model 2. For every 1 pound increase in weight, there is a -8.6556 decrease in mpg, compared to a -3.1910 decrease in model 2.

Interestingly, the reported intercept t value declines from model 2 to model 3, although it remains extremely significant, and the ta values for cyl and wt remain largely the same, although cyl becomes more significant in model 3.

Non-Linear Regression

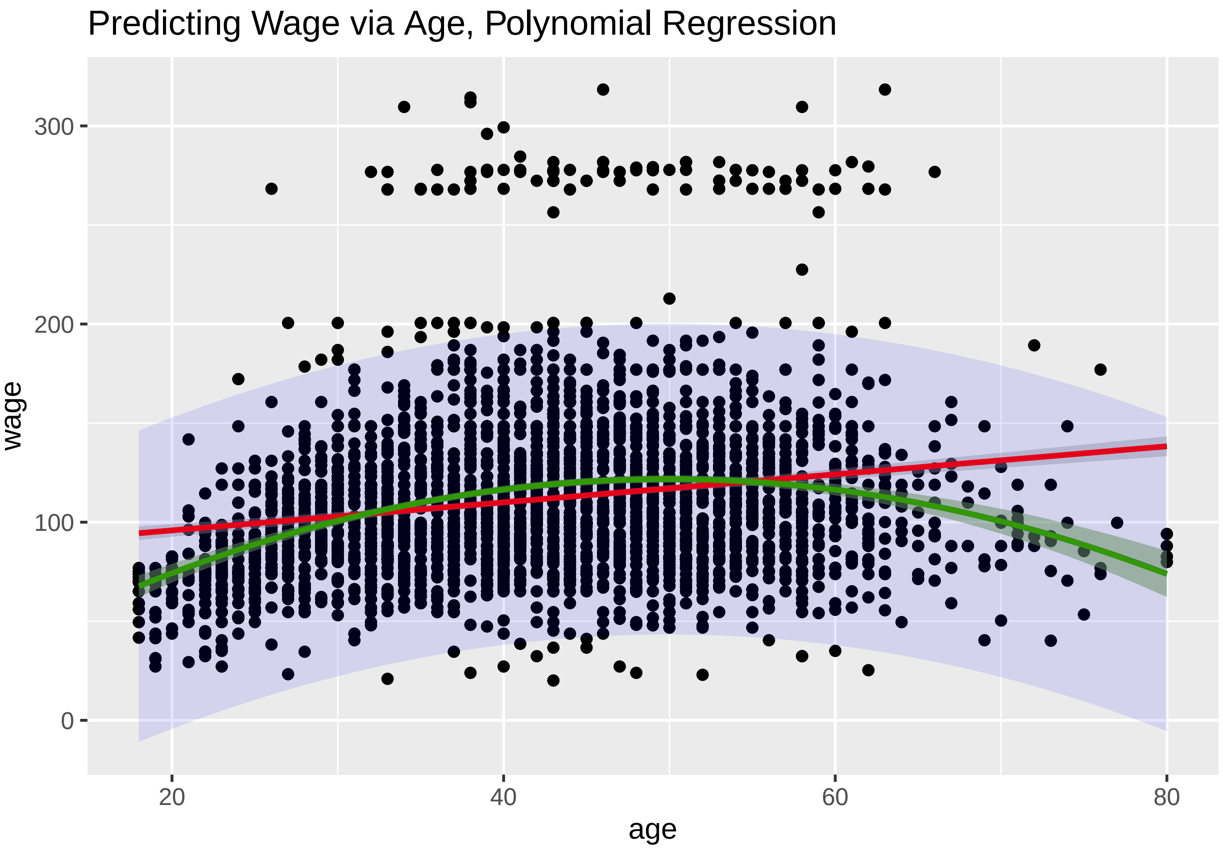
a. Fit a polynomial regression, predicting wage as a function of a second order polynomial for age. Report the results and discuss the output (hint: there are many ways to fit polynomials in R, e.g., I, ^, poly(), etc.).

```r
wage_data<- read.csv("/Users/cemallon/Documents/Machine Learning/PS-1/wage_data.csv")
model1b<- lm(wage~poly(age, 2),data=wage_data, raw=TRUE)
summary(model1b)
```

```
##
## Call:
## lm(formula = wage ~ poly(age, 2), data = wage_data, raw = TRUE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.126 -24.309  -5.017  15.494 205.621
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    111.7036     0.7302  152.99   <2e-16 ***
## poly(age, 2)1  447.0679    39.9926   11.18   <2e-16 ***
## poly(age, 2)2 -478.3158    39.9926  -11.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.99 on 2997 degrees of freedom
## Multiple R-squared:  0.08209,    Adjusted R-squared:  0.08147
## F-statistic:   134 on 2 and 2997 DF,  p-value: < 2.2e-16
```

Intercept: When age is 0 (a newborn), wage is 111.7036.
In this model incorporating the quadratic term, the linear term is still significant. Additionally, the R-squared value indicates that 8.2% of the variability is predicted with a polynomial degree 2. The residual standard error is approximatey 40 years. It appears that both the linear model and polynomaial of the quadratic term are not good fits for this data.
\textbf{b.Plot the function with 95% confidence interval bounds.}

```r
library(ggplot2)
conf.dist<- predict(model1b,newdata=wage_data, interval="confidence",level=0.95)
pred.dist<-predict(model1b,newdata=wage_data,interval="prediction",level=0.95)
wage_data[c("fit","lwr.conf","upr.conf")]<- conf.dist
wage_data[c("lwr.pred","upr.pred")]<- pred.dist[,2:3]
ggplot(wage_data,aes(age,wage))+
  geom_point()+
  geom_smooth(method="lm",formula=y~poly(x,2), col='blue')+
  geom_smooth(method="lm",formula=y~x, col='red')+
  geom_ribbon(data=wage_data, aes(x=age, ymin=lwr.pred, ymax=upr.pred), alpha=0.1, inherit.aes=F, fill=
  geom_ribbon(data=wage_data, aes(x=age, ymin=lwr.conf, ymax=upr.conf), alpha=0.2, inherit.aes=F, fill=
  geom_line(data=wage_data, aes(x=age, y=fit), colour="#339900", size=1)+
  labs(title="Predicting Wage via Age, Polynomial Regression")
```



c. Describe the output. What do you see substantively? What are we asserting by fitting a polynomial regression?

Most of the data in the lower portion of the plot are within the bounds of a 95% interval using the polynomial regression fit. If the data stopped there, the polynomial regression would be an extremely effective fit for our data. However, the data above $250,000 is not accounted for at all by the regression model or the confidence interval, it is totally ignored. By fitting a polyomial regression, therefore, we are asserting that this data is simply outliers that should not matter for our analysis.We are also still asserting that the data and its parameters are linear, however, based on the statistical output and graphical output, it is clear that linear models, even modified polynomial ones, are insufficient to explain the variation in our data; we should be using nonlinear models.

d. How does a polynomial regression differ both statistically and substantively from a linear regression (feel free to also generalize to discuss broad differences between non-linear and linear regression)?

A linear regression is based on a linear model, where each term is either a constant or a parameter multiplied by a predictor variable. When the parameters are linear, the model is linear. A polynomial regression is a modified linear regression model in which the predictor variable is squared. The model is still linear in its parameters, but the parameter being squared produces a curvature to the line of best fit, allowing you to better fit your data. Instead of a constant slope to the data, as is the case in a linear model, a polynomial model allows us to account for a changing slope in our data.

Even though a polynomial regression produces curves, it is still a linear regression. A nonlinear regression is based on a nonlinear function and allows for more flexible curve-fitting regressions. Nonlinear regression functions can have more than one parameter per predictor variable, unlike linear regressions, which can only have one parameter per predictor variable.