# Team Assignment 8
# Linear Regression
# STAT 6021

November 11, 2015


| | |
|---|---|
| Team members: | Don Chesworth |
| | Marcus Rosti |
| | Katherine Schinkel |
| | Mike Voltmer |
| Instructor: | Professor Holt |


## 1 Objective

To analyze the effectiveness of merely subsetting parameters versus subsetting and transforming variables in a least squares regression setting.

## 2 Question

### 2.1 Only Subsets

To subset the data meaningfully, we used a mix of anova and the t test for significance. We started by using regsubsets from the package leaps to identify the most important variables. This yielded age, PPE and DFA as the most important variables based on the exhaustive algorithm. Using those as a starting point, we next used anova to start removing variables. We dropped variables based on their significance and then verified there predictive power using 10 fold cross validation. After getting to the point where we could no longer remove significant variables via ANOVA we were left with,

```
lm2 <- lm(motor_UPDRS ~ age + sex + test_time + Jitter.Abs. + NHR +
            Jitter.PPQ5 + Jitter.DDP + Shimmer.APQ3 + Shimmer.APQ5
           + Shimmer.APQ11 + HNR + DFA + PPE, data=p)
```

### 2.2 Subsets and transformations

This was a more difficult task. The variables had little meaningful transformations. We used a mix of visual inspection and cross validation to confirm the predictive and inferential power of the model. Going off of only our subsets, we started adding transformations on the data and did cross validation to measure the fit. This ultimately yielded an interesting model. Based on cross validation, I raised age to the 13th and PPE to the 7th while removing NHR, PPE and sex.

```
lm3 <- lm(motor_UPDRS ~ poly(age,13) + test_time + Jitter.Abs. +
            Jitter.PPQ5 + Jitter.DDP + Shimmer.APQ5 +
            Shimmer.APQ11 + HNR + DFA + poly(PPE,7), data=p)
```