Group Project
Framingham Heart Study

April 4, 2014

```
frm <- read.csv("frmgham2.csv")
require(GGally)

## Loading required package:  GGally
## Loading required package:  ggplot2
## Loading required package:  reshape
## Loading required package:  plyr
##
## Attaching package:  'reshape'
##
## The following objects are masked from 'package:plyr':
##
##    rename, round_any

require(ggplot2)
```

1 Repeat Observations

The Framingham Heart Study has individuals with observations. For the sake of simplicity we will first remove extra observation such that each individual has only one observation.

```
frm$RANDID <- as.factor(frm$RANDID)
frm1 <- frm[which(!duplicated(frm$RANDID)), ]
```

We are left with 4434 observation, which is still a fairly hefty dataset.

2 Candidate Continuous Variables for Linear Regression

To evaluate continuous variables I will create a subset of the data containing only coninuous variables

```
frm2 <- frm1[, c(3, 4, 5, 6, 8, 9, 12, 13)]
```

In the following example I utilized the GGally package to create a matrix of plots and correlational ceofficients for candidate continuous predictor and outcome variables, after removing observations with missing data.

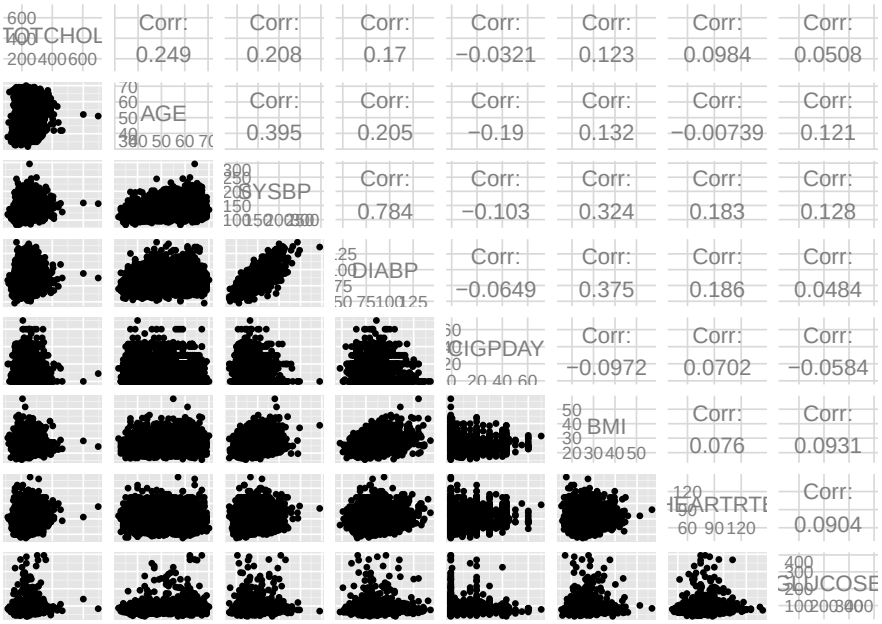


Figure 1: Awesome Image

From the figure we can see that the most consistent correlation between one variable and the others seems to be Systolic Blood pressure. This is a good outcome variable because it intuitively makes sense as an outcome, whereas other variables such as BMI occur in a large part due to environmental factors that aren't factored into this study such as diet. Originally, I was thinking of using Total Cholestrol, however it has relatively weak correlation with many of the factors of interest. Systolic Blood Pressure has the strongest correlation with Diabolic Blood Pressure, however this is a relatively uninteresting relationship to study as it they can be considered more or less colinear. We can move on instead to the next highest correlation which is BMI.

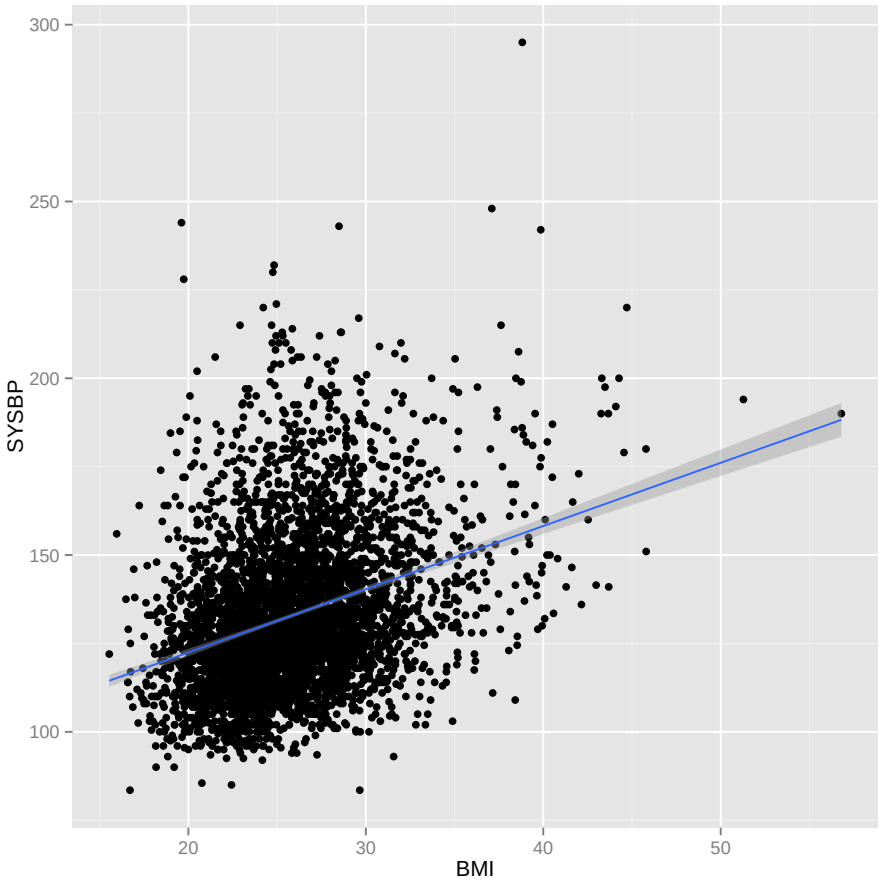
2.1 Simple Liinear Regression Analysis

```
m1 <- lm(SYSBP ~ BMI, data = frm2)
summary(m1)

##
## Call:
## lm(formula = SYSBP ~ BMI, data = frm2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56.21 -14.78  -3.83  10.42 138.90
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  86.6668     2.0286    42.7   <2e-16 ***
## BMI           1.7885     0.0775    23.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.1 on 4413 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.108,Adjusted R-squared:  0.107
## F-statistic:  532 on 1 and 4413 DF,  p-value: <2e-16

qplot(BMI, SYSBP, data = frm2) + geom_smooth(method = "lm", se = TRUE)

## Warning:  Removed 19 rows containing missing values (stat_smooth).
## Warning:  Removed 19 rows containing missing values (geom_point).
```



From our simple linear modeling summary we can see a conclude that there is an association between systolic blood pressure and BMI. The adjusted R^2 is relatively low. It would be interesting to see then the effect of adding new variables to our model.