**More Money, More Problems?**

**An analysis into possible correlation between socio-economic status and hypertension**

Over the ages many poets and philosophers, including the Notorious B.I.G., have pondered whether an increase in wealth corresponds with an associated increase in troubles. While it may be difficult to answer this question in a broad sense, thanks to modern day medical studies and analysis tools we may be able to get closer to an answer in some limited scope. Specifically, if we define a person’s “money” by their reported socio-economic status, and define how much a person’s “problems” are affecting them by the presence or absence of hypertension we should be able to test for correlation between these two factors.

**Problem Statement:**

We will test the hypothesis that a difference in socio-economic status corresponds with a difference in Systolic Blood Pressure after accounting for other health factors such as age, height, weight, and cholesterol levels. Since hypertension is medically defined as having a Systolic Blood Pressure of 140 mm Hg or higher, any factor that is seen to be associated with an increase in Systolic Blood Pressure can be seen as also having an association with hypertension. Since hypertension can also be identified as having a Diastolic Blood Pressure of 80 mm Hg or higher we will also check to see if the findings are affected when using this different variable, but the primary focus will be on Systolic Blood Pressure as it is typically given more attention within the medical community.

**Data Set Used:**

The data set that we will use is taken from an epidemiological heart disease study that was performed on Los Angeles County employees. This included testing these subjects for heart disease and other heart health factors in 1950, and then following up with additional testing of those same subjects again in 1962. Of particular interest is the reported Systolic Blood Pressure, which we will be using this as our response variable. Interestingly enough, the socio economic status was only reported from 1950 and did not have an updated value provided in 1962. Therefore most information gathered by the researchers during their follow up in 1962 will be irrelevant in regards to answering our question of interest.

After reducing the number of variables down to the testing performed in 1950 that is relevant to our study we are left with the following explanatory variables:

1. Age of the subject in 1950 when the recordings were taken and diagnosis given
2. The M.D. who examined the subject and made the diagnosis
3. Height of the subject
4. Weight of the subject
5. Blood Cholesterol Level, recorded as a percentage value rounded to the nearest integer
6. Socio-Economic Status, rated as a level between 1 (highest) and 5 (lowest)

The explanatory variable we are most interested in is the socio-economic status, but since we expect some collinearity between this value and other health variables we want to account for the other variables in our linear regression model. This is to ensure that any correlation found is caused directly be the wealth of the subject itself and not its effect on other health factors. Possible collinearity that we are cautious of may include access to superior health care, superior diet, and other lifestyle differences.

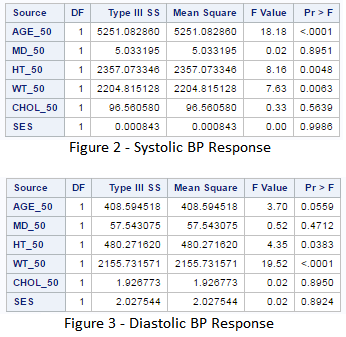
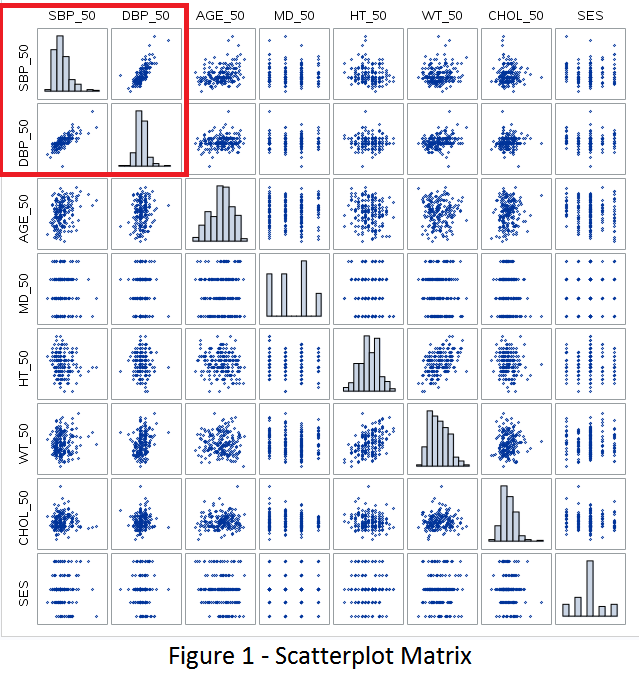
**Scope:**

This was an observational study, and as such any inference that we may draw can only be one of association and not causation. Furthermore, the heart disease study itself was not performed on subjects randomly drawn from the greater population. Rather, it was performed specifically on employees of Los Angeles County. Therefore any inference we make can only be assumed to apply to the Los Angeles County employees on whom the study was performed.

**Exploratory Data Analysis:**

Upon examining the scatter plot matrix of these variables seen in Figure 1 we observe exactly the strong positive linear correlation that we would expect between Systolic Blood Pressure values and Diastolic Blood Pressure values. This will not affect our analysis as these two values will never both be in the same model at the same time. Furthermore, the correlation between each other variable with the two blood pressure variables appears to be largely the same with the one exception of Age showing more of an upward trend with Systolic BP than it does with Diastolic BP. This seems to fit with general medical knowledge that Systolic BP is expected to steadily rise with age in a way that Diastolic BP does not.

Aside from that, there are no other obvious items of note in the matrix. There doesn’t appear to be any extreme outliers. Aside from a weak positive correlation between height and weight there doesn’t appear to be any other collinearity to be concerned with. Finally, while there is some right skew in a few variables there is no evidence against normality that the Central Limit Theorem could not be relied upon to correct given the 200 observations recorded in the dataset.



**Initial Findings Prior to Variable Selection:**

When running a model with all chosen variables that are relevant to our question of interest, we can already see that the F statistic for Socio-Economic Status is almost vanishingly small. As seen in figure 2, for the Systolic Blood Pressure that is of primary interest the F statistic for Socio-Economic Status (SES) rounds to 0.00, giving us an F statistic of 0.9986. This is very convincing evidence that we must fail to reject the null hypothesis that a difference in Socio-Economic status has no impact on a subject’s Systolic BP, and as such would not contribute to hypertension. As seen in figure 3, when we run this same model to explain Diastolic BP we see nearly the same F value for Socio-Economic status and as such have no reason to suspect that our findings will be different if we use Diastolic BP rather than Systolic BP as our response variable.

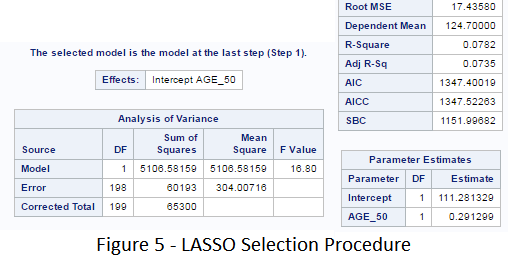
While our initial question of interest already appears to be answered, I am going to continue with variable selection techniques and formal analysis to find out exactly what variables have an effect on Systolic BP, and as such which variables can contribute to hypertension.

**Variable Selection:**

Previously we had chosen a set of explanatory variables that seemed to be relevant to answering our question of interest, creating a model that appeared as follows:

Y­Systolic BP = β0 + βAge + βM.D. Performing Test + βHeight + βWeight + βCholesterol + βSocio-Economic Status

Based upon our initial observation in Figure 2 above it seems that Age is certainly going to be relevant, with Height and Weight potentially relevant as well. Indeed, when we run a LASSO procedure using SAS we find that only Age is necessary for a good predictive model as seen in Figure 5 below:

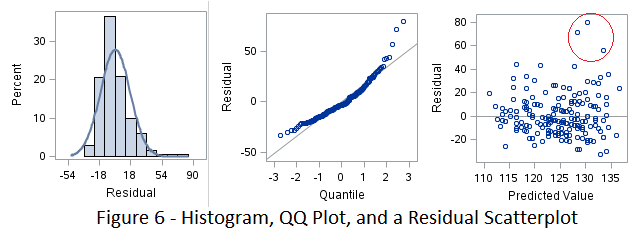


As a parsimonious model is a desirable one, and the findings by LASSO generally agree with our suspicions based upon running our hand-picked model previously, we will now proceed with examining residual plots and otherwise checking assumptions for following simple linear regression model:

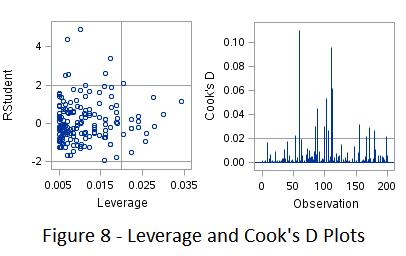
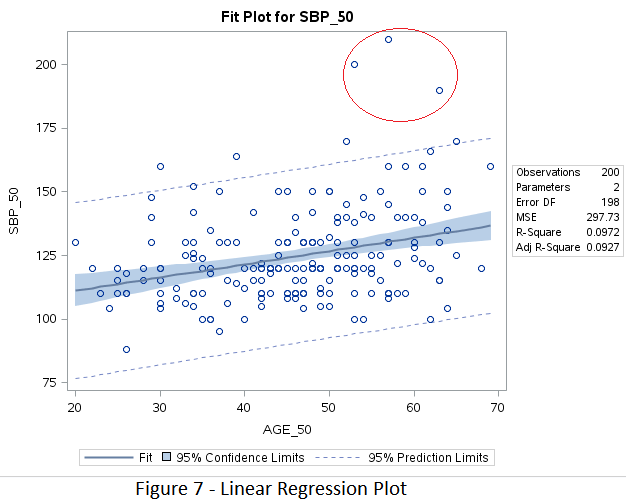
Y­Systolic BP = β0 + βAge

**Residual Analysis:**

Looking at the Histogram, QQ Plot, and Scatter plot of the residuals seen below in Figure 6 there is very little evidence against normality. The histogram is a little right skewed, but close enough to normal for a 200 observation dataset thanks to the Central Limit Theorem. The QQ plot appears to be a straight line for the most part, except for perhaps three outliers in the top right area. Those same three outliers show up in the scatter plot as well, which otherwise appears to be a random cloud.



Investigating those three outliers further, I examined the linear regression itself and found that those three values correspond to observed Systolic BP values of 210, 200, and 190 as seen in the linear regression in Figure 7 below. These values seem suspicious to me since according to modern medical understanding any Systolic BP value of 180 or higher is considered a hypertensive crisis that requires emergency care. That being said, I have no additional evidence to suggest that these data points were erroneous and based upon the leverage plot and Cook’s D values seen in Figure 8 below these points are not influential (No single point has a Cook’s D of 1 or higher, with the highest value being less than 0.12) and so I will choose to leave them in the model.



**Conclusion:**

Based upon the dataset analyzed there is absolutely no evidence at the alpha 0.05 level to reject the null hypothesis that Socio-Economic Status is not associated with Systolic Blood Pressure after accounting for age, diagnosing practitioner, height, weight, and cholesterol (P value 0.99). Furthermore, there is no evidence at the alpha 0.05 level to reject the null hypothesis that Socio-Economic Status is not associated with Diastolic Blood Pressure when accounting for these other factors (P value 0.89), nor is there evidence to reject the null hypothesis that Socio-Economic Status is not associated with Systolic Blood Pressure even when not accounting for any other factors whatsoever (P value 0.78). Therefore we must conclude that there is no evidence that more money is associated with more problems for the Los Angeles County employees that participated in this study.

We did, however, find that there is convincing evidence at the alpha 0.05 level that age is associated with Systolic Blood Pressure (P value <0.0001). It is estimated that for every year older a person is at time of testing, the mean expected Systolic Blood Pressure will be 0.52 points higher (a 95% confidence interval for this change lies between 0.30 and 0.75). This was an observational study on a non-random sample and so we cannot infer a causal relationship and any inferences made will only apply to the Los Angeles County employees that participated in the study. You can find further details on this P value and confidence interval in Figure 9 below.

