Regression Analysis student project: An analysis of *Framingham Heart Study* data

David Saul Davis

July 9, 2012

1 Introduction

The *Framingham Heart Study* (FHS) is a long-term, ongoing cardiovascular study. It is named after the town in Massachusetts in which its subjects live. The study has been in operation since 1948, and is the source of much of our "common knowledge" about the factors that lead to heart disease[1]. The original purpose of the study was to determine which life-style and physical characteristics could predict the occurrence of various types of heart disease. For this project, I will use a subset of the FHS data to study the factors that determine systolic blood pressure (SBP).

2 Data

My dataset originally consisted of the entire FHS "2.20" dataset[2]. This dataset consisted of 40 columns of data for 11,628 individuals. It was decided to focus on a limited number of variables. These are

Sex: Coded as 0 for male and 1 for female.

Total Cholesterol: Total blood serum cholesterol (i.e., HDL + LDL) measured in mg/L. Hereafter referred to a **TCH**.

Age: Integer values of age as of last birthday.

Cigarettes smoker per day: The average number of cigarettes smoked per day. Hereafter referred to as **CPD**.

Body-mass index: Measured in units of kg/cm^2. Here after referred to as **BMI**.

[1] source: http://en.wikipedia.org/wiki/Framingham_Heart_Study

[2] download: http://biostat.mc.vanderbilt.edu/dupontwd/wddtext/index.html#datasets

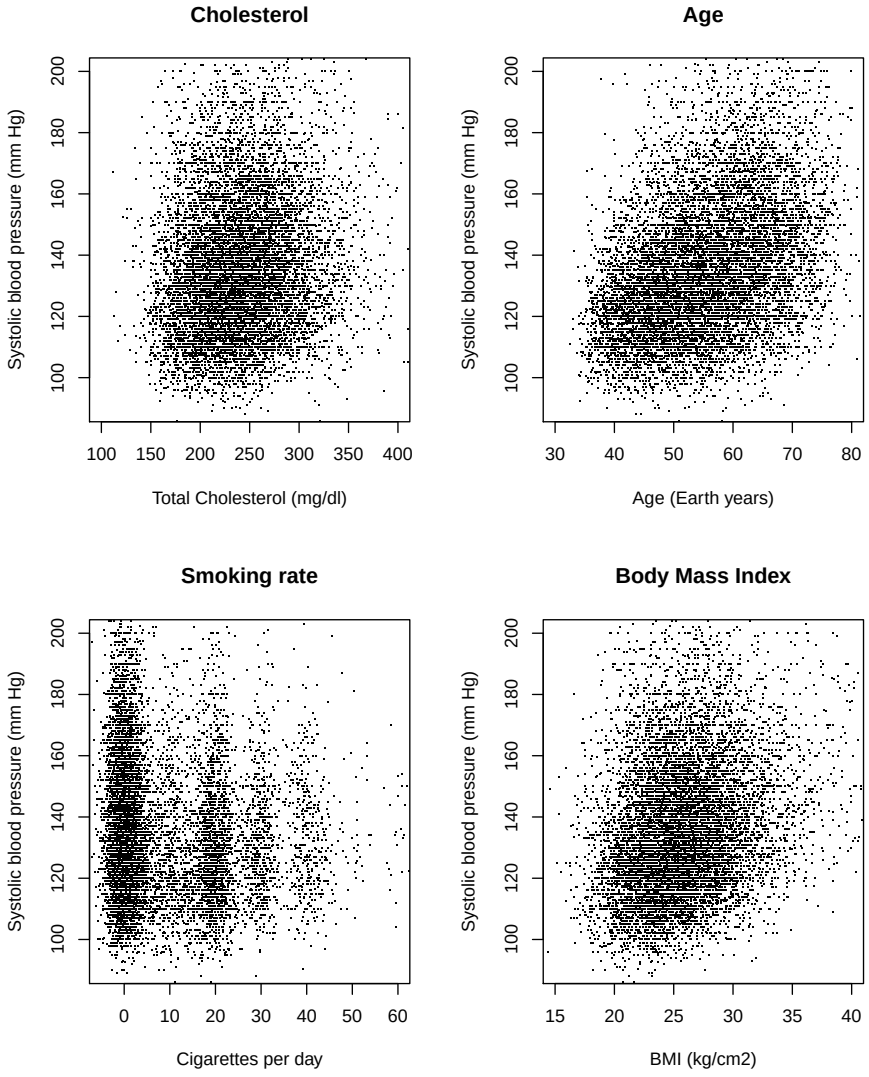

Figure 1: The influence of cholesterol (top left), age (top right), cigarettes per day (bottom left), and BMI (bottom right) on SBP.

Systolic blood pressure: The peak measured blood pressure in units of mm/Hg. Hereafter referred to as **SBP**.

The data was checked for completeness, and any record that did not have valid measurements for every variable listed above was removed from the dataset. After this process, 11,101 individuals remained with measurements in all six variables.

All the variables, except for sex, are numeric. In order to analyze the influence of sex, it was coded as 0 for men and 1 for women. Furthermore, the following two adjustments were made for visual inspection of the data, but were not included during the data analysis. First, the majority of people do not smoke, and therefore have a measurements for the number of cigarettes smoked per day equal to 0. There are also many people that report a multiple of 10 cigarettes (i.e., a "half pack"). Thus, many points were plotted on top of one another. While this does not affect the regression analysis, it makes visual inspection less informative. A random number drawn from a normal distribution with a mean of zero and standard deviation of two was added to the CPD variable in order to spread out the data. Second, a similar adjustment was made to age. The age was recorded at an integer value. A random number drawn from a uniform distribution from zero to one was added to ages to convert them from integers to floating-point numbers. Figure 1 shows all the variables (except for sex) plotted with SBP. All these variables show that SBP is a variable with a large variance. The values for SBP appear more as a cloud than one with strong linear trend. Due to the distribution of the data, we expect a relatively low R^2 value for our relationship. However, due to the large number of data points, we expect a strongly significant relationship between all the variables and SBP.

In order to determine if a transformation should be applied to SBP, a histogram of the data were plotted. The results are shown in the top panel of Figure 2. The distribution of SBP clearly has a positive skew. This suggests we should perform a "lowering" transformation. After some experimentation, it was determined that a log transformation was the most effective. The lower panel of Figure 2 shows the transformed data. The distribution is clearly less skewed. Hence forth, any reference to "SBP" will actually refer to the natural logarithm of SBP.

3 Analysis

3.1 Simple Regression

In order to get a sense of the data, SBP was regressed on every other variable individually. The regression was performed using the `lm()` function in `R`. The significance of the β term was recorded, and the variables were ranked from most significant to least significant. The results are shown in Table 1. Because all the variables are highly significant, I have tabulated the logarithm of the P

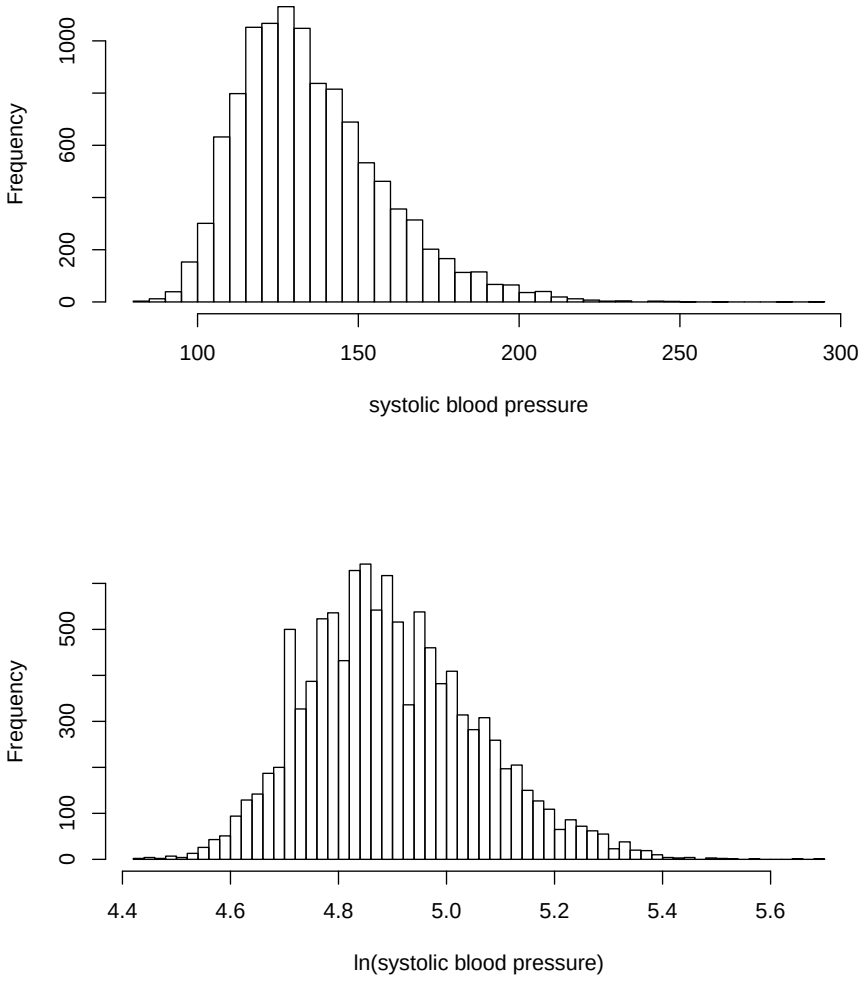

Figure 2: The distribution of SBP before (top panel) and after (bottom panel) a logarithmic transformation was applied to the data. Note the clear positive skew has been largely removed by the transformation.

Variable	log(P)	R^2 (%)
Age	-∞	15.2
BMI	-193	7.6
TCH	-61	2.4
CPD	-27	1.1
Sex	-3	0.1

Table 1: The significance and R^2 of SBP regressed individually on a given variable

value. Furthermore, I have included the R^2 value.

Clearly age plays an important role in determining an individuals SBP, accounting for 15% of the variance. Sex is clearly independent from age, so even though the R^2 is small, we expect it to remain significant when we perform multiple regression. Similarly, while there could be a weak correlation between CPD and age, we expect CPD to remain significant as well. However, both BMI and TCH could be strongly correlated with age. Both these variables could be found to be insignificant when multiple regression is performed.

3.2 Multiple Regression

My approach to multiple regression will be as follows. I will add variables one by one, starting with the most significant, moving to the least significant. If the adjusted R^2 value is greater than in the previous step, I will consider the model improved, and retain the extra variable. If however, the adjusted R^2 worsens, I will discard that variable, and move will regress SBP first on age, then age and BMI, and so forth.

Using the process described above, every variable except CPD added to the quality of the model. Equations representing the complete set of models tested are shown below.

$$\begin{aligned}
\text{Model A}: &\ SBP = \alpha + \beta_1 \cdot Age \\
\text{Model AB}: &\ SBP = \alpha + \beta_1 \cdot Age + \beta_2 \cdot BMI \\
\text{Model ABC}: &\ SBP = \alpha + \beta_1 \cdot Age + \beta_2 \cdot BMI + \beta_3 \cdot TCH \\
\text{Model ABCC}: &\ SBP = \alpha + \beta_1 \cdot Age + \beta_2 \cdot BMI + \beta_3 \cdot TCH + \beta_4 \cdot CPD \\
\text{Model ABCS}: &\ SBP = \alpha + \beta_1 \cdot Age + \beta_2 \cdot BMI + \beta_3 \cdot TCH + \beta_4 \cdot sex
\end{aligned}$$

The coefficients for each of these models (including the adjusted R^2 and error terms were determined using the `lm()` routine. The results are reported in Table 2. Examining Table 2, one can see that the majority of the explanatory power is derived from two variables—BMI and Age. While the ABCS model doe indeed have more explanatory power than the A model, it has many more parameters. Using the principles of parsimony, one might prefer the AB model. This has almost all the explanatory power of the higher order models, but without the extra complexity.

Model	α	β_1	β_2	β_3	β_4	Adjusted R^2
A	4.5402	.0065	NA	NA	NA	15.20
AB	4.2970	.0063	.0099	NA	NA	21.62
ABC	4.2447	.0061	.0097	.0029	NA	22.26
ABCC	4.2412	.0061	.0098	.0029	.0001	22.26
ABCS	4.2405	.0061	.0098	.0027	.0099	22.35

Table 2: The results of multiple regression.

Having settled on a simple model that uses only age and BMI to predict SPB, the final step is the test for higher-order cross terms. We do this in the model AB2, which has to following form:

$$SBP = \alpha + \beta_1 \cdot Age + \beta_2 \cdot BMI + \beta_3 \cdot age \cdot BMI$$

The adjusted R^2 of Model AB2 is, 20.83%, lower than that of the Model AB. The age/BMI cross-term has a high p-value of 0.415, and is therefore *not* statistically significant. Furthermore, if low-order cross-terms are not significant, higher-order will not be significant either. Thus, no models were constructed beyond AB2. The Model AB is preferred over AB2, and will be considered the final model.

3.3 Residuals

The final step to testing the quality of our fit is to examine the fitted values, and in particular the differences between the fitted values and the true values of SBP. The basic premise is that, if the fitted model is perfect, the only reason a value is not predicted perfectly is due to the random noise term. Theory predicts that this noise should be distributed normally. Thus, we want to examine the residuals for trends, and if we find that the residuals are trend-free, we want to examine their distribution to determine it they are consistent with a normal distribution.

A very basic test is to plot the residuals versus the fitted values. This is shown in the upper panel of Figure 3. There is no obvious trend here. The residuals were then examined using a quantile-quantile (QQ) plot, shown in the bottom panel of Figure 3. This plot shows that the residuals are quite close to normally distributed, though slightly positively skewed.

4 Conclusion

We have shown that roughly 20% of the variance of people's blood pressure can be explained using a model that references their age and BMI only. Due to the large sample size, the model components are extremely statistically significant. Several factors (TCH, CPD, sex) we not included in the final model because they added significantly to the complexity of the model for only marginal

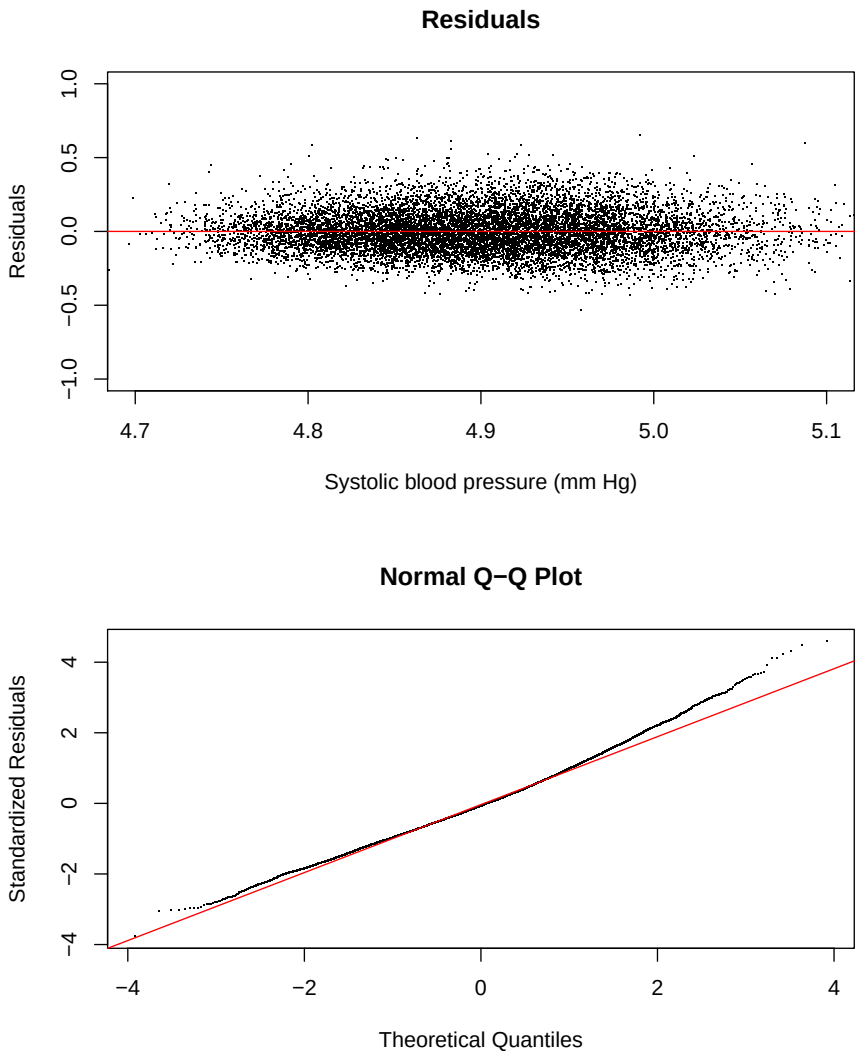

Figure 3: Analysis of model residuals. (upper) The residuals vs. the fitted values. (lower) A QQ plot showing that the residuals are slightly positively skewed, but close to normally distributed.

gains in its explanatory power. Cross-terms of age and BMI were tested, but found to be unimportant. The final model is:

$$SBP = e^{4.2970 + .0063 * age + .0099 * BMI.}$$

Using my age (33.5) and my BMI (23.4), this formula predicts my systolic blood pressure to be 114.4.

The point that I have not yet addressed is that 80% of the variance is *not* explained by this model. The source of this variance is currently unknown. There are likely strong genetic factors that contribute to blood pressure. Perhaps parents blood pressure could account for much of the unexplained variance. Furthermore a more refined measure of how "in shape" an individual is would be a better predictor than BMI.