**LOGISTIC REGRESSION ANALYSIS OF**

**FRAMINGHAM HEART STUDY**

Team #4:

Jasmine Barrera, Gabriel Cunningham, Seyoung Jung,

Travis Klipp, Joe Sanchez, Nathaniel Siriwattana

6 December, 2018

San Diego State University

Stat 680A. Advanced Biostatistical Methods

Fall 2018

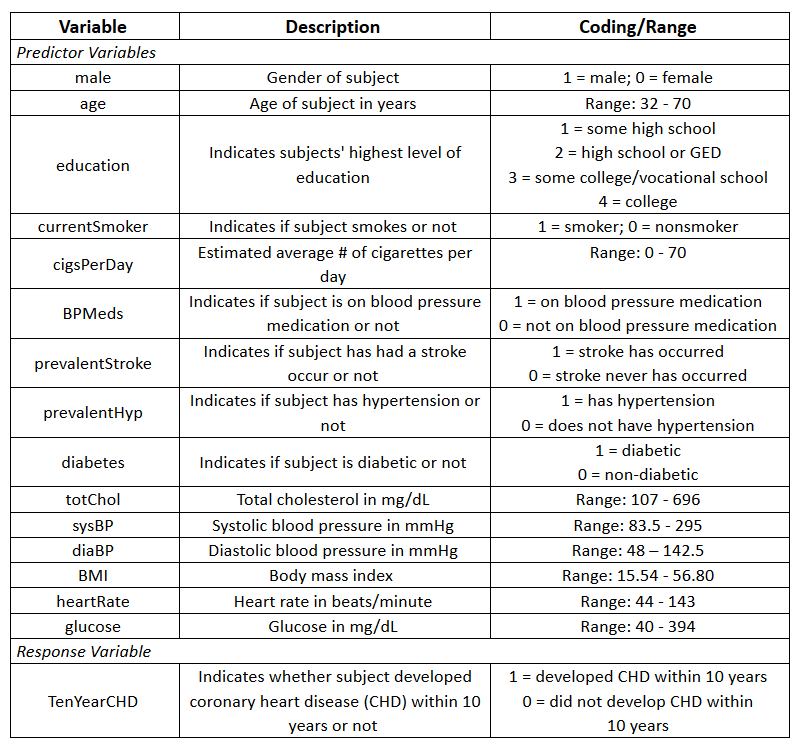
Dr. Kung-Jong Lui

**Introduction**

Coronary heart disease remains one of the top causes of mortality in first world countries. It is for this reason that there has been much interest in identifying and reducing risk factors in order to reduce the prevalence of Coronary heart disease.

The data analyzed in this report comes from the Framingham Heart Study, an ongoing cardiovascular cohort study that began in 1948. The dataset includes 15 predictor variables and a single response variable (Table 1):

Table 1: Framingham heart study predictor and response variable descriptions.



The aim of this project is to develop a logistic regression model which can accurately determine whether or not a person is likely to develop coronary heart disease within 10 years, based on their risk factors. The other objective includes demonstrating other biostatistical analysis methods, such as stratified data analysis and Wald hypothesis testing.

**Theory**

The relationships between explanatory variables and categorical outcomes can be described using logistic regression, where the probability of response is modeled for categorical data. In the case of binary logistic regression, there are multiple explanatory factors Xi and a single dependent variable Y that follows a Bernoulli distribution, such that it takes on a value of either 0 or 1. The model can be summarized as follows, with an error term 𝜀 that follows the standard normal distribution:



The corresponding probability parameter for the distribution of Y is *p*, where the probabilities of the two outcomes are P(Y=1) = *p* and P(Y=0) = *1-p.* The mean value of Y is *p*, which can undergo a logit transformation, where logit {*p*} = ln{*p / (1-p)*} is equal to the log odds parameter ( 𝜆) and can be written as the following:

**

The inverse logit transformation results in the equation



which allows a linear logistic regression model of the following form to produce a probability between 0 and 1 (*Fleiss, 2003, p. 284-293*):

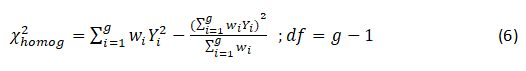


It is important to note that there are underlying assumptions which must be true in order to use a binary logistic regression model. First, a binary logistic regression model requires that the dependent variable Y be binary, as illustrated above. Second, the observations must be independent of each other. Third, there must be none or hardly any multicollinearity among the independent variables. There must also be a linear relationship between the independent variables and the log odds parameter (𝜆) and a large sample size is also required.

Sometimes it can be useful to use stratified data analysis as a means of determining whether or not the measure of association is the same across all strata. There are multiple measures of association (Yi) that can be used in this type of analysis, some of which are risk difference (RD), natural log of risk ratio (lnRR), and natural log of odds ratio (lnOR); for this report, the focus will be on lnOR. In stratified data analysis, the chi square of homogeneity (𝜒2 homog) is calculated in order to test the homogeneity of the measure of association across all stratum. If the measure of association is determined to be homogenous, then the chi square of association (𝜒2 assoc) can be calculated in order to determine the significance of the mean measure of association. Alternatively, a confidence interval can be constructed for the mean measure of association (Yio) in order to test whether or not there is significant association. For example, if the measure of association is lnOR, then the following equation would be used to test the homogeneity of two stratum:



If the measure of association is not equal across all stratum (test of homogeneity null hypothesis is rejected), then the significance of the stratum must be calculated individually in order to determine whether or not an interaction term is necessary. The following equations are used to find the mentioned chi square statistics (*Fleiss, 2003, p. 236-237*):





Another method used to determine whether or not a predictor variable is significant and should be included in the model is the Wald test. This method relies on the assumption that the data is large. The hypothesis test can be written as Ho: 𝛽i = 0 vs. Ha: 𝛽i ≠ 0 , with respect to the construction of a model. The test statistic which is used to test the hypothesis is the Wald test statistic *z*, where



If we fail to reject the null hypothesis, then the parameter can be eliminated from the model. If the null hypothesis is rejected, then the parameter is considered significant and will remain in the model (*Fleiss, 2003, p. 300-304*).

**Exploratory Data Analysis**

This dataset has 16 columns (variables) and 4,240 rows (observations). In this part, we will perform EDA, or Exploratory Data Analysis, to find patterns or outliers, so that we can analyze this dataset more efficiently. We will compare 10 of independent variables to our target variable (CHD) which is the 10-year risk of CHD, or coronary heart disease.

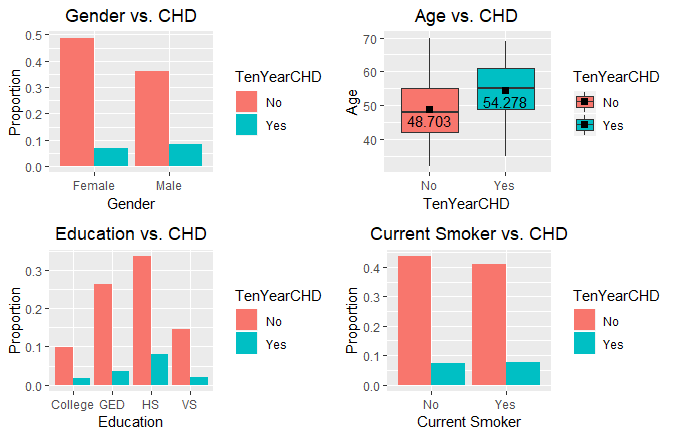


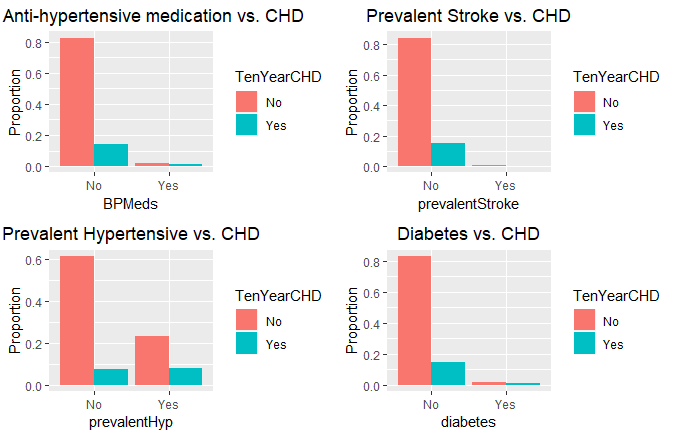
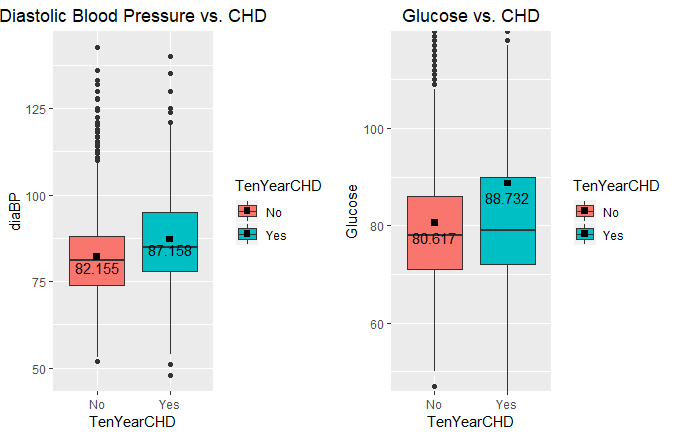
Figure 1 on the right shows us four different graphs. First, the bar plot of Gender vs. CHD tells us that there are relatively more women than men in our observations. But both genders have nearly the same CHD rates. And the top right graph indicates that the people with CHD is roughly 5.58 years older than the people without CHD in average. Also, the bottom left bar plot has four different x-axis labels: College, GED (High School/GED), HS (some High School), and VS (Some College/Vocational School). As you can see, the highest level of education that the participants have is high school, followed by GED, vocational school, and college. But the CHD rate between each education level doesn’t show any significant difference. And surprisingly, the bottom right graph shows that almost half of the participants are current smoker, but the CHD rates between the two are also almost the same.

Figure 2 on the left also contains four different bar plots. The upper left graph is a bar plot of the use of anti-hypertensive medication at exam versus CHD. One can easily notice that approximately 80% of participants do not use the medication at the exam and do not have CHD, whereas 15% of the participants use the medication but have CHD. And the upper right bar plot indicates that almost none of the participants have prevalent stroke. Also, the bottom left graph tells us that there are much more participants who don’t have prevalent hypertensive than who have it, but interestingly, the CHD rates are almost the same for both. And the bottom right bar plot shows us that more than 90% of participants do not have diabetes. However, it also shows that if one has diabetes, they are much more likely to have CHD.



In Figure 3, we have two box plots for Diastolic Blood Pressure versus CHD, and Glucose versus CHD. The left box plot shows us that the participants who do not have CHD have 82.155 of diastolic blood pressure in average, whereas the participants who have CHD have 87.158 in average. Also, the participants who do not have CHD have 80.617 of glucose level in average, and the participants who do have CHD have 88.732 of glucose level. It is worth noting that the mean value of glucose level for the participants who have CHD is much higher than the middle quantile and it is very close to the upper quantile, unlike the participants who don’t have CHD.

Before finalizing our main model, we created male and female stratum and calculated the chi square statistic for homogeneity to determine if an interaction term is needed. In addition, we performed a chi square test for association to determine if developing CHD within 10 years is associated with having a stroke in the past.

*χ*12 = 3.841 > .00217, so we fail to reject the null hypothesis that the measure of association are equal for the male and female stratum. Therefore, we do not need an interaction term with gender in our main model.

Since we found the measure of association to be equal for the male and female stratum, we calculated the chi square statistic for association to determine whether developing CHD within 10 years is associated with having a stroke in the past. 

*χ*12 = 3.841 < 7.74, so we find evidence to support the alternative hypothesis that developing CHD within 10 years is positively associated with having a stroke in the past. While this simple model yields a strong association, we are interested in creating a more complex model to predict developing CHD more accurately.

Our next method will be to fit a logistic regression model to our data. Recall, our model assumptions rely on a linear relationship between the log odds relationship of the predictor variables and response variable, a large sample size, independent responses, and parameters that are not highly correlated. We found there’s a high correlation between prevalentHyp and sysBP (please see correlation table below), diaBP and sysBP, and Current Smoker and CigsPerDay. From prior analysis, we found prevalentHyp, diaBP, and currentSmoker aren’t too relevant to our analysis so we dropped these variables. It’s also important to note that our response variable is a categorical variable with 2 values (disease and non-disease). This will be a binomial logistic regression model. For comparison, we first examine the rate of CHD and non-CHD in our dataset. Fitting a baseline model by dividing the CHD rate by the total number of observations, we find 84.77% of the data has a response rate of 0 for the 10 year CHD value. We first run a model including all of the variables:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.7793409 0.7480617 -11.736 < 2e-16 \*\*\*

gender 0.5075773 0.1300270 3.904 9.48e-05 \*\*\*

age 0.0625616 0.0078726 7.947 1.91e-15 \*\*\*

education2 -0.0615935 0.1445117 -0.426 0.6699

education3 -0.3579743 0.1901681 -1.882 0.0598 .

education4 -0.0363587 0.1988856 -0.183 0.8549

cigsPerDay 0.0200458 0.0049845 4.022 5.78e-05 \*\*\*

BPMeds 0.4620016 0.2726924 1.694 0.0902 .

prevalentStroke 1.4047263 0.5943422 2.363 0.0181 \*

diabetes -0.0009784 0.4334187 -0.002 0.9982

totChol 0.0016792 0.0013475 1.246 0.2127

sysBP 0.0176370 0.0028173 6.260 3.85e-10 \*\*\*

BMI 0.0177480 0.0148615 1.194 0.2324

heartRate -0.0048721 0.0050807 -0.959 0.3376

glucose 0.0064825 0.0027651 2.344 0.0191 \*

We can suspect that many of these variables can be dropped by looking at the p-values. Many of them have values greater than .05, suggesting that they’re not too significant in this model. It’s likely we will only need a combination of gender, age, cigsperday, totChol, sysBP, and glucose in our model since the p-values are very small. Evaluating this model towards the test data set portion yields an 84.6% prediction rate, which is relatively decent, but since there are so many variables in the model, we can probably simplify and perhaps optimize our prediction rate.

We now perform stepwise selection to optimize our model. We notice that 5 of our variables were dropped, as expected:  
Initial Model: TenYearCHD ~ gender + age + education + cigsPerDay + BPMeds + prevalentStroke + diabetes + totChol + sysBP + BMI + heartRate + glucose  
Final Model:  
TenYearCHD ~ gender + age + cigsPerDay + BPMeds + prevalentStroke + sysBP + glucose  
  
Education, BMI, heartrate, diabetes, and totChol were dropped from our model. We’re left with gender, age, cigsperday, BPMeds prevalentstroke, sysBP, and glucose as parameters with the following coefficients:  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) -8.573136 0.511246 -16.769 < 2e-16 \*\*\*  
gender 0.530746 0.126714 4.189 2.81e-05 \*\*\*  
age 0.064883 0.007622 8.512 < 2e-16 \*\*\*  
cigsPerDay 0.019261 0.004920 3.915 9.04e-05 \*\*\*  
BPMeds 0.488001 0.272026 1.794 0.07282 .   
prevalentStroke 1.348480 0.588618 2.291 0.02197 \*   
sysBP 0.018497 0.002662 6.949 3.67e-12 \*\*\*  
glucose 0.006270 0.002080 3.015 0.00257 \*\*   
  
The prediction error rate is 85.15%, which is slightly higher than our first model using all of the variables. We gain over .5% prediction rate, and since this newer model after stepwise selection is a lot simpler and easier to understand, this model is better for our analysis. We also see nearly all of the variables in the model are highly relevant to the model since they have very low p-values.  
Wald test for BPMeds  
 in glm(formula = TenYearCHD ~ gender + age + cigsPerDay + BPMeds +   
 prevalentStroke + sysBP + glucose, family = binomial(link = "logit"),   
 data = train\_data)  
F = 3.218258 on 1 and 2525 df: p= 0.072941   
 Wald’s tests suggest that removing BPMeds from the model may not be significant, so we run a model with only gender, age, cigsPerDay, prevalentStroke, sysBP, and glucose. Removing it for sake of simplicity results in an increase in the prediction rate to 85.7%, so we remove it from our model, and our final model becomes the following:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.712661 0.504705 -17.263 < 2e-16 \*\*\*

gender 0.520345 0.126346 4.118 3.81e-05 \*\*\*

age 0.065216 0.007615 8.564 < 2e-16 \*\*\*

cigsPerDay 0.019152 0.004917 3.895 9.83e-05 \*\*\*

prevalentStroke 1.425864 0.581526 2.452 0.01421 \*

sysBP 0.019577 0.002588 7.563 3.93e-14 \*\*\*

glucose 0.006279 0.002057 3.053 0.00227 \*\*

**Conclusion**  
It should also be noted that since we determined there’s homogeneity between male and females for prevalence of stroke as it affects CHD, we didn’t include those terms as interaction terms in our model. Our log odds equation is thus -8.712661 + .520345Gender + .065216Age + .0192152cigsPerDay + 1.425864prevalentStroke +.019577sysBP + .006279glucose. All of the odds ratios for the parameters in our model suggest at least a slightly positive association with CHD, and all of the variables are highly relevant to our model given our p-values. The odds ratios suggest prevalentStroke has a high association with our model, which is consistent with our prior analysis. We also see that gender and age are notable in our model, and that older people are more likely to suffer from CHD (which is unsurprising given our intuition) and males are slightly more likely to suffer from CHD. Studies like this are important for obvious reasons as to prevent CHD and for doctors to address the people who are predisposed to these kind of diseases.

**Reference**

1. Joseph L. Fleiss et al., *STATISTICAL METHODS FOR RATES AND PROPORTIONS*, by JOSEPH L. FLEISS, 3rd ed., JOHN WILEY & SONS, 2003, pp. 236-237, 284–293, 300-304.