A Logistic Regression of Coronary Heart Disease

Statistics 138
Project II
XXX

Soloman Wong
SID: XXX
XXXX
SID: XXX
XXXX
SID: XXXXXX
XXXX
SID: XXX

Abstract

The goal of this study is explore the relationship between age and the presence or absence of Coronary Heart Disease in this sample population. We expect to see the higher the age, the more likely that person has Coronary Heart Disease. As a result, we conduct a logistic regression to to determine the association between CHD evidence and an individual's age with use of SAS.

I. Introduction

This study analyzes the presence of Coronary Heart Disease (CHD) from 2013. This dataset is included from Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X.: *Applied Logistic Regression (Third Edition)*. These data are copyrighted by John Wiley & Sons Inc. It contains two category variables: Age and CHD. In this study, the total of 100 participants from the age of 20 to 69 are invited. They have to do a physical test in order to identify if they are either absent (0) or present (1) to the Coronary Heart Disease by using binary numbers. These data have been used in an introduction to logistic regression. These data illustrate the differences between logistic regression and other regression techniques as well as demonstrate the usefulness and purpose of logistic regression.

II. Methodology

While age is a continuous variable and CHD is not, so that the linear regression methods seems to be not appropriate. Logistic regression can describe the relationship between a categorical outcome and a predictor variable. Therefore, because there is a general trend that CHD rates increase with age, it is reasonable to use logistic regression to model the effect of age on CHD rates. In our case, we will model the presence or absence of coronary heart disease using age as predictor variable (x). We obtain the logit function to be:

$$logit(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1$$

which we calculated β0 to be -5.3095 with a standard error of 1.1337 and and β1 is 0.1109 with standard error of 0.0241. The confidence interval for β0 is [-7.7256, -3.2459] and β1 is [0.0669, 0.1620]. Hence, we get the estimation logit(π(x)) is :

$$logit(\pi(x)) = -5.3095 + 0.1109x$$

The results are significant and it says that the log-odds of CHD per one year increase in age is 0.1109 and the change could be as little as 0.0669 or as much as 0.162 with 95% confidence.

The estimated probability of age given the explanatory variable(presence or absence of CHD):

$$\pi(x) = \frac{e^{-5.3095+0.1109x}}{1+e^{-5.3095+0.1109x}}$$

The odd ratio is exp(0.1109) = 1.117, which means that the chance that a person will have CHD increases 11.7% for every year of age. Thus there is a strong association between age and CHD, and the model seems to fit well.

We use PROC CATMOD and PROC LOGISTIC to generalized logits. The parameter estimates from the CATMOD procedure are the same as those from a logistic regression program such as PROC LOGISTIC. The chi-square statistics and the predicted values are also identical. In the binary response, PROC CATMOD can be made to model the probability of the maximum value by organizing the input data so that the maximum value occurs first in the PROC CATMOD statement.

III. Results

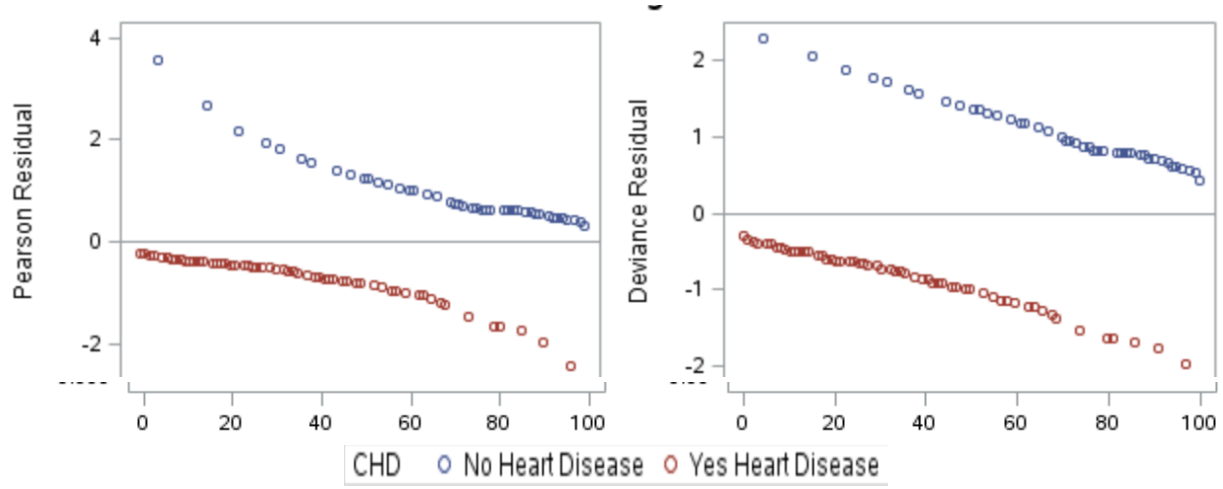
The odds ratio estimates produce two set of confidence intervals, based on the profile-likelihood and Wald test approach:

- 95% Profile-Likelihood CI: [1.069, 1.176]
- 95% Wald CI: [1.066, 1.171]

Both confidence intervals are too close to each other. Since the value of 1 is not within the interval and the both estimates are greater than 1, we reject the null hypothesis to conclude that age and CHD have association between each other.

The standardized Pearson residuals and deviance residuals demonstrated on the next page show a problem a overestimation or underestimation. The diagnostics contain some outliers such as 4 and -3, with all other variables being less than plus or minus 2.5. After more examination of the residuals, we concluded that for CHD absent, there is an overestimation as all residuals are positive. Additionally, the there is a pattern of overestimation in that residual values decrease for no heart disease. In contrast, the Pearson residuals are negative for those with CHD, implying a degree of underestimation. Although AGE is correlated and slightly predicts CHD, our residual analysis suggests that AGE may not be the best indicator of having CHD.

Standardized Pearson versus Age



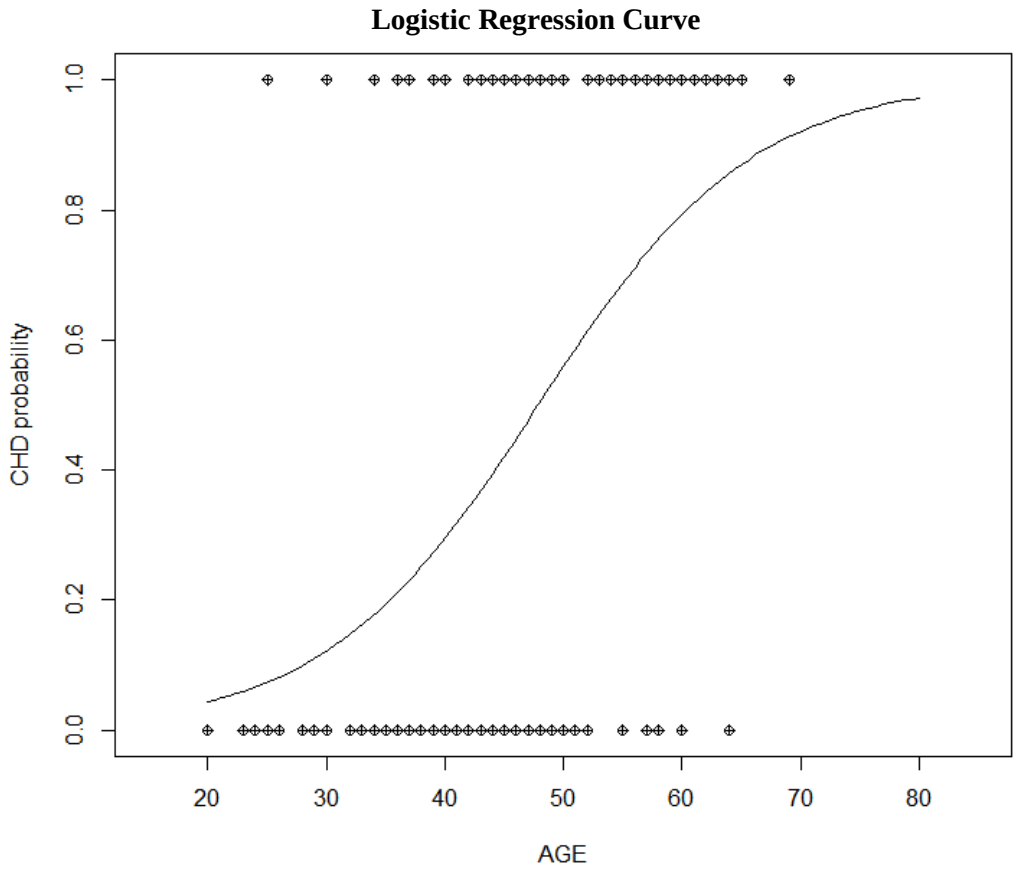
Goodness of Fit

In analyzing the logistic regression through PROC LOGISTIC, the R-squared value is 0.2541 and max rescaled R-squared is 0.3410. The formula for R-squared for a logistic regression is defined as

$$\square^2 = 1 - [\frac{\square(\square_0)}{\square(\square_\square)}]^{2/\square} = 0.2541$$

where $\square$ is the sample size, $\square(\square_0)$is the likelihood for a model with only intercept, and $\square(\square_1)$is the likelihood for the selected model with age as a predictor variable. The Cox Snell R-squared value of 0.2541 suggests that there is approximately 25.41% variation in response variable explained by the model. It is important to note that low R-squared values are generally the norm when dealing with logistic regressions since R-squared is based on comparisons of predicted values from the fitted model to those from the base model. Thus, R-squared is does not accurately assess the goodness-of-fit in the CHD and AGE model.

The criterion for goodness of fit will be based on both analysis of AIC and the Hosmer and Lemeshow goodness of fit test at the 0.05 significance level. The Akaike Information Criterion (AIC) of the intercept only model is 138.663 and the AIC of the model with the age variable is 111.353. The decrease in AIC after taking age into account shows that age has a better fit effect on the the model and thus will be included in our analysis thus far. The Hosmer and Lemeshow Chi-Square test statistic was 0.697 (df = 7) with a p-value of 0.9984. Therefore at α= 0.05, we cannot reject the null hypothesis that the model is fitting the data.



IV. Conclusion and Findings

In conclusion, we find that age and CHD have associations to each other which corroborates our initial hypothesis. As time progresses and a person's age increases, there is a highly likelihood that he or she will be get coronary heart disease. This is a logical conclusion in that a decreasing rate in metabolism is strongly associated with an increase in age, and a decreasing rate in metabolism is also strongly associated with an increased chance in receiving coronary heart disease.

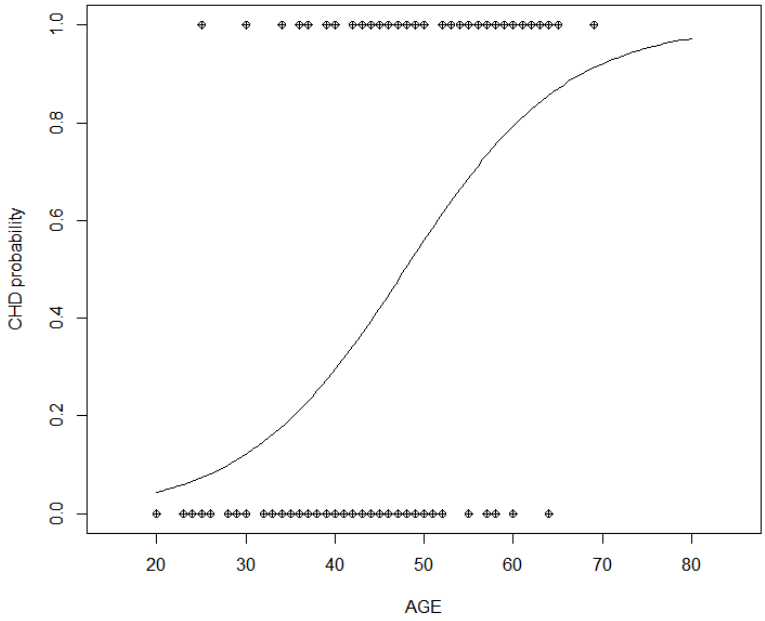
Based on the result from the research that we have conducted, we can check how different components can reduce the rate of CHD as ages increase. In the logistic regression curve shown in the Appendix, we can tell that as age increases, the predicted probability of receiving CHD increases.

Next Steps

In the residual plot, we notice that there is presence of outliers. Specifically, there is an outlier for CHD present in a twenty-five year old and CHD absent in a sixty-four year old. It is reasonable to remove the outliers and compare to check if the result is consistent with the initial

findings. We can consider adding other variables, binary or continuous, and then implement a multiple logistic regression to better predict the best indicators of receiving coronary heart disease.

R Appendix:




```
chdage <- read.csv("C:/Users/Soloman/Desktop/uc davis/2016 fall quarter/sta 138/chdage.csv")

model=glm(CHD~AGE,data=chdage,family=binomial(link=logit))
b0=coef(model)[1]       # extract the estimate of intercept #
b1=coef(model)[2]

estprob=function(x){
  z=exp(b0+b1*x)/(1+exp(b0+b1*x))
  return(z)
}

curve(estprob,from=20,to=80,xlim=c(15,85),ylim=c(0,1),ylab="CHD probability",xlab="AGE")
points(chdage$AGE,chdage$CHD,pch=10)
```

SAS Code:

```
/* Generated Code (IMPORT) */
/* Source File: chdage.xls */
/* Source Path: /folders/myfolders */
/* Code generated on: 11/18/16, 4:41 PM */

%web_drop_table(mydata);


FILENAME REFFILE '/folders/myfolders/chdage.xls';

PROC IMPORT DATAFILE=REFFILE
        DBMS=XLS
        OUT=mydata;
        GETNAMES=YES;
RUN;

PROC CONTENTS DATA=mydata; RUN;


%web_open_table(mydata);

proc format;
        value chd 0 = 'No Heart Disease' 1 = 'Yes Heart Disease';
run;

proc plot;
    plot CHD*Age;

proc catmod order=data;
        direct AGE;
        model CHD = AGE / ml nogls pred = prob;
        format CHD chd.;
        title 'Logitic Regression of Coronary Heart Disease Data Using CATMOD Procedure';
run;

proc logistic DESCENDING;
        model CHD = Age / link=logit rsq lackfit plcl plrl risklimits influence iplots scale = none
aggregate;
        output out = tasko p = predprob reschi=pearson resdev=deviance ;
```

```
        format CHD chd.;
        title 'Logistic Regression of the Coronary Heart Disease Data Using LOGISTIC
Procedure';
run;
```