**BIST 0535 Biometrics Computing Fall 2017**

**Group Project (Part II)**

**INTRODUCTION**

In this project II, we compare different variables with response(chd) in order to figure out whether there is a difference between group who has heart attack in the past five years and the group who doesn’t, with respect to the specified variables, which has potential factors to lead to the response.

The set of data (6 files) we use on this project are obtain from in clean version of the data in project Part I. the intention of this project is to identify risk factors for CHD.

In our data set, many subjects have multiple records. Multiple records from a subject are not considered independent because they are from the same subject. To satisfy the independency requirement, we use subtract the first observation of each subject after sorting the combined dataset by subject and date of visit. There are 4500 observations in total.

**METHOD SELECTION**

During the process, we compare different variables with response(chd) one by one for the intention of finding out the risk factors. To compare the discrete variables with response(chd), we compare frequency or distribution with plot. Basically we use Chi-Square to test whether there is a difference between different groups. For the continuous part, we compare the mean or median of the variables, and use ttest or Wilcoxon two-sample test to find out the difference between group with or without heart attack in the past five years.

**RESULTS**

**Univariate Analysis**

A brief association between predictors and response:

Note: √ means predictors are different between group chd = 0 or chd = 1,

× means they are not difference.

The results are as follows:

Response : chd

Variables :

discreate

×smoke (Table1-2)

×drink (Table3-4)

×marital (Table5-6)

×education (Table7-8)

√glucose\_status (Table9-10)

√pregnancy (Table11-12)

×exercise (Table13-14)

√site (Table15-16)

×gender (Table17-18)

continuous

√age (Table19-21,Graph1)

×height (Table22-24,Graph2)

×weight (Table25-27,Graph3)

√hdl (Table28-30,Graph4)

√ldl (Tablle31-33,Graph5)

√tg (Table34-36,Graph6)

√tc (Table37-39,Graph7)

√glucose (Table40-42,Graph8)

√dbp (Table43-45,Graph9)

√sbp (Table46-48,Graph10)

**Logistic Regression:**

We first roughly select two models for the analysis:

1. If we put all of the variables into selection, the results would be glucose, age, tg, and smoke.(Table49-54)
2. If we put the associated variables into selection, the results would be glucose, age, tg, and without smoke.(Table55-60)

However, as we have already confirmed that there is no difference between group who has heart attack in the past five years and group who doesn’t, specified by smoking status. It might be that smoke affects the chd together with other factors, but itself alone won’t affect chd. Then we choose the second model as our final model, interpretations are:

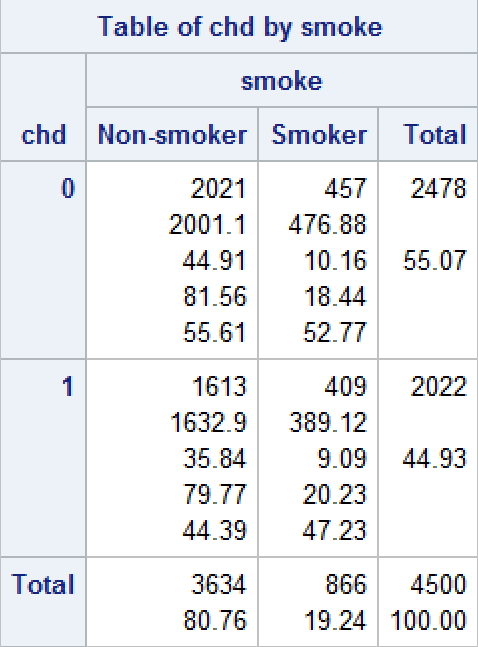
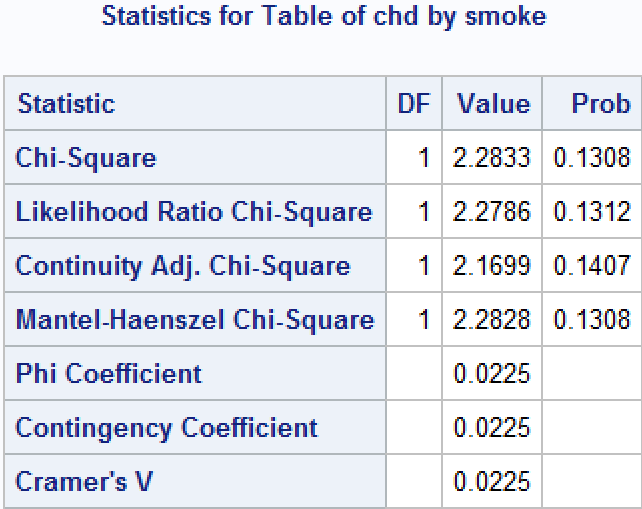
1. High glucose are more likely to have chd than low glucose due to high glucose risk.
2. Elder observations are more likely to have chd than youngers.
3. High triglyceride observations are more likely to have chd than lower observations.

**Conclusion**

the variables site, glucose\_status, glucose, pregnancy, age, hdl, ldl, tg, tc, dbp, sbp are different between two groups of response(chd).

Best model: chd = glucose age tg.

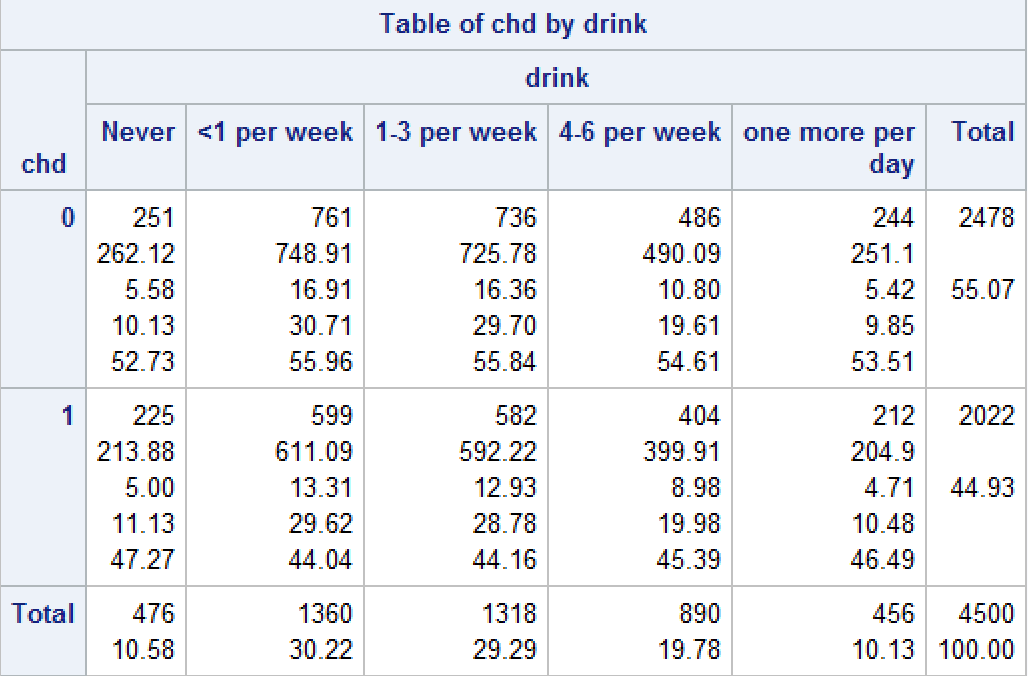
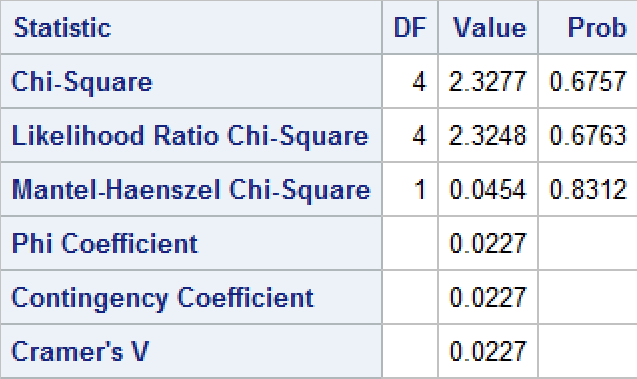
Tables 1-2

The frequencies are all greater than 5, then we use the Chi-Square test.

Chi-Square Prob. is 0.1308, then fail to reject hypotheses, there is no difference between people who has a chd or not in the past few years, with respect to smoking status.

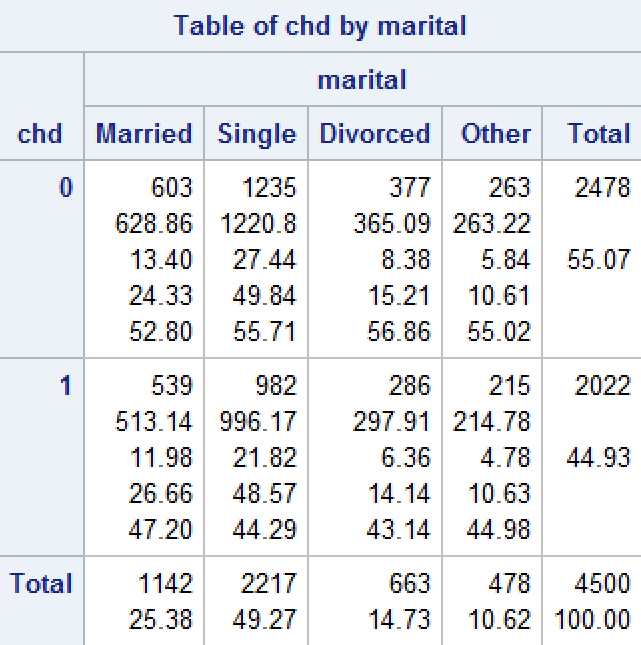
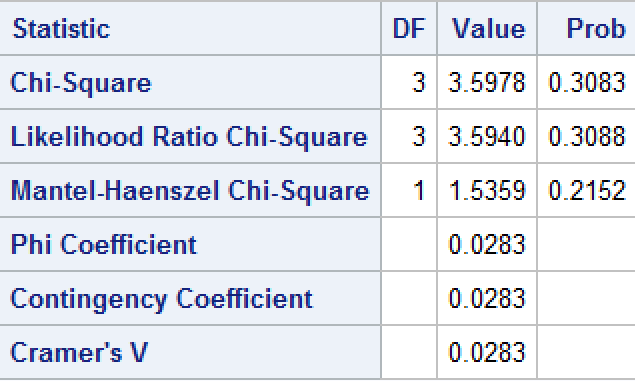
Tables 3-4

The frequencies are all greater than 5, then we use the Chi-Square test.

The Chi-Square Prob. is 0.6757, then fail to reject hypotheses, there is no difference between people who has a chd or not in the past few years, with respect to drinking status

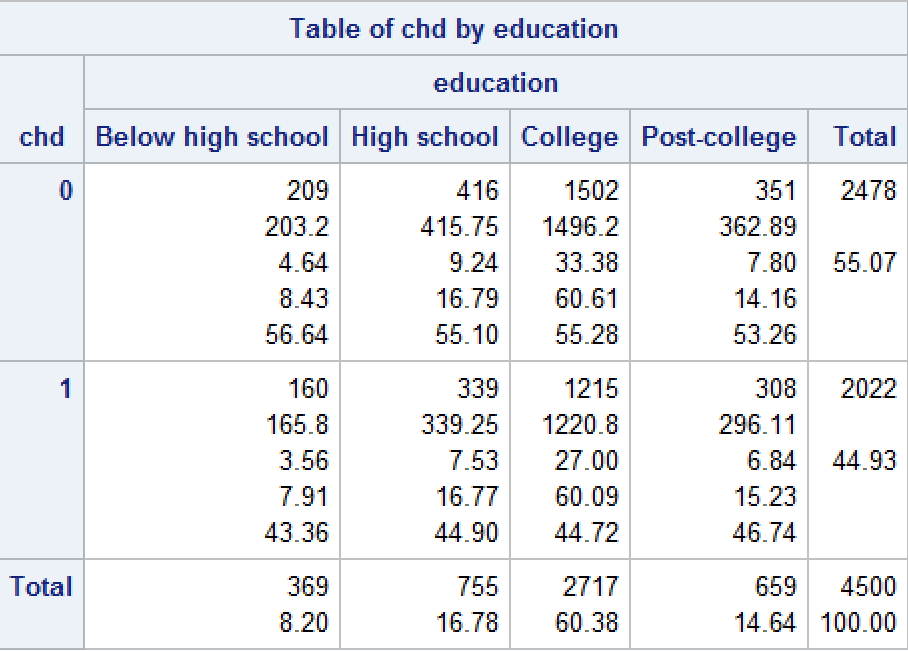
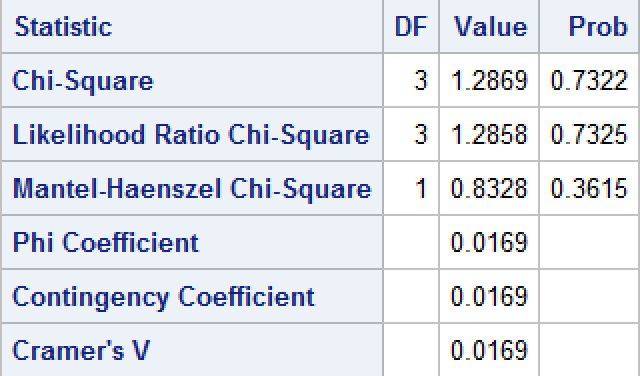
Table 5-6

The frequencies are all greater than 5, then we use the Chi-Square test.

The Chi-Square Prob. is 0.3038, then fail to reject hypotheses, there is no difference between people who has a chd or not in the past few years, with respect to marital status(1 = ‘Single’, 2 = ‘Married’, 3 = ‘Divorced’, 4 = ‘Other’).

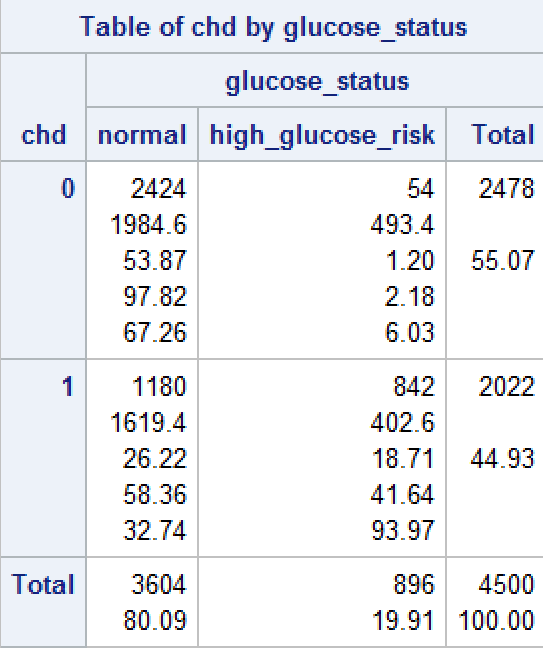
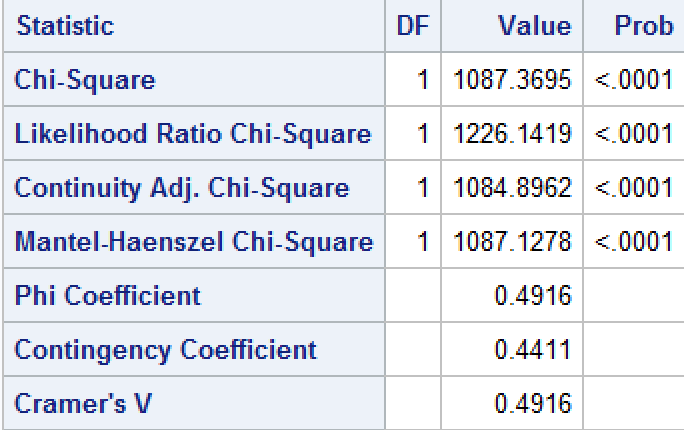
Tables 7-8

The frequencies are all greater than 5, then we use the Chi-Square test.

The Chi-Square Prob. is 0.7322, then fail to reject hypotheses, there is no difference between people who has a chd or not in the past few years, with respect to education status(1 = ‘’Below high school’, 2 = ‘High school graduate’, 3 = ‘College graduate’, 4 = ‘Post-college’).

Tables 9-10

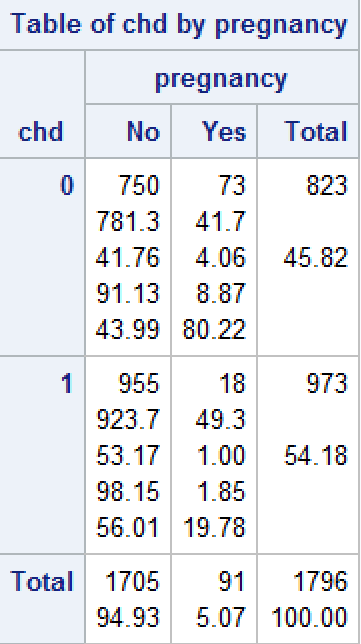
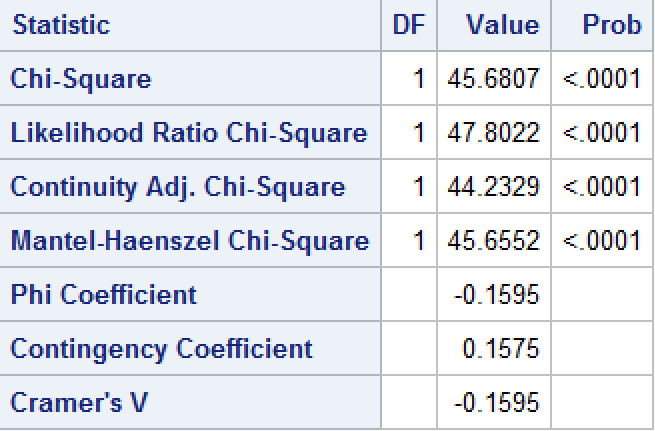
 

If we specify the observations whose glucose is greater or equal to 120 as high risk of glucose group and the others as the standard glucose group, then we try the frequency test.

The frequencies are all greater than 5, then we use the Chi-Square test.

The Chi-Square Prob. <.0001, then we reject null, meaning that there is a difference between people who has a chd or not in the past few years, with respect to glucose group, which means the chd has something to do with high risk of glucose or not.

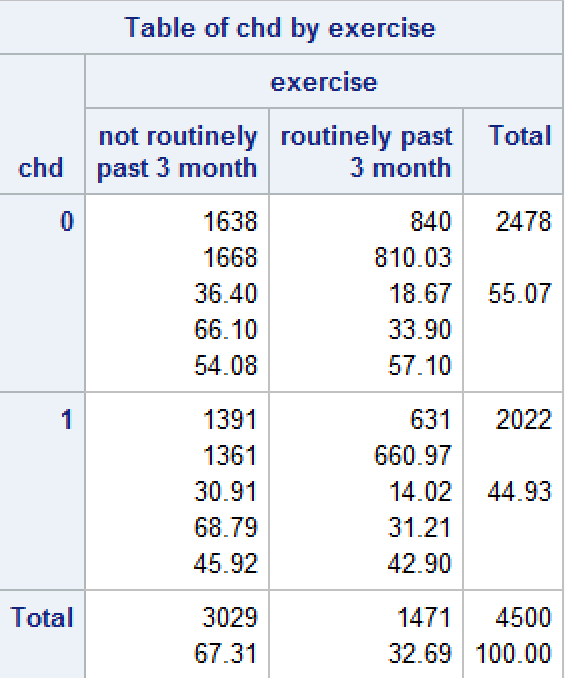
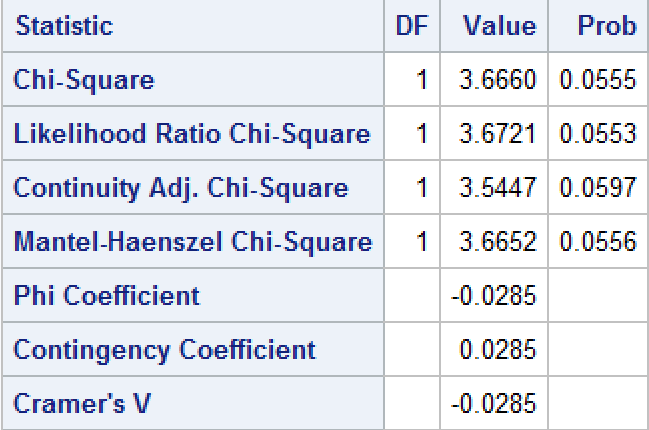
Tables 11-12

First, we delete the subject = 9, which means ‘not answer’, the frequencies are all greater than 5, then we use the Chi-Square test.

The Chi-Square Prob. <.0001, then we reject null, meaning that there is a difference between people who has a chd or not in the past few years, with respect to the situation of pregnancy or not, which means the chd has something to do with pregnancy or not.

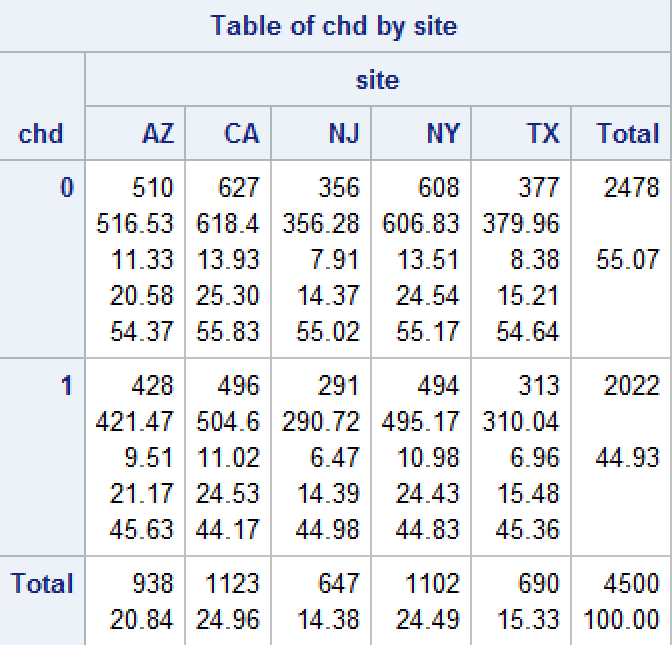
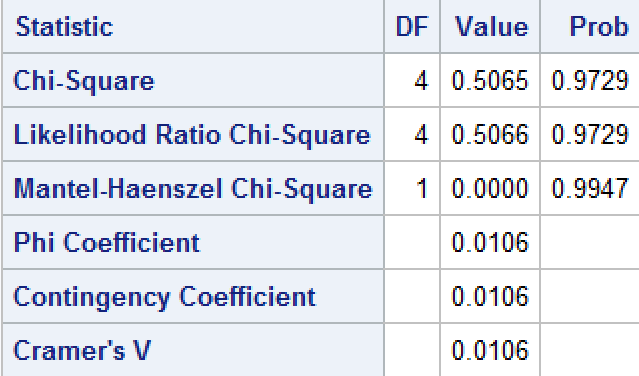
Tables 13-14

The frequencies are all greater than 5, then we use the Chi-Square test.

The Chi-Square Prob. is 0.0555, then fail to reject hypotheses, there is no difference between people who has a chd or not in the past few years, with respect to exercise.

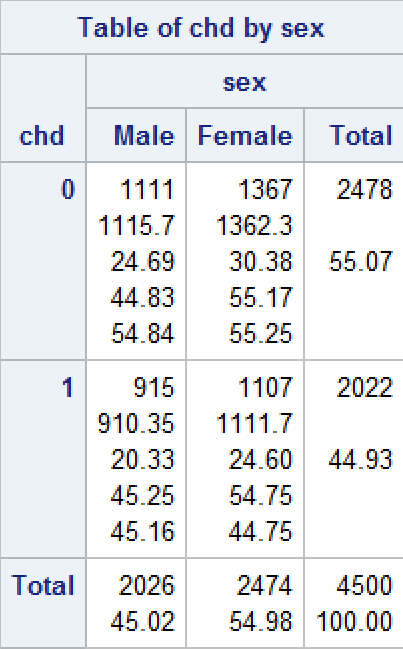
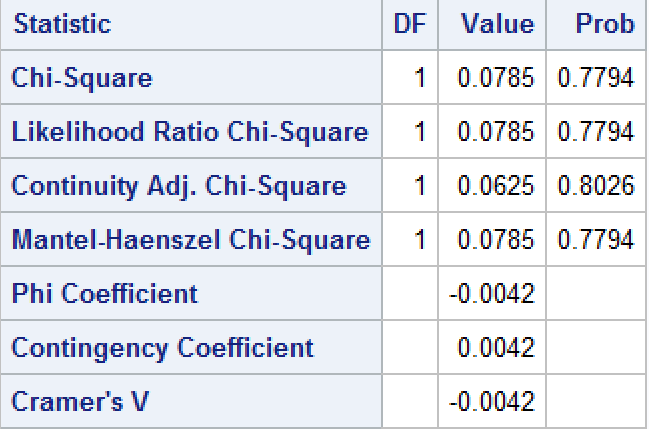
Tables 15-16

The frequencies are all greater than 5, then we use the Chi-Square test.

The Chi-Square Prob. is 0.9729, then fail to reject hypotheses, there is no difference between people who has a chd or not in the past few years, with respect to sites.

Tables 17-18

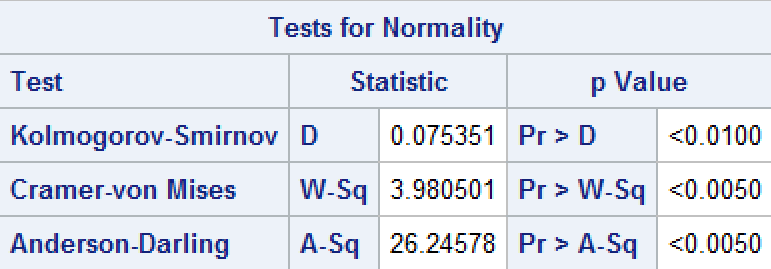
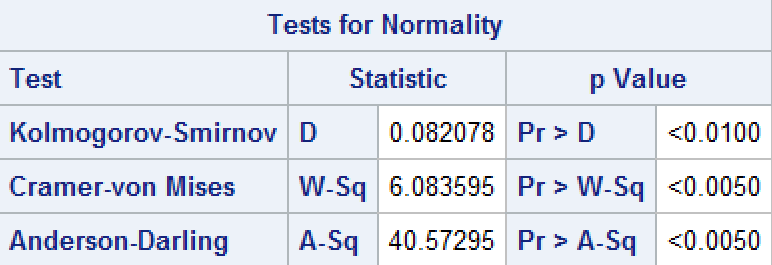
The frequencies are all greater than 5, then we use the Chi-Square test.

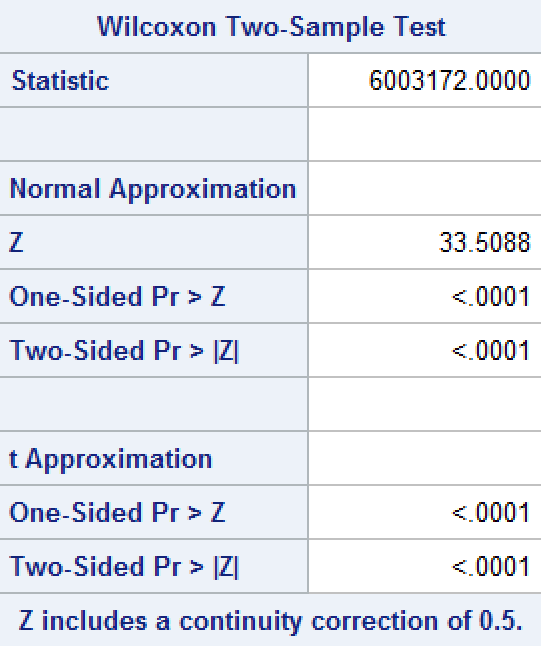
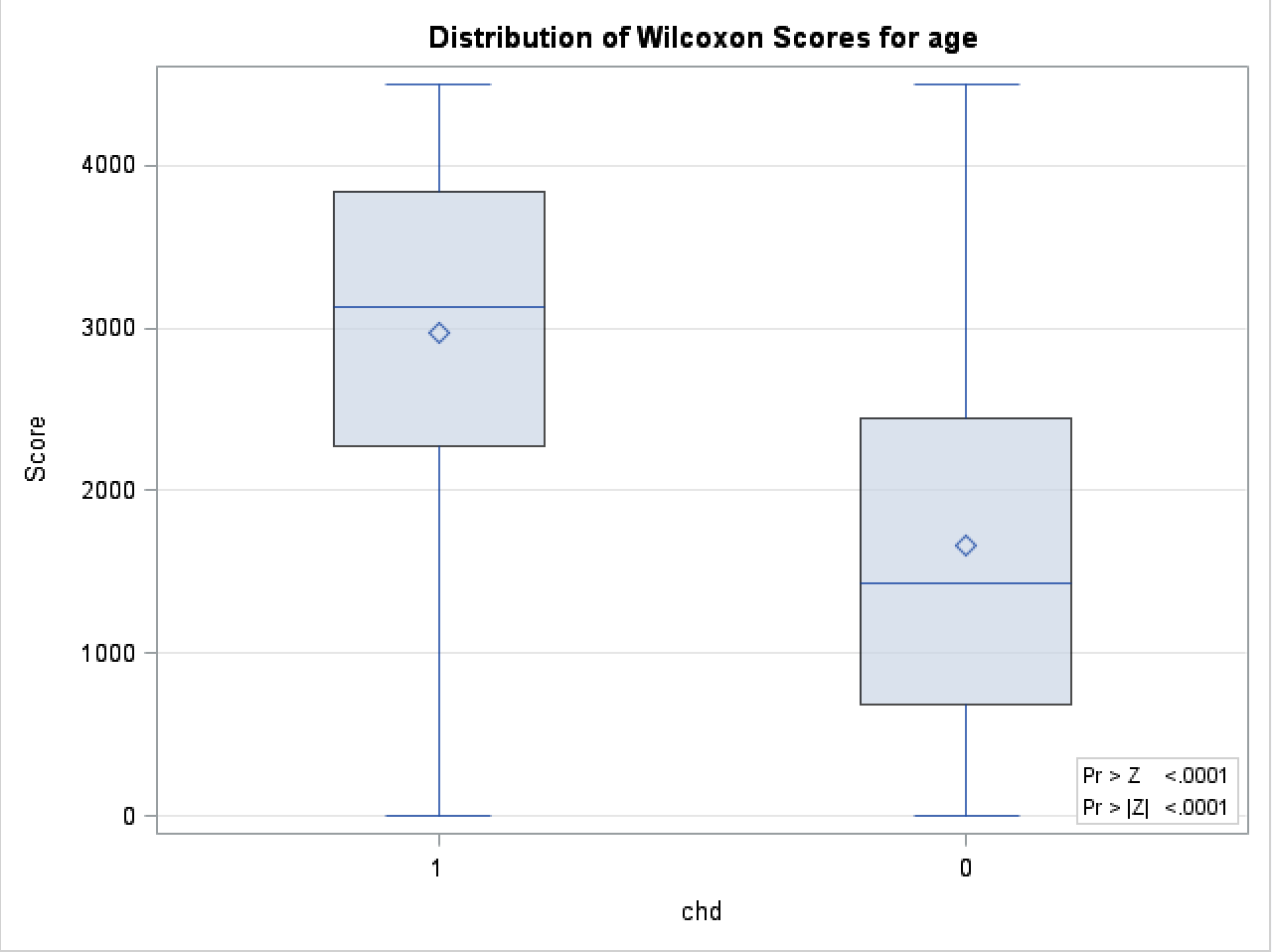
The Chi-Square Prob. Is 0.7794, fail to reject null, there is no difference between people who has a chd or not in the past few years, with respect to gender.

Tables 19-21, Graph 1

**Since population sample are larger than 2000, we use Anderson-Darling test for testing normality**

chd = 0 chd = 1



Pr > A-Sq < 0.0050 for both two groups , reject null, age is not normally distributed, with respect to chd.

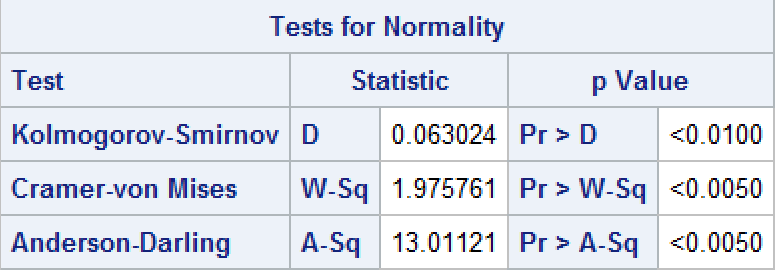
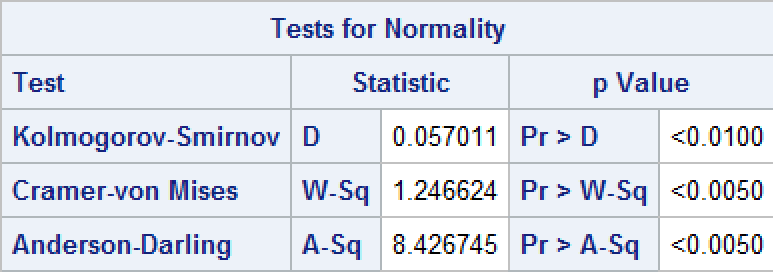
Then we use Wilcoxon Two-Sample Test:

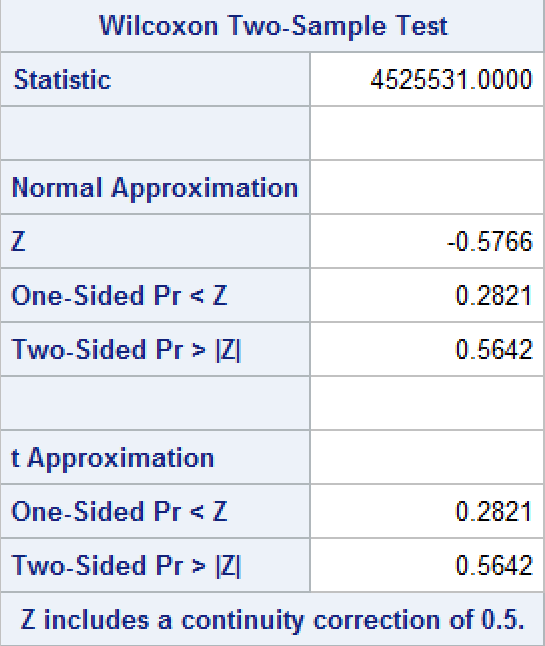
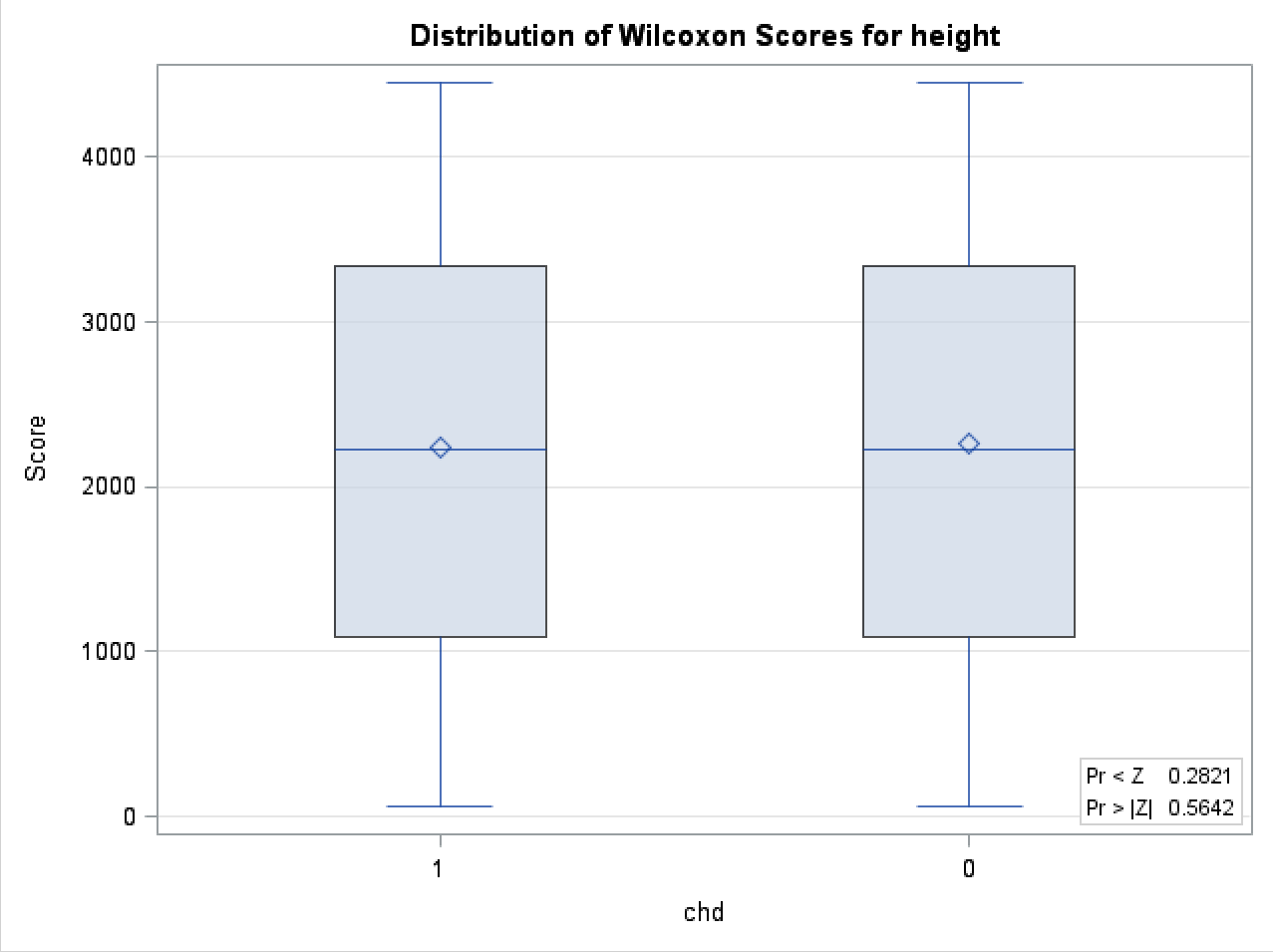
Two-Sided Pr > |Z| < .0001, then we reject null, two groups doesn’t have same median, meaning there is a difference of age, with respect to chd, which means that chd has association with age.

We can also see that through the graph, which two groups of response are significantly difference in age.

Table 22-24, Graph 2

chd = 0 chd = 1

Pr > A-Sq < 0.0050 for both two groups , reject null, height is not normally distributed, with respect to chd.

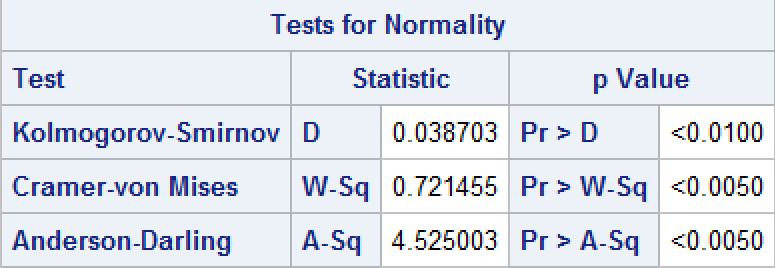
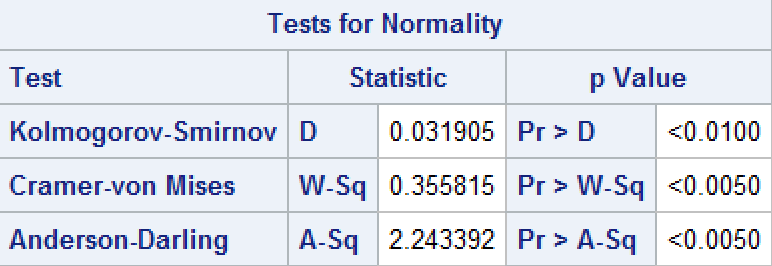
Then we use Wilcoxon Two-Sample Test:

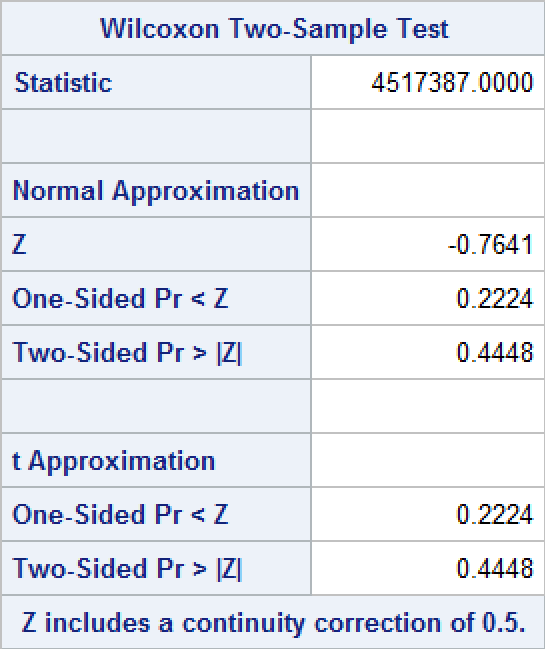
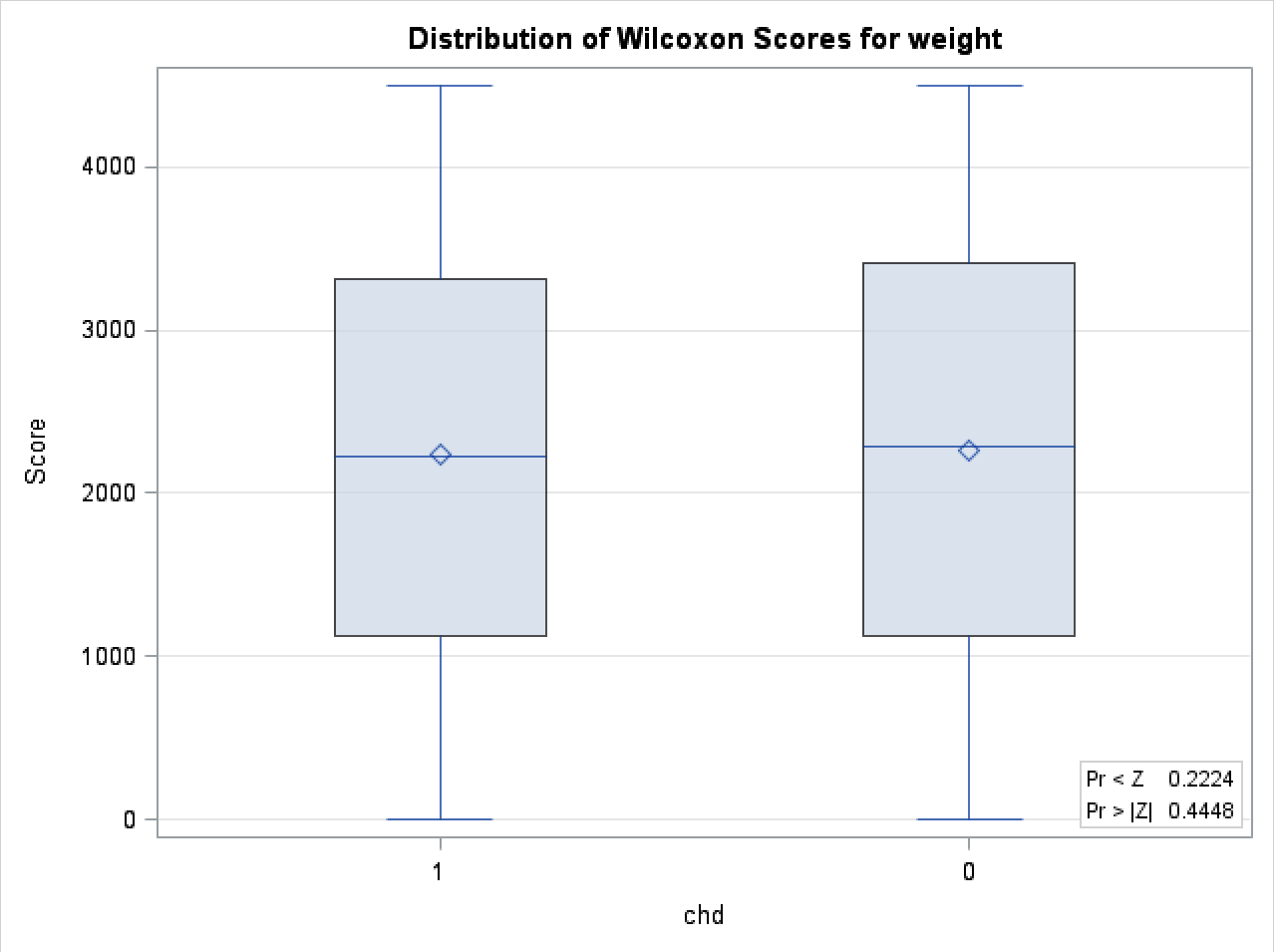
Two-Sided Pr > |Z| = 0.5642, then we fail to reject null, two groups have same median, meaning there is no difference of height, with respect to chd, which means that chd has no association with age.

We can also see that through the graph, which the two groups have no difference of height.

Table 25-27,Graph 3

chd = 0 chd = 1

Pr > A-Sq < 0.0050 for both two groups , reject null, weight is not normally distributed, with respect to chd.

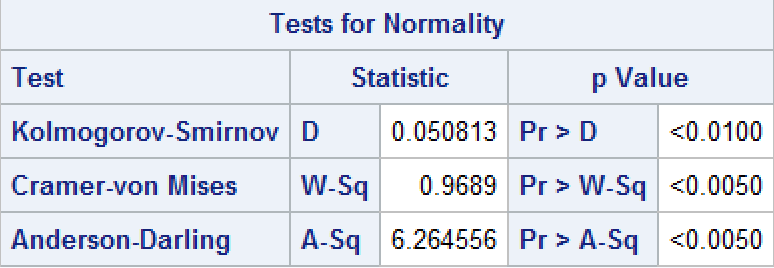
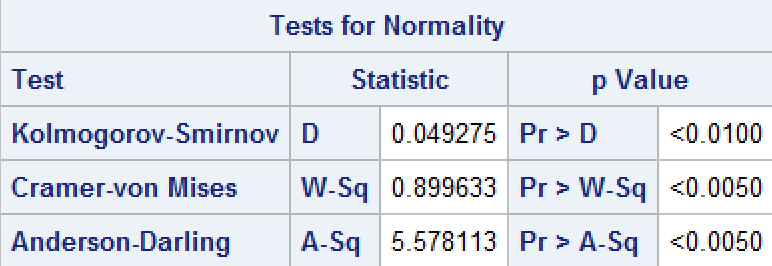
Then we use Wilcoxon Two-Sample Test:

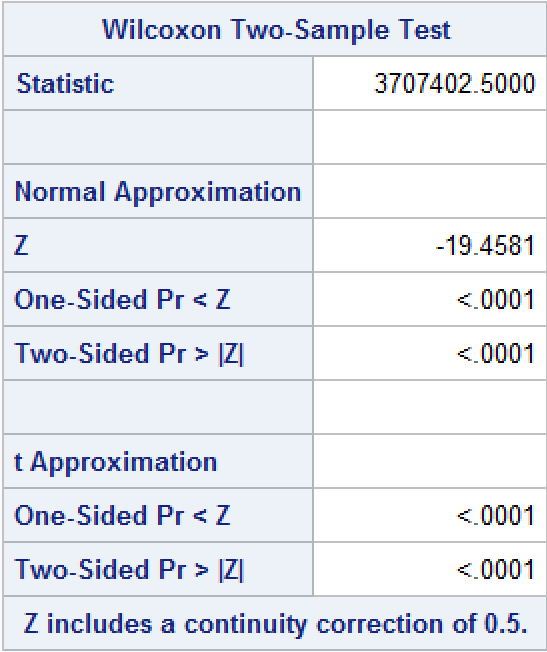
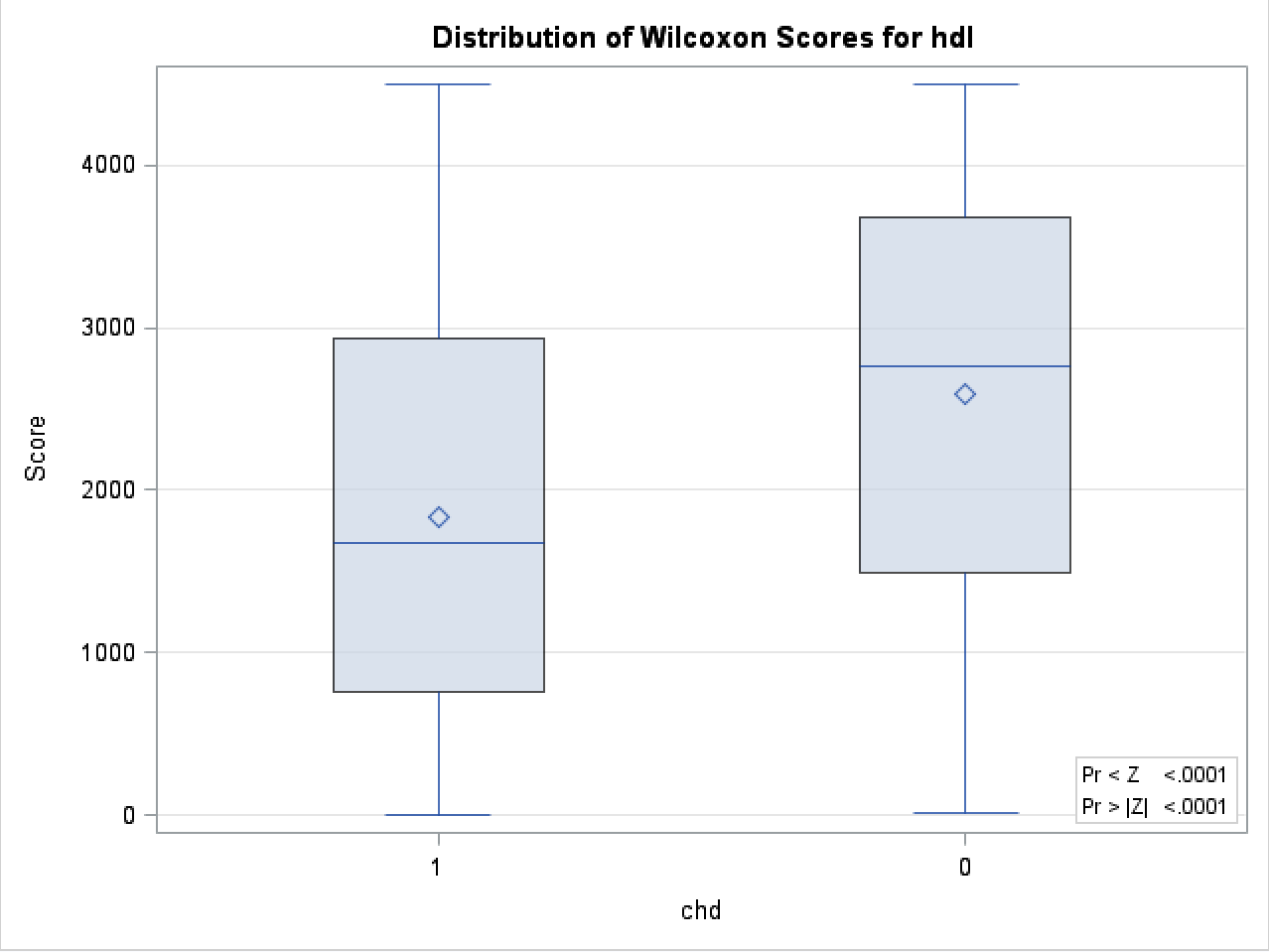
Two-Sided Pr > |Z| < .5642, then we fail to reject null, two groups have same median, meaning there is no difference of weight, with respect to chd, which means that chd has no association with weight.

We can also see that through the graph, which there is no difference of weight between two group in response.

Table28-30, Graph4

chd = 0 chd = 1

Pr > A-Sq < 0.0050 for both two groups , reject null, weight is not normally distributed, with respect to chd.

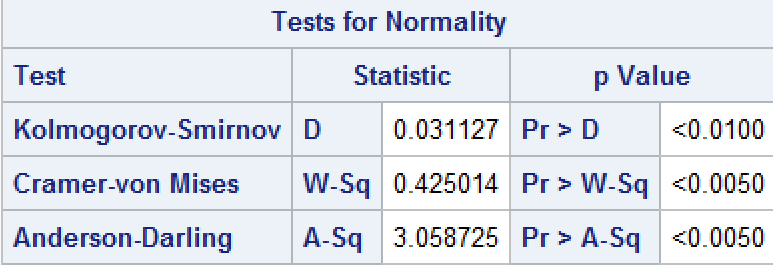
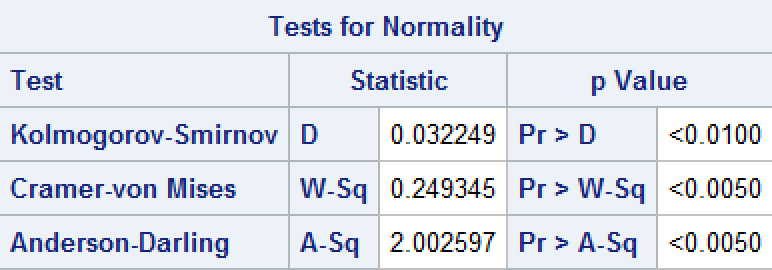
Then we use Wilcoxon Two-Sample Test:

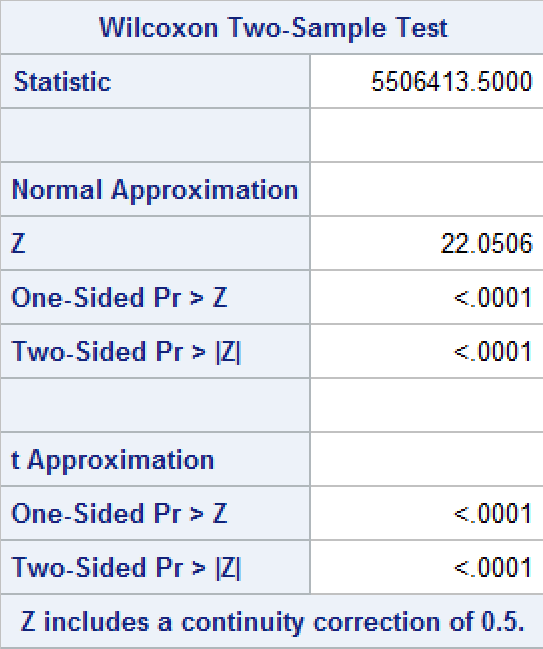
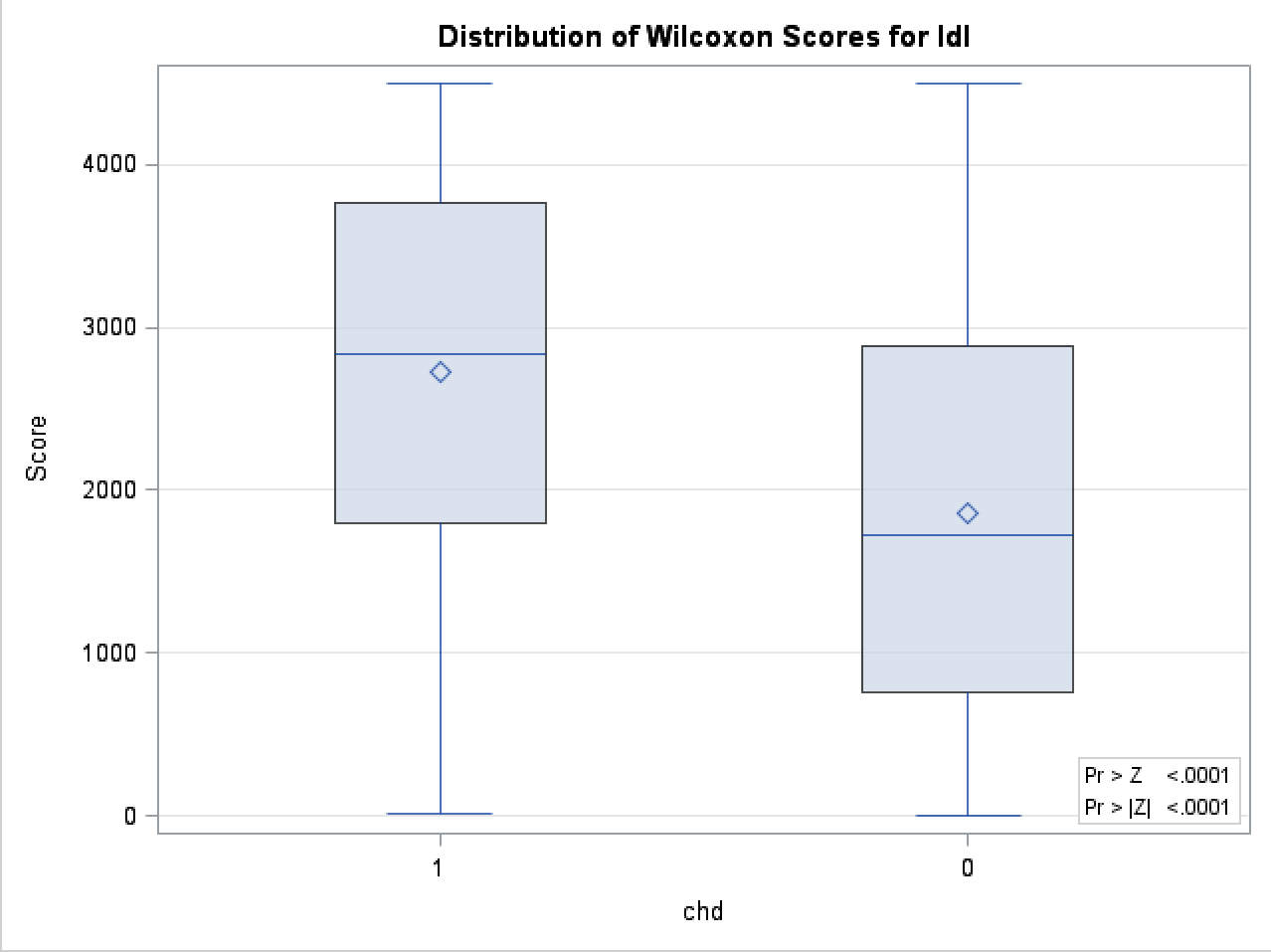
Two-Sided Pr > |Z| < .0001, then we reject null, two groups doesn’t have same median, meaning there is a difference of hdl, with respect to chd, which means that chd has association with high density cholesterol.

We can also see that through the graph, which two groups of response are significantly difference in hdl.

Table31-33, Graph5

chd = 0 chd = 1

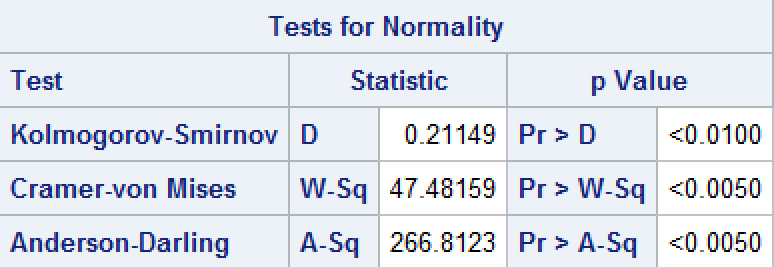
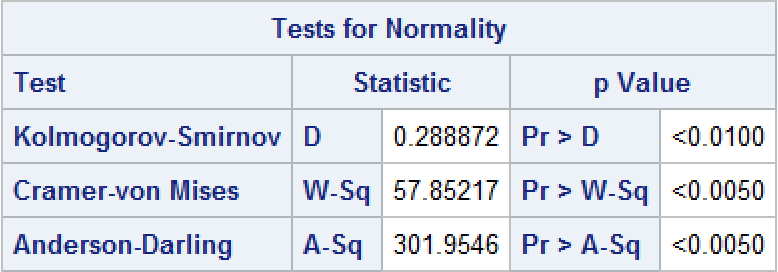
Pr > A-Sq < 0.0050 for both two groups , reject null, ldl is not normally distributed, with respect to chd.

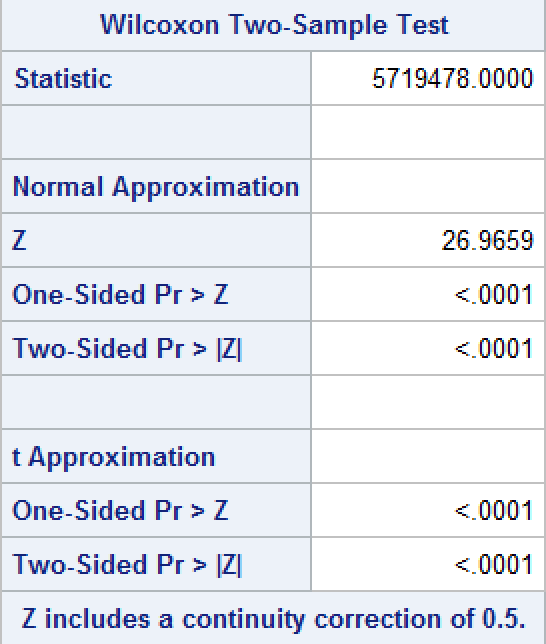
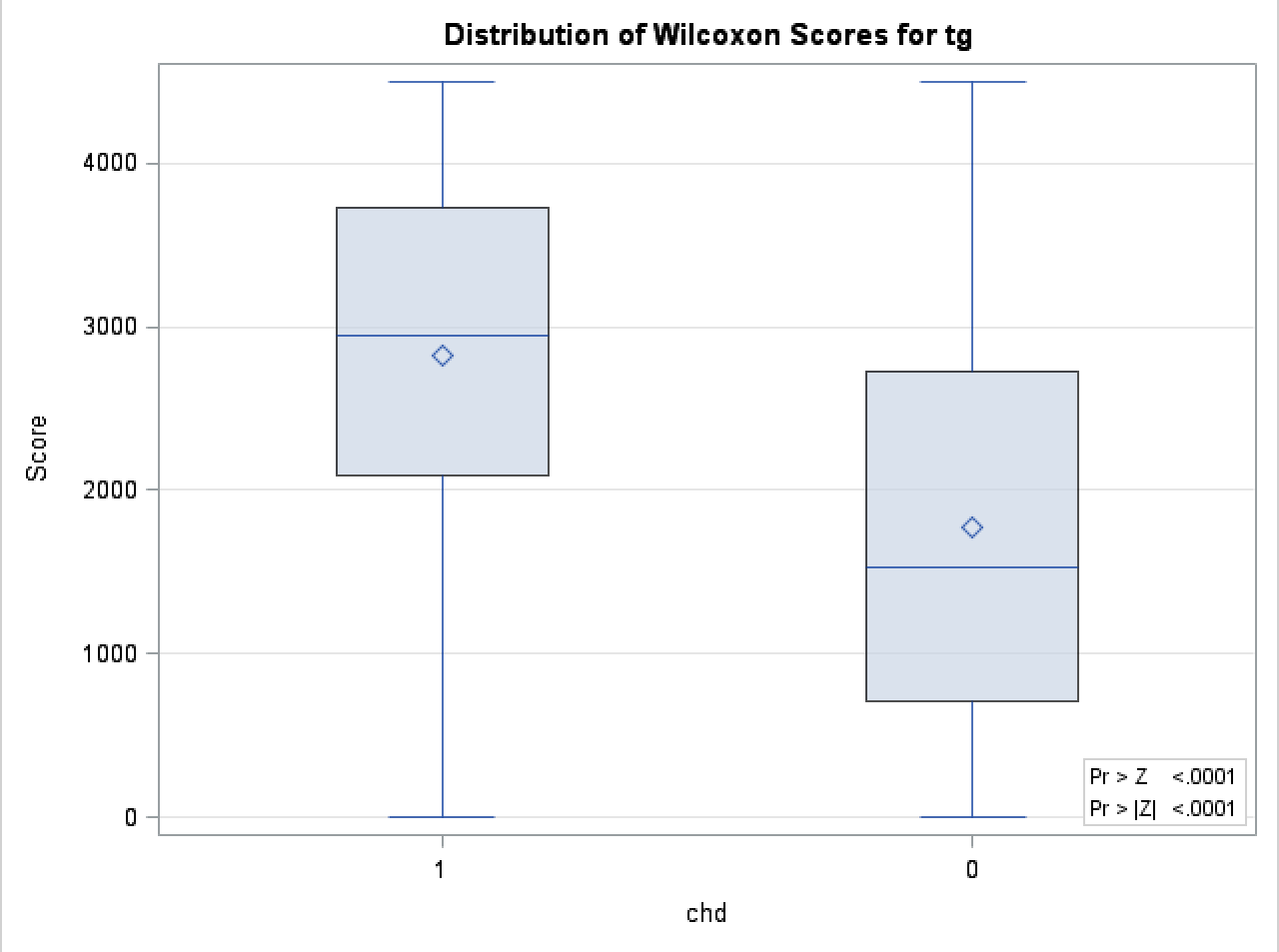
Then we use Wilcoxon Two-Sample Test:

Two-Sided Pr > |Z| < .0001, then we reject null, two groups doesn’t have same median, meaning there is a difference of ldl, with respect to chd, which means that chd has association with low density cholesterol.

We can also see that through the graph, which two groups in response are significantly difference in ldl.

Table34-36, Graph6

Pr > A-Sq < 0.0050 for both two groups , reject null, tg is not normally distributed, with respect to chd.

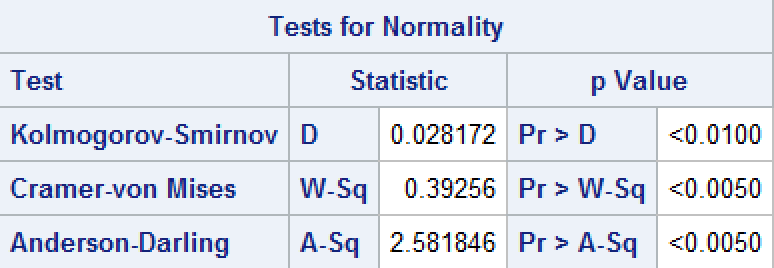
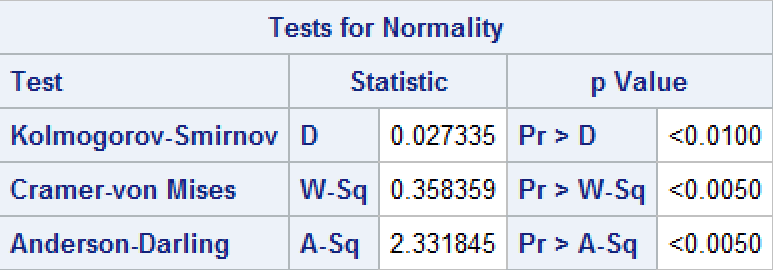
Then we use Wilcoxon Two-Sample Test:

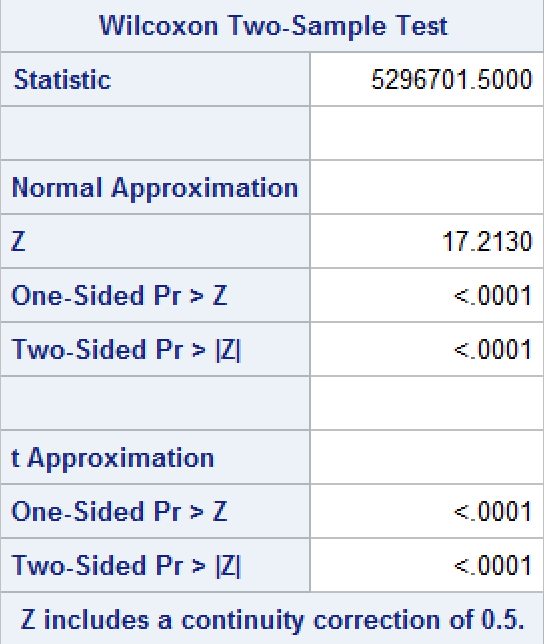
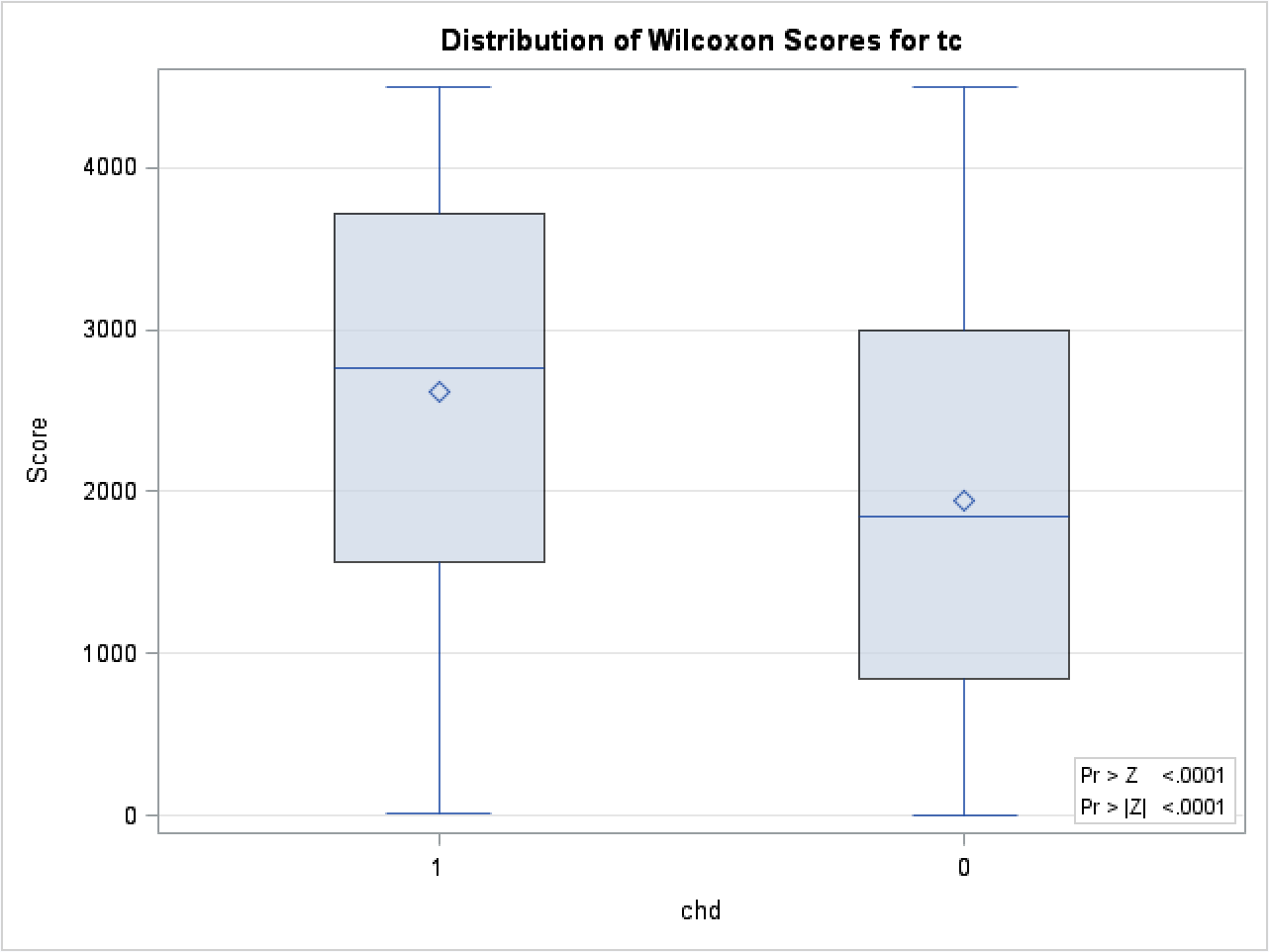
Two-Sided Pr > |Z| < .0001, then we reject null, two groups doesn’t have same median, meaning there is a difference of tg, with respect to chd, which means that chd has association with triglyceride.

We can also see that through the graph, which two groups in response are significantly difference in tg.

Table37-39 Graph7

chd = 0 chd = 1

Pr > A-Sq < 0.0050 for both two groups , reject null, tc is not normally distributed, with respect to chd.

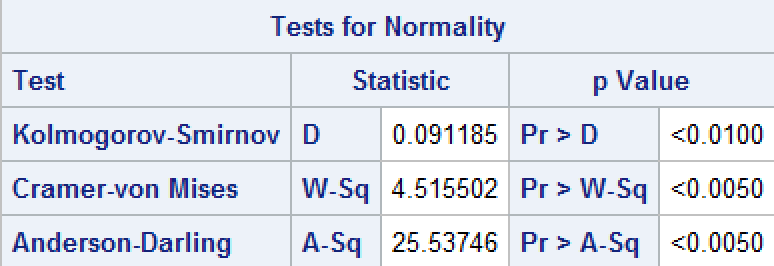
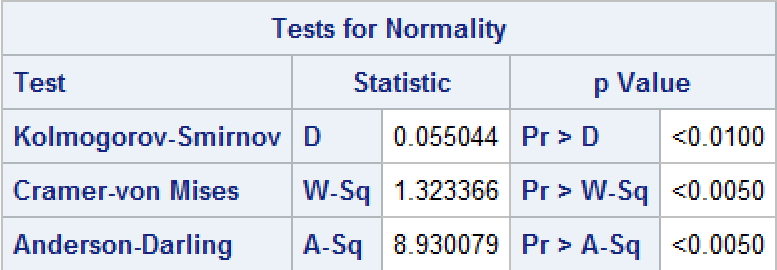
Then we use Wilcoxon Two-Sample Test:

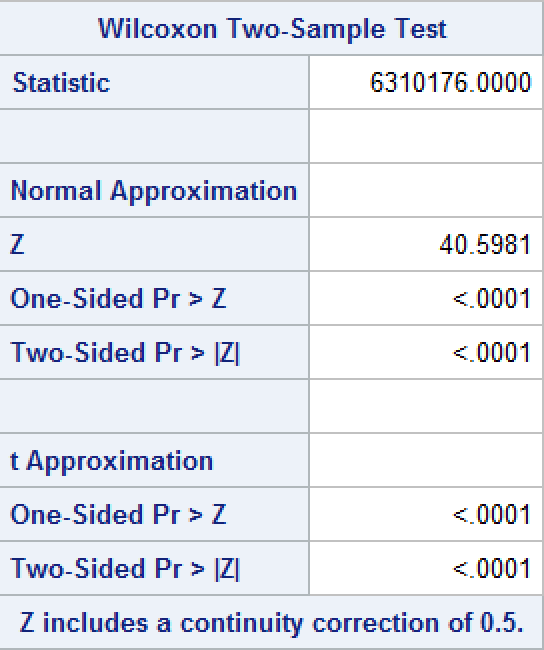
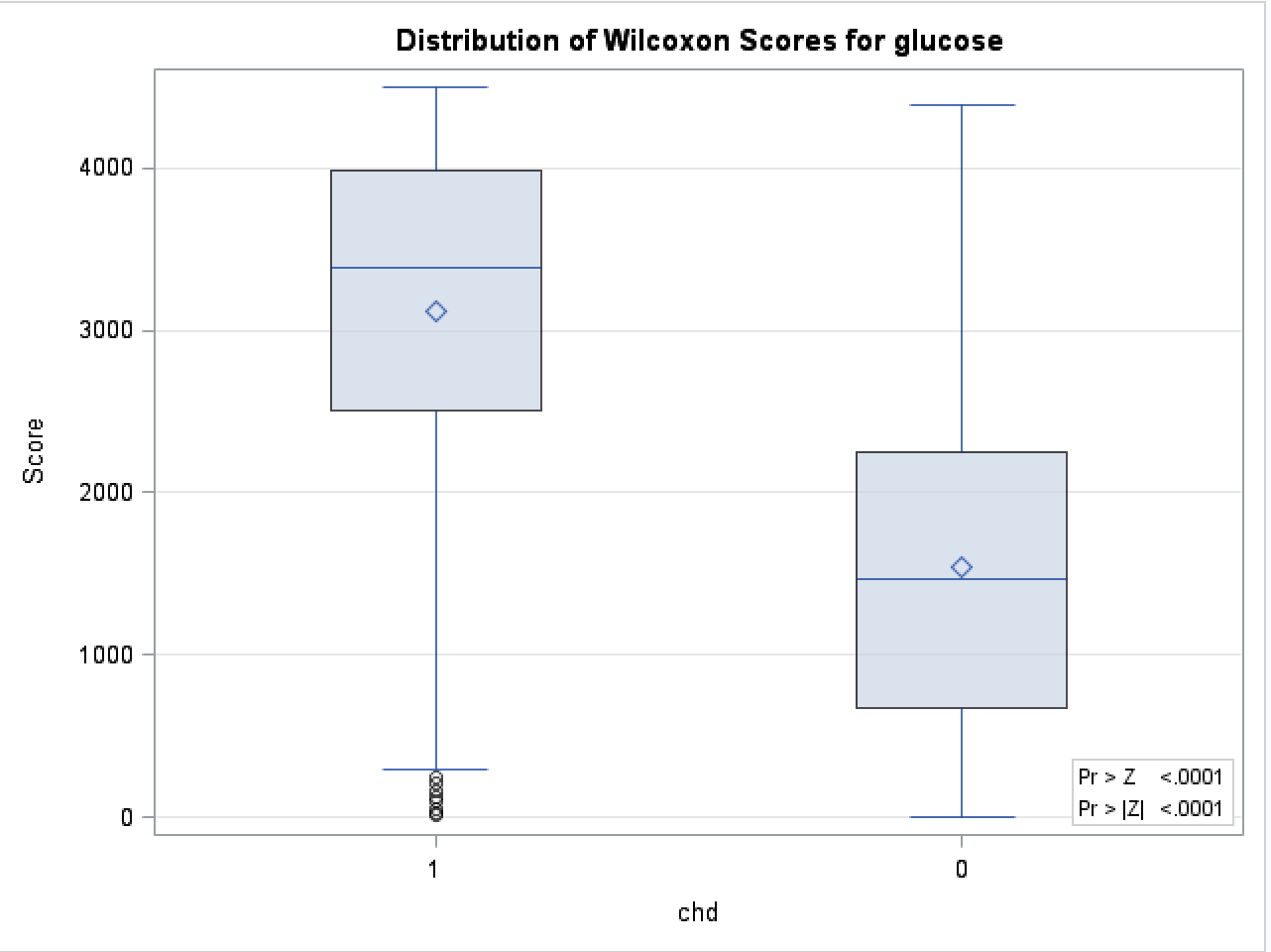
Two-Sided Pr > |Z| < .0001, then we reject null, two groups doesn’t have same median, meaning there is a difference of tc, with respect to chd, which means that chd has association with total cholesterol.

We can also see that in the graph, which tc is significantly different for the two groups in response.

Table40-42, Graph8

chd = 0 chd = 1

Pr > A-Sq < 0.0050 for both two groups , reject null, glucose is not normally distributed, with respect to chd.

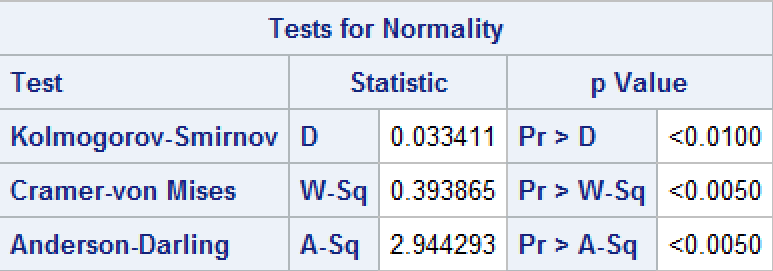
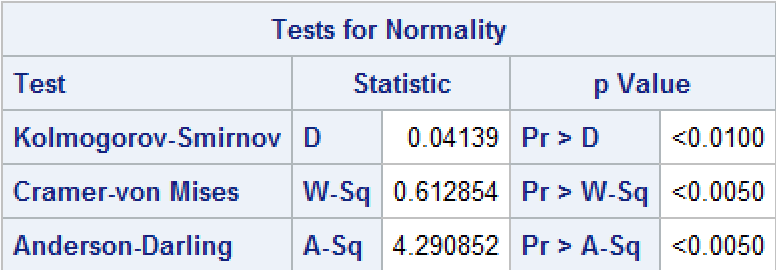
Then we use Wilcoxon Two-Sample Test:

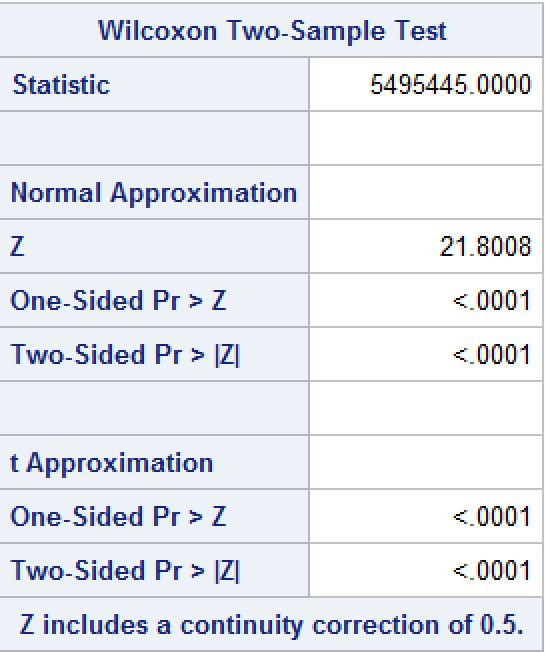
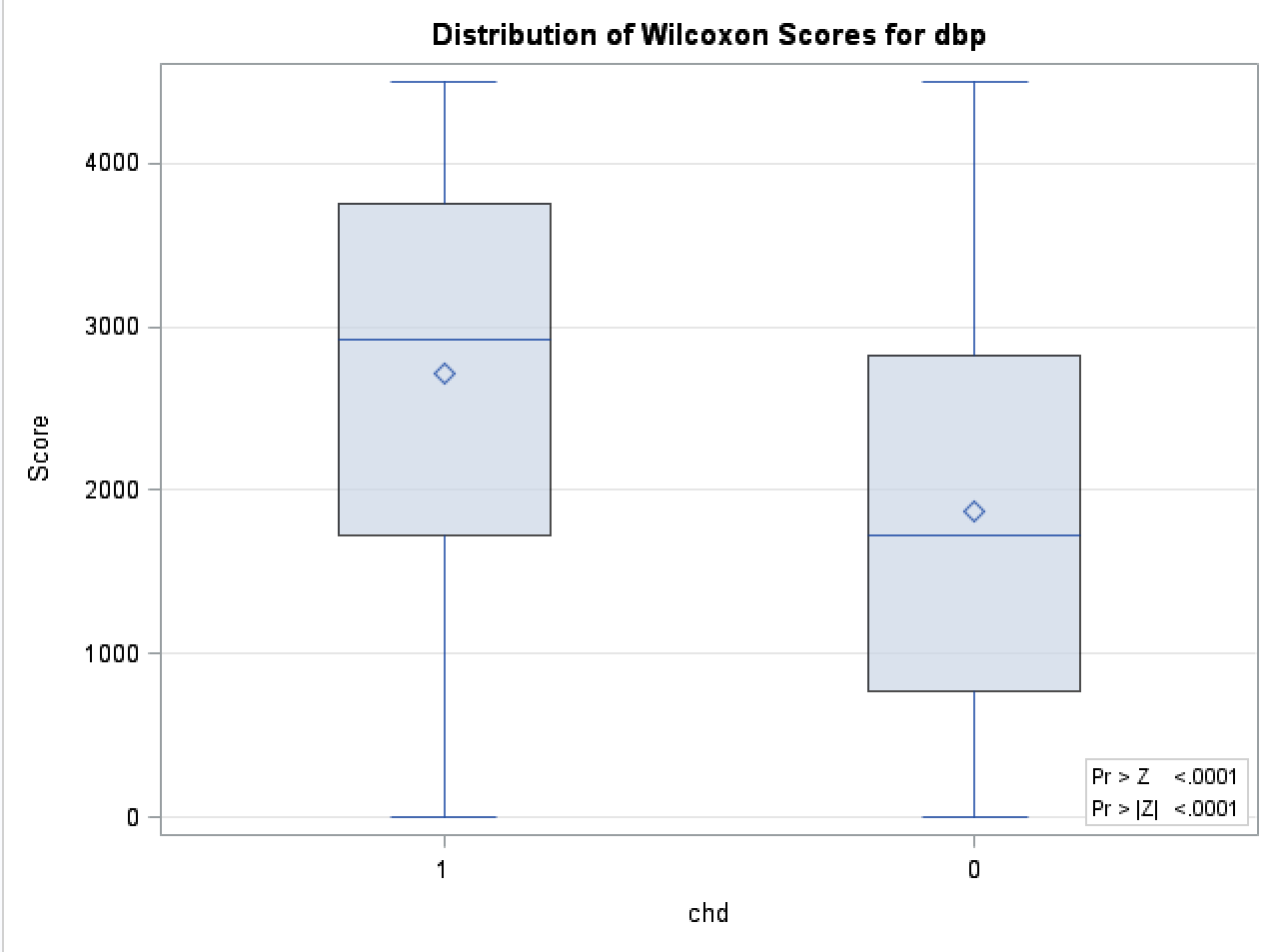
Two-Sided Pr > |Z| < .0001, then we reject null, two groups doesn’t have same median, meaning there is a difference of glucose, with respect to chd, which means that chd has association with glucose.

We can also see that through the graph, which two groups in response are significantly difference in glucose.

Table43-45, Graph9

chd = 0 chd = 1

Pr > A-Sq < 0.0050 for both two groups , reject null, dbp is not normally distributed, with respect to chd.

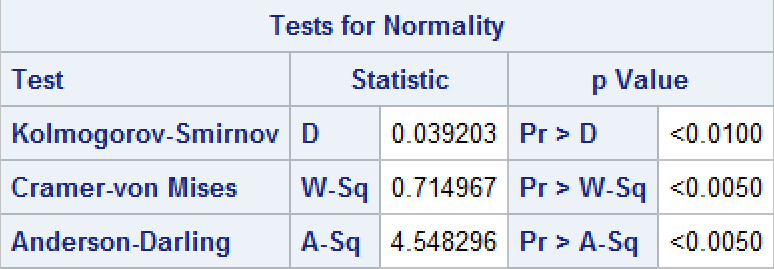
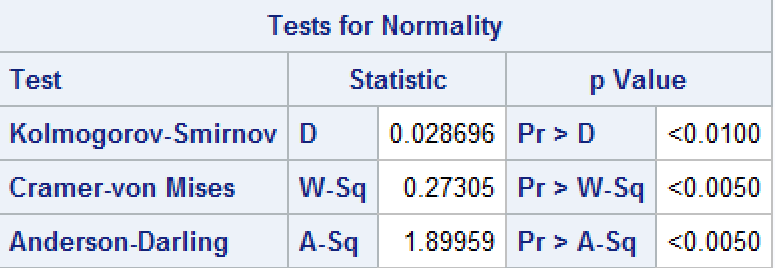
Then we use Wilcoxon Two-Sample Test:

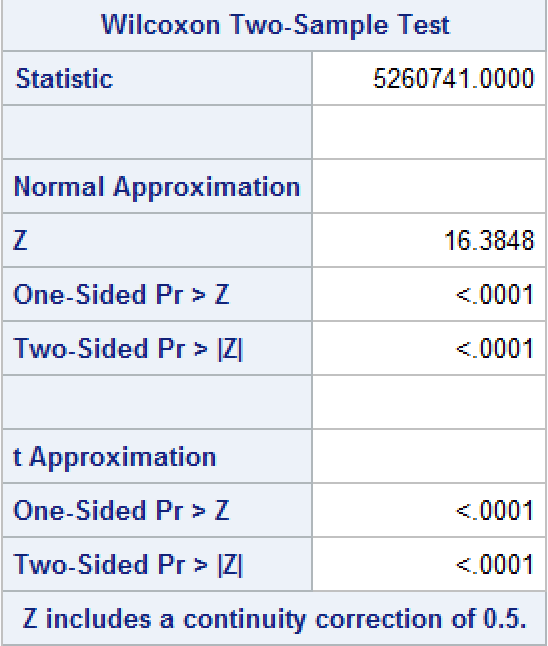
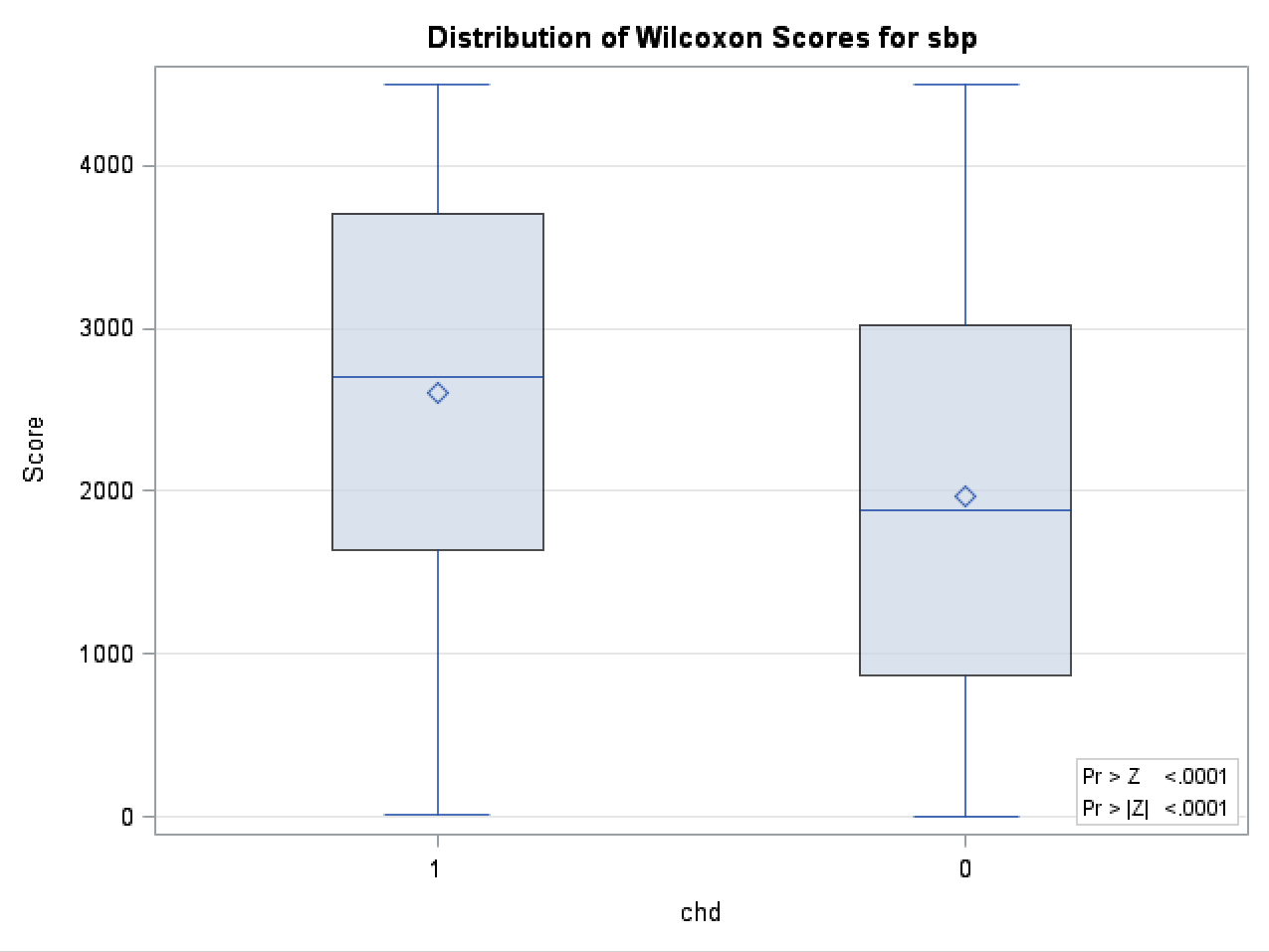
Two-Sided Pr > |Z| < .0001, then we reject null, two groups doesn’t have same median, meaning there is a difference of dbp, with respect to chd, which means that chd has association with diastolic blood pressure.

We can also see that through the graph, which two groups in response are significantly difference in dbp.

Table46-48, Graph10

chd = 0 chd = 1

Pr > A-Sq < 0.0050 for both two groups , reject null, sbp is not normally distributed, with respect to chd.

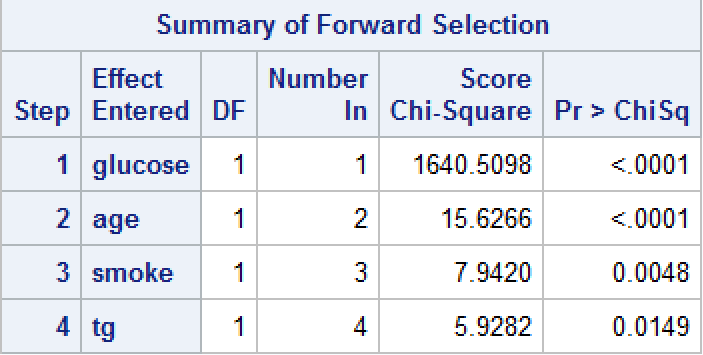
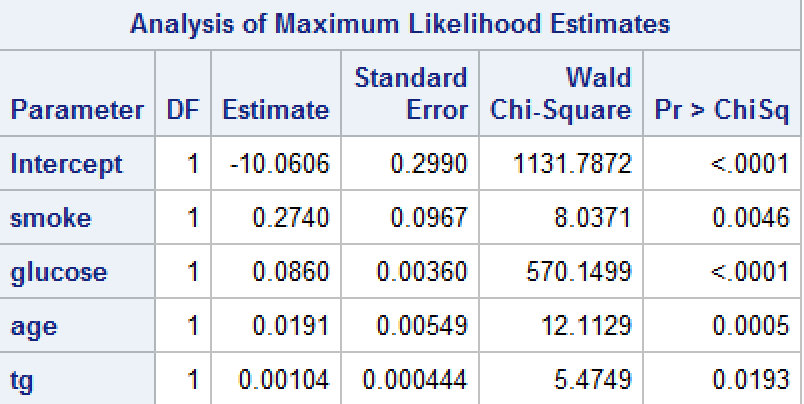
Then we use Wilcoxon Two-Sample Test:

Two-Sided Pr > |Z| < .0001, then we reject null, two groups doesn’t have same median, meaning there is a difference of sbp, with respect to chd, which means that chd has association with systolic blood pressure.

We can also see that through the graph, which two groups in response are significantly difference in sbp.

Forward selection

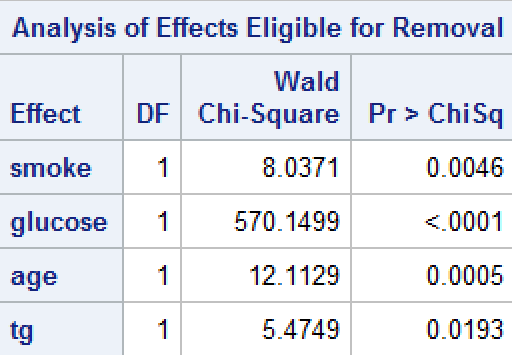
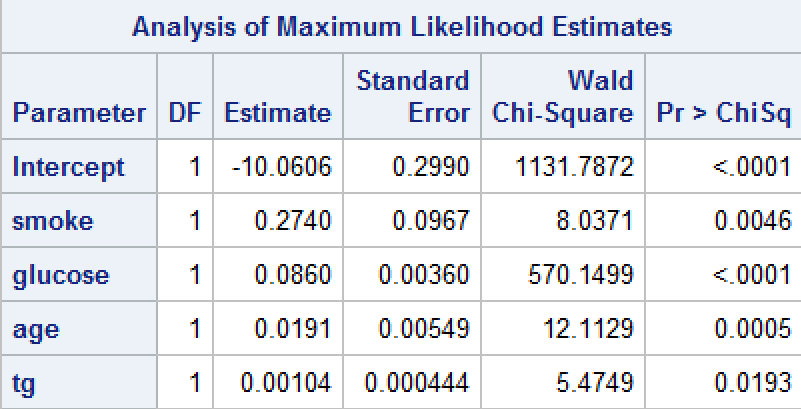
Table49-50

Variables fit for the model are: glucose age smoke tg

Backward selection

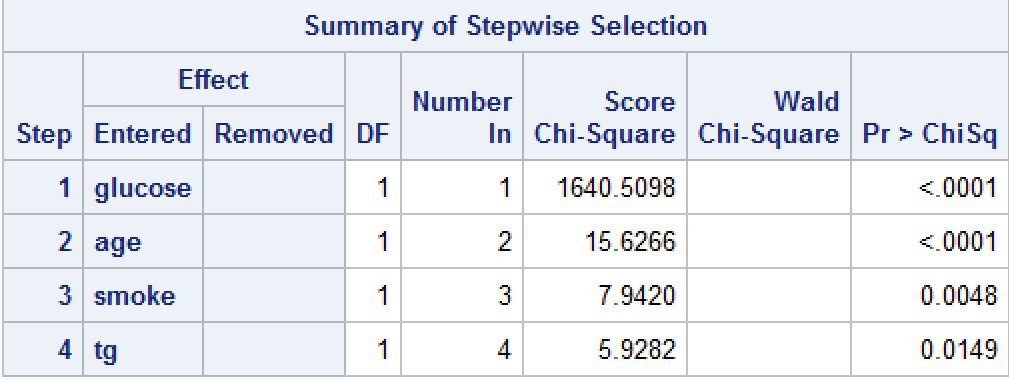
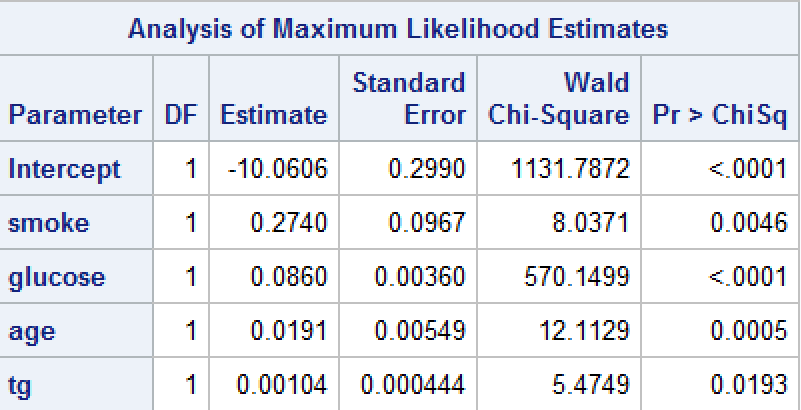
Table51-52

Variables fit for the model are: glucose age smoke tg

Stepwise selection

Table53-54

Variables fit for the model are: glucose age smoke tg

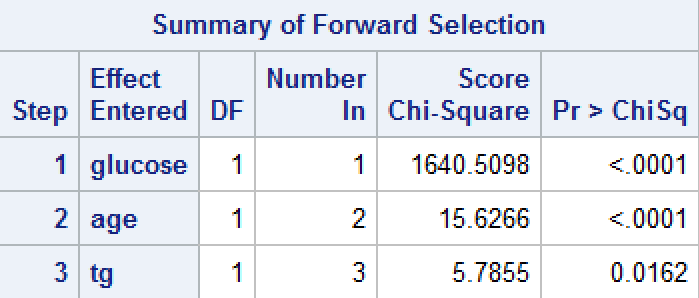
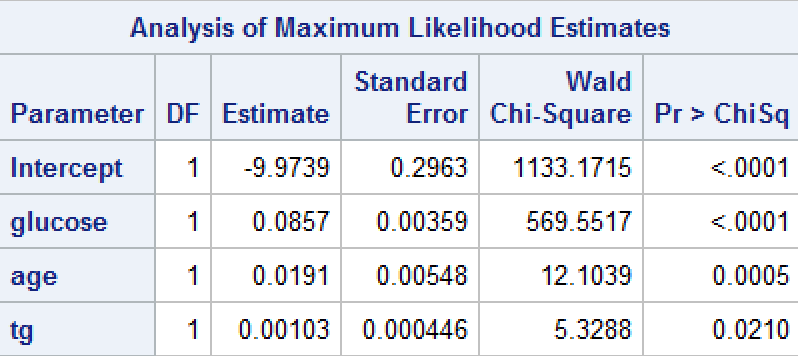
As we can see that all the variables selected are positive associated with chd, which means:

1. Smokers are more likely to have chd than non-smokers.
2. High glucose are more likely to have chd than low glucose due to high glucose risk.
3. Elder observations are more likely to have chd than youngers.
4. High triglyceride observations are more likely to have chd than lower observations.

Three selections using variables associated with chd:

Forward selection

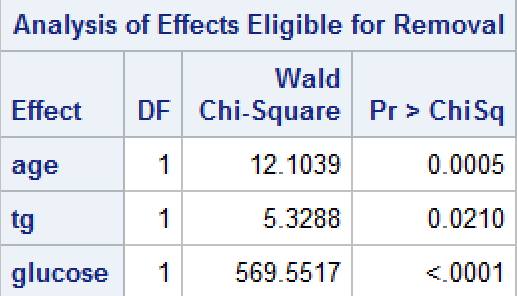
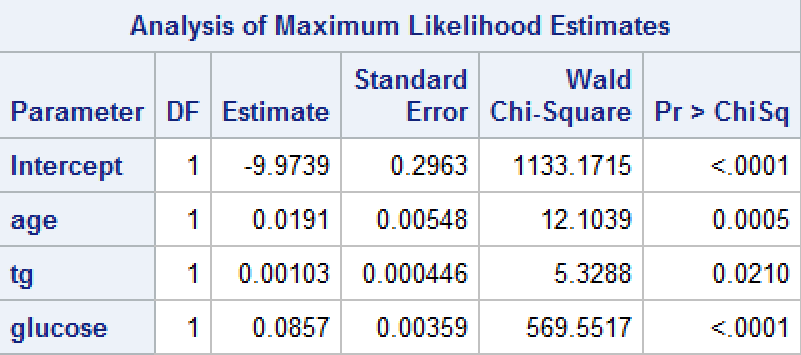
Table55-56

Variables fit for the model are: glucose age tg

Backward selection

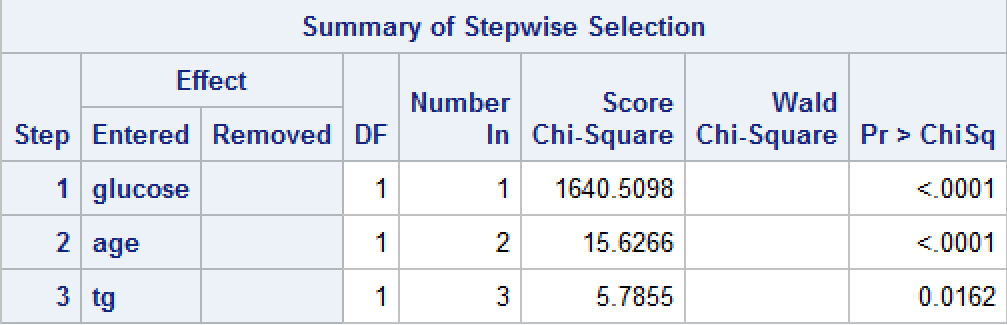
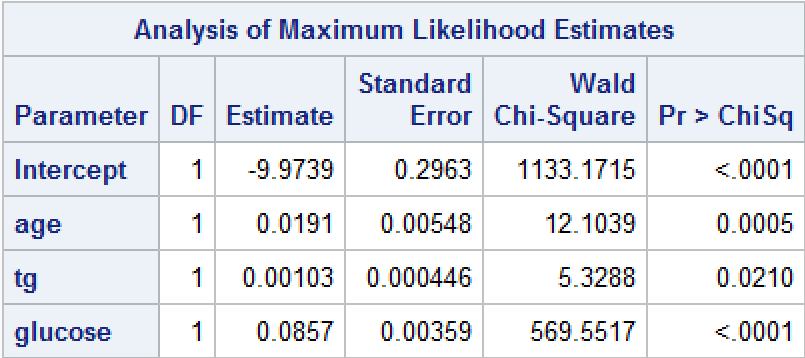
Table57-58

Variables fit for the model are: glucose age tg

Stepwise selection

Table59-60

Variables fit for the model are: glucose age tg

As we can see that all the variables selected are positive associated with chd, which means:

1. High glucose are more likely to have chd than low glucose due to high glucose risk.
2. Elder observations are more likely to have chd than youngers.
3. High triglyceride observations are more likely to have chd than lower observations.