data project;

set project.BRFSS;

keep AVEDRNK3 \_TOTINDA \_VEGESU1 CVDINFR4 \_AGE\_G SEXVAR \_RACE BPHIGH6 TOLDHI3 \_SMOKER3 \_BMI5CAT \_BMI5;

run;

data d\_freq;

set project;

where AVEDRNK3<77 and AVEDRNK3 ~=. and \_TOTINDA~=9 and \_VEGESU1<99998 and CVDINFR4~=7 or 9 and CVDINFR4~=. and \_RACE~=9 and BPHIGH6~=7 AND BPHIGH6~=9 AND BPHIGH6~=. and TOLDHI3~=7 OR 9 and TOLDHI3~=. and \_SMOKER3~=9 and \_BMI5CAT~=. and \_BMI5~=.;

run;

Chi Square

proc freq data=d\_freq;

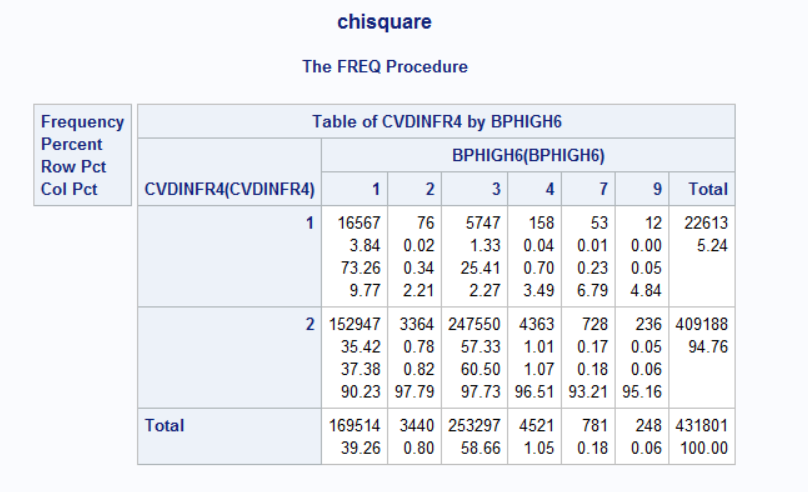
table CVDINFR4\*BPHIGH6/CHISQ measures;

where CVDINFR4 in (1,2) and BPHIGH6 ne .;

title 'chisquare';

footnote '1=Yes 2=Yes,during pregnancy(F) 3=No 4=borderline 7=Dontknow 9= Refused';

run;





Logistic Regression:’

data logistic4;

set natality\_data (keep=RF\_PHYPE CIG\_0 MAGER);

if CIG\_0 = 0 then SMOKE = 0; \*SMOKE='N';

else if 0 < CIG\_0 < 98 then SMOKE = 1; \*SMOKE='Y';

else SMOKE=.; \*SMOKE = 'NA';

run;

proc logistic data=logistic4;

title 'Pre\_Pregnancy Hypertension as a Function of Smoking

Status';

model RF\_PHYPE(event='Y')= SMOKE MAGER / details lackfit;

where RF\_PHYPE~='U' and SMOKE~=.;

run;

data logisticRecoded;

do MAGER=15 to 55 by 0.1;

CurveSmoke0=exp(-6.3821 + 0.5144\*0 + 0.0791\*MAGER)/(1 + exp(-

6.3821 + 0.5144\*0 + 0.0791\*MAGER));

CurveSmoke1=exp(-6.3821 + 0.5144\*1 + 0.0791\*MAGER)/(1 + exp(-

6.3821 + 0.5144\*0 + 0.0791\*MAGER));

output;

end;

run;

proc sgplot data=logisticRecoded;

title 'Estimated Probability of Pre-Pregnancy Hypertension for

Smokers vs Non-Smokers for Different Ages';

series x=MAGER y=CurveSmoke0;

series x=MAGER y=CurveSmoke1;

run;

**IMPORT DATA**

/\*import dataset\*/

libname project "\\apporto.com\dfs\CLT\Users\swestfa2\_clt\Downloads";

proc import datafile="\\apporto.com\dfs\CLT\Users\swestfa2\_clt\Downloads\BRFSS.xlsx"

out=BRFSS DBMS= xlsx replace;

getnames=yes;

run;

**1. Handling missing data for each analysis**

data project;

set BRFSS;

keep AVEDRNK3 \_TOTINDA \_VEGESU1 CVDINFR4 \_AGE\_G SEXVAR \_RACE BPHIGH6 TOLDHI3 \_SMOKER3 \_BMI5;

where AVEDRNK3<77 and AVEDRNK~=. and \_TOTINDA~=9 and \_VEGESU1<99998 and CVDINFR<7 and CVDINFR4~=. and \_RACE~=9 and \_RACE~=. and BPHIGH<7 AND BPHIGH6~=. and TOLDHI2<7 and TOLDHI3~=. and \_SMOKER3~=9 and \_BMI5~=.;

run;

* **Interpretation:** Like the natality dataset we have worked with this semester, each variable in the BRFSS codebook contains its own values for missing, refused, etc. In this step, we removed the individual “null” values for each variable used throughout the analysis.

**2. Descriptive statistics:**

* **One continuous variable. Includes histogram and boxplot, numerical measures of location and variation**

**Code:**

/\*Descriptive statistics for one continuous variable\*/

data alcohol;

set project;

where AVEDRNK3~=77 OR 99; /\*77= dont know, 99= refused\*/

Run;

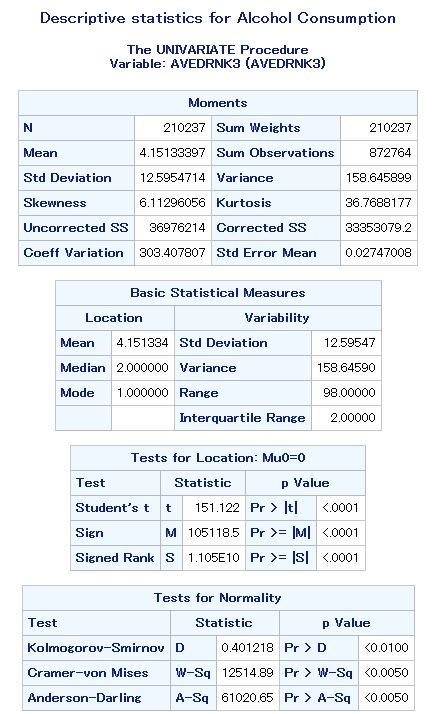
proc univariate data=alcohol normal plot; /\*we want a histogram\*/

title 'Descriptive statistics for Alcohol Consumption';

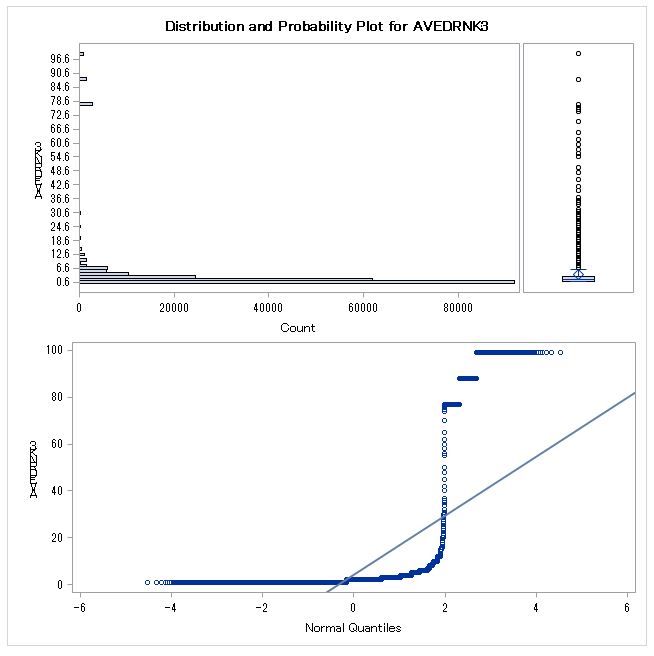
var AVEDRNK3;

Run;

**Results:**

****

****

****

**Interpretation:**

The average reporter has 4.15 drinks on a night where they are drinking. However, there is large variability/skewness in these data. This is expected, since drinking habits are highly personal. The distribution graphs demonstrate these results.

* **Descriptive statistics for one categorical variable. Includes pie graph/bar chart, frequencies and percentages.**

**Code:**

/\*Descriptive statistics for one categorical variable\*/

data physicalactivity;

set project;

keep \_TOTINDA;

where \_TOTINDA~=9;

run;

proc FREQ data=physicalactivity;

title 'Desciptive Statistics for Physical Activity';

tables \_TOTINDA;

run;

proc gchart data=physicalactivity;

title 'Physcial Activity';

vbar \_TOTINDA / midpoints= 1 to 2 by 1;

run;

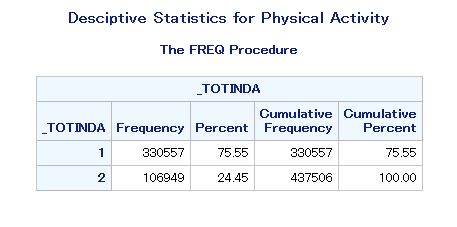
proc gchart data=physicalactivity;

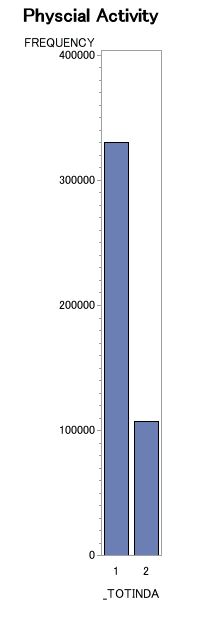
pie \_TOTINDA / discrete

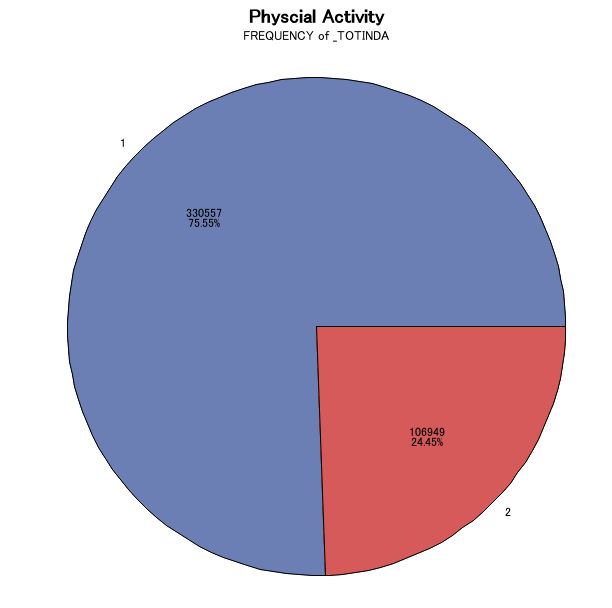
value=inside percent=inside slice=outside;

Run;

**Results:**

****

****

****

**Interpretation**:

TOTINDA=1 means they reported physical activity within the past 30 days, TOTINDA=2 means no physical activity. 75.55% of people (330,557) reported yes to physical activity, and 24.45% (106, 409) reported no.

* **Descriptive statistics for one continuous variable grouped by a categorical variable**

**Code:**

/\*Descriptive Statistics for continuous variable grouped by a categorical variable\*/

data BMIbySmoking;

set project;

keep \_BMI5 \_SMOKER3;

where \_SMOKER3 ~=9;

run;

proc means data=BMIbySmoking;

title 'Descriptive Statistics for BMI by Smoking Status';

class \_SMOKER3;

var \_BMI5;

run;

proc gchart data=BMIbySmoking;

title 'BMI by Smoking Status';

vbar \_SMOKER3 / midpoints= 1 to 4 by 1 type=mean;

Run;

How to relabel the pie chart

data age\_edu;

set natality\_data;

keep MEDUC MAGER;

if MEDUC=>1 and MEDUC=<8;

run;

proc means data=age\_edu;

title 'Average Age During Pregnancy with Some College Edu';

class MEDUC;

var MAGER;

run;

data age\_edu;

set natality\_data;

keep MEDUC MAGER;

if MEDUC=>4 and MEDUC=<8;

run;

proc means data=age\_edu;

title 'Average Age During Pregnancy with Some College Edu';

class MEDUC;

var MAGER;

run;

proc gchart data=age\_edu\_relabel;

title 'Average Age of Pregnant Mothers with Differnt Edu Levels';

pie EDU\_LEVELS / type=percent discrete plabel=(h=1.5 color=depk);

where EDU\_LEVELS~='NOT APPLICABLE';

data age\_edu\_relabel;

set age\_edu;

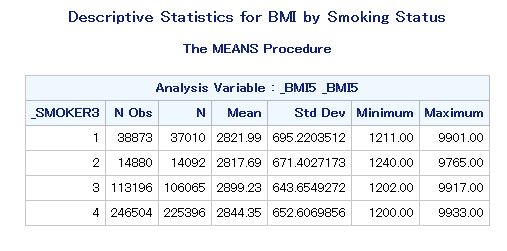
**Results:**

\_SMOKER3=1: Current everyday smoker

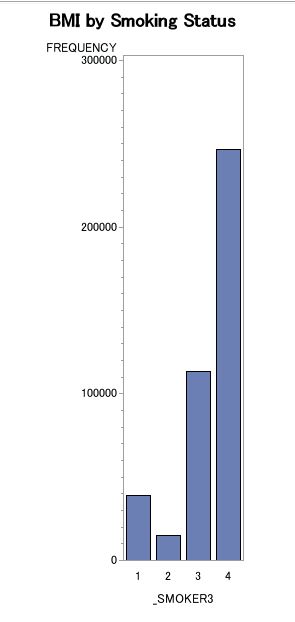
\_SMOKER3=2: Current smoker sometimes

\_SMOKER3=3: Former smoker

\_SMOKER3=4: Never smoked



Bar chart:



\*\*I wanted the y-axis to be mean BMI but I can’t figure out how to change it

**Interpretation**:

The mean BMI does not vary heavily depending on whether a person smokes. Current smokers have slightly lower BMIs, which is expected since nicotine suppresses the appetite.

**3. Test of Normality- includes skewness, histogram w normal curve overlay, normal qq plot, significance test for normality**

**Code:**

/\*Test of normality\*/

data drinknormality;

set project;

where AVEDRNK3~=9 and AVEDRNK3 ne .;

run;

/\*Histogram with normal curve overlay\*/

ods select Histogram ParameterEstimates GoodnessofFit FitQuantiles Bins;

proc univariate data=drinknormality;

title 'Normality for Binge Drinking';

var AVEDRNK3;

histogram AVEDRNK3 / normal;

run;

/\* QQ plot\*/

proc univariate data=drinknormality;

title 'Normality for Average Drinks';

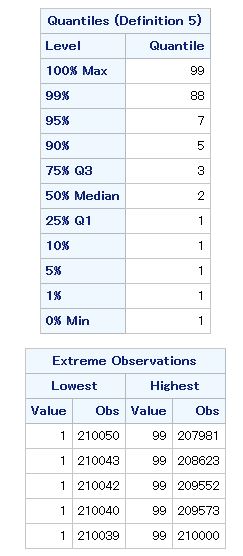
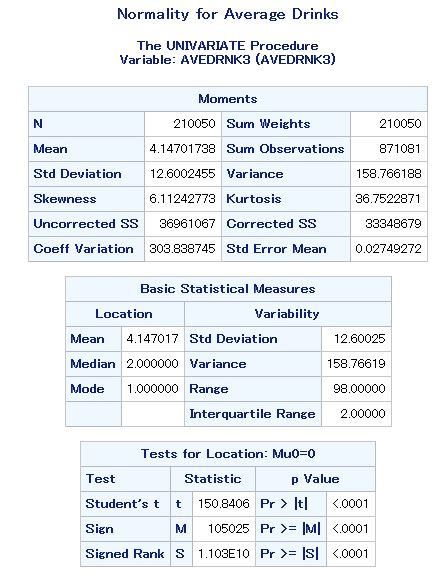
var AVEDRNK3;

qqplot AVEDRNK3 /normal (MU=EST SIGMA=EST COLOR=RED L=1);

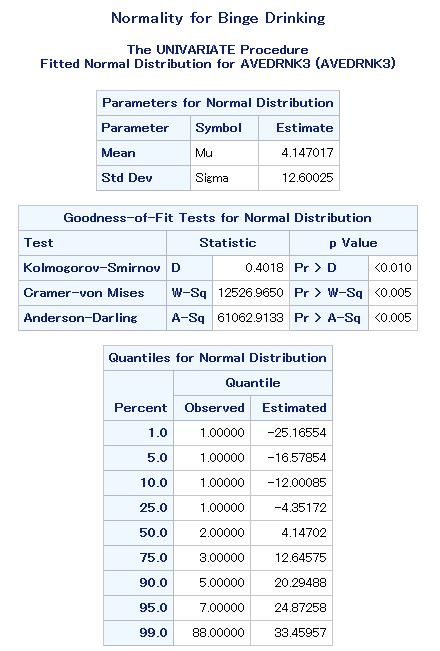
run;

**Results + Interpretation :**

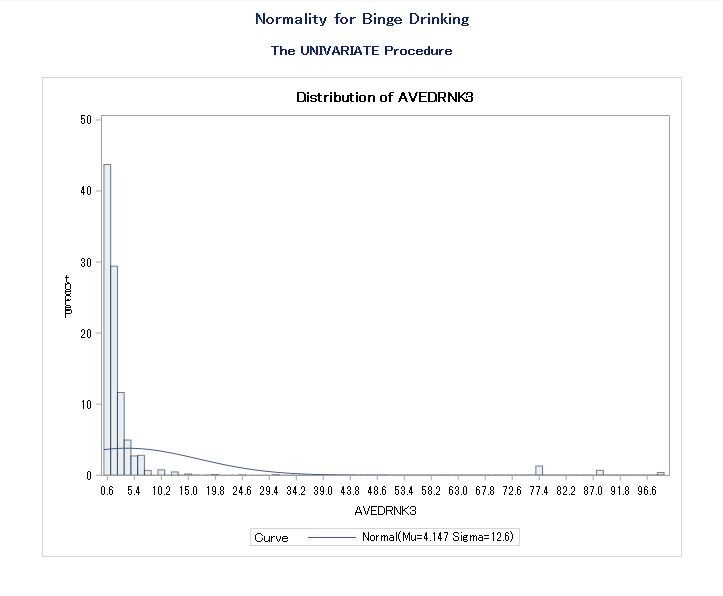
Basics- probably don’t need

****

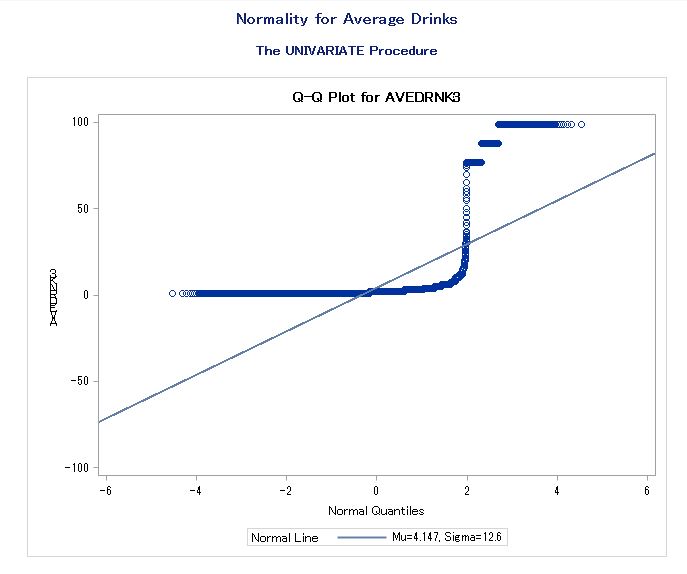
Normality tests- this data is not normal (P-Values for each goodness of fit test is <0.05, meaning it is significantly different, or not normal). However, it is a drinking variable so we expect some skew.

****

Histogram w/ normal curve overlay- again, not normal but we don’t expect normality here

****

QQ-plot - same as everything above. If data were normal, it would be a straight line.

****

**4. Test for H0 with 95% confidence interval**

**Code:**

/\* Test for h0 with 95% CI\*/

data BMIheartattack;

set project;

keep CVDINFR4 \_BMI5;

where CVDINFR4=1;

run;

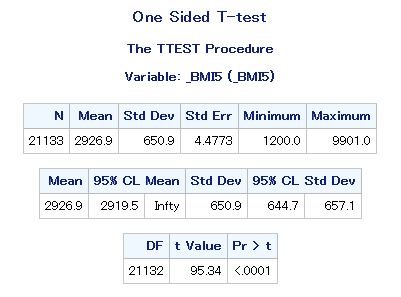
proc ttest data=BMIheartattack h0=2500 sides=U alpha=0.05 plots= none;

title 'One Sided T-test';

var \_BMI5;

run;

**Results:**



**Interpretation:** Here, we are hypothesizing that an “overweight” (>25%) BMI makes a person more likely to experience a heart attack. Therefore, the H0 = 2500 since our codebook assumes two decimal places in the BMI variable. We get a p-value of<0.0001, meaning there is a statistically higher chance of a person with a BMI >25% having a heart attack compared to someone with BMI<25%. However, the hypothesized value is not included in the 95% CI [29.19-infinity]. This means we can only say with 95% confidence that someone who has had a heart attack will fall within the 29.19%+ BMI range (I think).

**5. Contingency table with Chi-square test for H0. Includes OR and RR.**

/\* Chi-square w contingency table\*/

data ageMI;

set project;

keep \_age\_g CVDINFR4;

where CVDINFR4<7;

run;

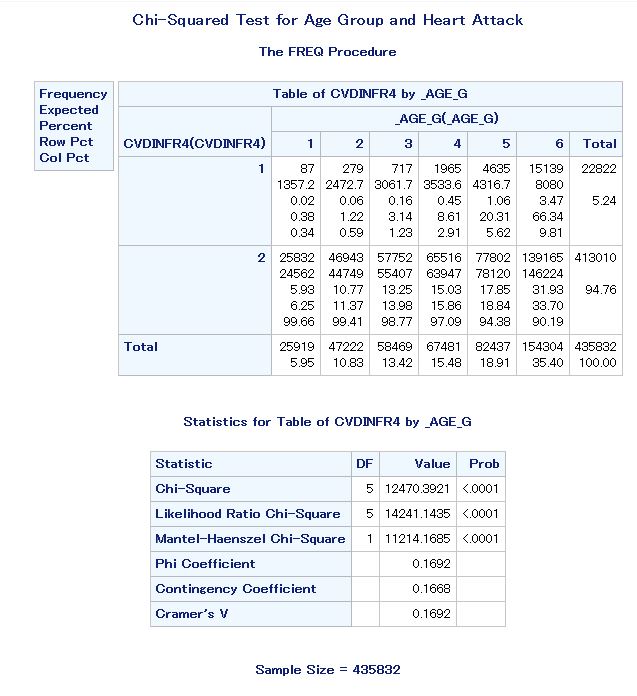
proc freq data=ageMI;

TITLE 'Chi-Squared Test for Age Group and Heart Attack';

tables CVDINFR4\* \_AGE\_G / chisq expected relrisk;

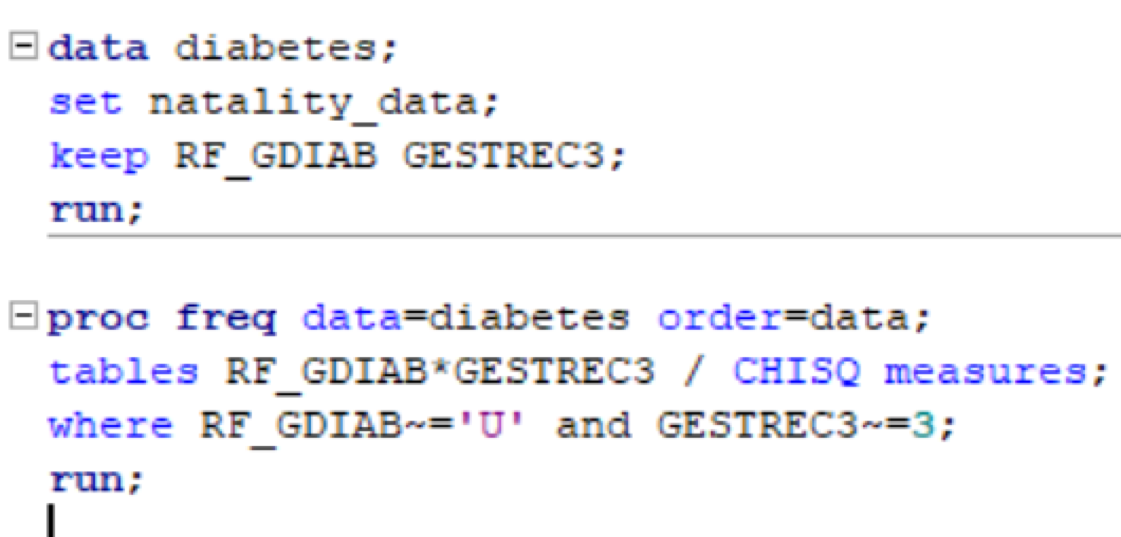
run;

**Results:**

****

\*\*idk why im not getting odds ratio and relative risk here

**Interpretation:** CVDINFR4= heart attack, chi square value is significant, meaning as age increases, likelihood of heart attack increases aka they are not independent of each other(I think).



**6. Independent T-test. Includes 95% CI.**

**Code:**

/\*Independent T-test\*/

data ageveg;

set project;

keep \_AGE\_G \_VEGESU1;

where \_VEGESU1<99998;

run;

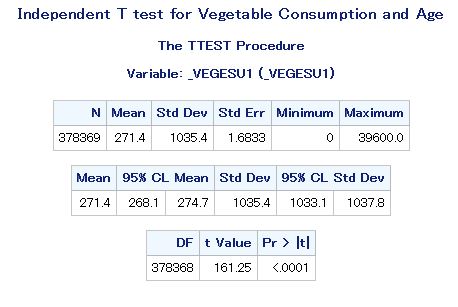
proc ttest data=ageveg;

title 'Independent T test for Vegetable Consumption and Age';

var \_VEGESU1 \_AGE\_G;

run;

**Results:**

****

**Interpretation:**

i have no clue tbh idk if these are good variables to test against each other

**7. Scatter plot of 2 continuous variables and linear correlation.**

**Code:**

/\* Scatter Plot/Correlation \*/

data BMIbyVegetables;

set project;

run;

proc sgplot data=BMIbyVegetables (firstobs=1 obs=500);

title 'BMI by Vegetable Consumption';

scatter x=\_BMI5 y=\_VEGESU1;

Ellipse x=\_BMI5 y=\_VEGESU1;

RUN;

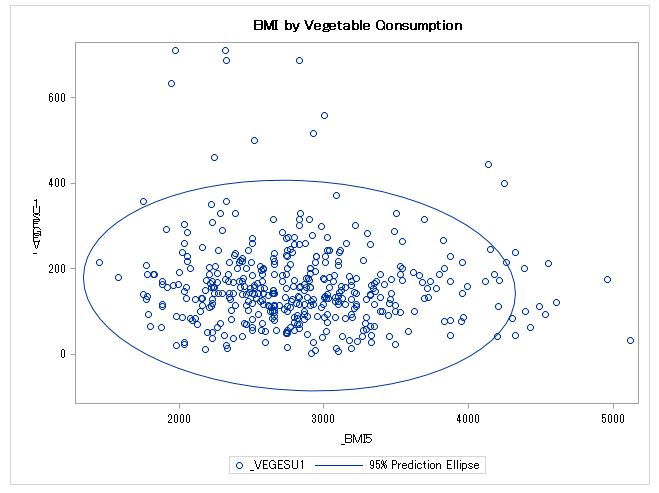
proc corr data=BMIbyVegetables;

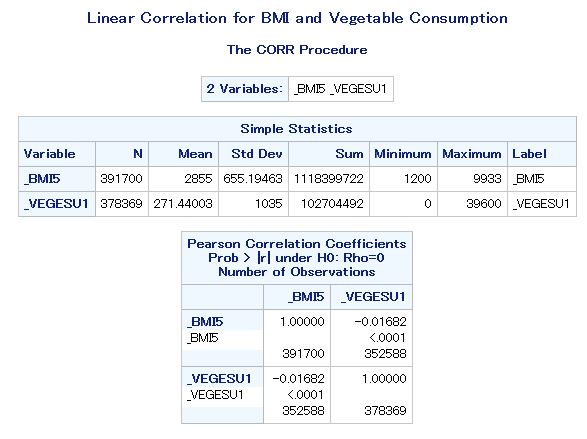
title 'Linear Correlation for BMI and Vegetable Consumption';

var \_BMI5 \_VEGESU1;

run;

**Results:**

****

****

**Interpretation:** \_VEGESU1 and \_BMI5 both assume 2 decimal places, so mean BMI is 28.55% and mean vegetable consumed in a day is 2.71. These values have a loose, negative (but still statistically significant) correlation. This means that as BMI increases, vegetable consumption decreases. This is expected.

proc univariate data=d\_freq normal plot;

title 'Normality for Average alcohol Drinking';

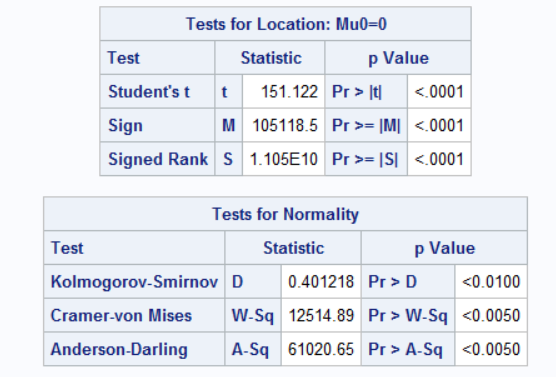
var AVEDRNK3;

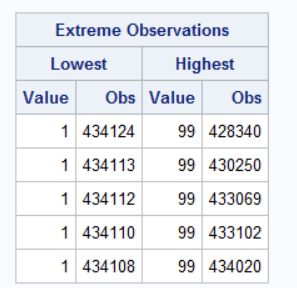
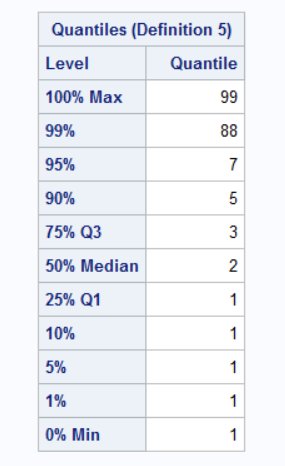
where AVEDRNK3 ne .;

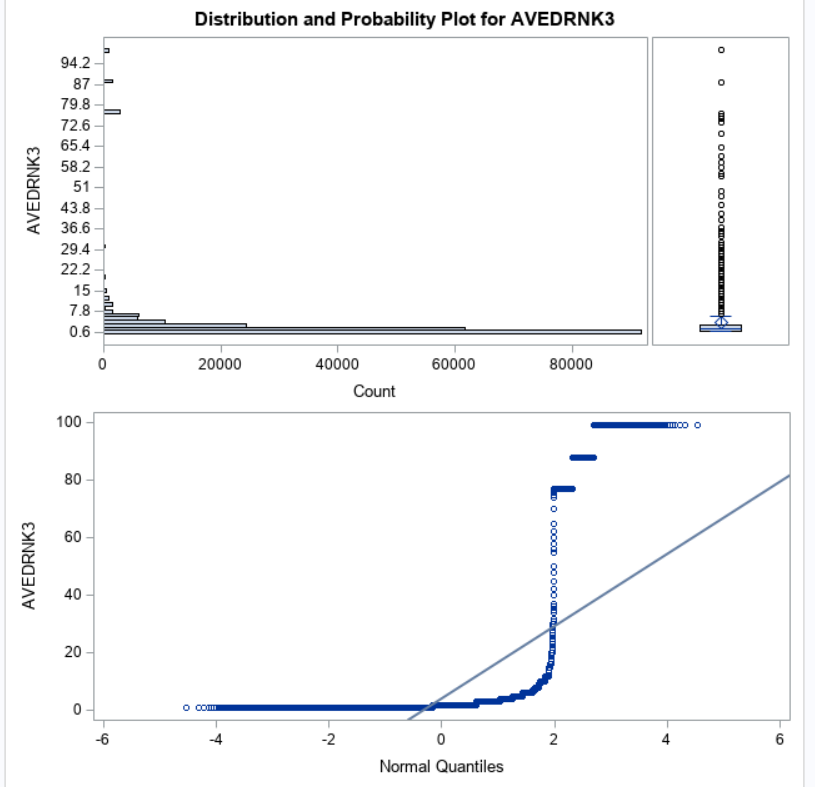
qqplot AVEDRNK3 /normal(MU=EST SIGMA=EST COLOR=RED L=1);

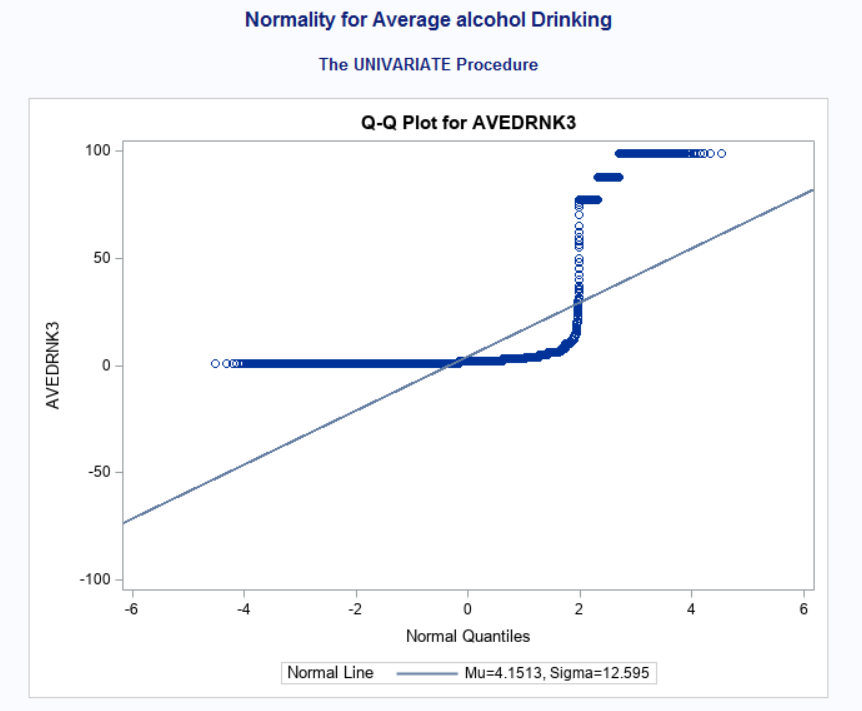
run;

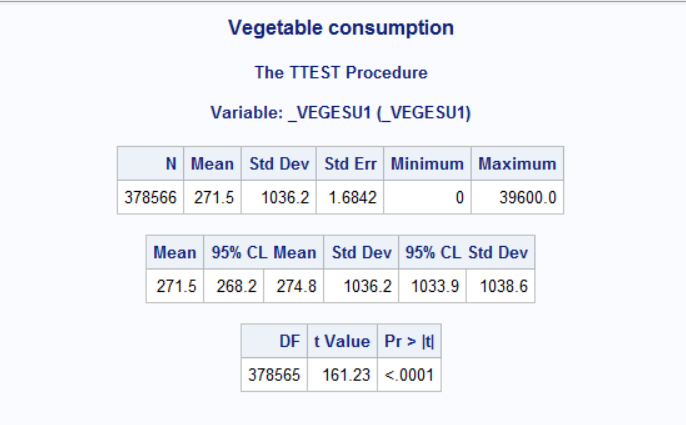












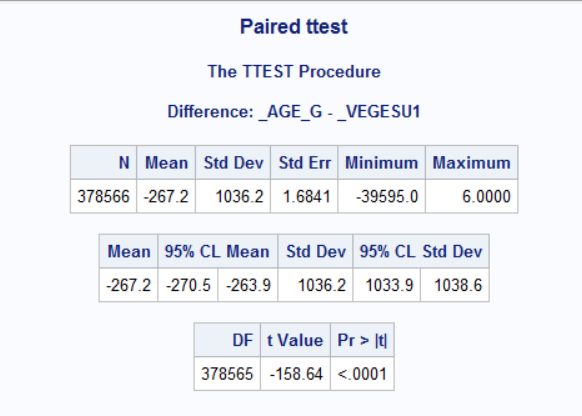
**/\*paired test\*/**

proc ttest data=project.q6;

paired \_AGE\_G \* \_VEGESU1;

title 'Paired ttest';

Run;



**10. Logistic regression analysis. (without selecting variables)**

**data logistic1;**

**set project;**

**where CVDINFR4<7 and CVDINFR4~=.;**

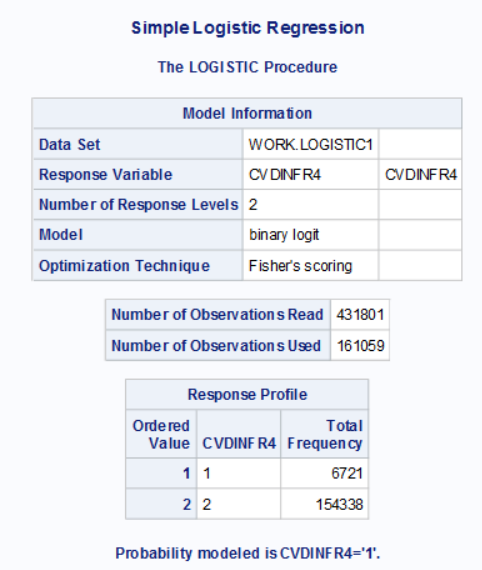
**run;**

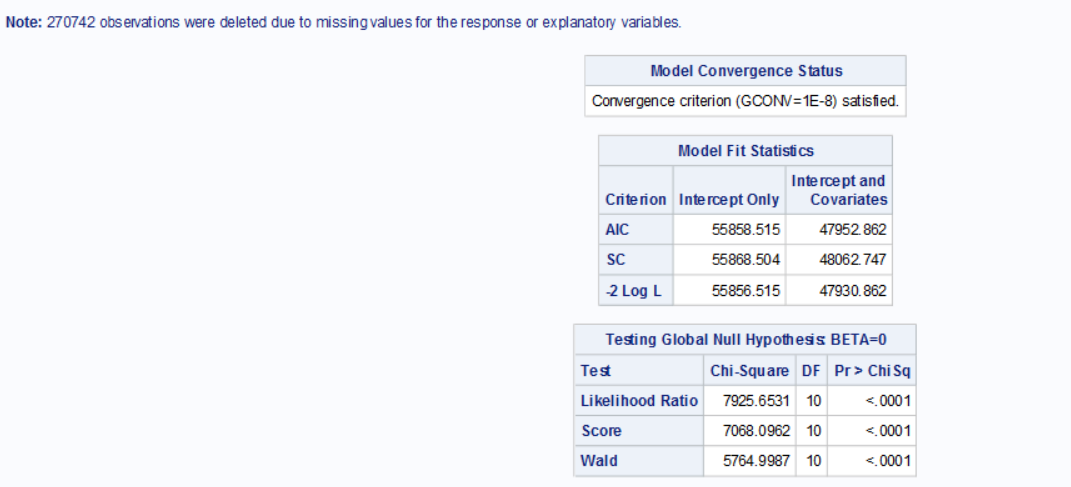
**proc logistic data=logistic1;**

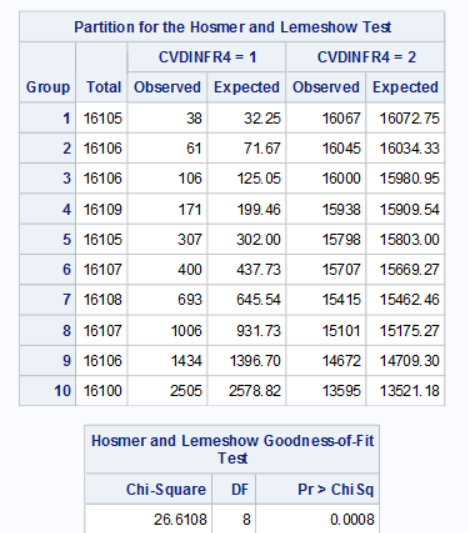
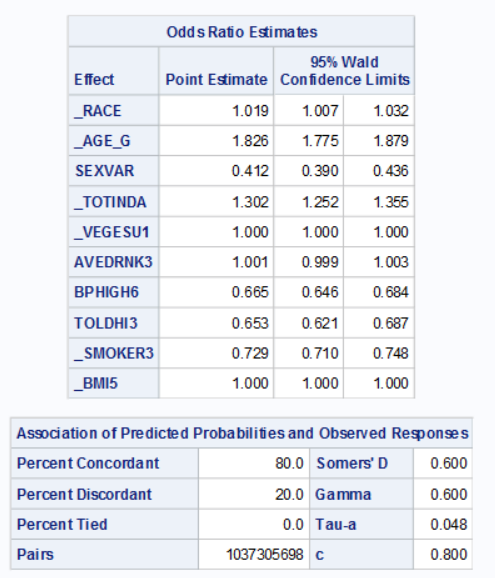
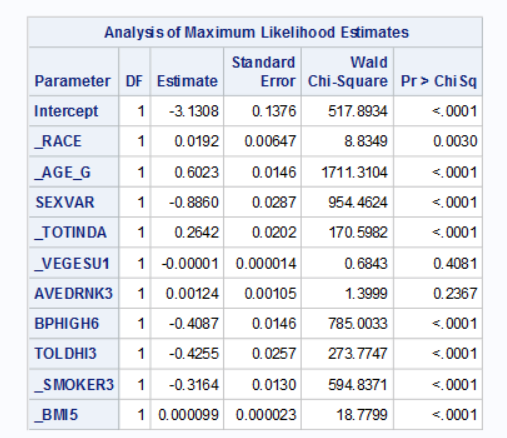
**title 'Simple Logistic Regression';**

**model CVDINFR4 (event=’1’)= \_RACE \_AGE\_G SEXVAR \_TOTINDA \_VEGESU1 AVEDRNK3 BPHIGH6 TOLDHI3 \_SMOKER3 \_BMI5 / lackfit;**

**Run;**

****

****

****

**/\*Logestic regression with selected variables\*/**

**data logistic1;**

**set d\_freq;**

**where CVDINFR4<7 and CVDINFR4~=. and \_AGE\_G= 6 and \_RACE=2 and**

**SEXVAR=1 and \_TOTINDA=2 and BPHIGH6=1 and TOLDHI3=1 and \_SMOKER3=1;**

**run;**

**data logistic2;**

**set logistic1;**

**where CVDINFR4<7 and CVDINFR4~=.;**

**run;**

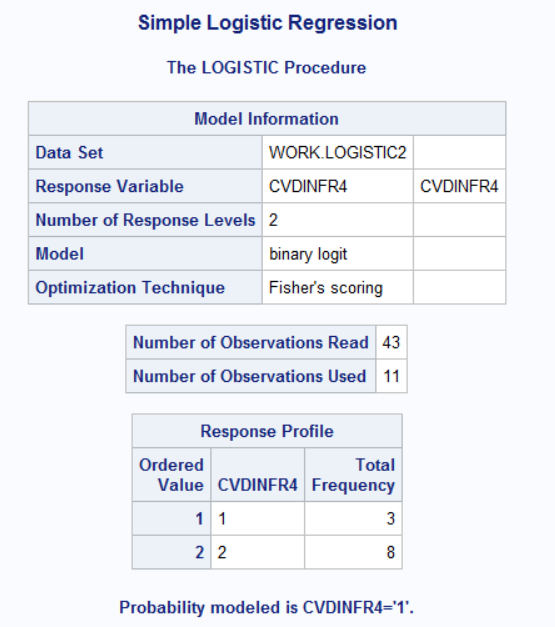
**proc logistic data=logistic2;**

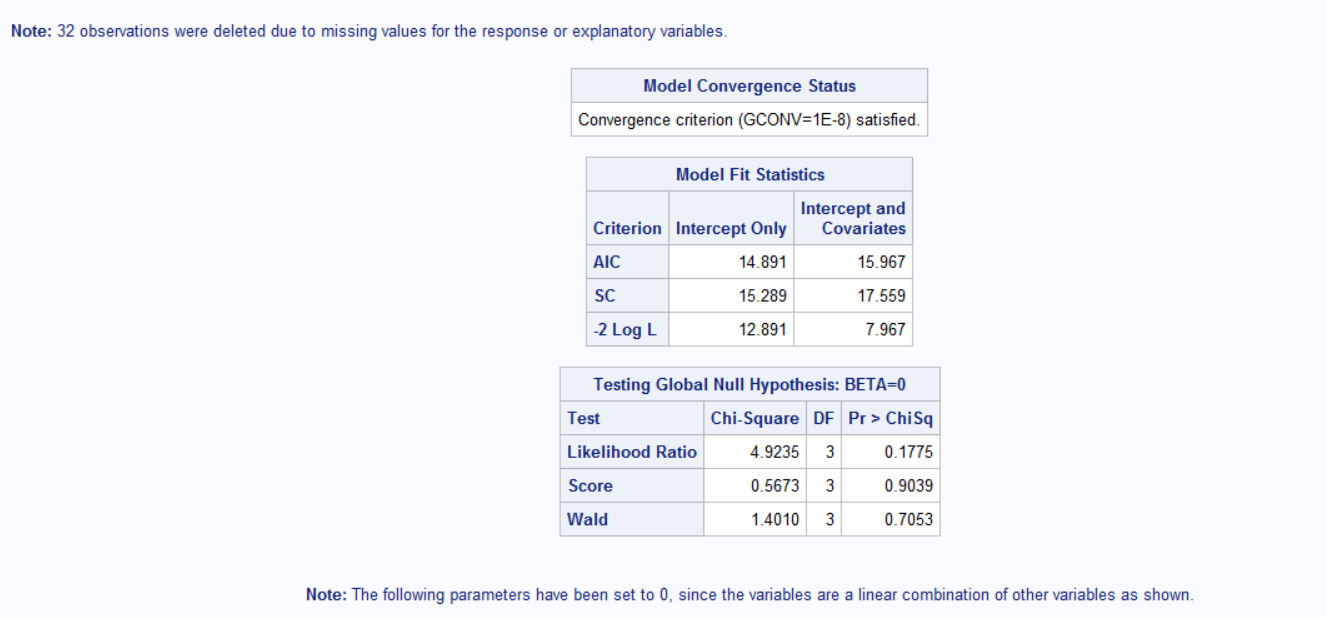
**title 'Simple Logistic Regression';**

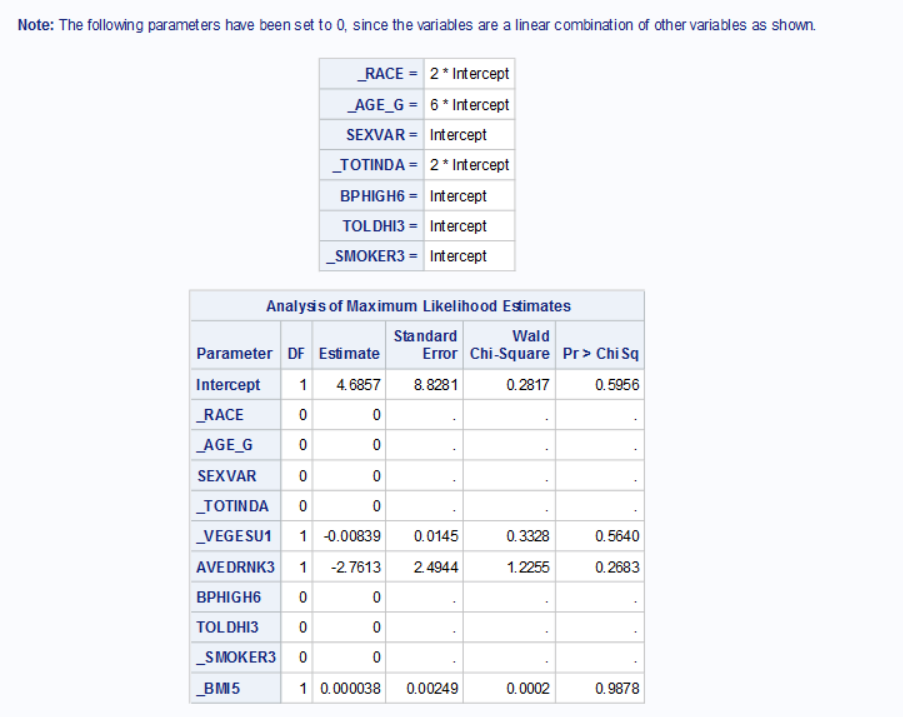
**model CVDINFR4 (event='1')= \_RACE \_AGE\_G SEXVAR \_TOTINDA \_VEGESU1**

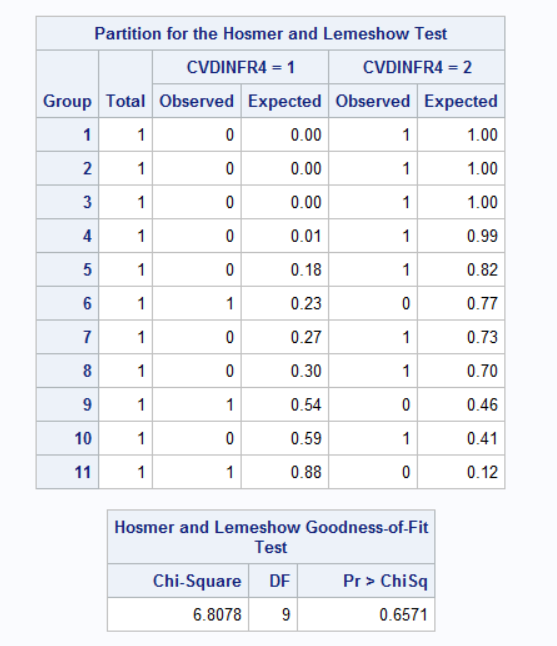
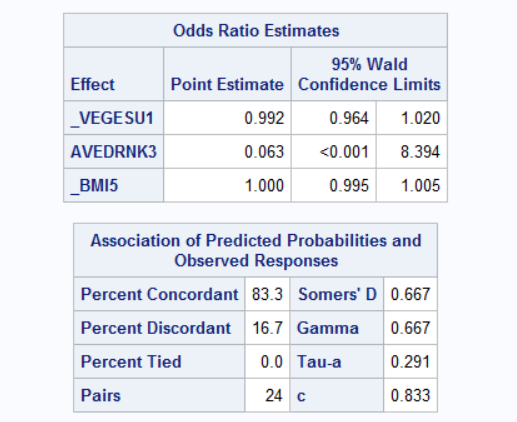
**AVEDRNK3 BPHIGH6 TOLDHI3 \_SMOKER3 \_BMI5 / lackfit;**

**Run;**

****

****

****

****

**Data biglogistic;**

**Set project;**

**If \_BMI5=0 THEN \_BMI5=0; else if \_BMI5=0 <\_BMI5<9999 then \_BMI5=1;**

**If \_RACE=1 then \_RACE1=1; else \_RACE2=0;**

**If \_RACE=2 then \_RACE2=1; else \_RACE2=0;**

**If \_RACE=3 then \_RACE3=1; else \_RACE3=0;**

**If \_RACE=4 then \_RACE4=1; else \_RACE4=0;**

**If \_RACE=5 then \_RACE5=1; else \_RACE5=0;**

**If \_RACE=6 then \_RACE6=1; else \_RACE6=0;**

**If \_RACE=7 then \_RACE7=1; else \_RACE7=0;**

**If \_RACE=8 then \_RACE8=1; else \_RACE8=0;**

**If AVEDRNK3=0 then AVEDRINK=0; else if 0<AVEDRNK3<76 then AVEDRNK3=1; else AVEDRNK3=0;**

**IF \_TOTINDA=1 then \_TOTINDA=1; else if \_TOTINDA =0 then \_TOTINDA=0;**

**If CVDINFR4=1 then CVDINFR4=1; elseif CVDINFR4=0 then CVDINFR4=0;**

**If SEXVAR=1 then SEXVAR=1; else if SEXVAR=0 then SEXVAR=0;**

**If \_VEGESU1=0 then \_VEGESU1=0; else if 0< \_VEGESU1<99998 then \_VEGESU1=1;**

**If BPHIGH6=1 then BPHIGH6=1; else BPHIGH6=0;**

**If TOLDHI3=1 then TOLDHI3=1; else TOLDHI3=0;**

**If \_AGE\_G=1 then \_AGE\_G1=1; else \_AGE\_G1=0;**

**If \_AGE\_G=2 then \_AGE\_G2=1; else \_AGE\_G2=0;**

**If \_AGE\_G=3 then \_AGE\_G3=1; else \_AGE\_G3=0;**

**If \_AGE\_G=4 then \_AGE\_G4=1; else \_AGE\_G4=0;**

**If \_AGE\_G=5 then \_AGE\_G5=1; else \_AGE\_G5=0;**

**If \_AGE\_G=6 then \_AGE\_G6=1; else \_AGE\_G6=0;**

**If \_SMOKER3=1 then \_SMOKER31=1; else \_SMOKER31=0;**

**If \_SMOKER3=2 then \_SMOKER32=1; else \_SMOKER32=0;**

**If \_SMOKER3=3 then \_SMOKER33=1; else \_SMOKER33=0;**

**If \_SMOKER3=4 then \_SMOKER34=1; else \_SMOKER34=0;**

**Run;**

**Proc logistic data=biglogistic;**

**Title ‘Multiple Logistic Regression’;**

**Model CVDINFR4(event=’1’) = AVEDRNK3 \_TOTINDA )VEGESU1 \_AGE\_G SEXVAR \_RACE BPHIGH6**

**TOLDHI3 \_SMOKER3 \_BMI5 / lackfit;**

**Where CVDINFR~=7 and CVDINFR4~=9 and CVDINFR~=.;**

**run;**