Midterm Lipid5238 Dataset.

NOTE: In the interest of fairness, questions about the midterm will only be answered in class.

Submit this document with your answers by the due date.

You will be asked to append all code used at the end.

Your name: chenchen zhou

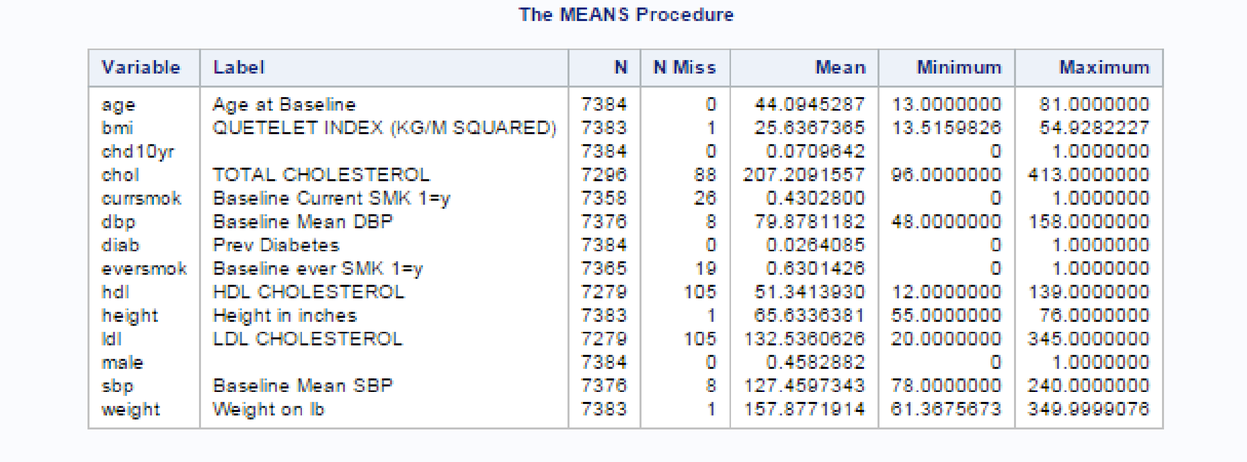
I certify that the following is solely my own work.

The SAS dataset lipid5238 contains 14 variables:

age	Age at baseline exam in years
bmi	Body mass index (kg/m^2)
chd10yr	0=did not develop CHD, 1=developed CHD
Chol	Total cholesterol (mg/dl)
Currsmok	0=does not smoke cigarettes, 1=smokes cigarettes
Dbp	Diastolic blood pressure (mmHg)
diab	0=Non-diabetic, 1=diabetic
Eversmok	0=Never smoked cigarettes, 1=smoked cigarettes (past or present)
Hdl	High density lipoproteins (mg/dl)
Height	Height in inches
Ldl	Low density lipoproteins
Male	0=female, 1=male
Sbp	Systolic blood pressure (mmHg)
Weight	Weight in pounds

Your job is to develop a model that uses logistic regression to predict the development of CHD (chd10yr). The development is totally DIY, you may not use algorithmic variable selection. It should follow the development given in class.

1. Conduct an exploratory analysis of the data to determine how many missing values exist. Also examine whether there is collinearity present in the data. Write a paragraph here describing how you plan to handle missing data (either a complete case analysis or imputation). If there is collinearity among the variables, describe it and select one for your candidate model. Explain why you selected the particular variable for inclusion.



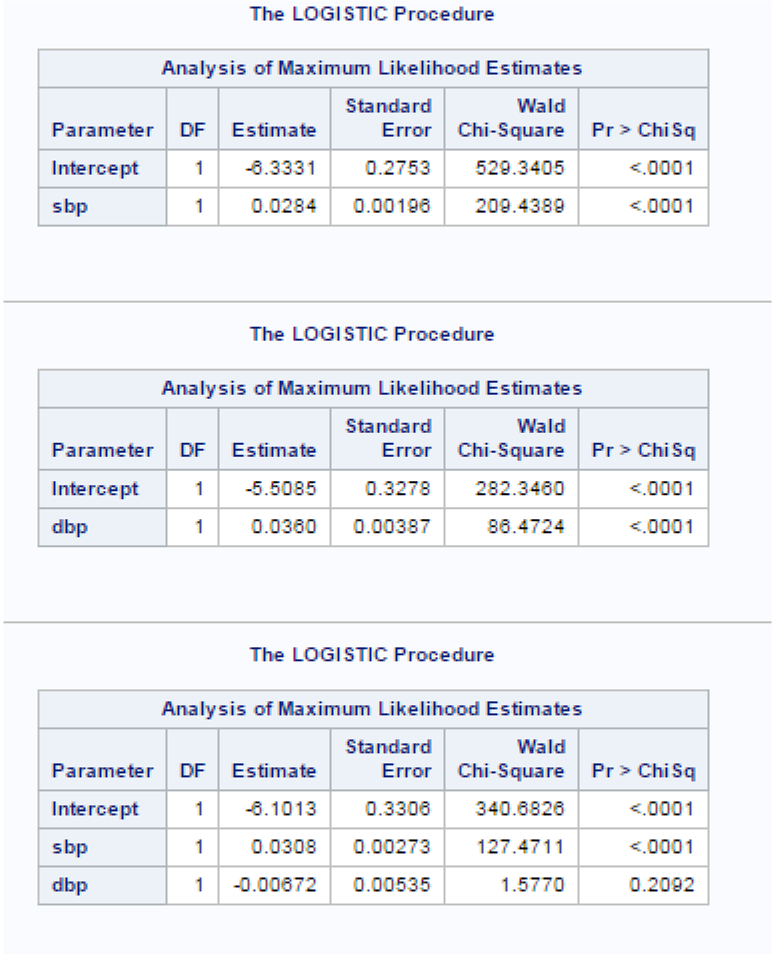
The MEANS Procedure

Variable	Label	N	N Miss	Mean	Minimum	Maximum
age	Age at Baseline	7384	0	44.0945287	13.0000000	81.0000000
bmi	QUETELET INDEX (KG/M SQUARED)	7383	1	25.6367365	13.5159826	54.9282227
chd10yr		7384	0	0.0709642	0	1.0000000
chol	TOTAL CHOLESTEROL	7296	88	207.2091557	96.0000000	413.0000000
currsmok	Baseline Current SMK 1=y	7358	26	0.4302800	0	1.0000000
dbp	Baseline Mean DBP	7376	8	79.8781182	48.0000000	158.0000000
diab	Prev Diabetes	7384	0	0.0264085	0	1.0000000
eversmok	Baseline ever SMK 1=y	7365	19	0.6301426	0	1.0000000
hdl	HDL CHOLESTEROL	7279	105	51.3413930	12.0000000	139.0000000
height	Height in inches	7383	1	65.6336381	55.0000000	76.0000000
ldl	LDL CHOLESTEROL	7279	105	132.5360626	20.0000000	345.0000000
male		7384	0	0.4582882	0	1.0000000
sbp	Baseline Mean SBP	7376	8	127.4597343	78.0000000	240.0000000
weight	Weight on lb	7383	1	157.8771914	61.3675673	349.9999076

From the above table, we could see there are at most 105 observations containing missing values. Comparing to a total number of 7384 observations, 105 is an acceptable number and it has not much effect in developing our model if we delete them. To be more careful about this argument, I examine the effect of unknown chol, hdl, ldl determinations(the variables have most missing values). The Fisher's exact test shows that the missing data is independent of chd10yr. So I will delete those observations which contain missing values in variables. I use a macro to generate a data set that has no missing data for any variables.

To check the existence of collinearity among those variables, I first check correlation coefficients between any two variables (generate correlation coefficients matrix). Select those pairs of two variables which are highly and positive correlated (above 0.6), and apply the chi square test for coefficients for individual variables in a univariate model and in a bivariate model. If coefficient is significant in univariate model for both variables but not significant in bivariate model, then collinearity exists.

Take variables (bmi, weight) as an example,



The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.3331	0.2753	529.3405	<.0001
sbp	1	0.0284	0.00196	209.4389	<.0001

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.5085	0.3278	282.3460	<.0001
dbp	1	0.0360	0.00387	86.4724	<.0001

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.1013	0.3306	340.6826	<.0001
sbp	1	0.0308	0.00273	127.4711	<.0001
dbp	1	-0.00672	0.00535	1.5770	0.2092

From the p value, we conclude that both sbp and dbp are significant in the respective univariate model. However, dbp becomes not significant if both sbp and dbp are included in the model. So there is collinearity here. Dpb seems to have little effect because it overlaps considerably with sbp in the model. Deleting dps can reduce standard errors of other estimated effects.

There exists collinearity for the following paris of variables.
(bmi weight), (sbp dbp), (currsmok eversmok), (male height), (male weight).

I will select the variables that are significant in both the univariate and bivariate model and drop those variables that are not significant in either univariate model or bivariate model.

Among those variables, height is not significant. So this variable must be dropped.

I do the above test for the following paris.

	variables	Decision
1	bmi weight	choose bmi over weight
2	ldl chol	Both ldl chol seleted
3	sbp dbp	choose sbp over dbp
4	weight height	choose weight over height
5	currsmok eversmok	choose currsmok over eversmok
6	male height	choose male over height
7	male weight	choose male over weight

For (ldl, chol), they are both significant in both univariate and bivariate models. Besides that, I use proc logistic to check the deviances and c statistics.

Model	-2 Log L	c
ldl	3575.174	0.667
chol	3564.691	0.671
ldl chol	3559.777	0.674

Both models have a great reduction in -2logL and have a high c value. So I decide to keep ldl chol in the first step.

For weight, dbp, eversmok ,height, they are not significant in bivariate models and models with weight,eversmok,height have a very high -2logL and low c. So I choose to drop these 4 variables.

So in the first step, I select my candidate model to contain predictors:

age sbp chol diab currsmok bmi male hdl ldl

2. Create an analytic file to be used in the analysis.
   Write a paragraph here describing the analytic file.

The analytic file I created contains no missing values with 7245 observations. Except the dependent variable chd10yr, it contains 6 continuous variables, age, sbp, chol, bmi, hdl, ldl and 3 categorical variables diab, currsmok, male.

3. Using the analytic file, conduct a complete univariate analysis of the univariate relationships of the variables to chd10yr.
   Write a paragraph here summarizing what you learned from the analysis. That is, which variables might be candidates for a multivariate model.

To conduct a univariate analysis of univariate relationship to the dependent variable, I do the following analysis.

a. Proc freq for categorical variables comparing incidence in the differing level of each categorical variable, and check chi-square statistics to have a quick check whether the variables have relationship with dependent variables.
b. Draw the Logit graphs of all continuous variables
c. Conduct a t-test for continuous variable comparing the average level in those who did and those who did not develop CHD in ten years.
d. Use proc logistic to check the reduction in deviance values and check c value.

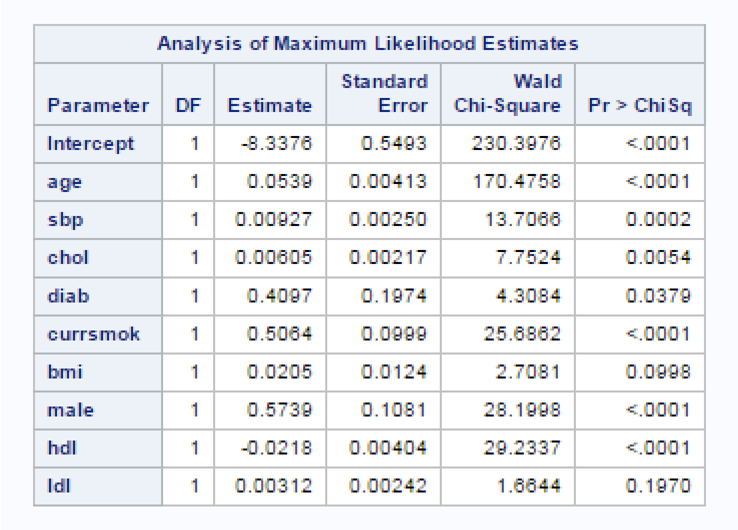
After I complete the above analysis, I see all continuous variables having a nice graph of logits, their logit graphs fit a line very well for all continuous variables except bmi. But the graph still show that there is a relationship between bmi and chd10yr. The slopes are all positive except for hdl. T tests show that all continuous variables a significant relationship between continuous variables and dependent variable chd10yr. For categorical variables, chi-square statistics (also Fisher's exact test)show that they have a significant effect in predicting the occurrence of chd10yr. Continuous variables all have a huge reduction in -2logL and a high c value. So until here, all variables selected from the first step should be included. That is

age sbp chol diab currsmok bmi male hdl ldl

4. Select your first candidate model based on the previous analysis and obtain the appropriate logistic model and obtain the Hosmer-Lemeshow goodness of fit statistic.

Write a paragraph here, summarizing the candidate model.

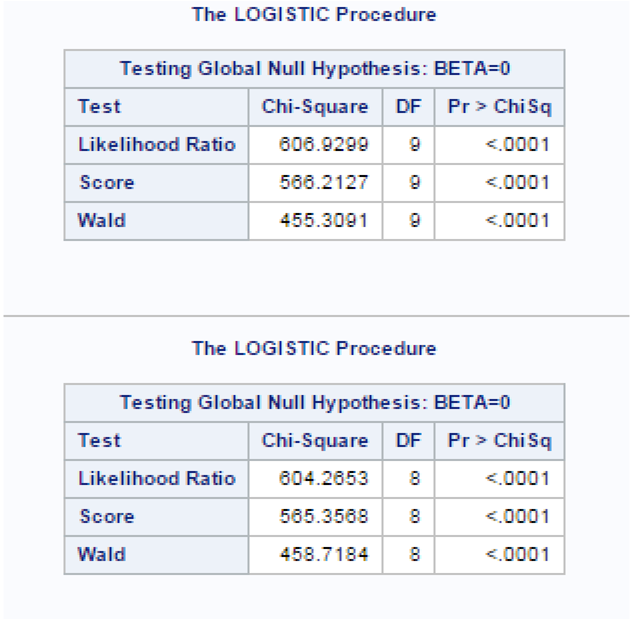
The candidate model I choose contains independent variables
age sbp chol diab currsmok bmi male hdl ldl



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.3376	0.5493	230.3976	<.0001
age	1	0.0539	0.00413	170.4758	<.0001
sbp	1	0.00927	0.00250	13.7066	0.0002
chol	1	0.00605	0.00217	7.7524	0.0054
diab	1	0.4097	0.1974	4.3084	0.0379
currsmok	1	0.5064	0.0999	25.6862	<.0001
bmi	1	0.0205	0.0124	2.7081	0.0998
male	1	0.5739	0.1081	28.1998	<.0001
hdl	1	-0.0218	0.00404	29.2337	<.0001
ldl	1	0.00312	0.00242	1.6644	0.1970

When I use proc logistic to check the estimates, I see bmi and ldl becomes non significant in this multivariate model and odds ratio confidences for those two variables contain one which implies that they don’t have a significant impact on chd10yr when other variables are included. I do the likelihood ration test for models with bmi and without bmi, and I also do it for models with ldl and without ldl as follows.

bmi



The LOGISTIC Procedure

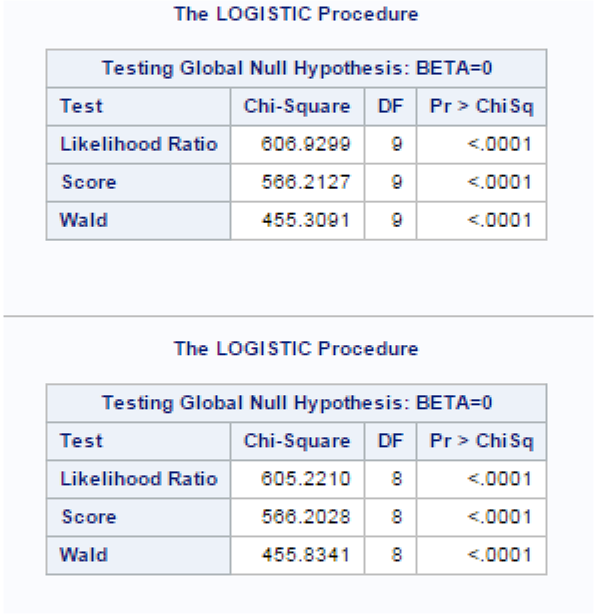
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	606.9299	9	<.0001
Score	566.2127	9	<.0001
Wald	455.3091	9	<.0001

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	604.2653	8	<.0001
Score	565.3568	8	<.0001
Wald	458.7184	8	<.0001

606.9299-604.2653=2.6646 (chi-square 1 df)

ldl



The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	606.9299	9	<.0001
Score	566.2127	9	<.0001
Wald	455.3091	9	<.0001

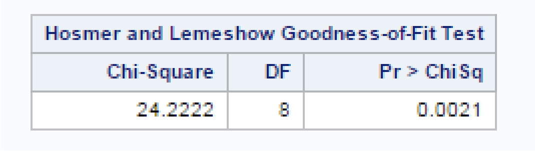
The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	605.2210	8	<.0001
Score	566.2028	8	<.0001
Wald	455.8341	8	<.0001

606.9299-605.2210=1.7089 (chi-square 1 df)

The chi-square critical value at 0.05 is 3.84, so we can't reject the null hypothesis of likelihood ratio test. So I could consider to drop those two variables.

Even though we keep those two variables, the result for H-L goodness of fit test is as follows, from the p value, we see the model we choose has a poor fit with our data.



Hosmer and Lemeshow Goodness-of-Fit Test

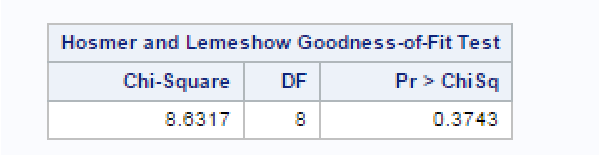
Chi-Square	DF	Pr > ChiSq
24.2222	8	0.0021

As it has a poor fit, so for now I am still very careful about dropping bmi and ldl. In conclude, the candidate model I choose contains independent variables

age sbp chol diab currsmok bmi male hdl ldl

5. Examine whether a transformation of any of the variables may be necessary. Write a paragraph here, reporting what you did and what you conclude from your analysis.

At first, from the goodness test see we see the model in step 4 does not fit well with our data. To examine whether a transformation is needed, I check the plots of loess smooths of the chd10yr by each continuous independent variable. From the logit smooth plots, I don't see linear lines. So a transformation of variables is suggested. For example, logit smooth plots for age is not linear. To fix this, I add age*age variable in our model and then I do the H-L goodness of fit test. I see much improvement when combining this term.



Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
8.6317	8	0.3743

The p-value indicates that we can't reject the null hypothesis which implies that there is no significant difference between the expected events and observed events.

Apart from age*age, I also tried add quadratic terms of each other continuous variables, like sbp*sbp, chol*chol, bmi*bmi hdl*hdl ldl*ldl, but adding none of them has a better fit than adding age*age. So I just add age*age into my model. Now in this step, the new candidate model is containing

age age*age sbp chol diab currsmok male hdl ldl

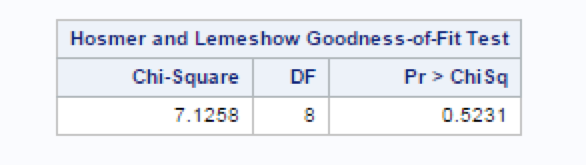
as our independent variables.

6. Examine whether you should include interaction terms in your model. Restrict the examination to interaction between age and male with the other variables.
   Write a paragraph here, reporting what you did and what you conclude from your analysis.

To examine the interaction terms, I just check every possible interaction between age and male with the other variables. There are too many combinations, I just list those several combinations which give out a great performance in H-L goodness of fit test.

First I tried the interaction between male with other variables, then I found when I add male*diab variable into my model, it gives highest p value in H-L test as follows. It's a good improvement, so I keep male*diab.

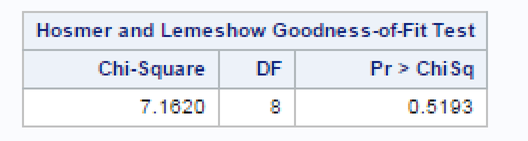
male*diab



Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
7.1258	8	0.5231

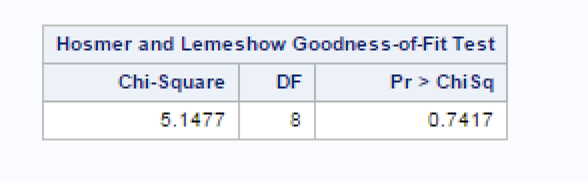
After adding male*diab, I tried all interaction terms between age and other variables. I just list several results which gives high p value.

age*sbp



Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
7.1620	8	0.5193

age*hdl



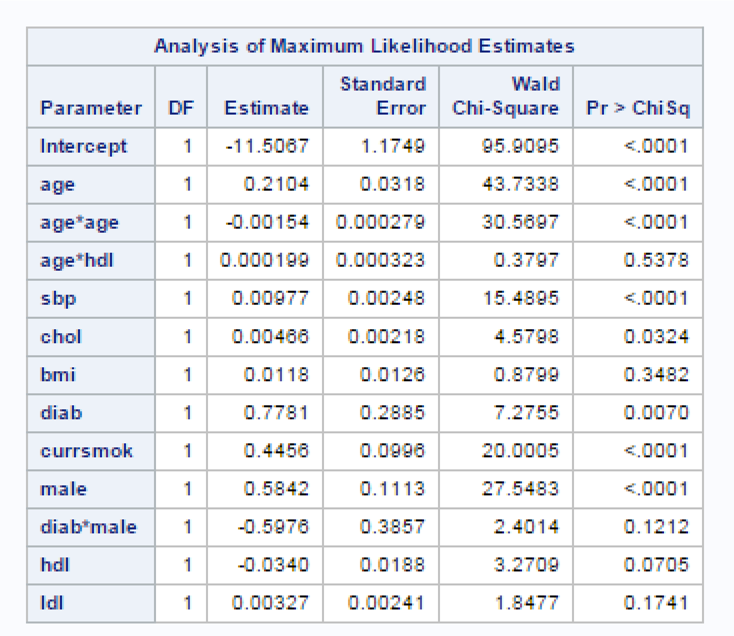
Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
5.1477	8	0.7417

I tried all interaction terms with age and male. The candidate model I think is best in this step is
age age*age age*hdl sbp chol bmi male diab currsmok male male*diab hdl ldl

7. Decide on the final model and obtain the appropriate logistic model including the Hosmer-Lemeshow goodness of fit statistic.

Write a paragraph here describing your final model and whether you would conclude that it provides an adequate fit to the data.

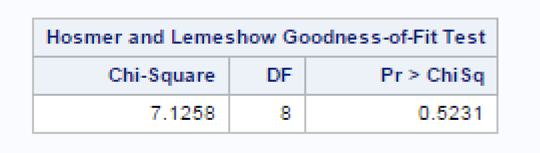
In step 6, the model I choose has the following independent variables:



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-11.5067	1.1749	95.9095	<.0001
age	1	0.2104	0.0318	43.7338	<.0001
age*age	1	-0.00154	0.000279	30.5697	<.0001
age*hdl	1	0.000199	0.000323	0.3797	0.5378
sbp	1	0.00977	0.00248	15.4895	<.0001
chol	1	0.00466	0.00218	4.5798	0.0324
bmi	1	0.0118	0.0126	0.8799	0.3482
diab	1	0.7781	0.2885	7.2755	0.0070
currsmok	1	0.4456	0.0996	20.0005	<.0001
male	1	0.5842	0.1113	27.5483	<.0001
diab*male	1	-0.5976	0.3857	2.4014	0.1212
hdl	1	-0.0340	0.0188	3.2709	0.0705
ldl	1	0.00327	0.00241	1.8477	0.1741

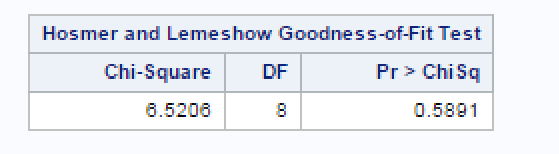
From above table, I see coefficients for age*hdl, bmi, diab*male,hdl,ldl are not significant. So I tried to delete some of them, and check the -2logL, c, H-L goodness of fit to decide whether to remain them in the model or delete them. For example, I list the following result.
delete age*hdl



Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
7.1258	8	0.5231

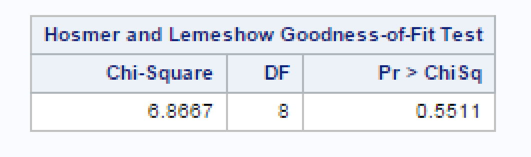
delete ldl



Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
6.5206	8	0.5891

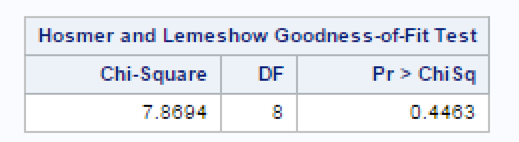
delete diab*male



Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
6.8667	8	0.5511

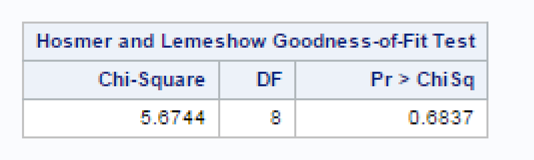
delete age*hdl and ldl



Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
7.8694	8	0.4463

delete bmi



Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
5.6744	8	0.6837

So there is much decrease if I delete some of them. In conclude, the model I think has a good fitness with least variables is the following.

age age*age age*hdl sbp chol bmi diab currsmok male male*diab hdl ldl

There are 12 independent variables in my model,

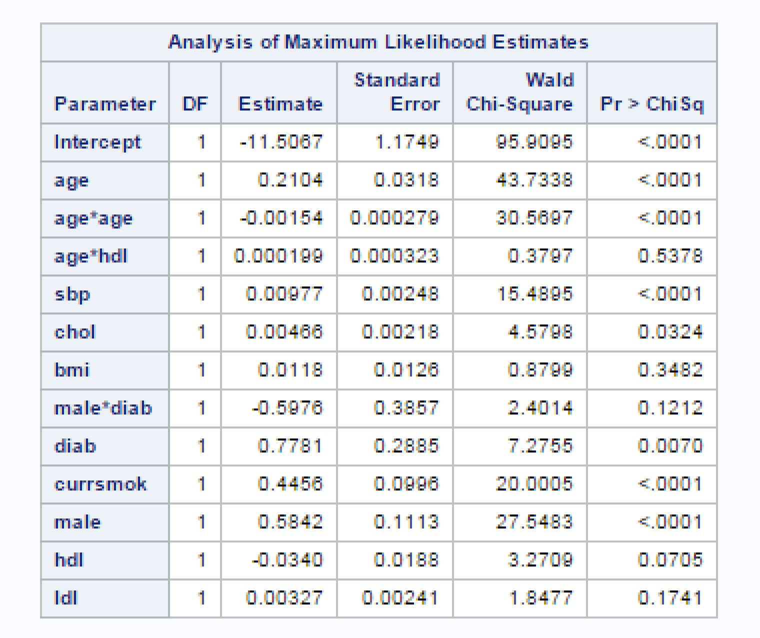
main factor: age sbp chol bmi male diab currsmok hdl ldl

transformation: age*age

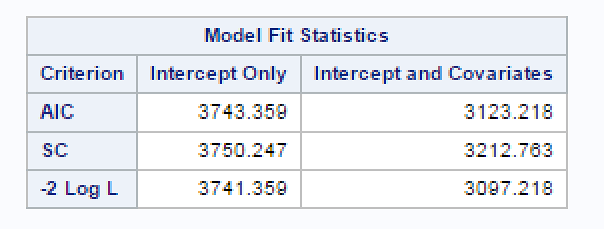
interaction: age*hdl male*diab

The estimates, -2log L, c ,H-L goodness of fit are listed as follows.

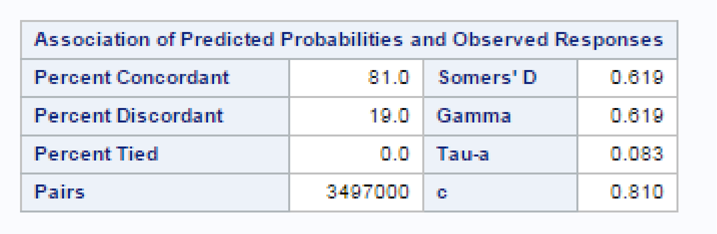
They all show that the model provides an adequate fit to the data.



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-11.5067	1.1749	95.9095	<.0001
age	1	0.2104	0.0318	43.7338	<.0001
age*age	1	-0.00154	0.000279	30.5697	<.0001
age*hdl	1	0.000199	0.000323	0.3797	0.5378
sbp	1	0.00977	0.00248	15.4895	<.0001
chol	1	0.00466	0.00218	4.5798	0.0324
bmi	1	0.0118	0.0126	0.8799	0.3482
male*diab	1	-0.5976	0.3857	2.4014	0.1212
diab	1	0.7781	0.2885	7.2755	0.0070
currsmok	1	0.4456	0.0996	20.0005	<.0001
male	1	0.5842	0.1113	27.5483	<.0001
hdl	1	-0.0340	0.0188	3.2709	0.0705
ldl	1	0.00327	0.00241	1.8477	0.1741



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3743.359	3123.218
SC	3750.247	3212.763
-2 Log L	3741.359	3097.218



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	81.0	Somers' D	0.619
Percent Discordant	19.0	Gamma	0.619
Percent Tied	0.0	Tau-a	0.083
Pairs	3497000	c	0.810

8. Copy all of your code here.
   Ps:There are warning messages when use proc loess. This is for accelerating the running speed. We can use PLOTS(MAXPOINTS=NONE) option to eliminate these warning messages.
   In the last step, when I add or delete some variables to check their fitness. Because there are too many cases, when I do this step, I modify the model in one proc. So the code I presented here doesn't include every try.

```
%let path=/courses/d0f434e5ba27fe300;

libname s5238 "&path/s5238";
```

```
proc contents data=s5238.lipid5238;
run;
proc means data=s5238.lipid5238 n nmiss mean min max;
run;

data tmp1;
   set s5238.lipid5238;
   nochol=(chol=.);
   nohdl=(hdl=.);
   noldl=(ldl=.);
run;

proc freq data=tmp1;
  tables (nochol nohdl noldl)*chd10yr/nocol nopercent exact;
run;



%let inputs=age bmi chd10yr chol currsmok dbp diab eversmok hdl height ldl
male sbp weight;
%let inputs1=age bmi chol dbp hdl height ldl sbp weight;
%let inputs2=chd10yr currsmok diab eversmok male;
%macro completecase(outdat=tmp,indat=,vars=);
data &outdat (drop=nummiss);
```

```
   set &indat (keep=&vars);
   nummiss=nmiss(of &vars);
   if nummiss=0;
run;
%mend completecase;

%completecase(indat=s5238.lipid5238,vars=&inputs);

proc corr data=tmp;
var &inputs;
run;

ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=bmi;
run;
ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=weight;
run;
ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=bmi weight;
run;
```

```
ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=ldl;
run;
ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=chol;
run;
ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=ldl chol;
run;

ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=sbp;
run;
ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=dbp;
run;
ods select parameterestimates;
proc logistic data=tmp;
```

```
model chd10yr(event="1")=sbp dbp;
run;


ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=height;
run;
ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=weight;
run;
ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=height weight;
run;



ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=eversmok;
run;
ods select parameterestimates;
```

```
proc logistic data=tmp;
model chd10yr(event="1")=currsmok;
run;
ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=eversmok currsmok;
run;


ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=male;
run;
ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=height;
run;
ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=male height;
run;

ods select parameterestimates;
proc logistic data=tmp;
```

```
model chd10yr(event="1")=male;
run;
ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=weight;
run;
ods select parameterestimates;
proc logistic data=tmp;
model chd10yr(event="1")=male weight;
run;


proc logistic data=tmp;
model chd10yr(event="1")=ldl;
run;


proc logistic data=tmp;
model chd10yr(event="1")=chol;
run;

proc logistic data=tmp;
model chd10yr(event="1")=ldl chol;
run;
```

```
data analyfile;
set tmp(keep=chd10yr age sbp chol diab currsmok bmi male hdl ldl);
run;
proc contents data=analyfile;
run;
proc means data=analyfile n nmiss mean min max;
run;


proc freq data=analyfile;
tables (male currsmok diab)*chd10yr/nocol nopercent chisq;
run;

%macro PlotLogits(indata=,numgrp=5,indepvar=,depvar=);
proc rank data=&indata groups=&numgrp out=Ranks;
   var &indepvar;
   ranks Bin;
run;
proc sql;
  create table toplot as
  select
          avg(&indepvar) as mean label="Mean of group",
          sum(&depvar) as num_chd label="Number of Events",
          count(*) as binsize label="Number at Risk",
          log((calculated num_chd+1)/
       (calculated binsize-calculated num_chd+1)) as logit
          from ranks
          group by bin;
quit;
proc sgscatter data=toplot;
   plot Logit*mean /
        reg markerattrs=(symbol=asterisk color=blue size=15);
   title "Estimated Logit Plot";
run;
title;
%mend PlotLogits;

%PlotLogits(indata=analyfile,
           numgrp=10,
           indepvar=age,
           depvar=chd10yr);
%PlotLogits(indata=analyfile,
           numgrp=10,
           indepvar=sbp,
           depvar=chd10yr);
%PlotLogits(indata=analyfile,
           numgrp=10,
           indepvar=chol,
           depvar=chd10yr);
%PlotLogits(indata=analyfile,
           numgrp=10,
           indepvar=bmi,
           depvar=chd10yr);

%PlotLogits(indata=analyfile,
           numgrp=10,
           indepvar=hdl,
           depvar=chd10yr);

%PlotLogits(indata=analyfile,
           numgrp=10,
           indepvar=ldl,
           depvar=chd10yr);



ods select statistics ttests;
proc ttest data= analyfile;
class chd10yr;
var chol sbp bmi age hdl ldl;
run;
```

```
%macro logreg(test=);
ods select fitstatistics parameterestimates;
ods select Association;
proc logistic data=tmp;
model chd10yr(event="1")=&test;
run;
%mend logreg;
%logreg(test =age);
%logreg(test =sbp);
%logreg(test =chol);
%logreg(test =diab);
%logreg(test =currsmok);
%logreg(test =bmi);
%logreg(test =male);
%logreg(test =hdl);
%logreg(test =ldl);

proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age sbp chol diab currsmok bmi male hdl ldl
        /clparm=both clodds=both;
run;
```

```
ods select globaltests fitstatistics;
proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age sbp chol diab currsmok bmi male hdl ldl;
run;
ods select globaltests fitstatistics;
proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age sbp chol diab currsmok male hdl ldl;
run;

ods select globaltests fitstatistics;
proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age sbp chol diab currsmok bmi male hdl ldl;
run;
ods select globaltests fitstatistics;
proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age sbp chol diab currsmok bmi male hdl;
run;


proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age sbp chol diab currsmok male hdl
        /lackfit;
run;
proc logistic  data=analyfile plots=none;
```

```
model chd10yr(event="1")=age sbp chol diab currsmok male hdl ldl
          /lackfit;
run;
proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age sbp chol diab bmi currsmok male hdl
          /lackfit;
run;



proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age sbp chol diab currsmok male hdl ldl
          /lackfit;
run;

proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age sbp chol diab currsmok bmi male hdl
          /lackfit;
run;

proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age sbp chol diab currsmok male hdl
          /lackfit;
run;
```

```
%macro plotloess(outsm= ,indata=,vars=);

proc loess data=&indata plots=none;

  model chd10yr=&vars/smooth=.25 .5 .75 1 1.25 1.5;

  output out=&outsm predicted=phat;

run;

proc freq data=&outsm;

 tables smoothingparameter;

run;

proc sort data=&outsm;

by smoothingparameter &vars;

run;

proc sgplot data=&outsm;

series x=&vars y=phat/group=smoothingparameter

lineattrs=(thickness=3);

run;


%mend plotloess;


%plotloess(outsm=smooth1,indata=analyfile, vars=age);

%plotloess(outsm=smooth2,indata=analyfile, vars=sbp);

%plotloess(outsm=smooth3,indata=analyfile, vars=chol);

%plotloess(outsm=smooth4,indata=analyfile, vars=bmi);

%plotloess(outsm=smooth5,indata=analyfile, vars=hdl);
```

```
%plotloess(outsm=smooth6,indata=analyfile, vars=ldl);

%macro smoothlog(outsm= ,indata=,vars=);
proc loess data=&indata;
  model chd10yr=&vars/smooth=.25 .5 .75 1 1.25 1.5;
  output out=&outsm predicted=phat;
run;
proc sort data=&outsm;
by smoothingparameter &vars;
run;
data &outsm;
set &outsm;
where 0<phat<1;
logit=log(phat/(1-phat));
run;
proc sgplot data=&outsm;
series x=&vars y=logit/group=smoothingparameter
lineattrs=(thickness=3);
run;

%mend smoothlog;


%smoothlog(outsm=smooth11,indata=analyfile, vars=age);
```

```
%smoothlog(outsm=smooth22,indata=analyfile, vars=sbp);
%smoothlog(outsm=smooth33,indata=analyfile, vars=chol);
%smoothlog(outsm=smooth44,indata=analyfile, vars=bmi);
%smoothlog(outsm=smooth55,indata=analyfile, vars=hdl);
%smoothlog(outsm=smooth66,indata=analyfile, vars=ldl);

ods output LackFitPartition =partition;
proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age age*age sbp chol bmi diab currsmok male hdl ldl
        /lackfit;
run;

proc sgplot data=partition;
series x=group y=eventsobserved /markers;
series x=group y=eventsexpected /markers;
run;

ods output LackFitPartition =partition;
proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age age*age sbp chol diab currsmok male hdl ldl
        /lackfit;
run;

/*add male*diab*/
```

```
proc logistic  data=analyfile plots=none;

model chd10yr(event="1")=age age*age sbp chol bmi diab currsmok male
male*diab hdl ldl

        /lackfit;

run;


/*add male*bmi*/

proc logistic  data=analyfile plots=none;

model chd10yr(event="1")=age age*age sbp chol bmi male*bmi diab currsmok
male hdl ldl

        /lackfit;

run;


/*age*sbp*/

proc logistic  data=analyfile plots=none;

model chd10yr(event="1")=age age*age age*sbp sbp chol bmi diab currsmok
male male*diab hdl ldl

        /lackfit;

run;

/*age*hdl*/

proc logistic  data=analyfile plots=none;

model chd10yr(event="1")=age age*age age*hdl sbp chol bmi diab currsmok
male male*diab hdl ldl

        /lackfit;

run;
```

```
/*delete bmi*/

proc logistic  data=analyfile plots=none;

model chd10yr(event="1")=age age*age age*hdl sbp chol diab currsmok male
male*diab hdl ldl

        /lackfit;

run;
```

```
/*delete age*hdl*/

proc logistic  data=analyfile plots=none;

model chd10yr(event="1")=age age*age sbp bmi chol diab currsmok male
male*diab hdl ldl

        /lackfit;

run;
```

```
/*delete ldl*/

proc logistic  data=analyfile plots=none;

model chd10yr(event="1")=age age*age age*hdl sbp bmi chol diab currsmok
male male*diab hdl

        /lackfit;

run;
```

```
/*delete diab*male*/
proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age age*age age*hdl sbp bmi chol diab currsmok
male hdl ldl
            /lackfit;
run;


/*delete age*hdl and ldl*/
proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age age*age sbp bmi chol diab currsmok male
male*diab hdl
            /lackfit;
run;
/*delete bmi*/
proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age age*age age*hdl sbp chol male*diab diab
currsmok male hdl ldl
            /lackfit;
run;
/*delete ldl*/
proc logistic  data=analyfile plots=none;
model chd10yr(event="1")=age age*age age*hdl sbp chol bmi male*diab diab
currsmok male hdl
            /lackfit;
run;
```

```
/*delete bmi and ldl*/

proc logistic  data=analyfile plots=none;

model chd10yr(event="1")=age age*age age*hdl sbp chol male*diab diab
currsmok male hdl

            /lackfit;

run;



proc logistic  data=analyfile plots=none;

model chd10yr(event="1")=age age*age age*hdl sbp chol bmi male*diab diab
currsmok male hdl ldl

            /lackfit;

run;
```