Portfolio of demonstrated skills

for the

Certificate in SAS Programming and Data Analysis

Ibrahim A. Weaver

STA 5066
Data Management Analysis with SAS

Assignment with output from the course:

Code:

```
libname prg '/home/u59231693/my_shared_file_links/jhshows0/STA5066';

proc contents data=prg.discount2016;
proc print data=prg.discount2016(obs=7);
run;

data work.extended;
set prg.discount2016;
where Start_Date = '01DEC2016'd;
Promotion = 'Holidays Bonus';
drop Unit_Sales_Price;
Season ='Winter';
output;
Start_Date ='01JUL2017'd;
End_Date= '31JUL2017'd;
Season ='Summer';
output;
run;

proc print data=work.extended;
run;
```

Output:

The CONTENTS Procedure

Data Set Name	PRG.DISCOUNT2016	**Observations**	697
Member Type	DATA	**Variables**	5
Engine	V9	**Indexes**	0
Created	07/28/2016 09:36:18	**Observation Length**	40
Last Modified	07/28/2016 09:36:18	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	1632
Obs in First Data Page	697
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	/home/u59231693/my_shared_file_links/jhshows0/STA5066/discount2016.sas7bdat
Release Created	9.0401M3
Host Created	X64_8PRO
Inode Number	1635242012
Access Permission	rw-r--r--
Owner Name	jhshows0
File Size	128KB
File Size (bytes)	131072

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Label
5	Discount	Num	8	PERCENT.	Discount as Percent of Normal Retail Sales Price
3	End_Date	Num	8	DATE9.	End Date
1	Product_ID	Num	8	12.	Product ID
2	Start_Date	Num	8	DATE9.	Start Date
4	Unit_Sales_Price	Num	8	DOLLAR13.2	Discount Retail Sales Price per Unit

Obs	Product_ID	Start_Date	End_Date	Unit_Sales_Price	Discount
1	210100100027	01MAY2016	31MAY2016	$17.99	70%
2	210100100030	01AUG2016	31AUG2016	$32.99	70%
3	210100100033	01AUG2016	31AUG2016	$161.99	70%
4	210100100034	01AUG2016	31AUG2016	$187.99	70%
5	210100100035	01MAY2016	31MAY2016	$172.99	70%
6	210100100038	01JUL2016	31JUL2016	$59.99	60%
7	210100100039	01JUN2016	31AUG2016	$21.99	70%

Obs	Product_ID	Start_Date	End_Date	Discount	Promotion	Season
1	210200100007	01DEC2016	31DEC2016	50%	Holidays Bonus	Winter
2	210200100007	01JUL2017	31JUL2017	50%	Holidays Bonus	Summer
3	210200300013	01DEC2016	31DEC2016	50%	Holidays Bonus	Winter
4	210200300013	01JUL2017	31JUL2017	50%	Holidays Bonus	Summer
5	210200300025	01DEC2016	31DEC2016	50%	Holidays Bonus	Winter
6	210200300025	01JUL2017	31JUL2017	50%	Holidays Bonus	Summer
7	210200300032	01DEC2016	31DEC2016	50%	Holidays Bonus	Winter
8	210200300032	01JUL2017	31JUL2017	50%	Holidays Bonus	Summer
9	210200300061	01DEC2016	31DEC2016	50%	Holidays Bonus	Winter
10	210200300061	01JUL2017	31JUL2017	50%	Holidays Bonus	Summer
11	210200400002	01DEC2016	31DEC2016	50%	Holidays Bonus	Winter
12	210200400002	01JUL2017	31JUL2017	50%	Holidays Bonus	Summer
13	210200400039	01DEC2016	31DEC2016	50%	Holidays Bonus	Winter
14	210200400039	01JUL2017	31JUL2017	50%	Holidays Bonus	Summer
15	210200400092	01DEC2016	31DEC2016	50%	Holidays Bonus	Winter

The output goes on to 332 observations, This was our first SAS course and it focused mainly on teaching us the basics (formats, syntax, the different “proc” functions, etc..).

```
libname q5 '/home/u59231693/my_shared_file_links/jhshows0/STA5066';

proc format;
   value $gender  'F' = 'Female'
                  'M' = 'Male'
                  other= 'Invalid Code';
run;

proc format;
   value salrange  20000-99999 = 'Below $100,000'
                  100000-499999 = '$100,000 or more'
                  . = 'Missing';
                  other= 'Invalid Code'
run;

proc print data=q5.nonsales (obs=10);
   var Employee_ID Job_Title Salary Gender;
   title1 'Distribution of Salary and Gender Values';
   title2 'for Non-Sales Employees';
   format Gender $gender.;
   format Salary salrange.;
run;
```

The output:

Distribution of Salary and Gender Values
for Non-Sales Employees

Obs	Employee_ID	Job_Title	Salary	Gender
1	120101	Director	$100,000 or more	Male
2	120104	Administration Manager	Below $100,000	Female
3	120105	Secretary I	Below $100,000	Female
4	120106	Office Assistant II	Missing	Male
5	120107	Office Assistant III	Below $100,000	Female
6	120108	Warehouse Assistant II	Below $100,000	Female
7	120108	Warehouse Assistant I	Below $100,000	Female
8	120110	Warehouse Assistant III	Below $100,000	Male
9	120111	Security Guard II	Below $100,000	Male
10	120112		Below $100,000	Female

Overall Average:'

STA 5067
Advanced Data Management and Analysis with SAS

```
%let path= /courses/d649d56dba27fe300/STA5067/SAS Data;
libname orion "&path/orion";
proc sql ;
select Employee_ID, Job_Title,
case
             when scan(Job_Title,-1,' ')='Manager'
                  then 'Manager'
             when scan(Job_Title,-1,' ')='Director'
                  then 'Director'
             when scan(Job_Title,-1,' ')='Officer'
                 then 'Executive'
             when scan(Job_Title,-1,' ')='President'
                 then 'Executive'
          end as Level,
Salary,
case
     When scan(Job_Title,-1,' ')='Manager' and Salary <
52000 then 'Low'
     When scan(Job_Title,-1,' ')='Manager' and Salary
between 52000 and 72000 then 'Medium'
     When scan(Job_Title,-1,' ')='Manager' and Salary >
72000 then 'High'
     When scan(Job_Title,-1,' ')='Director' and Salary <
108000 then 'Low'
     When scan(Job_Title,-1,' ')='Director' and Salary
between 108000 and 135000 then 'Medium'
     When scan(Job_Title,-1,' ')='Director' and Salary >
135000 then 'High'
     When scan(Job_Title,-1,' ')='Officer'or 'President' and
Salary < 240000 then 'Low'
     When scan(Job_Title,-1,' ')='Officer'or 'President' and
Salary between 240000 and 300000 then 'Medium'
     When scan(Job_Title,-1,' ')='Officer'or 'President' and
Salary > 300000 then 'High'
     end as Salary_Range
from orion.Staff
;
quit;
```

Output:

Distribution of Salary and Gender Values for Non-Sales Employees

Employee ID	Employee Job Title	Level	Employee Annual Salary	Salary_Range
120101	Director	Director	$163,040	High
120102	Sales Manager	Manager	$108,255	High
120103	Sales Manager	Manager	$87,975	High
120104	Administration Manager	Manager	$46,230	Low
120105	Secretary I		$27,110	Low
120106	Office Assistant II		$26,960	Low
120107	Office Assistant III		$30,475	Low
120108	Warehouse Assistant II		$27,660	Low
120109	Warehouse Assistant I		$26,495	Low
120110	Warehouse Assistant III		$28,615	Low
120111	Security Guard II		$26,895	Low
120112	Security Guard I		$26,550	Low
120113	Security Guard II		$26,870	Low
120114	Security Manager	Manager	$31,285	Low
120115	Service Assistant I		$26,500	Low
120116	Service Assistant II		$29,250	Low
120117	Cabinet Maker III		$31,670	Low
120118	Cabinet Maker II		$28,090	Low
120119	Electrician IV		$30,255	Low
120120	Electrician II		$27,645	Low
120121	Sales Rep. II		$26,600	Low
120122	Sales Rep. II		$27,475	Low

*The output goes on for a few more pages

```
%let path= /courses/d649d56dba27fe300/STA5067/SAS Data;
libname orion "&path/orion";


proc means data=orion.order_fact nway noprint;
   var Total_Retail_Price;
   class Customer_ID;
   output out=customer_sum sum=CustTotalPurchase;
run;
proc sort data=customer_sum;
   by descending CustTotalPurchase;
run;


data _null_;
set customer_sum (obs=1);
call symputx('top', Customer_ID);
run;

proc print data=orion.orders noobs;
where Customer_ID =&top;
var Order_ID Order_Type Order_Date Delivery_Date;
title "Orders for Customer &top - Orion's Top Customer";
run;


data _null_;
set customer_sum (obs=1);
call symputx('top', Customer_ID);
run;

data _null_;
set orion.customer_dim;
where Customer_ID = &top;
call symputx('topname', Customer_Name);
run;

proc print data=orion.orders noobs;
where Customer_ID =&top;
var Order_ID Order_Type Order_Date Delivery_Date;
title "Orders for Customer &topname - Orion's Top Customer”;
```

Output:

Orders for Customer 16 - Orion's Top Customer

Order_ID	Order_Type	Order_Date	Delivery_Date
1230450371	2	24MAR2003	26MAR2003
1231305521	2	27AUG2003	04SEP2003
1231305531	2	27AUG2003	29AUG2003
1234538390	2	12JAN2005	14JAN2005
1234588648	2	17JAN2005	19JAN2005
1234659163	2	24JAN2005	26JAN2005
1234972570	2	24FEB2005	26FEB2005
1235591214	2	25APR2005	27APR2005
1235611754	2	27APR2005	29APR2005
1235744141	2	10MAY2005	12MAY2005
1236128456	2	17JUN2005	19JUN2005
1239713046	2	19JUL2006	20JUL2006
1239932984	2	16AUG2006	17AUG2006
1240599396	2	07NOV2006	08NOV2006
1240961599	2	22DEC2006	23DEC2006

Overall Average:'

```
%let path= /courses/d649d56dba27fe300/STA5067/SAS Data;
libname orion "&path/orion";
proc sql;
title "Countries with more Female than Male Customers";
select Country 'Country',
sum(Gender="M") as M "Male Customers",
sum(Gender="F") as F "Female Customers"
from orion.Customer
group by Country
having F gt M
order by F desc;
```

Output:

Countries with more Female than Male Customers

Country	Male Customers	Female Customers
CA	7	8
ZA	1	3

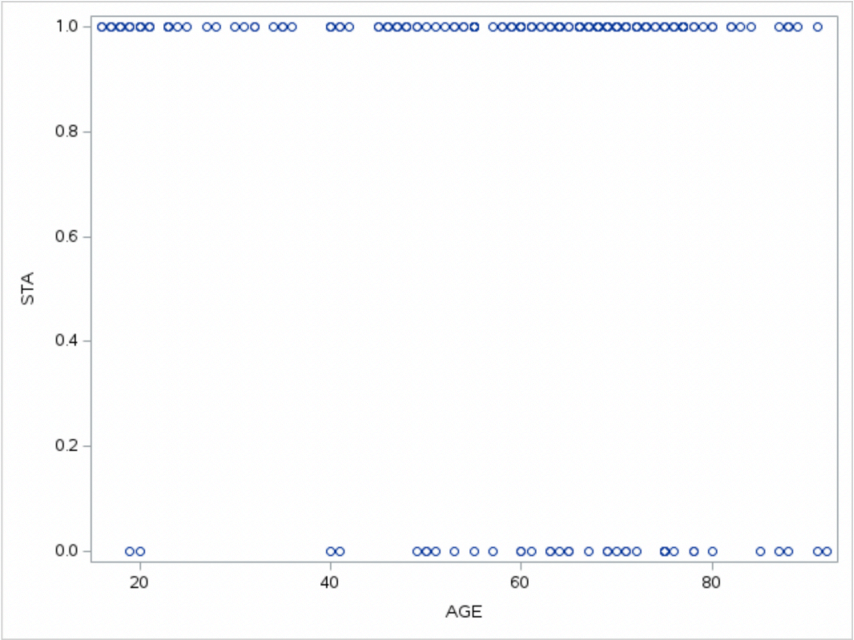
STA 5238
Applied Logistic Regression

Our first assignment for this course:

```
filename rawdata '/home/u59231693/my_shared_file_links/huffer/
5238/icu.txt';

data icu;
infile rawdata;
input ID STA AGE GENDER;
if 15 <= age <= 24 then do ; grp = 1 ; midpt = 19.5 ; end ;
else if 25 <= age <= 34 then do ; grp = 2 ; midpt = 29.5 ; end ;
else if 35 <= age <= 44 then do ; grp = 3 ; midpt = 39.5 ; end ;
else if 45 <= age <= 54 then do ; grp = 4 ; midpt = 49.5 ; end ;
else if 55 <= age <= 64 then do ; grp = 5 ; midpt = 59.5 ; end ;
else if 65 <= age <= 74 then do ; grp = 6 ; midpt = 69.5 ; end ;
else if 75 <= age <= 84 then do ; grp = 7 ; midpt = 79.5 ; end ;
else if 85 <= age <= 94 then do ; grp = 8 ; midpt = 89.5 ; end ;
run;

/*Q1A problem 1(b)*/
proc sgplot data=icu;
scatter y=STA x=age;
run;

/*Q1A problem 1(c)*/
proc means data=icu mean ;
var sta midpt ;
class grp ;
types grp ;
output out=grpmeans mean= mean_sta age ;
run;

proc sgplot data=grpmeans;
scatter y=mean_sta x=age;
run;

/*Q1b 1c*/
proc logistic data=icu;
model sta(event="1") = age;
run;

/*Q1d*/
proc logistic data=icu ;
model sta(event="1") = age / clparm=both covb ;
output out=Prediction p=pred lower=L upper=U xbeta=etahat
stdxbeta=SEetahat ;
run;
```

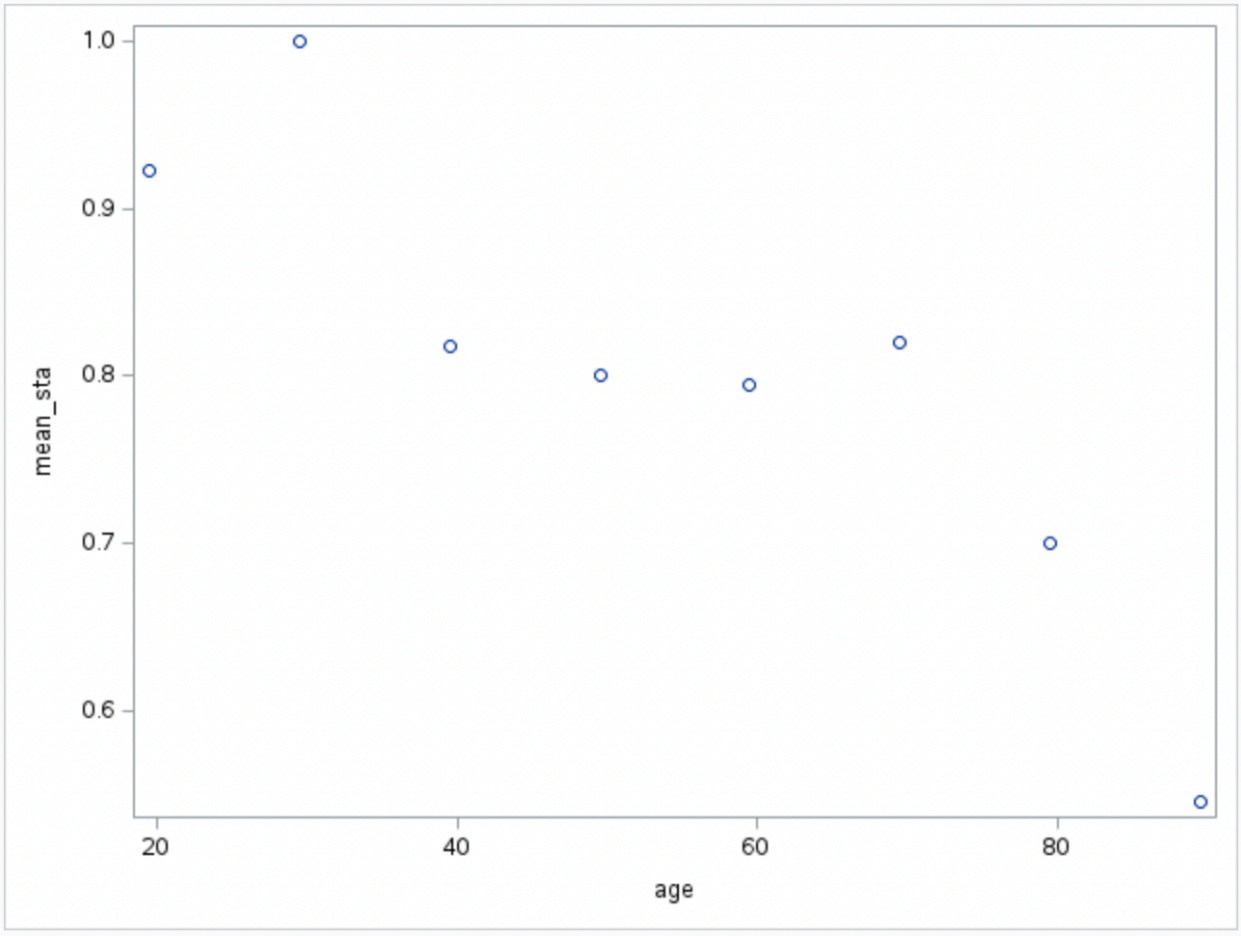
```
/*Q1e*/
proc print data=Prediction (obs=10);
run;

proc sgplot data=Prediction;
scatter y=pred x=age;
run;

data combined ;
set Prediction grpmeans ;
run ;

proc sgplot data=combined ;
scatter x=age y=sta ;
scatter x=age y=pred ;
scatter x=age y=mean_sta / markerattrs=(symbol=plus color=green
size=20) ;
run ;

/*Q1(f)*/
proc sql;
   insert into work.icu
        set age = 86,
            sta=1;
quit;
proc logistic data=icu ;
model sta(event="1") = age / clparm=both covb ;
output out=Prediction p=pred lower=L upper=U xbeta=etahat
stdxbeta=SEetahat ;
run;

proc print data=Prediction;
run;
```

Output:



The MEANS Procedure

grp	N Obs	Variable	Mean
1	26	STA	0.9230769
		midpt	19.5000000
2	8	STA	1.0000000
		midpt	29.5000000
3	11	STA	0.8181818
		midpt	39.5000000
4	25	STA	0.8000000
		midpt	49.5000000
5	39	STA	0.7948718
		midpt	59.5000000
6	50	STA	0.8200000
		midpt	69.5000000
7	30	STA	0.7000000
		midpt	79.5000000
8	11	STA	0.5454545
		midpt	89.5000000



The LOGISTIC Procedure

Model Information	
Data Set	WORK.ICU
Response Variable	STA
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	200
Number of Observations Used	200

Response Profile		
Ordered Value	STA	Total Frequency
1	0	40
2	1	160

Probability modeled is STA=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	202.161	196.306
SC	205.459	202.903
-2 Log L	200.161	192.306

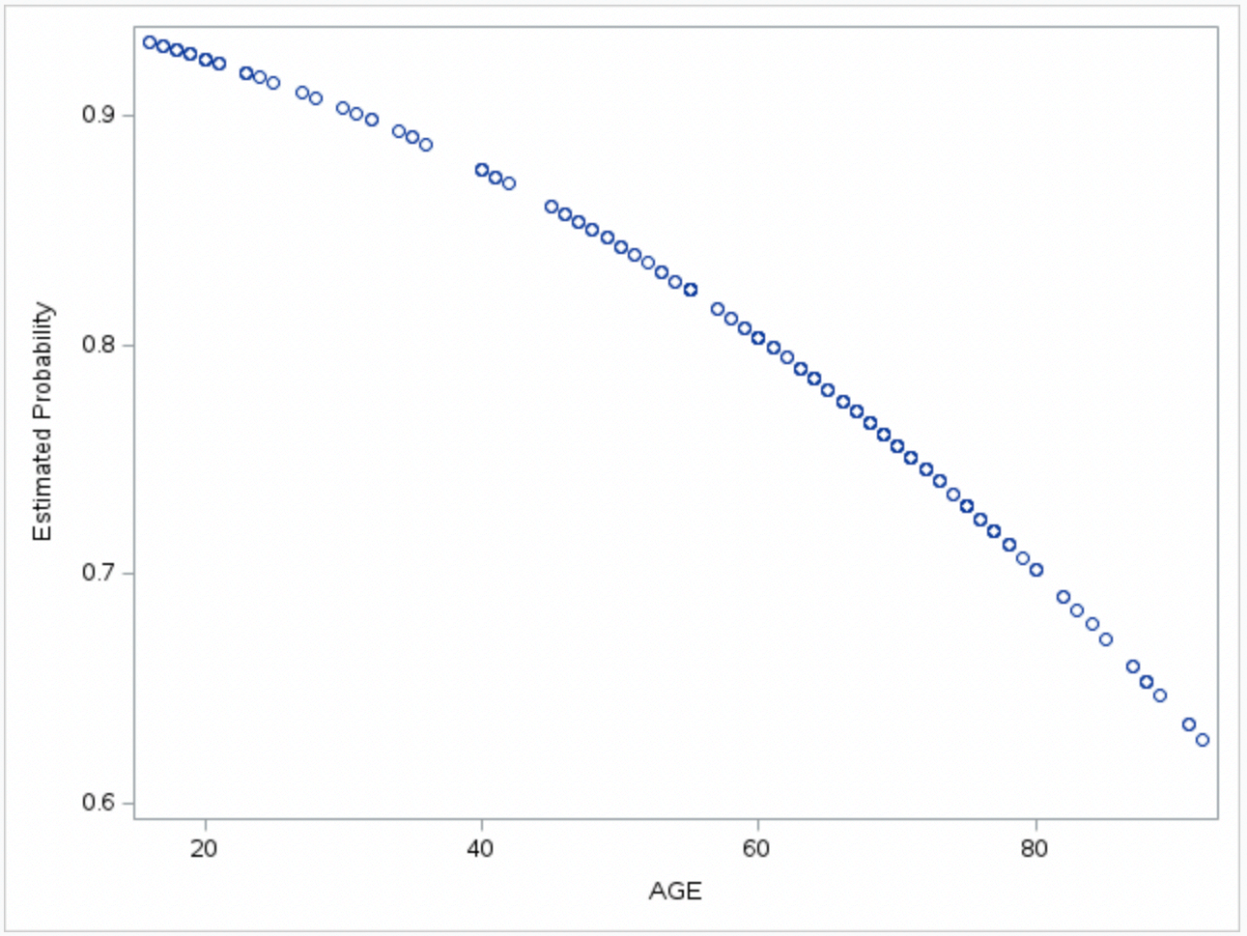
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.8546	1	0.0051
Score	7.1789	1	0.0074
Wald	6.7963	1	0.0091

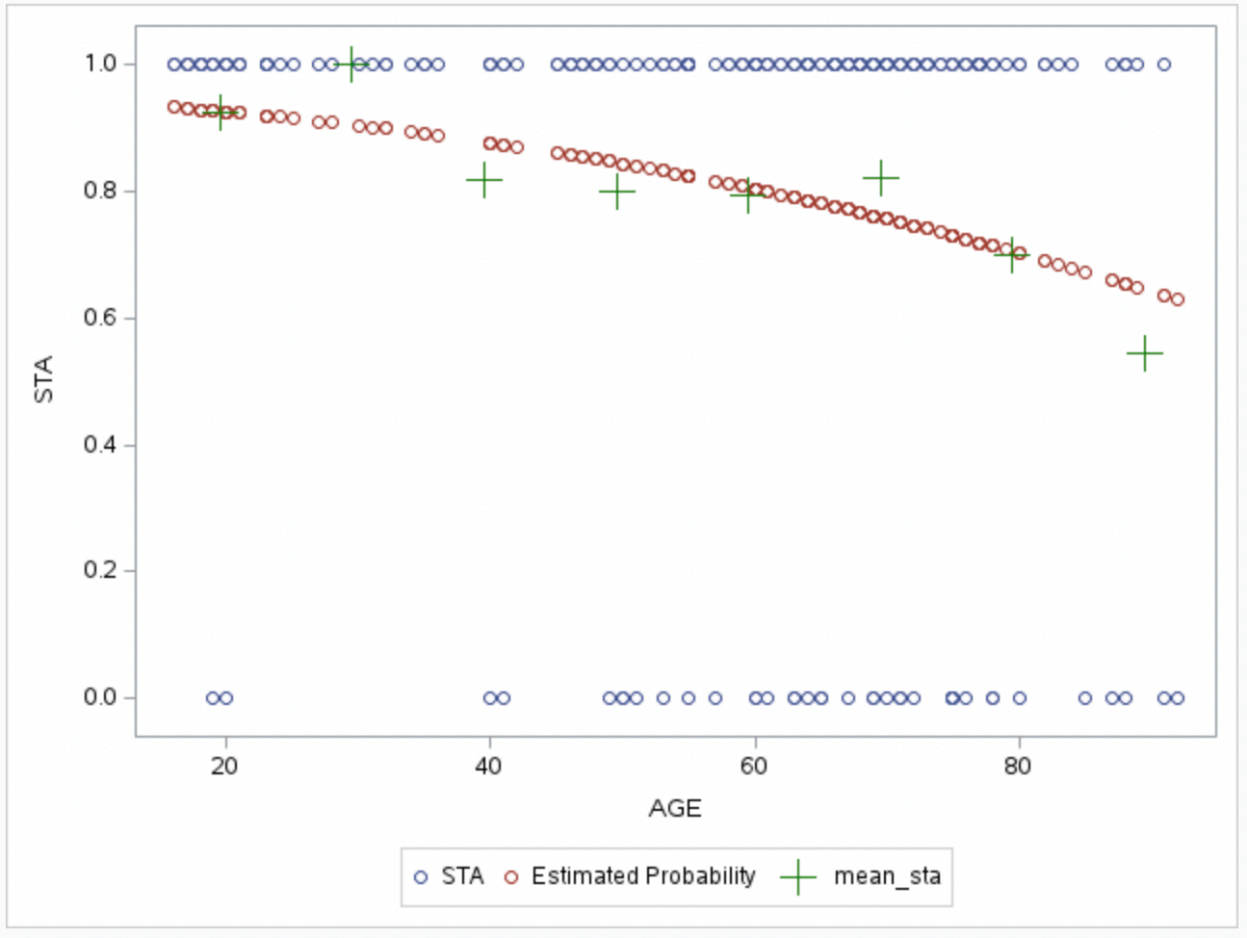
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.0584	0.6961	19.3036	<.0001
AGE	1	-0.0275	0.0106	6.7963	0.0091

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
AGE	0.973	0.953	0.993

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	62.1	Somers' D	0.260
Percent Discordant	36.1	Gamma	0.265
Percent Tied	1.8	Tau-a	0.084
Pairs	6400	c	0.630

Obs	ID	STA	AGE	GENDER	grp	midpt	_LEVEL_	etahat	SEetahat	pred	L	U
1	4	0	87	1	8	89.5	1	0.66232	0.30549	0.65978	0.51589	0.77921
2	8	1	27	1	2	29.5	1	2.31480	0.42701	0.91010	0.81425	0.95898
3	12	1	59	0	5	59.5	1	1.43348	0.18671	0.80744	0.74413	0.85808
4	14	1	77	0	7	79.5	1	0.93774	0.22881	0.71864	0.61994	0.79998
5	27	0	76	1	7	79.5	1	0.96528	0.22245	0.72418	0.62932	0.80239
6	28	1	54	0	4	49.5	1	1.57119	0.20697	0.82795	0.76234	0.87834
7	32	1	87	1	8	89.5	1	0.66232	0.30549	0.65978	0.51589	0.77921
8	38	1	69	0	6	69.5	1	1.15807	0.18881	0.76098	0.68740	0.82173
9	40	1	63	0	5	59.5	1	1.32332	0.18028	0.78973	0.72512	0.84246
10	41	1	30	1	2	29.5	1	2.23218	0.39850	0.90310	0.81017	0.95317


Estimated Probability
0.9
0.8
0.7
0.6
20
40
60
80
AGE


STA
1.0
0.8
0.6
0.4
0.2
0.0
20
40
60
80
AGE
STA
Estimated Probability
mean_sta

Obs	ID	STA	AGE	GENDER	grp	midpt	_LEVEL_	etahat	SEetahat	pred	L	U
1	4	0	87	1	8	89.5	1	0.69211	0.30319	0.66644	0.52445	0.78353
2	8	1	27	1	2	29.5	1	2.29558	0.42383	0.90851	0.81228	0.95796
3	12	1	59	0	5	59.5	1	1.44040	0.18640	0.80852	0.74556	0.85885
4	14	1	77	0	7	79.5	1	0.95935	0.22758	0.72299	0.62558	0.80304
5	27	0	76	1	7	79.5	1	0.98608	0.22132	0.72831	0.63467	0.80532
6	28	1	54	0	4	49.5	1	1.57402	0.20640	0.82836	0.76305	0.87853
7	32	1	87	1	8	89.5	1	0.69211	0.30319	0.66644	0.52445	0.78353
8	38	1	69	0	6	69.5	1	1.17315	0.18830	0.76371	0.69085	0.82378
9	40	1	63	0	5	59.5	1	1.33350	0.18000	0.79142	0.72725	0.84374
10	41	1	30	1	2	29.5	1	2.21541	0.39563	0.90162	0.80845	0.95216
11	42	1	35	0	3	39.5	1	2.08179	0.34984	0.88912	0.80157	0.94089
12	47	0	78	0	7	79.5	1	0.93263	0.23414	0.71761	0.61627	0.80084
13	50	1	70	1	6	69.5	1	1.14643	0.19165	0.75886	0.68369	0.82084
14	51	1	55	1	5	59.5	1	1.54729	0.20147	0.82452	0.75995	0.87459
15	52	0	63	0	5	59.5	1	1.33350	0.18000	0.79142	0.72725	0.84374
16	53	1	48	0	4	49.5	1	1.73437	0.24343	0.84997	0.77855	0.90127
17	58	1	66	1	6	69.5	1	1.25332	0.18150	0.77787	0.71045	0.83328
18	61	1	61	1	5	59.5	1	1.38695	0.18203	0.80010	0.73695	0.85116
19	73	1	66	0	6	69.5	1	1.25332	0.18150	0.77787	0.71045	0.83328
20	75	1	52	0	4	49.5	1	1.62747	0.21743	0.83582	0.76876	0.88631
21	82	1	55	0	5	59.5	1	1.54729	0.20147	0.82452	0.75995	0.87459
22	84	1	59	0	5	59.5	1	1.44040	0.18640	0.80852	0.74556	0.85885
23	92	1	63	0	5	59.5	1	1.33350	0.18000	0.79142	0.72725	0.84374
24	96	1	72	0	6	69.5	1	1.09298	0.19985	0.74894	0.66847	0.81528
25	98	1	60	0	5	59.5	1	1.41367	0.18393	0.80434	0.74138	0.85497
26	100	1	78	0	7	79.5	1	0.93263	0.23414	0.71761	0.61627	0.80084
27	102	1	16	1	1	19.5	1	2.58955	0.53031	0.93019	0.82494	0.97414
28	111	1	62	0	5	59.5	1	1.36022	0.18072	0.79580	0.73224	0.84741
29	112	1	61	0	5	59.5	1	1.38695	0.18203	0.80010	0.73695	0.85116
30	127	0	19	0	1	19.5	1	2.50938	0.50089	0.92480	0.82166	0.97043
31	136	1	35	0	3	39.5	1	2.08179	0.34984	0.88912	0.80157	0.94089
32	137	1	74	1	6	69.5	1	1.03953	0.20982	0.73876	0.65210	0.81012

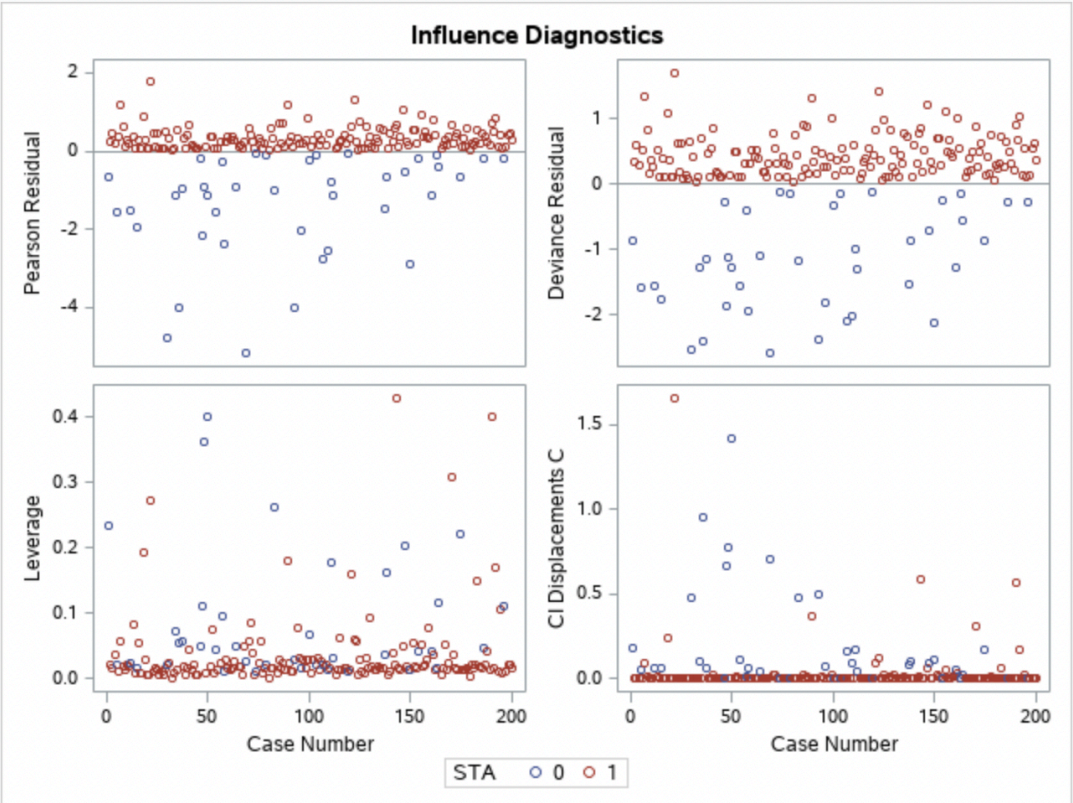
*Prints to approximately 200 observations

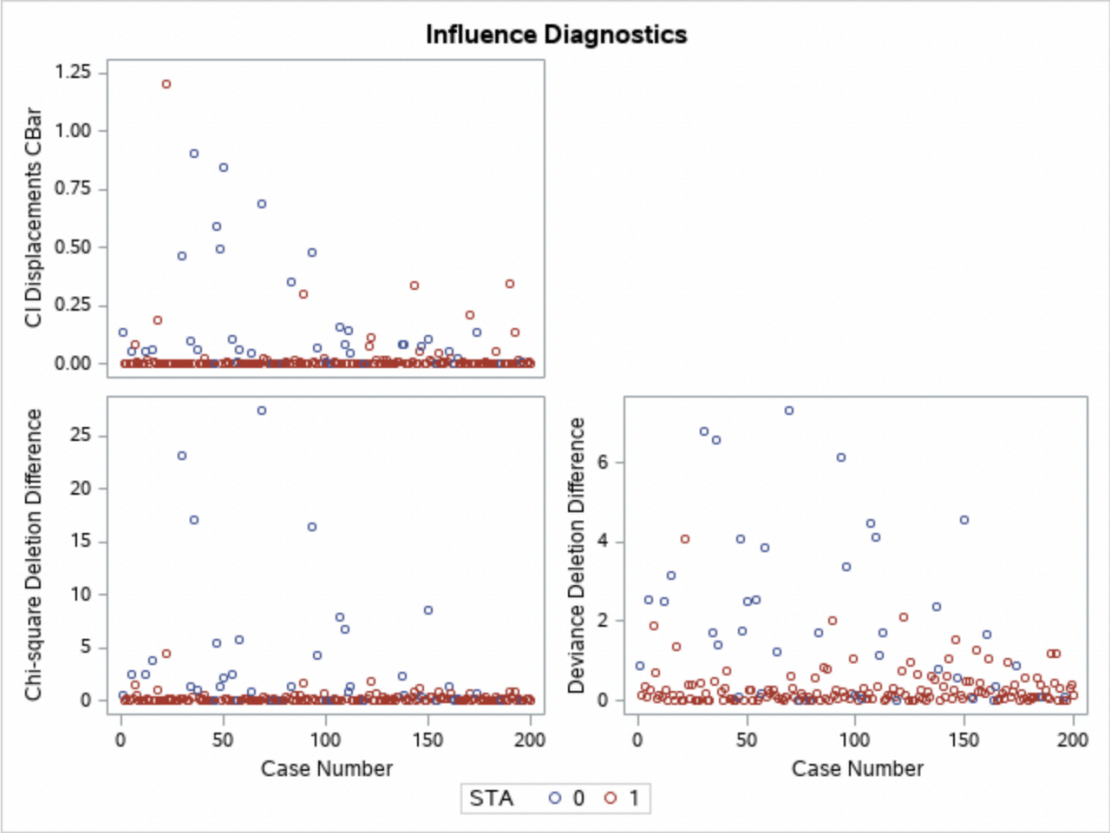
```
filename rawdata '/home/u59231693/my_shared_file_links/huffer/
5238/icu.txt';

data icu;
infile rawdata;
input ID STA AGE GENDER RACE SER CAN CRN INF CPR SYS HRA PRE
TYPE FRA PO2 PH PCO BIC CRE LOC;
nloc=0;
if LOC = 1 or LOC = 2 then nloc=1;
run;
ods graphics on;
proc logistic data=icu;
model sta(event="1") = age sys age*sys can type ph pco nloc/
influence;
output out=Prediction p=pred lower=L upper=U xbeta=etahat
stdxbeta=SEetahat ; run;
ods graphics off;
proc print data=Prediction;
run;
```

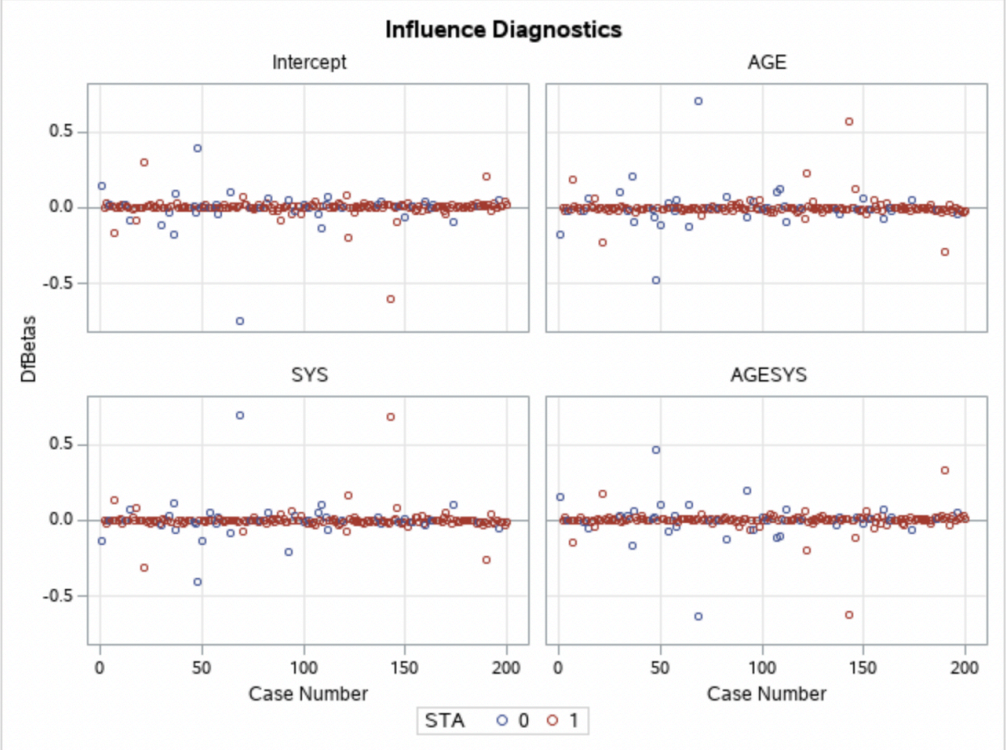
Output:

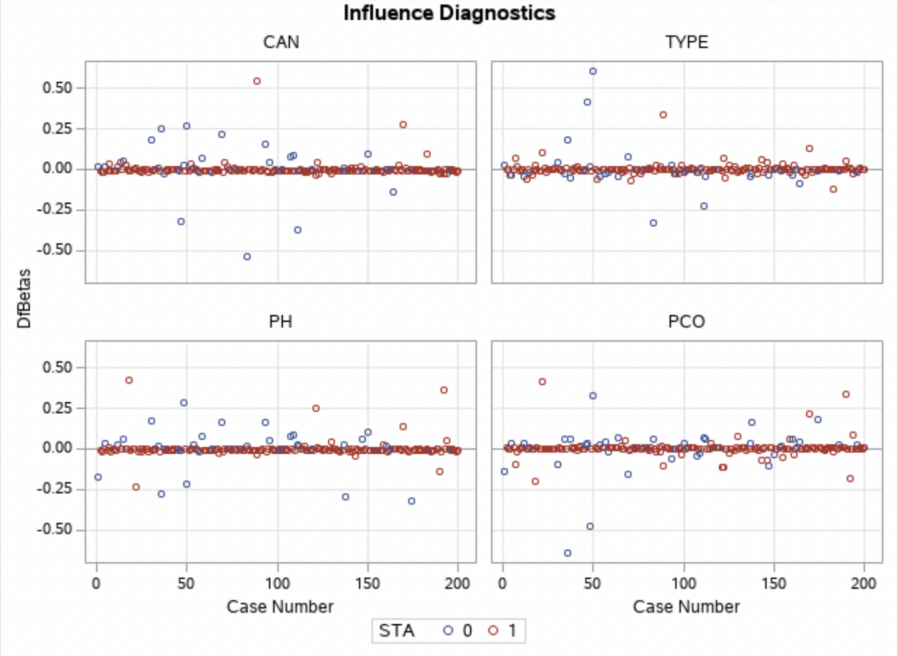
Model Fit Statistics

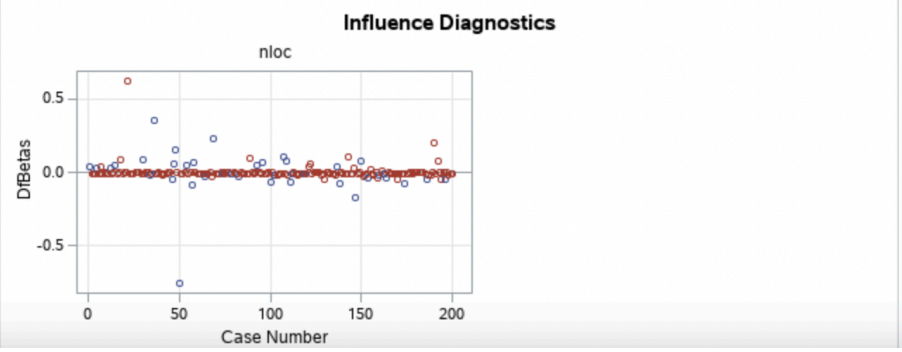
Criterion	Intercept Only	Intercept and Covariates
AIC	202.161	140.234
SC	205.459	169.919
-2 Log L	200.161	122.234


Influence Diagnostics
Pearson Residual
Deviance Residual
Leverage
CI Displacements C
Case Number
Case Number
STA 0 1


Influence Diagnostics
CI Displacements CBar
Chi-square Deletion Difference
Deviance Deletion Difference
Case Number
Case Number
STA 0 1


Influence Diagnostics
Intercept
AGE
SYS
AGESYS
DfBetas
0.5
0.0
-0.5
0
50
100
150
200
Case Number
STA 0 1


Influence Diagnostics
CAN
TYPE
PH
PCO
DfBetas
0.50
0.25
0.00
-0.25
-0.50
0
50
100
150
200
Case Number
STA 0 1


Influence Diagnostics
nloc
DfBetas
0.5
0.0
-0.5
0
50
100
150
200
Case Number

We can see from looking at the plots that there are a few cases with large residuals and deletion differences, in cases which fit the model poorly, and can have a large influence on estimated parameters.

Regression Diagnostics

Case Number	Covariates								Pearson Residual	Deviance Residual	Hat Matrix Diagonal	Intercept DfBeta	AGE DfBeta	SYS DfBeta	AGESYS DfBeta	CAN DfBeta	TYPE DfBeta	PH DfBeta	PCO DfBeta	nloc DfBeta	Confidence Interval Displacement C	Confidence Interval Displacement CBar	Delta Deviance	Delta Chi-Square
	AGE	SYS	AGE * SYS	CAN	TYPE	PH	PCO	nloc																
30	19.0000	140.0	2660.0	0	1.0000	0	0	0	-4.7641	-2.5161	0.0202	-0.1134	0.1017	-0.0355	0.0302	0.1804	0.0420	0.1754	-0.0937	0.0903	0.4767	0.4671	6.7977	23.1634
31	35.0000	150.0	5250.0	0	1.0000	0	0	0	0.3055	0.4224	0.0227	-0.00518	0.00485	0.0147	-0.0126	-0.0108	-0.00024	-0.0119	0.00357	-0.00452	0.00222	0.00217	0.1806	0.0955
32	74.0000	170.0	12580.0	0	0	0	1.0000	0	0.0198	0.0279	0.00107	0.000275	-0.00023	-0.00018	0.000228	-0.00033	-0.00046	-0.00023	0.000504	-0.00036	4.208E-7	4.203E-7	0.000782	0.000391
36	53.0000	148.0	7844.0	0	1.0000	1.0000	1.0000	0	-4.0230	-2.3850	0.0530	-0.1801	0.2058	0.1119	-0.1716	0.2533	0.1820	-0.2760	-0.6434	0.3518	0.9567	0.9060	6.5940	17.0907

We can see here that case 30 is an outlier, with its Pearson residual and deviance residual extremely high compared to other cases, along with its delta deviance and delta chi square.
Similarly 36 has very high values and residuals. So we can say these cases fit the model poorly, along with cases 69 and 93.

STA 5939
Introduction to Statistical Consulting

This is from my final project I completed in this course titled "Statistical Analysis of Student Exam Scores and the Influence of Different Factors", where I used logistic regression in SAS to determine what factors are the most influential on student exam scores.

Code used in my final project:

```
FILENAME REFFILE '/home/u59231693/sasuser.v94/
examsupdated2.csv';

PROC IMPORT DATAFILE=REFFILE
     DBMS=CSV
     OUT=WORK.IMPORT2;
     GETNAMES=YES;
RUN;


/*status 1 = pass 0 = fail, gender 1= male 0=female,
testprep 1 = completed 0 = none*/

/*testing correlation between independant variables*/

proc freq data=work.import2 order=formatted;

tables testprep*lunch / chisq;

run;

/* Exploring Data */
proc univariate data=WORK.IMPORT2;
     ods select Histogram;
     var status mathScore readingScore writingScore total;
     histogram status mathScore readingScore writingScore total;
run;

ods select CollinDiag CollinDiagNoInt;
 proc reg data=work.import2;
  model status = Race ParentsEdu TestPreptext Sex lunch / collin
collinoint;
  run; quit;
```

```
/*Splitting data into training and test data */
ods noproctitle;
ods graphics / imagemap=on;
data temp;
set work.import2;
n=ranuni(8);
proc sort data=temp;
  by n;
  data training testing;
   set temp nobs=nobs;
   if _n_<=.7*nobs then output training;
    else output testing;
   run;

PROC CONTENTS DATA=WORK.TESTING; RUN;
PROC CONTENTS DATA=WORK.TRAINING; RUN;



proc logistic data=WORK.TRAINING;
     class Race ParentsEdu lunch / param=glm;
     model status(event='1')=Race ParentsEdu gender testprep
lunch / link=logit
          selection=backward slstay=0.05 hierarchy=single
details technique=fisher;
run;


proc logistic data=WORK.TESTING;
     class Race ParentsEdu TestPreptext Sex lunch / param=glm;
     model status(event='1')=Race ParentsEdu TestPreptext Sex
lunch / link=logit lackfit risklimits iplots
          selection=backward slstay=0.05 hierarchy=single
details technique=fisher;
run;
```

Output:

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of testprep by lunch		
	lunch		
testprep	free/reduced	standard	Total
0	238 23.80 35.79 68.39	427 42.70 64.21 65.49	665 66.50
1	110 11.00 32.84 31.61	225 22.50 67.16 34.51	335 33.50
Total	348 34.80	652 65.20	1000 100.00

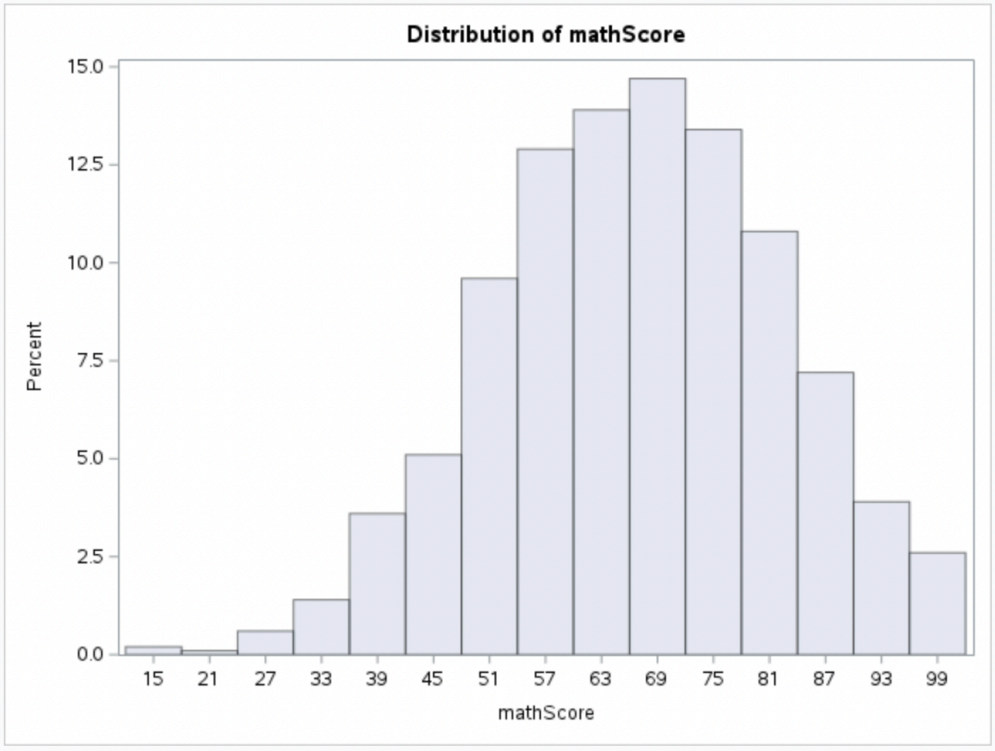
Statistics for Table of testprep by lunch

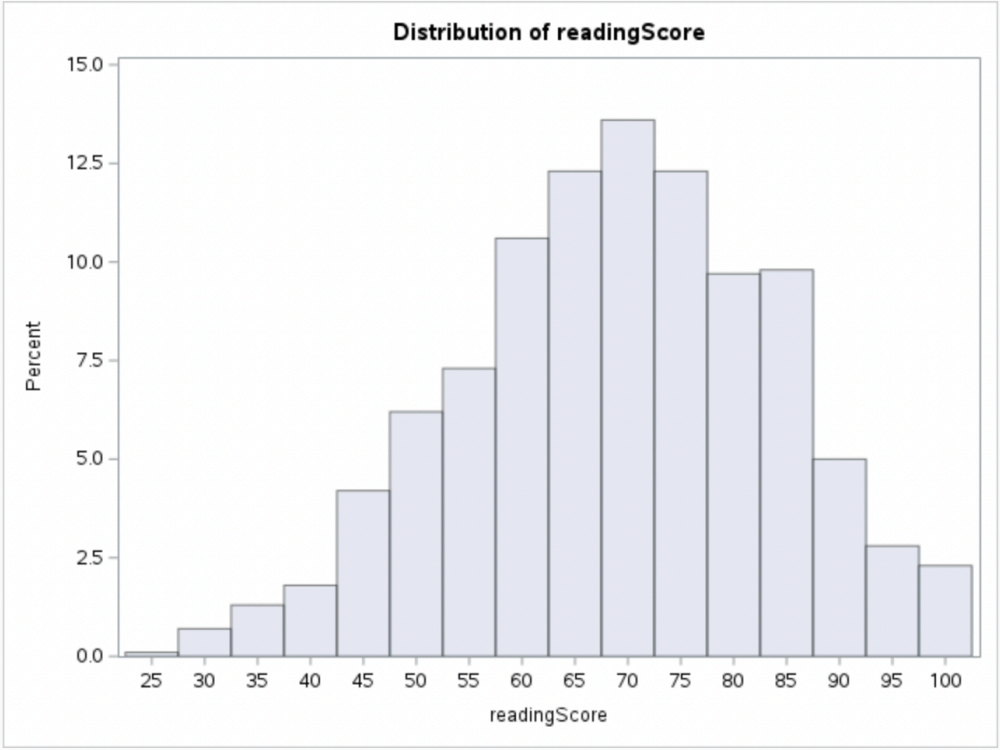
Statistic	DF	Value	Prob
Chi-Square	1	0.8566	0.3547
Likelihood Ratio Chi-Square	1	0.8606	0.3536
Continuity Adj. Chi-Square	1	0.7313	0.3925
Mantel-Haenszel Chi-Square	1	0.8557	0.3549
Phi Coefficient		0.0293	
Contingency Coefficient		0.0293	
Cramer's V		0.0293	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	238
Left-sided Pr <= F	0.8403
Right-sided Pr >= F	0.1964
Table Probability (P)	0.0367
Two-sided Pr <= P	0.3616

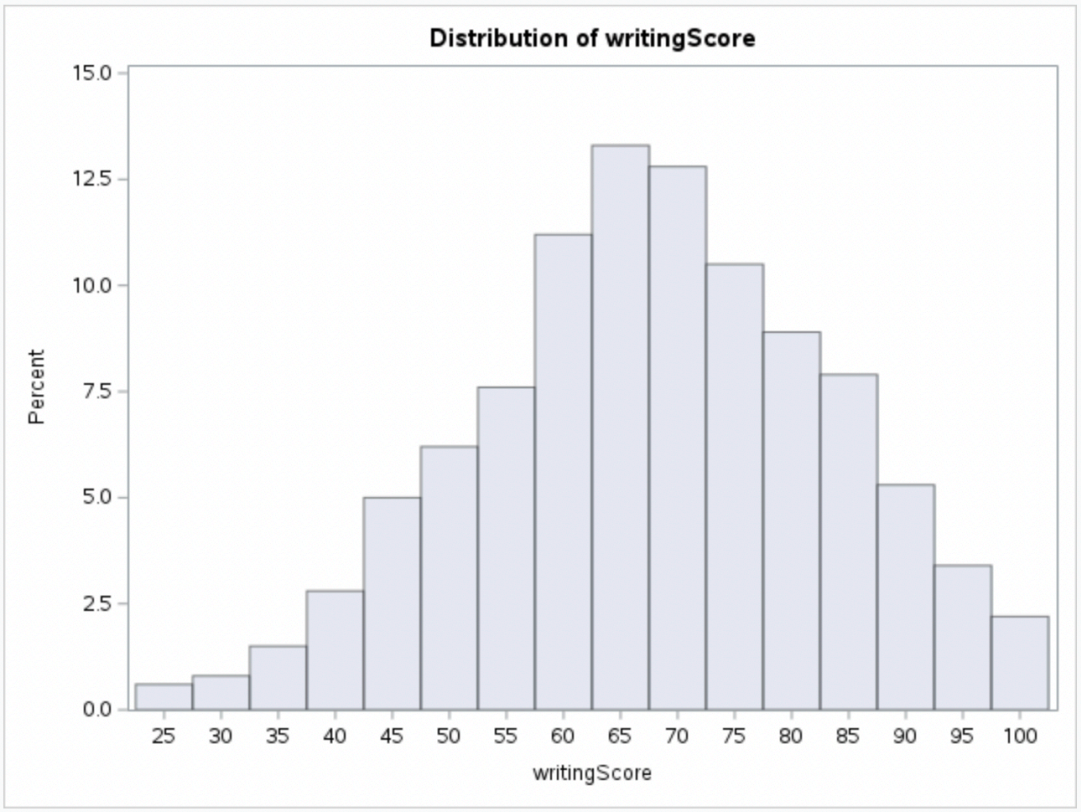
Sample Size = 1000

The UNIVARIATE Procedure

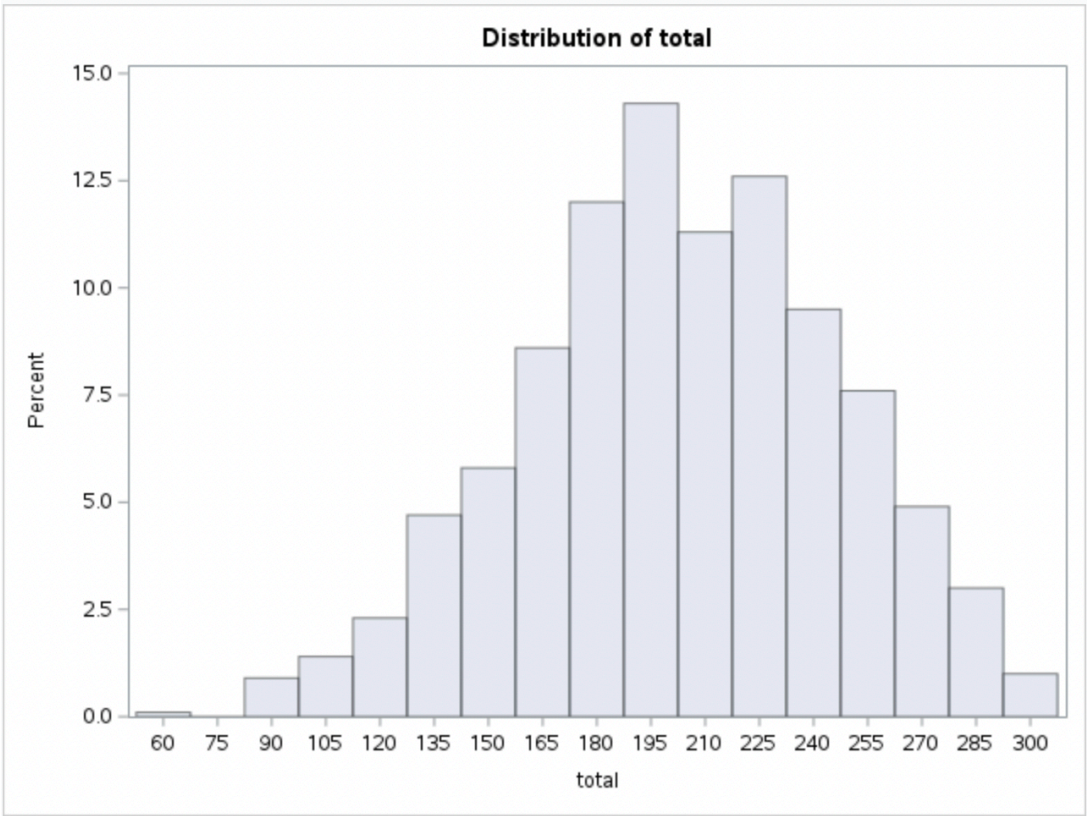




The UNIVARIATE Procedure



The UNIVARIATE Procedure



Model Information	
Data Set	WORK.TRAINING
Response Variable	status
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	700
Number of Observations Used	700

Response Profile		
Ordered Value	status	Total Frequency
1	0	206
2	1	494

Probability modeled is status='1'.

Backward Elimination Procedure

Class Level Information							
Class	Value	Design Variables					
Race	group A	1	0	0	0	0	
	group B	0	1	0	0	0	
	group C	0	0	1	0	0	
	group D	0	0	0	1	0	
	group E	0	0	0	0	1	
ParentsEdu	associate's degree	1	0	0	0	0	0
	bachelor's degree	0	1	0	0	0	0
	high school	0	0	1	0	0	0
	master's degree	0	0	0	1	0	0
	some college	0	0	0	0	1	0
	some high school	0	0	0	0	0	1
lunch	free/reduced	1	0				
	standard	0	1				

Step 0. The following effects were entered:

Intercept Race ParentsEdu gender testprep lunch

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	850.322	764.173
SC	854.873	823.337
-2 Log L	848.322	738.173

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	110.1495	12	<.0001
Score	102.6629	12	<.0001
Wald	87.8963	12	<.0001

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Race	4	18.6104	0.0009
ParentsEdu	5	19.3420	0.0017
gender	1	2.4041	0.1210
testprep	1	25.5135	<.0001
lunch	1	45.9702	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.4070	0.3692	14.5216	0.0001
Race	group A	1	-1.1148	0.4040	7.6125	0.0058
Race	group B	1	-0.9943	0.3410	8.5023	0.0035
Race	group C	1	-0.9999	0.3200	9.7666	0.0018
Race	group D	1	-0.2787	0.3405	0.6699	0.4131
Race	group E	0	0	.	.	.
ParentsEdu	associate's degree	1	1.0445	0.2888	13.0766	0.0003
ParentsEdu	bachelor's degree	1	1.0458	0.3517	8.8445	0.0029
ParentsEdu	high school	1	0.4272	0.2765	2.3871	0.1223
ParentsEdu	master's degree	1	1.0372	0.4239	5.9864	0.0144
ParentsEdu	some college	1	0.3295	0.2704	1.4847	0.2230
ParentsEdu	some high school	0	0	.	.	.
gender		1	-0.2859	0.1844	2.4041	0.1210
testprep		1	1.0629	0.2104	25.5135	<.0001
lunch	free/reduced	1	-1.2570	0.1854	45.9702	<.0001
lunch	standard	0	0	.	.	.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Race group A vs group E	0.328	0.149	0.724
Race group B vs group E	0.370	0.190	0.722
Race group C vs group E	0.368	0.197	0.689
Race group D vs group E	0.757	0.388	1.475
ParentsEdu associate's degree vs some high school	2.842	1.613	5.006
ParentsEdu bachelor's degree vs some high school	2.846	1.428	5.669
ParentsEdu high school vs some high school	1.533	0.892	2.636
ParentsEdu master's degree vs some high school	2.821	1.229	6.475
ParentsEdu some college vs some high school	1.390	0.818	2.362
gender	0.751	0.523	1.078
testprep	2.895	1.916	4.373
lunch free/reduced vs standard	0.285	0.198	0.409

Results

The first thing I do after making sure my test and train set are working properly is run my data set through a backwards selection elimination process. Backwards selection is a method for selecting the predictors (also known as independent variables or features or variables or even factors in this case) to include in a logistic regression model. It is a form of model selection, which is the process of choosing the appropriate set of predictors to use in a model.
In backwards selection, the process starts by fitting a logistic regression model with all the available predictors. Then, the least significant predictor is removed from the model, and the model is refit. This process is repeated until all the remaining predictors are significant at a predetermined level (e.g. $p < 0.05$).

```
proc logistic data=WORK.TRAINING;
    class Race ParentsEdu lunch / param=glm;
    model status(event='1')=Race ParentsEdu gender testprep lunch / link=logit
        selection=backward slstay=0.05 hierarchy=single details technique=fisher;
run;
```

This is the code we use to run the backwards selection, we've added all the variables as should be done in this case (so you can start with the full model), and as u could see the significance level we've chosen is 0.05, which is what is normally chosen in most cases. However this can change depending on the study.

Analysis of Effects Eligible for Removal			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Race	4	18.6104	0.0009
ParentsEdu	5	19.3420	0.0017
gender	1	2.4041	0.1210
testprep	1	25.5135	<.0001
lunch	1	45.9702	<.0001

Note: No (additional) effects met the 0.05 significance level for removal from the model.

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	gender	1	4	2.4041	0.1210

The output you get from running the backwards selection proc logistic actually outputs a lot of information that can be a little confusing at times. However, what is shown above is pretty much the only thing that is important to us at this step.

In the first table titled “Analysis of Effects Eligible for Removal” we are given all our variables along with p values. The only P-value above 0.05 is for “gender”. This tells us that gender is not statistically significant on the response variable (does not really effect a student passing or failing) based on our significance level of 0.05. However it is the only factor with a p- value above that number, so gender is the only variable removed. Now we want to re run it with our completed model, so we remove gender from the code and run it again.

```
55 proc logistic data=WORK.TRAINING;
56     class Race ParentsEdu lunch testprep / param=glm;
57     model status(event='1')=Race ParentsEdu testprep lunch ;
58 run;
```

We get the following from the results:

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Race	4	18.0807	0.0012
ParentsEdu	5	18.5870	0.0023
testprep	1	25.7921	<.0001
lunch	1	45.9573	<.0001

This tells us the model works and that all the variables are statistically significant, with testprep and lunch being more significant than race and parentsEdu.

After that runs and works, we use our test set to run it again and evaluate our model. Starting with our backwards elimination we get the same result and remove gender as a variable. Now running the full model on the test set we get:

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	370.197	321.080
SC	373.901	365.526
-2 Log L	368.197	297.080

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	71.1171	11	<.0001
Score	65.6137	11	<.0001
Wald	52.1508	11	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Race	4	10.5958	0.0315
ParentsEdu	5	21.7850	0.0006
testprep	1	9.1740	0.0025
lunch	1	23.4137	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.3835	0.5333	6.7292	0.0095
Race	group A	1	0.5096	0.9137	0.3110	0.5770
Race	group B	1	-0.6168	0.5112	1.4562	0.2275
Race	group C	1	-0.8378	0.4626	3.2794	0.0702
Race	group D	1	0.2102	0.5009	0.1761	0.6747
Race	group E	0	0	.	.	.
ParentsEdu	associate's degree	1	1.6207	0.4696	11.9102	0.0006
ParentsEdu	bachelor's degree	1	1.7777	0.6092	8.5142	0.0035
ParentsEdu	high school	1	1.5419	0.4475	11.8702	0.0006
ParentsEdu	master's degree	1	2.0502	0.6713	9.3268	0.0023
ParentsEdu	some college	1	1.0739	0.4074	6.9496	0.0084
ParentsEdu	some high school	0	0	.	.	.
testprep	0	1	-1.0409	0.3437	9.1740	0.0025
testprep	1	0	0	.	.	.
lunch	free/reduced	1	-1.4487	0.2994	23.4137	<.0001
lunch	standard	0	0	.	.	.

From the Type 3 Analysis of Effects table we can see that again all the variables that we kept after backwards elimination are statistically significant. However, we do see some difference from our training set, as in this one “lunch” is showing as the most statistically significant, followed by parents education, then testprep, and finally race. Race being very close to 0.05 (our significance level) and seeming to be a lot less significant than our other variables, although still statistically significant.