```
SAS Examples SASLR
==================

ssh abc@glue.umd.edu, tap sas913, sas
https://www.statlab.umd.edu/sasdoc/sashtml/onldoc.htm


a. Logistic regression.
b. Probit regression.
c. Compare SAS and S-Plus logistic regression.


Logistic Regression
====================
OPTION PS=45 LS=70;

DATA LOGISTIC;
INPUT ACCIDENT AGE VISION DRIVE_ED;
DATALINES;
1 17 1 1
1 44 0 0
1 48 1 0
1 55 0 0
1 75 1 1
0 35 0 1
0 42 1 1
0 57 0 0
0 28 0 1
0 20 0 1
0 38 1 0
0 45 0 1
0 47 1 1
0 52 0 0
0 55 0 1
1 68 1 0
1 18 1 0
1 68 0 0
1 48 1 1
1 17 0 0
```

```
1 70 1 1
1 72 1 0
1 35 0 1
1 19 1 0
1 62 1 0
0 39 1 1
0 40 1 1
0 55 0 0
0 68 0 1
0 25 1 0
0 17 0 0
0 45 0 1
0 44 0 1
0 67 0 0
0 55 0 1
1 61 1 0
1 19 1 0
1 69 0 0
1 23 1 1
1 19 0 0
1 72 1 1
1 74 1 0
1 31 0 1
1 16 1 0
1 61 1 0
;

Better to use: INFILE and give the complete path on wam
'/homes/bnk/driver' where the data are in "driver".

OPTION PS=45 LS=70;

DATA LOGISTIC;
INFILE '/homes/abc/driver';
INPUT ACCIDENT AGE VISION DRIVE_ED;
DATALINES;


First simple logistic (not stepwise)
------------------------------------
```

```
PROC LOGISTIC DATA=LOGISTIC DESCENDING;
MODEL ACCIDENT=AGE VISION DRIVE_ED;
RUN;
QUIT;
```

The LOGISTIC Procedure

Model Information

Data Set	WORK.LOGISTIC
Response Variable	ACCIDENT
Number of Response Levels	2
Number of Observations	45
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	ACCIDENT	Total Frequency
1	1	25
2	0	20

Probability modeled is ACCIDENT=1.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.1883	0.9945	0.0359	0.8498
AGE	1	0.00656	0.0183	0.1290	0.7195
VISION	1	1.7096	0.7056	5.8708	0.0154
DRIVE_ED	1	-1.4937	0.7046	4.4949	0.0340

We see that AGE is not significant. So, now run FORWARD selection.

Second FORWARD selection and options

```
PROC LOGISTIC DATA=LOGISTIC DESCENDING;
MODEL ACCIDENT=AGE VISION DRIVE_ED/
                  selection=forward
                  ctable pprob = (0 to 1 by .1)
                  lackfit
                  risklimits;
RUN;
QUIT;
```

The LOGISTIC Procedure

Model Information

Data Set	WORK.LOGISTIC
Response Variable	ACCIDENT
Number of Response Levels	2
Number of Observations	45
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	ACCIDENT	Total Frequency
1	1	25
2	0	20

Probability modeled is ACCIDENT=1.

Forward Selection Procedure

Step 0. Intercept entered:

Model Convergence Status

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

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
10.7057	3	0.0134

Step 1. Effect VISION entered:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	63.827	59.244
SC	65.633	62.857
-2 Log L	61.827	55.244

Note: AIC = -2 Log L + 2*p. E.g. 55.244+2*2=59.244
Note: SC=Schwartz Criterion=BIC
NOTE: -2 Log (L1-L2) = 61.827 - 55.244 = 6.583 WITH 1 DF!!!
The p-value is 0.0103 and so VISION is sginificant!!

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq

Likelihood Ratio	6.5830	1	0.0103
Score	6.4209	1	0.0113
Wald	6.0756	1	0.0137

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
4.9818	2	0.0828

Step 2. Effect DRIVE_ED entered:

Model Convergence Status

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

Criterion	Intercept Only	Intercept and Covariates
AIC	63.827	56.287
SC	65.633	61.707
-2 Log L	61.827	50.287

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.5391	2	0.0031
Score	10.5976	2	0.0050
Wald	8.5949	2	0.0136

Note: -2 Log (L1-L2)=61.827-50.287=11.54 tests 2 coefficients of both VISION and DRIVE_ED. Thus df=3-1=2 !!!

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
0.1293	1	0.7191

NOTE: No (additional) effects met the 0.05 significance level for entry into the model. Thus only VISION and DRIVE_ED are significant.

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	VISION	1	1	6.4209	0.0113
2	DRIVE_ED	1	2	4.8680	0.0274

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.1110	0.5457	0.0414	0.8389
VISION	1	1.7137	0.7049	5.9113	0.0150
DRIVE_ED	1	-1.5000	0.7037	4.5440	0.0330

Model:
p=Prob. of an accident

logit=log(p/(1-p))=log(Odds of an accident) = z'b =
= 0.1110 + 1.7137*VISION - 1.5000*DRIVE_ED

Thus, If VISION =0, and DRIVE_ED=0, then

1

log(p/(1-p))= 0.1110, and p= $\frac{1}{1 + \exp(-0.1110)}$ = 0.5277

In general p= $\frac{1}{1+\exp(-z'b)}$ = Odds/(1+Odds)!!!

Note: p/(1-p)=exp(0.1110)=1.117395 <---- odds

If VISION =1, and DRIVE_ED=0, then

log(p/(1-p))= 0.1110 + 1.7137 = 1.8247

$$p = \frac{Odds}{1+Odds} = \frac{\exp(1.8247)}{1+\exp(1.8247)} = \frac{6.200934}{1+6.200934} = 0.8611291$$

Thus, Odds Ratio = 6.200934/1.117395=5.549456

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
VISION	5.550	1.394	22.093
DRIVE_ED	0.223	0.056	0.886

Association of Predicted Probabilities and Observed Responses

Percent Concordant	67.2	Somers' D	0.532
Percent Discordant	14.0	Gamma	0.655
Percent Tied	18.8	Tau-a	0.269
Pairs	500	c	0.766

Wald Confidence Interval for Adjusted Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits

VISION	1.0000	5.550	1.394	22.093
DRIVE_ED	1.0000	0.223	0.056	0.886

The LOGISTIC Procedure

Partition for the Hosmer and Lemeshow Test

		ACCIDENT = 1		ACCIDENT = 0	
Group	Total	Observed	Expected	Observed	Expected
1	11	2	2.20	9	8.80
2	11	6	5.80	5	5.20
3	10	6	5.80	4	4.20
4	13	11	11.19	2	1.81

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
0.0756	2	0.9629

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	25	0	20	0	55.6	100.0	0.0	44.4	.
0.100	25	0	20	0	55.6	100.0	0.0	44.4	.
0.200	23	0	20	2	51.1	92.0	0.0	46.5	100.0
0.300	23	9	11	2	71.1	92.0	45.0	32.4	18.2
0.400	23	9	11	2	71.1	92.0	45.0	32.4	18.2
0.500	17	9	11	8	57.8	68.0	45.0	39.3	47.1
0.600	11	14	6	14	55.6	44.0	70.0	35.3	50.0
0.700	11	18	2	14	64.4	44.0	90.0	15.4	43.8
0.800	11	18	2	14	64.4	44.0	90.0	15.4	43.8
0.900	0	18	2	25	40.0	0.0	90.0	100.0	58.1

```
   1.000     0    20     0    25    44.4    0.0  100.0     .    55.6
```

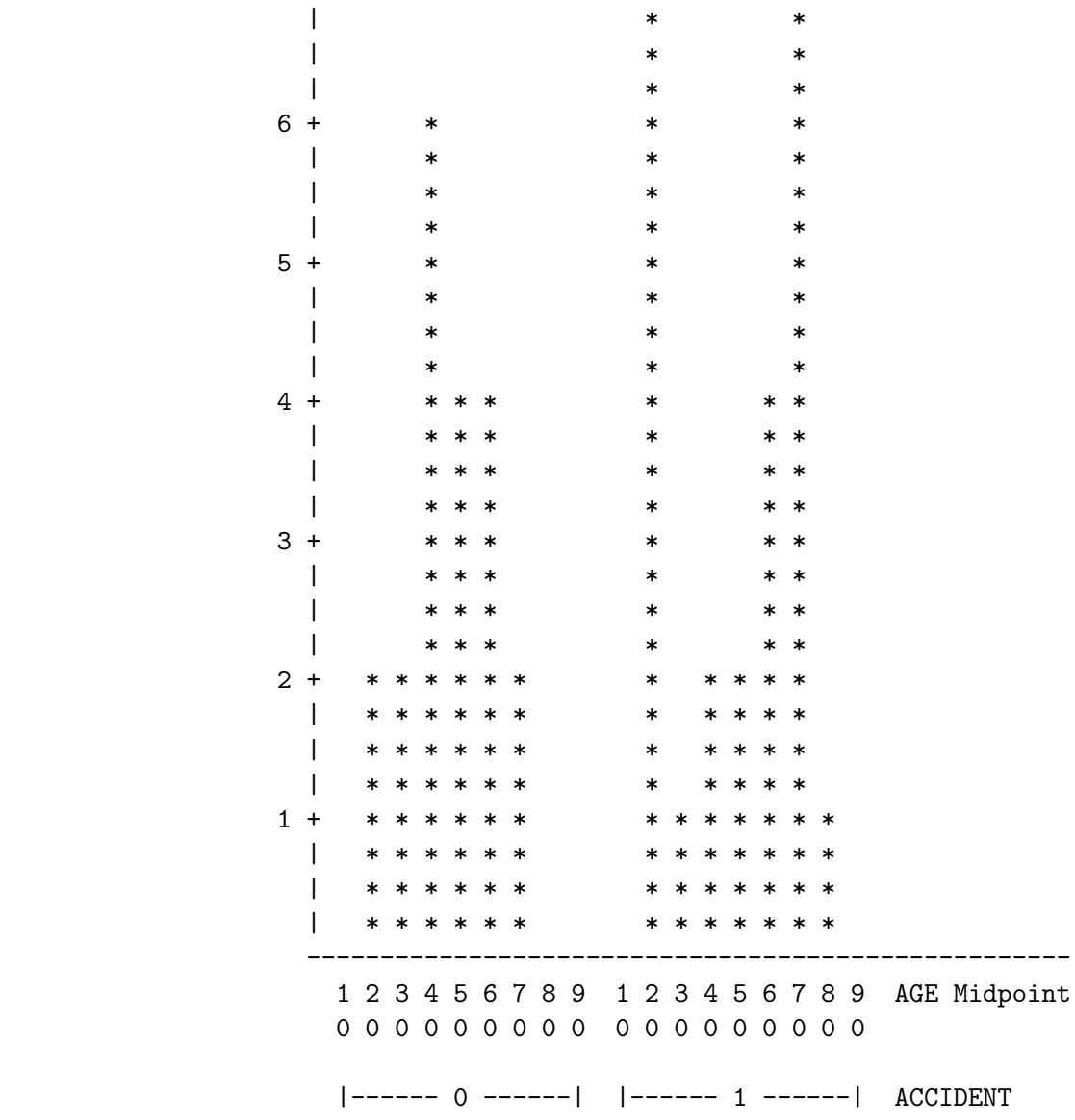
Assume Prob level = 0.3. This means we follow the prediction rule

```
 p >= 0.3 -----> Predict Accident
   <  0.3 -----> Predict No Accident
```

Recall: there were 25 accidents, 20 no accidents.

Sensitivity=P(Correct|Accident Occured)= 23/25=0.92
Specifity=P(Corredct|Accident Did'nt Occur)=9/20=0.45
False POS=P(Incorrrect|Occurrence was predicted)=11/(11+23)=11/34=0.3235
False NEG=P(Incortrect|Non-Occurrence was predicted)=2/(2+9)=2/11=0.1818
P(Correct)=(23+9)/45 = 32/45=0.7111111

Now, we are surprised that AGE did not enter. Let's look at the AGE distribution for those who had and did not have an accident.



We see the non-accident class contains mainly "middle age" people, but the accident class consists mainly of very young or very old people. This suggets that we code AGE differently by categorizing it as AGEGROUP as follows:

AGEGROUP=0 if AGE in [20,65]

```
AGEGROUP=1 otherwise

OPTION PS=45 LS=70;

DATA LOGISTIC;
INPUT ACCIDENT AGE VISION DRIVE_ED;
IF AGE < 20 OR AGE > 65 THEN AGEGROUP=1;
ELSE AGEGROUP=0;
DATALINES;
1 17 1 1
1 44 0 0
1 48 1 0
1 55 0 0
1 75 1 1
0 35 0 1
0 42 1 1
0 57 0 0
0 28 0 1
0 20 0 1
0 38 1 0
0 45 0 1
0 47 1 1
0 52 0 0
0 55 0 1
1 68 1 0
1 18 1 0
1 68 0 0
1 48 1 1
1 17 0 0
1 70 1 1
1 72 1 0
1 35 0 1
1 19 1 0
1 62 1 0
0 39 1 1
0 40 1 1
0 55 0 0
0 68 0 1
0 25 1 0
0 17 0 0
```

```
0 45 0 1
0 44 0 1
0 67 0 0
0 55 0 1
1 61 1 0
1 19 1 0
1 69 0 0
1 23 1 1
1 19 0 0
1 72 1 1
1 74 1 0
1 31 0 1
1 16 1 0
1 61 1 0
;
```

Or use INFILE:

```
DATA LOGISTIC;
INFILE '/homes/bnk/driver';
INPUT ACCIDENT AGE VISION DRIVE_ED;
IF AGE < 20 OR AGE > 65 THEN AGEGROUP=1;
ELSE AGEGROUP=0;
DATALINES;
```

Now, replace AGE by AGEGROUP:

```
PROC LOGISTIC DATA=LOGISTIC DESCENDING;
MODEL ACCIDENT=AGEGROUP VISION DRIVE_ED/
                    selection=forward
                    ctable pprob = (0 to 1 by .1)
                    lackfit
                    risklimits;
RUN;
QUIT;
```

This time we get a very different model including AGEGROUP!!!

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.3334	0.5854	5.1886	0.0227
AGEGROUP	1	2.1611	0.8014	7.2711	0.0070
VISION	1	1.6258	0.7325	4.9265	0.0264

log(p/(1-p))=-1.3334 + 2.1611*AGEGROUP + 1.6258*VISION

Summary:

Model	AIC
Accident = Vision+Dr.Ed	56.2874
Accident = Age+Vision+Dr.Ed	58.1583
Accident = AGEGROUP+Vision	52.434 <---- Better model!!!

==========

Now use PROBIT on model ACCIDENT=AGEGROUP+VISION for comparison!!!

Note on Binary response models:

By default for a binary response, Y with levels 0 and 1 in the data, PROBIT sorts the levels in ascending order and assigns Ordered Value 1 to the lowest level (0) and Ordered Value 2 to the next lowest level (1). Consequently, PROBIT models Pr(Y=0). To model Pr(Y=1), switch the response levels by doing a descending sort, and then specify the ORDER=DATA option in PROBIT to tell the procedure to order the response levels as they are encountered in the data:

```
proc sort;
   by descending y;
   run;
proc probit order=data;
   class y;
```

```
          model y = <your model effects>;
          run;
```

Thus, in our case:

```
proc sort;
by descending ACCIDENT;
run;

PROC PROBIT DATA=LOGISTIC order=data;;
class ACCIDENT;
MODEL ACCIDENT=AGEGROUP VISION;
RUN;
```

The SAS System 1

Probit Procedure

Model Information

Data Set	WORK.LOGISTIC
Dependent Variable	ACCIDENT
Number of Observations	45
Name of Distribution	Normal
Log Likelihood	-23.05899761

Class Level Information

Name	Levels	Values
ACCIDENT	2	1 0

Response Profile

Ordered Value	ACCIDENT	Total Frequency

1	1	25
2	0	20

PROC PROBIT is modeling the probabilities of levels of ACCIDENT having LOWER Ordered Values in the response profile table.

Algorithm converged.

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
AGEGROUP	1	8.0957	0.0044
VISION	1	5.4002	0.0201

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.8063	0.3316	-1.4562	-0.1564	5.91	0.0150
AGEGROUP	1	1.3273	0.4665	0.4130	2.2416	8.10	0.0044
VISION	1	0.9997	0.4302	0.1565	1.8428	5.40	0.0201

Phi^{-1}(p) = -0.8063 + 1.3273*AGEGROUP + 0.9997*VISION

p=Phi(-0.8063 + 1.3273*AGEGROUP + 0.9997*VISION)

(AGEGROUP, VISION)=(0,0), p=Phi(-0.8063)=0.2100349
(AGEGROUP, VISION)=(0,1), p=Phi(-0.8063+0.9997)=0.5766771

Compare with logistic reg:

log(p/(1-p))=-1.3334 + 2.1611*AGEGROUP + 1.6258*VISION
1/(1+exp(1.3334))=0.2085975

```
(AGEGROUP, VISION)=(0,0), p=1/(1+exp(1.3334))=0.2085975
(AGEGROUP, VISION)=(0,1), p=1/(1+exp(1.3334-1.6258))=0.5725836

Almost the same!!!
============================================================

Compare with Splus
--------------------
Data in stat/STAT430/driver

> A <- matrix(scan("driver", what=numeric()), ncol=4, byrow=T)
> A[1:3,]
     [,1] [,2] [,3] [,4]
[1,]    1   17    1    1
[2,]    1   44    0    0
[3,]    1   48    1    0

> dimnames(A) <- list(NULL,c('Accident','Age','Vision','Dr_Ed'))

> A <- data.frame(A)  ##Convenient: get names for columns.
> A
   Accident Age Vision Dr.Ed
 1        1  17      1     1
 2        1  44      0     0
 3        1  48      1     0
 4        1  55      0     0
 5        1  75      1     1
 6        0  35      0     1
 7        0  42      1     1
 8        0  57      0     0
 9        0  28      0     1
10        0  20      0     1
11        0  38      1     0
12        0  45      0     1
13        0  47      1     1
14        0  52      0     0
15        0  55      0     1
16        1  68      1     0
17        1  18      1     0
```

```
18        1  68      0     0
19        1  48      1     1
20        1  17      0     0
21        1  70      1     1
22        1  72      1     0
23        1  35      0     1
24        1  19      1     0
25        1  62      1     0
26        0  39      1     1
27        0  40      1     1
28        0  55      0     0
29        0  68      0     1
30        0  25      1     0
31        0  17      0     0
32        0  45      0     1
33        0  44      0     1
34        0  67      0     0
35        0  55      0     1
36        1  61      1     0
37        1  19      1     0
38        1  69      0     0
39        1  23      1     1
40        1  19      0     0
41        1  72      1     1
42        1  74      1     0
43        1  31      0     1
44        1  16      1     0
45        1  61      1     0

> A$Accident
[1] 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0
    0 0 1 1 1
[39] 1 1 1 1 1 1 1

LR1 <- glm(A$Accident ~ A$Age+A$Vision+A$Dr.Ed, family=binomial)

Better to change notation:
Accident <- A$Accident
Age <- A$Age
Vision <- A$Vision
```

```
Dr.Ed <- A$Dr.Ed

First: Accident ~ Age+Vision+Dr.Ed

LR1 <- glm(Accident ~ Age+Vision+Dr.Ed, family=binomial)

> LR1
Call:
glm(formula = Accident ~ Age + Vision + Dr.Ed, family = binomial)

Coefficients:
 (Intercept)         Age   Vision     Dr.Ed
  -0.1883224 0.006556125 1.709513 -1.493708

Degrees of Freedom: 45 Total; 41 Residual
Residual Deviance: 50.15832


> summary(LR1)

Call: glm(formula = Accident ~ Age + Vision + Dr.Ed, family = binomial)
Deviance Residuals:
       Min         1Q    Median        3Q      Max
 -1.963448 -0.7141648 0.5117001 0.9790089 1.835269

Coefficients:
                   Value Std. Error    t value
(Intercept) -0.188322357 0.99157996 -0.1899215
        Age  0.006556125 0.01819121  0.3604007
     Vision  1.709512725 0.70187097  2.4356510
      Dr.Ed -1.493708417 0.70095875 -2.1309505

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 61.82654 on 44 degrees of freedom

Residual Deviance: 50.15832 on 41 degrees of freedom

Number of Fisher Scoring Iterations: 3
```

```
Correlation of Coefficients:
       (Intercept)        Age     Vision
   Age -0.8346803
Vision -0.2654521   0.0042028
 Dr.Ed -0.2837496   0.0079376 -0.2094491


Second: Accident ~ Vision+Dr.Ed


LR1 <- glm(Accident ~ Vision+Dr.Ed, family=binomial)


> LR1
Call:
glm(formula = Accident ~ Vision + Dr.Ed, family = binomial)

Coefficients:
 (Intercept)   Vision     Dr.Ed
   0.1109733 1.713708 -1.499912

Degrees of Freedom: 45 Total; 42 Residual
Residual Deviance: 50.28742


To test the significance of AGE, take the differenmce in Residual Deviance:


Model                           Resid Dev  df
-----------------------------------------------------------
Accident =     Vision+Dr.Ed    50.28742   42=N-p=45-3
Accident = Age+Vision+Dr.Ed    50.15832   41=N-p=54-4


> 50.28742-50.15832 at df=42-41=1
[1] 0.1291
> 1 - pchisq(0.1291,1)
[1] 0.719367 <------- p-value with df=1. Large!!! Similar to SAS!!!

Thus AGE is not significant!!! So, Accident = Vision+Dr.Ed is a
```

```
sensible model. Now let's see what step() gives using AIC.

Now performs stepwise model selection using "step()".

> step(LR1)
Start:  AIC= 58.1583
 Accident ~ Age + Vision + Dr.Ed

Single term deletions

Model:
Accident ~ Age + Vision + Dr.Ed

scale:  1

       Df Sum of Sq      RSS       Cp
<none>              44.39405 52.39405
   Age  1  0.129889 44.52394 50.52394
Vision  1  5.932396 50.32645 56.32645
 Dr.Ed  1  4.540950 48.93500 54.93500

Step:  AIC= 56.2874
 Accident ~ Vision + Dr.Ed

Single term deletions

Model:
Accident ~ Vision + Dr.Ed

scale:  1

       Df Sum of Sq      RSS       Cp
<none>              44.83992 50.83992
Vision  1  5.972585 50.81250 54.81250
 Dr.Ed  1  4.589943 49.42986 53.42986
Call:
glm(formula = Accident ~ Vision + Dr.Ed, family = binomial)

Coefficients:
 (Intercept)   Vision     Dr.Ed
```

```
   0.1109733 1.713708 -1.499912

Degrees of Freedom: 45 Total; 42 Residual
Residual Deviance: 50.28742

Get the same selected model without AGE as from SAS and as above!!!!

logit=log(p/(1-p))=log(Odds of an accident) = z'b =
                         = 0.1110 + 1.7137*VISION - 1.5000*DRIVE_ED

Summary:

Model                           Resid Dev  df           AIC
-------------------------------------------------------------------
Accident =     Vision+Dr.Ed     50.28742   42=N-p=45-3  56.2874
Accident = Age+Vision+Dr.Ed     50.15832   41=N-p=54-4  58.1583
Accident = AGEGROUP+Vision                              52.434
```