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 0 35 0 1 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
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

<table>
<thead>
<tr>
<th>Ordered Value</th>
<th>ACCIDENT</th>
<th>Total Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>20</td>
</tr>
</tbody>
</table>

Probability modeled is ACCIDENT=1.

Model Fit Statistics

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Intercept Only</th>
<th>Intercept and Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>63.827</td>
<td>58.158</td>
</tr>
<tr>
<td>SC</td>
<td>65.633</td>
<td>65.385</td>
</tr>
<tr>
<td>-2 Log L</td>
<td>61.827</td>
<td>50.158</td>
</tr>
</tbody>
</table>

Testing Global Null Hypothesis: BETA=0=beta1=beta2=beta3
<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>11.6682</td>
<td>3</td>
<td>0.0086</td>
</tr>
<tr>
<td>Score</td>
<td>10.7057</td>
<td>3</td>
<td>0.0134</td>
</tr>
<tr>
<td>Wald</td>
<td>8.6681</td>
<td>3</td>
<td>0.0340</td>
</tr>
</tbody>
</table>

Note: $P(-2(\log L_1 - \log L_2) > 11.6682) = 0.0086$, df=3

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-0.1883</td>
<td>0.9945</td>
<td>0.0359</td>
<td>0.8498</td>
</tr>
<tr>
<td>AGE</td>
<td>1</td>
<td>0.00656</td>
<td>0.0183</td>
<td>0.1290</td>
<td>0.7195</td>
</tr>
<tr>
<td>VISION</td>
<td>1</td>
<td>1.7096</td>
<td>0.7056</td>
<td>5.8708</td>
<td>0.0154</td>
</tr>
<tr>
<td>DRIVE_ED</td>
<td>1</td>
<td>-1.4937</td>
<td>0.7046</td>
<td>4.4949</td>
<td>0.0340</td>
</tr>
</tbody>
</table>

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  Total
   Value   ACCIDENT Frequency
           1     25
           2     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

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.7057</td>
<td>3</td>
<td>0.0134</td>
</tr>
</tbody>
</table>
Step 1. Effect VISION entered:

Model Fit Statistics

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Intercept Only</th>
<th>Intercept and Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>63.827</td>
<td>59.244</td>
</tr>
<tr>
<td>SC</td>
<td>65.633</td>
<td>62.857</td>
</tr>
<tr>
<td>-2 Log L</td>
<td>61.827</td>
<td>55.244</td>
</tr>
</tbody>
</table>

Note: AIC = -2 Log L + 2p. E.g. 55.244+2*2=59.244
Note: SC=Schwartz Criterion=BIC
NOTE: -2(LogL1-LogL2) = 61.827 - 55.244 = 6.583 WITH 1 DF!!

The p-value is 0.0103 and so VISION is significant!!

Testing Global Null Hypothesis: BETA=0

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>6.5830</td>
<td>1</td>
<td>0.0103</td>
</tr>
<tr>
<td>Score</td>
<td>6.4209</td>
<td>1</td>
<td>0.0113</td>
</tr>
<tr>
<td>Wald</td>
<td>6.0756</td>
<td>1</td>
<td>0.0137</td>
</tr>
</tbody>
</table>

Residual Chi-Square Test

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.9818</td>
<td>2</td>
<td>0.0828</td>
</tr>
</tbody>
</table>

Step 2. Effect DRIVE_ED entered:
Model Convergence Status

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

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Intercept Only</th>
<th>Intercept and Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>63.827</td>
<td>56.287</td>
</tr>
<tr>
<td>SC</td>
<td>65.633</td>
<td>61.707</td>
</tr>
<tr>
<td>-2 Log L</td>
<td>61.827</td>
<td>50.287</td>
</tr>
</tbody>
</table>

Testing Global Null Hypothesis: BETA=0

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>11.5391</td>
<td>2</td>
<td>0.0031</td>
</tr>
<tr>
<td>Score</td>
<td>10.5976</td>
<td>2</td>
<td>0.0050</td>
</tr>
<tr>
<td>Wald</td>
<td>8.5949</td>
<td>2</td>
<td>0.0136</td>
</tr>
</tbody>
</table>

Note: \(-2(\text{LogL}_1-\text{LogL}_2)=61.827-50.287=11.54\) tests 2 coefficients of both VISION and DRIVE_ED. Thus df=3-1=2 !!!

Residual Chi-Square Test

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1293</td>
<td>1</td>
<td>0.7191</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Effect</th>
<th>Number</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step</td>
<td>Entered</td>
<td>DF</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>----</td>
</tr>
<tr>
<td>1</td>
<td>VISION</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>DRIVE_ED</td>
<td>1</td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>0.1110</td>
<td>0.5457</td>
<td>0.0414</td>
<td>0.8389</td>
</tr>
<tr>
<td>VISION</td>
<td>1</td>
<td>1.7137</td>
<td>0.7049</td>
<td>5.9113</td>
<td>0.0150</td>
</tr>
<tr>
<td>DRIVE_ED</td>
<td>1</td>
<td>-1.5000</td>
<td>0.7037</td>
<td>4.5440</td>
<td>0.0330</td>
</tr>
</tbody>
</table>

Model:
p=Prob. of an accident

\[
\text{logit}=\log(p/(1-p))=\log(\text{Odds of an accident}) = z'\beta = 0.1110 + 1.7137\times\text{VISION} - 1.5000\times\text{DRIVE}_ED
\]

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

\[
\frac{1}{1 + \exp(-0.1110)} = 0.5277
\]

In general \( p = \frac{\text{Odds}}{1+\text{Odds}} \)

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

If VISION =1, and DRIVE_ED=0, then

\[
\log(p/(1-p)) = 0.1110 + 1.7137 = 1.8247
\]
Odds exp(1.8247) = 6.200934
p = -------- = ----------------- = ----------- = 0.8611291
1+Odds 1+ exp(1.8247) 1+6.200934

Thus, Odds Ratio = 6.200934/1.117395 = 5.549456

Odds Ratio Estimates

<table>
<thead>
<tr>
<th>Effect</th>
<th>Point Estimate</th>
<th>95% Wald Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>VISION</td>
<td>5.550</td>
<td>1.394</td>
</tr>
<tr>
<td>DRIVE_ED</td>
<td>0.223</td>
<td>0.056</td>
</tr>
</tbody>
</table>

Association of Predicted Probabilities and Observed Responses

<table>
<thead>
<tr>
<th>Percent Concordant</th>
<th>Somers' D</th>
<th>Percent Discordant</th>
<th>Gamma</th>
<th>Percent Tied</th>
<th>Tau-a</th>
<th>Pairs</th>
<th>c</th>
<th>Tau-a</th>
</tr>
</thead>
<tbody>
<tr>
<td>67.2</td>
<td>0.532</td>
<td>14.0</td>
<td>0.655</td>
<td>18.8</td>
<td>0.269</td>
<td>500</td>
<td>0.766</td>
<td></td>
</tr>
</tbody>
</table>

Wald Confidence Interval for Adjusted Odds Ratios

<table>
<thead>
<tr>
<th>Effect</th>
<th>Unit</th>
<th>Estimate</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>VISION</td>
<td>1.0000</td>
<td>5.550</td>
<td>1.394</td>
</tr>
<tr>
<td>DRIVE_ED</td>
<td>1.0000</td>
<td>0.223</td>
<td>0.056</td>
</tr>
</tbody>
</table>

The LOGISTIC Procedure

Partition for the Hosmer and Lemeshow Test

<table>
<thead>
<tr>
<th>Group</th>
<th>ACCIDENT = 1</th>
<th>ACCIDENT = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>Total</td>
<td>Observed</td>
</tr>
</tbody>
</table>

Hosmer and Lemeshow Goodness-of-Fit Test

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0756</td>
<td>2</td>
<td>0.9629</td>
</tr>
</tbody>
</table>

Classification Table

<table>
<thead>
<tr>
<th>Prob Level</th>
<th>Event Correct</th>
<th>Event Incorrect</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>False Positive</th>
<th>False Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>25</td>
<td>0</td>
<td>55.6</td>
<td>100.0</td>
<td>0.0</td>
<td>44.4</td>
</tr>
<tr>
<td>0.100</td>
<td>25</td>
<td>0</td>
<td>55.6</td>
<td>100.0</td>
<td>0.0</td>
<td>44.4</td>
</tr>
<tr>
<td>0.200</td>
<td>23</td>
<td>0</td>
<td>51.1</td>
<td>92.0</td>
<td>0.0</td>
<td>46.5</td>
</tr>
<tr>
<td>0.300</td>
<td>23</td>
<td>9</td>
<td>71.1</td>
<td>92.0</td>
<td>45.0</td>
<td>32.4</td>
</tr>
<tr>
<td>0.400</td>
<td>23</td>
<td>9</td>
<td>71.1</td>
<td>92.0</td>
<td>45.0</td>
<td>32.4</td>
</tr>
<tr>
<td>0.500</td>
<td>17</td>
<td>9</td>
<td>57.8</td>
<td>68.0</td>
<td>45.0</td>
<td>39.3</td>
</tr>
<tr>
<td>0.600</td>
<td>11</td>
<td>14</td>
<td>55.6</td>
<td>44.0</td>
<td>70.0</td>
<td>35.3</td>
</tr>
<tr>
<td>0.700</td>
<td>11</td>
<td>18</td>
<td>64.4</td>
<td>44.0</td>
<td>90.0</td>
<td>15.4</td>
</tr>
<tr>
<td>0.800</td>
<td>11</td>
<td>18</td>
<td>64.4</td>
<td>44.0</td>
<td>90.0</td>
<td>15.4</td>
</tr>
<tr>
<td>0.900</td>
<td>0</td>
<td>18</td>
<td>40.0</td>
<td>0.0</td>
<td>90.0</td>
<td>100.0</td>
</tr>
<tr>
<td>1.000</td>
<td>0</td>
<td>20</td>
<td>44.4</td>
<td>0.0</td>
<td>100.0</td>
<td>55.6</td>
</tr>
</tbody>
</table>

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

\[ p \geq 0.3 \Rightarrow \text{Predict Accident} \]
\[ < 0.3 \Rightarrow \text{Predict No Accident} \]

Recall: there were 25 accidents, 20 no accidents.

Sensitivity = \( \frac{\text{Correct}}{\text{Accident Occurred}} = \frac{23}{25} = 0.92 \)

Specificity = \( \frac{\text{Correct}}{\text{Accident Did'nt Occur}} = \frac{9}{20} = 0.45 \)
False POS = P(Incorrect | Occurrence was predicted) = 11 / (11 + 23) = 11 / 34 = 0.3235
False NEG = P(Incorrect | Non-Occurrence was predicted) = 2 / (2 + 9) = 2 / 11 = 0.1818
P(Correct) = (23 + 9) / 45 = 32 / 45 = 0.711111

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.

To get separate vbars for the accident classes 0 and 1 use GCHART:

```plaintext
proc gchart data=logistic;
  vbar age / midpoints=10 to 90 by 10;
  group=accident;
run;
```
<table>
<thead>
<tr>
<th>Frequency</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>AGE</th>
<th>Midpoint</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>--------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>-----</td>
<td>----------</td>
</tr>
<tr>
<td>8 +</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 +</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 +</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 +</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 +</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 +</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 +</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 +</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>-----</td>
<td>----------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

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 suggests that we code AGE differently by categorizing it as AGEGROUP as follows:

\[
\text{AGEGROUP}=0 \text{ if } \text{AGE in } [20,65] \\
\text{AGEGROUP}=1 \text{ otherwise}
\]

```
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
```
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, in the model 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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-1.3334</td>
<td>0.5854</td>
<td>5.1886</td>
<td>0.0227</td>
</tr>
<tr>
<td>AGEGROUP</td>
<td>1</td>
<td>2.1611</td>
<td>0.8014</td>
<td>7.2711</td>
<td>0.0070</td>
</tr>
<tr>
<td>VISION</td>
<td>1</td>
<td>1.6258</td>
<td>0.7325</td>
<td>4.9265</td>
<td>0.0264</td>
</tr>
</tbody>
</table>

\[
\log(p/(1-p)) = -1.3334 + 2.1611 \times \text{AGEGROUP} + 1.6258 \times \text{VISION}
\]

Summary:

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident = Vision+Dr.Ed</td>
<td>56.2874</td>
</tr>
<tr>
<td>Accident = Age+Vision+Dr.Ed</td>
<td>58.1583</td>
</tr>
<tr>
<td>Accident = AGEGROUP+Vision</td>
<td>52.434  &lt;---- Better model!!!</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

16
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

<table>
<thead>
<tr>
<th>Effect</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGEGROUP</td>
<td>1</td>
<td>8.0957</td>
<td>0.0044</td>
</tr>
<tr>
<td>VISION</td>
<td>1</td>
<td>5.4002</td>
<td>0.0201</td>
</tr>
</tbody>
</table>

Analysis of Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
<th>95% Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-0.8063</td>
<td>0.3316</td>
<td>-1.4562 -0.1564</td>
<td>5.91</td>
<td>0.0150</td>
</tr>
<tr>
<td>AGEGROUP</td>
<td>1</td>
<td>1.3273</td>
<td>0.4665</td>
<td>0.4130 2.2416</td>
<td>8.10</td>
<td>0.0044</td>
</tr>
<tr>
<td>VISION</td>
<td>1</td>
<td>0.9997</td>
<td>0.4302</td>
<td>0.1565 1.8428</td>
<td>5.40</td>
<td>0.0201</td>
</tr>
</tbody>
</table>

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:

\[ \log\left(\frac{p}{1-p}\right) = -1.3334 + 2.1611 \times \text{AGEGROUP} + 1.6258 \times \text{VISION} \]
\[ \frac{1}{1+\exp(1.3334)}=0.2085975 \]

(AGEGROUP, VISION)=(0,0), p=\frac{1}{1+\exp(1.3334)}=0.2085975
(AGEGROUP, VISION)=(0,1), p=\frac{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,] 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

<table>
<thead>
<tr>
<th>Accident</th>
<th>Age</th>
<th>Vision</th>
<th>Dr.Ed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>44</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>48</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>75</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>35</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>42</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>57</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
> A$Accident
[1] 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
   0 0 1 1 1
[39] 1 1 1 1 1 1 1
LR1 <- glm(A$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 -0.714165  0.511700  0.979009  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 difference in Residual Deviance:

<table>
<thead>
<tr>
<th>Model</th>
<th>Resid Dev</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident = Vision+Dr.Ed</td>
<td>50.28742</td>
<td>42=N-p=45-3</td>
</tr>
<tr>
<td>Accident = Age+Vision+Dr.Ed</td>
<td>50.15832</td>
<td>41=N-p=54-4</td>
</tr>
</tbody>
</table>
> 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:  

<table>
<thead>
<tr>
<th>Model</th>
<th>Resid Dev</th>
<th>df</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident = Vision+Dr.Ed</td>
<td>50.28742</td>
<td>42=N-p=45-3</td>
<td>56.2874</td>
</tr>
<tr>
<td>Accident = Age+Vision+Dr.Ed</td>
<td>50.15832</td>
<td>41=N-p=54-4</td>
<td>58.1583</td>
</tr>
<tr>
<td>Accident = AGEGROUP+Vision</td>
<td></td>
<td></td>
<td>52.434</td>
</tr>
</tbody>
</table>

23