**Analysis of the Titanic Data Set**

LOGISTIC REGRESSION VARIABLES SURVIVE

/METHOD=FSTEP(LR) CLASS AGE SEX

/CONTRAST (CLASS)=Indicator

/CONTRAST (AGE)=Indicator

/CONTRAST (SEX)=Indicator

/CLASSPLOT

/PRINT=ITER(1) CI(95)

/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).

**Logistic Regression**

|  |  |  |  |
| --- | --- | --- | --- |
| **Case Processing Summary** | | | |
| Unweighted Casesa | | N | Percent |
| Selected Cases | Included in Analysis | 2201 | 100.0 |
| Missing Cases | 0 | .0 |
| Total | 2201 | 100.0 |
| Unselected Cases | | 0 | .0 |
| Total | | 2201 | 100.0 |

|  |
| --- |
| a. If weight is in effect, see classification table for the total number of cases. |

|  |  |
| --- | --- |
| **Dependent Variable Encoding** | |
| Original Value | Internal Value |
| No | 0 |
| Yes | 1 |

This table indicates that the survivors are indicated by a value of 1, and those who did not survive are represented by a value of 0. These values determine how the output will be interpreted.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Categorical Variables Codings** | | | | | |
|  | | Frequency | Parameter coding | | |
| (1) | (2) | (3) |
| CLASS | Crew | 885 | 1.000 | .000 | .000 |
| First | 325 | .000 | 1.000 | .000 |
| Second | 285 | .000 | .000 | 1.000 |
| Third | 706 | .000 | .000 | .000 |
| SEX | Female | 470 | 1.000 |  |  |
| Male | 1731 | .000 |  |  |
| AGE | Child | 109 | 1.000 |  |  |
| Adult | 2092 | .000 |  |  |

The variables of Class, Sex, and Age are all categorical, and this table explains how each is coded. Sex and Age are binary, so their coding is simple: 1 for female, 1 for child, 0 for male, 0 for adult. Again, the coding factors into the interpretation of the results.

The variable of Class has 4 categories, and the coding indicates that all categories will be compared to Third Class. Because Third Class has all zeroes in the coding, it has been designated as the “baseline” category.

**Block 0: Beginning Block**

|  |  |  |  |
| --- | --- | --- | --- |
| **Iteration Historya,b,c** | | | |
| Iteration | | -2 Log likelihood | Coefficients |
| Constant |
| Step 0 | 1 | 2769.951 | -.708 |
| 2 | 2769.457 | -.740 |
| 3 | 2769.457 | -.740 |

|  |
| --- |
| a. Constant is included in the model. |
| b. Initial -2 Log Likelihood: 2769.457 |
| c. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001. |

Step 0 -2 Log Likelihood (-2LL) provides a baseline assessment of how well cases can be classified simply on the basis of classifying everyone into the category (survived or didn’t survive) that has the largest number of cases. In this case, everyone was classified into the “didn’t survive” category (see table below). The -2LL is analogous to the residual sum of squares in linear regression. It is an indicator of how much unexplained information there is after the model has been fitted. When comparing -2LL from one step to the next, the magnitude of -2LL should decrease, and the amount of decrease provides some sense of predictive improvement with each added variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classification Tablea,b** | | | | | |
|  | Observed | | Predicted | | |
| SURVIVE | | Percentage Correct |
| No | Yes |
| Step 0 | SURVIVE | No | 1490 | 0 | 100.0 |
| Yes | 711 | 0 | .0 |
| Overall Percentage | |  |  | 67.7 |

|  |
| --- |
| a. Constant is included in the model. |
| b. The cut value is .500 |

With only the constant in the equation, prediction places all cases in the Did Not Survive category.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | | |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 0 | Constant | -.740 | .046 | 263.472 | 1 | .000 | .477 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables not in the Equation** | | | | | |
|  | | | Score | df | Sig. |
| Step 0 | Variables | CLASS | 190.401 | 3 | .000 |
| CLASS(1) | 47.177 | 1 | .000 |
| CLASS(2) | 158.585 | 1 | .000 |
| CLASS(3) | 12.398 | 1 | .000 |
| AGE(1) | 20.956 | 1 | .000 |
| SEX(1) | 456.874 | 1 | .000 |
| Overall Statistics | | 556.727 | 5 | .000 |

After Step 0, these are the variables that are not in the equation. The Wald Statistic is analogous to the t-statistic in linear regression: It tells whether variable will make a significant contribution to the equation. In this case, it describes the Constant, which has little value.

Roa’s Score statistic is analogous to the Wald statistic but is easier to calculate, and can be used to determine whether a variable will make a significant contribution to the equation if entered. In a forward stepwise analysis, the variable with the largest Score (assuming that it is significant) will be the next entered into the equation.

**Block 1: Method = Forward Stepwise (Likelihood Ratio)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Iteration Historya,b,c,d,e** | | | | | | | | |
| Iteration | | -2 Log likelihood | Coefficients | | | | | |
| Constant | SEX(1) | CLASS(1) | CLASS(2) | CLASS(3) | AGE(1) |
| Step 1 | 1 | 2343.253 | -1.152 | 2.080 |  |  |  |  |
| 2 | 2335.001 | -1.306 | 2.309 |  |  |  |  |
| 3 | 2334.988 | -1.313 | 2.317 |  |  |  |  |
| 4 | 2334.988 | -1.313 | 2.317 |  |  |  |  |
| Step 2 | 1 | 2251.508 | -1.544 | 1.989 | .450 | 1.155 | .460 |  |
| 2 | 2229.232 | -1.949 | 2.367 | .721 | 1.591 | .668 |  |
| 3 | 2228.913 | -2.010 | 2.420 | .776 | 1.657 | .705 |  |
| 4 | 2228.913 | -2.011 | 2.421 | .777 | 1.658 | .706 |  |
| 5 | 2228.913 | -2.011 | 2.421 | .777 | 1.658 | .706 |  |
| Step 3 | 1 | 2235.729 | -1.618 | 1.963 | .525 | 1.227 | .483 | .725 |
| 2 | 2210.506 | -2.074 | 2.356 | .847 | 1.697 | .713 | 1.010 |
| 3 | 2210.061 | -2.152 | 2.418 | .918 | 1.776 | .758 | 1.060 |
| 4 | 2210.061 | -2.154 | 2.420 | .920 | 1.778 | .760 | 1.062 |
| 5 | 2210.061 | -2.154 | 2.420 | .920 | 1.778 | .760 | 1.062 |

|  |
| --- |
| a. Method: Forward Stepwise (Likelihood Ratio) |
| b. Constant is included in the model. |
| c. Initial -2 Log Likelihood: 2769.457 |
| d. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001. |
| e. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001. |

This table provides all of the steps that are used in creating the models. In Step 1, Sex was entered into the equation, and the -2LL decreased from 2769.951, and adding the first variable into the equation reduced it to 2334.988, so there was substantial improvement. Class was the second variable added into the equation, and the addition of this variable further reduced the -2LL to 2228.913. Finally, adding Age as a third variable in the equation reduced the -2LL to 2210.061.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Omnibus Tests of Model Coefficients** | | | | |
|  | | Chi-square | df | Sig. |
| Step 1 | Step | 434.469 | 1 | .000 |
| Block | 434.469 | 1 | .000 |
| Model | 434.469 | 1 | .000 |
| Step 2 | Step | 106.075 | 3 | .000 |
| Block | 540.544 | 4 | .000 |
| Model | 540.544 | 4 | .000 |
| Step 3 | Step | 18.852 | 1 | .000 |
| Block | 559.396 | 5 | .000 |
| Model | 559.396 | 5 | .000 |

From: <http://psych.unl.edu/psycrs/statpage/logistic_eg1.pdf>

**Step**--tests the contribution of the specific variable(s)entered on each step

**Block**--tests the contribution of all the variables entered with this block

**Model**--tests the fit of the whole model. The model Chi-square is an analog of the F-test for regression.

In this case, we can see that the addition of each variable, starting in Step 1, makes a significant contribution to the equation. Step, Block, and Model will all be the same for a model with a single set of predictors that are entered simultaneously.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Summary** | | | |
| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
| 1 | 2334.988a | .179 | .250 |
| 2 | 2228.913b | .218 | .304 |
| 3 | 2210.061b | .224 | .314 |

|  |
| --- |
| a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001. |
| b. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001. |

Cox and Snell R2 and Nagelkerke R2 are analogous to the linear regression R2 values. Cox and Snell ranges from 0 to less than 1, so this is not the preferred approach to providing this estimate. The Nagelkerke estimate is calculated in such a way as to be constrained between 0 and 1. It can be evaluated as indicating model fit; with a better model displaying a value closer to 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classification Tablea** | | | | | |
|  | Observed | | Predicted | | |
| SURVIVE | | Percentage Correct |
| No | Yes |
| Step 1 | SURVIVE | No | 1364 | 126 | 91.5 |
| Yes | 367 | 344 | 48.4 |
| Overall Percentage | |  |  | 77.6 |
| Step 2 | SURVIVE | No | 1364 | 126 | 91.5 |
| Yes | 367 | 344 | 48.4 |
| Overall Percentage | |  |  | 77.6 |
| Step 3 | SURVIVE | No | 1364 | 126 | 91.5 |
| Yes | 362 | 349 | 49.1 |
| Overall Percentage | |  |  | 77.8 |

|  |
| --- |
| a. The cut value is .500 |

The classification table indicates how well each model performed when used to predict the outcome variable. While the improvements in classification at each step appear small, they represent a significant improvement over the baseline model (constant only).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | | | | |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) | |
| Lower | Upper |
| Step 1a | SEX(1) | 2.317 | .120 | 375.438 | 1 | .000 | 10.147 | 8.027 | 12.827 |
| Constant | -1.313 | .059 | 498.414 | 1 | .000 | .269 |  |  |
| Step 2b | CLASS |  |  | 98.246 | 3 | .000 |  |  |  |
| CLASS(1) | .777 | .142 | 29.841 | 1 | .000 | 2.176 | 1.646 | 2.876 |
| CLASS(2) | 1.658 | .168 | 97.664 | 1 | .000 | 5.250 | 3.779 | 7.294 |
| CLASS(3) | .706 | .174 | 16.486 | 1 | .000 | 2.025 | 1.441 | 2.847 |
| SEX(1) | 2.421 | .139 | 303.037 | 1 | .000 | 11.261 | 8.574 | 14.790 |
| Constant | -2.011 | .119 | 283.739 | 1 | .000 | .134 |  |  |
| Step 3c | CLASS |  |  | 108.243 | 3 | .000 |  |  |  |
| CLASS(1) | .920 | .149 | 38.344 | 1 | .000 | 2.510 | 1.875 | 3.358 |
| CLASS(2) | 1.778 | .172 | 107.370 | 1 | .000 | 5.917 | 4.227 | 8.282 |
| CLASS(3) | .760 | .176 | 18.556 | 1 | .000 | 2.138 | 1.513 | 3.020 |
| AGE(1) | 1.062 | .244 | 18.924 | 1 | .000 | 2.891 | 1.792 | 4.664 |
| SEX(1) | 2.420 | .140 | 297.068 | 1 | .000 | 11.247 | 8.541 | 14.809 |
| Constant | -2.154 | .127 | 288.037 | 1 | .000 | .116 |  |  |

|  |
| --- |
| a. Variable(s) entered on step 1: SEX. |
| b. Variable(s) entered on step 2: CLASS. |
| c. Variable(s) entered on step 3: AGE. |

Again, the Wald Statistic is analogous to the t-statistic in linear regression. Tells whether variable makes a significant contribution to the equation. In each step, all variables in the equation make a significant contribution.

Exp(B) is the odds ratio, which is an indicator of the change in odds resulting from a unit change in the predictor. With categorical variables, this is the change in odds relative to the base category. Odds = P(event) / P(no event). An Exp(B) value greater that 1 indicates that as the predictor variable increases, so do the odds of survival. Conversely, and Exp(B) value less than 1 indicates that as the predictor variable increases, the odds of survival decrease.

If we look at the results from Step 3, we see that women were 11.247 times more likely to survive than males. Likewise, children were 2.891 times more likely to survive than adults. When examining Class, each of the class designations is compared to the baseline class, or 3rd class. We see that Class1 (Crew) was 2.51 times more likely to survive than 3rd class, 2nd class was 2.138 times more likely to survive than 3rd class, and 1st class passengers were 5.917 times more likely to survive than 3rd class. Consequently, an adult male in 3rd class had little chance of survival. If you are on a cruise ship that hits an iceberg, your best chance of survival would be to present yourself as a rich little girl.

The 95% C.I.for EXP(B) are important to examine, and it’s important that this interval does not cross 1. If the 95% CI values cross 1 (ie, one value less than 1, one value greater than 1) then the relationship between the predictor and outcome becomes muddied, meaning that sometimes it can increase the odds ratio, and sometimes it can decrease the odds ratio. My preference is to remove variables with this characteristic from the equation.

The column labeled “B” represents the regression weights used to create the prediction. In binary logistic regression, the equation predicts the probability of group membership. A prediction ranging between 0.0-0.49 would place the subject in the group designated as “0” (in this case, the Did Not Survive group), and a prediction between 0.5-1.0 would place the subject in the group designated with a “1”.

The equation to predict probability is: .

The equation developed from Step 3 would be:



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model if Term Removed** | | | | | |
| Variable | | Model Log Likelihood | Change in -2 Log Likelihood | df | Sig. of the Change |
| Step 1 | SEX | -1384.728 | 434.469 | 1 | .000 |
| Step 2 | CLASS | -1167.494 | 106.075 | 3 | .000 |
| SEX | -1294.278 | 359.643 | 1 | .000 |
| Step 3 | CLASS | -1164.547 | 119.034 | 3 | .000 |
| AGE | -1114.456 | 18.852 | 1 | .000 |
| SEX | -1281.486 | 352.911 | 1 | .000 |

This table indicates whether removing a variable from the equation would have a significant detrimental impact on the model. In this case, removing any of the variables from the equation in Step 3 would result in a loss of predictive power.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables not in the Equation** | | | | | |
|  | | | Score | df | Sig. |
| Step 1 | Variables | CLASS | 104.698 | 3 | .000 |
| CLASS(1) | 2.010 | 1 | .156 |
| CLASS(2) | 73.936 | 1 | .000 |
| CLASS(3) | .142 | 1 | .707 |
| AGE(1) | 6.025 | 1 | .014 |
| Overall Statistics | | 123.072 | 4 | .000 |
| Step 2 | Variables | AGE(1) | 19.417 | 1 | .000 |
| Overall Statistics | | 19.417 | 1 | .000 |

The Variables Not in the Equation table provides an indication of the order in which variables were selected for entry. Again, Roa’s Score statistic is analogous to the Wald statistic but is easier to calculate, and can be used to determine whether a variable will make a significant contribution to the equation if entered. In a forward stepwise analysis, the variable with the largest Score (assuming that it is significant) will be the next entered into the equation. This table indicates that Class would be entered into the equation before Age, which is what the output reflects.

**Step number: 1**

Observed Groups and Predicted Probabilities

2000 + +

I I

I Y I

F I Y I

R 1500 + Y +

E I N I

Q I N I

U I N I

E 1000 + N +

N I N I

C I N I

Y I N I

500 + N Y +

I N Y I

I N Y I

I N N I

Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------

Prob: 0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1

Group: NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

Predicted Probability is of Membership for Yes

The Cut Value is .50

Symbols: N - No

Y - Yes

Each Symbol Represents 125 Cases.

**Step number: 2**

Observed Groups and Predicted Probabilities

1600 + +

I I

I I

F I I

R 1200 + +

E I I

Q I I

U I Y I

E 800 + Y +

N I N I

C I N I

Y I Y N I

400 + N N +

I N N I

I N NN Y Y I

I N NN N N Y Y I

Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------

Prob: 0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1

Group: NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

Predicted Probability is of Membership for Yes

The Cut Value is .50

Symbols: N - No

Y - Yes

Each Symbol Represents 100 Cases.

**Step number: 3**

Observed Groups and Predicted Probabilities

1600 + +

I I

I I

F I I

R 1200 + +

E I I

Q I I

U I Y I

E 800 + Y +

N I N I

C I N I

Y I Y N I

400 + N N +

I N N I

I N N N Y Y Y I

I N N N N N Y Y I

Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------

Prob: 0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1

Group: NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

Predicted Probability is of Membership for Yes

The Cut Value is .50

Symbols: N - No

Y - Yes

Each Symbol Represents 100 Cases.

With one binary categorical variable in the equation (Step 1), cases are all categorized into one of two probabilities. As more variables are entered into the equation, the dispersion of predictions becomes more apparent. Even though the overall accuracy of the equation does not appear to improve between steps 1 and 3 (as determined by the percent correctly classified), the ability to differentiate between cases improves. These plots can sometimes provide insight into setting the cutoff value to a value that can improve classification. (The cutoff value is the point at which cases transition from being classified in group 0 to group 1. The default is 0.5). In this example, 0.5 happens to work fairly well.