Hi all,

I’m using Python to get a feel for the data. This is all I have so far as I am having trouble getting the age data. Clearly gender and class matter. I think age and family members will also matter but I don’t have that data yet. I’ll add more as I get it.

The number of first class women was 94 and 91 survived. (97%)

The number of first class men was 122 and 45 survived. (37%)

The number of second class women was 76 and 70 survived. (92%)

The number of second class men was 108 and 17 survived. (16%)

The number of third class women was 144 and 72 survived. (50%)

The number of third class men was 347 and 47 survived. (13.5%)

The number of females on board was 314.

The number of males on board was 577.

644 passengers embarked from S and 217 survived.

77 passengers embarked from Q and 30 survived.

168 embarked from C and 93 survived.

Test set statistics

The number of passengers in the test set is 418

The number of women is 152

The number of men is 266

The percent of passengers who were women is 36.3636363636 in the test set, 35.24% in the training set.

The number of first class passengers was 107 and the percent was 25.59 in the test set and 24.24% in the training set.

The number of second class passengers was 93 and the percent was 22.24 in the test set and 20.65 % in the training set.

The number of third class passengers was 218 and the percent was 52.15 in the test set and 55.1% in the training set.

Test set Training set

pclass sex pclass sex number of passengers

1 M 57 1 1 122

F 50 2 94

2 M 63 2 M 108

F 30 F 76

3 M 146 3 M 347

F 72 F 144

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Test set |  |  |  |  | Training set |  |  |  |  |
| Pclass | Sex | Number | ratio | % of total | Pclass | Sex | Number | ratio | % of total |
| 1 | M | 57 | 1.14 | 13.6 | 1 | M | 122 | 1.3 | 13.7 |
| 1 | F | 50 |  | 12 | 1 | F | 94 |  | 10.5 |
| 2 | M | 63 | 1.05 | 15 | 2 | M | 108 | 1.42 | 12.1 |
| 2 | F | 30 |  | 7.2 | 2 | F | 76 |  | 8.5 |
| 3 | M | 146 | 2 | 35 | 3 | M | 347 | 2.4 | 38.9 |
| 3 | F | 72 |  | 17 | 3 | F | 144 |  | 16.6 |

Using the Newton Raphson method for survived vs pclass, sex, age, embarked and fare I get:

The accuracy of the model is 0.784511784512

The confusion matrix is

True0 is 454 and false0 is 97

False1 is 95 and true1 is 245

Using the Newton Raphson method for survived vs pclass, sex, age, embarked and fare and all the sibling data I get:

The accuracy of the model is 0.811447811448

The confusion matrix is

True0 is 486 and false0 is 105

False1 is 63 and true1 is 237

**————— 11/2/2012 3:21:17 PM ————————————————————**

Welcome to Minitab, press F1 for help.

**Binary Logistic Regression: survived versus sex**

Link Function: Logit

Response Information

Variable Value Count

survived 1 342 (Event)

0 549

Total 891

Logistic Regression Table

Odds 95% CI

Predictor Coef SE Coef Z P Ratio Lower Upper

Constant 1.05659 0.128987 8.19 0.000

sex

male -2.51371 0.167178 -15.04 0.000 0.08 0.06 0.11

Log-Likelihood = -458.902

Test that all slopes are zero: G = 268.851, DF = 1, P-Value = 0.000

\* NOTE \* No goodness of fit test performed.

\* NOTE \* The model uses all degrees of freedom.

Measures of Association:

(Between the Response Variable and Predicted Probabilities)

Pairs Number Percent Summary Measures

Concordant 109044 58.1 Somers' D 0.53

Discordant 8829 4.7 Goodman-Kruskal Gamma 0.85

Ties 69885 37.2 Kendall's Tau-a 0.25

Total 187758 100.0

**Delta Chi-Square versus P**

**Delta Deviance versus P**

**Binary Logistic Regression: survived versus sex, pclass**

Step Log-Likelihood

0 -593.328

1 -420.581

2 -413.751

3 -413.598

4 -413.598

5 -413.598

Link Function: Logit

Response Information

Variable Value Count

survived 1 342 (Event)

0 549

Total 891

Factor Information

Factor Levels Values

sex 2 female, male

Logistic Regression Table

Odds 95% CI

Predictor Coef SE Coef Z P Ratio Lower Upper

Constant 3.29464 0.297434 11.08 0.000

sex

male -2.64340 0.183826 -14.38 0.000 0.07 0.05 0.10

pclass -0.960553 0.106055 -9.06 0.000 0.38 0.31 0.47

Log-Likelihood = -413.598

Test that all slopes are zero: G = 359.459, DF = 2, P-Value = 0.000

Goodness-of-Fit Tests

Method Chi-Square DF P

Pearson 26.8200 3 0.000

Deviance 29.0987 3 0.000

Hosmer-Lemeshow 26.5123 3 0.000

Brown:

General Alternative 25.6666 2 0.000

Symmetric Alternative 6.5029 1 0.011

Table of Observed and Expected Frequencies:

(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

Group

Value 1 2 3 4 5 Total

1

Obs 47 17 45 72 161 342

Exp 33.7 23.7 51.6 86.7 146.3

0

Obs 300 91 77 72 9 549

Exp 313.3 84.3 70.4 57.3 23.7

Total 347 108 122 144 170 891

Measures of Association:

(Between the Response Variable and Predicted Probabilities)

Pairs Number Percent Summary Measures

Concordant 143877 76.6 Somers' D 0.67

Discordant 18892 10.1 Goodman-Kruskal Gamma 0.77

Ties 24989 13.3 Kendall's Tau-a 0.32

Total 187758 100.0

**Delta Chi-Square versus P**

**Delta Deviance versus P**

**Binary Logistic Regression: survived versus sex, pclass, age**

Step Log-Likelihood

0 -592.358

1 -412.064

2 -403.368

3 -403.114

4 -403.113

5 -403.113

Link Function: Logit

Response Information

Variable Value Count

survived 1 342 (Event)

0 547

Total 889

Factor Information

Factor Levels Values

sex 2 female, male

\* NOTE \* 889 cases were used

\* NOTE \* 2 cases contained missing values

Logistic Regression Table

Odds 95% CI

Predictor Coef SE Coef Z P Ratio Lower Upper

Constant 4.68521 0.449551 10.42 0.000

sex

male -2.60488 0.186392 -13.98 0.000 0.07 0.05 0.11

pclass -1.15955 0.118748 -9.76 0.000 0.31 0.25 0.40

age -0.0324900 0.0073580 -4.42 0.000 0.97 0.95 0.98

Log-Likelihood = -403.113

Test that all slopes are zero: G = 378.489, DF = 3, P-Value = 0.000

Goodness-of-Fit Tests

Method Chi-Square DF P

Pearson 358.694 279 0.001

Deviance 352.301 279 0.002

Hosmer-Lemeshow 24.593 7 0.001

Brown:

General Alternative 14.885 2 0.001

Symmetric Alternative 4.322 1 0.038

Table of Observed and Expected Frequencies:

(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

Group

Value 1 2 3 4 5 6 7 8 9 Total

1

Obs 19 14 15 9 35 46 49 79 76 342

Exp 13.9 9.2 12.1 20.4 34.3 50.4 57.3 72.1 72.3

0

Obs 163 78 74 79 54 49 39 9 2 547

Exp 168.1 82.8 76.9 67.6 54.7 44.6 30.7 15.9 5.7

Total 182 92 89 88 89 95 88 88 78 889

Measures of Association:

(Between the Response Variable and Predicted Probabilities)

Pairs Number Percent Summary Measures

Concordant 157454 84.2 Somers' D 0.69

Discordant 27691 14.8 Goodman-Kruskal Gamma 0.70

Ties 1929 1.0 Kendall's Tau-a 0.33

Total 187074 100.0

**Delta Chi-Square versus P**

**Delta Deviance versus P**

**Binary Logistic Regression: survived versus sex, pclass, age, embarked**

Step Log-Likelihood

0 -590.443

1 -408.739

2 -399.531

3 -399.244

4 -399.244

5 -399.244

Link Function: Logit

Response Information

Variable Value Count

survived 1 340 (Event)

0 547

Total 887

Factor Information

Factor Levels Values

sex 2 female, male

embarked 3 C, Q, S

\* NOTE \* 887 cases were used

\* NOTE \* 4 cases contained missing values

Logistic Regression Table

Odds 95% CI

Predictor Coef SE Coef Z P Ratio Lower Upper

Constant 5.03634 0.477557 10.55 0.000

sex

male -2.56641 0.187615 -13.68 0.000 0.08 0.05 0.11

pclass -1.15143 0.124806 -9.23 0.000 0.32 0.25 0.40

age -0.0332755 0.0074399 -4.47 0.000 0.97 0.95 0.98

embarked

Q 0.0002567 0.368253 0.00 0.999 1.00 0.49 2.06

S -0.520822 0.229437 -2.27 0.023 0.59 0.38 0.93

Tests for terms with more than 1 degree of freedom

Term Chi-Square DF P

embarked 7.15381 2 0.028

Log-Likelihood = -399.244

Test that all slopes are zero: G = 382.400, DF = 5, P-Value = 0.000

Goodness-of-Fit Tests

Method Chi-Square DF P

Pearson 430.167 364 0.010

Deviance 418.100 364 0.026

Hosmer-Lemeshow 21.378 8 0.006

Brown:

General Alternative 15.266 2 0.000

Symmetric Alternative 3.871 1 0.049

Table of Observed and Expected Frequencies:

(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

Group

Value 1 2 3 4 5 6 7 8 9 10 Total

1

Obs 14 11 8 20 16 36 56 63 87 29 340

Exp 8.3 8.5 10.5 14.8 24.7 39.9 61.1 64.7 79.0 28.6

0

Obs 113 83 83 68 73 53 46 25 2 1 547

Exp 118.7 85.5 80.5 73.2 64.3 49.1 40.9 23.3 10.0 1.4

Total 127 94 91 88 89 89 102 88 89 30 887

Measures of Association:

(Between the Response Variable and Predicted Probabilities)

Pairs Number Percent Summary Measures

Concordant 157675 84.8 Somers' D 0.70

Discordant 27212 14.6 Goodman-Kruskal Gamma 0.71

Ties 1093 0.6 Kendall's Tau-a 0.33

Total 185980 100.0

**Delta Chi-Square versus P**

**Delta Deviance versus P**