Titanic survivor rates statistical analysis
Name: rtm Spring 2011

1. Goal

We want to examine the probability of survival of a titanic passenger as a function of applicable variables.
The variables of greatest interest are:
- gender
- cabin class (from best to worst: 1st, 2nd, 3rd)
- age (limited to child VS adult)
- crew/passenger status

We can also examine the relevance of the following variables:
- member of a family present on board VS single
- place of embarkation (Southampton, Cherbourg, Queenstown (Cobh) in Ireland)

Data acquired from: http://lib.stat.cmu.edu/S/Harrell/data/descriptions/titanic.html

2. Method

Problems of the given type are in the field of Categorical Data Analysis. Alan Agresti has comprehensive works on the subject. The approach to use in the everything-binary case is Loglinear Regression. The basic (2-dimensional table of independent RVs) model for the loglinear regression is:

$$\ln(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

Label the marginal probability of being in the i-th column of the table p_i; marginal probability of being in the j-th row p_j; probability of being in box i, j: p_{ij}. In the case of independence

$$p_{ij} = p_i p_j.$$

Using frequencies instead of probabilities, if there are E_{ij} entries in each column, and N total entries, then $E_{ij} = np_{ij} = Np_i p_j$. Take logs to obtain

$$\ln(E_{ij}) = \ln(N) + \ln(p_i) + \ln(p_j)$$

Set $\mu_{ij} = E_{ij}, \lambda = \ln(N), \lambda_i^X = \ln(p_i), \lambda_j^Y = \ln(p_j)$ to get the original equation.
To account for dependence, one can add terms of the type λ_{ij}^{XY}, which counts the number of successes for the $(X, Y) = (i, j)$ case.

Also applicable is the Logistic Regression technique, which is tailored for the binary outcome model (where $Y_i \in \{0, 1\}$). The normality of error assumption is impractical when the outcomes are $0, 1$.
The logistic model transforms the usual linear regression equation

$$Y = X\beta + \epsilon$$

to

$$Y = \frac{\exp(X\beta)}{1 + \exp(X\beta)} + \epsilon.$$

The choice of function stems from the fact that the domain of $f(x) = \frac{e^x}{1+e^x}$ is $(0, 1)$, with 0 the infimum to the left and 1 the supremum to the right.
We can rewrite the model as

$$logit(\pi) = \ln(\frac{\pi(x)}{1 - \pi(x)}) = X\beta + \hat{\epsilon}$$

One can always resort to the Linear Probability Model

$$\pi(x) = \alpha + \beta x.$$

An obvious caveat for this model is that probabilities lie in the $[0, 1]$ interval, while linear functions range over the real line.

3. Background

The $269.1m$ long, $46,328$ ton decadent Titanic embarked on her maiden voyage at Southampton on 10 April 1912, and sank on the night of 14-15 April 1912 after colliding with an iceberg. Only 706 of the 2223 people on board survived.
Notes on the sinking:
-The iceberg hit the liner at 23:40pm. The Titanic sank 2hrs 40mins later at 2:20am. The officers, unwilling to cause undue alarm, took a half hour to order the dropping of the lifeboats. The first lifeboat was dropped at 00:40am. Thus, there wasn't enough time for a proper evacuation. Even the insufficiently many lifeboats on board were not filled to capacity.
-Swimming in the cold Atlantic waters meant death by hypothermia in 15-20 minutes.
-In its final moments, the ship broke in two under its own weight.
-Titanic's band played during the sinking. None of the members survived.
-Captain Edward Smith sank with his ship on what some claim was his last voyage before retirement.
-The Carpathia arrived in the area at 4:10am.

Could the disaster have been avoided?
-The outlook lacked binoculars after a quarters assignment mix-up.
-Due to its sturdy design, the Titanic would have withstood a head-on collision with the iceberg. In his unsuccessful attempt to evade the obstacle, the Captain compromised five of the ship's watertight compartments. The ship had been designed to survive up to 4 exposed compartments.
-There weren't enough lifeboats. (Though due to the rapidity of the sinking, even the available boats did not fill.)
-Titanic had no distress rockets. At the time of the sinking, the lookouts on the nearby SS Californian noticed lights, but since none of the lights were red, the Californian crew assumed the rockets were fired in celebration rather than distress.
-Californian's operator was asleep during the sinking.

4. Statistical results

It is convenient to construct a table for the data. The four-way table is:

```
, , ageliteral = child, sexliteral = female

        classliteral
survived crew 1st 2nd 3rd
       0    0   0   0  17
       1    0   1  13  14

, , ageliteral = adult, sexliteral = female

        classliteral
survived crew 1st 2nd 3rd
       0    3   4  13  89
       1   20 140  80  76

, , ageliteral = child, sexliteral = male

        classliteral
survived crew 1st 2nd 3rd
       0    0   0   0  35
       1    0   5  11  13

, , ageliteral = adult, sexliteral = male

        classliteral
survived crew 1st 2nd 3rd
       0  670 118 154 387
       1  192  57  14  75
```

Survival according to age group table:

```
        age
survived child adult
       0    52  1438
       1    57   654
```

Survival VS gender:

```
        gender
survived female male
       0    126 1364
       1    344  367
```

Survival VS Class

```
        class
survived crew 1st 2nd 3rd
       0  673 122 167 528
       1  212 203 118 178
```

One can immediately establish the general trends in the data from the tables alone. The regression will then associate numerics to the relationships between the variables.
Running a Generalized Linear Model right away yields:

```
Call:  glm(formula = titanic2$survived ~ titanic2$pclass + titanic2$age +
    titanic2$sex)

Coefficients:
       (Intercept)  titanic2$pclass1st  titanic2$pclass2nd  titanic2$pclass3rd
           0.89877             0.17555            -0.01053            -0.13118
 titanic2$ageadult    titanic2$sexmale
          -0.18130            -0.49068
Degrees of Freedom: 2200 Total (i.e. Null);  2195 Residual
Null Deviance:      481.3
Residual Deviance: 359.6        AIC: 2273
```

Running logit yields

```
Call:
glm(formula = survived ~ class1st + class2nd + class3rd + classcrew +
    agechild + ageadult + sexmale + sexfemale, family = binomial(logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0812  -0.7149  -0.6656   0.6858   2.1278

Coefficients: (3 not defined because of singularities)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1862     0.1586   7.481 7.40e-14 ***
class1st      0.8577     0.1573   5.451 5.00e-08 ***
class2nd     -0.1604     0.1738  -0.923    0.356
class3rd     -0.9201     0.1486  -6.192 5.93e-10 ***
classcrew         NA         NA      NA       NA
agechild      1.0615     0.2440   4.350 1.36e-05 ***
ageadult          NA         NA      NA       NA
sexmale      -2.4201     0.1404 -17.236  < 2e-16 ***
sexfemale         NA         NA      NA       NA
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2769.5  on 2200  degrees of freedom
Residual deviance: 2210.1  on 2195  degrees of freedom
AIC: 2222.1
```

```
Number of Fisher Scoring iterations: 4
```

The 'NA's stem from the linear dependency of the respective columns. The default case (to which the intercept is associated) is crew-member-adult-female. As one can see, in the general no inter-variable-dependencies case, everything except 2nd class belonging is significant. By allowing various correlations, we can obtain better fits. For example

```
Call:
glm(formula = survived ~ -1 + class1st * sexmale + class2nd *
    sexmale + class3rd + agechild + sexmale, family = binomial(logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6780  -0.7038  -0.5905   0.5336   2.0228

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
class1st          3.55788    0.50706   7.017 2.27e-12 ***
sexmale          -1.26918    0.07535 -16.843  < 2e-16 ***
class2nd          1.87728    0.29737   6.313 2.74e-10 ***
class3rd         -0.38895    0.10788  -3.605 0.000312 ***
agechild          1.05304    0.22904   4.598 4.27e-06 ***
class1st:sexmale -2.96391    0.53631  -5.527 3.27e-08 ***
sexmale:class2nd -2.51558    0.37582  -6.694 2.18e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3051.2  on 2201  degrees of freedom
Residual deviance: 2157.8  on 2194  degrees of freedom
AIC: 2171.8

Number of Fisher Scoring iterations: 6
```

Testing with all class-age and class-sex dependencies included produces a poor fit:

```
Call:
glm(formula = survived ~ class1st * sexmale + class2nd * sexmale +
    class3rd * sexmale + classcrew * sexmale + class1st * agechild +
    class2nd * agechild + class3rd * agechild + classcrew * agechild,
    family = binomial(logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6771  -0.7099  -0.6057   0.2374   2.2293
```

```
Coefficients: (4 not defined because of singularities)
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.89712    0.61914   3.064 0.002183 **
class1st             1.65823    0.80030   2.072 0.038264 *
sexmale             -3.14690    0.62453  -5.039 4.68e-07 ***
class2nd            -0.08004    0.68757  -0.116 0.907325
class3rd            -2.11453    0.63702  -3.319 0.000902 ***
classcrew                 NA         NA      NA       NA
agechild             0.33791    0.26920   1.255 0.209391
class1st:sexmale    -1.13608    0.82048  -1.385 0.166162
sexmale:class2nd    -1.06807    0.74658  -1.431 0.152539
sexmale:class3rd     1.76160    0.65159   2.704 0.006860 **
sexmale:classcrew         NA         NA      NA       NA
class1st:agechild   16.51217  858.44954   0.019 0.984654
class2nd:agechild   17.28628  367.06861   0.047 0.962439
class3rd:agechild         NA         NA      NA       NA
classcrew:agechild        NA         NA      NA       NA
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2769.5  on 2200  degrees of freedom
Residual deviance: 2099.2  on 2190  degrees of freedom
AIC: 2121.2

Number of Fisher Scoring iterations: 15
```

Running a simple loglin

```
titanic.loglin<-loglin(titanic.table, margin=list(1,2,3,4), param=TRUE)
```

gives us

```
$lrt
[1] 1243.663

$pearson
[1] 1637.445

$df
[1] 25

$margin
```

```
$margin[[1]]
[1] "survived"

$margin[[2]]
[1] "classliteral"

$margin[[3]]
[1] "ageliteral"

$margin[[4]]
[1] "sexliteral"


$param
$param$`(Intercept)`
[1] 3.015185

$param$survived
         0          1
 0.3699295 -0.3699295

$param$classliteral
      crew        1st        2nd        3rd
 0.5902083 -0.4115541 -0.5428901  0.3642359

$param$ageliteral
    child     adult
-1.477264  1.477264

$param$sexliteral
    female       male
-0.6518609  0.6518609
```

Attempting to get the dependencies (

```
titanic.loglin<-loglin(titanic.table, margin=list(c(1,2,3,4)), param=TRUE)
```

) gives me errors.

5. Analysis

The results indicate that:
-The most significant predictors are class belonging and gender. Having a ticket to a first class cabin markedly increases chances of survival, while being male results in a slightly lower but still large in magnitude decrease in the probability of a positive outcome.
-Next, a second class ticket increases one's chance of survival compared to the average; as does being a child.

-While women and children in 1st and 2nd classes have probability of survival close to 1, the men fare poorly. Second class men survived at a ration lower than that of 3rd class or crew member men.
-2nd class men had the worst survival rates (about 0.08), while 1st and 2nd class children and 1st class women survived at a rate of almost 1.
-3rd class children faced better odds than their parents, but the class factor meant that only about a third of them survived.
-Of the crew, most women (20/23) survived, while most men (about 4/5) perished.

6. Conclusions

One was more likely to survive the Titanic debacle if one was 1) a woman, 2) first class, 3) a child. Follows a brief analysis of each factor.
All children and most of the women in first and second class survived. Furthermore, the women and children among the third class passengers and the crew fared significantly better than the grown men in third class and the crew.

The reason for the high survival rate among women and children is the "Women and children first" doctrine that was a semi-official part of maritime law in 1912. This chivalric practice arose from the events of the sinking of HMS Birkenhead in 1852. The Birkenhead sinking is interesting in itself. The ship was the first British ironclad. At the time of its sinking it was a troopship. The soldiers on board stood firm and sank with the ship to allow the women and children to evacuate. Only 193 of 643 people survived.
One should mention that for all its chivalry, the "Women and children first" motto was isolated to high-profile disasters such as the sinking of the Titanic. British society in 1912 was anti-women and anti-children. Women only gained the vote in 1928. British children, including members of the upper class, were routinely beaten up and otherwise abused, including at school, well into the second half of the 20th century.
Another interesting sinking was that of the Lusitania in 1915 by a German U-boat. The Lusitania sank in 18 minutes. There was no "Women and children first" at the Lusitania. (Reference:

```
 http://www.sciencenews.org/view/generic/id/56817/title/
Titanic_study_It_takes_time_to_do_the_right_thing
```

).

The other major predictor for survivability of the Titanic disaster is class membership. To be fair to the ship's officers, they did not follow an explicit policy of favor toward the upper class passengers. The policy they did follow, other than "Women and children first", was "First come first serve". (Gleicher talks briefly about this.) Since the lifeboats were located on the first and second class decks, the adopted policy amounted to upper class preference. The third class passengers, down in the bowels of the ship, took longer to understand the gravity of the situation. The stewards' efforts to keep 3rd class passengers in line with the idea of conducting an orderly evacuation prevented many from reaching the lifeboats.
Survivability among 3rd class passengers and the crew was about the same. The crewmembers were working class and many of them were, like the 3rd class passengers, relegated

to the deep recesses of the ship, from which escape was unlikely (the doomed engine room workers spring to mind).

7. References

Data: http://lib.stat.cmu.edu/S/Harrell/data/descriptions/titanic.html
Categorical Data Analysis books: Agresti - An Introduction to Categorical Data Analysis, Second Edition, Wiley 2007; and Categorical Data Analysis, Wiley 2002.
General book on regression: Freud, Wilson, Sa - Regression Analysis, Academic Press 2006
Example of doing log-linear fits on R:

`http://www.stat.washington.edu/quinn/classes/536/S/loglinexample.html`

Paper on the subject: David Gleicher - Who Survived the Titanic? A Logistic Regression Analysis, Published in International Review of Maritime History, XVI(2), December 2004, pp. 61-94