The sinking of the *Titanic*

The logistic regression model is a member of a general class of models called *log–linear models*. These models are particularly useful when studying contingency tables (tables of counts). Such tables occur when observations are cross–classified using several categorical variables (contingency tables are sometimes called cross–classifications). The logistic regression form is then appropriate if one of the categorical variables takes on two values and can be viewed as a target variable. For example, in clinical trials, whether the patient lives or dies is a reasonable target variable, and different categorical variables could be potential predictors (for example, gender, membership in the treatment or control group, presence or absence of certain symptoms, etc.).

Here's an example of a contingency table of this form. One of the most famous maritime disasters occurred during the maiden voyage of the ocean liner *Titanic*, which struck an iceberg in the North Atlantic and sank on April 15, 1912. Many articles, books, and movies have told the story of this disaster, but a relatively straightforward statistical analysis tells the story in a remarkably evocative way. The table below summarizes the mortality experiences of the 2201 people on board the ocean liner, given as survival percentages of the number of people of certain subgroups at risk. People are separated by Gender, Age (child or adult) and Economic status (first class, second class, third class [steerage], or crew).

Gender		**Age**	
Male			
		Adult	*Child*
	First class	32.6% of 175	100% of 5
Economic status	*Second class*	8.3% of 168	100% of 11
	Third class	16.2% of 462	27.1% of 48
	Crew	22.3% of 862	—
Female			
	First class	97.2% of 144	100% of 1
Economic status	*Second class*	86.0% of 93	100% of 13
	Third class	46.1% of 165	45.2% of 31
	Crew	87.0% of 23	—

Note that there were no children among the crew.

Since the predictors are all categorical, tables summarize marginal relationships with survival, as follows (these take the place of the side–by–side boxplots used for continuous predictors; note that this also applies to 0/1 predictors).

Economic status	Percent survived
First class	62.5% of 325
Second class	41.4% of 285
Third class	25.2% of 706
Crew	24.0% of 885

Age	Percent survived
Child	52.3% of 109
Adult	31.3% of 2092

Gender	Percent survived
Female	73.2% of 470
Male	21.2% of 1731

The chance of survival was apparently related to all three of these factors. Mortality was much higher for men than for women, and higher for adults than for children. This is of course consistent with the "rule of the sea," which says that women and children should be saved first in a disaster. The observed survival percentage is directly related to economic status, with higher status associated with higher survival probability, and the crew having survival rates similar to those in steerage.

Three–dimensional contingency tables allow us to assess the possibility of interaction effects among the predictors. These can be presented as two–way tables, with survival percentages given in each cell. First, the interaction of Economic status and Gender:

	Gender	
Economic status	*Female*	*Male*
First class	97.2% of 145	34.4% of 180
Second class	87.7% of 106	14.0% of 179
Third class	45.9% of 196	17.3% of 510
Crew	87.0% of 23	22.3% of 862

The most striking pattern here is the difference between Third class and the others. While for the other three status levels mortality was much higher among men than among women, for steerage passengers the difference is much smaller, with less than half of the

women surviving (that is, steerage female survival percentage is considerably lower than would be expected from the main effects alone).

The following table summarizes the interaction of Economic status and Age:

	Age	
Economic status	*Child*	*Adult*
First class	100.0% of 6	61.8% of 319
Second class	100.0% of 24	36.0% of 261
Third class	34.2% of 79	24.1% of 627
Crew	—	24.0% of 885

This interaction also makes clear the different nature of Third class compared with the others; while no children of the other classes died, almost two–thirds of those in steerage did.

Finally, the following table represents the interaction of Age and Gender:

	Age	
Gender	*Child*	*Adult*
Female	62.2% of 45	74.4% of 425
Male	45.3% of 64	20.3% of 1667

Adult women had a higher survival rate than girl children did, but for men the survival rate was twice as high for children than for adults.

We can use logistic regression to try to decide which of these potential effects are useful to build a model predicting survival probability accurately. The following table summarizes the properties of the models considered. All of the models are hierarchical, in that the presence of an interaction effect in the model implies that the associated main effects are also present. Such a requirement is usually sensible from an intuitive point of view. It also has advantages in estimation and testing, since for nonhierarchical models estimates and tests for interaction terms can change depending on how effects are coded (as indicator variables or effect codings). Since there were no children in the crew, the interaction between economic status and age (EA) is fit using only two of the codings corresponding to pairwise products of those for the main effects, rather than three.

As an example, consider the output from a model using only Economic status as a predictor:

```
Binary Logistic Regression: Survived versus Economic status

Method

Link function                 Logit
Categorical predictor coding  (-1, 0, +1)
Residuals for diagnostics     Pearson
Rows used                     14


Response Information

Variable  Value      Count  Event Name
Survived  Event        711  Survived
          Non-event   1490
At risk   Total       2201


Deviance Table

Source             DF  Adj Dev  Adj Mean  Chi-Square  P-Value
Regression          3    180.9     60.30      180.90    0.000
  Economic status   3    180.9     60.30      180.90    0.000
Error              10    491.1     49.11
Total              13    672.0


Model Summary

Deviance   Deviance
    R-Sq  R-Sq(adj)      AIC
  26.92%     26.47%  2596.56


Coefficients

Term                 Coef  SE Coef   VIF
Constant          -0.5201   0.0508
Economic status
  Crew            -0.6350   0.0754  1.66
  First class      1.0293   0.0956  1.85
```

```
  Second class    0.1728   0.0991  1.89


Odds Ratios for Categorical Predictors

Level A          Level B       Odds Ratio       95% CI
Economic status
  First class    Crew             5.2822  (4.0224, 6.9366)
  Second class   Crew             2.2431  (1.6923, 2.9731)
  Third class    Crew             1.0702  (0.8507, 1.3463)
  Second class   First class      0.4246  (0.3067, 0.5880)
  Third class    First class      0.2026  (0.1529, 0.2685)
  Third class    Second class     0.4771  (0.3568, 0.6380)

Odds ratio for level A relative to level B


Regression Equation

P(Survived)  =  exp(Y')/(1 + exp(Y'))


Y' = -0.5201 - 0.6350 Economic status_Crew + 1.0293 Economicstatus_First class
     + 0.1728 Economic status_Second class - 0.5672 Economicstatus_Third class


Goodness-of-Fit Tests

Test             DF  Chi-Square  P-Value
Deviance         10      491.06    0.000
Pearson          10      446.38    0.000
Hosmer-Lemeshow   2        0.00    1.000


Measures of Association

Pairs        Number  Percent  Summary Measures       Value
Concordant   539216     50.9  Somers D                0.28
Discordant   239042     22.6  Goodman-Kruskal Gamma   0.39
Ties         281132     26.5  Kendalls Tau-a          0.12
Total       1059390    100.0

Association is between the response variable and predicted probabilities
```

Economic status is entered as a factor variable, being fit using three effect coding variables (recall that the default in `Minitab` is to use indicator variables, but this can be changed by clicking on the `Coding` button). Estimated odds ratios are given for each of the pairs of economic status levels, with 95% confidence intervals for the odds ratio based on Wald tests provided (an interval not including 1 indicates statistically significant differences in the odds of survival of the two levels). The simultaneous test of the significance of economic status (corresponding to a partial F–test in ANOVA modeling) is highly statistically significant, equaling 180.9 on 3 degrees of freedom ($p < .001$).

The deviance and Pearson goodness-of-fit tests, which certainly are valid here, both strongly indicate lack of it, an unsurprising result, given that age and gender are not in the model. We would like to compare the different possible models, but since we have categorical predictors here we can't even use least squares best subsets to make it easier. The solution is to bite the bullet and fit the different possible models by hand, recording summary statistics in order to compare them. Corrected AIC (AIC_C) is technically not valid in the logistic regression context, but it still seems to be a useful way of trading off fit versus complexity; as always, $AIC_C = AIC + 2(\nu + 1)(\nu + 2)/(n - \nu - 2)$, where ν is the number of estimated regression coefficients in the model.

The table below summarizes exploration of different models. It gives summary statistics for every possible model, including the LR statistic assessing the strength of the regression, the number of predictors in the model, the Somers' D statistic, the Deviance goodness–of–fit p–value, and the AIC value (with $n = 2201$ the AIC and AIC_C values are virtually identical). A "good" model would have higher LR (implying strong fit), lower number of predictors (implying a simple model), higher D (implying stronger association with survival), high enough deviance p (implying a good fit), and lower AIC. The models are listed based on decreasing order of LR within models of the same type (one main effect, two main effects and one interaction, and so on).

Model	LR	Predictors	D	Deviance p	AIC
G	434.5	1	.40	.000	2339.0
E	180.9	3	.28	.000	2596.6
A	19.6	1	.05	.000	2753.9
E, G	540.5	4	.49	.000	2238.9
A, G	440.4	2	.42	.000	2335.1
E, A	206.5	4	.30	.000	2573.0
E, A, G	559.4	5	.52	.000	2222.1
E, G, EG	605.7	7	.50	.000	2179.7
A, G, AG	456.7	3	.43	.000	2320.8
E, A, EA	235.7	6	.30		2547.7
E, A, G, EG	626.1	8	.53	.000	2161.4
E, A, G, EA	595.1	7	.52	.000	2222.1
E, A, G, AG	577.4	6	.52	.000	2206.0
E, A, G, EA, EG	670.3	10	.53	.021	2121.2
E, A, G, EG, AG	634.7	9	.53	.000	2154.8
E, A, G, EA, AG	606.9	8	.53	.000	2180.5
E, A, G, EA, EG, AG	672.0	11	.54	.037	2121.5

According to the deviance test none of the models fit the table adequately, but the models {E, A, G, EA, EG} and {E, A, G, EA, EG, AG} come closest. The four models with smallest values of AIC include those two and {E, A, G, EG} and {E, A, G, EG, AG}. According to AIC the best model is {E, A, G, EA, EG}. Another way of choosing among these models is to compare their fitted values. These are given on the next page for the three simplest models of the four mentioned here. The fitted survival percentages are similar for the three models for the adult classes, but differ for the child classes. Since children represent less than 5% of the total population at risk, the simple model {E, A, G, EG} might be adequate to describe most of the important associations with survival in the data, although it can be noted that the {E, A, G, EA, EG} model also provides a better fit to survival percentages of adult men from Second class and adult women from Third class than do the other models.

		E, A, G, EG	
Gender		**Age**	
Male			
		Adult	*Child*
	First class	33.7% of 175	59.4% of 5
Economic status	*Second class*	12.9% of 168	29.9% of 11
	Third class	15.5% of 462	34.4% of 48
	Crew	22.3% of 862	—
Female			
	First class	97.2% of 144	99.0% of 1
Economic status	*Second class*	86.7% of 93	94.9% of 13
	Third class	41.9% of 165	67.4% of 31
	Crew	87.0% of 23	—

		E, A, G, EA, EG	
Gender		**Age**	
Male			
		Adult	*Child*
	First class	32.6% of 175	100% of 5
Economic status	*Second class*	8.3% of 168	100% of 11
	Third class	16.8% of 462	22.0% of 48
	Crew	22.3% of 862	—
Female			
	First class	97.2% of 144	100% of 1
Economic status	*Second class*	86.0% of 93	100% of 13
	Third class	44.6% of 165	53.0% of 31
	Crew	87.0% of 23	—

		E, A, G, EG, AG	
Gender		**Age**	
Male			
		Adult	*Child*
	First class	33.4% of 175	70.0% of 5
Economic status	*Second class*	12.3% of 168	39.5% of 11
	Third class	14.5% of 462	44.1% of 48
	Crew	22.3% of 862	—
Female			
	First class	97.2% of 144	97.7% of 1
Economic status	*Second class*	87.5% of 93	89.4% of 13
	Third class	45.2% of 165	49.7% of 31
	Crew	87.0% of 23	—

Here is a summary of the results of fitting the {E, A, G, EG} model:

Binary Logistic Regression: Survived versus Economic status, Age group, Gender

Method

Link function	Logit
Categorical predictor coding	(-1, 0, +1)
Residuals for diagnostics	Pearson
Rows used	14

Response Information

Variable	Value	Count	Event Name
Survived	Event	711	Survived
	Non-event	1490	
At risk	Total	2201	

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	8	626.06	78.258	626.06	0.000
Economic status	3	155.50	51.832	155.50	0.000
Age group	1	20.34	20.339	20.34	0.000
Gender	1	290.49	290.493	290.49	0.000
Economic status*Gender	3	66.67	22.222	66.67	0.000
Error	5	45.90	9.180		
Total	13	671.96			

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
93.17%	91.98%	2161.39

Coefficients

Term	Coef	SE Coef	VIF
Constant	0.711	0.153	
Economic status			
Crew	0.139	0.249	16.58

```
  First class            1.257    0.220   7.11
  Second class          -0.199    0.174   3.90
Age group
  Adult                 -0.527    0.115   1.08
Gender
  Female                 1.567    0.115   2.78
Economic status*Gender
  Crew Female            0.007    0.249  14.05
  First class Female     0.550    0.220   7.84
  Second class Female    0.325    0.174   4.52


Odds Ratios for Categorical Predictors

Level A          Level B    Odds Ratio       95% CI
Economic status
  Any level      Any level           *  (     *,      *)
Age group
  Child          Adult          2.8681  (1.8259, 4.5053)
Gender
  Any level      Any level           *  (     *,      *)

Odds ratio for level A relative to level B
Odds ratios are not calculated for predictors that are included in interaction
     terms because these ratios depend on values of the other predictors in the
     interaction terms.


Regression Equation

P(Survived)  =  exp(Y')/(1 + exp(Y'))


Y' = 0.711 +0.139Economicstatus_Crew +1.257Economicstatus_First class
     -0.199Economicstatus_Second class -1.197Economicstatus_Third class
     -0.527Agegroup_Adult +0.527Agegroup_Child +1.567Gender_Female
     -1.567Gender_Male +0.007Economicstatus*Gender_Crew Female
     -0.007Economicstatus*Gender_Crew Male
     +0.550Economicstatus*Gender_First class Female
     -0.550Economicstatus*Gender_First class Male
     +0.325Economicstatus*Gender_Second class Female
     -0.325Economicstatus*Gender_Second class Male
     -0.882Economicstatus*Gender_Third class Female
     +0.882Economicstatus*Gender_Third class Male
```

```
Goodness-of-Fit Tests

Test             DF  Chi-Square  P-Value
Deviance          5       45.90    0.000
Pearson           5       42.77    0.000
Hosmer-Lemeshow   3        0.40    0.941


Measures of Association

Pairs        Number  Percent  Summary Measures       Value
Concordant   722426     68.2  Somers D               0.53
Discordant   161300     15.2  Goodman-Kruskal Gamma   0.63
Ties         175664     16.6  Kendalls Tau-a         0.23
Total       1059390    100.0

Association is between the response variable and predicted probabilities
```

The fitted coefficients correspond to the patterns noted earlier: on the *Titanic* it was better to be female, better to be a child, and better to be of higher economic status; given this, women in steerage did worse than expected.

The model that adds EA does noticeably better than either of the other models for male children, correctly noting that all of the boys in first class and second class survived, while few of those in third class did; thus, in order to correctly summarize the pattern for these 64 passengers, the {E, A, G, EA, EG} model is best.

You might wonder about regression diagnostics for these data. As was noted earlier, when a logistic regression model is based only on categorical predictors (including a model where all predictors are indicator variables), the data actually take the form of a contingency table. In that context diagnostics for individual observations (for the Titanic data, individual passengers) are not relevant, since it is the number of observations that fall into a particular cell of the table that is either unusual or not, and hence it is a cell, not an observation, that might be outlying. This means that if the data are given at the level of individual observations regression diagnostic are not meaningful. If the data are given

at the level of cells of the table (that is, combinations of the predictor values), as is the case here, regression diagnostics can be meaningful. Comparison of the original table of observed proportions (as given on page 1) to a table of estimated probabilities (as in the three given on page 8) corresponds to an examination of residuals in this situation.

Minitab commands

Including an interaction term in a logistic regression model is done the same way it is done for ANOVA and ANCOVA models, by clicking on `Model` and then highlighting and adding interactions as needed. Note that, just in the case of analysis of variance and covariance models, interactions are only sensibly defined if *at most* one term in the interaction is a numerical variable.

If you have data at the 0/1 response level, and wish to construct tables of the relationships between categorical predictors and the response, click on `Stat` → `Tables` → `Cross Tabulation and Chi-Square`. Enter the variables under `Categorical variables:`, with the 0/1 response variable under `For rows`, with the predictor under `For columns`. Click on `Counts` and `Column percents`.

If your data are already summarized into covariate patterns, with target variables representing number of successes and number of trials for each covariate pattern, constructing tables of the relationships is more complicated. Say you have three categorical predictors `A`, `B`, and `C`, success count variable `Successes`, and trials count variable `Trials`. To construct the table corresponding to any effect (main effect or interaction effect), first fit a logistic regression model to the table based on that main effect or interaction, and save the estimated probabilities; these will be the observed proportions of successes. The estimated probabilities will correspond to the observed proportions. Note that this won't work if any of the cells in the table of covariate patterns is empty, as was the case in the Titanic data (there were no crew who were children, so the Economic status × Age group proportions can't be found this way). To get the total number of trials for each cell of the subtable you're interested in, click on `Stat` → `Tables` → `Cross Tabulation and Chi-Square`. Enter the variables determining the table under `Categorical variables:`

(`A B` in the example above). Click in the box next to `Frequencies are in:`, and enter the variable with the number of trials in the box (`Trials` in the example above).

If you want to fit a logistic regression model to data where some combinations of factor variables are missing, you need to fit the model manually using indicator or effect coding variables, just as is true for ANOVA and ANCOVA models. To do this, first create all of the needed variables for main effects and interaction effects. Then, go to `Stat` → `Regression` → `Regression`, and fit an ordinary least squares regression using those variables as predictors. `Minitab` will fit the regression, and note that some variables were dropped out (corresponding to the empty cells). Now, go back to `Stat` → `Regression` → `Binary Logistic Regression`, and fit the logistic regression using the same variables as were used in the least squares fit. You'll need to construct the tests of significance for each effect manually as well, just as you would need to do for partial F–tests in an ANOVA fit. To get an effect, fit the model that includes all of the predictors except those that correspond to the effect you're interested in. Subtract the `LR` statistic for this model from the `LR` statistic for the model that includes all of the variables. This statistic (which must be nonnegative) is the likelihood ratio statistic for testing the effect; compare it to a χ^2 distribution, with the degrees of freedom being the number of variables that define the effect, to get a tail probability for testing the hypothesis that the effect adds nothing given the others.