Multinomial Logistic Regression Models

Polytomous responses. Logistic regression can be extended to handle responses that are *polytomous*, i.e. taking $r > 2$ categories. (Note: The word *polychotomous* is sometimes used, but this word does not exist!) When analyzing a polytomous response, it's important to note whether the response is *ordinal* (consisting of ordered categories) or *nominal* (consisting of unordered categories). Some types of models are appropriate only for ordinal responses; other models may be used whether the response is ordinal or nominal. If the response is ordinal, we do not necessarily have to take the ordering into account, but it often helps if we do. Using the natural ordering can

- lead to a simpler, more parsimonious model and
- increase power to detect relationships with other variables.

If the response variable is polytomous and all the potential predictors are discrete as well, we could describe the multiway contingency table by a loglinear model. But fitting a loglinear model has two disadvantages:

- It has many more parameters, and many of them are not of interest. The loglinear model describes the joint distribution of all the variables, whereas the logistic model describes only the conditional distribution of the response given the predictors.
- The loglinear model is more complicated to interpret. In the loglinear model, the effect of a predictor X on the response Y is described by the XY association. In a logit model, however, the effect of X on Y is a main effect.

If you are analyzing a set of categorical variables, and one of them is clearly a "response" while the others are predictors, I recommend that you use logistic rather than loglinear models.

Grouped versus ungrouped. Consider a medical study to investigate the long-term effects of radiation exposure on mortality. The response variable is

$$Y = \begin{cases} 1 & \text{if alive,} \\ 2 & \text{if dead from cause other than cancer,} \\ 3 & \text{if dead from cancer other than leukemia,} \\ 4 & \text{if dead from leukemia.} \end{cases}$$

The main predictor of interest is level of exposure (low, medium, high). The data could arrive in ungrouped form, with one record per subject:

```
low  4
med  1
med  2
high 1
 .
 .
 .
```

Or it could arrive in grouped form:

Exposure	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$
low	22	7	5	0
medium	18	6	7	3
high	14	12	9	9

In ungrouped form, the response occupies a single column of the dataset, but in grouped form the response occupies r columns. Most computer programs for polytomous logistic regression can handle grouped or ungrouped data.

Whether the data are grouped or ungrouped, we will imagine the response to be multinomial. That is, the "response" for row i,

$$y_i = (y_{i1}, y_{i2}, \ldots, y_{ir})^T,$$

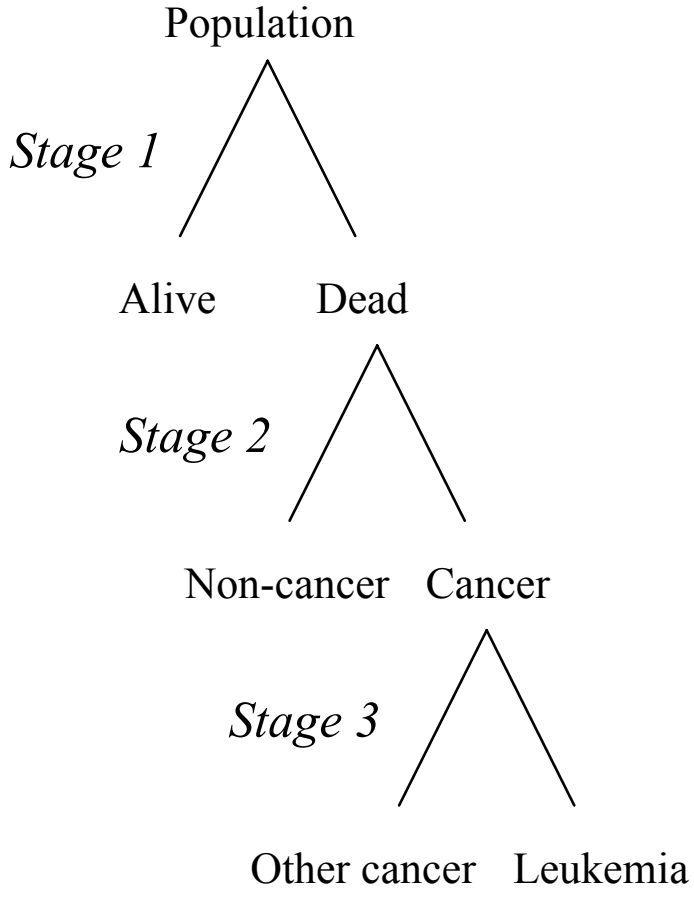
is assumed to have a multinomial distribution with index $n_i = \sum_{j=1}^{r} y_{ij}$ and parameter

$$\pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{ir})^T.$$

If the data are grouped, then n_i is the total number of "trials" in the ith row of the dataset, and y_{ij} is the number of trials in which outcome j occurred. If the data are ungrouped, then y_i has a 1 in the position corresponding to the outcome that occurred and 0's elsewhere, and $n_i = 1$. Note, however, that if the data are ungrouped, we do not have to actually create a dataset with columns of 0's and 1's; a single column containing the response level $1, 2, \ldots, r$ is sufficient.

Describing polytomous responses by a sequence of binary models. In some cases, it makes sense to "factor" the response into a sequence of binary choices and model them with a sequence of ordinary logistic models.

For example, consider the study of the effects of radiation exposure on mortality. The four-level response can be modeled in three stages:



The stage 1 model, which is fit to all subjects, describes the log-odds of death.

The stage 2 model, which is fit only to the subjects that die, describes the log-odds of death due to cancer versus death from other causes.

The stage 3 model, which is fit only to the subjects who die of cancer, describes the log-odds of death due to leukemia versus death due to other cancers.

Because the multinomial distribution can be factored into a sequence of conditional binomials, we can fit these three logistic models separately. The overall likelihood function factors into three independent likelihoods.

This approach is attractive when the response can be naturally arranged as a sequence of binary choices. But in situations where arranging such a sequence is unnatural, we should probably fit a single multinomial model to the entire response.

Baseline-category logit model. Suppose that

$$y_i = (y_{i1}, y_{i2}, \ldots, y_{ir})^T$$

has a multinomial distribution with index $n_i = \sum_{j=1}^{r} y_{ij}$ and parameter

$$\pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{ir})^T.$$

When the response categories $1, 2, \ldots, r$ are **unordered**, the most popular way to relate π_i to covariates is through a set of $r-1$ baseline-category logits. Taking j^* as the baseline category, the model is

$$\log\left(\frac{\pi_{ij}}{\pi_{ij^*}}\right) = x_i^T \beta_j, \quad j \neq j^*.$$

If x_i has length p, then this model has $(r-1) \times p$ free parameters, which we can arrange as a matrix or a vector. For example, if the last category is the baseline ($j^* = r$), the coefficients are

$$\beta = [\beta_1, \beta_2, \ldots, \beta_{r-1}]$$

or

$$\text{vec}(\beta) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{r-1} \end{bmatrix}.$$

Comments on this model

- The kth element of β_j can be interpreted as: the increase in log-odds of falling into category j versus category j^* resulting from a one-unit increase in the kth covariate, holding the other covariates constant.
- Removing the kth covariate from the model is equivalent to simultaneously setting $j-1$ coefficients to zero.
- Any of the categories can be chosen to be the baseline. The model will fit equally well, achieving the same likelihood and producing the same fitted values. Only the values and interpretation of the coefficients will change.

- To calculate π_i from β, the back-transformation is

$$\pi_{ij} = \frac{\exp(x_i^T \beta_j)}{1 + \sum_{k \neq j^*} \exp(x_i^T \beta_k)}$$

for the non-baseline categories $j \neq j^*$, and the baseline-category probability is

$$\pi_{ij^*} = \frac{1}{1 + \sum_{k \neq j^*} \exp(x_i^T \beta_k)}.$$

Model fitting. This model is not difficult to fit by Newton-Raphson or Fisher scoring. PROC LOGISTIC can do it.

Goodness of fit. If the estimated expected counts $\hat{\mu}_{ij} = n_i \hat{\pi}_{ij}$ are large enough, we can test the fit of our model versus a saturated model that estimates π independently for $i = 1, \ldots, N$. The deviance for comparing this model to a saturated one is

$$G^2 = 2 \sum_{i=1}^{N} \sum_{j=1}^{r} y_{ij} \log \frac{y_{ij}}{\mu_{ij}}.$$

The saturated model has $N(r-1)$ free parameters and the current model has $p(r-1)$, where p is the

length of x_i, so the degrees of freedom are

$$df = (N-p)(r-1).$$

The corresponding Pearson statistic is

$$X^2 = \sum_{i=1}^{N}\sum_{j=1}^{r} r_{ij}^2,$$

where

$$r_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$$

is the Pearson residual. If the model is true, both are approximately distributed as χ^2_{df} provided that

- no more than 20% of the μ_{ij}'s are below 5.0, and
- none are below 1.0.

In practice this is often not satisfied, so there may be no way to assess the overall fit of the model.

However, we may still apply a χ^2 approximation to ΔG^2 and ΔX^2 to compare nested models, provided that $(N-p)(r-1)$ is large relative to Δdf.

Overdispersion

Overdispersion means that the actual covariance matrix of y_i exceeds that specified by the multinomial

model,

$$V(y_i) = n_i \left[\text{Diag}(\pi_i) - \pi_i \pi_i^T \right].$$

It is reasonable to think that overdispersion is present if

- the data are grouped (n_i's are greater than 1),
- x_i already contains all covariates worth considering, and
- the overall X^2 is substantially larger than its degrees of freedom $(N-p)(r-1)$.

In this situation, it may be worthwhile to introduce a scale parameter σ^2, so that

$$V(y_i) = n_i \sigma^2 \left[\text{Diag}(\pi_i) - \pi_i \pi_i^T \right].$$

The usual estimate for σ^2 is

$$\hat{\sigma}^2 = \frac{X^2}{(N-p)(r-1)},$$

which is approximately unbiased if $(N-p)(r-1)$ is large. Introducing a scale parameter does not alter the estimate of β (which then becomes a quasilikelihood estimate), but it does alter our

estimate of the variability of $\hat{\beta}$. If we estimate a scale parameter, we should

- multiply the estimated ML covariance matrix for $\hat{\beta}$ by $\hat{\sigma}^2$ (SAS does this automatically);
- divide the usual Pearson residuals by $\hat{\sigma}$; and
- divide the usual X^2, G^2, ΔX^2 and ΔG^2 statistics by $\hat{\sigma}^2$ (SAS reports these as scaled statistics).

These adjustments will have little practical effect unless the estimated scale parameter is substantially greater than 1.0 (say, 1.2 or higher).

Example. The table below, reported by Delany and Moore (1987), comes from a study of the primary food choices of alligators in four Florida lakes. Researchers classified the stomach contents of 219 captured alligators into five categories: Fish (the most common primary food choice), Invertebrate (snails, insects, crayfish, etc.), Reptile (turtles, alligators), Bird, and Other (amphibians, plants, household pets, stones, and other debris).

Let's describe these data by a baseline-category model, with Primary Food Choice as the outcome and Lake, Sex, and Size as covariates.

			Primary Food Choice				
Lake	*Sex*	*Size*	*Fish*	*Inv.*	*Rept.*	*Bird*	*Other*
Hancock	M	small	7	1	0	0	5
		large	4	0	0	1	2
	F	small	16	3	2	2	3
		large	3	0	1	2	3
Oklawaha	M	small	2	2	0	0	1
		large	13	7	6	0	0
	F	small	3	9	1	0	2
		large	0	1	0	1	0
Trafford	M	small	3	7	1	0	1
		large	8	6	6	3	5
	F	small	2	4	1	1	4
		large	0	1	0	0	0
George	M	small	13	10	0	2	2
		large	9	0	0	1	2
	F	small	3	9	1	0	1
		large	8	1	0	0	1

Because the usual primary food choice of alligators appears to be fish, we'll use fish as the baseline category; the four logit equations will then describe the log-odds that alligators select other primary food types instead of fish.

Entering the data. When the data are grouped, as

they are in this example, SAS expects the response categories $1, 2, \ldots, r$ to appear in a single column of the dataset, with another column containing the frequency or count. That is, the data should look like this:

```
Hancock  male     small fish     7
Hancock  male     small invert   1
Hancock  male     small reptile  0
Hancock  male     small bird     0
Hancock  male     small other    5
Hancock  male     large fish     4
Hancock  male     large invert   0
Hancock  male     large reptile  0
Hancock  male     large bird     1
Hancock  male     large other    2
    --lines omitted--
George   female   large bird     0
George   female   large other    1
```

The lines that have a frequency of zero are not actually used in the modeling, because they contribute nothing to the loglikelihood. You can include them if you want to, but it's not necessary.

Specifying the model. In the `model` statement, you need to tell SAS about the existence of a count or frequency variable; otherwise SAS will assume that the data are ungrouped, with each line representing a single alligator.

You also need to specify which of the categories is the baseline. The link function is `glogit`, for generalized logit.

To get fit statistics, include the options `aggregate` and `scale=none`.

```
options nocenter nodate nonumber linesize=72;
data gator;
   input lake $ sex $ size $ food $ count;
   cards;
Hancock  male     small fish     7
Hancock  male     small invert   1
Hancock  male     small reptile  0
Hancock  male     small bird     0
      -- lines omitted --
George   female   large other    1
;

proc logist data=gator;
   freq count;
   class lake size sex / order=data param=ref ref=first;
   model food(ref='fish') = lake size sex / link=glogit
      aggregate scale=none;
   run;
```

Here is the output pertaining to the goodness of fit.

Response Profile

Ordered Value	food	Total Frequency
1	bird	13
2	fish	94
3	invert	61
4	other	32
5	reptile	19

Logits modeled use food='fish' as the reference category.

NOTE: 24 observations having zero frequencies or weights were excluded since they do not contribute to the analysis.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > ChiSq
Deviance	40	50.2637	1.2566	0.1282
Pearson	40	52.5643	1.3141	0.0881

Number of unique profiles: 16

There are $N = 16$ profiles (unique combinations of lake, sex and size) in this dataset. The saturated model, which fits a separate multinomial distribution to each profile, has $16 \times 4 = 64$ free parameters. The current model has an intercept, three lake coefficients, one sex coefficient and one size coefficient for each of the four logit equations, for a total of 24 parameters. Therefore, the overall fit statistics have $64 - 24 = 40$ degrees of freedom.

Output pertaining to the significance of covariates:

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	66.4974	20	<.0001
Score	59.4616	20	<.0001
Wald	51.2336	20	0.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
lake	12	36.2293	0.0003
size	4	15.8873	0.0032
sex	4	2.1850	0.7018

The first section (global null hypothesis) tests the fit of the current model against a null or intercept-only model. The null model has four parameters (one for each logit equation). Therefore the comparison has $24 - 4 = 20$ degrees of freedom. This test is highly significant, indicating that at least one of the covariates has an effect on food choice.

The next section (Type III analysis of effects) shows the change in fit resulting from discarding any one of the covariates—lake, sex or size—while keeping the others in the model. For example, consider the test for lake. Discarding lake is equivalent to setting three coefficicients to zero in each of the four logit

equations; so the test for lake has $3 \times 4 = 12$ degrees of freedom. Judging from these tests, we see that

- lake has an effect on food choice;
- size has an effect on food choice; and
- sex does not have a discernible effect.

This suggests that we should probably remove sex from the model. We also may want to look for interactions between lake and size.

Here are the estimated coefficients:

Analysis of Maximum Likelihood Estimates

Parameter		food	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		bird	1	-2.4633	0.7739	10.1310	0.0015
Intercept		invert	1	-2.0744	0.6116	11.5025	0.0007
Intercept		other	1	-0.9167	0.4782	3.6755	0.0552
Intercept		reptile	1	-2.9141	0.8856	10.8275	0.0010
lake	Oklawaha	bird	1	-1.1256	1.1924	0.8912	0.3452
lake	Oklawaha	invert	1	2.6937	0.6692	16.2000	<.0001
lake	Oklawaha	other	1	-0.7405	0.7422	0.9956	0.3184
lake	Oklawaha	reptile	1	1.4008	0.8105	2.9872	0.0839
lake	Trafford	bird	1	0.6617	0.8461	0.6117	0.4341
lake	Trafford	invert	1	2.9363	0.6874	18.2469	<.0001
lake	Trafford	other	1	0.7912	0.5879	1.8109	0.1784
lake	Trafford	reptile	1	1.9316	0.8253	5.4775	0.0193
lake	George	bird	1	-0.5753	0.7952	0.5233	0.4694
lake	George	invert	1	1.7805	0.6232	8.1623	0.0043
lake	George	other	1	-0.7666	0.5686	1.8179	0.1776
lake	George	reptile	1	-1.1287	1.1925	0.8959	0.3439
size	large	bird	1	0.7302	0.6523	1.2533	0.2629
size	large	invert	1	-1.3363	0.4112	10.5606	0.0012

size	large	other	1	-0.2906	0.4599	0.3992	0.5275
size	large	reptile	1	0.5570	0.6466	0.7421	0.3890
sex	female	bird	1	0.6064	0.6888	0.7750	0.3787
sex	female	invert	1	0.4630	0.3955	1.3701	0.2418
sex	female	other	1	0.2526	0.4663	0.2933	0.5881
sex	female	reptile	1	0.6275	0.6852	0.8387	0.3598

Odds Ratio Estimates

Effect	food	Point Estimate	95% Wald Confidence Limits	
lake Oklawaha vs Hancock	bird	0.324	0.031	3.358
lake Oklawaha vs Hancock	invert	14.786	3.983	54.893
lake Oklawaha vs Hancock	other	0.477	0.111	2.042
lake Oklawaha vs Hancock	reptile	4.058	0.829	19.872
lake Trafford vs Hancock	bird	1.938	0.369	10.176
lake Trafford vs Hancock	invert	18.846	4.899	72.500
lake Trafford vs Hancock	other	2.206	0.697	6.983
lake Trafford vs Hancock	reptile	6.900	1.369	34.784
lake George vs Hancock	bird	0.563	0.118	2.673
lake George vs Hancock	invert	5.933	1.749	20.125
lake George vs Hancock	other	0.465	0.152	1.416
lake George vs Hancock	reptile	0.323	0.031	3.349
size large vs small	bird	2.076	0.578	7.454
size large vs small	invert	0.263	0.117	0.588
size large vs small	other	0.748	0.304	1.842
size large vs small	reptile	1.745	0.492	6.198
sex female vs male	bird	1.834	0.475	7.075
sex female vs male	invert	1.589	0.732	3.449
sex female vs male	other	1.287	0.516	3.211
sex female vs male	reptile	1.873	0.489	7.175

How do we interpret them? Recall that there are four logit equations to predict the log-odds of

- birds versus fish,
- invertebrates versus fish,
- other versus fish, and

- reptiles versus fish.

The intercepts give the estimated log-odds for the reference group lake=Hancock, size=small, sex=male. For example, the estimated log-odds of birds versus fish in this group is -2.4633; the estimated log-odds of invertebrates versus fish is -2.0744; and so on.

The lake effect is characterized by three dummy coefficients in each of the four logit equations. The estimated coefficient for the Lake Oklawaha dummy in the bird-versus-fish equation is -1.1256. This means that alligators in Lake Oklawaha are less likely to choose birds over fish than their colleagues in Lake Hancock are. In other words, fish appear to be less common in Lake Oklawaha than in Lake Hancock. The estimated odds ratio of $\exp(-1.1256) = 0.32$ is the same for alligators of all sex and sizes, because this is a model with main effects but no interactions.