

Models for Two-Way Contingency Tables

Recall in the last section Generalized Linear Models (GLMs) were introduced as an extension of the traditional linear model, it eases the assumptions in the following ways:

1. Drops the normality assumption
 - The response variable is allowed to follow any distribution of the exponential family (binomial, Poisson, negative binomial, gamma, multi-nomial, etc)
2. Assumes the variance of the response depends on a function of the mean, called a variance function
3. The mean of the population is allowed to depend on the linear combination of the predictors through a **link function, g**
 - The following link functions are used most often
 - **Logit** $\left[g(x) = \log\left(\frac{x}{1-x}\right) \right]$,
 - **Inverse** $[f(x) = 1/x]$,
 - **Inverse squared** $[f(x) = 1/x^2]$,
 - **Log** $[f(x) = \log(x)]$,
 - **Cumulative logit** (e.g. $[f(x_1, x_2, x_3) = [\log\left(\frac{x_1}{1-x_1}\right), \log\left(\frac{x_1+x_2}{1-x_1-x_2}\right)]]$),
 - **Identity** $[f(x) = x]$.
 - The below table specifies which function to use with each distribution

Distribution	Default Link Function
Normal	Identity
Poisson	Log
Binomial	Logit
Multinomial	Cumulative Logit
Gamma	Inverse

In SAS

- PROC GENMOD
 - The GENMOD is a procedure for analyzing generalized linear models
- PROC LOGISTIC
 - The LOGISTIC procedure is constructed for logistic regression and provides useful information as diagnostic plots, odds ratios and other measures specific to logistic regression models.
- PROC CATMOD

- The CATMOD procedure is a procedure designed to fit models to functions of categorical response variables.

All of these procedures report the deviance. PROC LOGISTIC reports AIC and BIC, and it can be calculated with information from PROC CATMOD.

Estimation in Generalized Linear Models

GLMs are estimated with the Maximum Likelihood (ML) method. **This chooses the value which makes the observed data the most probable (equivalent to the least squares method).**

Example: Let τ be the prevalence of a disease in some population. Suppose that a random sample of size 100 is selected and we observe $Y = 40$ individuals with the disease.

Use the data(Y) to obtain an estimate of $\tau_{\text{hat}}(Y)$, assuming τ has good statistical properties. By "good" it means the estimate has little to no bias and small variance.

In this example, if $\tau = .5$ then we would write the likelihood function as:

$$P_{\tau}(Y = 40) = {}_{100}C_{40} .5^{40} (1 - .5)^{100 - 40}$$

As a function of τ , the function $P_{\tau}(Y = ?)$ is called the likelihood function

Log-linear Models/Contingency Tables

Log linear models for contingency tables specify how the cell counts depend on the levels of categorical variables defining that table. Loglinear models treat all variables as symmetrically and attempt to model all important associations among them. In this sense, it's very similar to correlation analysis of continuous variables where the goal is to determine the patterns of dependence and independence among a set of variables.

Loglinear models are generalized linear models with Poisson response distribution, Log link function, and Identify variance function (for Poisson: Expected value = variance)

Data are represented in contingency tables as cell counts. The counts in the cells are assumed to follow a Poisson distribution. Loglinear models are used to model association patterns among categorical variables.

Log linear models are analogous to correlation analysis for normally distributed data, and are most appropriate when there is no clear distinction between response and explanatory variables.

Example of Contingency tables:

# accidents	0	1	2	3	4 or more
# men	80	61	13	1	0

		Clumsy	
		Yes	No
Colicky	Yes	128	20
	No	45	232

		Liver		No		Yes	
		Spleen		No	Yes	No	Yes
Heart	Lung						
	No	No		4	1	0	3
	Yes			3	1	0	0
Yes	No			5	4	0	2
	Yes			4	2	0	3

X(rows)/Y(columns)	Col 1	Col 2	Total
Row 1	n_{11}	n_{12}	n_{1+}
Row 2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Data can be through of arising by sampling from a population and classify each individual in one of the cell of the two-way cross-classification of the two binary responses it falls in. Each count is assumed to follow a Poisson distribution with expected frequencies:

X(rows)/Y(columns)	Col 1	Col 2	Total
Row 1	m_{11}	m_{12}	
Row 2	m_{21}	m_{22}	
Total			

If we fix n (condition on n) the counts in the four cells follow a multinomial distribution.

$$m_{ij} = n\pi_{ij}$$

X(rows)/Y(columns)	Col 1	Col 2	Total
Row 1	π_{11}	π_{12}	π_{1+}
Row 2	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

Loglinear models are constricted using the expected values (m_{ij}) rather than π_{ij} . The main distributional assumption is that n_{ij} follow a Poisson distribution with expectancy m_{ij} .

There are several different kind of loglinear models we can fit to the data above.

Saturated Model

A model that is as complicated as the number of observations. Such a model is over-specified ('less-than-full-rank-coding' or 'GLM coding'). The result does not reduce the complexity of the data and will give a 'perfect prediction'.

The $\{\lambda^X_i\}$ are called row effects, $\{\lambda^Y_j\}$ are called column effects and $\{\lambda^{XY}_{ij}\}$ are called interaction effects.

To solve a saturated model, the practice is to impose linear constraints on the parameters to reduce the number of parameters.

- ‘effect coding’

1. $\lambda_1^X + \lambda_2^X = \lambda_1^Y + \lambda_2^Y = 0$
2. $\lambda_{11}^{XY} + \lambda_{12}^{XY} = \lambda_{21}^{XY} + \lambda_{22}^{XY} = \lambda_{11}^{XY} + \lambda_{21}^{XY} = 0$

- ‘reference coding’ or ‘treatment coding’ – the default in SAS and R

1. $\lambda_2^X = \lambda_2^Y = 0$
2. $\lambda_{12}^{XY} = \lambda_{21}^{XY} = \lambda_{22}^{XY} = 0.$

X(rows)/Y(columns)	Col 1	Col 2
Row 1	$\exp\{\mu + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}\}$	$\exp\{\mu + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}\}$
Row 2	$\exp\{\mu + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}\}$	$\exp\{\mu + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}\}$

Depending on the coding the above table reduces to the following tables:

Under effect coding

X(rows)/Y(columns)	Col 1	Col 2
Row 1	$\exp\{\mu + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}\}$	$\exp\{\mu + \lambda_1^X - \lambda_1^Y - \lambda_{11}^{XY}\}$
Row 2	$\exp\{\mu - \lambda_1^X + \lambda_1^Y - \lambda_{11}^{XY}\}$	$\exp\{\mu - \lambda_1^X - \lambda_1^Y + \lambda_{11}^{XY}\}$

Under treatment coding

X(rows)/Y(columns)	Col 1	Col 2
Row 1	$\exp\{\mu + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}\}$	$\exp\{\mu + \lambda_1^X\}$
Row 2	$\exp\{\mu + \lambda_1^Y\}$	$\exp\{\mu\}$

Testing Goodness of Fit in Loglinear Models

As with any other model, the GoF can be evaluated by comparing the observed to the predicted values - or the current model to the saturated model. It is **not** always appropriate to simply subtract the observed and fitted values!

For a loglinear model, two goodness of fit statistics are commonly used:

1. Deviance Statistic

$$G^2(M) = 2 \sum_{\text{cell}} n_{\text{cell}} [\log(n_{\text{cell}}) - \log(\hat{m}_{\text{cell}})] = 2 \sum_{\text{cell}} n_{\text{cell}} \log\left(\frac{n_{\text{cell}}}{\hat{m}_{\text{cell}}}\right)$$

2. Pearson chi-squared statistic

$$Q(M) = \sum_{\text{cell}} \frac{(n_{\text{cell}} - \hat{m}_{\text{cell}})^2}{\hat{m}_{\text{cell}}}$$

Where n_{cell} is the observed count for a cell and \hat{m}_{cell} is the fitted cell frequency in model M. Under the assumption that model M is the right model, both $G^2(M)$ and $Q(M) \sim \chi^2(n - p)$, with p being the number of parameters in the model M.

Properties of the Goodness of Fit Statistic

1. When the model M holds, both statistics follow a chi-squared with the degrees of freedom equal to the number of cells minus the number of estimated parameters
2. The deviance (likelihood ratio) statistic can be used to test the difference between two **nested** models, M1 and M2

Model M1 is said to be **nested** in M2 ($M1 \subset M2$) if the parameters are a subset of the parameters in M2.

3. The statistic $G^2(M1 | M2) = G^2(M1) - G^2(M2)$ will follow a chi-squared with $p2 - p1$ DF where $p1$ and $p2$ are the number of linearly independent parameters in models M1 and M2, respectively

Saturated Loglinear Model for SxR Table

$$\log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}; \quad i = 1, 2, \dots, S \text{ and } j = 1, 2, \dots, R$$

As above, the $\{\lambda_i^X\}$ are called row effects, $\{\lambda_j^Y\}$ are called column effects and λ_{ij}^{XY} are called interaction effects.

The number of parameters in the above model will be $1 + S + R + RS$. The number of observations is $S \times R$, hence the model is over-specified. To reduce the number of parameters linear constraints are imposed on the parameters. For example, *reference coding* would be represented as:

$$\lambda_S^X = \lambda_R^Y = 0, \text{ or}$$

$$\lambda_{Si}^{XY} = \lambda_{iR}^{XY} \text{ for every } i \text{ and } j, \text{ or:}$$

$$\sum_{i=1}^S \lambda_i^X = 0 \text{ and } \sum_{j=1}^R \lambda_j^Y = 0$$

$$\sum_{i=1}^S \lambda_{ij}^{XY} = 0 \text{ for each } j; \text{ and}$$

$$\sum_{j=1}^R \lambda_{ij}^{XY} = 0 \text{ for each } i$$

With either of these two coding constraints, the effective number of parameters for the saturated model is:

$$1 + (S - 1) + (R - 1) + (R - 1)(S - 1) = SR$$

For any number of dimensions the number of parameters in the saturated log-linear model equals the number of cells in the table. The saturated model give "perfect prediction" since it has the same number of observations as parameters.