Biostatistics 140.754
Advanced Methods in Biostatistics IV

Jeffrey Leek

Assistant Professor
Department of Biostatistics
jleek@jhsph.edu

Tip + Paper

Tip Meet with seminar speakers. When you go on the job market face recognition is priceless. I met Scott Zeger at UW when I was a student. When I came for an interview I already knew him (and Ingo, and Rafa, and ....).

Paper of the Day: "Equilibrium points in n-person games" `http://www.pnas.org/content/36/1/48.full.pdf+html`

Fill out the survey so I can get some feedback on how the class is going so far: `http://www.surveymonkey.com/s/5VR275W`

Outline For Today

- Interactions
- Some special cases of glms (case-control studies, log-linear models for tables)

Many slides borrowed from Ken Rice, Jon Wakefield, and Brian Caffo

Binary data models

Binary responses are very common in most disciplines. The fundamental quantity to model is the probability $Y = 1$ given covariates x, where $Y|x \sim \mathit{Bernoulli}\{p(x)\}$.

Linear models will clearly not be a good idea in general since they do not constrain the probabilities to lie in (0,1), and Bernoulli responses have variance $p(x)\{1 - p(x)\}$ which is non-constant. We will spend the next few slides considering *logistic regression models*. We won't talk about *relative risk regression* - based on a log link

Binary data models †

Applications concentrate on *logistic regression*, in particular **linear logistic** regression of the form:

$$\mathrm{logit}(p(x_i)) = \log\left(\frac{p(x_i)}{(1-p(x_i))}\right) = x_i^T\boldsymbol{\beta}$$

or equivalently

$$p(x_i) = \mathrm{expit}(x_i^T\boldsymbol{\beta}) = \frac{e^{x_i^T\boldsymbol{\beta}}}{1+e^{x_i^T\boldsymbol{\beta}}}$$

"Default" estimating equations are:

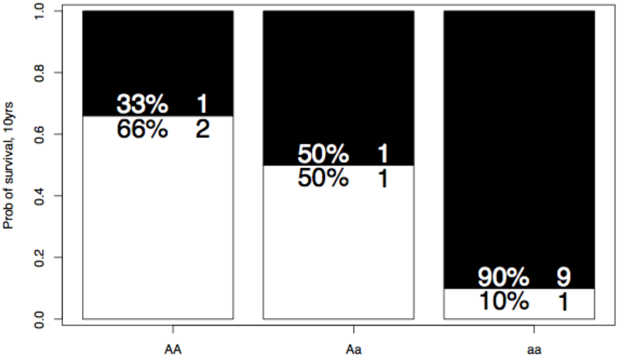
$$\sum_{i=1}^{n} X_i(Y_i - \mathrm{expit}(x_i^T\boldsymbol{\beta})) = \mathbf{0}$$

- these also define the MLE, assuming $Y_i \sim \textit{Bernoulli}(p(x_i))$. The β_j parameters estimates the **log odds ratio** associated with a one unit change in x_j (adjusted for all the other covariates). Also note this is one way to ensure that $p(x_i) \in [0,1]$.

But what are odds?

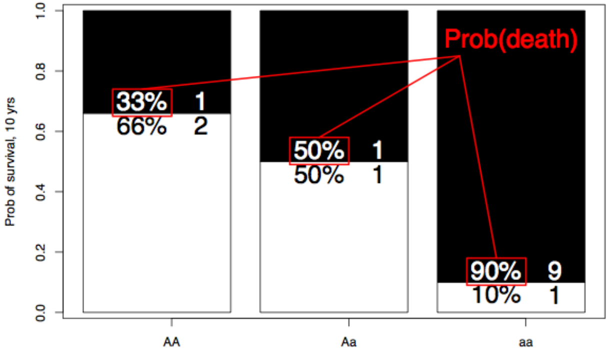
Odds †

Odds are a [gambling friendly] measure of chance:



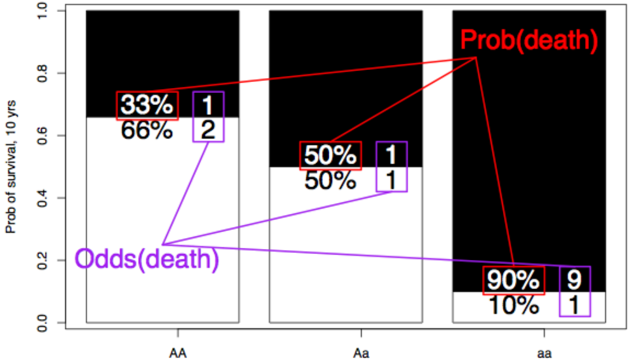
Odds †

Odds are a [gambling friendly] measure of chance:



Odds †

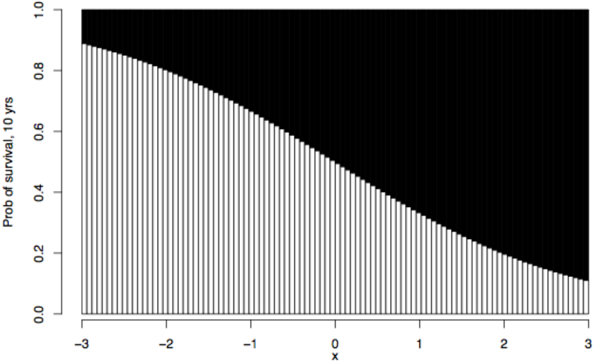
Odds are a [gambling friendly] measure of chance:



- so what are **odds ratios**?

Odds †

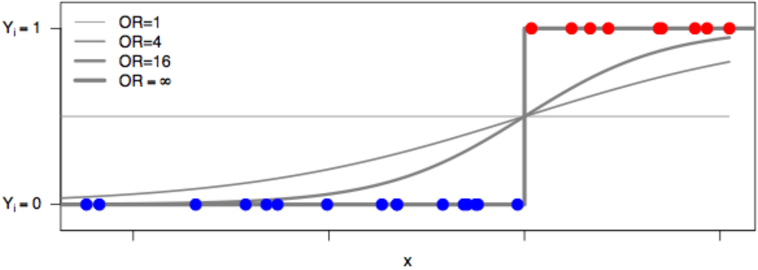
Odds ratio $= 2$, for each 1-unit difference in x;



Regression parameter $\hat{\beta}_x \approx \log(2)$, but report $e^{\hat{\beta}_x} \approx 2$

Odds †

Beware! The MLE exists, but need not be finite:



While rare in large samples, “perfect” or “complete separation” is not pathological. What does R do?

Linear model comparison

Suppose the "true" linear model is given by:

$$\mathbb{E}[Y|X,Z] = \beta_0 + \beta_1 X + \beta_2 Z$$

and we have the "reduced" model

$$\mathbb{E}[Y|X] = \beta_0^* + \beta_1^* X$$

- The addition of a covariate Z leads to a decrease in *bias* in the parameter of interest β_1
- If the covariate Z is independent of the covariate of interest X then its omission will not result in bias in $\hat{\beta}_1$.
- The addition of a covariate Z leads to a decrease in the *variance* of a parameter of interest, if Z is strongly associated with the outcome Y (by reducing the estimate of σ^2). If Z has a weak association then the variance may be increased.

Log-linear model comparison

For the log-linear model:

$$\mathbb{E}[Y|X,Z] = \text{exp}(\beta_0 + \beta_1 X + \beta_2 Z)$$

if we assume the model

$$\begin{aligned} \mathbb{E}[Y|X] &= \text{exp}(\beta_0^* + \beta_1^*) \\ &= \text{exp}(\beta_0 + \beta_1 X)\mathbb{E}[\text{exp}(\beta_2 Z|X)] \end{aligned}$$

then, in general, omission of Z will lead to bias, but if X and Z are independent then no bias results.

Note: if $Y|X,Z \sim \mathrm{Poisson}(\mathbb{E}[Y|X,Z])$ then $Y|X$ will no longer be Poisson, and in particular will display extra-Poisson variability.

Log-linear model comparison

For the logistic model

$$\mathbb{E}[Y|X,Z] = \frac{\exp(\beta_0 + \beta_1 X + \beta_2 Z)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 Z)}$$

if we assume the model:

$$\begin{aligned}\mathbb{E}[Y|X] &= \frac{\exp(\beta_0^* + \beta_1^* X)}{1 + \exp(\beta_0^* + \beta_1 X^*)} \\ &= \mathbb{E}_{z|x}\left[\frac{\exp(\beta_0 + \beta_1 X + \beta_2 Z)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 Z)}\right]\end{aligned}$$

then omission of Z will lead to a different estimate of β_1 even if X and Z are independent.

Summary measures for binary data

It is easier to think about *probability differences*, but unfortunately such differences are constrained (since probabilities $\in [0, 1]$.

Logistic regression models produce unconstrained parameters, but their interpretation is far more difficult since they produce summaries based on *odds ratios*

In the rare event situation, then odds ratios approximate ratios of probabilities and interpretation is more straightforward - Poisson approximation to binomial in rare disease case [1]

[1] Remember this fact from Rafa's course, I think?

Simpson's paradox

Consider the data in the tables below. For both men and women, the treatment appears beneficial, but when the data are collapsed over gender, the association is reversed. This occurs because of confounding.

	Men		Women	
	Diseased	Disease-free	Diseased	Disease-free
Control	8	2	9	21
Treatment	18	12	2	8
Odds Ratio	1.6		1.7	

	Diseased	Disease-free
Control	17	23
Treatment	20	20
Odds Ratio	0.7	

Simpson's paradox

Let

$$p_{xz} = \mathrm{Pr}(\mathrm{Y} = 1 | \mathrm{X} = \mathrm{x}, \mathrm{Z} = \mathrm{z})$$

$$q_x = \mathrm{Pr}(\mathrm{Z} = 1 | \mathrm{X} = \mathrm{x})$$

and

$$p_x^* = \mathrm{Pr}(\mathrm{Y} = 1 | \mathrm{X} = \mathrm{x})$$

for $x, z = 0, 1$. The "paradox" reflects the fact that it is possible to have:

$$p_{00} < p_{10}$$

and

$$p_{01} < p_{11}$$

but

$$p_{00}(1 - q_0) + p_{01} q_0 = p_0^* > p_1^* = p_{10}(1 - q_1) + p_{11} q_1$$

Simpson's paradox

In the example

$$\begin{aligned} Y &= \text{disease/disease} - \text{free} \\ X &= \text{control/treatment} \\ Z &= \text{male/female} \end{aligned}$$

$p_{00} = 2/10 = 0.20$
$p_{10} = 13/30 = 0.43$
$p_{01} = 21/30 = 0.7$
$p_{11} = 8/10 = 0.8$
$p_0^* = 23/40 = 0.58$
$p_1^* = 20/40 = 0.50$
$q_0 = 30/40$
$q_1 = 10/40$

- The paradox has nothing to do with small numbers
- The paradox cannot occur if $q_0 = q_1$ (i.e. there is no confounding)

Non-collapsibility

We now consider the situation where $q_0 = q_1$. This occurs, by construction in a randomized clinical trial in which (say) equal numbers of men and women receive the treatment.

	Men		Women	
	Diseased	Disease-free	Diseased	Disease-free
Control	95	5	10	90
Treatment	90	10	5	95
Odds Ratio	2.1		2.1	

	Diseased	Disease-free
Control	105	95
Treatment	95	105
Odds Ratio	1.2	

This displays the *non-collapsibility* of the odds ratio. Not to be confused with confounding.

Logistic regression analysis

Let p_{ij} represent the probability of disease for sex i and treatment j, we fit models

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + s_i + t_j$$

$i, j = 0, 1$ with $s_1, t_1 = 0$

```
> y <- c(95,90,10,5)
> z <- c(5,10,90,95)
> sex <- factor(c(0,0,1,1))
> tmt <- factor(c(0,1,0,1))
>options(contrasts = c("contr.treatment","contr.poly"))
>modss <- glm(cbind(y,z) ∼ sex + tmt, family = "binomial")

Estimate Std.  Error z value Pr(>|z|)
(Intercept) 2.9444 0.3763 7.824 5.12e-15 ***
sex1 -5.1417 0.4010 -12.821 < 2e-16 ***
tmt1 -0.7472 0.4010 -1.863 0.0624 .
---
Null deviance:  3.4508e+02 on 3 degrees of freedom
Residual deviance:  -1.4433e-14 on 1 degrees of freedom
> 1/exp(-0.747)
[1] 2.110659
> 1/exp(-5.14)
[1] 170.7158
```

The odds of disease are reduced by 2.1 for each gender when moving from control to treatment group, and are reduced by 170 (for each tmt group) when going from men to women.

Logistic regression analysis

Next we examine the marginal association

```
> modpa <- glm(cbind(y,z) ∼ tmt,family = "binomial")
> summary(modpa)
Coefficients:  Estimate Std.  Error z value Pr(>|z|)
(Intercept) 0.1001 0.1416 0.707 0.480
tmt1 -0.2002 0.2003 -1.000 0.318
Null deviance:  345.08 on 3 degrees of freedom
Residual deviance:  344.08 on 2 degrees of freedom
> 1/exp(-0.200)
[1] 1.221403
```

Now the odds of disease in the treatment group are 1.2 lower than in the control group.

Note: the standard error is smaller when gender is omitted (compare to linear regression)

Logistic regression analysis

In any table we have an “averaged” summary measure where the average is with respect to the population making up that table. So in the table for which we had equal numbers of men and women (which mimics a randomized trial) the odds ratio comparing treatment to control is the averaged effect, averaged across men and women (and any other variables that were unobserved). Such measures are known as *population averaged*

It has been argued that we should report the summary measure that is closest, in terms of covariates, to the individuals for whom we wish to make inference. Such a measure is known as *subject-specific*

When a logistic model, interpretation is more straightforward if we include additional covariates (that are associated with the response)

You know now, that the difference between PA and SS estimates shouldn’t really be referred to as “bias”, since the two models are estimating different quantities. Remember your superpopulation!!

Case-control studies

Logistic regression models offer advantages from a statistical perspective, but are also useful since odds ratios arise as summary measures from a range of sampling designs, and in particular from the case-control study.

Cohort studies investigate the causes of disease by proceeding from cause to effect; case-control studies proceed from effect to cause. In the simple case of a single binary exposure the following table demonstrates the form of the data that may be obtained. In a cohort study n_{1+} and n_{2+} are fixed.

	Diseased	Not-Diseased	
Exposed	n_{11}	n_{12}	n_{1+}
Not exposed	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n_{++}

It can also take a long time for the disease to develop (cancer is an obvious example) and so the study may have to run for a long period.

Case-control studies

The case-control (or retrospective) study provides a way of overcoming these difficulties. Study participants are now selected via their disease status. Those with the disease are cases, those without are controls. We then proceed backwards in time to determine the exposure level of the participants in the study. We are now fixing n_{+1} and n_{+2}

There are a number of disadvantages to the case-control study

- We can no longer estimate the probabilities of disease given exposure status without external information
- We have to select the participants carefully. The probability of selection for the study, for both cases and controls, must not depend on the exposure status, otherwise *selection bias* will be introduced

The great benefit is that, we can still estimate the strength of the relationship between exposure and disease.

Case-control studies

Consider the situation where we have a binary response Y taking the values $0/1$ corresponding to disease-free/diseased, and a vector of exposures/risk factors contained in X. We suppose that X is a vector of length $(k+1) \times 1$, where the first element is 1, and the remaining entries correspond to the k exposures/risk factors. In the case-control scenario, we select individuals on the basis of Y and the random variables are X. We are interested in

$$\mathrm{Pr(Y = 1|X) = Pr(disease|exposure) = p(X)}$$

These probabilities may be estimated directly with a cohort design where the $x's$ are treated as fixed.

Case-control studies

Suppose that Y and X are related via the logistic regression model:

$$\mathrm{Pr(Y = 1|X) = p(X)} = \frac{\exp(\beta_0 + \sum_{j=1}^{k} X_j\beta_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{k} X_j\beta_j)}$$

The *relative risk* of individuals having exposures X and X^* is defined as:

$$\mathrm{Relative\,Risk} = \frac{\mathrm{Pr(Y = 1|X)}}{P(Y = 1|X^*)}$$

which is approximately the odds ratio:

$$\frac{\mathrm{Pr(Y = 1|X)/Pr(Y = 0|X)}}{P(Y = 1|X^*)/\mathrm{Pr(Y = 0|X^*)}}$$

for rare diseases. With the logistic regression model above we have:

$$\frac{p(X)/(1 - p(X))}{p(X^*)/(1 - p(X^*)} = \exp\left\{\sum_{j=1}^{p} \beta_j(X_j - X_j^*)\right\}$$

so that, in particular, $\exp(\beta_j)$ represents the increase in odds of disease associated with a unit increase in X_j when all other

Case-control studies

We now show how analysis proceeds for case-control data. We first introduce an indicator variable Z which represents the event that an individual was selected for the study ($Z = 1$) or not ($Z = 0$). We now let

$$\pi_1 = \mathrm{Pr(Z = 1|Y = 1)}$$

denote the probability that a case was selected for the study and π_0 be the same for controls. Typically $\pi_1 \gg \pi_0$ Now consider the probability a person is diseased, given exposures X and selection

$$\begin{aligned}
\mathrm{Pr(Y = 1|Z = 1, X)} &= \frac{\mathrm{Pr(Z = 1|Y = 1)Pr(Y = 1|X)}}{\mathrm{Pr(Z = 1|Y = 1, X)Pr(Y = 1|X) + Pr(Z = 1|Y = 0)Pr(Y = 0|X)}} \\
&= \frac{\mathrm{Pr(Z = 1|Y = 1)Pr(Y = 1|X)}}{\mathrm{Pr(Z = 1|Y = 1)Pr(Y = 1|X) + Pr(Z = 1|Y = 0)Pr(Y = 0|X)}} \\
&= \frac{\pi_1 \exp(X\beta)/(1 + \exp(X\beta))}{\pi_1 \exp(X\beta)/(1 + \exp(X\beta)) + \pi_0/(1 + \exp(X\beta))} \\
&= \frac{\pi_1 \exp(\beta_0 + \sum_{j=1}^{k} X_j\beta_j)}{\pi_0 + \pi_1 \exp(\beta_0 + \sum_{j=1}^{k} X_j\beta_j)} \\
&= \frac{\exp(\beta_0^* + \sum_{j=1}^{k} X_j\beta_j)}{1 + \exp(\beta_0^* + \sum_{j=1}^{k} X_j\beta_j)}
\end{aligned}$$

where $\beta_0^* = \beta_0 + \log \pi_1/\pi_0$

Case-control studies

So the probabilities of disease in a case-control study also follow a logistic model but with an altered intercept. The usual case is that $\pi_1 \gg \pi_0$ so the intercept is increased to account for the "over-sampling" of cases.

An important assumption used in the derivation is:

$$\mathrm{Pr}(\mathrm{Z} = 1 | \mathrm{Y} = \mathrm{y}, \mathrm{X}) = \mathrm{Pr}(\mathrm{Z} = 1 | \mathrm{Y} = 1) = \pi_{\mathrm{y}}$$

in other words, the selection probabilities depend on the disease status but not the exposure. A a random sample of cases and controls is sufficient for this assumption to be valid.

In general the probabilities of selection are not constant, but depend on how many cases or controls have been selected. In this case a more general proof is required (Prentice and Pike 1979, Biometrika)

Relationship between Poisson and multinomial

Consider the log linear model: $Y \sim \mathrm{Poisson}(\mu)$ $\log(\mu) = \alpha\mathbf{1} + X\boldsymbol{\beta}$. You can show [for fun/practice] that the sufficient statistic for α is $\mathbf{1}'y = y_+ = \sum_{i=1}^n y_i$.

Since the sum of Poisson random variables is Poisson with mean = sum of the means. So

$$y_+ = \mathrm{Poisson}(\mu_+)$$

and

$$\mu_+ = \sum_{i=1}^n \mu_i = \sum_{i=1}^n \exp(\eta_i)$$

where $\eta_i = \log(\mu_i) = \alpha + X_i^t\beta$ so:

$$\begin{aligned} \mathrm{Pr(y|y_+)} &= \frac{\prod_{i=1}^n \exp(\eta_i y_i - \mu_i)/y_i!}{\exp(\log(\mu_+)y_+ - \mu_+)/y_+!} \\ &= \frac{y_+!}{\prod_{i=1}^n y_i!}\prod_{i=1}^n\left(\frac{\mu_i}{\mu_+}\right)^{y_i} \end{aligned}$$

which is multinomial with y_+ trials and success probability $\pi_i = \frac{\mu_i}{\mu_+}$

Relationship between Poisson and multinomial

Note that $\pi_i = \frac{\mu_i}{\mu_+} = \frac{\exp(\alpha+X_i'\beta)}{\sum_{i=1}^n \exp(\alpha+X_i'\beta)} = \frac{\exp(X_i'\beta}{\sum \exp(X_i\beta)}$
Now we can decompose the log likelihood:

$$\log(\mathrm{Pr}(\mathrm{y}|\pi_{\mathrm{i}}, \mu_+) = \underbrace{\log(\mathrm{Pr}(\mathrm{y}|\mathrm{y}_+, \pi_{\mathrm{i}}))}_{\text{multinomial log likelihood}} - \log\left[\mathrm{Pr}(\mathrm{y}_+|\mu_+)\right]$$

So that if we maximize the likelihood then we maximize the multinomial log-likelihood with respect to the π_i. Therefore, provided a multinomial can be written as: $\pi_i = \frac{\mu_i}{\mu_+}$, we can fit it via log-linear models.

Response and stimulus factors

This result is a big deal. Before this trick, it was very difficult to analyze multi-way contingency tables. Now it is a piece of cake with log-linear Poisson regression.

When using this trick, it is convenient to divide the factors in your contingency table into *response* and *stimulus* factors. Stimulus factors have their marginal totals fixed in advance (so you can use the trick). The main interest is the conditional probabilities of the response factors given the stimulus factors.

Identifying the multinomial model corresponding to any surrogate Poisson model is straightforward. The *minimum model* is the interaction of all stimulus factors, and must be included for the analysis to respect the fixed totals over the response factors. The minimum model is usually of no interest, corresponding to a uniform distribution over response factors independent of stimulus factors.

Interactions between response and stimulus factors indicate interesting structure.

A simple example

Consider the log-linear model $\mu_{ij} = \exp(\lambda + \lambda_i^x + \lambda_j^y)$ and the "independence" multinomial probabilities $\pi_{ij} = \pi_i + \pi_j$ for a two by two table. If you assume the log-linear model, prove to yourself that if we let $\pi_{ij} = \frac{\mu_{ij}}{\sum_{ij} \mu_{ij}}$ then the π_{ij} satisfy the independence multinomial probabilities.

So we can write the independence model as a log-linear model where:

$$\log(\mu_{ij}) = \lambda + \underbrace{\lambda_i^x}_{\text{row effect}} + \underbrace{\lambda_j^y}_{\text{column effect}}$$

Typically $\lambda_I^x = \lambda_I^y = 0$ for identifiability. This model requires $(I-1) + (I-1) + 1 = 2I - 1$ degrees of freedom. Leaving $I^2 - 2I + 1 = (I-1)^2$ residual degrees of freedom[2]

[2] Where have you seen this type of calculation before?

An example

Data set reported by Ries & Smith (1963), analyzed by Cox & Snell (1989) and described in *Modern Applied Statistics with S+*.

M user?	No				Yes			
Temperature	Low		High		Low		High	
Preference	X	M	X	M	X	M	X	M
Water softness								
Hard	68	42	42	30	37	52	24	43
Medium	66	50	33	23	47	55	23	47
Soft	63	53	29	27	57	49	19	29

We study how the proportion of users preferring one or other brand varies with the other factors. In our terminology brand (X/M) is the only response factor and M user/Temperature/Water softness are stimulus factors. So the minimum model has all interactions

$$\text{M user} \times \text{Temperature} \times \text{Water softness}$$

The example in R

```
# Load the data in, stacking the Brand, Temp, M.user, and Soft variables
> detg <- cbind(expand.grid(Brand = c("X","M"),Temp=c("Low","High"), M.user = c("N","Y"),Soft =
c("Hard","Medium","Soft")),Fr = (68,42,42,30,37,52,24,43, 66, 50, 33, 23, 47, 55,23,47, 63,63,
29,27,57,49,19,29))

# Tell R that the Soft variable is ordered
>detg$Soft <- ordered(detg$Soft,levels=c("Soft","Medium","Hard")

# Fit the model including only main effects for Soft/Brand/M.user
> detg.mod <- glm(terms(Fr   M.user*Temp*Soft + Brand*(M.user + Temp +
Soft,keep.order=T),family=poisson,data=detg)
> summary(detg.mod)
...
BrandM -0.39112 0.13590 -2.878 0.00400 **
M.userY:BrandM 0.56702 0.12775 4.439 9.05e-06 ***
TempHigh:BrandM 0.25665 0.13286 1.932 0.05340 .
SoftMedium:BrandM 0.04952 0.15529 0.319 0.74983
SoftSoft:BrandM -0.02111 0.15763 -0.134 0.89346
```

Equivalence to logistic analysis

```
> attach(detg)
> deterg <- cbind(detg[Brand=="X",-1],M=detg[Brand=="M","Fr"])
> detach()
> names(deterg)[4] <- "X"
> detg.lg <- glm(cbind(M,X) ~ M.user*Temp, family=binomial,data=deterg)

> summary(detg.lg, correlation=F)
Coefficients:
Estimate Std.  Error z value Pr(>|z|)
(Intercept) -0.30647 0.10942 -2.801 0.0051 **
M.userY 0.40757 0.15961 2.554 0.0107 *
TempHigh 0.04411 0.18463 0.239 0.8112
M.userY:TempHigh 0.44427 0.26673 1.666 0.0958
Null deviance:  32.826 on 11 degrees of freedom
Residual deviance:  5.656 on 8 degrees of freedom


> detg.m0 <- glm(Fr ~ M.user.Temp*Soft + Brand, family=poisson, data=detg)
> summary(detg.m0)
Null deviance:  118.627 on 23 degrees of freedom
Residual deviance:  32.826 on 11 degrees of freedom
```

Further Reading

More references on log-linear models can be found:

- http://data.princeton.edu/wws509/notes/c5.pdf
- http://www-m4.ma.tum.de/courses/GLM/lec9.pdf
- In Mcullagh and Nelder

Be careful, you are fitting a bunch of interactions, this can become computationally intensive. See `loglin()` and associated help files for more computationally efficient approaches.