**Section 4.2.4 – Poisson regression for contingency tables: loglinear models**

Poisson regression models are frequently used to model counts in a contingency table. For this setting, each cell corresponds to an independent Poisson random variable. The explanatory variables in the model are simply categorical explanatory variables representing the rows and columns!

Example: Larry Bird (BirdCh4.R)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Second |   |
|   |   | Made | Missed | Total |
| First | Made | 251 | 34 | 285 |
| Missed | 48 | 5 | 53 |
|   | Total | 299 | 39 | 338 |

The data can be converted to a data frame:

> c.table <- array(data = c(251, 48, 34, 5), dim = c(2,2),

 dimnames = list(First = c("made", "missed"),

 Second = c("made", "missed")))

> bird <- as.data.frame(as.table(c.table))

> bird

 First Second Freq

1 made made 251

2 missed made 48

3 made missed 34

4 missed missed 5

Using bird$Freq as the Poisson response variable, we can form the following model:

log(μ) = β0 + β1FirstMissed + β2SecondMissed

where

FirstMissed = 1 when the first free throw is missed and = 0 for made

SecondMissed = 1 when the second free throw is missed and = 0 for made

Notice that this model means there is “independence” between the row and column variable because they do not “interact” in how they affect the response.

Below is how we can estimate this model under independence:

> mod.fit1 <- glm(formula = Freq ~ First + Second, data =

 bird, family = poisson(link = log))

> summary(mod.fit1)

Call:

glm(formula = Freq ~ First + Second, family = poisson(link = log), data = bird)

Deviance Residuals:

 1 2 3 4

-0.0703 0.1623 0.1934 -0.4659

Coefficients:

 Estimate Std. Error z value Pr(>|z|)

(Intercept) 5.52989 0.06241 88.61 <2e-16 \*\*\*

Firstmissed -1.68220 0.14959 -11.25 <2e-16 \*\*\*

Secondmissed -2.03688 0.17025 -11.96 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

 Null deviance: 402.05553 on 3 degrees of freedom

Residual deviance: 0.28575 on 1 degrees of freedom

AIC: 28.212

Number of Fisher Scoring iterations: 3

The estimated model is

log() = 5.53 - 1.68FirstMissed - 2.04SecondMissed

The estimated mean counts are

> predict(object = mod.fit1, newdata = bird, type =

 "response")

 1 2 3 4

252.115385 46.884615 32.884615 6.115385

> Pearson.test <- chisq.test(x = c.table, correct = FALSE)

> Pearson.test$expected

 Second

First made missed

 made 252.11538 32.884615

 missed 46.88462 6.115385

Notice these counts are the same as the expected cell counts that we could calculate using Chapter 3 methods!

Below is how we can estimate the model that includes the interaction:

> mod.fit2 <- glm(formula = Freq ~ First + Second +

 First:Second, data = bird, family = poisson(link =

 log))

> summary(mod.fit2)

Call:

glm(formula = Freq ~ First + Second + First:Second, family = poisson(link = log), data = bird)

Deviance Residuals:

[1] 0 0 0 0

Coefficients:

 Estimate Std. Error z value Pr(>|z|)

(Intercept) 5.52545 0.06312 87.540 <2e-16 \*\*\*

Firstmissed -1.65425 0.15754 -10.501 <2e-16 \*\*\*

Secondmissed -1.99909 0.18275 -10.939 <2e-16 \*\*\*

Firstmissed:Secondmissed -0.26267 0.50421 -0.521 0.602

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

 Null deviance: 4.0206e+02 on 3 degrees of freedom

Residual deviance: 4.3077e-14 on 0 degrees of freedom

AIC: 29.926

Number of Fisher Scoring iterations: 3

The estimated model is

log() = 5.52 - 1.65FirstMissed - 2.00SecondMissed

 - 0.26FirstMissed×SecondMissed

Notice the residual deviance is 0! This is because we are estimating a saturated model! There are four parameters and four Poisson observations.

The estimated mean counts are

> predict(object = mod.fit2, newdata = bird, type =

 "response")

 1 2 3 4

251 48 34 5

> bird$Freq

[1] 251 48 34 5

The estimated values are the same as the observed!

Below is the information needed for a LRT involving the interaction (H0:β3 = 0 vs. Ha:β3 ≠ 0):

> library(package = car)

> Anova(mod.fit2, test = "LR")

Analysis of Deviance Table (Type II tests)

Response: Freq

 LR Chisq Df Pr(>Chisq)

First 174.958 1 <2e-16 \*\*\*

Second 226.812 1 <2e-16 \*\*\*

First:Second 0.286 1 0.593

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(mod.fit1, mod.fit2, test = "Chisq")

Analysis of Deviance Table

Model 1: Freq ~ First + Second

Model 2: Freq ~ First + Second + First:Second

 Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 1 0.28575

2 0 0.00000 1 0.28575 0.593

> library(package = vcd)

> assocstats(c.table)

 X^2 df P(> X^2)

Likelihood Ratio 0.28575 1 0.59296

Pearson 0.27274 1 0.60150

Phi-Coefficient : 0.028

Contingency Coeff.: 0.028

Cramer's V : 0.028

The hypothesis test for the interaction term results in
-2log(Λ) = 0.286 with a p-value of 0.593. Notice that this is exactly the same as we found in Chapter 1 and we would find using the methods in Chapter 3!

Loglinear models

It is very common in categorical data analysis books to differentiate between Poisson regression models when

* Only categorical explanatory variables are present (Larry Bird data) and
* At least one continuous explanatory variable is present (Horseshoe crab data)

For example, Agresti’s categorical books would give the following models for the Larry Bird data:

log(μij) = λ +  +  for i = 1, 2 and j = 1, 2

under independence and

log(μij) = λ +  +  +  for i = 1, 2 and j = 1, 2

under dependence, where

* μij is the mean response for row i and column j
* , , and  are parameters, and they have constraints such as , , and  for (i,j) = (1,1), (1,2), and (2,1) so that we do not have more parameters than Poisson observations (rows of our data set); in other courses, you may hear these constraints referred to as “identifiability” restrictions**.**

The models for the Larry Bird data are often referred to as loglinear models. A “Poison regression” model would generally refer to a model where at least one explanatory variable is continuous.

This differentiating between models with and without categorical explanatory variables is unnecessary. For example, the estimated model for the Larry Bird data under independence was

log() = 5.53 - 1.68FirstMissed - 2.04SecondMissed

Using Agresti’s notation, this model would be written instead as



where  and . Essentially, the loglinear model is an ANOVA model representation of a regression model. Remember that ANOVA models are just a special case of regression models!

**Section 4.2.5 – Large loglinear models**

Contingency tables can have additional “dimensions” to them. For example, a three-dimensional (three-way) contingency table would have a form:

 Z=1 Z = 2

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Y |  |  |  |  | Y |  |
|  |  | 1 | 2  |  |  |  |  | 1 | 2 |  |
| X | 1 | n111 | n121 | n1+1 |  | X | 1 | n112 | n122 | n1+2 |
| 2  | n211 | n221 | n2+1 |  | 2  | n212 | n222 | n2+2 |
|  |  | n+11 | n+21 | n++1 |  |  |  | n+12 | n+22 | n++2 |

where Z represents the “strata” or “layer” dimension. Similar to the last section, these dimensions are represented as explanatory variables in a Poisson regression model. Main effects, two-way interactions, and three-way interactions can be included.

Poisson and multinomial discussion

In contingency table settings, one can perform very similar analyses using a multinomial regression model (or a logistic regression model when there are only two response categories) and a Poisson regression model. With respect to the two dimensional table case, this occurs for the following reason:

The joint probability distribution of IJ independent Poisson random variables CONDITIONAL on their sum is a multinomial distribution.

The same holds true for three, four, ... dimensional cases as well by taking into account the larger number of independent Poisson random variables.

To see this result, consider a 2×2 contingency table using the notation at the beginning of Chapter 3:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Y |  |
|  |  | 1 | 2  |  |
| X | 1 | n11 | n12 | n1+ |
| 2  | n21 | n22 | n2+ |
|  |  | n+1 | n+2 | n++ |

Please note that I will use a little different notation here for the Poisson aspect to help make the connection to Chapter 3. Suppose nij ~ independent Po(μij) for i = 1, 2 and j = 1, 2. Then the joint probability distribution is



Let n++ = n11 + n12 + n21 + n22 (note: n++ = n in our more commonly used notation). Because each random variable is independent, we have n++ ~ . Then









This is a multinomial distribution for n++ trials and probabilities .

The BirdCh4.R program contains some additional code to help demonstrate the equivalence between the multinomial and Poisson approaches.

Is there an advantage to using the Poisson model over the multinomial model? There can be one – when you do not want to specify having ONLY one response variable. Below are two examples to explain this:

Example: One could think of the Larry Bird example as having TWO response variables: First free throw outcome and second free throw outcome.

Example: In the fiber example, the fiber source was an “input” and we were interested in the bloating “output” that resulted from it. Thus, bloating was the only response variable.

Interestingly, you will still obtain the same analysis results even if you used a Poisson regression model when a multinomial regression model may seem more appropriate and vice versa. This is because of the equivalence that occurs between the Poisson and multinomial distributions.

Odds ratios

Odds ratios can be helpful to interpret the fit of a model in contingency table settings. Simply, find the predicted counts for each cell and use odds ratios to interpret relationships between variables. For example, the estimated odds ratio for the Larry Bird data example was 1 when estimating the model without the interaction.

Ordinal categorical variables

One can take advantage of ordinal properties of a categorical explanatory variable in a Poisson regression setting. Please see my book for a discussion.