

Statistics 4590 GLM's for Tables of Counts

Consider the dataset we studied previously on gender bias in grad school admissions at Wabash Tech:

```
> admit.status

, , Business
  Admit Deny
Male  480  120
Female 180   20

, , Law
  Admit Deny
Male   10   90
Female 100  200
```

We analyzed this problem using the Mantel–Haenzel test to see if the odds of a male being admitted were greater than a female being admitted, conditional on the school to which they applied.

Now we'll analyze this data using (a) a logistic model; (b) a log–linear model.

Logistic Model

To analyze this using a logistic model, we need to have the data set up in a specific manner in R:

```
> admit.reg <- data.frame(School=c("B","B","L","L"),
  Gender=c("M","F","M","F"), Admit=c(480,180,10,100),Deny=c(120,20,90,200))

> admit.reg
  School Gender Admit Deny
1     B      M  480  120
2     B      F  180   20
3     L      M   10   90
4     L      F  100  200
```

Define

$Y_i =$ number of students admitted to grad school

Then $Y_i \sim \text{Bin}(m_i, \pi_i)$. The m_i values are simply the sum of those admitted and those not admitted for each (gender, school) combination.

We start with a logistic model to estimate $\text{logit}(\pi_i)$ from gender and school, including an interaction term:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

where

$x_1 = 1$ if applying to law school, 0 otherwise,

$x_2 = 1$ if applicant is male, 0 otherwise.

```
> admit.logisint <- glm(cbind(Admit,Deny)~School*Gender,family=binomial, data = admit.reg)
```

```

> summary(admit.logisint)

# some output deleted

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.1972     0.2357   9.322 < 2e-16 ***
SchoolL         -2.8904     0.2656 -10.881 < 2e-16 ***
GenderM          -0.8109     0.2569  -3.157 0.00159 **
School:GenderM  -0.6931     0.4383  -1.582 0.11375
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3.8845e+02 on 3 degrees of freedom
Residual deviance: 8.8818e-14 on 0 degrees of freedom
AIC: 29.232

```

Note that the residual deviance for this model is essentially 0, *i.e.* the model fits the data perfectly. This is because we are using 4 observations to estimate 4 parameters.

What does β_3 tell us in this model? The odds of being admitted are

$$\frac{\pi_i}{1 - \pi_i} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)$$

So the odds of a male being admitted: $\exp[(\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1]$

Odds of a female being admitted: $\exp[\beta_0 + \beta_1 x_1]$

Then the odds ratio (male vs. female) is $\exp[(\beta_2 + \beta_3 x_1)]$. This depends on the school.

What if $\beta_3 = 0$? Then the odds ratio is constant for each school. This is what we had to assume to use the Mantel-Haenzel test. Therefore, testing to see if we can drop the interaction term is equivalent to testing if the odds ratio is the same for each of our tables.

We test this with the drop in deviance test. We can use the `anova` output below to get the deviance values:

```

> anova(admit.logisint) # Some output deleted
Terms added sequentially (first to last)

      Df Deviance Resid. Df Resid. Dev
NULL                3      388.45
School              1    353.39
Gender              1     32.46
School:Gender      1     2.60 # Dev. of School+Gender model
School:Gender      0 8.882e-14 # Dev. of full model

```

Suppose we are able to drop the interaction term. Let's refit the model without it:

```

> admit.logis <- glm(cbind(Admit,Deny)~School+Gender,family=binomial, data = admit.reg)

```

```

> summary(admit.logis, dispersion=admit.logis$deviance/admit.logis$df.residual)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.4287      0.3288   7.387 1.5e-13 ***
SchoolL     -3.1794      0.3379  -9.409 < 2e-16 ***
GenderM      -1.0815      0.3360  -3.219 0.00129 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 2.596104)

Null deviance: 388.4465  on 3  degrees of freedom
Residual deviance:  2.5961  on 1  degrees of freedom
AIC: 29.828

```

There is some evidence for overdispersion in this model, which we would need to account for in doing inference with this model.

Log-linear Model

To use a log-linear model to analyze this data, we must set up the data in the following form:

```

> admit.reg1 <- data.frame(School = c("B","B","B","B","L","L","L","L"),
  Gender=c("M","F","M","F","M","F","M","F"),
  Admit=c("Y","Y","N","N","Y","Y","N","N"),
  Count = c(480, 180, 120, 20, 10,100,90,200))
> admit.reg1
  School Gender Admit Count
1      B      M      Y  480
2      B      F      Y  180
3      B      M      N  120
4      B      F      N   20
5      L      M      Y   10
6      L      F      Y  100
7      L      M      N   90
8      L      F      N  200

```

For the log-linear model, we need to use admission status, gender and school as the explanatory variables. Our response is

$$Y_i = \text{number of students at a (school, gender, admit) combination}$$

so $Y_i \sim \text{Poisson}(\lambda_i)$. Our model is

$$\log(\lambda_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 x_2 x_3$$

with x_1 and x_2 defined earlier, and $x_3 = 1$ if student is admitted, 0 otherwise.

Some of the R output follows:

```

> admit.logfit <- glm(Count ~ School * Gender * Admit, family = poisson,
  data = admit.reg1)

```

```
> summary(admit.logfit)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.9957     0.2236  13.397 < 2e-16 ***
SchoolL          2.3026     0.2345   9.818 < 2e-16 ***
GenderM          1.7918     0.2415   7.419 1.18e-13 ***
AdmitY           2.1972     0.2357   9.322 < 2e-16 ***
School:GenderM  -2.5903     0.2728  -9.494 < 2e-16 ***
School:AdmitY   -2.8904     0.2656 -10.881 < 2e-16 ***
GenderM:AdmitY  -0.8109     0.2569  -3.157 0.00159 **
School:GenderM:AdmitY -0.6931     0.4383  -1.582 0.11375
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 9.3598e+02 on 7 degrees of freedom
Residual deviance: 1.3323e-14 on 0 degrees of freedom
AIC: 66.591
```

Can we drop the 3-way interaction term?

```
> anova(admit.logfit, test= "Chi") # test = "Chi" tells R
                                     # to include the drop-in-deviance p-value
```

Analysis of Deviance Table

Model: poisson, link: log

Response: Count

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi)
NULL			7	935.98	
School	1	135.92	6	800.06	2.078e-31
Gender	1	33.49	5	766.57	7.165e-09
Admit	1	97.67	4	668.91	4.954e-23
School:Gender	1	280.46	3	388.45	5.963e-63
School:Admit	1	353.39	2	35.06	7.742e-79
Gender:Admit	1	32.46	1	2.60	1.217e-08
School:Gender:Admit	1	2.60	0	1.332e-14	0.11

Now, look at the `anova` output to see the deviance of the model with all terms except the 3-way interaction. It is 2.60, which is equal to the deviance from the logistic model that did not have the gender-school interaction term. The p-value, $P(\chi^2 \geq 2.60) = 0.11$ is also included in the output. Therefore, there is little evidence of a 3-way interaction. Recall that in the Mantel-Haenzel test we had to assume no 3-way interaction of the variables. We now have a way to test if this assumption is reasonable.

This is no coincidence. McCullagh and Nelder (1989, *Generalized Linear Models, 2nd edition*) discuss under what conditions a logistic and log-linear analysis can yield the same results.

To fit a model with no 3-way interaction, we would use the following notation in R:

```
> admit.logfit2 <- glm(Count ~ (School + Gender + Admit)^2, family = poisson, data = admit.reg1)
```

The $()^2$ notation in `glm` tells R to include all interaction terms up to, and including, second order.