Lecture 23: Generalized Linear Model

Ailin Zhang

2023-07-11

Moving Beyond Linearity

- The truth is never linear! Or almost never. But the linearity assumption is often good enough.

Moving Beyond Linearity

- The truth is never linear! Or almost never. But the linearity assumption is often good enough.
- Generalized linear model (GLM):
  - Structure of GLM
  - Statistical theory and regression diagnostic
  - Application for count data (Contingency Tables)
  - Design-based statistical inference

The Structure of Generalized Linear Models

$$Y_i = \mu_i + \epsilon_i$$

- A generalized linear model (or GLM) consists of three components:
  1. Random Component: variability or randomness in the response variable (Y_i). The random component is typically specified using a distribution from the exponential family, which includes several common distributions such as the Gaussian, Bernoulli, Poisson, and multinomial distributions.

The Structure of Generalized Linear Models

$$Y_i = \mu_i + \epsilon_i$$

- A generalized linear model (or GLM) consists of three components:
  1. Random Component: variability or randomness in the response variable (Y_i). The random component is typically specified using a distribution from the exponential family, which includes several common distributions such as the Gaussian, Bernoulli, Poisson, and multinomial distributions.
  2. A linear predictor: a linear function of regressors

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

The Structure of Generalized Linear Models

$$Y_i = \mu_i + \epsilon_i$$

- A generalized linear model (or GLM) consists of three components:

1. Random Component: variability or randomness in the response variable (Y_i). The random component is typically specified using a distribution from the exponential family, which includes several common distributions such as the Gaussian, Bernoulli, Poisson, and multinomial distributions.
2. A linear predictor: a linear function of regressors

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

3. A smooth and invertible linearizing link function $g(.)$. It transforms the expection of the response variable $\mu_i = E(Y_i)$ to the linear predictor:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

Common Link Functions

Table 15.1 Some Common Link Functions and Their Inverses

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$\log_e \mu_i$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Square-root	$\sqrt{\mu_i}$	η_i^2
Logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
Complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

NOTE: μ_i is the expected value of the response, η_i is the linear predictor, and $\Phi(\cdot)$ is the cumulative distribution function of the standard-normal distribution.

Canonical Link Functions

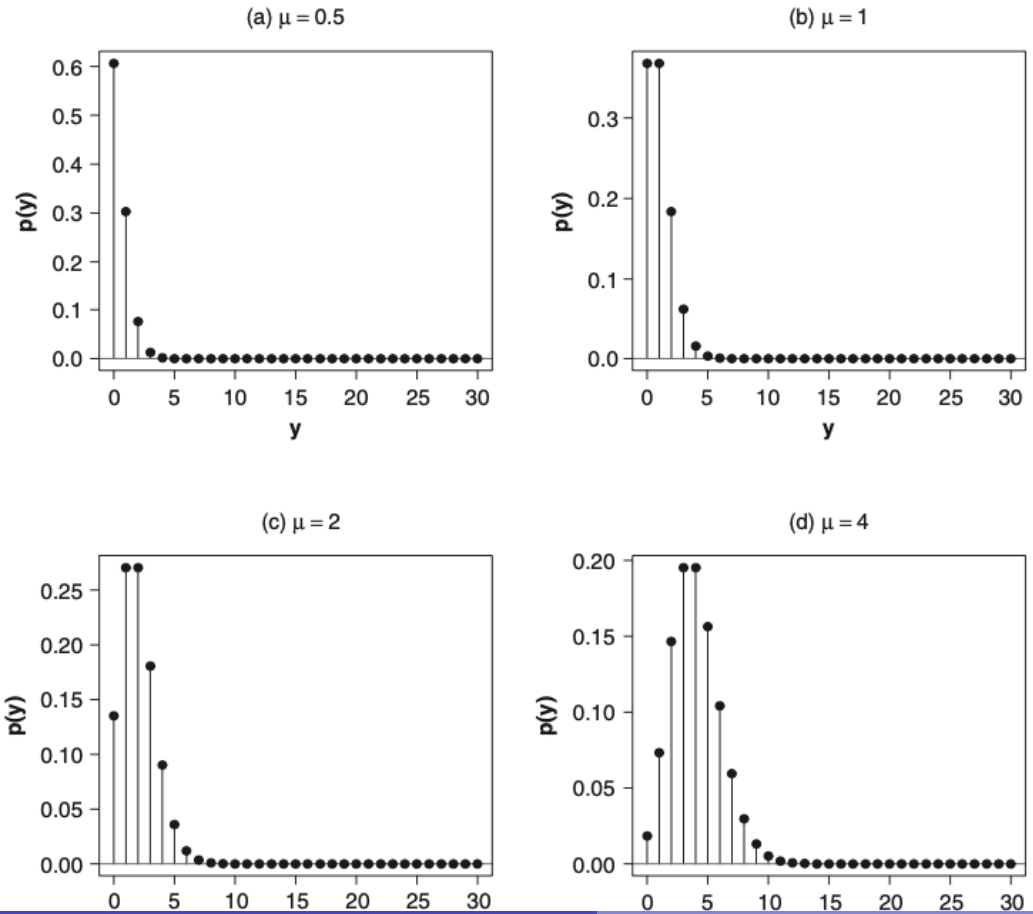
Table 15.2 Canonical Link, Response Range, and Conditional Variance Function for Exponential Families

Family	*Canonical Link*	*Range of* Y_i	$V(Y_i\|\eta_i)$
Gaussian	Identity	$(-\infty, +\infty)$	ϕ
Binomial	Logit	$\frac{0, 1, \ldots, n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poisson	Log	0,1,2,...	μ_i
Gamma	Inverse	$(0,\infty)$	$\phi\mu_i^2$
Inverse-Gaussian	Inverse-square	$(0,\infty)$	$\phi\mu_i^3$

NOTE: ϕ is the dispersion parameter, η_i is the linear predictor, and μ_i is the expectation of Y_i (the response). In the binomial family, n_i is the number of trials.

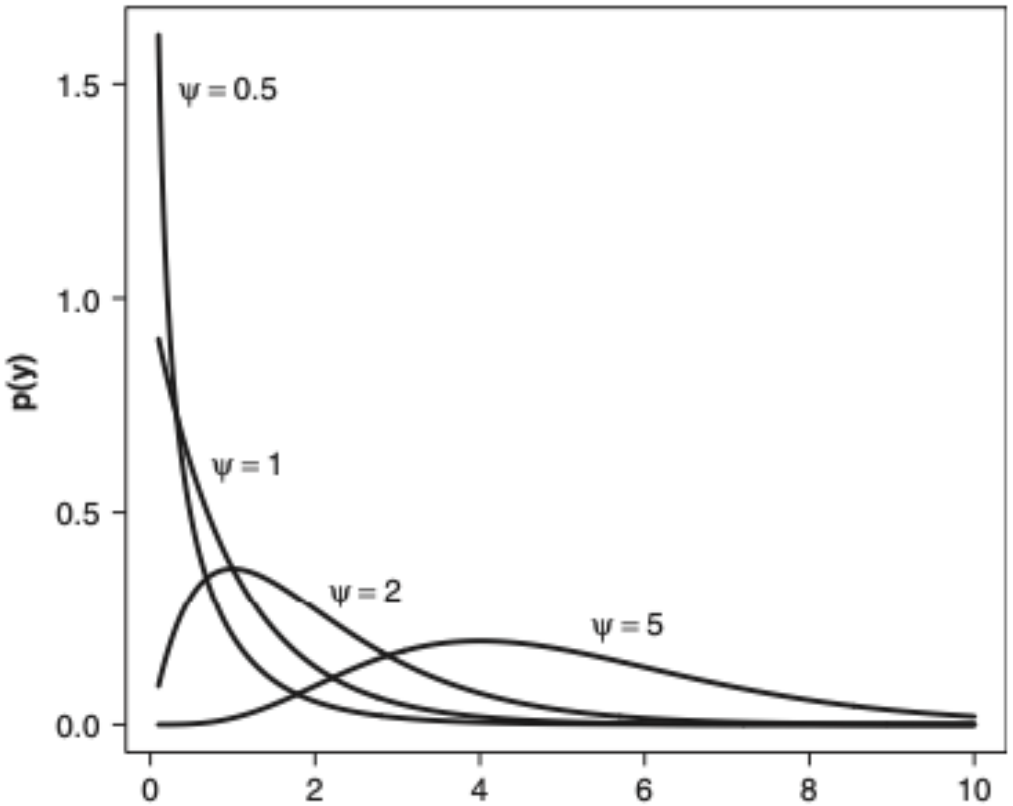
Poisson distribution

$p(y) = \mu^y \frac{e^{-\mu}}{y!}$: a discrete family with probability function indexed by the rate parameter $\mu > 0$



Gamma Distribution

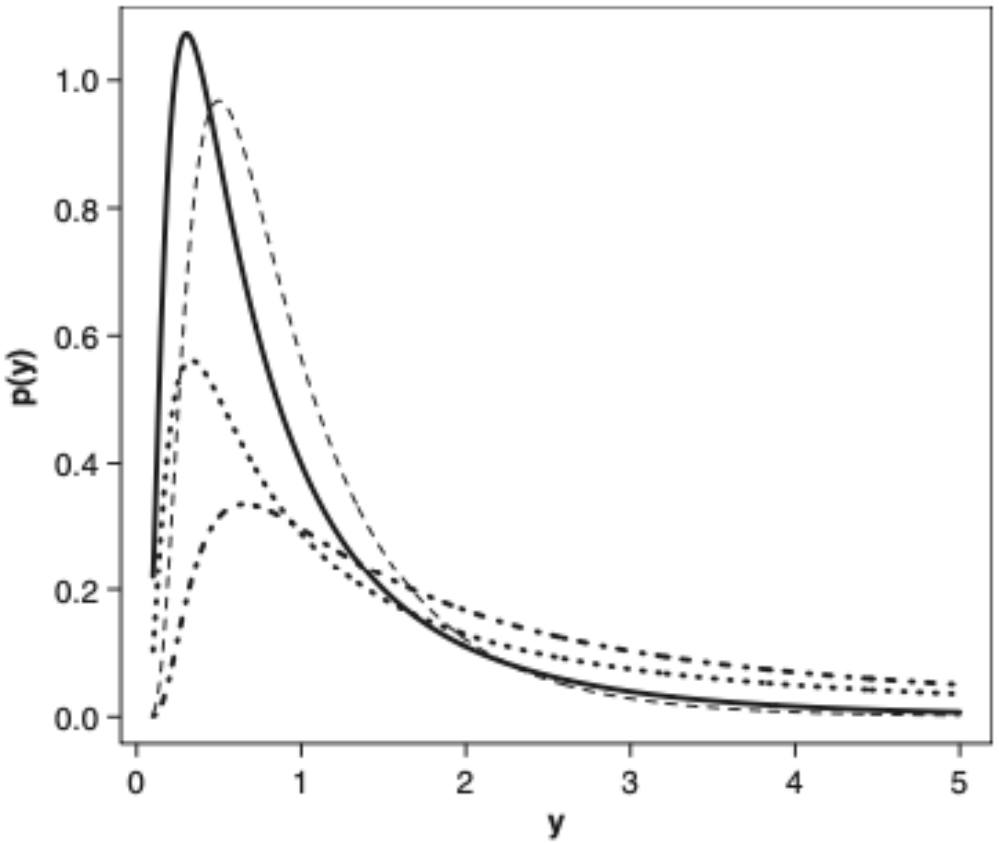
a continuous family with probability-density function indexed by the scale parameter $\omega > 0$ and shape parameter $\psi > 0$: $p(y) = (\frac{y}{\omega})^{\psi-1} \frac{\exp(-\frac{y}{\omega})}{\omega\Gamma(\psi)}$



Inverse-Gaussian Distribution

A continuous family indexed by two para- meters, μ and λ:

$$p(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[-\frac{\lambda(y-\mu)^2}{2y\mu^2}\right]$$



R Example

The "mtcars" dataset is a built-in dataset in R that contains information about various cars.

- mpg: Miles per gallon (numeric)
- cyl: Number of cylinders (numeric)
- disp: Displacement (cubic inches) (numeric)
- hp: Gross horsepower (numeric)
- drat: Rear axle ratio (numeric)
- wt: Weight (in thousands of pounds) (numeric)
- qsec: 1/4 mile time (seconds) (numeric)
- vs: Engine (0 = V/S, 1 = AM) (numeric)
- am: Transmission (0 = automatic, 1 = manual) (numeric)
- gear: Number of forward gears (numeric)
- carb: Number of carburetors (numeric)

R Example

```r
data(mtcars)
model <- glm(vs ~ mpg + cyl, data = mtcars, family = binomial(link = "logit"))
summary(model)
##
## Call:
## glm(formula = vs ~ mpg + cyl, family = binomial(link = "logit"),
##     data = mtcars)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5202  -0.2324  -0.1769   0.3465   1.3567
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  15.9714    11.0224   1.449   0.1473
## mpg          -0.1633     0.2399  -0.681   0.4961
## cyl          -2.1482     1.0811  -1.987   0.0469 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.860  on 31  degrees of freedom
## Residual deviance: 17.491  on 29  degrees of freedom
## AIC: 23.491
##
## Number of Fisher Scoring iterations: 6
```

Deviance

- The residual deviance for a GLM is $D_m = 2(\log(L_s) - \log(L_m))$ where L_m is the maximized likelihood under the model in question and L_s is the maximized likelihood under a saturated model
- The residual deviance is analogous to the residual sum of squares for a linear model.
- $R^2 = 1 - \frac{D_m}{D_{null}}$

Analysis of Deviance (ANOVA)

```
null_model <- glm(vs ~ 1, data = mtcars, family = poisson(link
lr_test <- anova(model, null_model, test = "LRT")
print(lr_test)
## Analysis of Deviance Table
##
## Model 1: vs ~ mpg + cyl
## Model 2: vs ~ 1
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        29     17.491
## 2        31     23.147 -2  -5.6559  0.05913 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Count data

$$\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

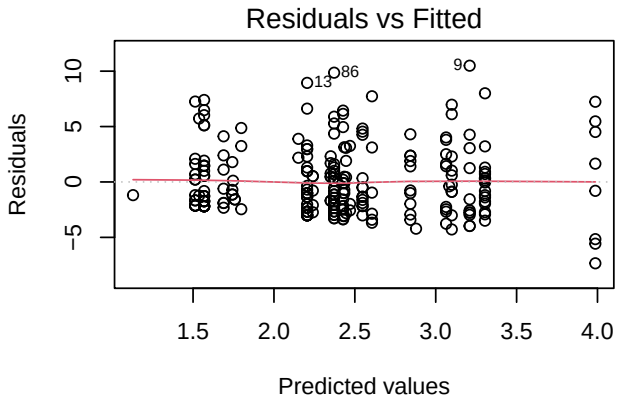
Michael Ornstein's data on interlocking directorates among 248 dominant Canadian firms:

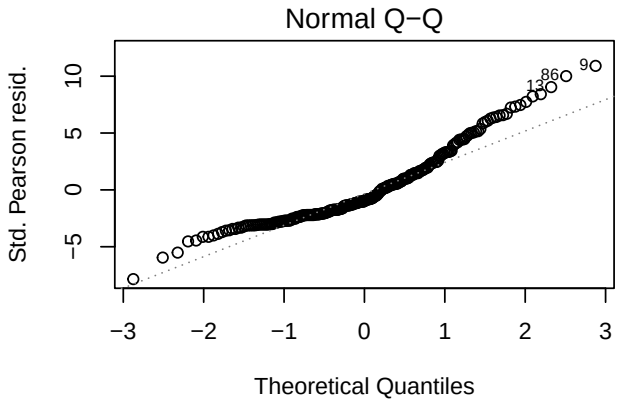
- assets: Assets in millions of dollars.
- sector: Industrial sector. A factor with levels: AGR, agriculture, food, light industry; BNK, banking; CON, construction; FIN, other financial; HLD, holding companies; MAN, heavy manufacturing; MER, merchandizing; MIN, mining, metals, etc.; TRN, transport; WOD, wood and paper.
- nation: Nation of control. A factor with levels: CAN, Canada; OTH, other foreign; UK, Britain; US, United States.
- interlocks: Number of interlocking director and executive positions shared with other major firms.

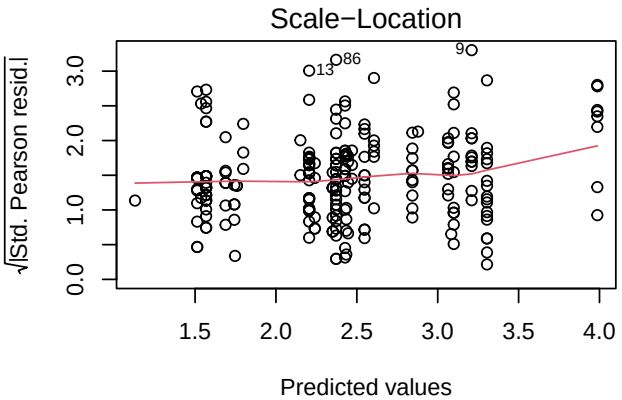
R example

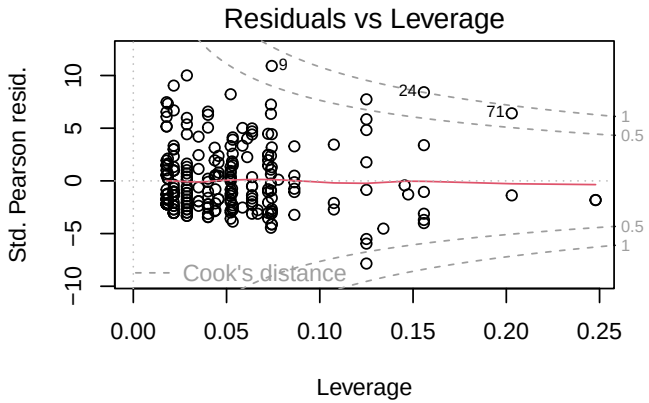
```
model1 <- glm(interlocks~.,data = Ornstein[,-1],family = "poisson")
summary(model1)
##
## Call:
## glm(formula = interlocks ~ ., family = "poisson", data = Ornstein[,
##     -1])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -10.3803   -2.8477   -0.9639    1.3847    8.3772
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.37195    0.05177  45.814  < 2e-16 ***
## sectorBNK    1.61471    0.07072  22.834  < 2e-16 ***
## sectorCON   -0.61381    0.21183  -2.898  0.00376 **
## sectorFIN    0.93258    0.06691  13.939  < 2e-16 ***
## sectorHLD    0.23398    0.11883   1.969  0.04896 *
## sectorMAN    0.05522    0.07547   0.732  0.46432
## sectorMER    0.17575    0.08652   2.031  0.04223 *
## sectorMIN    0.69211    0.06670  10.376  < 2e-16 ***
## sectorTRN    0.83791    0.07399  11.325  < 2e-16 ***
## sectorWOD    0.72715    0.07531   9.655  < 2e-16 ***
## nationOTH   -0.22080    0.07322  -3.016  0.00257 **
## nationUK    -0.62924    0.08892  -7.077 1.48e-12 ***
## nationUS    -0.85894    0.04856 -17.689  < 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: 3737.0  on 247  degrees of freedom
```

A model fit statistic $R^2 = 1 - (D_m)/(D_{null}) = 1 - \frac{1887.4}{3737.0} = 0.495$, shows that the model accounts for nearly half the deviance in number of interlocks.









Models for Overdispersed Count Data

- If the Poisson model fits the data reasonably, we would expect the residual deviance to be roughly equal to the residual degrees of freedom.
- Large residual deviance is so large suggests that the conditional variation of the expected number of interlocks exceeds the variation of a Poisson-distributed variable
- This common occurrence in the analysis of count data is termed overdispersion.
- Treament: quasi-Poisson , negative-binomial GLMs and zero-Inflated Poisson Regression

The Quasi-Poisson Model

A simple remedy for overdispersed count data is to introduce a dispersion parameter into the Poisson model, so that the conditional variance of the response is now $\mathrm{Var}(Y_i|\eta_i) = \phi\mu_i$

- The coefficient standard errors for the quasi-Poisson model are $\sqrt{\phi}$ times those for the Poisson model.
- The effect of introducing a dispersion parameter and obtaining quasi-likelihood estimates is (realistically) to inflate the coefficient standard errors

```r
model2 <- glm(interlocks~.,data = Ornstein[,-1],family = "quasipoisson")
summary(model2)
##
## Call:
## glm(formula = interlocks ~ ., family = "quasipoisson", data = Ornstein[,
##     -1])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -10.3803   -2.8477   -0.9639    1.3847    8.3772
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.37195    0.16122  14.713  < 2e-16 ***
## sectorBNK    1.61471    0.22020   7.333 3.62e-12 ***
## sectorCON   -0.61381    0.65959  -0.931 0.353024
## sectorFIN    0.93258    0.20833   4.476 1.18e-05 ***
## sectorHLD    0.23398    0.37003   0.632 0.527796
## sectorMAN    0.05522    0.23500   0.235 0.814416
## sectorMER    0.17575    0.26942   0.652 0.514824
## sectorMIN    0.69211    0.20770   3.332 0.001000 **
## sectorTRN    0.83791    0.23038   3.637 0.000339 ***
## sectorWOD    0.72715    0.23452   3.101 0.002167 **
## nationOTH   -0.22080    0.22800  -0.968 0.333832
## nationUK    -0.62924    0.27688  -2.273 0.023954 *
## nationUS    -0.85894    0.15120  -5.681 3.95e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 9.696081)
##
##     Null deviance: 3737.0  on 247  degrees of freedom
## Residual deviance: 2278.3  on 235  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

The Negative-Binomial Model

To adopt a Poisson model for the count Y_i by assuming that the expected count μ_i is an unobservable random variable that is gamma-distributed with mean μ_i and constant scale parameter ω

$$p(Y_i) = \frac{\Gamma(y_i + \omega)}{y!\Gamma(\omega)} \times \frac{\mu^{y_i}\omega^{\omega}}{(\mu_i + \omega)^{\mu_i + \omega}}$$

```r
library(MASS)
model3 <- glm.nb(interlocks~.,data = Ornstein[,-1])
summary(model3)
##
## Call:
## glm.nb(formula = interlocks ~ ., data = Ornstein[, -1], init.theta = 1.153576415,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9861  -1.0717  -0.3078   0.4258   1.9603
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.34086    0.15708  14.902  < 2e-16 ***
## sectorBNK    1.64581    0.36790   4.473 7.70e-06 ***
## sectorCON   -0.66053    0.49832  -1.326 0.184998
## sectorFIN    0.97179    0.25244   3.850 0.000118 ***
## sectorHLD    0.26530    0.39902   0.665 0.506130
## sectorMAN    0.11523    0.21136   0.545 0.585631
## sectorMER    0.20903    0.26633   0.785 0.432543
## sectorMIN    0.66281    0.20816   3.184 0.001452 **
## sectorTRN    0.88122    0.27110   3.251 0.001152 **
## sectorWOD    0.68439    0.26664   2.567 0.010266 *
## nationOTH   -0.09012    0.26667  -0.338 0.735396
## nationUK    -0.57447    0.27029  -2.125 0.033558 *
## nationUS    -0.83703    0.15076  -5.552 2.82e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.1536) family taken to be 1)
##
##     Null deviance: 398.30  on 247  degrees of freedom
## Residual deviance: 291.39  on 235  degrees of freedom
## AIC: 1740.1
##
```

Zero-Inflated Poisson Regression

- Too many 0s in the data than is consistent with a Poisson (or negative-binomial) distribution

Zero-Inflated Poisson Regression:

1. A binary logistic-regression model for membership in the latent class of individuals for whom the response variable is necessarily 0
2. a Poisson-regression model for the latent class of individuals for whom the response may be 0 or a positive count

```r
library(pscl)
model_zip <- zeroinfl(interlocks~.,data = Ornstein[,-1])
summary(model_zip)
##
## Call:
## zeroinfl(formula = interlocks ~ ., data = Ornstein[, -1])
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -3.8804 -1.5405 -0.6468  1.1539 10.8116
##
## Count model coefficients (poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.46761    0.05174  47.689  < 2e-16 ***
## sectorBNK    1.65259    0.07069  23.377  < 2e-16 ***
## sectorCON   -0.75189    0.21201  -3.546 0.000391 ***
## sectorFIN    0.81959    0.06691  12.249  < 2e-16 ***
## sectorHLD    0.29597    0.11882   2.491 0.012738 *
## sectorMAN    0.15745    0.07578   2.078 0.037734 *
## sectorMER    0.08238    0.08674   0.950 0.342270
## sectorMIN    0.58741    0.06719   8.743  < 2e-16 ***
## sectorTRN    0.71112    0.07419   9.585  < 2e-16 ***
## sectorWOD    0.66510    0.07544   8.816  < 2e-16 ***
## nationOTH   -0.11604    0.07468  -1.554 0.120209
## nationUK    -0.62519    0.08888  -7.034    2e-12 ***
## nationUS    -0.72009    0.04826 -14.921  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.4480     0.5607  -4.366 1.27e-05 ***
## sectorBNK      0.5021     1.2072   0.416  0.67745
## sectorCON    -16.7551  4709.1905  -0.004  0.99716
## sectorFIN    -16.6238  2221.8102  -0.007  0.99403
## sectorHLD      0.3281     1.2124   0.271  0.78668
## sectorMAN      0.2329     0.5653   0.412  0.68032
## sectorMER     -1.3054     1.1463  -1.139  0.25480
```

Loglinear Models for Contingency Tables

The joint distribution of several categorical variables defines a contingency table.

	Group A	Group B	Group C
Level 1	n11	n12	n13
Level 2	n21	n22	n23
Level 3	n31	n32	n33

Example: Happiness and Belief in Heaven

```
##    happy heaven count
## 1    not      no    32
## 2    not     yes   190
## 3 pretty      no   113
## 4 pretty     yes   611
## 5   very      no    51
## 6   very     yes   326
```

```
xtabs(count ~ happy + heaven, HappyHeaven)
##          heaven
## happy      no yes
##   not      32 190
##   pretty 113 611
##   very    51 326
```

```r
model4 <- glm(count ~ happy + heaven, family = poisson, data = HappyHeaven)
summary(model4)
##
## Call:
## glm(formula = count ~ happy + heaven, family = poisson, data = HappyHeaven)
##
## Deviance Residuals:
##       1         2         3         4         5         6
## -0.15570   0.06459   0.54947  -0.23152  -0.65897   0.27006
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.49313    0.09408   37.13  < 2e-16 ***
## happypretty  1.18211    0.07672   15.41  < 2e-16 ***
## happyvery    0.52957    0.08460    6.26 3.86e-10 ***
## heavenyes    1.74920    0.07739   22.60  < 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: 1019.87238  on 5  degrees of freedom
## Residual deviance:    0.89111  on 2  degrees of freedom
## AIC: 49.504
##
## Number of Fisher Scoring iterations: 3
```

Test for independence

$$\log(\mu) = \beta_0 + \beta_1 * \text{Happy pretty} + \beta_2 * \text{Happy very} + \beta_3 * \text{Heaven Yes}$$

```
anova(model4,test="LR")
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: count
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        5    1019.87
## happy    2   294.87          3     725.01 < 2.2e-16 ***
## heaven   1   724.11          2       0.89 < 2.2e-16 ***
##
```

Test for association between categorical variables

Null Hypothesis: there is no association between the two categorical variables.

```
chisq.test(xtabs(count ~ happy + heaven, HappyHeaven))
##
##  Pearson's Chi-squared test
##
## data:  xtabs(count ~ happy + heaven, HappyHeaven)
## X-squared = 0.88368, df = 2, p-value = 0.6429
```

- If the p-value is les than your significance level (commonly 0.05), you can reject the null hypothesis that there is no association between the two categorical variables.
- Otherwise you can not reject the null hypothesis.