Statistical Inference & Model Checking for Poisson Regression

Edps/Psych/Soc 589

Carolyn J. Anderson

Department of Educational Psychology

ILLINOIS

Outline

- Inference for model parameters (i.e., confidence intervals and hypothesis tests).
  - Wald Statistics
  - Likelihood ratio tests.
  - (Score tests).
- Assessing model fit.
  - Global fit statistics (X^2, G^2).
  - Residuals.
  - Confidence intervals for fitted values
  - Overdispersion.
- Extended example: Bullies
- Zero Inflated Models (ZIPs)
- Fitting GLMs, which provides further tools and insights into inference and model assessment.

Much of the basic logic and concepts for Poisson regression are the same as those for logistic regression, but we'll consider logistic regression in more detail when we cover Chapter 5.

Inference for model parameters

Suppose that we've fit the Poisson regression model

$$\log(\hat{\mu}_i) = \hat{\alpha} + \hat{\beta} x_i$$

For the AIDS data we obtained

$$\log(\hat{\mu}_i) = -1.9442 + 2.1748 x_i^*$$

where x_i^* is log(month period).
Typically, the kinds of hypothesis tests that we'ld be interested in performing are whether our explanatory variable(s) had an effect on the response variable; that is,
$H_O : \beta = 0$ versus one of the following:

- $H_A : \beta \neq 0$ (2-tailed test).
- $H_A : \beta > 0$ (right tailed test).
- $H_A : \beta < 0$ (left tailed test).

Wald Statistics

Maximum likelihood estimates are approximately normally distributed for large samples (i.e., MLEs are asymptotically normally distributed).
Which means that for "large" samples,

$$\hat{\beta} \approx \mathcal{N}(\beta, \sigma^2_{\hat{\beta}})$$

We can use this to

- Construct confidence intervals for parameters.
- Test hypotheses.

For a $(1-\alpha)100\%$ confidence interval estimate of β:

$$\hat{\beta} \pm z_{\alpha/2} ASE$$

where ASE is the asymptotic standard error of $\hat{\beta}$.
For the AIDs example, a 95% confidence interval for β is

$$2.1748 \pm 1.96(.2151) \longrightarrow (1.753, 2.596)$$

Wald Statistics for Hypothesis Testing

For the hypothesis test of $H_O : \beta = 0$ versus $H_A : \beta \neq 0$ (or $H_A : \beta > 0$ or $H_A : \beta < 0$),

$$z = \frac{\hat{\beta} - \beta_o}{ASE} = \frac{\hat{\beta}}{ASE}$$

where β_o is the hypothesized value of β under the null hypothesis (i.e., $\beta_o = 0$).

If the null hypothesis is *true*, then the statistic z is approximately standard normal (for large samples)

$$z = \frac{\hat{\beta}}{ASE} \approx \mathcal{N}(0, 1)$$

Wald Statistic

An estimated parameter divided by it's ASE and squared is a "Wald Statistic".

$$z^2 = \left(\frac{\hat{\beta}}{ASE}\right)^2 \approx \chi^2_1$$

z^2 has (asymptotically) a chi-squared distribution with $df = 1$.

- Wald statistics are usually provided on computer output, along with p–values (see SAS output).
- Wald statistics can be used to test 2-sided alternatives, while z can be used to test 1-sided as well as 2-sided alternatives.

  AIDS Example:

$$H_o : \beta = 0 \quad \text{versus} \quad H_o : \beta \neq 0$$

"Chi-square" $= (2.1748/.2151)^2 = (10.11)^2 = 102.23$, $df = 1$, $p < .01$.

Likelihood Ratio Test

For this test, we look at the ratio of

- The maximum value of the likelihood function over all possible parameter values assuming that the null hypothesis is true.
- The maximum value of the likelihood function over a larger set of possible parameter values (possible parameters for a "full" or more general model).

Suppose that we wish to test $H_O : \beta = 0$.

- Let $l_1 =$ the maximum value of the likelihood function for the full model:

$$\log(\mu_i) = \alpha + \beta x_i$$

- Let $l_0 =$ the maximum value of the likelihood function for the model when H_O is true:

$$\log(\mu_i) = \alpha$$

Model under H_O places more restrictions on the set of possible

Likelihood Ratio Test Statistic

The Likelihood ratio test statistic equals

$$\begin{aligned} -2\log(l_0/l_1) &= -2\{\log(l_0) - \log(l_1)\} \\ &= -2(L_0 - L_1) \end{aligned}$$

where $L_0 = \log(l_0)$ (and $L_1 = \log(l_1)$) are the "maximized log-likelihood functions".

- When $l_0 = l_1$ so $L_0 = L_1$, the test statistic equals 0.
- If H_O is true, then the likelihood ratio test statistic is approximately chi-squared distributed with degrees of freedom equal to $df = 1$ (for $H_O : \beta = 0$).
- When the null is false, $l_0 < l_1$ (so $L_0 < L_1$), the test statistic is > 0. The larger the statistic, the greater the evidence against the null hypothesis being true.

Example using Likelihood Ratio Test Statistic

AIDs example:

- $H_O : \beta = 0$ versus $H_A : \beta \neq 0$.
- $$-2(L_O - L_1) = -2(383.2532 - 478.3435) = 190.1806,$$

  which with $df = 1$ has a very small p–value.

Where to get these values come from?

- Under "Criteria For Assessing Goodness of Fit", SAS/GENMOD provides the maximized log-likelihood value (see "Log Likelihood")

Criterion	DF	Value	Value/DF
Deviance	12	17.0917	1.4243
Pearson Chi-Square	12	15.9884	1.3324
Log Likelihood		478.3435	

- Or using the "type3" option to the model statement:
  model count = lmonth / dist=count link=log type3;

. . . and in R

```
# Fit model with log(month)
aids$log.month ← log(aids$month)
poi2 ← glm(count ∼ log.month,data=aids,
family=poisson(link="log"))
summary(poi2)

# Compare models with and without predictors
anova(poi2)

          Df  Deviance  Resid.  Df  Resid.  Dev
 NULL                        13          207.272
 log.month  1    190.18       12           17.092
# or
lr ← poi2$null.deviance - poi2$deviance
df ← poi2$df.null - rpoi2$df.residual
pval ← 1 - pchisq(lr, df)
```

Assessing Model Fit to Data

In assessing whether a model fits the data well (i.e., the model provides a "good" or accurate description of the data), we should examine/consider

- Global Fit Statistics.
- Residuals.
- Confidence Intervals
- Overdispersion.

Global Fit Statistics

The null hypothesis for these statistics is

H_O : *The model fits the data.*
The lack of fit is not statistically large.

Let

- $i = 1, \ldots, N$ index the levels of the explanatory variable.
- $y_i =$ observed count for ith level of the explanatory variable.

The Pearson "goodness-of-fit" statistic is

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

The Likelihood ratio "goodness-of-fit" statistic is

$$G^2 = 2 \sum_{i=1}^{n} y_i \log(y_i / \hat{\mu}_i)$$

Distribution of Global Fit Statistics

If

1. The null hypothesis is true,
2. The fitted values are large $\longrightarrow \hat{\mu}_i \geq 5$,
3. The number of levels of the explanatory variable(s) is fixed,

Then the sampling distributions of X^2 and G^2 are approximately chi-squared with degrees of freedom (or "*residual df*") equal to

$$df = \text{number of counts} - \text{number of (unique) model parameters.}$$

If (1) is false (but (2) and (3) hold), then we expect X^2 and/or G^2 to be "large" (i.e., far out in the right tail of the proper chi-squared distribution) and this is evidence against the null hypothesis.

AIDs Example & Global Fit Statistics

For the AIDs example for the model $\log(\mu_i) = \alpha + \beta x_i$,

Statistic	df	Value	
X^2	12	15.99	
G^2	12	17.09	(R glm gives G^2, "Residual deviance")

Note: $N = 14$, we have just 1 observation (count) for each quarter (3 month period), and the number of parameters $= 2$, so

$$df = 14 - 2 = 12$$

Wonder why SAS and R don't automatically provide p–values for these statistics?

Problem with the AIDs Example

Problems with global fit statistics and Poisson regression:

1. Often there are (many) small "cell" counts.
   $\rightarrow$ In the AIDs data, 6 of the 14 counts are less than 5.
2. The chi-squared approximation for X^2 and G^2 is based on asymptotic theory, which means that as sample size increases for a fixed N (number of levels of the explanatory variable), the approximation becomes better.
   $\rightarrow$ If we add observations in the AIDs example, we also increase N, the number of levels of the explanatory variable.

Because of these two problems,

- X^2 and G^2 are not good measures of a model's "badness-" or "lack-of-fit"
- The approximation of the distributions of X^2 and G^2 by the chi-squared distribution is bad.

Lung Cancer Example

- Table doesn't increase as add more observations, because the Explanatory Variables are :
  - City in Denmark (Fredericia, Horsens, Kolding, Vejle).
  - Age class (40–54, 55–59, 60–64, 65–69, 70–74, $>$75).
- Of the 24 cells, smallest ones are 2 and 4; that is, 92% are "large".

Criterion	DF	Value	Value/DF	p
Deviance (G^2)	20	26.2815	1.3141	.16
Pearson Chi-Square (X^2)	20	24.2465	1.2123	.23

Ways to deal with Problem

1. Use small sample methods.
2. "Discretize" or collapse levels of the explanatory variable.
   There are two variants of this latter strategy:
   1. Sum observed counts and fitted values within the same category of the collapsed explanatory variable and recompute the test statistics.
   2. Sum observed counts within the same category of the explanatory variable and re-fit the model to the data and using scores for each category of the explanatory variable (e.g., means, or other).
3. Bayesian methods? (I put a quick little example of how fit models in SAS/GENMOD)

Grouping Observations

By grouping observations on the basis of the explanatory variable, X^2 and G^2 should be better approximated by chi-squared distributions.

By collapsing data, both problems are solved:

1. The observations/counts per level of the explanatory variable increases.
2. As you increase the number of observations per category/level of the explanatory variable, the number of levels of the explanatory variable is constant (i.e., N is fixed).

Will this strategy work for the AIDs example?

No, because of the nature of the study. If you collect more data, you necessarily increase N.

Collapsing strategies work with the horseshoe crab example (Agresti, 1996).

Collapsing Horseshoe Crab Data

Before collapsing, there are 66 different widths of female crabs (many with very small counts).

8 categories of widths were used and both of the two of the collapsing strategies were used.

Deciding on categories for explanatory variable:

- Often just take equal spacing (as done in horseshoe example, i.e., 1 cm) — easy, works well when observations are equally spread out over the range of the explanatory variable.
- Each of categories should have $\hat{\mu}_i \geq 5$.

Method I: Fit Model then Collapse

Both the observations and fitted values from the model (fit to the uncollapsed data) within the same width categories were summed. Using the summed counts and fitted values, the model test statistics were recomputed:

Statistic	df	Value	p–value
X^2	6	6.5	—
G^2	6	6.9	—

Note:
df = (# of categories)−(# of parameters)= $8 - 2 = 6$.

Method II: Collapse then Fit Model

Sum the observed counts within each of the width categories and re-fit the model to collapsed data using some scores for the width categories (mean width within category).

With this method, each observation within a category is treated as if it has the same width. (Method II is "easier" than Method I, but Method II is also "cruder" than Method I).

The Poisson regression with log link and the offset $t = \log$(number of cases per width category):

$$\log(\hat{\mu}_i/t) = -3.535 + .173x_i$$

and the estimated model from un-collapsed data was

$$\log(\hat{\mu}_i) = -3.305 + .164x_i$$

The fit statistics for the model fit to the collapsed data:

Statistic	df	Value	p–value
X^2	6	6.5	.37
G^2	6	6.9	.33

Method II: continued

Fitted counts look pretty close to the observed counts

Width	Mean Width	Number Cases	Number Counts	Fitted Count
< 23.25	22.69	14	14	20.5
23.25–24.25	23.84	14	20	25.1
24.25–25.25	24.77	28	67	58.9
25-25–26.25	25.84	39	105	98.6
26.25–27.25	26.79	22	63	65.6
27.25–28.25	27.74	24	93	84.3
28.25–29.25	28.67	18	71	74.2
> 29.25	30.41	14	72	77.9

Conclusion: Both Methods I & II yield indicate that the Poisson regression model looks pretty good for these data.

Residuals

Pearson Residuals are standardized differences between observed counts and fitted values:

$$\text{Pearson residual} = \frac{(\text{observed} \ - \ \text{fitted})}{\sqrt{\hat{\text{V}}\text{ar}(\text{observed})}}$$

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

Notes:

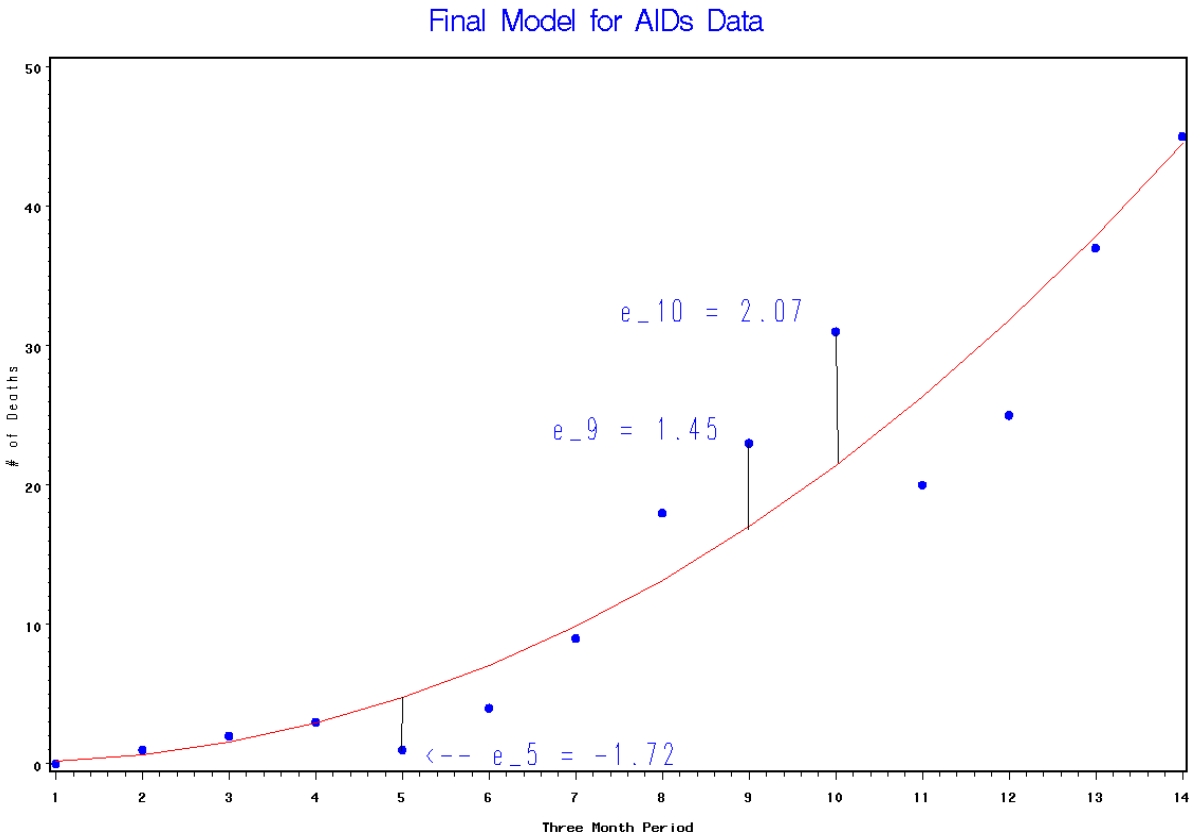
- The estimated standard deviation of a fitted value in Poisson regression is the square root of the fitted value.
- $\sum_i e_i^2 = X^2$
- Observations with larger Pearson residuals make larger contributions to X^2.

Residuals from AIDs example

Month x_i	# Deaths y_i	Fitted Value	Pearson residual	Contribution to X^2
1	0	.14	-.38	.14
2	1	.65	.44	.19
3	2	1.56	.35	.12
4	3	2.92	.05	.00
5	1	4.74	-1.72	2.95
6	4	7.05	-1.15	1.32
7	9	9.86	-.27	.07
8	18	13.17	1.33	1.77
9	23	17.02	1.45	2.10
10	31	21.40	2.07	4.30
11	20	26.33	-1.23	1.52
12	25	31.82	-1.29	1.46
13	37	37.87	-1.14	.02
14	45	44.49	.08	.01

Residuals from AIDs example (plot)

The largest residual $e_{10} = 2.07$, with the next largest $e_5 = -1.72$, which taken together contribute $4.30 + 2.95 = 7.25$ to X^2 for the model, which is about 45% of $X^2 = 15.99$.



Adjusted Residuals

When a model fits data, the Pearson residuals should be approximately normally distributed with mean 0 but the variance is slightly less than 1.
$(y_i - \hat{\mu}_i)$ tends to be smaller than $(y_i - \mu_i)$, because sample data are used to obtain $\hat{\mu}_i$.

$$\begin{aligned} \text{Adjusted residual} &= \frac{\text{Pearson residual}}{\text{Pearson residual's standard error}} \\ &= \frac{e_i}{\sqrt{(1-h_i)}} = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\mu}_i(1-h_i)}} \end{aligned}$$

where h_i is "*leverage*", which is a measure of how much an observation potentially influences the fit of the model.
Adjusted residuals are

- Approximately $\mathcal{N}(0,1)$ when the model holds.
- Good for finding "large" residuals.

AIDs Adjusted Residuals

Month x_i	# Deaths y_i	Fitted Value	Pearson residual	Adjusted Residual
1	0	.14	-.38	-.39
2	1	.65	.44	.46
3	2	1.56	.35	.38
4	3	2.92	.05	.05
5	1	4.74	-1.72	-1.86
6	4	7.05	-1.15	-1.24
7	9	9.86	-.27	-.29
8	18	13.17	1.33	1.41
9	23	17.02	1.45	1.53
10	31	21.40	2.07	2.19
11	20	26.33	-1.23	-1.32
12	25	31.82	-1.29	-1.33
13	37	37.87	-1.14	-1.16
14	45	44.49	.08	.10

Confidence intervals for fitted values

Just as in normal linear regression, we can put confidence intervals around out Poisson regression fitted values.

- $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \Sigma_{\hat{\beta}})$.
- The linear predictor is a linear combination of $\hat{\boldsymbol{\beta}}$:

$$\boldsymbol{x}_i' \hat{\boldsymbol{\beta}}$$

e.g., For one explanatory variable,

$$\boldsymbol{x}_i' \hat{\boldsymbol{\beta}} = (1, x_i) \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \hat{\alpha} + \hat{\beta} x_i$$

Therefore the linear predictor is normally distributed,

$$\boldsymbol{x}_i' \hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{x}_i' \boldsymbol{\beta}, \sigma^2)$$

where $\sigma^2 = \boldsymbol{x}_i' \Sigma_{\hat{\beta}} \boldsymbol{x}_i$.

Confidence intervals for Regression

- Since the linear predictor is normal, then a $(1-\alpha)100\%$ confidence interval for $\log(\mu_i)$ is

$$\log(\hat{\mu}_i) \pm z_{\alpha/2}\sqrt{\sigma^2}$$

- A $(1-\alpha)100\%$ confidence interval for μ_i is

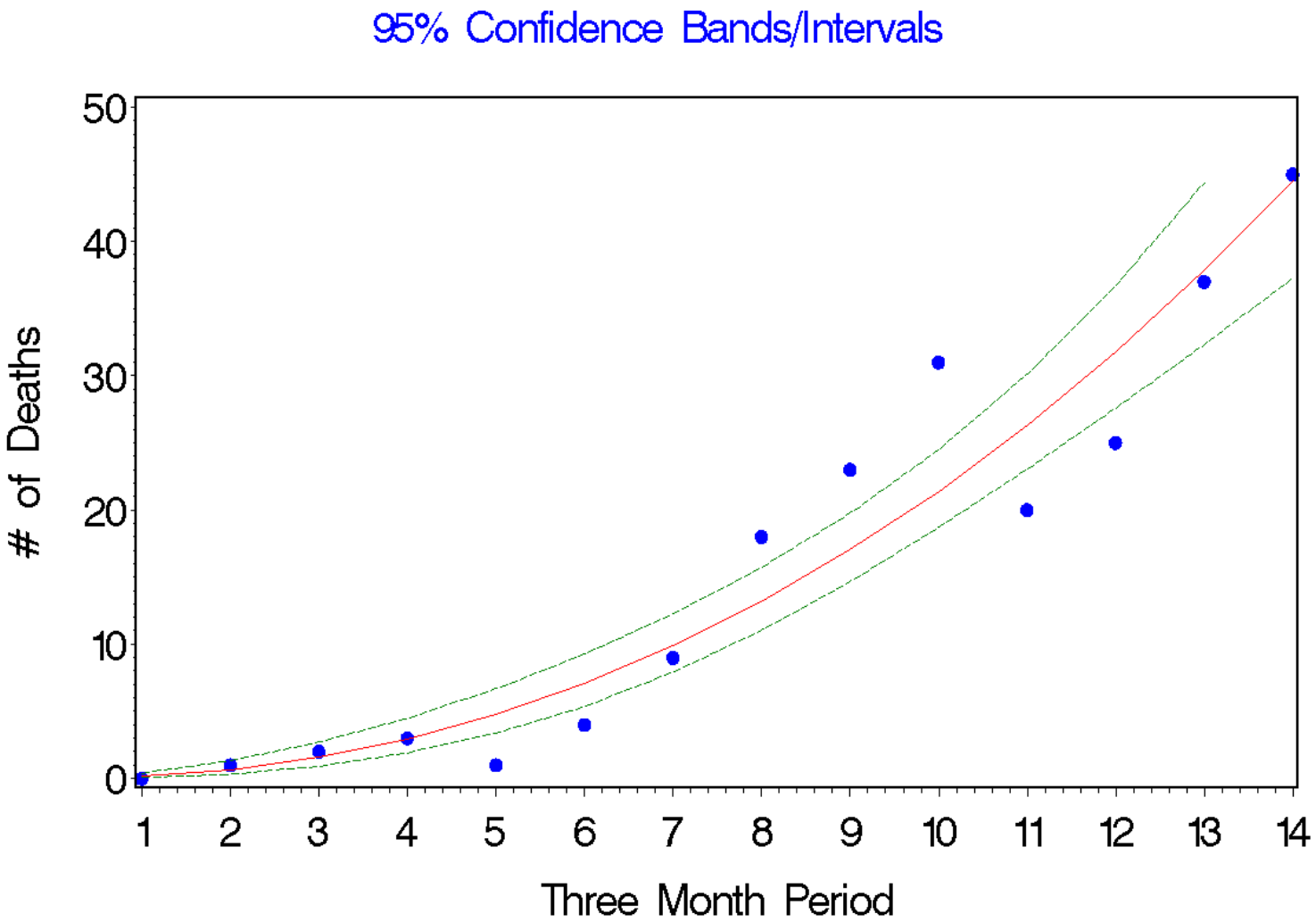
$$\exp\left[\log(\hat{\mu}_i) \pm z_{\alpha/2}\sqrt{\sigma^2}\right]$$

- On the SAS/GENMOD output under the observation statistics, UPPER and LOWER correspond to the upper and lower ends of confidence intervals for μ_i.
- In R...

Confidence intervals for Regression

```
# to get 95% bands on model fitted values
test ← predict(mod.poi1, newdata=NULL,
           type=c(‘‘response’’),
           se.fit=TRUE)
names(test)
upper ← test$fit + 1.96*test$se.fit
lower ← test$fit - 1.96*test$se.fit
```

AIDs Confidence Bands for Regression



Overdispersion

Observed count data often show greater variability than would be expected if the data were really from a Poisson (or binomial) distribution.

If data come from a Poisson distribution, then mean = variance

But often we find that mean < variance

This situation is referred to as overdispersion.

Most common causes of overdispersion: Heterogeneity (and lack of independence).

In Poisson regression, we assume that the randomness of observations on individuals with the same value on the explanatory variable(s) can be described by the same Poisson distribution (i.e., the same mean).

Example of Overdispersion

Consider the example of the number of violent incidences of individuals with mental illnesses who had been treated in ER of a psychiatric hospital.
If the Poisson model for counts is correct, then for each patient with the same age, concern score, and history of violent incidences, the expected count should equal

$$\hat{\mu}_i = t_i \exp\{\hat{\alpha} + \hat{\beta}_1 \text{age}_i + \hat{\beta}_2 \text{concern}_i + \hat{\beta}_3 \text{history}_i\}$$

where t_i is the offset (the number of days the individual was in the community during the 6 month period of the study).
And the estimated probability of observing counts equal to $y = 0, 1, 2, \ldots,$ should be given by

$$\hat{P}(Y_i = y) = \frac{e^{-\hat{\mu}_i} \hat{\mu}_i^y}{y!}$$

However, the observed variability is greater than $\hat{\mu}_i$.

Overdispersion: When to be Concerned

When the random component is a distribution where the mean and variance are related.

In particular, overdispersion is

- a concern with Poisson: $\sigma^2 = \mu_i$.
- a concern with Binomial: $\sigma^2 = N\mu_i(1 - \mu_i)$.
- not a concern with Normal: σ^2 is not a function of μ_i.

Detecting Overdispersion

For grouped data, you can compute the mean and sample variance of the counts. If the Poisson distribution is a good model for the data, then mean $=$ variance, but if

$$\text{sample mean} < \text{sample variance} \longrightarrow \text{overdispersion}$$

For an example, see Agresti. He illustrates this for the crab data.
For grouped (e.g., crab data grouped into 8 categories instead of 66 different values of width) or ungrouped data (e.g., AIDs and violent incidence examples), if the Poisson model is a good one, then Pearson's X^2 divided by df should equal 1, but if

$$X^2/df > 1 \longrightarrow \text{ overdispersion}$$

Examples: Detecting Overdispersion

$$X^2/df > 1 \longrightarrow \text{ overdispersion}$$

Data Set	df	X^2	X^2/df
Horseshoe crabs	64	174.3	2.7
Deaths due to AIDs	12	15.99	1.33
Violent incidences	793	12711.79	16.03
Hodgkin's disease	38	9956.24	276.94

Implication of Overdispersion

& How to deal with it.

If there is extra variation in the data, then estimates of variances and standard errors for the estimated model parameters are too small. When estimated standard errors are too small, test statistics for testing hypotheses such as $H_o : \beta = 0$ are too big (i.e., "inflated").

We'll discuss two ways to deal with this.

- Adjust estimated standard errors — When you're primarily concerned with testing hypotheses regarding parameter estimates.
- Use an alternative distribution as your random component (i.e., model the extra variability) — When you're concerned with prediction.

Adjusting estimated standard errors

An estimate of the extra variance is Pearson's X^2 for the model divided by it's degrees of freedom.

$$X^2/df$$

To adjust the ASE for parameter estimates we multiply them by

$$\text{Adjusted } ASE = \sqrt{X^2/df}(ASE)$$

Example: Violent incidences — multiply ASE by $\sqrt{16.03} = 4.00$.

	Est.	Uncorrected			Corrected		
Coefficent	Param	ASE	z	p-value	ASE	z	p-value
Intercept	-3.410	.0690	-49.29	$< .0001$	.2800	-12.31	$< .0001$
Age	-.045	.0023	-19.69	$< .0001$	.0091	-4.92	$< .0001$
Concern	.083	.0075	11.20	$< .0001$	.0300	2.80	$< .0051$
History	.420	.0380	11.26	$< .0001$	.0150	2.81	$< .0051$

Same conclusion (the effects are very strong in this case).

Modeling the Extra Variability

When you're concerned with prediction (as in the violent incidence of individuals with mental illnesses), simply adjusting ASE for hypothesis testing is not enough.

Model the extra variability: $\tilde{\mu}_i$ (the parameter of the Poisson distribution) is considered a random variable. Even after taking into account the linear predictor, there is still some variability in $\tilde{\mu}_i$ not accounted for, i.e.,

$$\tilde{\mu}_i = t_i \exp\{\alpha + \sum_j \beta_j x_{ij}\}\epsilon_i$$

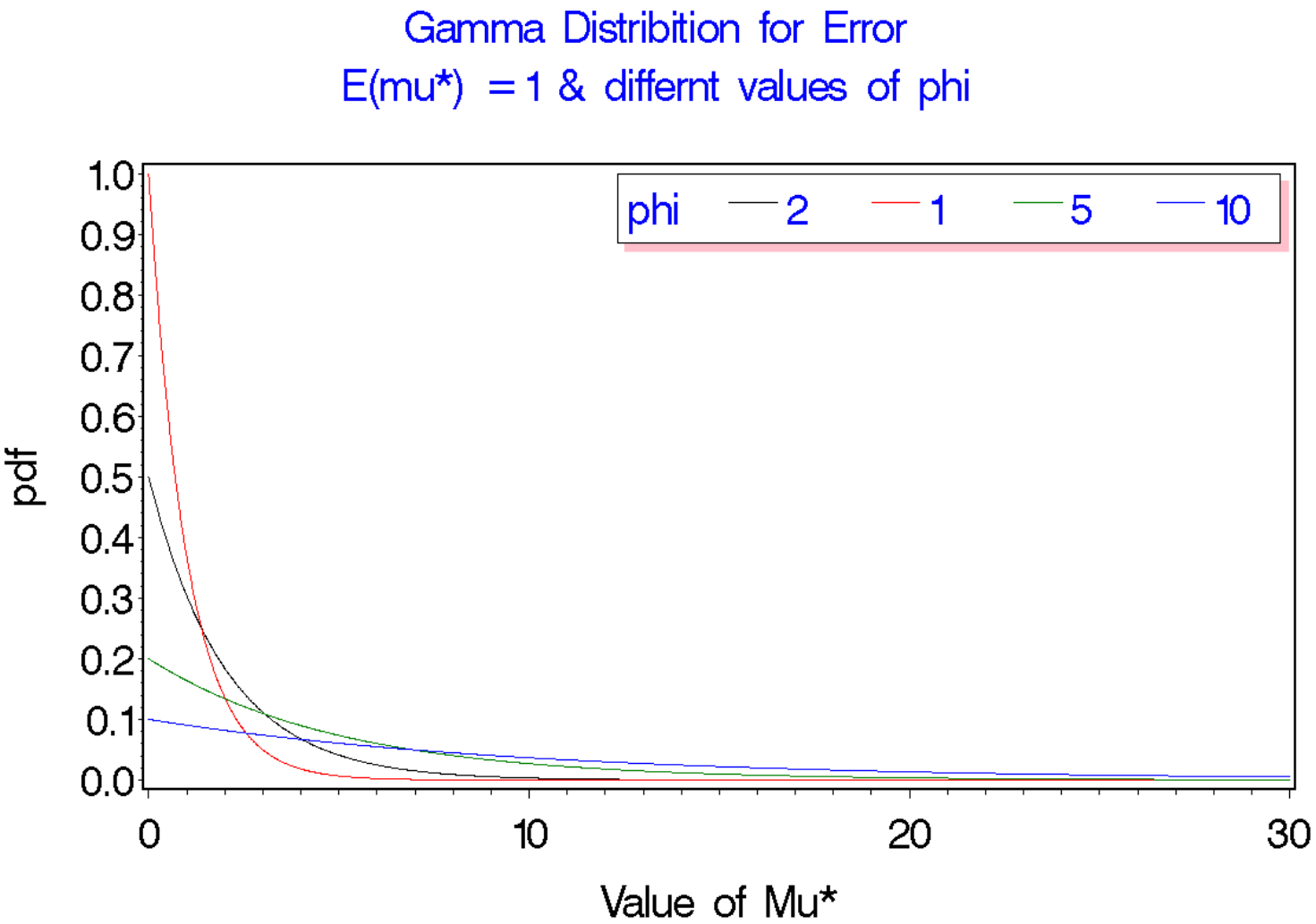
where $\epsilon_i > 0$ is an unobserved random variable, and t_i is the offset (if there is one).

The probability distribution assumed for ϵ_i is usually a Gamma distribution (for mathematical convenience) with

$$E(\epsilon_i) = 1 \qquad \text{and} \qquad \text{Var}(\epsilon_i) = 1/\phi$$

So,. . .

Gamma Distributions with $\mu = 1$



Modeling the extra variability

$$E(\tilde{\mu}_i) = E(t_i \exp\{\alpha + \sum_j \beta_j x_{ij}\}\epsilon_i) = t_i \exp\{\alpha + \sum_j \beta_j x_{ij}\}$$

and

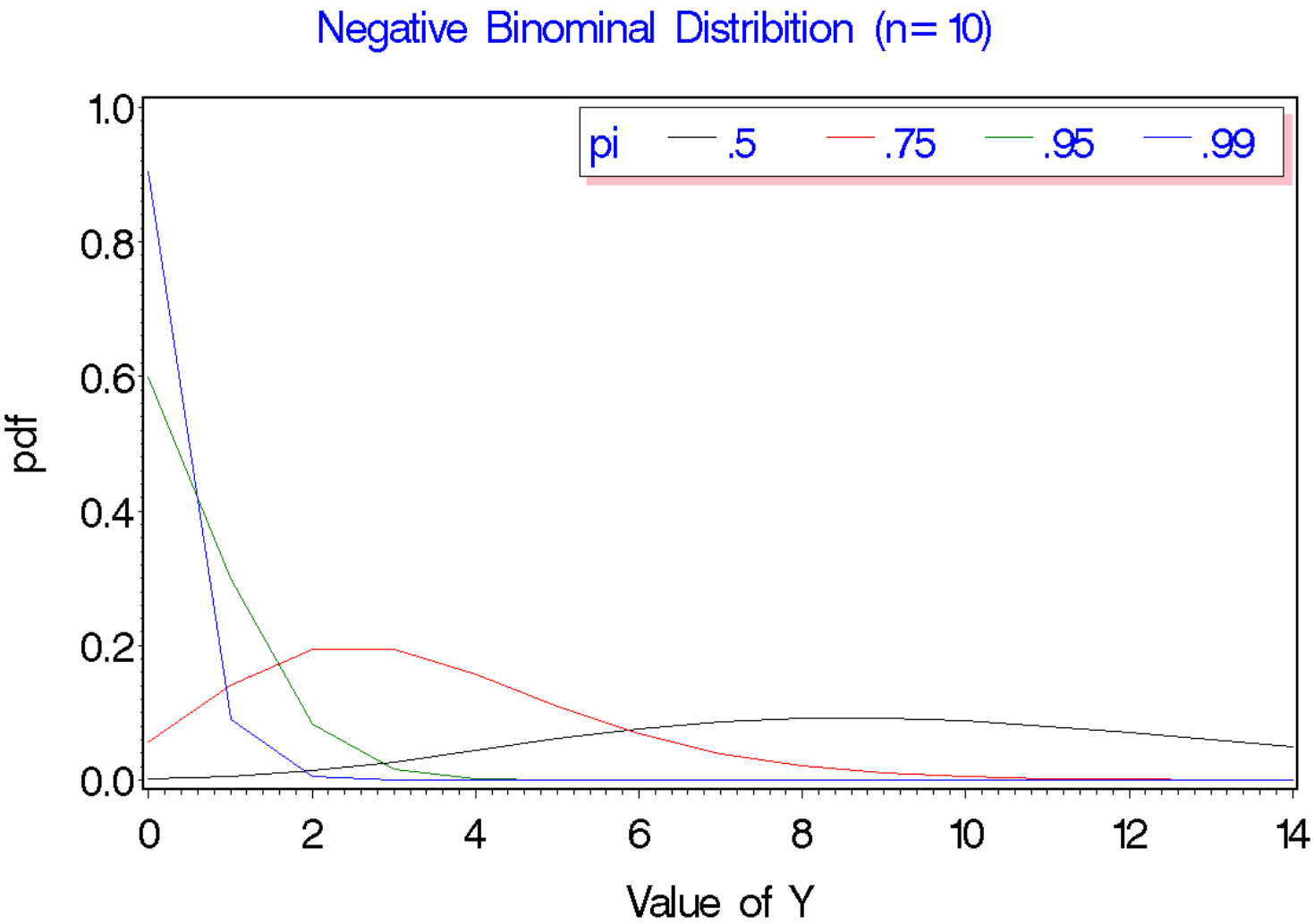
$$\text{Var}(\tilde{\mu}_i) = \mu_i^2/\phi$$

and the variance of observed counts $\mu_i + \mu_i^2/\phi$.

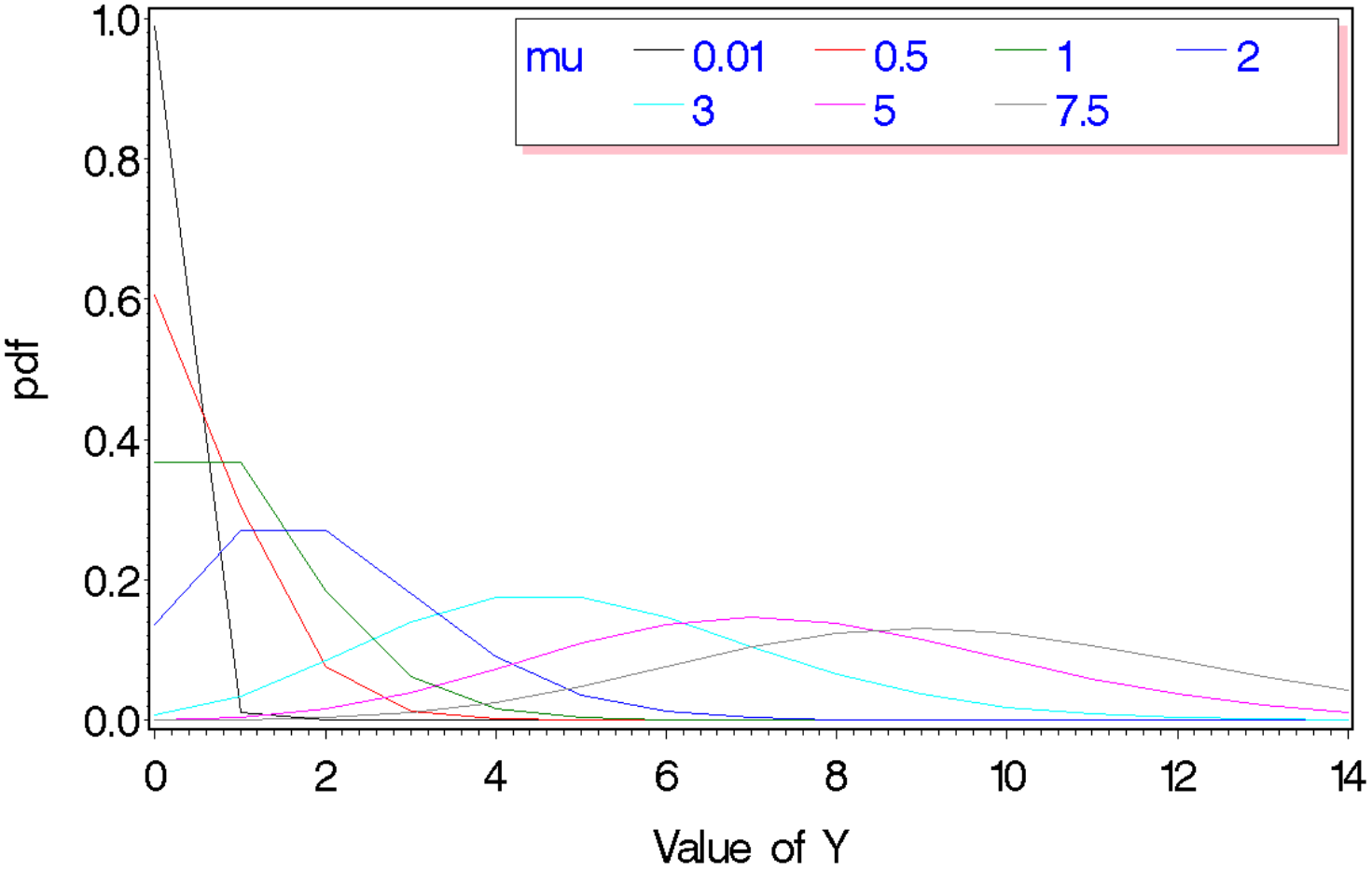
If $\tilde{\mu}_i$ is known, then the distribution of counts y_i would be Poisson with parameter $\tilde{\mu}_i$.

Since $\tilde{\mu}_i$ is not known and is a random variable, the distribution for y_i is a Negative Binomial Distribution.

Negative Binomial Distribution



Poisson Distribution



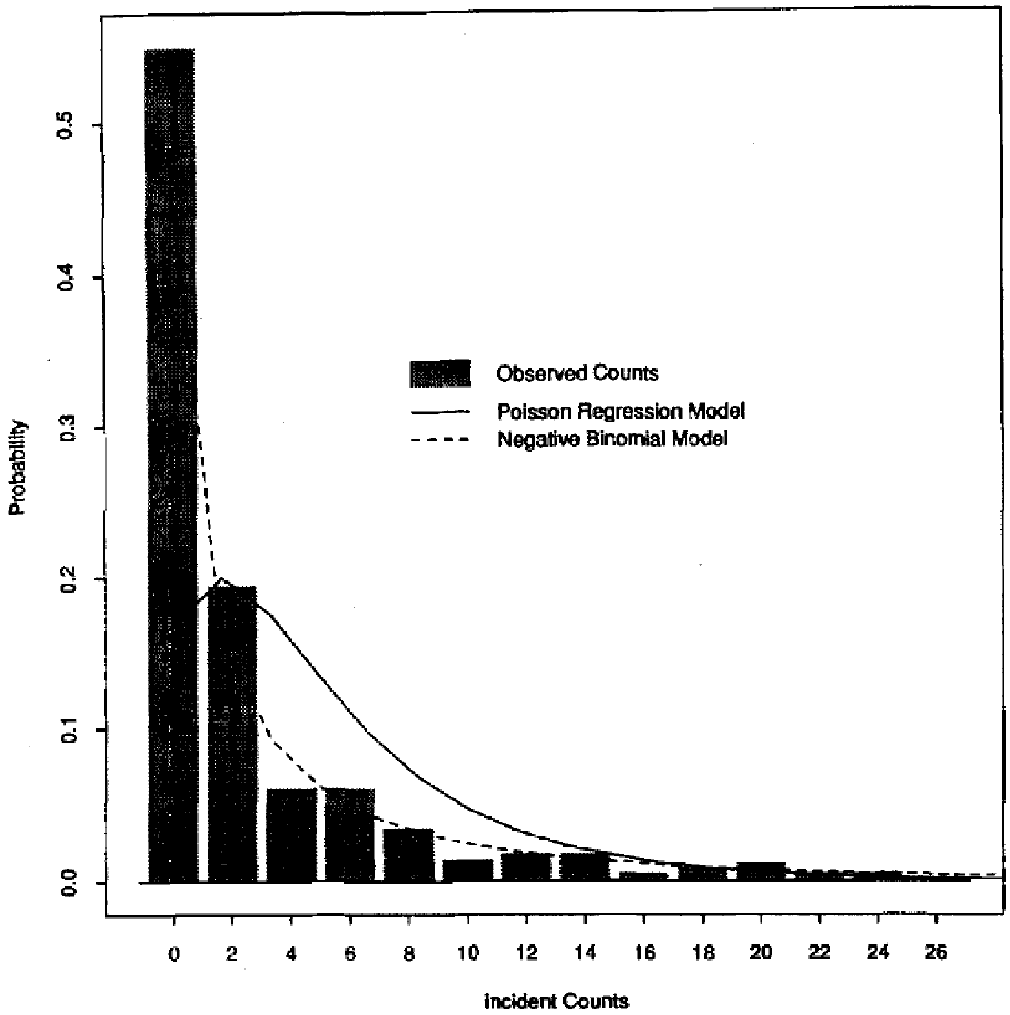
Example: Modeling the extra variability

For the violent incidents example, when the negative binomial distribution is used as the random component of the GLM, the following estimated parameters are obtained:

Coefficient	Estimate	ASE	z	p–value
Intercept	-3.5500	.26200	-13.55	$< .0001$
Age	-.0459	.00799	-5.74	$< .0001$
Concern	.0962	.03090	3.12	.0029
History	.5360	.11550	3.45	.0008

which is similar to before, but the probability distribution is much better approximated by the Negative Binomial.

Improvement when Modeling the extra variability



Overdispersion and SAS/GENMOD

- Built in MODEL options

  From the SAS/GENMOD documentation:

  - SCALE = number
  - PSCALE sets scaling parameter equal to 1 during estimation but standard errors and statistics are adjusted using Pearson's X^2/df.
  - DSCALE same as PSCALE but uses G^2/df.

- Using Negative Binomial distribution as the random component.
  model count = < linear predictor > / link=log dist=Negbin ;

Overdispersion Example from Lindsey

(1997); *Applying Generalized Linear Models*

The data consist of counts of T_4 cells/mm in blood samples from 20 patients in remission from Hodgkin's disease and 20 other patients in remission from disseminated malignancies:

Hodgkin's Disease		Non-Hodgkin's Disease	
396	568	375	375
1212	171	752	208
554	1104	151	116
257	435	736	192
295	397	315	1252
288	1004	657	700
431	795	440	771
1621	1378	688	426
902	958	410	979
1283	2415	377	503

Question: Is the average count of T_4 cells/mm the same or different for patients in remission from Hodgkin's disease as the average count from those in remission from disseminated malignancies?

Possible Model: Poisson regression

Hodgkin's disease example (continued)

- Random component: $Y =$ number of T_4 cells/mm. Assume Poisson and if this doesn't fit, we'll try Negative Binomial.
- Systematic component: In this example, the explanatory/predictor variables is discrete so we'll define

$$X = \begin{cases} 0 & \text{if non-Hodgkin's} \\ 1 & \text{if Hodgkin's disease} \end{cases}$$

So linear predictor is

$$\alpha + \beta x$$

- Link function: $\log$.

$$\log(\mu_y) = \begin{cases} \alpha & \text{if non-Hodgkin's} \\ \alpha + \beta & \text{if Hodgkin's disease} \end{cases}$$

Possible Model: Poisson regression

Hodgkin's disease example (continued)

Fit of this model to data yields

Model	df	G^2	p-value	X^2	p-value
Poisson regression	38	9956.23	$< .001$	10523.75	$< .001$
Negative binomial	38	42.41	.29	40.26	.37

Parameter Estimates from the two models:

		Poisson Distribution			Negative Binomial		
Parameter	df	Estimate	ASE	Wald	Estimate	ASE	Wald
α	1	6.2560	.0098	407935	6.2560	.1365	2101.53
β	1	.4572	.0125	1333.90	.4572	.1929	5.62

Note: dispersion parameter $= .3706$ and ASE $= .0787$ (95% CI: 0.2163 0.5248). This was estimated by MLE.

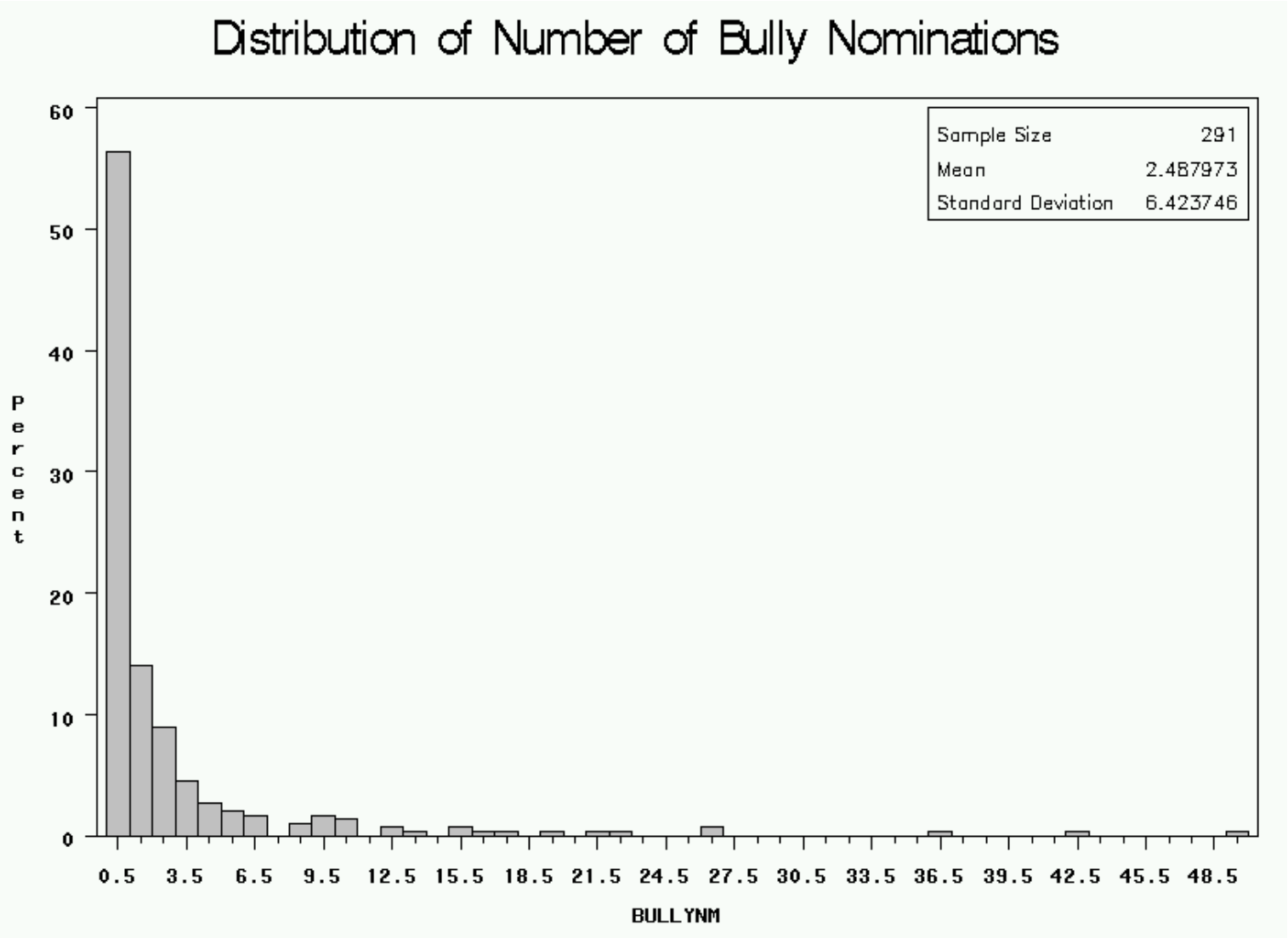
Outline of Bully Example

- The Problem
- A little exploratory analysis.
- Initial Modeling
- Revised models
- Conclusion
- Zero inflated models (something new)

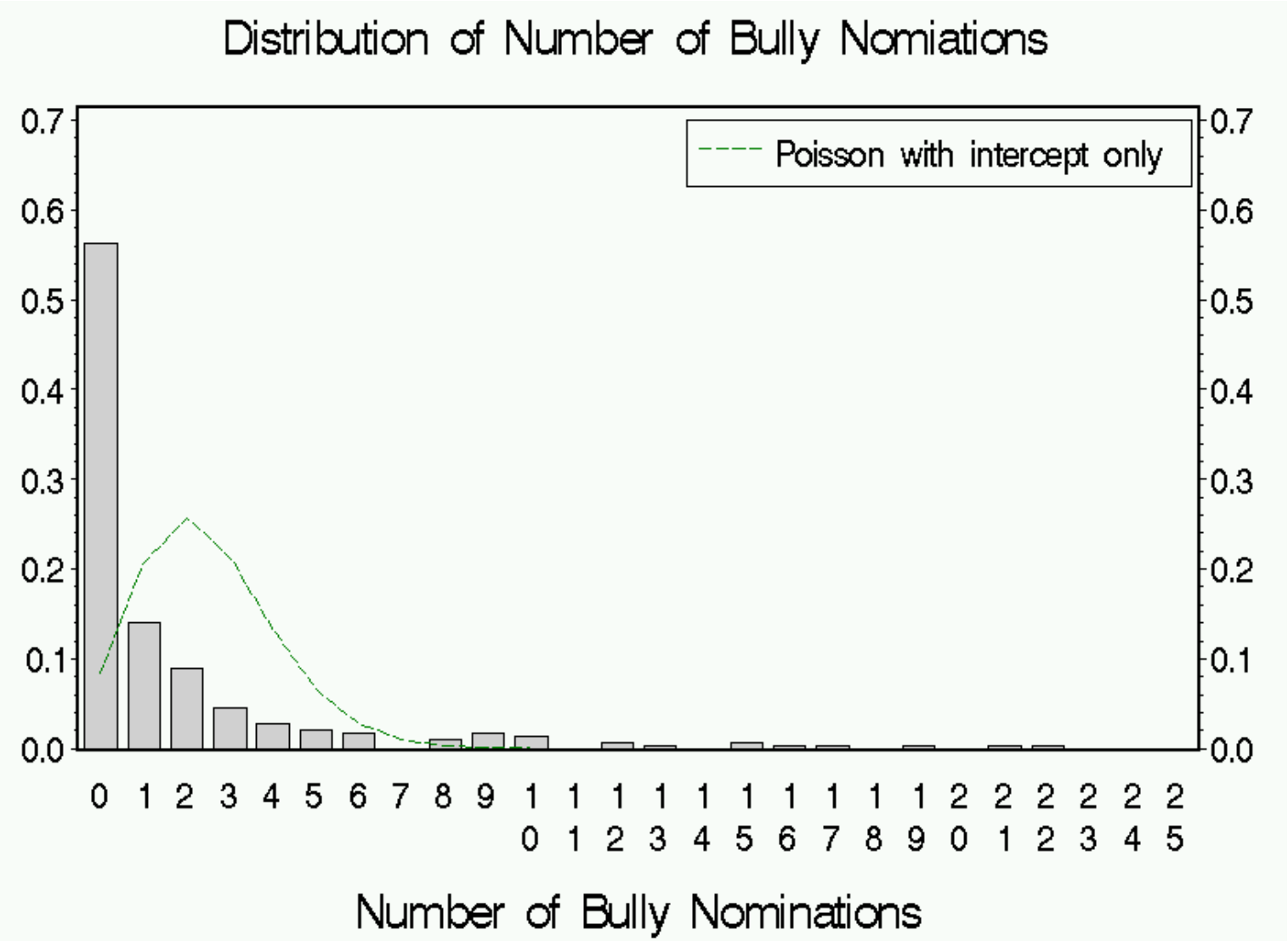
The Problem

- The data are from Espelage, D.L., Holt, M.K., & Henkel, R.R. (2004). Examination of peer-group contextual effects on aggression during early adolescence. *Child Development*, 74, 205–220.
- Two ways to measure bullying
  - Self Report: 9 item Illinois Bully Scale (Espelage & Holt, 2001).
  - Peer nominations: Kids list everyone who they view as a bully. The total number of nominations a child receives is a measure of bullying that child's bullying.
- Peer nominations more "objective" than self report and it's getting harder to obtain IRB approval of peer nominations.
- Model peer nominations (a count) with self report measure (bully scale) as a predictor variable. . . ignoring clustering. . .

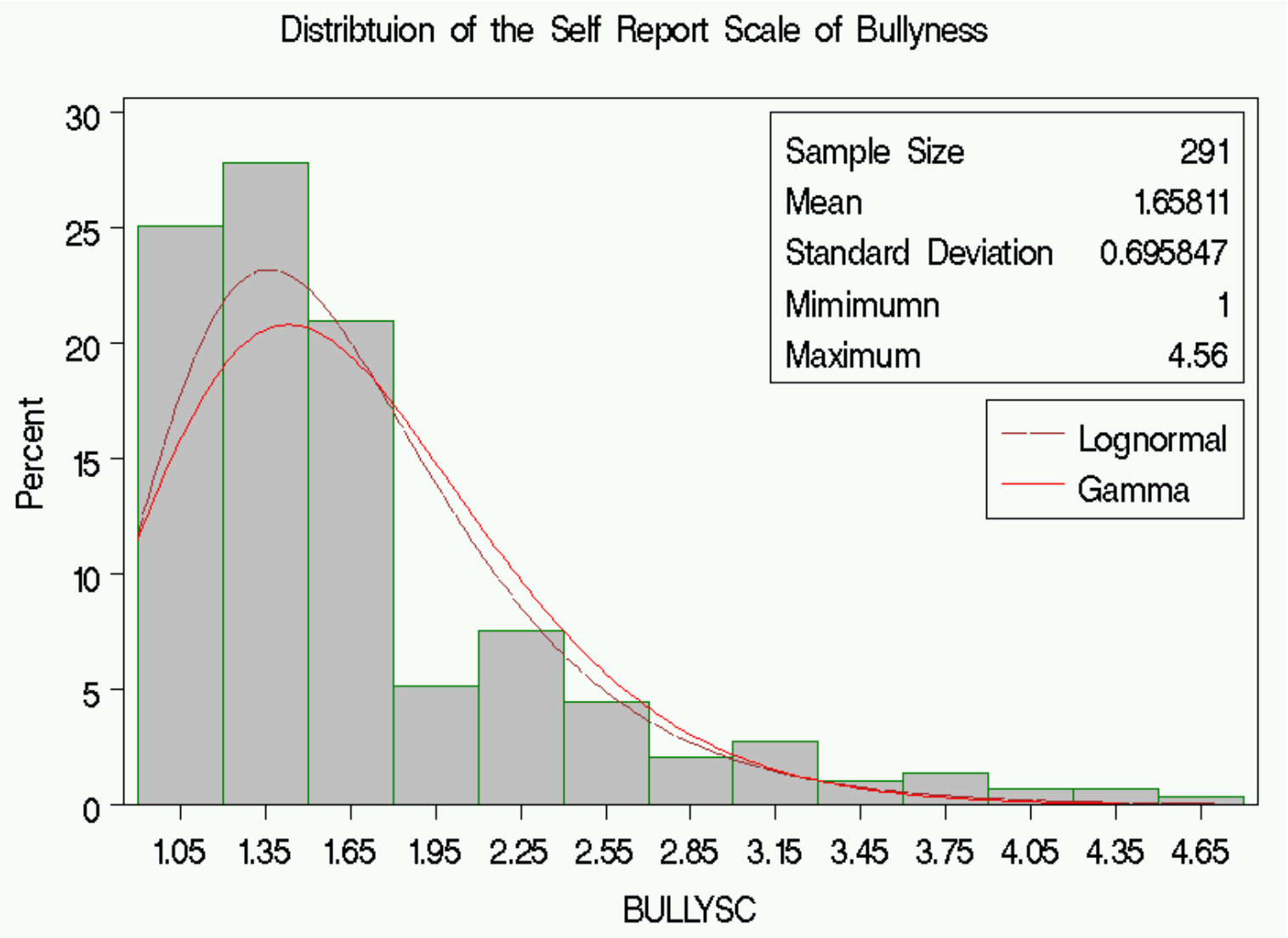
Exploratory Analysis



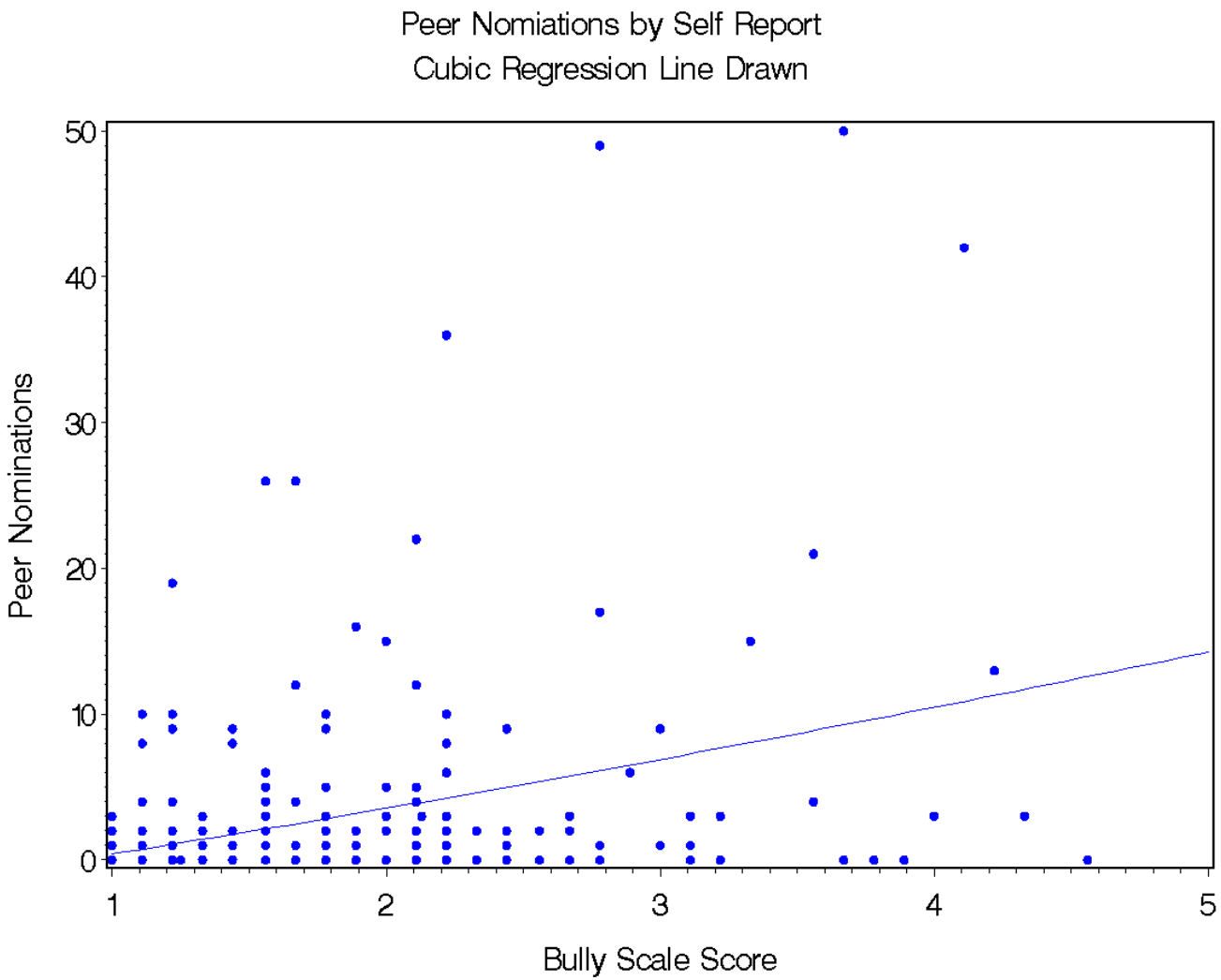
The Predictor Variable



The Predictor Variable



Relationship Between the Measures



We Know So Far That. . .

- Both variables are highly positively skewed.
- There are a lot of kids who did not receive any peer nominations.
- There does appear to be a relationship between peer nominations and scale score.
- Mean peer nominations is much smaller than the variance:

$$2.49 < 41.26$$

Initial Modeling

- Random Component:
  - Y_{ij} = the number of nominations received by kid i in peer group j.
  - Poisson distribution.
- Linear Predictor:

$$\beta_0 + \beta_1(\text{bullysc})_{ij} = \beta_0 + \beta_1 x_{ij}$$

- The Link is the Log, the canonical link.
- The initial models is a standard Poisson regression model

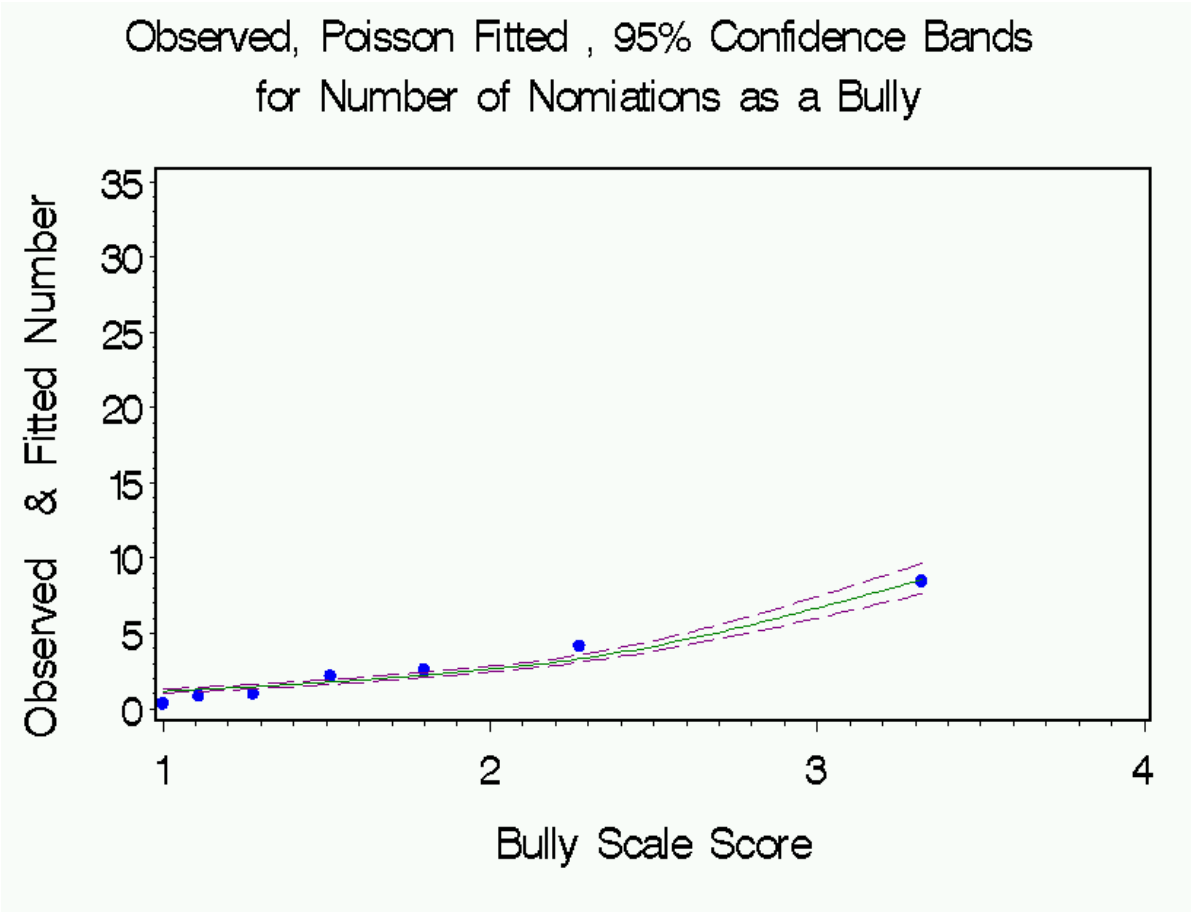
$$E(Y_{ij}) = \mu_{ij} = \exp[\beta_0 + \beta_1 x_{ij}]$$

where

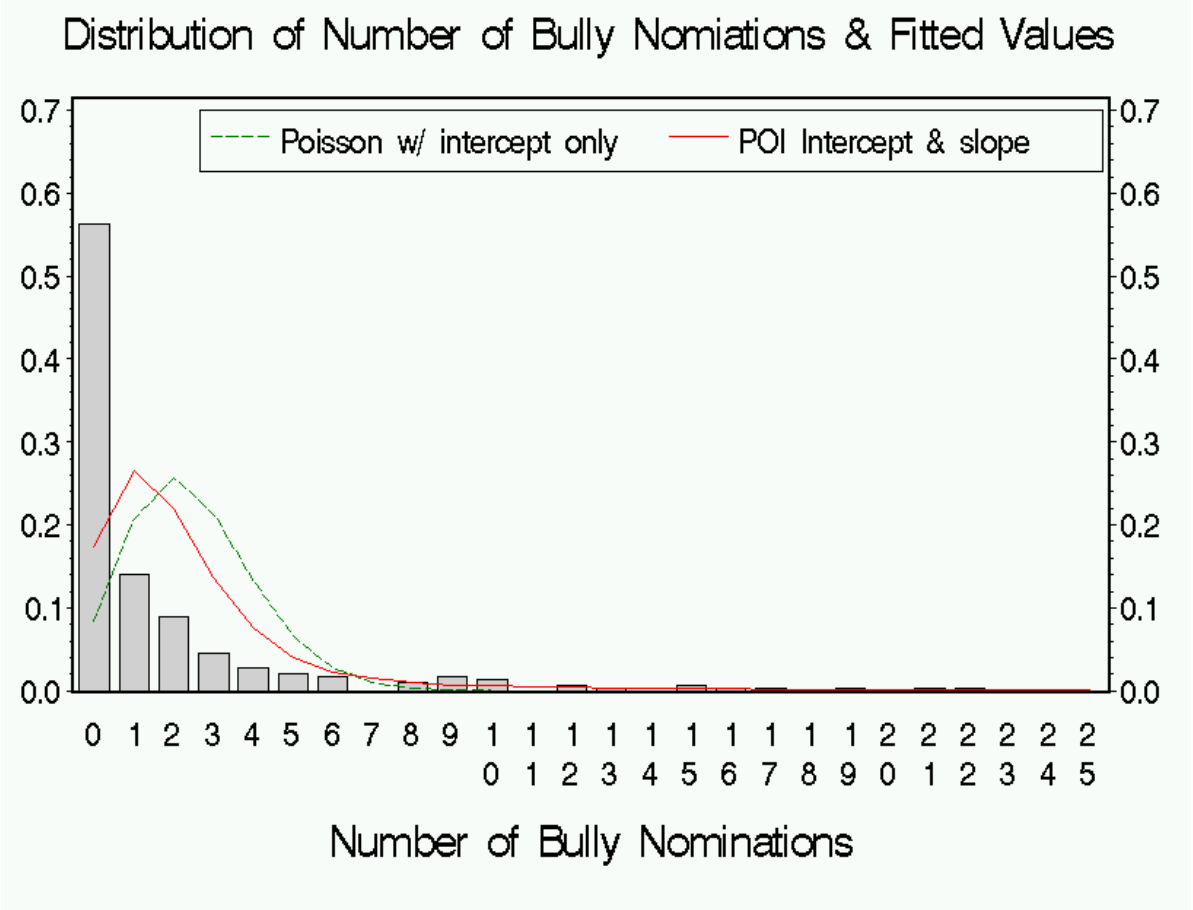
$$P(Y_{ij} = y) = \frac{e^{-\mu_{ij}}\,\mu_{ij}^{y}}{y_{ij}!}$$

Fit of Poisson Regression Model

Model was fit and then grouped to "look" at fit.



Fit of Marginal Distribution



Revised Model

To deal with the overdispersion, we'll change the random component to Negative Binomial:

- Random= Negative Binominal
- Linear predictor= $\beta_0 + \beta_1 x_{ij}$.
- log link
- Our next model is

$$Y_{ij} = \mu_{ij}\epsilon_{ij} = \underbrace{\exp[\beta_0 + \beta_1 x_{ij}]}_{\text{Poisson}} \quad \underbrace{\epsilon_{ij}}_{\text{Gamma}}$$

where

- $E(\epsilon_{ij}) = 1$
- $\text{var}(\epsilon_{ij}) = 1/\phi$ (ϕ is the "dispersion" parameter).
- $E(Y_{ij}|x_{ij}) = \mu_{ij} = \exp[\beta_0 + \beta_1 x_{ij}]$
- $\text{var}(Y_{ij}|x_{ij}) = \mu_{ij} + \mu_{ij}^2/\phi$
- and

$$P(Y_{ij} = y) = \frac{\Gamma(y+\phi)}{y!\Gamma(\phi)} \left(\frac{\phi}{\phi+\mu_{ij}}\right)^{\phi} \left(\frac{\mu_{ij}}{\phi+\mu_{ij}}\right)^{y}$$

Fit Statistics & Parameter Estimates

$df = 289$ for all of these

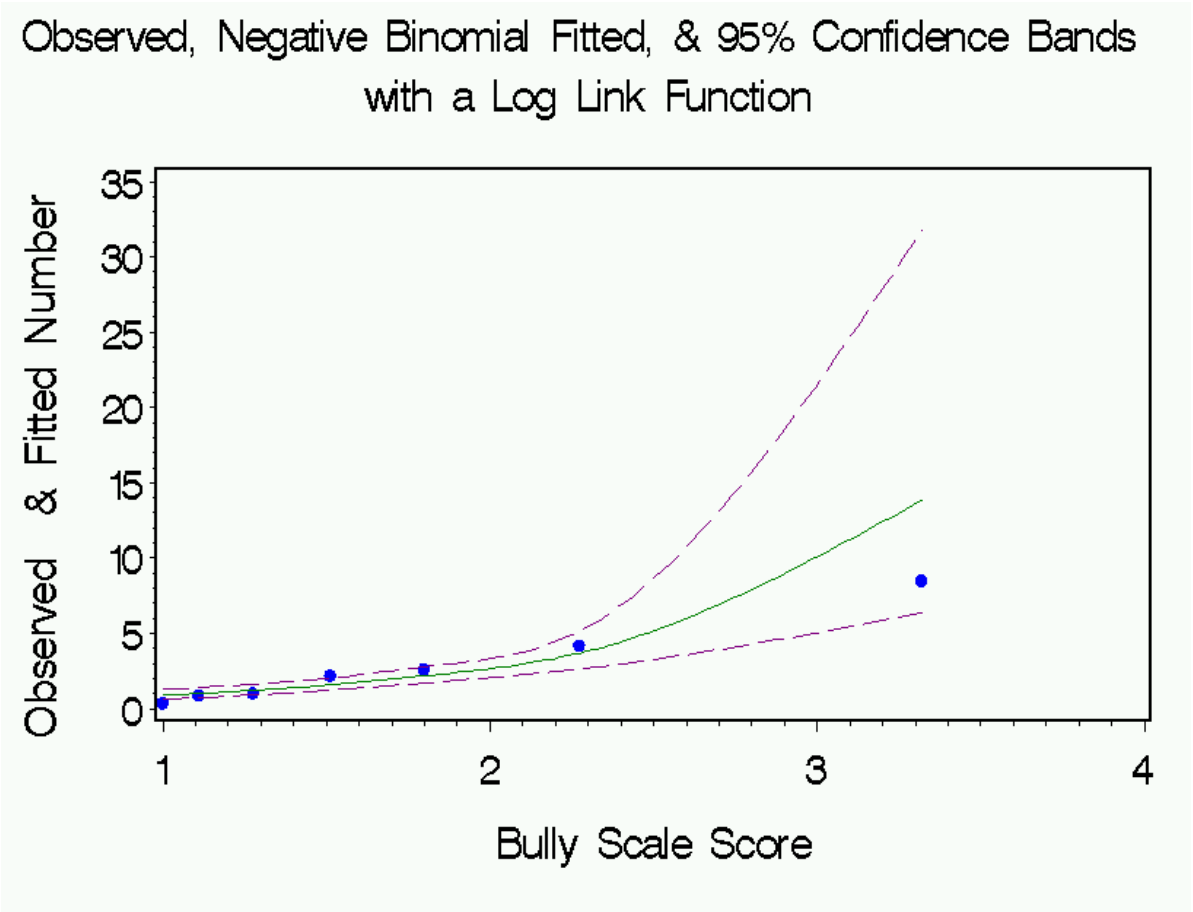
Dist	Link	G^2	X^2	X^2/df	AIC	BIC
Poisson	log	1774.60	2977.94	10.30	2160	2168
NegBin	log	244.08	351.13	1.22	998	1010

	Poisson				Negative Binomial			
Parm	est.	se	Wald	p	est.	se	Wald	p
β_0	-0.66	0.09	55.87	$< .01$	-1.19	0.36	11.30	$< .01$
β_1	0.81	0.04	540.05	$< .01$	1.09	0.20	30.66	$< .01$
$1/\phi$	—				3.50	0.44	—	

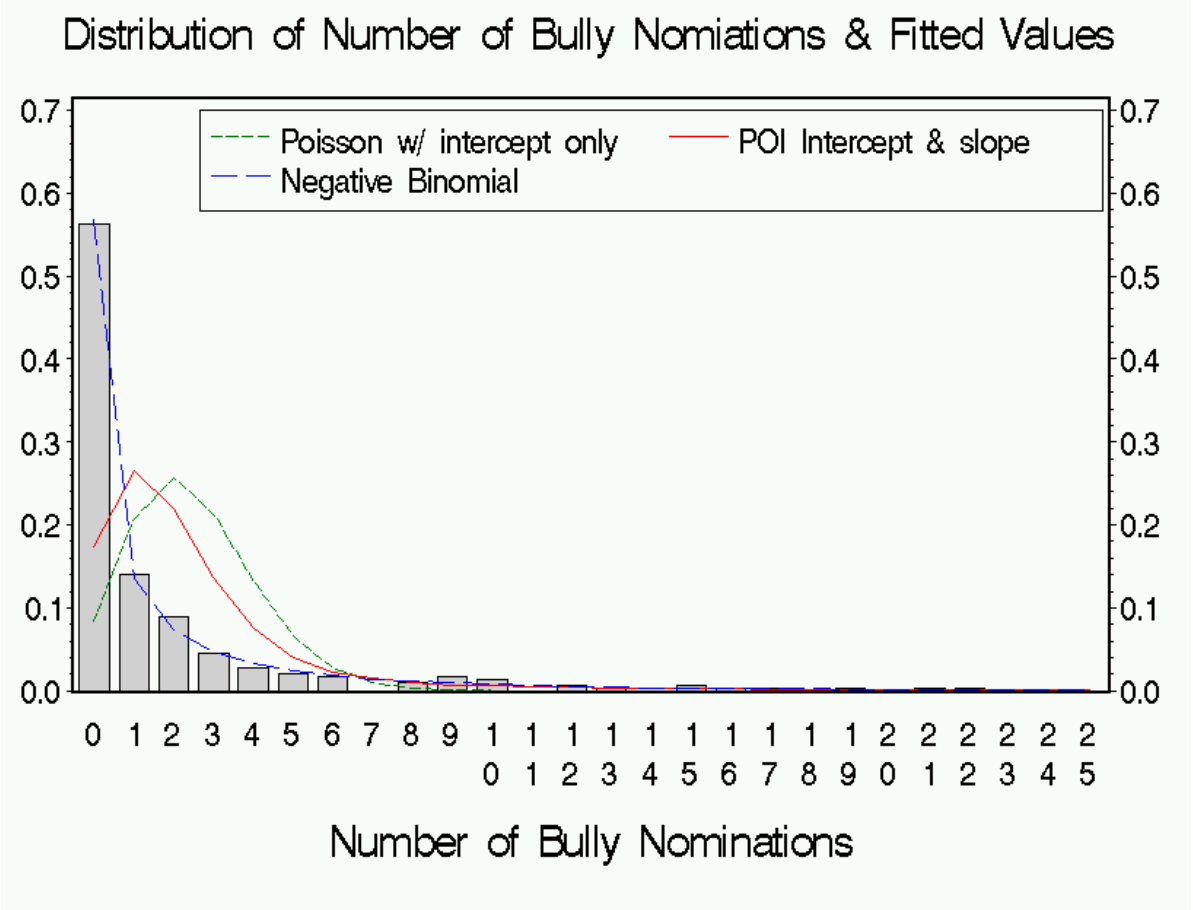
For interpretation,

$\exp(.81) = 2.25$ and $\exp(1.09) = 2.98$

Fit of Negative Binomial Model to Data



Fit of Marginal Distribution



Change of the Link Function

The relationship between Y_{ij} and x_{ij} looks like a straight line. . . The New GLM:

- Negative Binomial
- $\beta_0 + \beta_1 x_{ij}$
- Identity Link function

This model is

$$Y_{ij} = \mu_{ij}\epsilon_{ij} = \underbrace{(\beta_0 + \beta_1 x_{ij})}_{\text{Poisson}} \underbrace{\epsilon_{ij}}_{\text{Gamma}}$$

- $E(\epsilon_{ij}) = 1$
- $\text{var}(\epsilon_{ij}) = 1/\phi$
- $E(Y_{ij}|x_{ij}) = \mu_{ij} = \beta_0 + \beta_1 x_{ij}$
- and

$$P(Y_{ij} = y) = \frac{\Gamma(y+\phi)}{y!\Gamma(\phi)} \left(\frac{\phi}{\phi+\mu_{ij}}\right)^{\phi} \left(\frac{\mu_{ij}}{\phi+\mu_{ij}}\right)^{y}$$

Fit Statistics & Parameter Estimates

$df = 289$ for all of these

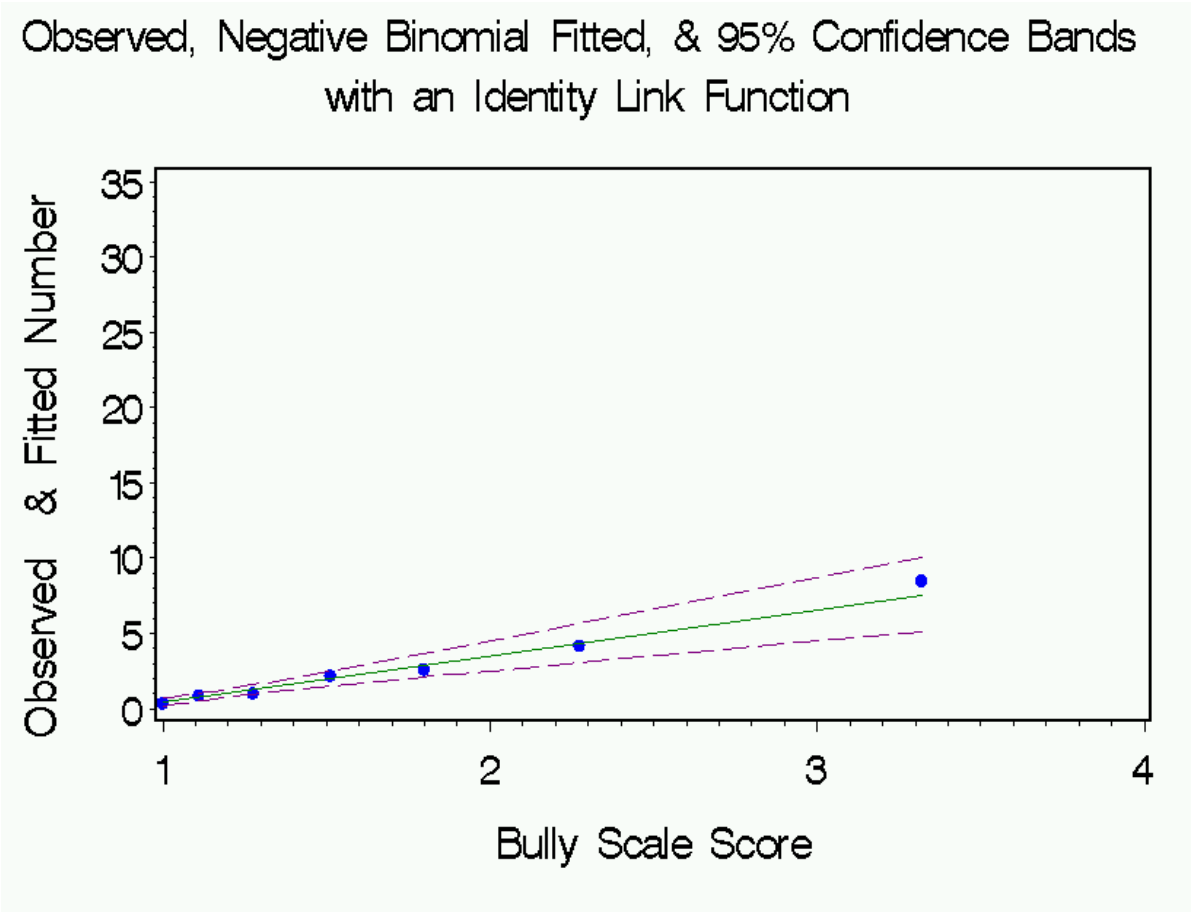
Dist	Link	G^2	X^2	X^2/df	AIC	BIC
Poisson	log	1774.60	2977.94	10.30	2160	2168
NegBin	log	244.08	351.13	1.22	998	1010
NegBin	Identity	244.33	337.09	1.17	992	1003

	Log Link				Identity Link			
Parm	est.	se	Wald	p	est.	se	Wald	p
β_0	-1.19	0.36	11.30	$< .01$	-2.64	0.65	16.44	$< .01$
β_1	1.09	0.20	30.66	$< .01$	3.07	0.57	29.50	$< .01$
$1/\phi$	3.50	0.44	—		3.35	0.43		

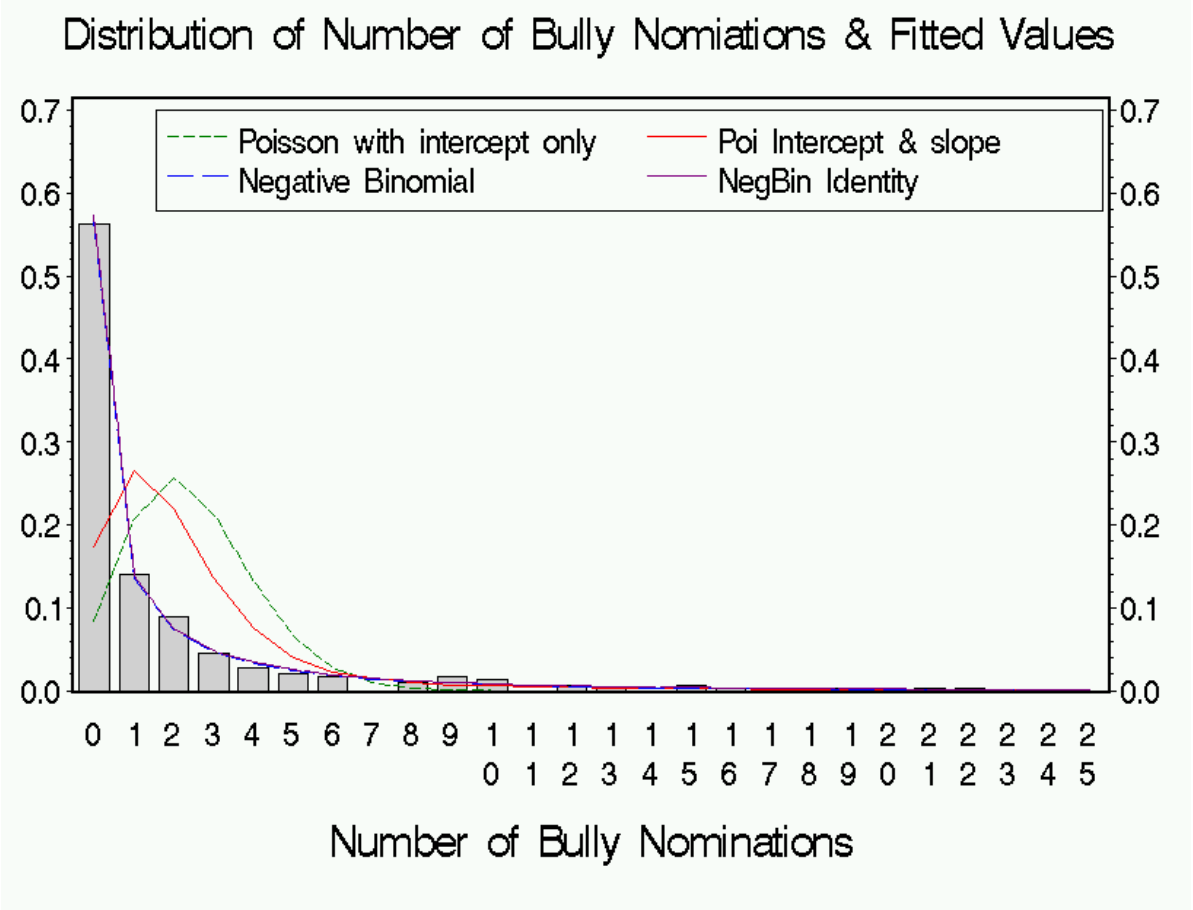
For interpretation, a one unit change in bully scale leads to

$\exp(1.09) = 2.98$ times or 3.07 more nominations

Fit of Negative Binomial Model w/ Identity



Fit of Marginal Distribution w/ Identity



Conclusion

Conclusion

The bully scale can reasonably be used lieu of the peer nominations(?)

Support from this comes from

- The similarity of the marginal distributions for the two measures (both positively skewed).
- Goodness of fit of the negative binomial regression with identity link function.

Qualifications (i.e., more to be done):

- Add in other variables known to be related to bullying (e.g., gender) to try to account for extra variability (i.e, systematic vs random).
- More modeling that takes into account peer groupings (i.e., see whether there are "errors" or systematic differences between peer groups).

Zero Inflated Models

Models for situations where there might be two underlying types or groups: one group that follows the regression model and the other that just gives 0's.

Recommended supplemental reading:

- Long, J.S. (1997). Regression Models for Categorical and Limited Dependent Variables.
- Donald Erdman, Laura Jackson, Arthur Sinko (2008). Zero-Inflated Poisson and Zero-Inflated Negative Binomial Models Using the COUNTREG Procedure (Paper 322-2008). SAS Institute Inc., Cary, NC.
  $\longrightarrow$ PROC COUNTREG is in SAS v9.2
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. NY: Chapman & Hall

Basic Zero Inflated Model (e.g., "ZIP")

- The basic model is essentially a latent class type model of the form

$$P(Y_{ij} = y|x_{ij}) = \begin{cases} \pi + (1-\pi)P(0|x_{ij}) & \text{for } y = 0 \\ (1-\pi)P(y|x_{ij}) & \text{for } y > 0 \end{cases}$$

where

  - $\pi =$ the probability of being in the "zero only" type or class.
  - $P(0|x_{ij})$ and $P(y|x_{ij})$ are based on some model, such as Poisson or Negative Binomial regression.

- ZIP model is a Zero Inflated Poisson usually with a log link:

$$P(Y_{ij} = y|x_{ij}) = \begin{cases} \pi + (1-\pi)\exp(-\mu_{ij}) & \text{for } y = 0 \\ (1-\pi)\dfrac{\exp(-\mu_{ij})\mu_{ij}^{y}}{y!} & \text{for } y > 0 \end{cases}$$

ZIP Model (continued)

$$P(Y_{ij} = y|x_{ij}) = \begin{cases} \pi + (1-\pi)\exp(-\mu_{ij}) & \text{for } y = 0 \\ (1-\pi)\frac{\exp(-\mu_{ij})\mu_{ij}^{y}}{y!} & \text{for } y > 0 \end{cases}$$

- Mean

$$E(Y_{ij}|x_{ij}) = (0 \times \pi) + \mu_{ij} \times (1-\pi) = \mu_{ij} - \mu_{ij}\pi$$

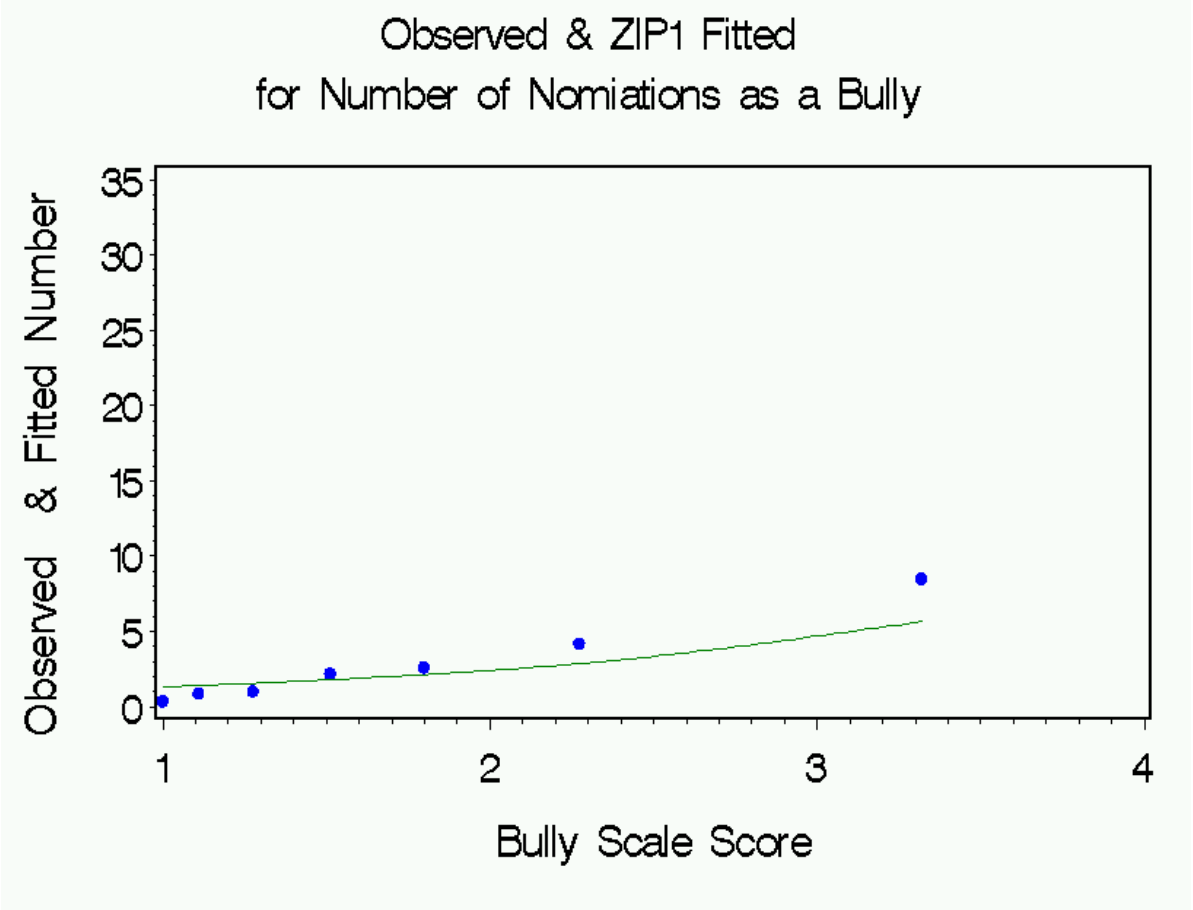
- Variance:

$$\text{var}(Y_{ij}|x_{ij}) = \mu_{ij}(1-\pi)(1+\mu_{ij}\pi)$$

- Note that if $\pi = 0$, we simply have a standard Poisson regression with log link.
- Extending the ZIP model by noting that class membership is dichotomous, so we can do a logistic regression (or other model for binary data) on the probability of class membership, e.g., a logit model,

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \gamma_o + \gamma_1 z_{1ij} + \ldots + \gamma_q z_{qij}$$

ZIP and Bully Nominations



Extending the ZIP

- Since class membership is dichotomous, we can do a logistic regression (or other model for binary data) on the probability of class membership
- For example,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \gamma_o + \gamma_1 z_{1i} + \ldots + \gamma_q z_{qi}$$

- For our Bully nominations, we could try

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \gamma_o + \gamma_1 (\text{bullyscale})_{ij}$$

Extending the ZIP

A comparison of how well various ZIP models fit the data:

Dist.	Model link	for π	df	G^2	X^2	AIC	BIC
Poi	log	none	288	1579.04	814.82	1585	1596
Poi	log	logit	287	1561.26	824.00	1569	1584
Poi	Ident	logit	287	1553.76	825.90	1562	1576

ZIP Model Parameter Estimates

and How to interpret them:

	ZIP w/o model for π				ZIP With Logit model for π			
parm	est	se	Wald	p	est	se	Wald	p
β_0	0.48	0.10	0.28	$< .01$	0.50	0.10	0.30	$< .01$
β_1	0.60	0.04	0.52	$< .01$	0.60	0.04	0.52	$< .01$
γ_0	0.21	0.12	-0.03	.09	1.50	0.35	0.82	$< .01$
γ_1					-0.77	0.20	-1.16	$< .01$

ZIP w/o model for π:

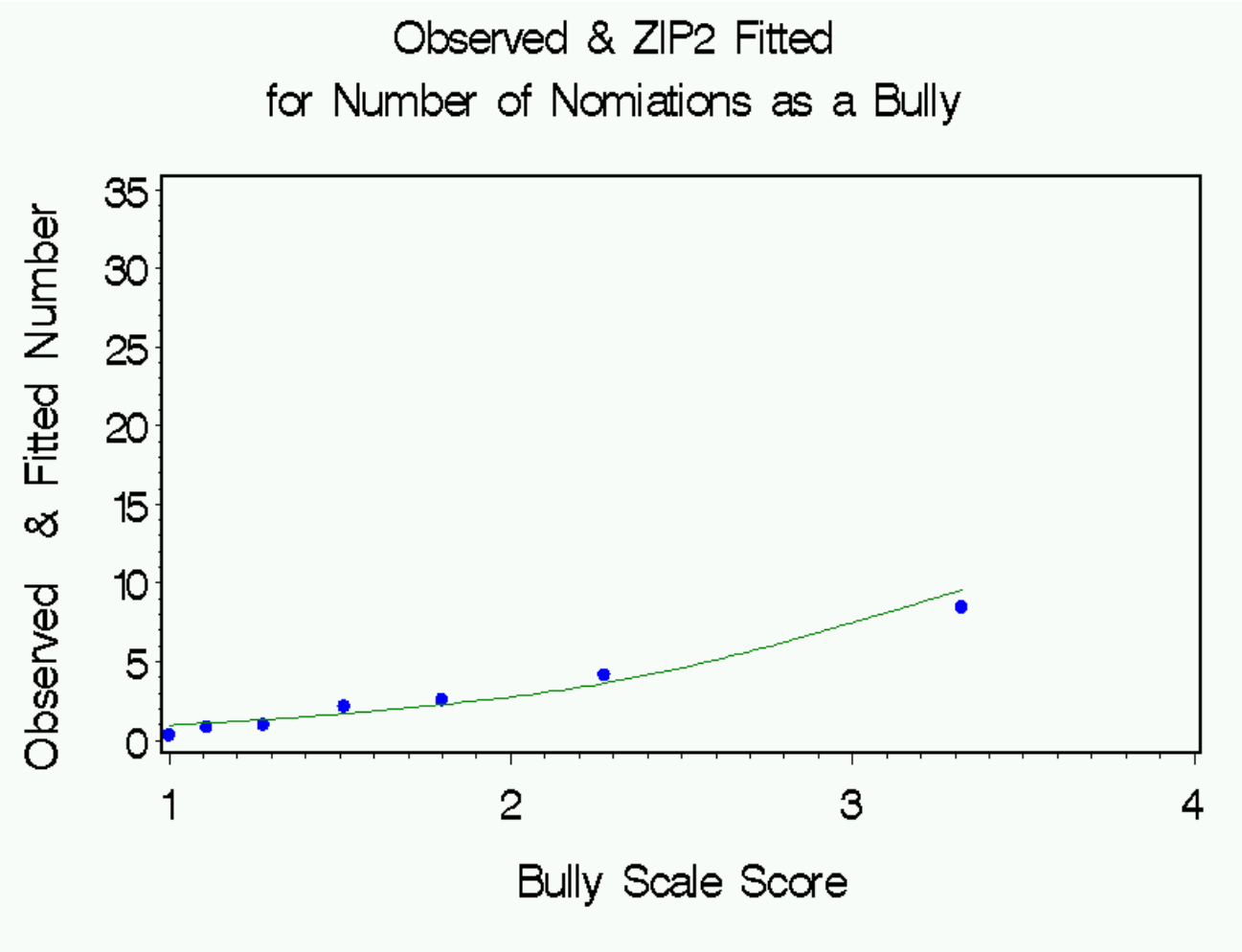
$$\exp(0.60) = 1.82 \qquad \text{and} \qquad \hat{\pi} = \frac{\exp(0.21)}{1 + \exp(0.21)} = .55$$

ZIP With Logit model for π:

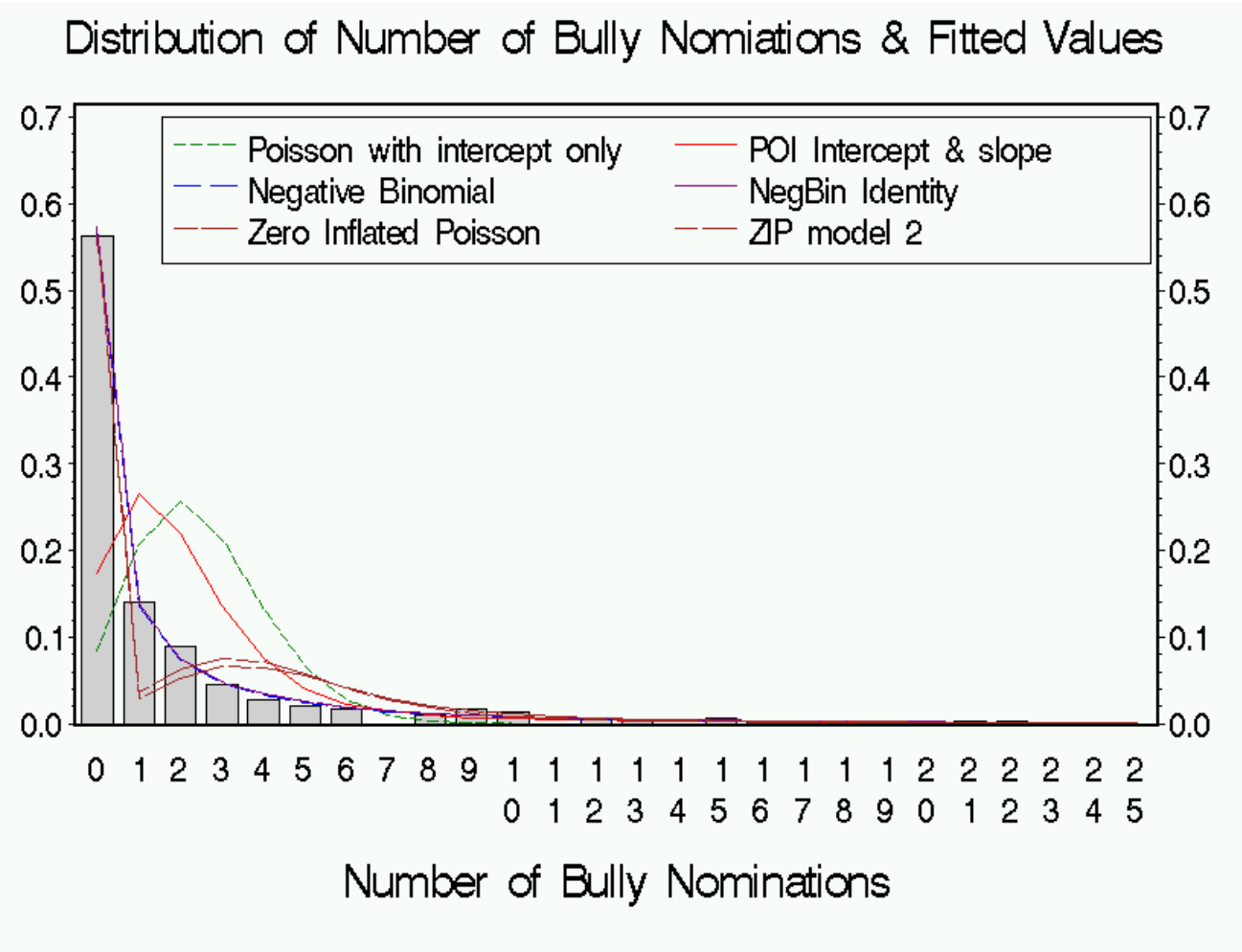
$$\exp(0.60) = 1.82 \qquad \text{and} \qquad \hat{\pi} = \frac{\exp(1.50 - 0.77(\text{bullysc})_{ij})}{1 + \exp(1.50 - 0.77(\text{bullysc})_{ij})}$$

Note that $\exp(-.77) = 0.46$.

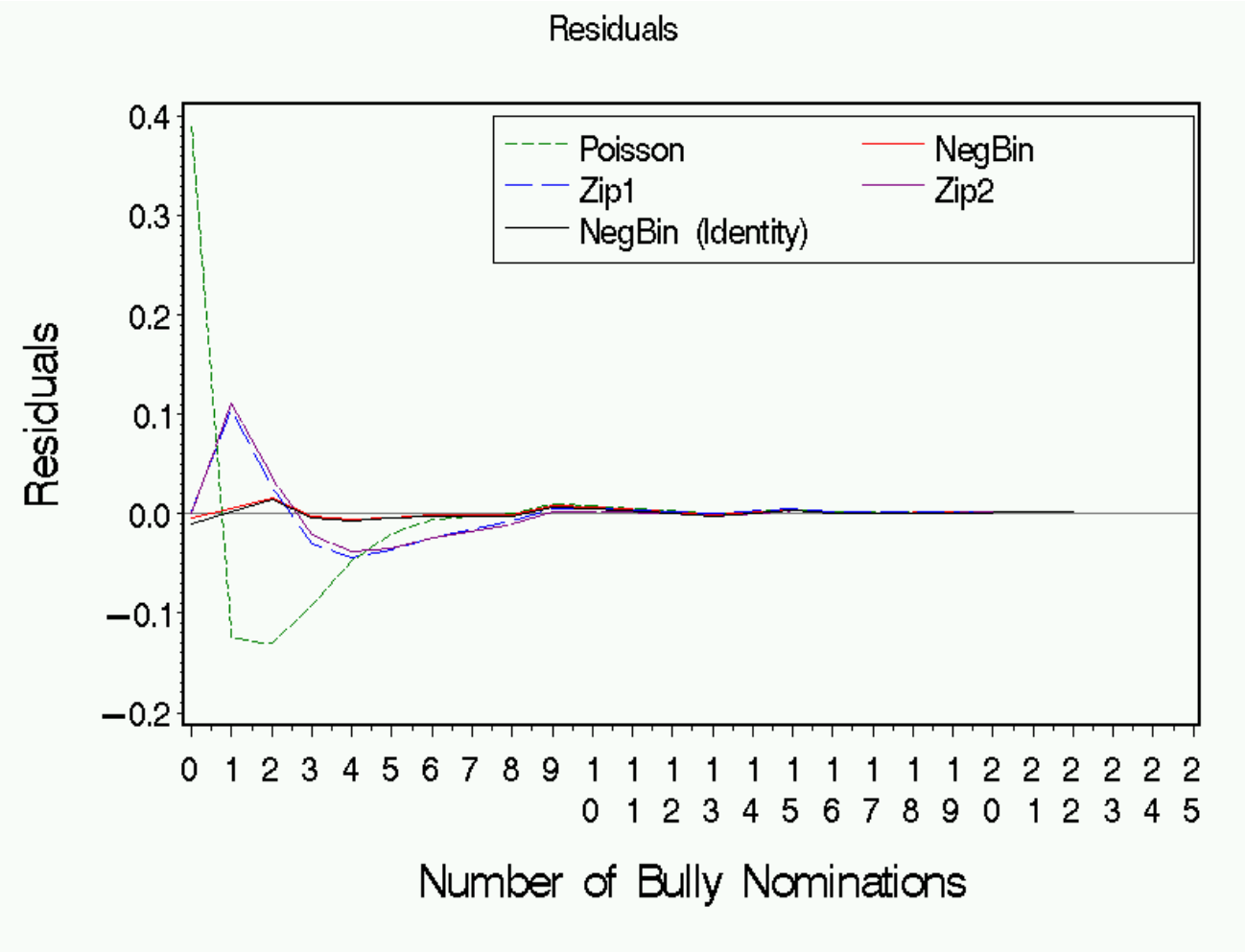
ZIP w/ logit model and Bully Nominations



Comparing all Fitted



Comparing all Fitted



Mix and Match

- You can also have a zero inflated Negative Binomial model.
- You can specify a model other than logit for the mixing probability.
- You can do all this as a multi-level (random effects) model.
- Change link functions and/or distributions

SAS: Data

```
data bullynom;
    input      BULLYSC   BULLYNM;
 datalines;
    1.56        0.00
    1.56        0.00
    1.11        0.00
    1.56        0.00
    1.22        4.00
      ⋮
    3.22        0.00
    1.22        0.00
    1.89        0.00
    run;
```

SAS for Models

```
/* Poisson Regression */
proc genmod data=bullynom;
    model bullynm = bullysc / link=log dist=poi type3;
    output out=genmodpoi pred=fitpoi upper=uppoi
        lower=lopoi stdreschi=respoi;
    title1 'Poisson Regression';
/* Negative Binomial Regression */
proc genmod data=bullynom;
    model bullynm = bullysc / link=log dist=negbin type3 ;
    output out=genmodnb pred=nbfit
            upper=nbup lower=nblo stdreschi=resnegbin;
    title1 'Negative Binomial';
```

SAS: models

Poisson Regression with identity link

```
proc genmod data=bullynom;
    model bullynm = bullysc / link=identity dist=poi type3;
run;
```

Poisson Regression with log link

```
proc genmod data=bullynom;
    model bullynm = bullysc / link=log dist=poi type3;
run;
```

Zero Inflated Poission Regression with predictor of inflation probability

```
proc genmod data=bullynom;
    model bullynm = bullysc / link=log dist=zip type3;
    zeromodel bullysc / link=logit;
run;
```

R: data

The data are in file bully_data.txt

BULLYSC	BULLYNM
1.56	0.00
1.56	0.00
1.11	0.00
1.56	0.00
1.22	4.00
⋮	
3.22	0.00
1.22	0.00
1.89	0.00

```
bully ← read.table("bully_data.txt",header=TRUE)
```

R: models

A standard poisson with log link:

```
summary(mod1 ← glm(nom ∼ bsc, data=bully, family=poisson))
```

A negative binomial with log link:

```
library(MASS)
summary(mod2 ← glm.nb(nom ∼ bsc, data=bully))
```

Note that "disperson" in R output is ϕ; whereas, SAS gives $1/phi$.

A zip model with predictors for probability

```
library(pscl)
summary(mod4 ← zeroinfl(nom ∼ bsc | bsc, data = bully))
```

SAS using NLMIXED – skip?

```
proc nlmixed data=bullynom;
/* Some starting values */
parm beta0= -2.4076 beta1= 1.0168 a0=1 a1=.01;
/* linear predictor for the inflation probability */
    linpinf = a0 + a1*bullysc;
/* infprob = inflation probability for zeros * /
/* = logistic transform of the linear predictor*/
    infprob = 1/(1+exp(-linpinf));
/* Poisson mean */
    mu = exp( beta0 + beta1*bullysc);
```

SAS v 9.1 using NLMIXED (continued)

```
/* Build the ZIP log likelihood */
if bullynm=0 then
     ll = log(infprob + (1-infprob)*exp(-mu));
else ll = log((1-infprob)) - mu + bullynm*log(mu)
     - lgamma(bullynm + 1);
model bullynm ~ general(ll);
title 'Zero Inflated Poisson regression';
```

Fitting GLMS

Important for understanding inferential procedures.

Unlike normal (ordinary) regression, there is no "closed" form equation from which we can obtain estimates of model parameters. We much use some sort of iterative algorithm.

Two commonly used algorithms are

1. Fisher scoring
2. Newton-Raphson

For Binomial logistic regression and Poisson log-linear models, Fisher scoring simplifies to Newton-Raphson.

SAS/GENMOD uses Newton-Raphson with ridge stabilization; whereas, R/glm uses Fisher. Should be same when use canonical link but . . . .

Newton-Raphson

Newton-Raphson is an iterative algorithm

- It requires initial estimates (educated guesses) for the parameter estimates.
- On each cycle the estimates are "up-dated" by approximating the log-likelihood function by a simpler polynomial function that has the shape of a concave parabola.
- The cycles are repeated until the fitted values (parameter estimates) change less than some specified criterion (a very small number).

Newton-Raphson is sometimes referred to as, "iteratively reweighted least squares".

Because it is a type of weighted least squares, the weights change from cycle to cycle and depend on variability (which is not constant with means in Binomial & Poisson distributions).

Statistic inference & the Likelihood function

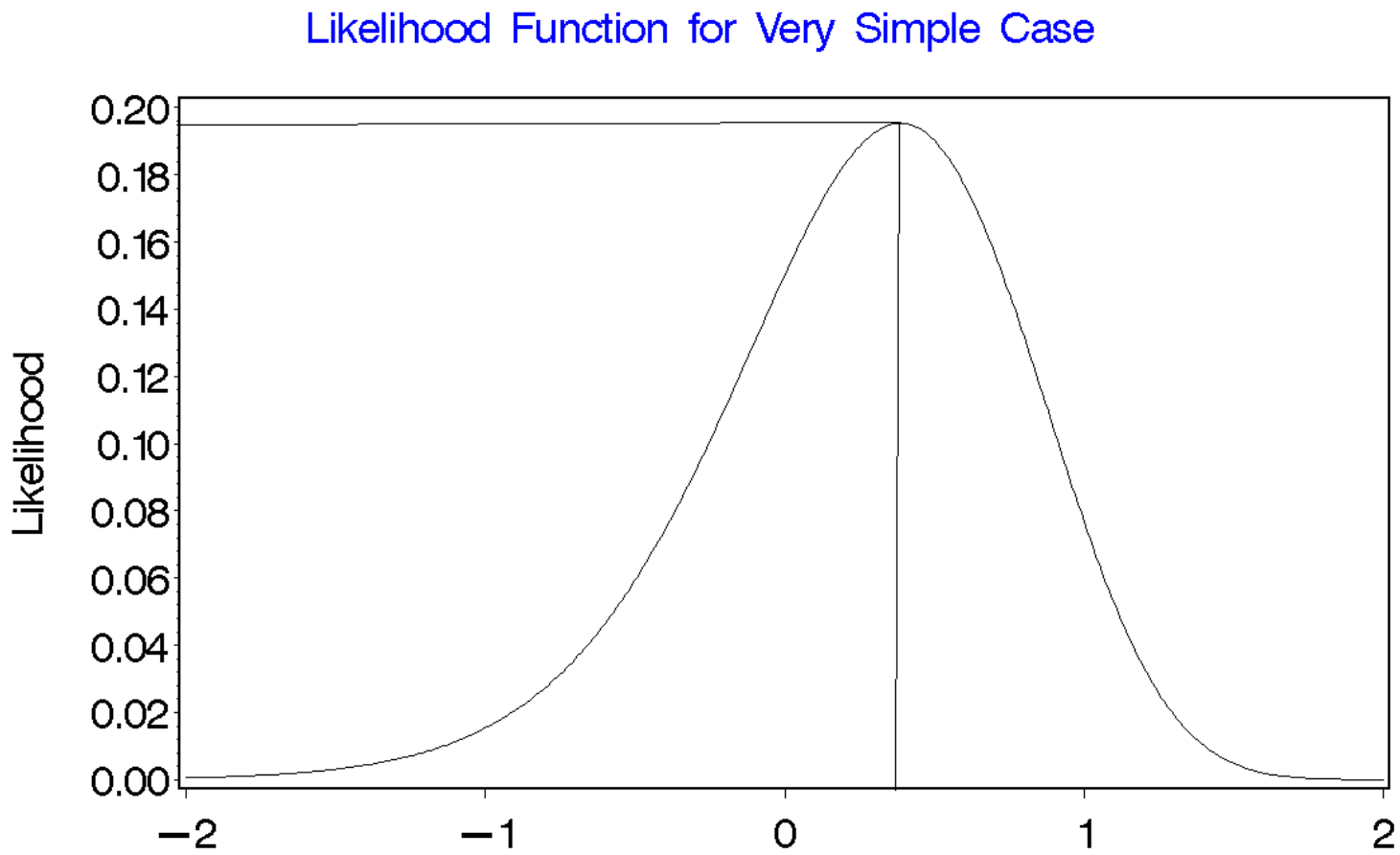
We can now consider each of the procedures we discussed for testing hypotheses about model parameters and see what information each of them uses (and therefore how they differ):

1. Wald tests.
2. Likelihood ratio tests.
3. Efficient score tests.

For purpose of illustration, consider the following log-likelihood function for the parameter β in a Poisson regression.

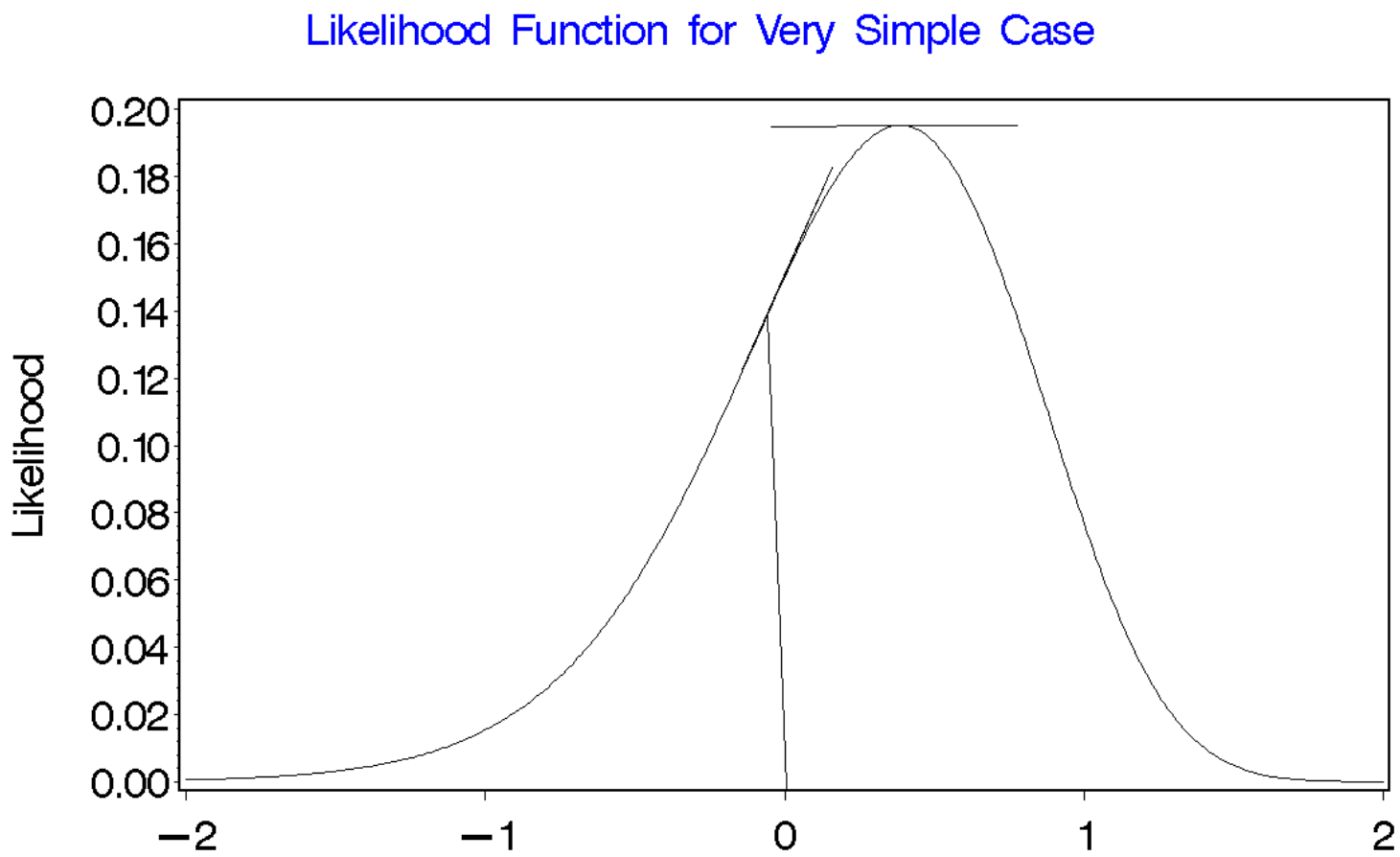
Wald Test

Only uses information about the log-likelihood function at the maximum likelihood estimated of $\hat{\beta}$.



Efficient Score Test

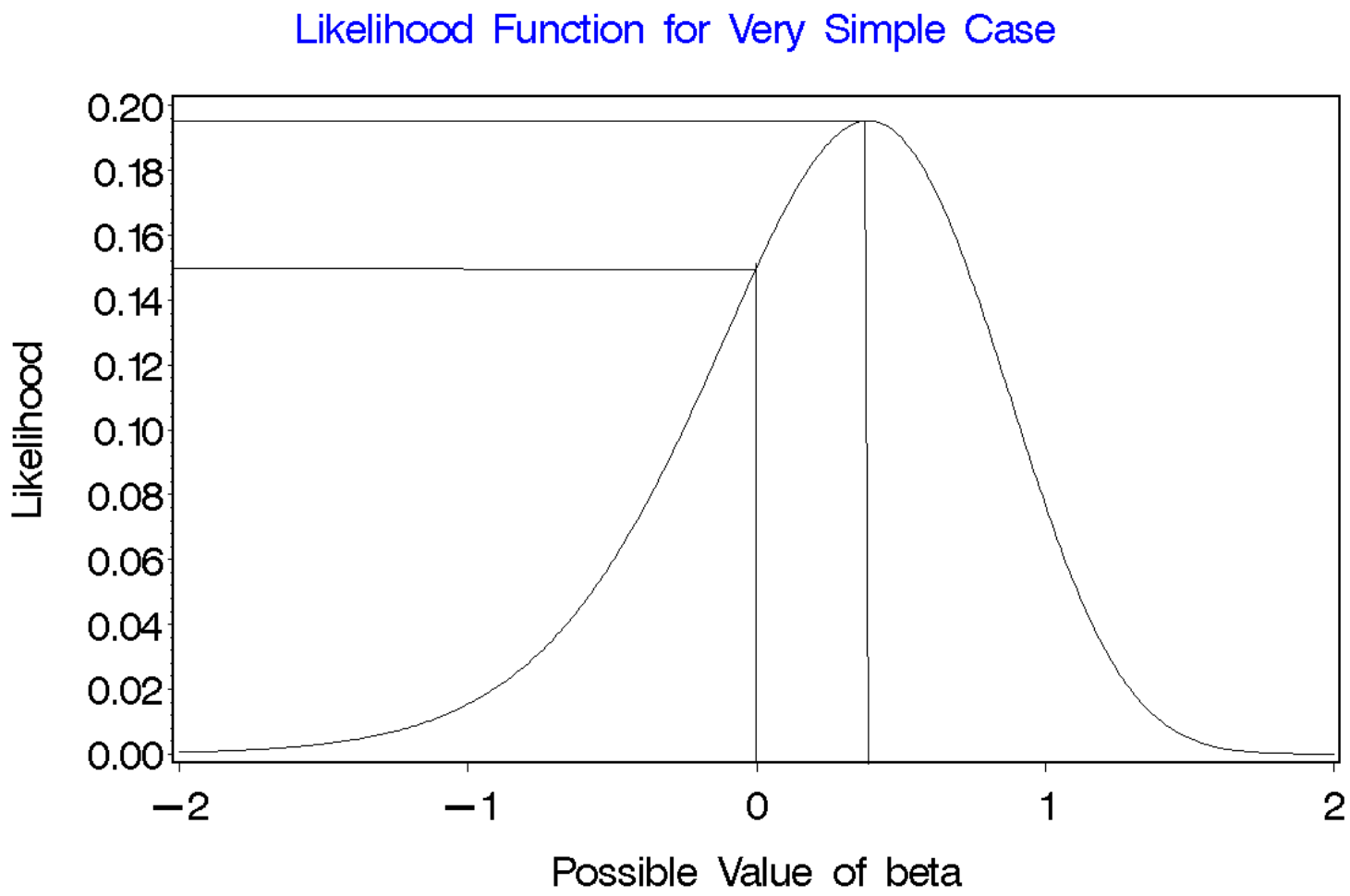
Uses information about the slope of the function at the null hypothesis value of $\beta = 0$.



Efficient Score Test (continued)

- It compares the slope at $\beta = 0$ to the slope at the maximum likelihood estimate, which is 0 (i.e., the derivative of the function equals 0 at the MLE for β).
- The further the MLE of β is from zero, the greater the slope at $\beta = 0$ tends to be.
- The score statistic equals the square of the ratio of the slope (derivative) at $\beta = 0$ to its ASE.
- The score statistic has an approximate chi-squared distribution with $df = 1$.
- You (usually) don't have to estimate the model to perform this test.
- Examples include CMH for conditional independence.
- This test can be performed even when $\hat{\beta}$ is infinite (whereas the Wald test cannot).

Likelihood Ratio



"Deviance"

This is an analog to regression models decomposition of sum of squares. Let

- L_S equal the maximum of the log-likelihood function of the most complex model; that is, the model with as many parameters as there are observations — the "**Saturated Model**".
- L_M equal the maximum of the log-likelihood function of a simpler model (of interest).

The deviance compares the log-likeihood value for the saturated and some simpler model.

$$\text{Deviance} = -2(L_M - L_S)$$

For Poisson loglinear (regression) models and Binomial logit models,

$$\text{Deviance} = G^2$$

Deviance

So, for lots of GLMs, Deviance has an approximate chi-squared distribution . . . and for some GLMs, it doesn't — which is another good reason why programs that fit GLMs don't automatically print out p–values.
When Deviance has an approximate chi-squared distribution, the "residual" degrees of freedom equal

$$df = \#\text{ responses} - \#\text{ nonredundant model parameters}$$

Deviance Residuals components of Deviance (and for binomial logit and Poisson regression G^2). They are alternatives to Pearson residuals. These also can be adjusted so that they are approximately $\mathcal{N}(0, 1)$.
Deviance & Model Comparison — a very useful property.

Deviance & Model Comparison

- Suppose that we have two models, M_0 and M_1, where M_0 is a special case of M_1 (i.e., they are "nested").
- Assuming that the more complex model M_1 fits the data, the likelihood ratio test that the simpler model fits (i.e., that you don't need the terms that are in M_1 but are not in M_0) is
  -.2in
-

$$\begin{aligned} -2(L_0 - L_1) &= -2(L_O - L_S) - \{-2(L_1 - L_S)\} \\ &= \text{Deviance}_0 - \text{Deviance}_1 \\ &= G^2_O = G^2_1 \end{aligned}$$

- with degrees of freedom

$$df = df_0 - df_1$$

where df_0 and df_1 are the residual df's for models M_0 and M_1,

Summary

GLM theory unifies important models for continuous and discrete responses.

Some common models as GLMs

Random Component	Link Function	Systematic Component	Model
Normal	Identity	Continuous	Regression
Normal	Identity	Categorical	ANOVA
Normal	Identity	Mixed	ANCOVA
Binomial	Logit	Mixed	Logistic regression
Poisson	Log	Mixed	Loglinear
Multinomial	Generalized Logit	Mixed	Multinomial response

The next section of the course will cover logistic regression, loglinear models, and multinomial response models in more detail.