An assessment of non-traditional regression models for count data

Alex Zajichek, Biostatistician
Quantitative Health Sciences
Cleveland Clinic

February 12, 2018

More methods?

- Assumptions for traditional models can be difficult to satisfy with real-life count data
- From observation, the acknowlegement of some subsequent methods for count data is underwhelming
  - Likely from the lack of application
- Utilization of non-traditional methods can give freedom to appropriately choose a model that better captures nuances in a particular dataset (more tools for the toolbox)

Typical models for types of count data

- **Linear regression:**
  - Can provide a good approximation when counts are relatively large
  - Assumes normally-distributed errors

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

- **Poisson regression:**
  - Very common approach when normal approximation does not appear to work
  - Assumes equi-dispersion

$$E[Y_i|\mathbf{X}_i] = V[Y_i|\mathbf{X}_i]$$

Typical models for types of count data

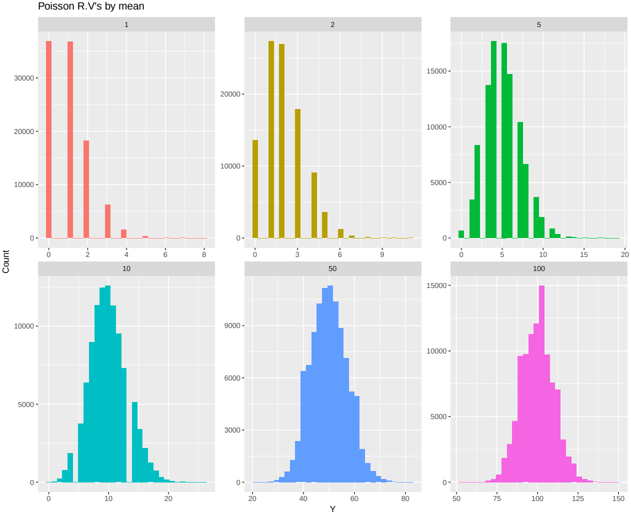

Figure 1: Distributions of randomly-generated Poisson R.V's varying by mean

Typical models for types of count data

- **Negative Binomial regression:**
  - Frequently used for over-dispersed count data

$$E[Y_i|\mathbf{X}_i] < V[Y_i|\mathbf{X}_i]$$

- What if we have...
  - Known restrictions on distribution of the outcome variable?
  - Underdispersion?

Let's first review Poisson regression!

Poisson regression

If the probability mass function (PMF) is

$$P(Y = y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$$

then

$$Y|\lambda \sim Poisson(\lambda)$$

- Defined for $Y = 0, 1, 2, ...; \lambda > 0$
- $E[Y] = V[Y] = \lambda \leftarrow$ mean equals variance (equidispersion)
- Implying $SD[Y] = \sqrt{\lambda}$

Poisson regression

Regression formulation:

- For $i = 1, 2, .., N$, assume the link

$$log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = \mathbf{X}_i \boldsymbol{\beta}$$

- Plug into PMF

$$P(Y_i = y_i | \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = \frac{e^{-e^{\mathbf{X}_i \boldsymbol{\beta}}} (e^{\mathbf{X}_i \boldsymbol{\beta}})^{y_i}}{y_i!}$$

- Find parameter estimates, $\hat{\boldsymbol{\beta}}$, via maximum likelihood estimation (MLE)
- R code:

```
model <- glm(Y~X, data, family = 'poisson')
```

- Interpretation of β_j:
  The average response multiplies by e^{β_j} for every unit increase in x_j, holding all other predictors fixed.

Zero-truncated Poisson regression

- Adjust Poisson PMF:

$$P(Y_i = y_i | Y_i > 0, \lambda_i) = \frac{P(Y_i = y_i, Y_i > 0)}{1 - P(Y_i = 0)} = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\frac{1}{1 - e^{-\lambda_i}}$$

- Simply redistributes the probability mass at 0 from the unconditional distribution
- Define the linear predictor the same as in Poisson regression

$$log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = \mathbf{X}_i\boldsymbol{\beta}$$

- Find the MLE's, $\hat{\boldsymbol{\beta}}$, using the zero-truncated PMF in the likelihood

How is inference and prediction affected when Poisson regression is used on zero-truncated data?

- **Hypothesis:**
  Coverage probabilities and predicted means are most inaccurate when the magnitude of counts are small. They will gradually improve as the data shifts away from 0.

- **Coverage probability of confidence intervals:**
  If the sampling process was repeated "many" times, it's expected that $100(1-\alpha)\%$ of confidence intervals will contain the parameter of interest. $\alpha =$ Type I error rate.

- We can get at the answer with simulation!

Simulation #1: Set-up

Define

$$log(\lambda_i) = \beta_0 + 0.01X_{i1} + 0.125X_{i2} + 0.20X_{i3}$$

where

- $X_{i1} \sim Uniform(0,1) \quad X_{i2} \sim N(0,0.5) \quad X_{i3} \sim Binomial(1,0.5)$
- $\beta_0 = \{0, .25, .5, ..., 3.75, 4\} \leftarrow$ Shifts magnitude of counts
- $N = \{10, 25, 50, 100, 500, 1000\} \leftarrow$ Sample size

Simulation #1: Set-up

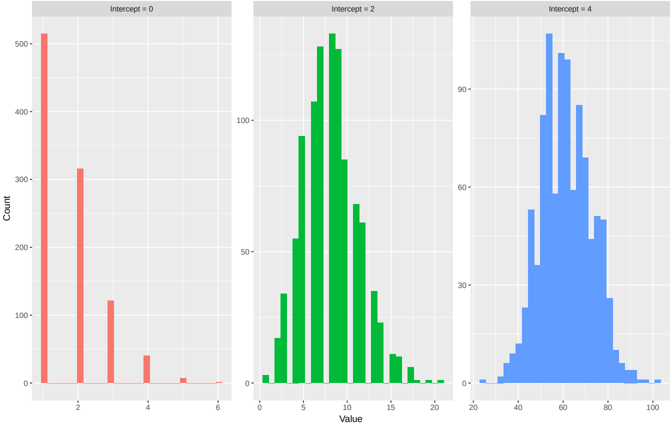

Figure 4: Example distributions of simulated data for varying intercept

Simulation #1: Process

Repeat the following for all combinations of N and β_0:

1. Randomly generate X_1, X_2, and X_3 of size N from their respective distributions
2. Calculate λ_i for all N observations using the defined linear predictor
3. Randomly generate zero-truncated Poisson realizations for each λ_i
   - This is the response variable Y
4. Fit standard Poisson regression model on Y using X_1, X_2, and X_3
5.
   - Compute a 95% confidence interval for each model parameter and indicate whether it contains the true coefficient.
   - Compute mean absolute difference (MAD) between the true and predicted λ_i's

$$MAD = \frac{\sum_{i=1}^{n} |\lambda_i - \hat{\lambda}_i|}{n}$$

6. Repeat 1-5 for $S = 1000$ samples
7. Calculate the proportion of intervals containing the true parameter

Simulation #1: Results

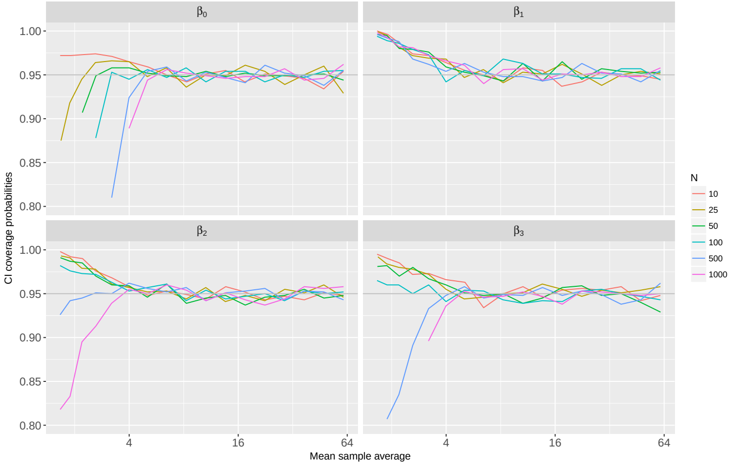

Figure 5: Sample average vs. coverage probabilities for model parameters by sample size

Simulation #1: Results

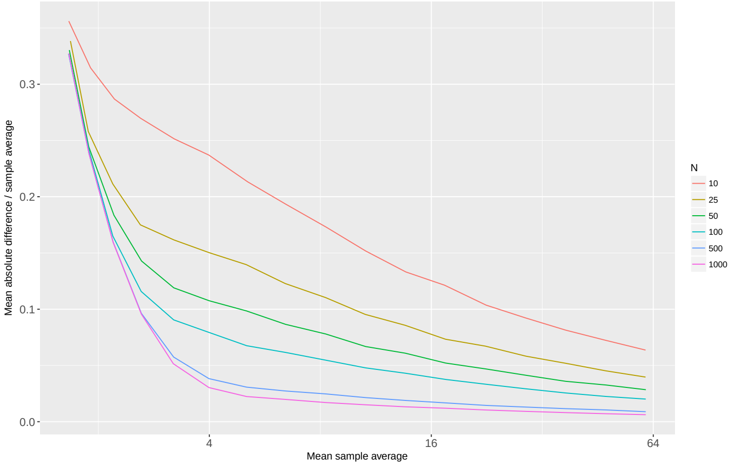

Figure 6: Sample average vs. (relative) MAD between predicted and true λ_i's by sample size

Observations from Simulation #1

- Coverage probabilities are unstable when counts are small
  - Overcoverage is not a good thing. Generally means large standard errors.
- Quickly converge to correct coverage probabilties as sample average get past 10 or so
- Sample size has largest effect on intercept term with respect to coverage
  - Focuses more on the wrong thing as sample size increase $\leftarrow$ Bias
- Larger sample size gives closer predictions to true λ_i across the board
- MAD dramatically decreases as data shifts away from zero
- MAD decreases at a faster rate as sample size increases

Remarks on zero-truncated count data

- The apparent underdispersion may be exaggerated if standard Poisson regression was used
- When counts are 'small', model misspecification (i.e. using Poisson regression) is prone to poor inference and prediction
- If there appears to be a 'cliff' at zero, stay away from Poisson regression

A model for *under*-dispersed count data

Recall:

- Poisson regression is appropriate for *equi*-dispersion

$$E[Y] = V[Y]$$

- Negative binomial regression is appropriate for *over*-dispersion

$$E[Y] < V[Y]$$

- Is there a model appropriate to handle *under*-dispersed count data?

$$E[Y] > V[Y]$$

Conway-Maxwell (COM) Poisson distribution

If

$$P(Y_i = y_i|\lambda_i, \nu) = \frac{\lambda_i^{y_i}}{(y_i!)^\nu Z(\lambda_i, \nu)}$$

for $Y_i = 0, 1, 2, \ldots$ and $\lambda_i, \nu > 0$, where

$$Z(\lambda_i, \nu) = \sum_{k=0}^{\infty} \frac{\lambda_i^k}{(k!)^\nu}$$

Then

$$Y_i|\lambda_i, \nu \sim CMP(\lambda_i, \nu)$$

is Conway-Maxwell (COM) Poisson random variable

Properties

- The dispersion parameter ν allows the traditional Poisson assumption of equi-dispersion to be relaxed
- When $\nu = 1$
  - $Z(\lambda_i, \nu) = \sum_{k=0}^{\infty} \frac{\lambda_i^k}{(k!)^\nu} = \sum_{k=0}^{\infty} \frac{\lambda_i^k}{k!} = e^{\lambda_i} \leftarrow$ Power series
  - Implies $Y_i|\lambda_i, \nu = 1 \sim Poisson(\lambda_i)$
- $E[Y] \approx \lambda^{1/\nu} - \frac{\nu-1}{2\nu}$
  - Approximation accurate if $\nu \leq 1$ (over-dispersion) or $\lambda > 10^\nu$ (large counts)

COM Poisson regression

- Again we assume the same relationship as the previous methods:

$$log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = \mathbf{X}_i \boldsymbol{\beta}$$

- Can optionally model ν in a similar way if it's suspected that different groups have different dispersion
  - Very cool feature!
- Use maximum likelihood estimation to find parameter estimates
- Likelihood-ratio (LR) test available to test for equidispersion

$$H_0 : \nu = 1 \quad H_A : \nu \neq 1$$

Simulation #2

- How well can the test for ν detect over/under dispersion?
- We can examine its *statistical power* with simulation

$$Power = P(\text{Reject } H_0 | \nu \neq 1)$$

- We'll 'reject' the null hypothesis if the p-value < 0.05

Simulation #2: Set-up

Define

$$log(\lambda_i) = \beta_0 + 0.01X_{i1} + 0.125X_{i2} + 0.20X_{i3}$$

where

- $X_{i1} \sim Uniform(0, 1) \quad X_{i2} \sim N(0, 0.5) \quad X_{i3} \sim Binomial(1, 0.5)$
- $\beta_0 = \{0, 1.33, 2.67\} \leftarrow$ Shifts magnitude of counts
- $N = \{10, 25, 50, 100, 250\} \leftarrow$ Sample size
- $\nu = \{.25, .50, ..., 1.75, 2.0\} \leftarrow$ Dispersion

Simulation #2: Process

Repeat the following for all combinations of β_0, N, and ν:

1. Randomly generate X_1, X_2, and X_3 of size N from their respective distributions
2. Calculate λ_i for all N observations using the defined linear predictor
3. Randomly generate COM Poisson realizations for each λ_i with ν
    - This is the response variable Y
4. Fit COM Poisson regression model on Y using X_1, X_2, and X_3
5. Compute p-value for equidispersion test, and indicate if < 0.05
6. Repeat 1-5 for $S = 1000$ samples
7. Calculate the proportion of tests that were rejected

Simulation #2: Results

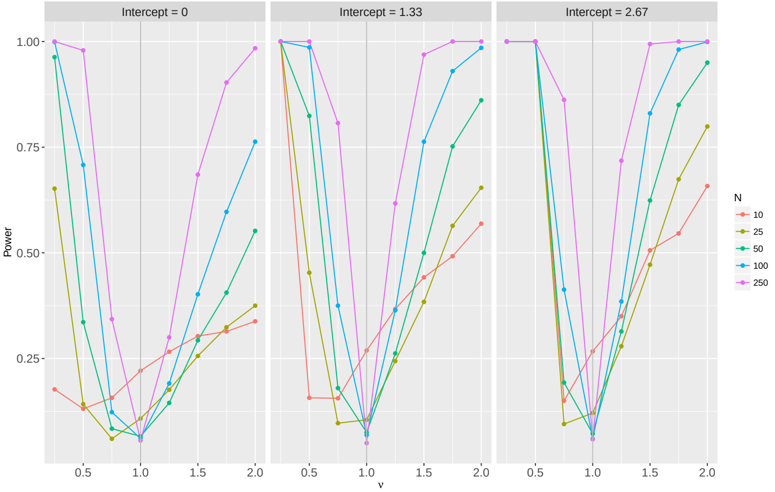

Figure 7: True dispersion, ν, vs. power of likelihood ratio test by sample size

Observations from Simulation # 2

- In general, increased sample size increases power, and accuracy of Type I error
- Power decreases as the data becomes more equidispersed
- When n is small, appears to be able to detect underdispersion ($\nu > 1$) with more power than overdispersion ($\nu < 1$) and vice-versa when n is larger
- More power as the magnitude of the counts increase

COM Poisson regression (toy) example

https://archive.ics.uci.edu/ml/datasets/Challenger+USA+Space+Shuttle+O-Ring

Note: The original data was bootstrapped for 500 samples for demonstration purposes

- Interested in modeling the number of O-rings that will experience thermal distress for a flight given the launch temperature

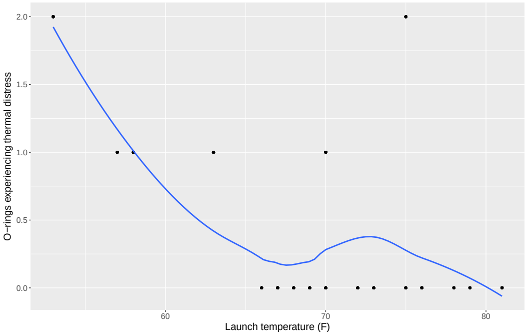

Figure 8: Launch temperature vs. number of O-rings experiencing thermal distress

COM Poisson regression: R code

```
> #Load package for COM Poisson regression
> library(COMPoissonReg)
> #Fit COM Poisson model
> mod_cmp <- glm.cmp(`Thermal distress` ~ `Launch temperature`,
      data = oring)
> summary(mod_cmp)$DF
                      Estimate     SE z.value   p.value
X:`Launch temperature`  -0.1405 0.0147 -9.5358 1.487e-21
```

- Launch temperature appears to be associated with the number of O-rings experiencing thermal distress

Testing hypothesis of equidispersion

```
> equitest(mod_cmp)$pvalue
[1] 0.0003879015 #Reject the null hypothesis
```

COM Poisson regression: R code

- Did the likelihood ratio test identify over or under dispersion?
  - 95% confidence interval for ν: (1.41, 2.67) $\leftarrow$ *under*
- Comparing fit with Poisson regression

```
> mod_p <- glm(‘Thermal distress‘ ~ ‘Launch temperature‘,
 data = oring, family = ’poisson’)
> data.frame("AIC_CMP" = AIC(mod_cmp), "AIC_P" = AIC(mod_p))
   AIC_CMP     AIC_P
  679.5021 690.0916
```

- Even with additional complexity of accounting for the dispersion, AIC shows a better fit for the CMP model

Limitations of COM Poisson regression

R package: **COMPoissonReg**

- Doesn't appear to optimized for robust performance
- Often runs into convergence issues when estimating parameters; sensitive to nuances in datasets
- Takes a long time to run as sample sizes get large

Interpretation/prediction

- Model coefficients do not have 'nice' interpretation
- Distribution average is messy. Either need to use approximation (metioned above), or use median of conditional distribution for count predictions.

Nevertheless, the methodology itself is sound!

Additional models for count data

- Zero-inflated Poisson regression
  - When a distribution has an excessive number of zeros than what would arise in a standard Poisson distribution
- Zero-inflated COM Poisson regression
  - Same as above, but also accounts for over/under dispersion simultaneously
- Quasi-Poisson models
  - Can adjust standard errors for more accurate inference when over/under dispersion is present
  - Doesn't have properties of the standard *generalized linear models* (linear, logistic, poisson, etc.) because it doesn't use the full likelihood to get estimates. This doesn't allow model comparisons with likelihood measures like AIC.

References

Kimberly Sellers ,Thomas Lotze thomas.lotze@thomaslotze.com, Andrew Raim andrew.raim@gmail.com (2017). COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression. R package version 0.4.1. https://CRAN.R-project.org/package=COMPoissonReg

Kimberly F. Sellers, Georgetown University and Galit Shumeli, University of Maryland, A Flexible Regression Model For Count Data, The Annals of Applied Statistics, 2010, Vol. 4, No. 2, 943-961, DOI: 10.1214/09-AOAS306

Thomas W. Yee (2015). Vector Generalized Linear and Additive Models: With an Implementation in R. New York, USA: Springer.

Thomas W. Yee and C. J. Wild (1996). Vector Generalized Additive Models. Journal of Royal Statistical Society, Series B, 58(3), 481-493.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Questions?