Regression Analysis of Coarsely Grouped Counts and Frequencies Using the Generalized Linear Model

Stefany Coxe[1], Leona S. Aiken[2], and Stephen G. West[2]

[1] Florida International University and [2] Arizona State University

Introduction

Categorical outcome variables are frequently used in behavioral research

Categorical variables may take the form of **grouped counts or frequencies**: underlying count or frequency that is coarsely grouped for response purposes

For example:

Alcohol consumption: Monitoring the Future (Johnston et al., 1993)

- On how many occasions have you had alcohol to drink in the past 30 days?
- **0, 1-2, 3-5, 6-9, 10-19, 20-39, 40 or more**

Psychiatric symptoms: Inventory of Complicated Grief (Prigerson, 1995)

- In the past month, have you ever felt lonely?
- **Not at all, once or twice, weekly, daily, several times a day**

Unlike common categorical outcomes(e.g., binary or count outcomes), there is **no statistical model** designed for grouped counts and frequencies

Several regression models in the **generalized linear model** (GLiM; McCullagh & Nelder, 1989) family are designed for categorical outcomes and may prove useful in analyzing **grouped count and frequency** outcomes

Method

Monte carlo simulations were used to assess the statistical performance of several GLiMs that may be used to analyze **grouped counts**

For all conditions, **one continuous variable** predicted the grouped count
The regression coefficient associated with this predictor was of primary interest

Several factors were thought **likely to affect statistical performance**

- **Mean structure** (linear or exponential)
- **Variance structure** (homoscedastic or Poisson-like increasing)
- **Effect size** (0, small, small to medium, large)
- **Sample size** (100, 250, 500, 1000)

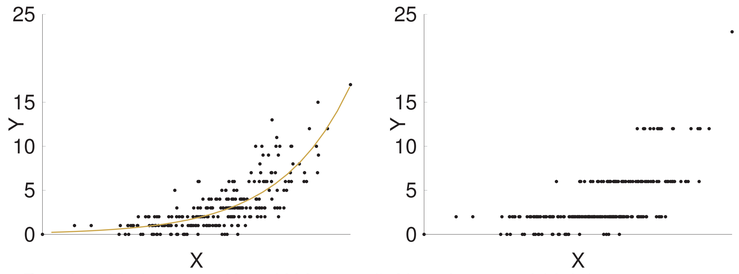
Outcomes were simulated as raw counts / frequencies, then **coarsely grouped** into ranges typical of this outcome type (Johnston et al., 1993; Prigerson, 1995)
The value of the grouped outcome is the **mean** of the range:

Raw outcome value	0	1 to 3	4 to 8	9 to 15	16 to 30
Grouped outcome value	0	2	6	12	23

SAS 9.2 was used to generate and analyze **1000 replications** per condition in a $2\times2\times4\times4$ full factorial design

Grouped count outcomes were analyzed with **three GLiM analysis models**: **linear regression**, **ordinal logistic regression**, and **Poisson regression**

Creating grouped counts



Relationship between X and Y for raw (left) and grouped (right) outcome
(n = 250, d = 0.80, exponential mean structure, increasing variance structure)

Generalized Linear Models (GLiMs)

Generalized linear models expand on the linear regression framework to better deal with categorical and non-normal outcome variables

- Familiar **linear combination** of coefficients and predictors
- Link function **transforms predicted outcome value**
- **Variance structures** beyond homoscedastic and normal
- Link function and variance structure depend on specific model

Linear regression (OLS regression)

- **Continuous, conditionally normally distributed outcome**
- Linear effect: 1 unit increase in X associated with *b* unit increase in Y
- Homoscedastic, conditionally normally distributed variance

$$\hat{Y} = b_0 + b_1X_1 + \cdots + b_pX_p$$

Ordinal logistic regression

- **Ordered categorical outcomes**
- Odds ratio: compares probability of being in a certain category (or higher) to probability of being in all lower categories
- Logistic distribution for variance structure

$$ln(\tfrac{\hat{\pi}_3}{\hat{\pi}_1+\hat{\pi}_2}) = b_{0.3} + b_1X_1 + \cdots + b_pX_p$$

$$ln(\tfrac{\hat{\pi}_2+\hat{\pi}_3}{\hat{\pi}_1}) = b_{0.23} + b_1X_1 + \cdots + b_pX_p$$

Poisson regression

- **Count outcome: integer values greater than 0, often right skewed**
- Multiplicative effect: 1 unit increase in X associated with multiplying Y by *b*
- Heteroscedastic, conditionally Poisson distributed variance

$$ln(\hat{\mu}) = b_0 + b_1X_1 + \cdots + b_pX_p$$

Results

Relative bias: (estimate - population value)/population value

- Relative bias **less than 5%** in all conditions; no differences across models

Type I error rate: probability of finding a significant effect, given that a significant effect is NOT present in the population (nominal value = .05)

- **Linear regression**: appropriate type I error in all conditions
- **Ordinal logistic regression**: appropriate type I error in all conditions
- **Poisson regression**: appropriate type I error only in conditions that parallel Poisson regression assumptions

Statistical power: probability of finding a significant effect, given that a significant effect is present in the population (acceptable value $>.80$)

- **Linear regression**: acceptable statistical power in all conditions
- **Ordinal logistic regression**: acceptable statistical power in all conditions
- **Poisson regression**: acceptable statistical power only in conditions that parallel Poisson regression assumptions

Confidence interval coverage: proportion of replications in which the population value is captured by the 95% CI (nominal value = .95)

- **Linear regression**: CI coverage good for most conditions, but low coverage for large effect size, increasing variance, especially with exponential mean structure
- **Ordinal logistic regression**: CI coverage good in all conditions
- **Poisson regression**: CI coverage good in few conditions

Conclusion

- **Ordinal logistic regression** performed well in terms of relative bias, type I error, statistical power, and confidence interval coverage, regardless of sample size, effect size, mean structure, or variance structure

- **Linear regression** performed well in many conditions, but did not provide adequate CI coverage for several conditions that are particularly relevant to the analysis of **count and frequency outcomes** (i.e., exponential mean structure and heteroscedastic variance)

- **Poisson regression** only performed well for conditions in which the ungrouped outcome followed the assumptions of Poisson regression (i.e., exponential mean structure and Poisson variance)

Coarse categorization of counts potentially loses much of the metric information contained in the data. This loss can lead to violation of assumptions associated with OLS and poisson regression. **Ordinal logistic regression**, which only considers the rank order of the categories, performed well in all conditions studied. We recommend its use for coarsely categorized count data.

Department of Psychology - Florida International University - Miami, FL
stefany.coxe@fiu.edu
http://faculty.fiu.edu/~scoxe