Modeling zero-inflated and overdispersed count data: An empirical study of school suspensions

Christopher David Desjardins

Research Problem

- The purpose of this study was to develop a model that best explains variability in number of school days suspended
  - Examined Poisson, negative binomial (NB), Poisson hurdle (PH), and negative binomial hurdle (NBH) models
- Secondarily, I examined the probability of a student being suspended at least one day using a binomial logistic regression model

School Suspensions

- School suspensions are a serious concern for school districts
- Suspensions appear to be associated with trouble with the law, school dropout, unemployment and substance abuse
- It is important to understand predictors of school suspension
- A relatively intractable statistical problem

Difficulties Modeling Suspensions

- School days suspended as a realization of the Poisson model
- This may be too simplistic
  - Zero-inflation
    - More zeros occur than would be expected if the data-generating process was a Poisson
    - *Structural* and *sampling* zeros.
    - Unreasonable fit for the zeros and possibly biased parameter estimates and standard errors
  - Overdispersion
    - Dispersion $\equiv \frac{\text{Var(Y)}}{\text{E(Y)}}$, where Y is a random variable
    - In the Poisson, $Var(Y) = E(Y)$ and overdispersion is said to occur when dispersion is > 1
    - Overly optimistic standard errors and lack-of-fit based on tests of deviance

Models for Zero-inflated & Overdispersed Count Data

- Zero-Inflation
  - Zero-Inflated Models
    - Mixture models that assume the existence of *structural* and *sampling* zeros
    - Model the probability of a *structural* zero separately from *sampling* zeros and non-zero counts
  - Hurdle Models
    - Two-part models that make no distinction between *structural* and *sampling* zeros
    - Model the probability of a zero separately from the non-zero count
- Overdispersion
  - Zero-inflated negative binomial or a NBH model

Negative Binomial Hurdle (NBH) Model

Probability Mass Function

$$\Pr(Y = y) = \begin{cases} p & I_{(y=0)} \\ \frac{1-p}{1-\left(\frac{k}{\mu+k}\right)^k} \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{k}{\mu+k}\right)^k \left(1 - \frac{k}{\mu+k}\right)^y & I_{(y>0)} \end{cases}$$

- μ is the mean of the NB model, p is the probability of a zero, and k is the dispersion parameter
- $\mathrm{E}(Y) = \frac{1-p}{1-\left(\frac{k}{\mu+k}\right)^k} \mu$
- $\mathrm{Var}(Y) = \frac{(1-p)}{1-\left(\frac{k}{\mu+k}\right)^k} \left(\mu^2 + \mu + \frac{\mu^2}{k}\right) - \left(\frac{(1-p)}{1-\left(\frac{k}{\mu+k}\right)^k} \mu\right)^2$

Regression in a NBH model

$$\log\left(\frac{p}{1-p}\right) = \mathbf{G}\boldsymbol{\gamma} \implies \begin{pmatrix} \log\left(\frac{p_1}{1-p_1}\right) \\ \log\left(\frac{p_2}{1-p_2}\right) \\ \vdots \\ \log\left(\frac{p_n}{1-p_n}\right) \end{pmatrix} = \begin{pmatrix} 1 & g_{12} & g_{13} & \cdots & g_{1p} \\ 1 & g_{22} & g_{23} & \cdots & g_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & g_{n2} & g_{n3} & \cdots & g_{np} \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{pmatrix}$$

$$\log(\boldsymbol{\mu}) = \mathbf{B}\boldsymbol{\beta} \implies \begin{pmatrix} \log \mu_1 \\ \log \mu_2 \\ \vdots \\ \log \mu_n \end{pmatrix} = \begin{pmatrix} 1 & b_{12} & b_{13} & \cdots & b_{1p} \\ 1 & b_{22} & b_{23} & \cdots & b_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & b_{n2} & b_{n3} & \cdots & b_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Where,

$$\mathbf{p} = \frac{\exp(\mathbf{G}\boldsymbol{\gamma})}{\exp(1 + \mathbf{G}\boldsymbol{\gamma})}$$

$$\exp(\mathbf{X}\boldsymbol{\beta})$$

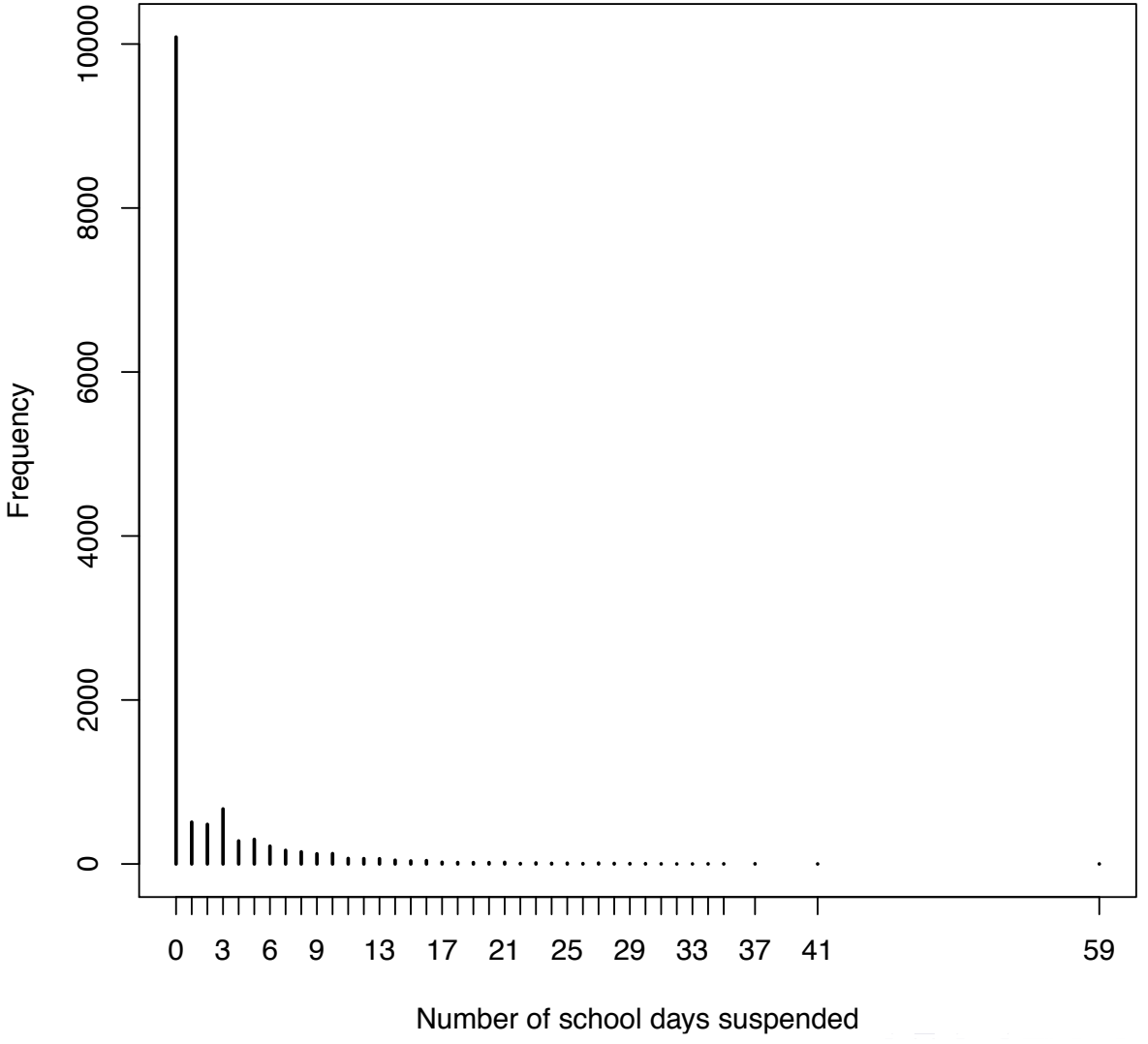
Detecting and Testing for Zero-Inflation and Overdispersion

- Marginal plot of the response
- Zero-Inflation
  - Compare Poisson to a Poisson hurdle (PH) via Vuong's test or Akaike's Information Criterion (AIC)
- Overdispersion
  - Fit Quasi-Poisson and examine estimated dispersion parameter
  - Compare Poisson to a NB via LRT, Vuong's test, AIC
- Zero-Inflation and Overdispersion
  - Compare NBH to a PH, NB, Poisson via Vuong's test, AIC

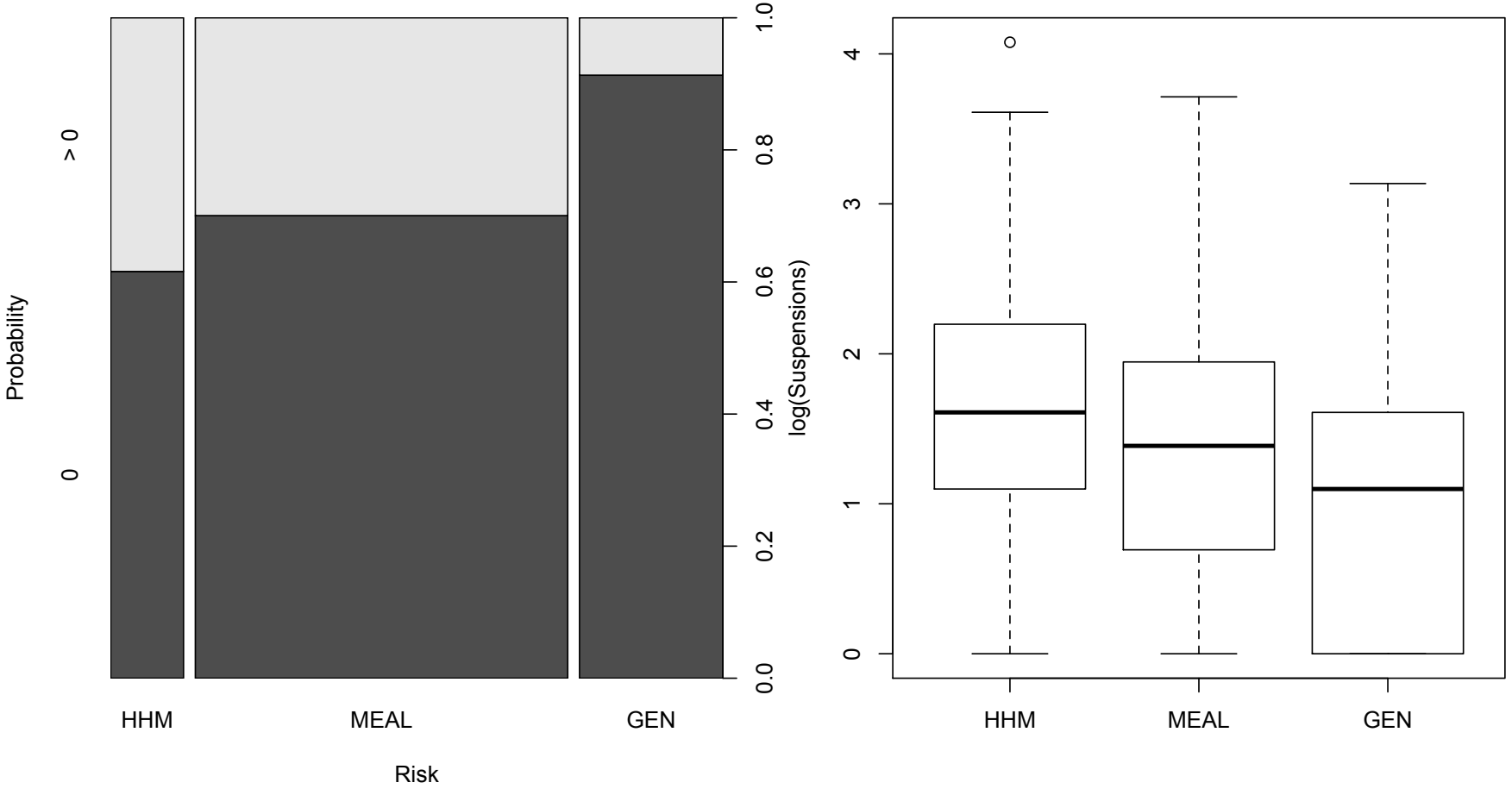
Data

- Data were collected on 13,606 students in grade 8 at Minneapolis Public Schools from 2003-2004 through 2007-2008 school years
- The response was number of school days suspended.
- The predictors:
  - Risk - homeless or highly mobile (HHM), on free or reduced price meals (MEAL), or neither HHM nor MEAL (GEN)
  - Ethnicity - Non-Hispanic White, African American, Asian American, Hispanic, American Indian
  - Gender
  - Special Education

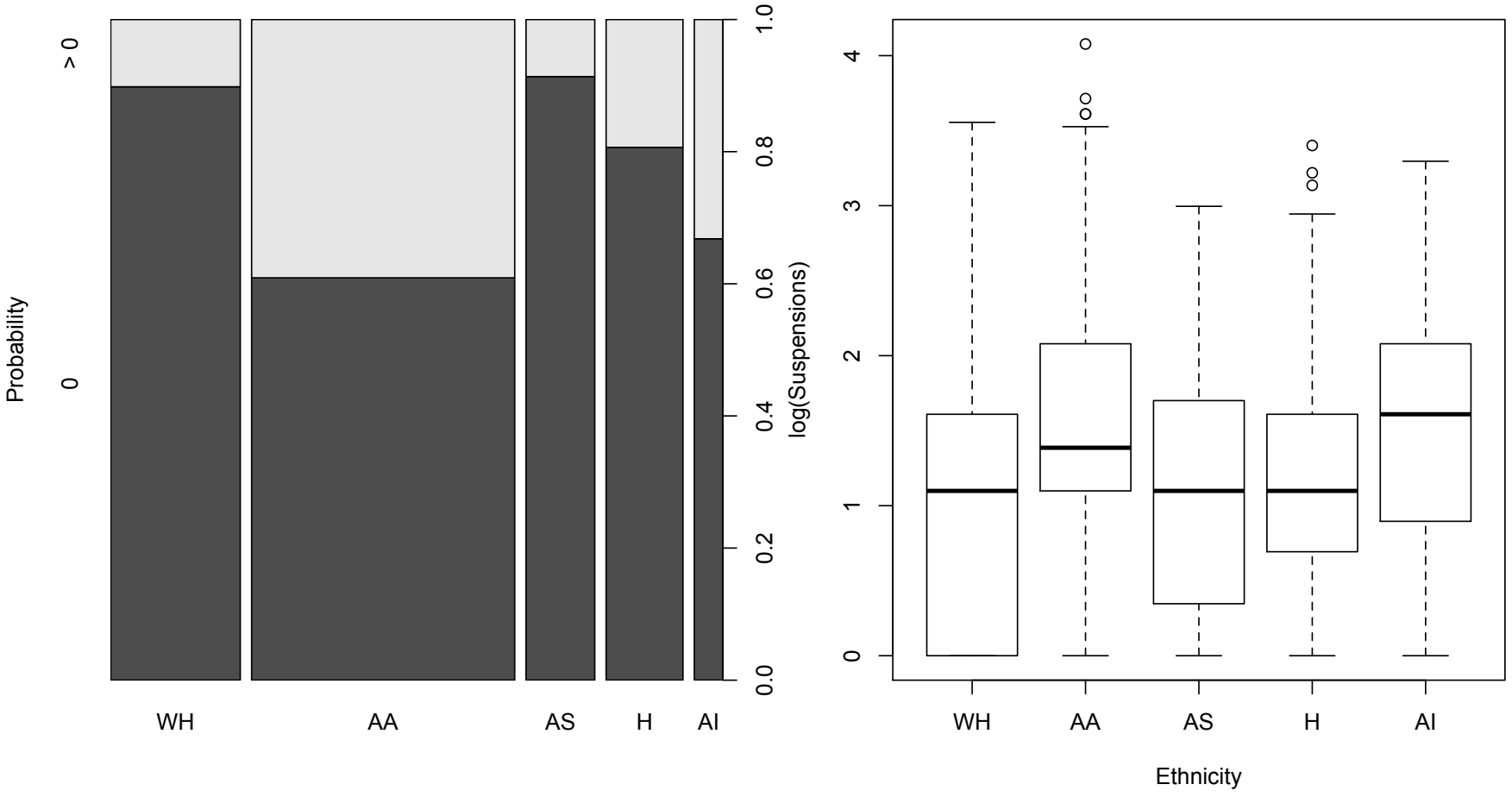
Marginal Distribution of School Suspensions



The Effect of Risk on School Days Suspended



The Effect of Ethnicity on School Days Suspended



Data Analysis

- Data split into a training (66%) and a test set (33%)
- Poisson, NB, PH, and NBH models were examined
- Binomial logistic regression model used to model the probability of being suspended at least one day
- Test for overdispersion - Quasi-Poisson and NB vs. Poisson via LRT
- Model comparison via AIC and Vuong's test
- Absolute model fit assessed by standardized Pearson residuals, residual plot, and predicted distribution of the model
- Percent relative agreement and sum of squared differences between observed and predicted response

Testing for Overdispersion and Model Comparison

- Quasi-Poisson estimated the dispersion parameter at 7.11
- NB model significant improvement in fit over Poisson ($\chi^2_1 = 21786, p < .001$)

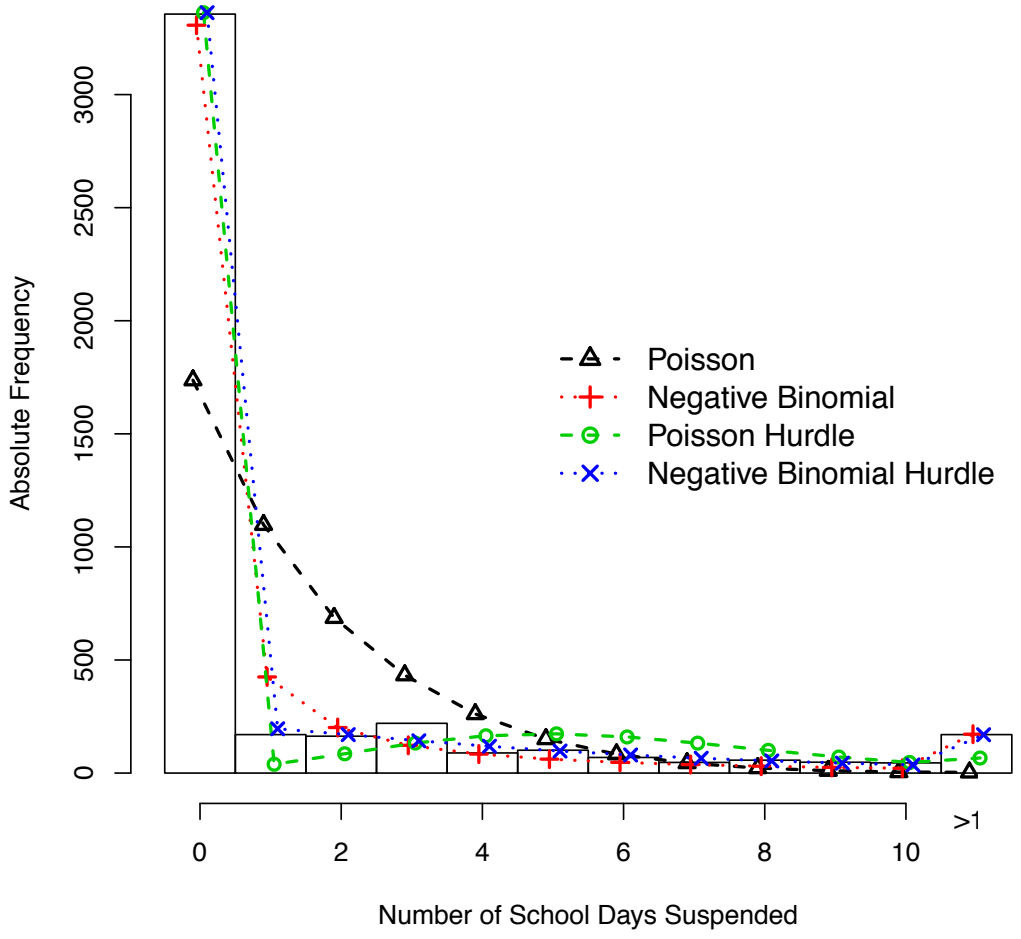
Model	AIC	Ranking based on Vuong's test
Poisson	43940	4
Negative Binomial	22152	2
Poisson Hurdle	25277	3
Negative Binomial Hurdle	21287	1

Percent Agreement and Sum of Squared Differences by Model (Test)

Model	Percent Agreement	Sum of Squared Differences
Poisson	31%	57005
NB	34%	57085
PH	33%	57024
NBH	33%	57022

- Standardized Pearson residuals greater than 3 for all models and patterns in residual plot

Observed and Predicted Distribution for School Days Suspended (Test)



Comparison of Parameter Estimates

		Poi	NB	PH	NBH
Count	Intercept	-0.44	-0.65	1.45	1.25
	MEAL	-0.17	-0.14	-0.10	-0.12
	GEN	-1.40	-1.48	-0.46	-0.56
	African American	1.10	1.22	0.31	0.37
	Asian	-0.60	-0.54	0.01	0.05
	Hispanic	0.22	0.26	0.06	0.07
	American Indian	0.95	1.14	0.32	0.38
	Special Education	0.58	0.62	0.34	0.39
	Male	0.40	0.57	0.04	0.05
		PH/NBH			
Zero	Intercept	-1.81			
	MEAL	-0.13			
	GEN	-1.31			
	African American	1.13			
	Asian	-0.74			
	Hispanic	0.20			
	American Indian	0.91			
	al Education	0.44			
	...	0.60			

Summary of the Binomial Logistic Regression Model

- Percent agreement 75% & no lack-of-fit evidence

	Estimate	SE	Wald	Pr(>|z|)
Intercept	-2.14	0.17	-12.73	0.00
MEAL	-0.12	0.09	-1.44	0.15
GEN	-1.41	0.13	-11.21	0.00
African American	1.56	0.16	9.96	0.00
Asian	-0.90	0.27	-3.36	0.00
Hispanic	0.40	0.19	2.07	0.04
American Indian	1.33	0.21	6.35	0.00
Special Education	0.40	0.14	2.95	0.00
Male	1.10	0.17	6.51	0.00
MEAL:Special Education	-0.01	0.15	-0.04	0.96
GEN:Special Education	0.46	0.23	2.03	0.04
Male:African American	-0.67	0.18	-3.70	0.00
Male:Asian	0.20	0.31	0.66	0.51
Male:Hispanic	-0.28	0.23	-1.25	0.21
Male:American Indian	-0.66	0.27	-2.47	0.01

Discussion

- NBH model fit best based on AIC and Vuong's statistic
- Large standardized Pearson residuals indicate poor fit for some of the observations
- Predicted distribution of NBH showed that model fit the data well
- Researchers using these models emphasize predicted distribution over residuals
- Binomial logistic regression with interaction good fit based on test of deviance, Pearson residuals, percent agreement, and sum of squared differences between observed and predicted responses

Discussion - Cont.

- Risk, ethnicity, special education, and gender associated with the probability a student is suspended at least one day
- Risk, ethnicity, and special education associated with the number of school days suspended given that a student was suspended at least one day
- These findings agree with extant literature

Limitations and Future Work

- 3-component Poisson, non-parametric, or cumulative logit models
- Zero-inflated models would have nearly identical fit to the hurdle models
- Simulation work
  - What conditions of overdispersion and zero-inflation these models might fail to provide adequate fit?

Conclusions

- NBH provide best fit to the count data
  - Researchers with data similar to the school suspension data and interest in knowing factors related to the probability of a non-zero count and the number of counts conditional on observing a count, should consider a hurdle model
- Binomial logistic regression model with interactions adequate
  - If interest is solely in modeling the probability of a student being suspended one day