Modeling Zero-Inflated Data

Brian Neelon

Department of Public Health Sciences,
Medical University of South Carolina

July 8, 2015

Common Count Distributions

Poisson Distribution:

$$
\begin{aligned}
\Pr(Y = y) &= \frac{\mu^y e^{-\mu}}{y!}, \quad \mu > 0;\ y = 0, 1, \ldots \\
E(Y) = V(Y) &= \mu
\end{aligned}
$$

$$\Longrightarrow \qquad \textit{equidispersion}$$

Common Count Distributions

Negative Binomial:

$$\Pr(Y = y) = \frac{\Gamma(y+r)}{\Gamma(r)y!}\left(\frac{\mu}{\mu+r}\right)^y\left(\frac{r}{\mu+r}\right)^r$$

$$r,\ \mu > 0;\ \ y = 0, 1, 2, \ldots$$

$$\begin{aligned} E(Y) &= \mu \\ V(Y) &= \mu(1+\mu/r) \\ &= \mu(1+\alpha\mu),\ \text{ where } \alpha = 1/r \end{aligned}$$

α = measure of *overdispersion*

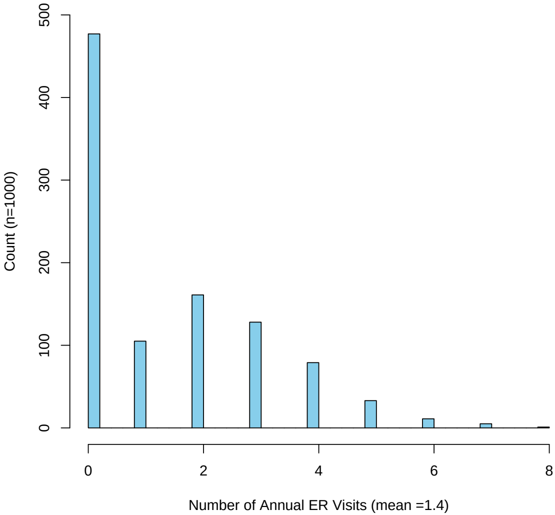
$\alpha > 0 \Rightarrow V(Y) > E(Y)$

HW: Show that as $\alpha \to 0$, NB $\overset{\text{dist}}{\Longrightarrow}$ Poisson

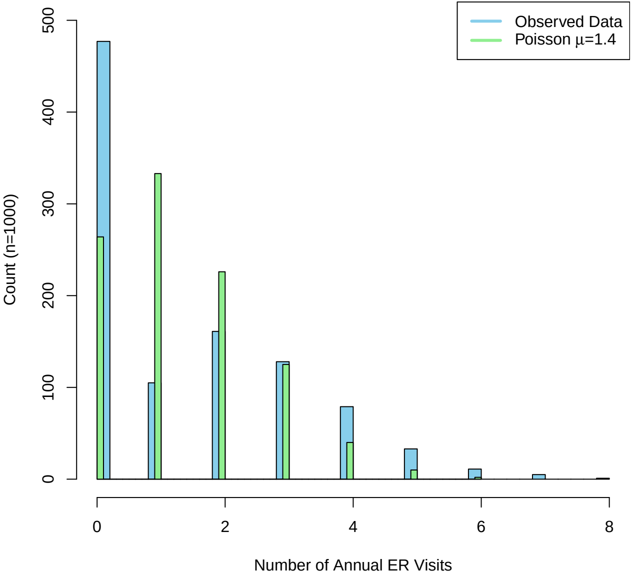
Generalized Poisson distribution[1] allows for both over- and underdispersion

[1] Consul and Jain, 1973

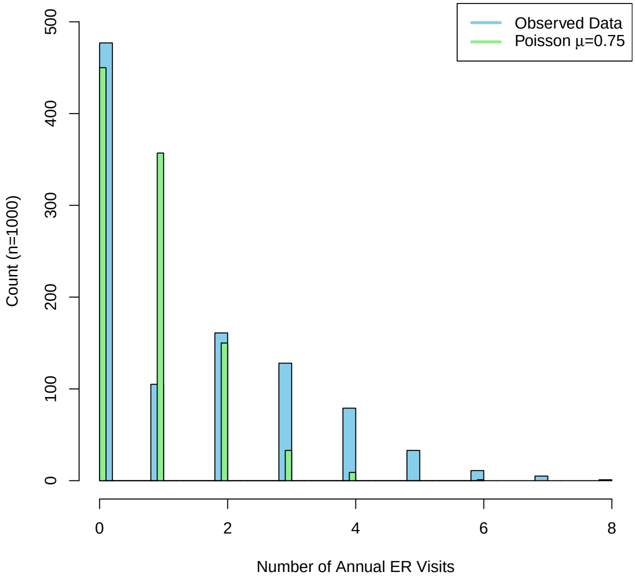
Illustrative Example: Annual ER Visits



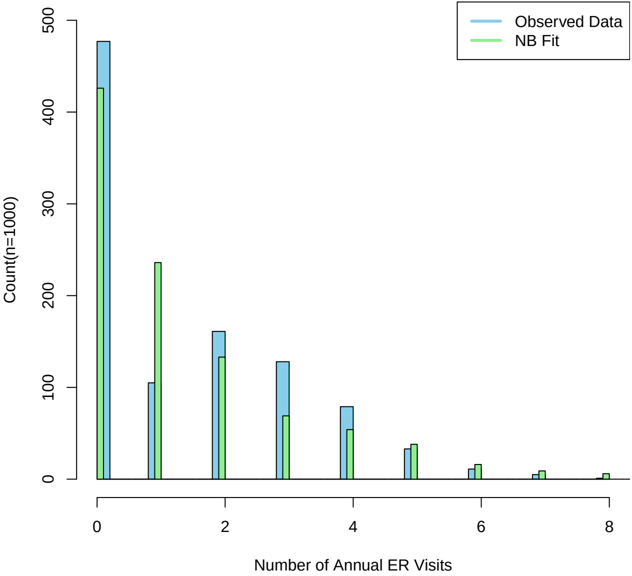
Poisson Fit


Observed Data
Poisson μ=1.4
Count (n=1000)
0
100
200
300
400
500
0
2
4
6
8
Number of Annual ER Visits

Poisson Fit with $\mu = 0.75$



Negative Binomial Fit



Zero Inflation

Zero inflation: When data contain more zeros than expected under a standard count model, the data are said to be zero inflated relative to the count distribution

Zero deflation: Fewer than expected zeros

In such cases, two-part mixtures models are often needed to assure adequate fit

These include:

1) Hurdle models: model zeros and nonzeros separately
2) Zero-inflated models: divide zeros into two types and model "extra" zeros separately

Hurdle Model

The hurdle model[2] is a two-part mixture distribution consisting of a point mass at zero followed by a zero-truncated count distribution for the positive observations:

$$\begin{aligned} \Pr(Y=0) &= 1-\pi, \quad 0 \leq \pi \leq 1 \\ \Pr(Y=y|Y>0) &= \frac{\pi p(y;\boldsymbol{\theta})}{1-p(0;\boldsymbol{\theta})}, \quad y=1,2,\ldots, \end{aligned}$$

where

$\pi = \Pr(Y>0)$ is the probability of a nonzero response

$p(y;\boldsymbol{\theta})$ is a count distribution with parameter vector $\boldsymbol{\theta}$

$p(0;\boldsymbol{\theta})$ is the count distribution evaluated at 0

[2] Cragg, 1971; Mullahy, 1986

Hurdle Model

The hurdle model can be written more compactly as

$$Y \sim (1-\pi)\mathrm{I}_{(y=0)} + \pi \frac{p(y;\boldsymbol{\theta})}{1-p(0;\boldsymbol{\theta})}\mathrm{I}_{(y>0)},$$

where $\mathrm{I}_{(\cdot)}$ is the indicator function.

Hurdle Model

The hurdle model can be written more compactly as

$$Y \sim (1-\pi)\mathrm{I}_{(y=0)} + \pi \frac{p(y;\boldsymbol{\theta})}{1-p(0;\boldsymbol{\theta})}\mathrm{I}_{(y>0)},$$

where $\mathrm{I}_{(\cdot)}$ is the indicator function.

What happens when $\pi = 0$?

Hurdle Model

The hurdle model can be written more compactly as

$$Y \sim (1-\pi)\mathrm{I}_{(y=0)} + \pi \frac{p(y;\boldsymbol{\theta})}{1-p(0;\boldsymbol{\theta})}\mathrm{I}_{(y>0)},$$

where $\mathrm{I}_{(\cdot)}$ is the indicator function.

What happens when $\pi = 0$? all zeros

Hurdle Model

The hurdle model can be written more compactly as

$$Y \sim (1-\pi)\mathrm{I}_{(y=0)} + \pi \frac{p(y;\boldsymbol{\theta})}{1-p(0;\boldsymbol{\theta})}\mathrm{I}_{(y>0)},$$

where $\mathrm{I}_{(\cdot)}$ is the indicator function.

What happens when $\pi = 0$? all zeros

When $\pi = 1$?

Hurdle Model

The hurdle model can be written more compactly as

$$Y \sim (1 - \pi)\mathrm{I}_{(y=0)} + \pi \frac{p(y; \boldsymbol{\theta})}{1 - p(0; \boldsymbol{\theta})}\mathrm{I}_{(y>0)},$$

where $\mathrm{I}_{(\cdot)}$ is the indicator function.

What happens when $\pi = 0$? all zeros

When $\pi = 1$? truncated count distribution

Hurdle Model

The hurdle model can be written more compactly as

$$Y \sim (1-\pi)\mathrm{I}_{(y=0)} + \pi \frac{p(y;\boldsymbol{\theta})}{1-p(0;\boldsymbol{\theta})}\mathrm{I}_{(y>0)},$$

where $\mathrm{I}_{(\cdot)}$ is the indicator function.

What happens when $\pi = 0$? all zeros

When $\pi = 1$? truncated count distribution

When $\pi = 1 - p(0;\boldsymbol{\theta})$?

Hurdle Model

The hurdle model can be written more compactly as

$$Y \sim (1-\pi)\mathrm{I}_{(y=0)} + \pi \frac{p(y;\boldsymbol{\theta})}{1-p(0;\boldsymbol{\theta})}\mathrm{I}_{(y>0)},$$

where $\mathrm{I}_{(\cdot)}$ is the indicator function.

What happens when $\pi = 0$? all zeros

When $\pi = 1$? truncated count distribution

When $\pi = 1 - p(0;\boldsymbol{\theta})$? ordinary count distribution

Hurdle Model

The hurdle model can be written more compactly as

$$Y \sim (1-\pi)\mathrm{I}_{(y=0)} + \pi \frac{p(y;\boldsymbol{\theta})}{1-p(0;\boldsymbol{\theta})}\mathrm{I}_{(y>0)},$$

where $\mathrm{I}_{(\cdot)}$ is the indicator function.

What happens when $\pi = 0$? all zeros

When $\pi = 1$? truncated count distribution

When $\pi = 1 - p(0;\boldsymbol{\theta})$? ordinary count distribution

$\pi > 1 - p(0,\boldsymbol{\theta}) \Rightarrow$ zero inflation

$\pi < 1 - p(0;\boldsymbol{\theta}) \Rightarrow$ zero deflation

Poisson Hurdle Model

$$\begin{aligned}
\Pr(Y=0) &= 1-\pi, \quad 0 \le \pi \le 1 \\
\Pr(Y=y|Y>0) &= \pi \frac{\mu^y \mathrm{e}^{-\mu}}{y!(1-\mathrm{e}^{-\mu})}, \quad \mu > 0;\ y = 1, 2, \ldots \\
\mathrm{E}(Y) &= \frac{\pi\mu}{1-\mathrm{e}^{-\mu}}
\end{aligned}$$

HW: Derive $\mathrm{V}(Y)$.

Interpreting μ:

- Not as straightforward as for ordinary Poisson
- For fixed π, as μ increases, $\mathrm{E}(Y)$ increases

Negative Binomial Hurdle Model

$$\begin{aligned}
\Pr(Y=0) &= 1-\pi, \quad 0 \leq \pi \leq 1 \\
\Pr(Y=y|Y>0) &= \frac{\pi}{1-\left(\frac{r}{\mu+r}\right)^r} \frac{\Gamma(y+r)}{\Gamma(r)y!} \left(\frac{\mu}{\mu+r}\right)^y \left(\frac{r}{\mu+r}\right)^r \\
\mathrm{E}(Y) &= \frac{\pi\mu}{1-\left(\frac{r}{\mu+r}\right)^r}
\end{aligned}$$

Zero-Inflated Models

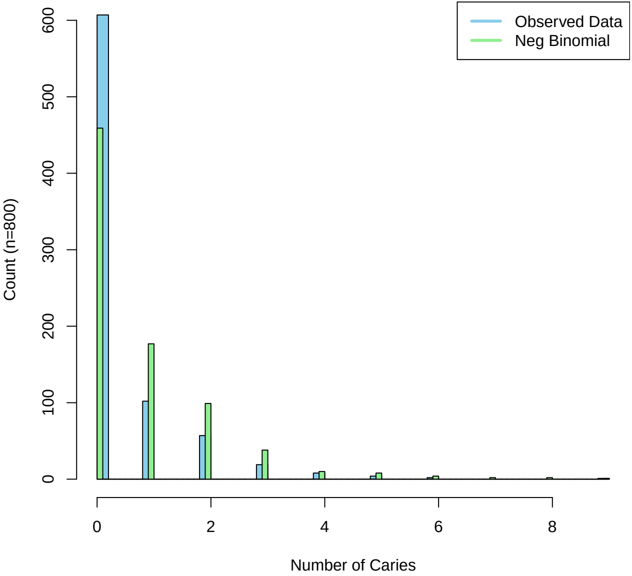
Zero-inflated models[3] partition the zeros into two types

Structural zeros: Zeros that arise due to some "structural" reason that prevents a positive count (ineligibility, not at risk, etc.)

Chance zeros: Zeros that occur "by chance" *among those at risk* – i.e., who don't have a structural zero

[3] Lambert, 1992

Example: Dental Caries



Zero-Inflated Model

The zero-inflated model is a mixture of a point mass at zero and an *untruncated* count distribution.

For example, the zero-inflated Poisson (ZIP) model is:

$$
\begin{aligned}
\Pr(Y=0) &= (1-\phi)+\phi e^{-\mu}, \quad 0 \le \phi \le 1 \\
\Pr(Y=y) &= \phi \frac{\mu^y e^{-\mu}}{y!}, \quad \mu > 0;\; y = 1,2,\ldots;\; \text{or, alternatively,} \\
Y &\sim (1-\phi) I_{(Z=0)} + \phi \mathrm{Poi}(y;\mu) I_{(Z=1)},
\end{aligned}
$$

where:

$\phi =$ "At-risk" probability (not same as π in hurdle model!)

$Z =$ Latent (unobserved) "at-risk" indicator

$\mu =$ Mean count among at-risk population

ZINB formed by choosing NB rather than Poisson

Comments

When $\phi = 1$ the model reduces to the ordinary Poisson

Otherwise, the zeros are inflated relative to the Poisson

For ZI model, $\mathsf{E}(Y) = \phi\mu$

HW1: Find $\mathsf{V}(Y)$ and show that $\mathsf{V}(Y) > \mathsf{E}(Y)$ when $\phi < 1$

- Hence zero-inflated models are overdispersed

HW2: Show that $\Pr(Y > 0) = \pi = \phi[1 - p(0; \boldsymbol{\theta})]$

- Hence ZI model can be written as a type of hurdle model in which only zero inflation and overdispersion are premitted

Testing for Zero Inflation

If no covariates, can use boundary-adjusted LR test[4] for ZI vs ordinary count distribtuion

For regression models, can use Vuong's test[5]

However, recent controversy over appropriateness of Vuong's test for comparing ZI vs ordinary count models[6]

- Ordinary model is limiting distribution – not strictly nested nor non-nested

Vuong's test okay for hurdle models vs. ordinary count, but requires two-stage approach[7]

Alternatively, use AIC as less formal comparison measure

[4] Chernoff, 1954
[5] Vuong, 1989
[6] Wilson, 2015
[7] Winkelmann, 2008

Deciding Between ZI and Hurdle Models

Suppose there's evidence of zero-inflation

How do we choose b/w hurdle and ZI models?

1) Subject matter considerations:

- ZI model: Zeros composed of two types – zeros among those not at risk, and zeros among those at risk who, by chance, have a zero count
  - Model $\Pr(Z = 1)$ and ordinary count given $Z = 1$
- Hurdle Model: only one type of zero
  - Model $\Pr(Y > 0)$ and truncated count given $Y > 0$

Deciding Between ZI and Hurdle Models (Cont'd)

2) Model fit considerations:

- Use model selection criteria or Vuong's test to choose b/w hurdle and ZI model
- Sometimes, all you care about is an appropriate model for Y that accounts for zero-inflation
- Target of inference is marginal mean $E(Y)$, not $E(Y|Y > 0)$ or $E(Y|Z > 0)$
- Can use previous formulas for $E(Y)$ to predict mean response

Regression Models for Zero-Inflated Data

Suppose we have a simple random sample (SRS) of size n from a zero-inflated population

Poisson Hurdle Regression Model:

$$\begin{aligned} Y_i &\sim (1-\pi_i)\mathrm{I}_{(y_i=0)} + \pi_i \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!(1-e^{-\mu_i})}\mathrm{I}_{(y_i>0)} \\ \operatorname{logit}(\pi_i) &= \operatorname{logit}\left[\Pr(Y_i > 0)\right] = \boldsymbol{x}_i'\boldsymbol{\beta}_1 \\ \ln(\mu_i) &= \boldsymbol{x}_i'\boldsymbol{\beta}_2, \qquad i = 1, \ldots, n, \end{aligned}$$

where $\boldsymbol{x}_i$ is a vector of covariates (can vary across components)

Regression Models for Zero-Inflated Data

Suppose we have a simple random sample (SRS) of size n from a zero-inflated population

Poisson Hurdle Regression Model:

$$
\begin{aligned}
Y_i &\sim (1-\pi_i)\mathrm{I}_{(y_i=0)} + \pi_i \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!(1-e^{-\mu_i})}\mathrm{I}_{(y_i>0)} \\
\mathrm{logit}(\pi_i) &= \mathrm{logit}\left[\Pr(Y_i > 0)\right] = \boldsymbol{x}_i'\boldsymbol{\beta}_1 \\
\ln(\mu_i) &= \boldsymbol{x}_i'\boldsymbol{\beta}_2, \qquad i = 1, \ldots, n,
\end{aligned}
$$

where $\boldsymbol{x}_i$ is a vector of covariates (can vary across components)

$\boldsymbol{\beta}_1$ = Log-odds of observing a positive response

$\boldsymbol{\beta}_2$ = Harder to interpret directly

$\boldsymbol{\beta}_2 > 0 \Rightarrow$ positive association b/w x and counts *among those with positive response*

Similar set-up for NB hurdle model

Example: ER Visits

Recall, we had 1000 patients and we wish to model the number of ER visits

Suppose we want to model Y as a function of insurance status (non-private vs. private)

Propose a Poisson hurdle model:

$$
\begin{aligned}
Y_i &\sim (1-\pi_i)\mathrm{I}_{(y_i=0)} + \pi_i \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!(1-e^{-\mu_i})} \mathrm{I}_{(y_i>0)} \\
\text{logit}(\pi_i) &= \beta_{10} + \beta_{11} x_i \\
\ln(\mu_i) &= \beta_{20} + \beta_{21} x_i, \qquad i = 1, \ldots, 1000,
\end{aligned}
$$

where $x_i = 1$ if non-private insurance and 0 if private

R Code for Poisson Hurdle Model

Can fit in R using pscl package:

library(pscl) # To fit hurdle and ZI regression

poisfit<−glm($y \sim x$, family=poisson(link="log"))

hurdfit<−hurdle($y \sim x$, dist = "poisson", link="logit")

Maximum likelihood estimates obtained via Fisher scoring

Model Comparison

Vuong's test:

vuong(poisfit,hurdfit) # Vuong's test from pscl

p-value< 0.0001 in favor of hurdle model

[8] Burnham and Anderson, 2004

Model Comparison

Vuong's test:

vuong(poisfit,hurdfit) # Vuong's test from pscl

p-value< 0.0001 in favor of hurdle model

AIC:

library(bbmle) # For AIC table

AICtab(poisfit,hurdfit)

AIC Difference: 541 in favor of hurdle

Rule of thumb: AIC difference of 10 or more strongly favors model with lower AIC[8]

[8] Burnham and Anderson, 2004

Parameter Estimates

Table 1: Poisson hurdle parameter estimates and SEs

Component	Parameter	Estimate	SE	p-val
Binary	β_{10}	-1.11	0.12	< 0.0001
	β_{11}	1.41	0.14	< 0.0001
Count	β_{20}	0.73	0.08	< 0.0001
	β_{21}	0.24	0.09	0.007

Interpretations of β_{11} and β_{21}?

Parameter Estimates

Table 1: Poisson hurdle parameter estimates and SEs

Component	Parameter	Estimate	SE	p-val
Binary	β_{10}	-1.11	0.12	< 0.0001
	β_{11}	1.41	0.14	< 0.0001
Count	β_{20}	0.73	0.08	< 0.0001
	β_{21}	0.24	0.09	0.007

Interpretations of β_{11} and β_{21}?

β_{11}: log-odds of a positive count (some ER use) for non-private vs. private

Parameter Estimates

Table 1: Poisson hurdle parameter estimates and SEs

Component	Parameter	Estimate	SE	p-val
Binary	β_{10}	-1.11	0.12	< 0.0001
	β_{11}	1.41	0.14	< 0.0001
Count	β_{20}	0.73	0.08	< 0.0001
	β_{21}	0.24	0.09	0.007

Interpretations of β_{11} and β_{21}?

β_{11}: log-odds of a positive count (some ER use) for non-private vs. private

β_{21}: Not so easy. Given at least one visit, non-private patients tend to have more visits

Predictions

Suppose we want to predict the mean number of visits for subjects with and without private insurance:

$$
\begin{aligned}
E(Y_i) &= \pi_i \frac{\mu_i}{1 - \exp(-\mu_i)} \\
&= \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)} \times \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2)}{1 - \exp[-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2)]}
\end{aligned}
$$

Predictions

Suppose we want to predict the mean number of visits for subjects with and without private insurance:

$$
\begin{aligned}
E(Y_i) &= \pi_i \frac{\mu_i}{1-\exp(-\mu_i)} \\
&= \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)} \times \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2)}{1-\exp[-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2)]}
\end{aligned}
$$

Use `predict statement` in R:

```
yhat<-predict(hurdfit,type="response")
```

Predictions

Suppose we want to predict the mean number of visits for subjects with and without private insurance:

$$
\begin{aligned}
E(Y_i) &= \pi_i \frac{\mu_i}{1-\exp(-\mu_i)} \\
&= \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)} \times \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2)}{1-\exp[-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2)]}
\end{aligned}
$$

Use predict statement in R:

yhat<−predict(hurdfit,type="response")

For patient with private insurance, $\hat{E}(Y_i) = 0.59$

For patient with non-private insurance, $\hat{E}(Y_i) = 1.64$

Predictions

Suppose we want to predict the mean number of visits for subjects with and without private insurance:

$$\begin{aligned} E(Y_i) &= \pi_i \frac{\mu_i}{1 - \exp(-\mu_i)} \\ &= \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta}_1)}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_1)} \times \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta}_2)}{1 - \exp[-\exp(\boldsymbol{x}_i' \boldsymbol{\beta}_2)]} \end{aligned}$$

Use predict statement in R:

yhat<−predict(hurdfit,type="response")

For patient with private insurance, $\hat{E}(Y_i) = 0.59$

For patient with non-private insurance, $\hat{E}(Y_i) = 1.64$

Risk ratio: $1.64/0.59 = 2.78$ times more visits on average

SEs and 95% CIs obtained via delta method or bootstrap

Example 2: ZINB Model for Dental Caries

In example 2, we had 800 dental caries patients. Suppose we want to assess efficacy of new flouride treatment (x)

ZINB Model:

$$
\begin{aligned}
Y_i &\sim (1-\phi_i)\mathrm{I}_{(Z_i=0)} + \phi_i \mathrm{NB}(y_i;\mu_i,\alpha)\mathrm{I}_{(Z_i=1)} \\
\mathrm{logit}(\phi_i) &= \mathrm{logit}\left[\Pr(Z_i=1)\right] = \beta_{10} + \beta_{11}x_i \\
\ln(\mu_i) &= \beta_{20} + \beta_{21}x_i, \qquad i = 1,\ldots,800.
\end{aligned}
$$

Interpretations of β_{11} and β_{21}?

Example 2: ZINB Model for Dental Caries

In example 2, we had 800 dental caries patients. Suppose we want to assess efficacy of new flouride treatment (x)

ZINB Model:

$$
\begin{aligned}
Y_i &\sim (1-\phi_i)\mathrm{I}_{(Z_i=0)} + \phi_i \mathrm{NB}(y_i; \mu_i, \alpha)\mathrm{I}_{(Z_i=1)} \\
\mathrm{logit}(\phi_i) &= \mathrm{logit}\left[\Pr(Z_i = 1)\right] = \beta_{10} + \beta_{11} x_i \\
\ln(\mu_i) &= \beta_{20} + \beta_{21} x_i, \qquad i = 1, \ldots, 800.
\end{aligned}
$$

Interpretations of β_{11} and β_{21}?

$\boldsymbol{\beta}_1$ = log-odds of being "at risk" for caries

Example 2: ZINB Model for Dental Caries

In example 2, we had 800 dental caries patients. Suppose we want to assess efficacy of new flouride treatment (x)

ZINB Model:

$$
\begin{aligned}
Y_i &\sim (1-\phi_i)\mathrm{I}_{(Z_i=0)} + \phi_i \mathrm{NB}(y_i; \mu_i, \alpha)\mathrm{I}_{(Z_i=1)} \\
\mathrm{logit}(\phi_i) &= \mathrm{logit}\left[\Pr(Z_i = 1)\right] = \beta_{10} + \beta_{11} x_i \\
\ln(\mu_i) &= \beta_{20} + \beta_{21} x_i, \qquad i = 1, \ldots, 800.
\end{aligned}
$$

Interpretations of β_{11} and β_{21}?

$\boldsymbol{\beta}_1$ = log-odds of being "at risk" for caries

$\boldsymbol{\beta}_2$ = log incidence ratio (IDR) for at-risk group

Specifically, $\exp(\beta_{21})$ = multiplicative increase in $E(Y)$ for *at-risk* patients given treatment vs those without treatment

R Code for ZINB Model

ML estimation proceeds via Newton Raphson or EM algorithm by treating latent at-risk indicator, Z, as a type of missing data

library(pscl)

NBfit<−glm.nb($y \sim x$) # not part of pscl

ZINBfit<−zeroinfl($y \sim x$, dist = "negbin", EM = TRUE)

Note: `pscl` parameterizes in terms of $1 - \phi = \Pr(Z_i = 0)$

Model Comparison

Vuong's test widely used but some recent controversy[9]

Let's use AIC instead:

AICtab(nbfit,ZINBfit)

AIC Difference: 10.5 in favor of ZINB

ZI models often require large sample sizes to distinguish models

[9] Wilson, 2015

Model Estimates

Table 2: ZINB parameter estimates and SEs.

Component	Parameter	Estimate	SE	p-value
Binary	β_{10}	-0.66	0.30	0.03
	β_{11}	-1.39	0.33	< 0.0001
Count	β_{20}	0.35	0.21	0.10
	β_{21}	-0.46	0.31	0.14
	$\log(\theta) = \log(1/\alpha)$	0.67	0.74	0.37

Interpretations of β_{11} and β_{21}?

Model Estimates

Table 2: ZINB parameter estimates and SEs.

Component	Parameter	Estimate	SE	p-value
Binary	β_{10}	-0.66	0.30	0.03
	β_{11}	-1.39	0.33	< 0.0001
Count	β_{20}	0.35	0.21	0.10
	β_{21}	-0.46	0.31	0.14
	$\log(\theta) = \log(1/\alpha)$	0.67	0.74	0.37

Interpretations of β_{11} and β_{21}?

β_{11}: log-odds of being "at risk" for trt group

$\exp(-1.39) = 0.25$ times lower odds of being at risk

Model Estimates

Table 2: ZINB parameter estimates and SEs.

Component	Parameter	Estimate	SE	p-value
Binary	β_{10}	-0.66	0.30	0.03
	β_{11}	-1.39	0.33	< 0.0001
Count	β_{20}	0.35	0.21	0.10
	β_{21}	-0.46	0.31	0.14
	$\log(\theta) = \log(1/\alpha)$	0.67	0.74	0.37

Interpretations of β_{11} and β_{21}?

β_{11}: log-odds of being "at risk" for trt group

$\exp(-1.39) = 0.25$ times lower odds of being at risk

β_{21}: log IDR for at-risk group

At-risk patients with treatment have $\exp(-.46) = 0.63$ time fewer caries on average ($p = 0.14$)

p-value for $\log(\theta)$ doesn't seem to have much use

Predictions

Marginal predictions often more meaningful:

$$
\begin{aligned}
\mathrm{E}(Y_i) &= \phi_i \mu_i \\
&= \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta}_1)}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_1)} \times \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_2)
\end{aligned}
$$

Predictions

Marginal predictions often more meaningful:

$$
\begin{aligned}
E(Y_i) &= \phi_i \mu_i \\
&= \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta}_1)}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_1)} \times \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_2)
\end{aligned}
$$

```
yhat<-predict(ZINBfit,type="response")
```

Predictions

Marginal predictions often more meaningful:

$$\begin{aligned} \mathrm{E}(Y_i) &= \phi_i \mu_i \\ &= \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta}_1)}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_1)} \times \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_2) \end{aligned}$$

```
yhat<-predict(ZINBfit,type="response")
```

For treatment patient, $\hat{\mathrm{E}}(Y_i) = 0.10$

For control patient, $\hat{\mathrm{E}}(Y_i) = 0.48$

Predictions

Marginal predictions often more meaningful:

$$\begin{aligned} \mathrm{E}(Y_i) &= \phi_i \mu_i \\ &= \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta}_1)}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_1)} \times \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_2) \end{aligned}$$

```
yhat<-predict(ZINBfit,type="response")
```

For treatment patient, $\hat{\mathrm{E}}(Y_i) = 0.10$

For control patient, $\hat{\mathrm{E}}(Y_i) = 0.48$

A nearly five-fold reduction in incidence of caries.

Current Work

- Marginalized ZIP and ZINB: Parameterizes E(Y) as a function of covariates so that β's have more intuitive interpretation (Long et al., 2014; Preisser et al., 2015)
- Longitudinal models: Min and Agresti (2005)
- Semicontinuous models: two-part mixtures of mass at zero and *continuous* distribution for positive values (e.g., medical costs)
- Marginalized semicontinuous model: Smith (2014, 2015)
- Bayesian and spatial approaches: Neelon et al. (2010, 2011, 2015)
- Many other directions – hypothesis testing, etc.

Helpful References

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Springer.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*.

Wilson, P. (2015). The misuse of the Vuong test for non-nested models to test for zero-inflation. *Economics Letters*.

Long, D. et al. (2014). A marginalized zero-inflated Poisson regression model with overall exposure effects. *Statistics in Medicine*.