Models for Count Data

Many dependent variables are count number – non-negative integer.

crimes a person has committed in lifetime

children living in a household

new companies founded in a year (in an industry)

of social protests per month in a city

Models for Count Data

Count variables can be modeled using linear regression model. Problems include:

- Possible of negative prediction.
- Count variables are often highly skewed.

Example:

crimes committed. Most people are zero or very low while few people are very high

Extreme skew violates normality assumption of OLS.

Models for Count Data

Common models for count data include:

- Poisson Regression Model
- FE Poisson Regression Model
- RE Poisson Regression Model
- Negative Binomial Regression Model
- Zero Inflated Poisson Regression Model
- Zero Inflated Negative Binomial Model
- Truncated Poisson Regression Model
- Zero Truncated Poisson Model
- Zero Truncated Negative Binomial Model

Poisson Regression Model

When dependent variable is non-negative integer or count number, the appropriated model is poisson regression model.

Dependent variable is a count variable taking small values (less than 100).

The model assumes poisson distribution.

$$\Pr[Y=y]=\frac{e^{-\mu}\mu^{y}}{y!},\quad y=0,1,2,\ldots,$$

where: μ is the intensity or rate parameter.

$$E[Y]=\mu \quad \text{and} \quad Var[Y]=\mu$$

Poisson Regression Model

By assuming relationship between μ and x as exponential mean parameterization:

$$\mu_i = \exp(x_i\beta), \;\; i = 1, 2, \ldots, N$$

Then, $$\operatorname{var}\left[y_i \mid x_i\right] = \exp(x_i\beta)$$

Log of μ_i as function of x_i can be stated as:

$$\ln \mu_i = x_i\beta$$

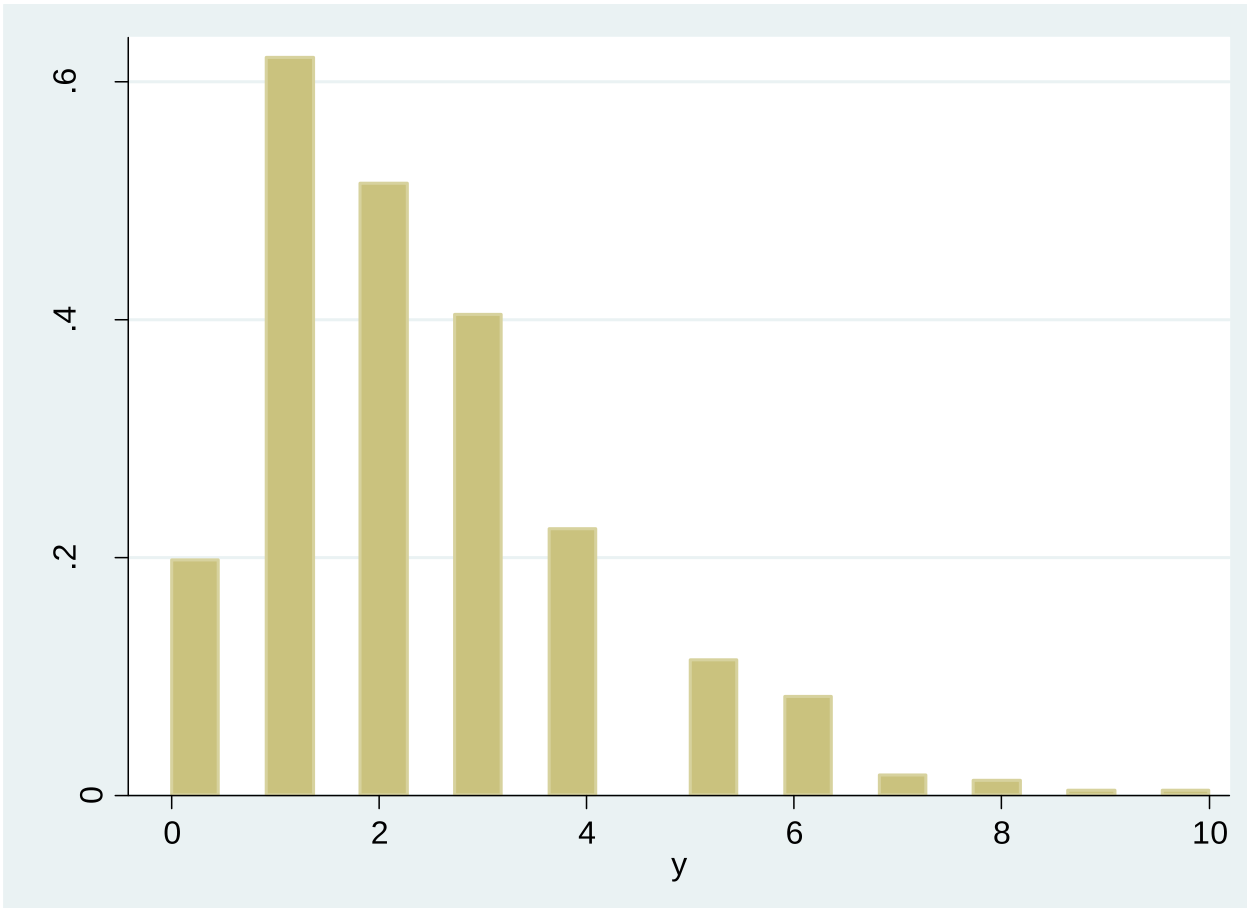
<u>Major Assumption</u>

$$E\left[y_i \mid x_i\right] = \mu_i = \exp(x_i\beta) = \operatorname{var}\left[y_i \mid x_i\right]$$

Mean equal Variance or equidispersion assumption .

Poisson Regression Model

Example: Number of times respondent going out to watch movie at the theater a month.



Poisson Regression Model

Poisson distribution function:

$$\Pr[Y = y] = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0, 1, 2, \ldots,$$

where: μ is average number of occurrence in a specified interval.

<u>Assumptions</u>:

- Independence
- Prob. of occurrence in short interval is proportional to the length of the interval
- Prob. of another occurrence in such a short interval is zero

Poisson Regression Model

The model can be estimated using MLE.

The log-likelihood Poisson function is

$$\ln L(\beta) = \sum_{i=1}^{n} w_i \{ y_i x_i \beta - \exp(x_i \beta) - \ln y_i ! \}$$

where: w_i is weight for MLE optimization.

Poisson Regression Model

Goodness of Fit Test

To test whether Poisson is appropriated, gof test compares log-likelihood of equation-level score and log-likelihood of Poisson.

Equation-level score: $score(x_i\beta) = y_i - \exp(x_i\beta)$

The log-likelihood function of equation-level score is

$$\ln L_{\max} = \sum_{i=1}^{n} w_i \left\{ -y_i \left(\ln y_i - 1 \right) - \ln y_i ! \right\}$$

The log-likelihood Poisson function is

$$\ln L(\beta) = \sum_{i=1}^{n} w_i \left\{ y_i x_i \beta - \exp(x_i\beta) - \ln y_i ! \right\}$$

Poisson Regression Model

Goodness of Fit Test

Deviance GOF test:

$$\chi^2_{Deviance} = -2\left[\ln L(\beta) - \ln L_{\max}\right] \sim \chi^2_{(n-k)}$$

Pearson GOF test:

$$\chi^2_{Pearson} = -\sum_{i=1}^{n} \frac{w_i\left(y_i - \exp(x_i\beta)\right)^2}{\exp(x_i\beta)} \sim \chi^2_{(n-k)}$$

Rejection of the goodness of fit test means the data are not poisson distributed.

Failed to reject means the data is poisson distributed.

Poisson Regression Model

Evaluation:

GOF Test to ensure Poisson is appropriated

Interpretation:

1. Sign & Meaning

Marginal Effect: $$\frac{\partial E\left[y|x\right]}{\partial x_j} = \hat{\beta}_j \exp(x_i \beta)$$

Incidence Rate Ratio: $$IRR = \exp\left(\hat{\beta}_j\right)$$

2. Overall LR Chi-squares Test
3. Pseudo R^2
4. Individual Test – z-test

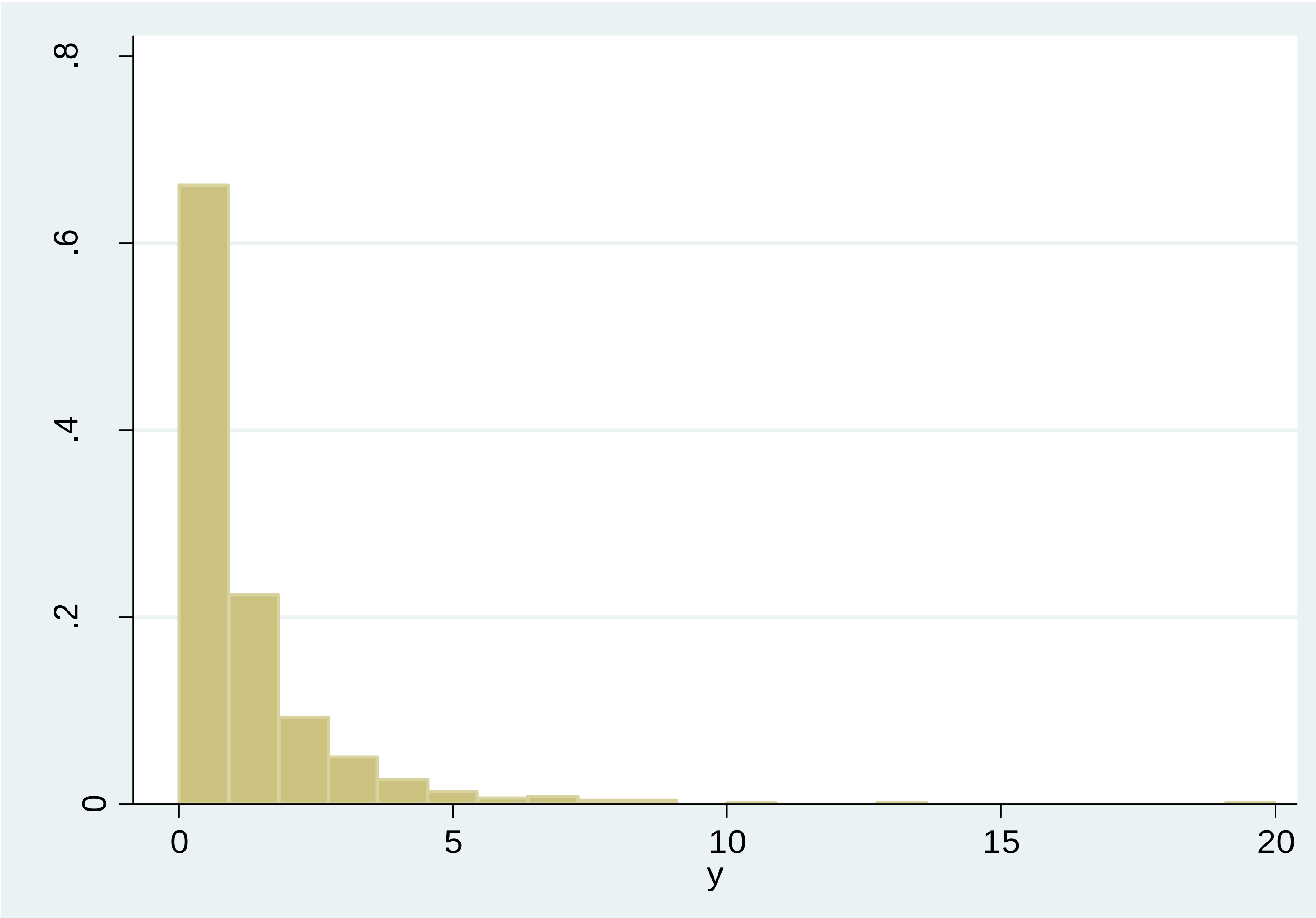
Negative Binomial Model

Poisson regression model is too restrictive since the distribution is parameterized a single parameter μ.

In some cases, counted data might have variance exceeds the mean. This problem is called ***overdispersion.*** The model should be ***negative binomial model.***

Negative Binomial Model

Overdispersion



Negative Binomial Model

Poisson with overdispersion can be addressed like adding error term to the Poisson model:

$$\mu_i = \exp(x_i\beta + \varepsilon_i),\ \ i = 1, 2, \ldots, N$$

Then, $\operatorname{var}\left[y_i \middle| x_i \right] > \mu_i$

<u>Additional Assumptions</u>

$E\left[\exp(\varepsilon_i)\right] = 1$ and $Var\left[\exp(\varepsilon_i)\right] = \dfrac{1}{\delta} = \alpha$

$\exp(\varepsilon_i)$ is Gamma distributed $F_{Gamma}(\cdot)$

$$F_{Gamma}\left(\exp(\varepsilon_i)\right) = \frac{\alpha^{\alpha} \exp(-\alpha\varepsilon_i)\varepsilon_i^{\alpha-1}}{\Gamma(\alpha)}$$

Negative Binomial Model

Negative Binomial distribution function:

$$\Pr\left[Y = y\right] = \frac{\Gamma\left(\alpha^{-1} + y\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y+1\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{y}$$

where: $\alpha = \frac{1}{\delta}$

If $\alpha = 0$, the model is Poisson model.

α represents the extent of overdispersion.

<u>Overdispersion Test</u>:

Test $H_0 : \alpha = 0$ using LR-Chi-square Bar Test.

Rejection means Negative Binomial is appropriated.

By Tatre Jantarakolica

Zero-inflated Poisson Model

If there are more zero in the data, ***excess zeros problem*** occurs. The model should be ***zero-inflated poisson model.***

ZIP combines Logit and Poisson. Logit $(i \in S)$ determines 0 or not, if not, Poisson $(i \notin S)$ determines number.

Log-likelihood function:

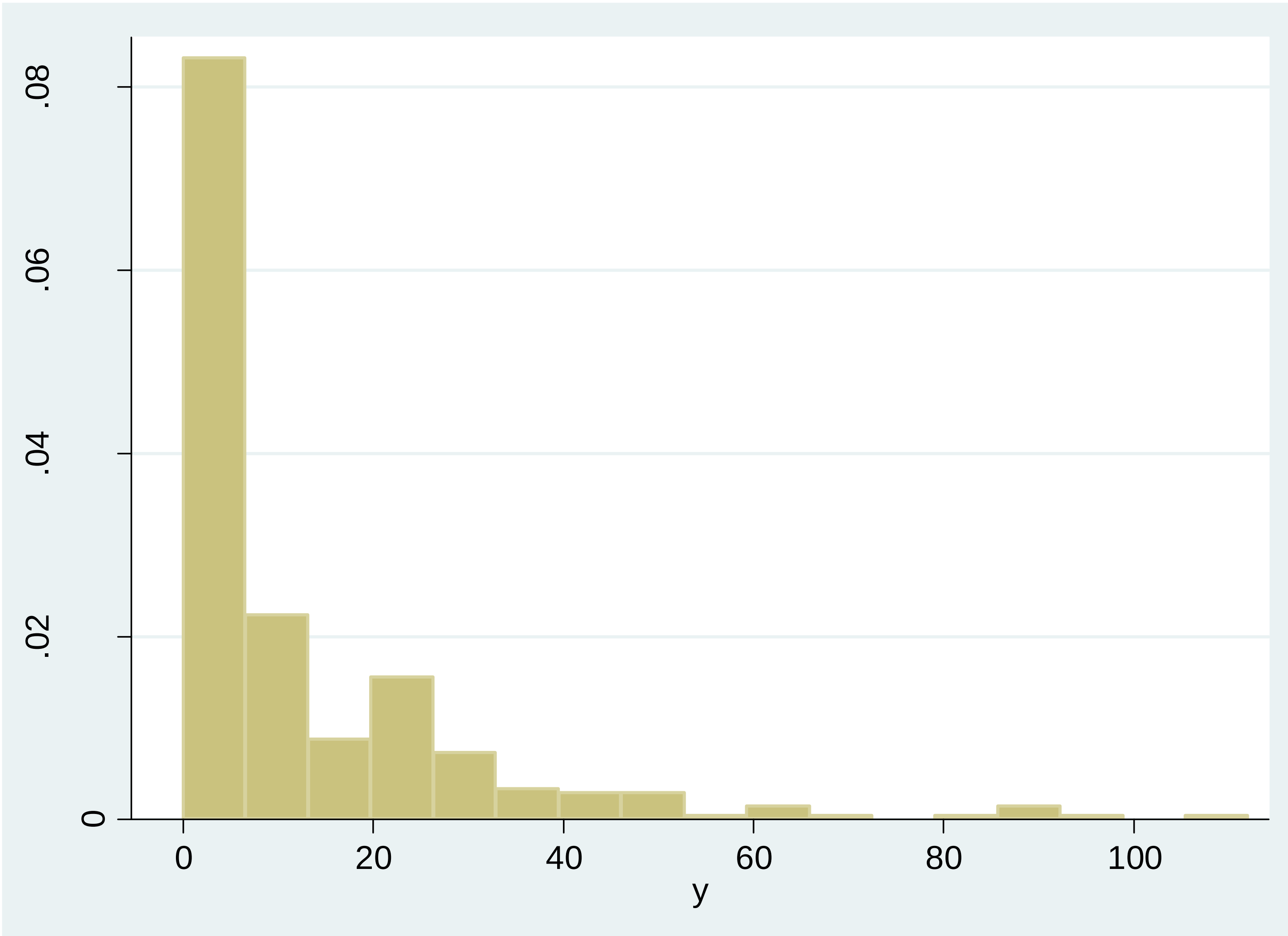
$$\ln L = \sum_{i \in S} w_i \ln\left[F\left(\exp(x_i \beta\right) + \left\{1 - F\left(\exp(x_i \beta)\right\} \exp\left(-\mu_i\right)\right] +$$

$$\sum_{i \notin S} w_i \left[\ln\left\{1 - F\left(\exp(x_i \beta)\right\} - \mu_i + \exp\left(x_i \beta\right) y_i - \ln(y_i !)\right]$$

where: $F(\cdot)$ is inverse of Logit link.

Zero-inflated Poisson Model

Excess Zeros Problem



Zero-inflated Poisson Model

Vuong test:

Test whether ZIP vs Poisson.

Rejection implies ZIP is more appropriated.

Generalized Linear Model (GLM)

<u>Assume Probability Distribution</u>:

- Gaussian
- Binomial
- Poisson
- Negative Binomial
- Gamma