Topics

Limited Dependent Variables

- Limited Distribution
  - Truncated Distribution
  - Censored Distribution
- Nonnormal Distribution
  - Generalized Linear Models
  - Models for Counted Data
    - Poisson Regression Models
    - Negative Binomial Regression Model
    - Zero Inflated Poisson Regression Model

By Tatre Jantarakolica

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
- 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
- FE Poisson Regression Model
- RE Poisson Regression Model
- FE Negative Binomial Regression Model
- RE Negative Binomial Regression 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.

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

where: μ is the intensity or rate parameter.

$E[Y]=\mu$ and $Var[Y]=\mu$

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

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

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

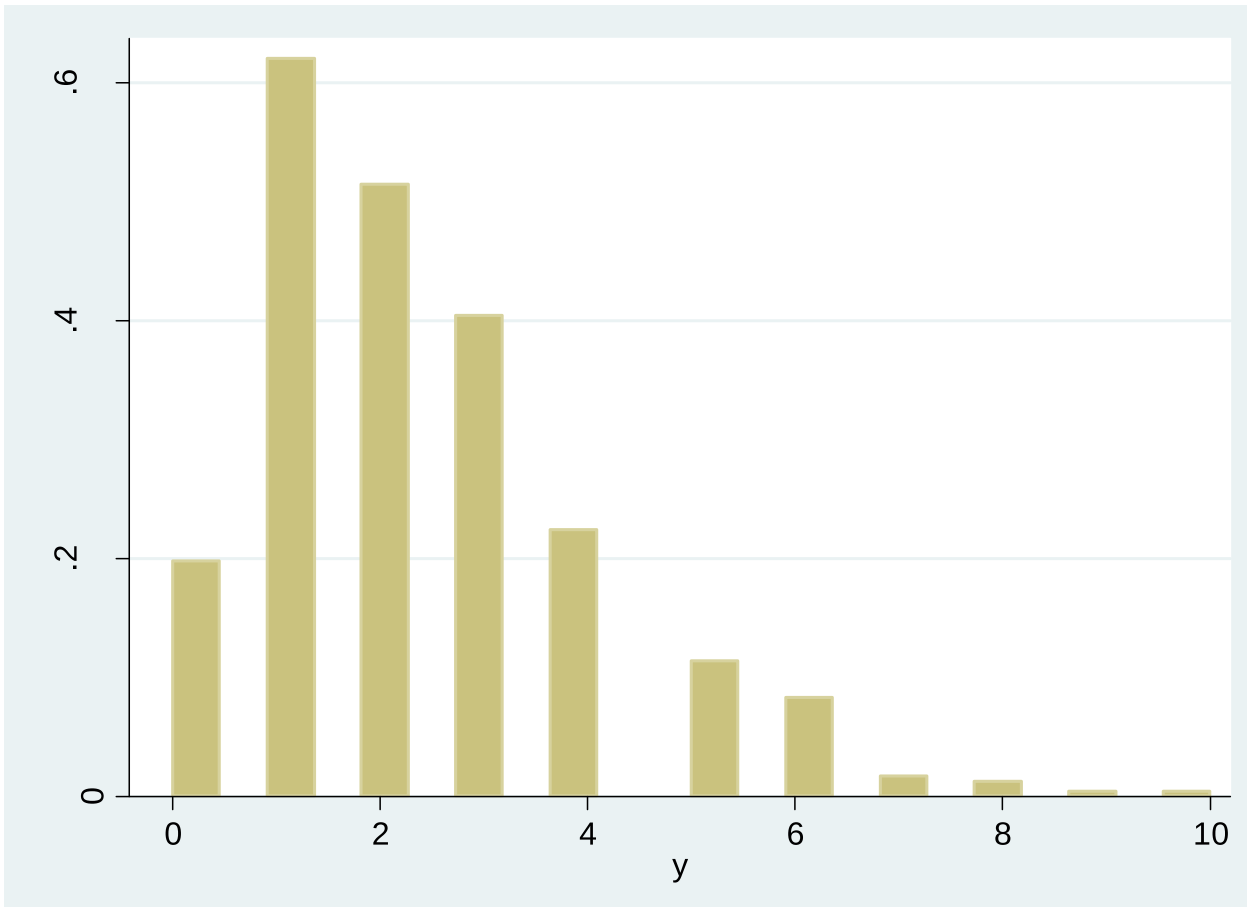
<u>Major Assumption</u>

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

Mean equal Variance or equidispersion assumption .

Poisson Regression Model

<u>Example</u>: 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

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

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

1. Sign & Meaning

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

   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

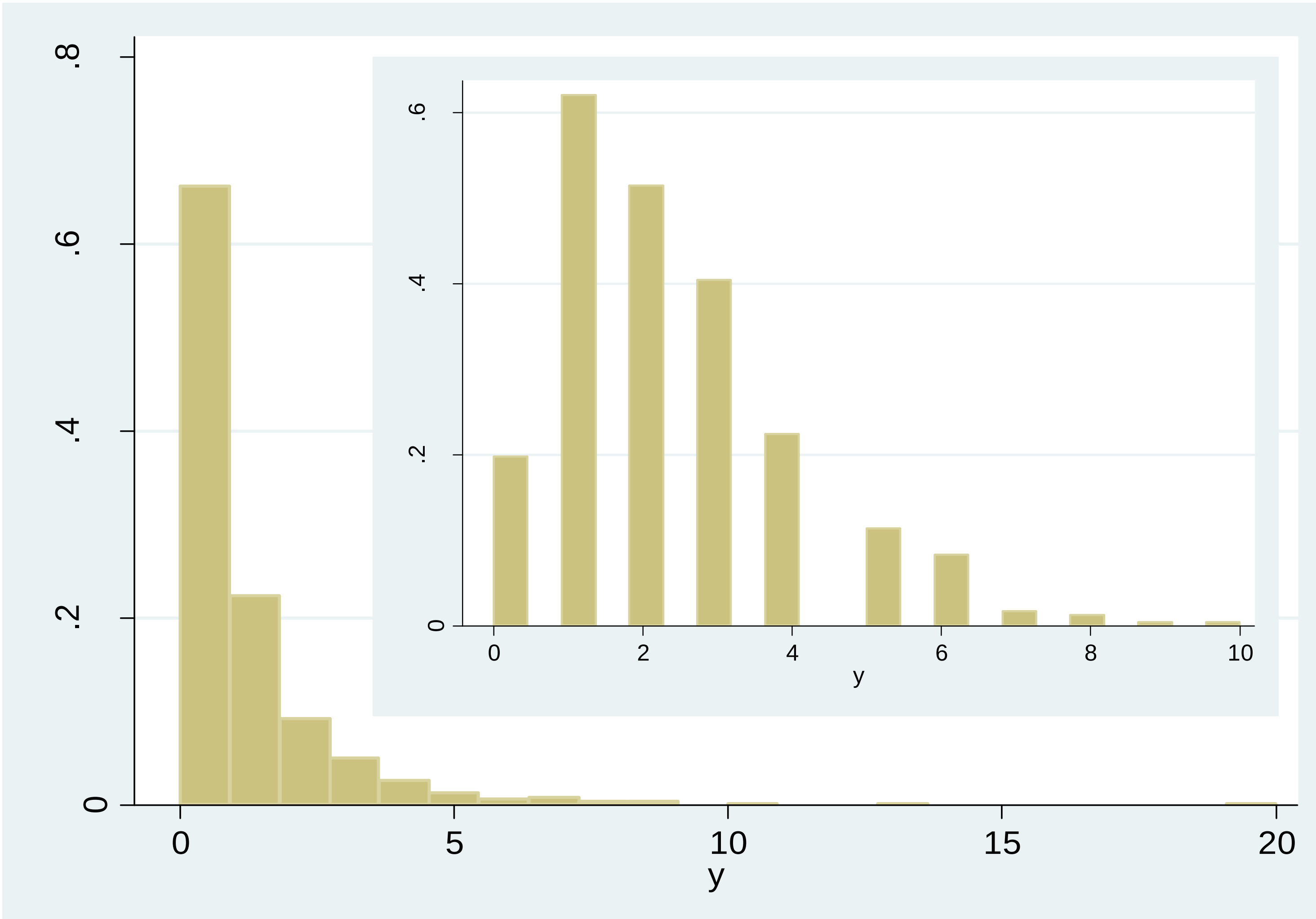
Negative Binomial Model

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

<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:

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

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.

Negative Binomial Model

Evaluation:

Overdispersion Test $\alpha = 0$ to ensure Negative Binomial model is appropriated

1. Sign & Meaning

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

   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

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:

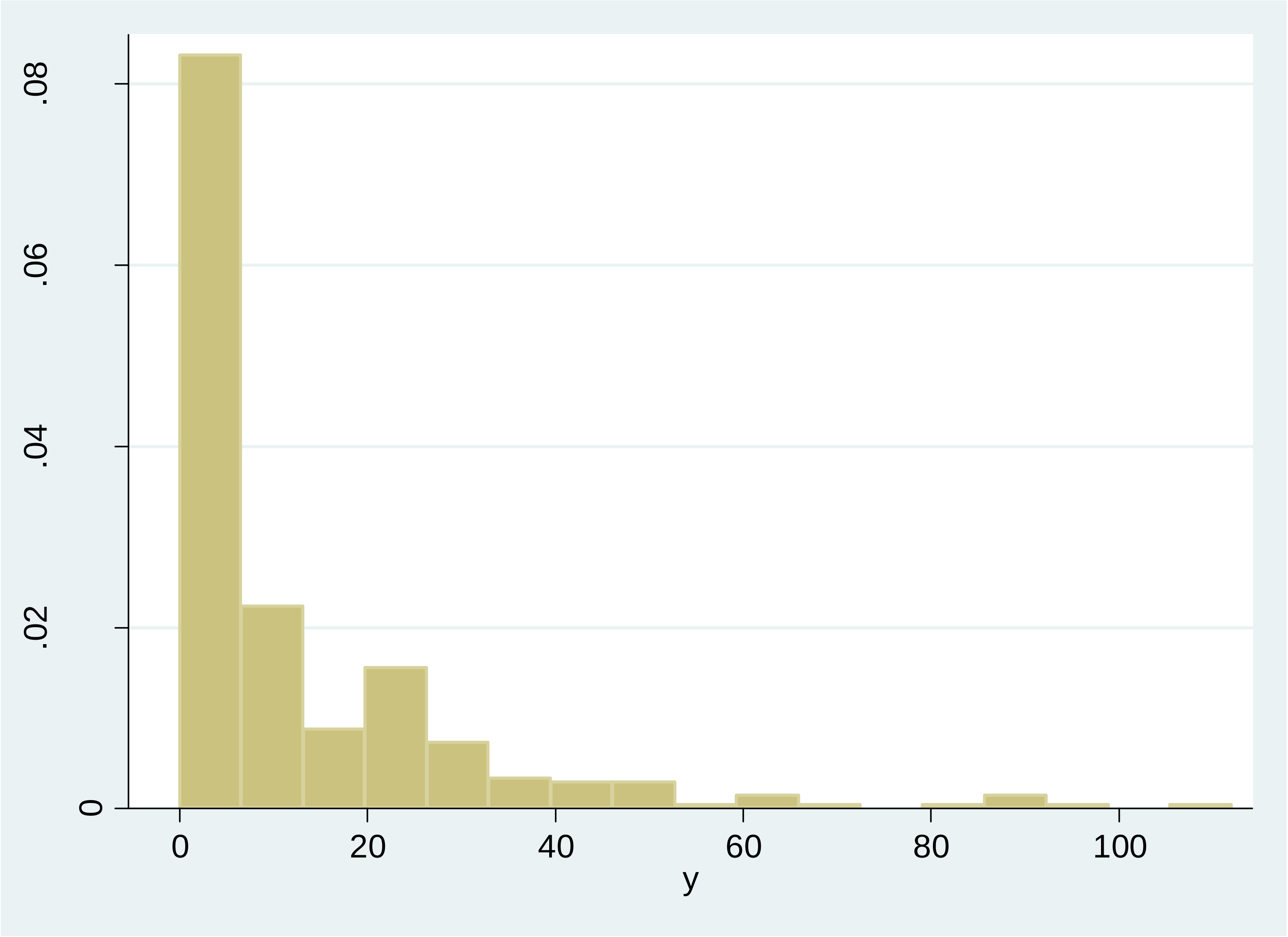
$$\ln L = \sum_{i \in S} w_i \ln\left[F\left(\exp(x_i\beta\right) + \left\{1 - F\left(\exp(x_i\beta\right)\right\}\exp\left(-\mu_i\right)\right] +$$

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

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.

Zero-inflated Poisson Model

Evaluation:

Vuong test to ensure Zero-inflated Poisson model is appropriated

1. Sign & Meaning

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

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

2. Overall LR Chi-squares Test

3. Log-likelihood value

4. Individual Test – z-test

Poisson Regression Model using Panel Data

Fixed Effects Poisson regression model

- Maximize conditional log-likelihood function.

Random Effects Poisson regression model

- Maximize log-likelihood function with random effects using Guass-Hermite Quadrature algorithm.

FE Poisson Regression Model

- Poisson regression model using Panel data might have fixed-effects.
- FE Poisson can be estimated by maximize conditional log-likelihood.

Conditional likelihood function:

$$\Pr\left(Y_{it}=y_{it}\middle|x_{it},\sum_t Y_{it}=\sum_t y_{it}\right)=\prod_{i=1}^{n}\left(\sum_t y_{it}\right)!\prod_{t=1}^{T_i}\frac{\exp\left(x_{it}\beta\right)^{y_{it}}}{y_{it}!\left\{\sum_s \exp\left(x_{is}\beta\right)\right\}^{y_{it}}}$$

FE Poisson Regression Model

Likelihood function:

$$\Pr(Y_{it}=y_{it}\,|\,x_{it})=\prod_{i=1}^{n}\prod_{t=1}^{T_i}\frac{1}{y_{it}!}\exp\{-\exp(\alpha_i)\exp(x_{it}\beta)+\alpha_i y_{it}\}\exp(x_{it}\beta)^{y_{it}}$$

$$=\prod_{i=1}^{n}\left(\prod_{t=1}^{T_i}\frac{\exp(x_{it}\beta)^{y_{it}}}{y_{it}!}\right)\exp\left\{-\exp(\alpha_i)\sum_{t=1}^{T_i}\exp(x_{it}\beta)+\alpha_i\sum_{t=1}^{T_i}y_{it}\right\}$$

$$\Pr\left(\sum_t Y_{it}=\sum_t y_{it}\,\middle|\,x_{it}\right)=\frac{1}{\left(\sum_t y_{it}\right)!}\exp\left\{-\exp(\alpha_i)\sum_{t=1}^{T_i}\exp(x_{it}\beta)+\alpha_i\sum_{t=1}^{T_i}y_{it}\right\}\left\{\sum_{t=1}^{T_i}\exp(x_{it}\beta)\right\}^{\sum_t y_{it}}$$

Conditional likelihood function is free of α_i:

$$\Pr\left(Y_{it}=y_{it}\,\middle|\,x_{it},\sum_t Y_{it}=\sum_t y_{it}\right)=\prod_{i=1}^{n}\left(\sum_t y_{it}\right)!\prod_{t=1}^{T_i}\frac{\exp(x_{it}\beta)^{y_{it}}}{y_{it}!\left\{\sum_s\exp(x_{is}\beta)\right\}^{y_{it}}}$$

RE Poisson Regression Model

RE Poisson can be estimated by maximize log-likelihood assuming Gamma distribution for RE using adaptive Guass-Hermite Quadrature `(default in STATA)`.

Log-likelihood function:

$$\ln L = \sum_{i=1}^{n} w_i \left\{ \begin{array}{l} \log\Gamma\left(\theta + \sum_{t=1}^{n_i} y_{it}\right) - \sum_{t=1}^{n_i} \log\Gamma(1+y_{it}) - \log\Gamma(\theta) + \theta \log\alpha_i \\ + \log(1-\alpha_i)\sum_{t=1}^{n_i} y_{it} + \sum_{t=1}^{n_i} y_{it}(x_{it}\beta) - \left(\sum_{t=1}^{n_i} y_{it}\right)\log\left(\sum_{t=1}^{n_i}\lambda_{it}\right) \end{array} \right\}$$

where: w_i is weight for panel i.

RE Poisson Regression Model

RE Poisson can be estimated by maximize log-likelihood assuming normal distribution for RE using nonadaptive Guass-Hermite Quadrature.

Log-likelihood function:

$$\ln L(\beta) = \sum_{i=1}^{n} w_i \log\left[\frac{1}{\sqrt{\pi}} \sum_{m=1}^{M} w_m^* \prod_{t=1}^{T_i} F\left\{ y_{it}, x_{it}\beta + \alpha_m^* \left(\frac{2\rho}{1-\rho} \right)^{1/2} \right\} \right]$$

where: $\rho = \frac{\sigma_\alpha^2}{\left(\sigma_\alpha^2 + 1\right)}$

RE Poisson Regression Model

Evaluation:

LR test of $\sigma_{\alpha} = 0$ to ensure RE Poisson model is appropriated

1. Sign & Meaning

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

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

2. Overall Wald Chi-squares Test
3. Log-likelihood Value
4. Individual Test – z-test

Negative Binomial Regression Model using Panel Data

Fixed Effects Negative Binomial regression model

- Maximize conditional log-likelihood function.

Random Effects Negative Binomial regression model

- Maximize log-likelihood function with random effects using Guass-Hermite Quadrature algorithm.

RE NB Regression Model

Negative Binomial regression model using Panel data can be stated as:

$$\Pr(Y_{it} = y_{it} \mid x_{it}, \delta_i) = \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it} + 1)} \left(\frac{1}{1+\delta_i}\right)^{\lambda_{it}} \left(\frac{\delta_i}{1+\delta_i}\right)^{y_{it}}$$

where: δ_i is the dispersion parameter which can be varied across groups.

$$1/(1+\delta_i) \sim Beta(r, s)$$

RE NB Regression Model

RE Negative Binomial can be estimated by maximize log-likelihood using Guass-Hermite Quadrature.

Log-likelihood function:

$$\ln L(\beta)=\sum_{i=1}^{n} w_i \left[\begin{array}{l} \ln\Gamma(r+s)+\ln\Gamma\left(r+\sum_{k=1}^{n_i}\lambda_{ik}\right)+\ln\Gamma\left(s+\sum_{k=1}^{n_i}y_{ik}\right) \\ -\ln\Gamma(r)-\ln\Gamma(s)-\ln\Gamma\left(r+s+\sum_{k=1}^{n_i}\lambda_{ik}+\sum_{k=1}^{n_i}y_{ik}\right) \\ +\sum_{t=1}^{n_i}\left\{\ln\Gamma(\lambda_{it}+y_{it})-\ln\Gamma(\lambda_{it})-\ln\Gamma(y_{it}+1)\right\} \end{array} \right]$$

where: w_i is weight for panel i.

FE NB Regression Model

Negative Binomial regression model using Panel data might have fixed-effects.

FE Negative Binomial can be estimated by maximize conditional log-likelihood.

Conditional likelihood function:

$$\ln L(\beta) = \sum_{i=1}^{n} w_i \left[\ln\Gamma\left(\sum_{t=1}^{n_i} \lambda_{it}\right) + \ln\Gamma\left(\sum_{t=1}^{n_i} y_{it} + 1\right) - \ln\Gamma\left(\sum_{t=1}^{n_i} \lambda_{it} + \sum_{t=1}^{n_i} y_{it}\right) + \sum_{t=1}^{n_i} \left\{ \ln\Gamma(\lambda_{it} + y_{it}) - \ln\Gamma(\lambda_{it}) - \ln\Gamma(y_{it} + 1) \right\} \right]$$

RE NB Regression Model

<u>Evaluation</u>:

LR test comparing panel vs pooled to ensure RE Negative Binomial model is appropriated.

1. Sign & Meaning

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

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

2. Overall Wald Chi-squares Test
3. Log-likelihood Value
4. Individual Test – z-test

Generalized Linear Model (GLM)

Assume Probability Distribution:

- Gaussian
- Binomial
- Poisson
- Negative Binomial
- Gamma