THE UNIVERSITY OF MEMPHIS

Dreamers. Thinkers. Doers.

Count Data Models

CIVL 7012/8012

In Today's Class

- Count data models
- Poisson Models
- Overdispersion
- Negative binomial distribution models
- Comparison
- Zero-inflated models
- R-implementation

Count Data

- In many a phenomena the dependent variable is of the count type, such as:
  - The number of patents received by a firm in a year
  - The number of visits to a dentist in a year
  - The number of speeding tickets received in a year
- The underlying variable is discrete, taking only a finite non-negative number of values.
  - In many cases the count is 0 for several observations
  - Each count example is measured over a certain finite time period.

Models for Count Data

- Poisson Probability Distribution: Regression models based on this probability distribution are known as Poisson Regression Models (PRM).

- Negative Binomial Probability Distribution: An alternative to PRM is the Negative Binomial Regression Model (NBRM), used to remedy some of the deficiencies of the PRM.

Can we apply OLS to count data?

Dependent Variable: P90
Method: Least Squares
Sample: 1 181
Included observations: 181

	Coefficient	Std. Error	t-Statistic	Prob.
C	−250.8386	55.43486	−4.524925	0.0000
LR90	73.17202	7.970758	9.180058	0.0000
AEROSP	−44.16199	35.64544	−1.238924	0.2171
CHEMIST	47.08123	26.54182	1.773851	0.0779
COMPUTER	33.85645	27.76933	1.219203	0.2244
MACHINES	34.37942	27.81328	1.236079	0.2181
VEHICLES	−191.7903	36.70362	−5.225378	0.0000
JAPAN	26.23853	40.91987	0.641217	0.5222
US	−76.85387	28.64897	−2.682605	0.0080

R-squared	0.472911	Mean dependent var	79.74586
Adjusted R-squared	0.448396	S.D. dependent var	154.2011
S.E. of regression	114.5253	Akaike info criterion	12.36791
Sum squared resid	2255959.	Schwarz criterion	12.52695
Log likelihood	−1110.296	Durbin–Watson stat	1.946344
F-statistic	19.29011	Prob(F-statistic)	0.000000

Note: P(90) is the number of patents received in 1990 and LR(90) is the log of R&D expenditure in 1990. Other variables are self-explanatory.

Patent data from 181firms

LR 90: log (R&D Expenditure)
Dummy categories
- AEROSP: Aerospace
- CHEMIST: Chemistry
- Computer: Comp Sc.
- Machines: Instrumental Engg
- Vehicles: Auto Engg.
- Reference: Food, fuel others

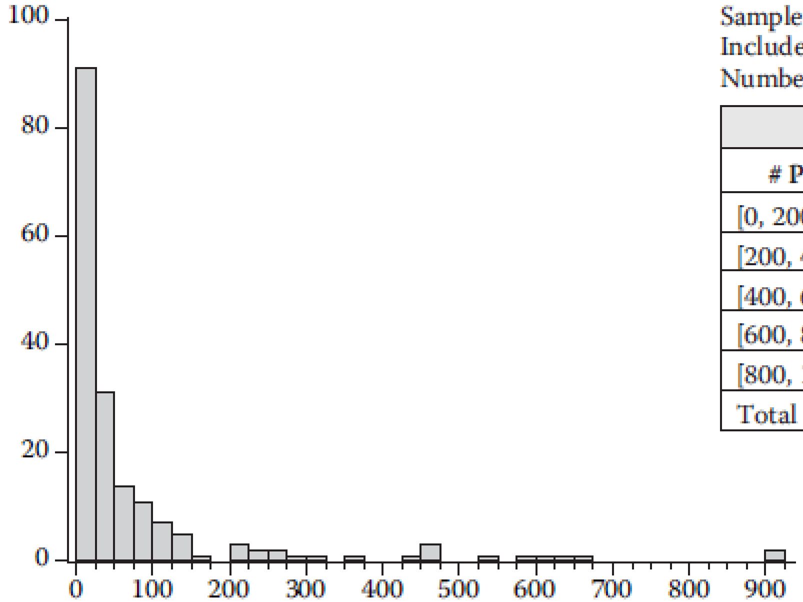
Dummy countries
- Japan:
- US:
- Reference: European countries

Inferences from the example *(1)*

- R&D have +ve influence
  - 1% increase in R&D expenditure increases the likelihood of patent increase by 0.73% ceteris paribus
- Chemistry has received 47 more patents compared to the reference category
- Similarly vehicles industry has received 191 lower patents compared to the reference category
- County dummy suggests that on an average US firms received 77 few patents compared to the reference category

Inferences from the example *(2)*

- OLS may not be appropriate as the number of patents received by firms is usually a small number



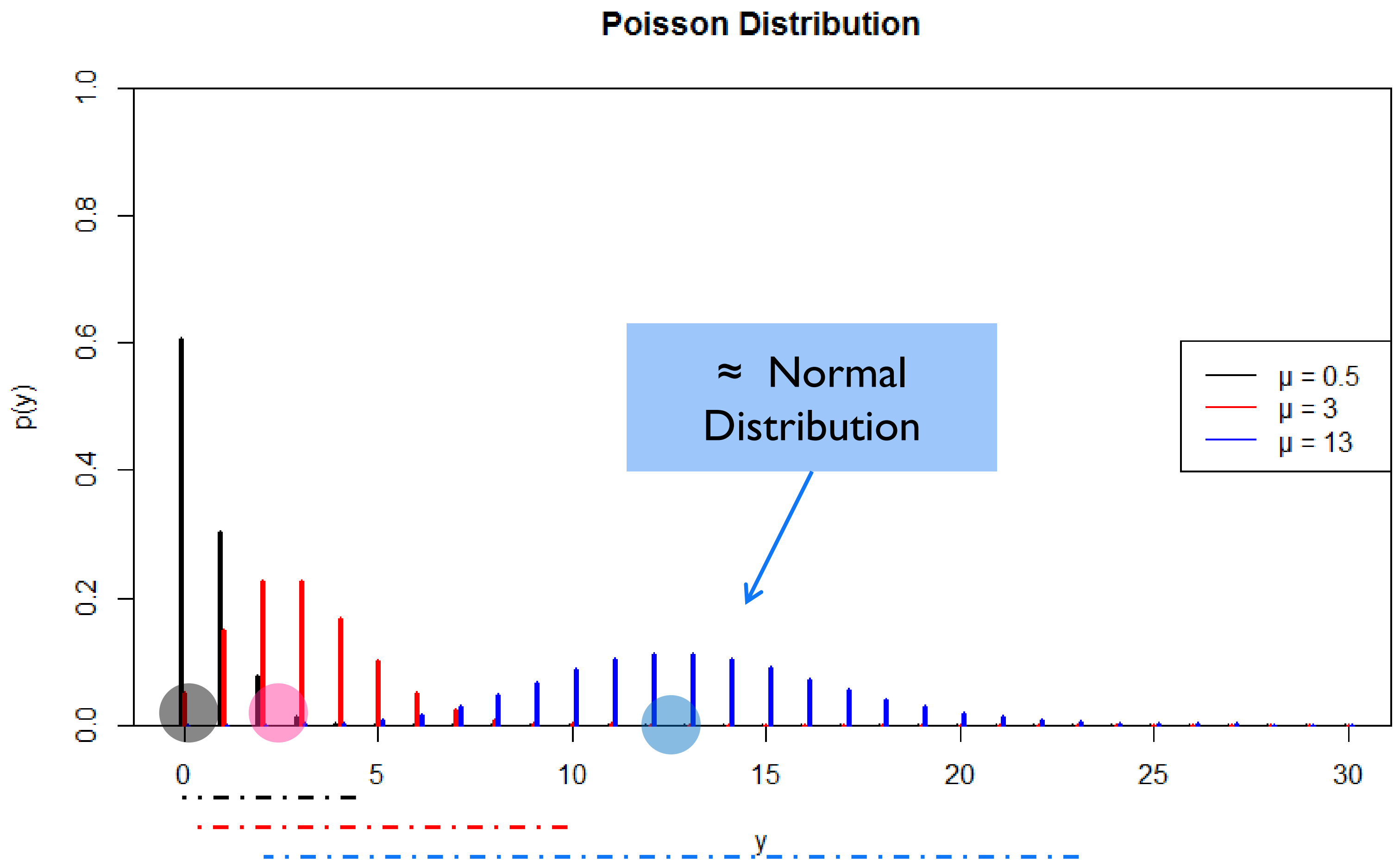
Tabulation of P90
Sample: 1 181
Included observations: 181
Number of categories: 5

# Patents	Count	Percent	Cumulative Count	Cumulative Percent
[0, 200)	160	88.40	160	88.40
[200, 400)	10	5.52	170	93.92
[400, 600)	6	3.31	176	97.24
[600, 800)	3	1.66	179	98.90
[800, 1000)	2	1.10	181	100.00
Total	181	100.00	181	100.00

Inferences from the example *(2)*

- The histogram is highly skewed to the right
- Coefficient of skewness: 3.3
- Coefficient of kurtosis: 14
- For a typical normal distribution
  - Skewness is 0 and kurtosis is 3
- We can not use OLS to work with count data

Poisson Distribution



Small mean ➡ Small count numbers ➡ Many zeroes ➡ Poisson Regression
Large mean ➡ Large count numbers ➡ Few/none zeroes ➡ OLS Regression

Poisson Regression Models *(1)*

- If a discrete random variable Y follows the Poisson distribution, its probability density function (PDF) is given by:

$$f(Y=y_i)=\Pr(Y=y_i)=\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!},\ \ y_i=0,1,2\ldots$$

where $f(Y|y_i)$ denotes the probability that the discrete random variable Y takes non-negative integer value y_i,

and λ is the parameter of the Poisson distribution.

Poisson Regression Models *(2)*

- **Equidispersion**: A unique feature of the Poisson distribution is that the mean and the variance of a Poisson-distributed variable are the same
- If variance > mean, there is **overdispersion**

Poisson Regression Models *(3)*

- The Poisson regression model can be written as:

$$y_i = E(y_i) + u_i = \lambda_i + u_i$$

- where the *y*s are independently distributed as Poisson random variables with mean λ for each individual expressed as:

- $\lambda_i = E(y_i|X_i) = \exp[B_1 + B_2X_{2i} + \ldots + B_kX_{ki}] = \exp(BX)$

- Taking the exponential of *BX* will guarantee that the mean value of the count variable, λ, will be positive.
- For estimation purposes, the model, estimated by ML, can be written as:

$$y_i = \frac{e^{-XB}\lambda_i^{y_i}}{y_i!} + u_i, y_i = 0,1,2\ldots$$

Solution

- Apply maximum likelihood approach

$$L(\beta) = \prod_i \frac{EXP[-EXP(\beta X_i)][EXP(\beta X_i)]^{y_i}}{y_i!}.$$

- Log of likelihood function

$$LL(\beta) = \sum_{i=1}^{n} \left[-EXP(\beta X_i) + y_i \beta X_i - LN(y_i!)\right]$$

Elasticity

- To provide some insight into the implications of parameter estimation results, elasticities are computed to determine the marginal effects of the independent variables.
- Elasticities provide an estimate of the impact of a variable on the expected frequency and are interpreted as the effect of a 1% change in the variable on the expected frequency λ_i

$$E_{x_{ik}}^{\lambda_i} = \frac{\partial \lambda_i}{\lambda_i} \times \frac{x_{ik}}{\partial x_{ik}} = \beta_k x_{ik}$$

Elasticity-Example

- For example, an elasticity of –1.32 is interpreted to mean that a 1% increase in the variable reduces the expected frequency by 1.32%.
- Elasticities are the correct way of evaluating the relative impact of each variable in the model.
- Suitable for continuous variables
- Calculated for each individual observation
- Can be calculated as an average for the sample

Pseudo Elasticity

- What happens for discrete (dummy variables)
- The pseudo-elasticity gives the incremental change in frequency caused by changes in the indicator variables.

$$E_{x_{ik}}^{\lambda_i} = \frac{EXP(\beta_k) - 1}{EXP(\beta_k)}$$

Poisson Regression Goodness of fit measures

- Likelihood ratio test statistics

$$X^2 = -2[LL(\boldsymbol{\beta}_R) - LL(\boldsymbol{\beta}_U)].$$

- Rho-square statistics

$$\rho^2 = 1 - \frac{LL(\beta)}{LL(0)}$$

Patent Data with Poisson Model

Dependent Variable: P90
Method: ML/QML – Poisson Count (Quadratic hill climbing)
Sample: 1 181
Included observations: 181
Convergence achieved after 6 iterations
Covariance matrix computed using second derivatives

	Coefficient	Std. Error	z-Statistic	Prob.
C	–0.745849	0.062138	–12.00319	0.0000
LR90	0.865149	0.008068	107.2322	0.0000
AEROSP	–0.796538	0.067954	–11.72164	0.0000
CHEMIST	0.774752	0.023126	33.50079	0.0000
COMPUTER	0.468894	0.023939	19.58696	0.0000
MACHINES	0.646383	0.038034	16.99479	0.0000
VEHICLES	–1.505641	0.039176	–38.43249	0.0000
JAPAN	–0.003893	0.026866	–0.144922	0.8848
US	–0.418938	0.023094	–18.14045	0.0000

R-squared	0.675516	Mean dependent var	79.74586
Adjusted R-squared	0.660424	S.D. dependent var	154.2011
S.E. of regression	89.85789	Akaike info criterion	56.24675
Sum squared resid	1388804.	Schwarz criterion	56.40579
Log likelihood	–5081.331	LR statistic	21482.10
Restr. log likelihood	–15822.38	Prob(LR statistic)	0.000000
Avg. log likelihood	–28.07365		

Note: LR90 is the logarithm of R&D expenditure in 1990.

LR90 coefficient suggests that 1%
Increase in R&D expenditure will
Increase the likelihood of patent
Receipt by 0.86%

For machines dummy
The number of patents received by
Machines category is
100(exp(0.6464)-1)= 90.86% compared
To the reference category

See the likelihood test statistics
2(-5081.331-(-15822.38))

Shows overall model significance

Poisson Regression Coefficient Interpretation

Example 1:

$y_i \sim \text{Poisson}(\exp(2.5 + 0.18X_i))$

$(e^{0.18}) = 1.19$

A one unit increase in X, will **increase** the average *number of y by 19%*

Example 2:

$y_i \sim \text{Poisson}(\exp(2.5 - 0.18X_i))$

$(e^{-0.18}) = 0.83$

A one unit increase in X, will **decrease** the average number of y by 17%

Safety Example *(1)*

Summary of Variables in California and Michigan Accident Data

Variable Abbreviation	Variable Description	Maximum/ Minimum Values	Mean of Observations	Standard Deviation of Observations
STATE	Indicator variable for state: 0 = California; 1 = Michigan	1/0	0.29	0.45
ACCIDENT	Count of injury accidents over observation period	13/0	2.62	3.36
AADT1	Average annual daily traffic on major road	33058/2367	12870	6798
AADT2	Average annual daily traffic on minor road	3001/15	596	679
MEDIAN	Median width on major road in feet	36/0	3.74	6.06
DRIVE	Number of driveways within 250 ft of intersection center	15/0	3.10	3.90

Safety Example (2)

Poisson Regression of Injury Accident Data

Independent Variable	Estimated Parameter	t Statistic
Constant	–0.826	–3.57
Average annual daily traffic on major road	0.0000812	6.90
Average annual daily traffic on minor road	0.000550	7.38
Median width in feet	– 0.0600	– 2.73
Number of driveways within 250 ft of intersection	0.0748	4.54
Number of observations	84	
Restricted log likelihood (constant term only)	–246.18	
Log likelihood at convergence	–169.25	
Chi-squared (and associated p-value)	153.85 (<0.0000001)	
R_p-squared	0.4792	
G^2	176.5	

$$E[y_i] = \lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i)$$

$$= EXP\begin{pmatrix} -0.83 + 0.00008(AADT1_i) \\ +0.0005(AADT2_i) - 0.06(MEDIAN_i) + 0.07(DRIVE_i) \end{pmatrix}$$

- Mathematical expression

Safety Example *(3)*

- The model contains a constant and four variables
  - two average annual daily traffic (*AADT*) variables, median width, and number of driveways.
- The mainline *AADT* appears to have a smaller influence than the minor road *AADT*, contrary to what is expected.
- Also, as median width increases, accidents decrease.
- Finally, the number of driveways close to the intersection increases the number of intersection injury accidents.
- The signs of the estimated parameters are in line with expectation.

Elasticity

Independent Variable	Elasticity
Average annual daily traffic on major road	1.045
Average annual daily traffic on minor road	0.327
Median width in feet	–0.228
Number of driveways within 250 ft of intersection	0.232

- 1% increase in AADT of the major road increases the expected frequency by 1.045%
- 1% increase in median width decreases the expected frequency by -0.228%

Limitations

- Poisson regression is a powerful tool
- But like any other model has limitations
- Three common analysis errors
  - Failure to recognize equidispersion
  - Failure to recognize if the data is truncated
  - If the data contains preponderance of zeros

Equidispersion Test *(1)*

Equidispersion can be tested as follows:

- 1. Estimate Poisson regression model and obtain the predicted value of Y.
- 2. Subtract the predicted value from the actual value of Y to obtain the residuals, e_i.
- 3. Square the residuals, and subtract from them from actual Y.
- 4. Regress the result from (3) on the predicted value of Y squared.
- 5. If the slope coefficient in this regression is statistically significant, reject the assumption of equidispersion.

Equidispersion Test *(2)*

- 6. If the regression coefficient in (5) is positive and statistically significant, there is **overdispersion**. If it is negative, there is **under-dispersion**. In any case, reject the Poisson model. However, if this coefficient is statistically insignificant, you need not reject the PRM.

- *Can correct standard errors by the method of quasi-maximum likelihood estimation (QMLE) or by the method of generalized linear model (GLM).*

Patent Example Equidispersion

Dependent Variable: (P90-P90F)^2-P90
Method: Least Squares
Sample: 1 181
Included observations: 181

	Coefficient	Std. Error	t-Statistic	Prob.
P90F^2	0.185270	0.023545	7.868747	0.0000

R-squared	0.185812	Mean dependent var	7593.204
Adjusted R-squared	0.185812	S.D. dependent var	24801.26
S.E. of regression	22378.77	Akaike info criterion	22.87512
Sum squared resid	9.01E+10	Schwarz criterion	22.89279
Log likelihood	–2069.199	Durbin–Watson stat	1.865256

Note: P90F is the predicted value of P90 from Table 12.4 and P90F^2 = P90F squared.

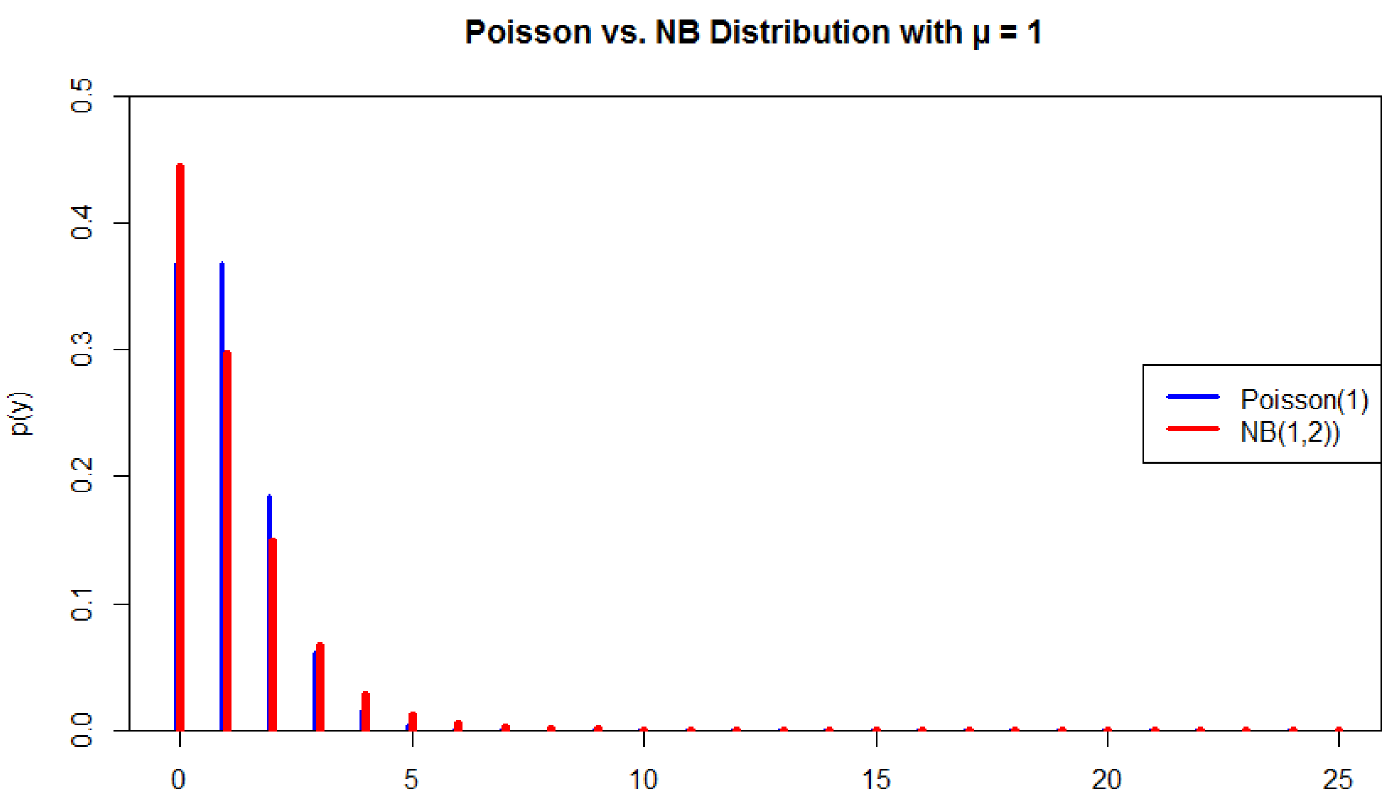
Overdispersion

- Observed variance > Theoretical variance
- The variation in the data is beyond Poisson model prediction

$$Var(Y) = \mu + \alpha * f(\mu), \quad (\alpha\text{: dispersion parameter})$$

- $\alpha = 0$, indicates standard dispersion (Poisson Model)
- $\alpha > 0$, indicates over-dispersion (Reality, Neg-Binomial)
- $\alpha < 0$, indicates under-dispersion (Not common)

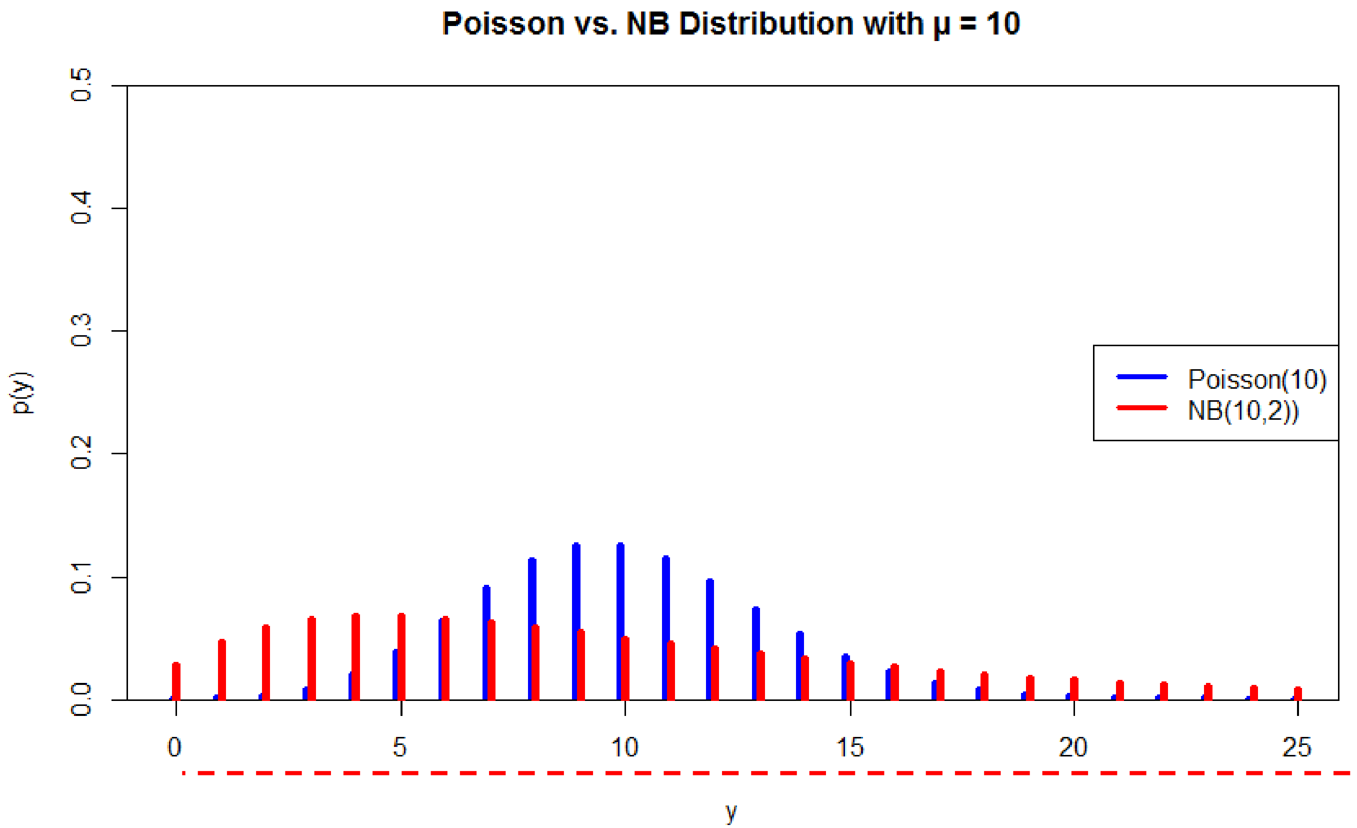
Negative Binomial vs. Poisson



Many zeroes ➡ Small mean ➡ Small count numbers ➡ Poisson Regression

Many zeroes ➡ Small mean ➡ more variability in count numbers ➡ NB Regression

Negative Binomial vs. Poisson



Many zeroes ➡ Large mean ➡ NB Regression

Few\none zeroes ➡ Large mean ➡ OLS Regression

Negative Binomial Regression Model

$$y_i \sim NB(\mu_i, \alpha)$$

- $\mu = E(y|x) = Exp(\beta X_i) = e^{\beta X_i}$
- α is the over dispersion parameter
- $Var(y|x) = \mu + \alpha\, \mu^2$ or ($\mu + \alpha\, \mu$, less used form)

When $\alpha = 0$, NB distribution is the same as a Poisson distribution

NB Probability Distribution

- One formulation of the negative binomial distribution can be used to model count data with over-dispersion

$$P(Y = y|\mu, \alpha) = \frac{\Gamma(y+\alpha^{-1})}{y!\,\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}, Where\ y = 0,1,2,..$$

Negative Binomial Regression Models

- For the Negative Binomial Probability Distribution, we have:

$$\sigma^2 = \mu + \frac{\mu^2}{r}; \mu > 0, r > 0$$

where σ^2 is the variance, μ is the mean and r is a parameter of the model.

- Variance is always larger than the mean, in contrast to the Poisson PDF.
- The NBPD is thus more suitable to count data than the PPD.
- As $r \rightarrow \infty$ and p (the probability of success) $\rightarrow$ 1, the NBPD approaches the Poisson PDF, assuming mean μ stays constant.

NB of the Patent Data

Dependent Variable: P90
Method: ML – Negative Binomial Count (Quadratic hill climbing)
Sample: 1 181
Included observations: 181
Convergence achieved after 6 iterations
Covariance matrix computed using second derivatives

	Coefficient	Std. Error	z-Statistic	Prob.
C	−0.407242	0.502841	−0.809882	0.4180
LR90	0.867174	0.077165	11.23798	0.0000
AEROSP	−0.874436	0.364497	−2.399022	0.0164
CHEMIST	0.666191	0.256457	2.597676	0.0094
COMPUTER	−0.132057	0.288837	−0.457203	0.6475
MACHINES	0.008171	0.276199	0.029584	0.9764
VEHICLES	−1.515083	0.371695	−4.076142	0.0000
JAPAN	0.121004	0.414425	0.291981	0.7703
US	−0.691413	0.275377	−2.510791	0.0120

Mixture Parameter

SHAPE:C(10)	0.251920	0.105485	2.388217	0.0169

R-squared	0.440411	Mean dependent var	79.74586
Adjusted R-squared	0.410959	S.D. dependent var	154.2011
S.E. of regression	118.3479	Akaike info criterion	9.341994
Sum squared resid	2395063.	Schwarz criterion	9.518706
Log likelihood	−835.4504	Hannan–Quinn criter.	9.413637
Restr. log likelihood	−15822.38	LR statistic	29973.86
Avg. log likelihood	−4.615748	Prob(LR statistic)	0.000000

NB of the Safety Example

Negative Binomial Regression of Injury Accident Data

Independent Variable	Estimated Parameter	t Statistic
Constant	–0.931	–2.37
Average annual daily traffic on major road	0.0000900	3.47
Average annual daily traffic on minor road	0.000610	3.09
Median width in feet	– 0.0670	–1.99
Number of driveways within 250 ft of intersection	0.0632	2.24
Overdispersion parameter, α	0.516	3.09
Number of observations	84	
Restricted log likelihood (constant term only)	–169.25	
Log likelihood at convergence	–153.28	
Chi-squared (and associated p-value)	31.95 (<0.0000001)	

Implementation in R

Poisson Model

glm(Y ~ X, family = poisson)

Negative Binomial Model

glm.nb(Y ~ X)

Hurdle-Poisson Model

hurdle(Y ~ X| X1, link = “logit”, dist = “poisson”)

hurdle(Y ~ X| X1, link = “logit”, dist = “negbin”)

Zero-Inflated Model

zip(Y ~ X| X1, link = “logit”, dist = “poisson”)

zinb(Y ~ X| X1, link = “logit”, dist = “negbin”)