NATIONAL RESEARCH
UNIVERSITY

Department of Sociology
Laboratory for Comparative Social Research

QUANTITATIVE DATA ANALYSIS

Models for Count Data

Violetta Korsunova

St. Petersburg, 2021

COUNT DATA

- The scale contains natural numbers (positive integers)
- The scale is *numeric* - the numbers are real
- The scale is *descrete* - only integer values are possible, fractional numbers make no sense

Examples:

- Number of children in household
- Number of cigarettes smoked daily
- Number of goals scored in a football match

COUNT DATA: WHAT TO DO?

The scale is *descrete* - OLS regression doesn't work

You can assume a Poisson distribution

POISSON DISTRIBUTION

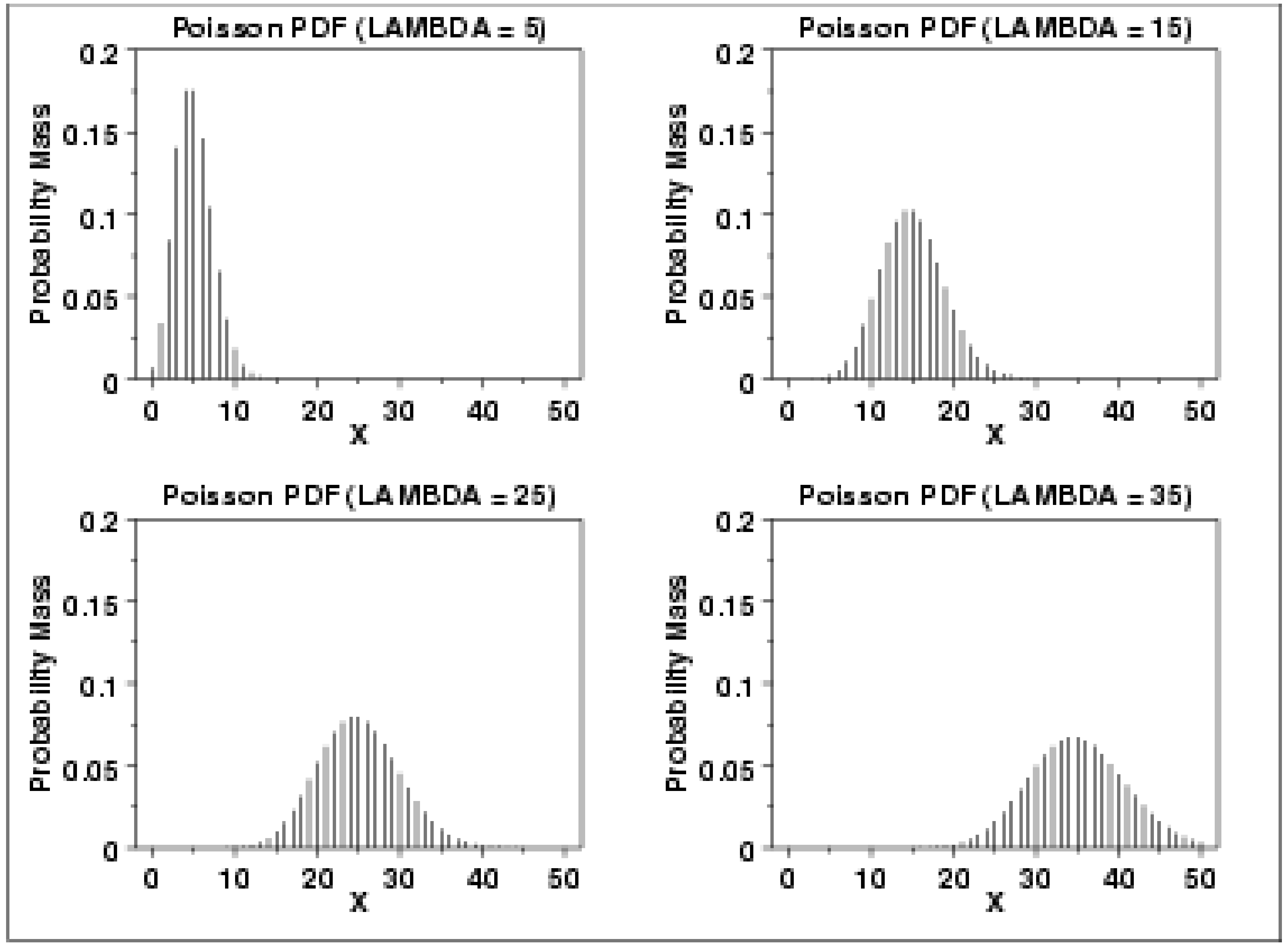
Expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate

The distribution is defined by only one parameter λ:

$$\lambda=\mu=\sigma^2$$

Hence, variance equals mean

POISSON DISTRIBUTION



POISSON REGRESSION

Predicts the probability of a certain number of events (k) to happen:

$$P_{(y=k)} = (\lambda^k/k!)e^{-\lambda}$$

Regression equation is:

$$log(y_i) = \beta_0 + \beta X^T{}_i$$

WHAT TO TAKE INTO ACCOUNT

- Over- or underdispersion: mean must equal variance
- Excessive zeros: ln0 is not defined
- Exposure: the units of observation differ in some dimension (area size, period of observation) and the outcome is proportional to that direction. E.g. salary per month or per week (exposure is time)

OVER- OR UNDERDISPERSSION

Use quazi-Poisson or negative-binomial regression

Qasi-Poisson consider the variance to be a linear function of the mean:

$$\sigma^2 = \mu + \alpha\mu$$

Negative-binomial consider the variance to be a quadratic function of the mean:

$$\sigma^2 = \mu + \alpha\mu^2$$

EXCESSIVE ZEROS

Use a zero-inflated model

Zero-inflated models predict separately the counts and the probability of getting zero

EXPOSURE

Specify the offcet in your model (e.g. time period, area size, population)

The offset variable should not contain zero as ln0 is non-defined

EXAMPLE IN R

Let’s use “bioChemists” data from package “pscl” on article production by graduate students in biochemistry Ph.D. programs

Variables:

art - N of articles produced during last 3 years of Ph.D.

fem - factor indicating gender of student

mar - factor indicating marital status of a student

kid5 - number of children aged 5 or younger

phd - prestige of Ph.D. department

ment - N of articles produced by PhD mentor during last 3 years

```
> library(pscl)
> data("bioChemists")
> head(bioChemists)
  art   fem     mar kid5  phd ment
1   0   Men Married    0 2.52    7
2   0 Women  Single    0 2.05    6
3   0 Women  Single    0 3.75    6
4   0   Men Married    1 1.18    3
5   0 Women  Single    0 3.75   26
6   0 Women Married    2 3.59    2
```

```
> mpois = glm(art~fem+mar+kid5+phd+ment, data = bioChemists, family = poisson)
> summary(mpois)

Call:
glm(formula = art ~ fem + mar + kid5 + phd + ment, family = poisson,
    data = bioChemists)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5672  -1.5398  -0.3660   0.5722   5.4467

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.304617   0.102981   2.958   0.0031 **
femWomen    -0.224594   0.054613  -4.112 3.92e-05 ***
marMarried   0.155243   0.061374   2.529   0.0114 *
kid5        -0.184883   0.040127  -4.607 4.08e-06 ***
phd          0.012823   0.026397   0.486   0.6271
ment         0.025543   0.002006  12.733  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1817.4  on 914  degrees of freedom
Residual deviance: 1634.4  on 909  degrees of freedom
AIC: 3314.1

Number of Fisher Scoring iterations: 5
```

```
> library(AER)
> dispersiontest(mpois, trafo = 1) #Checks linear relationship

	Overdispersion test

data:  mpois
z = 5.7825, p-value = 3.681e-09 #Mean and variance are not equal
alternative hypothesis: true alpha is greater than 0
sample estimates:
    alpha
0.8245398 #Overdisperssion is present

> dispersiontest(mpois, trafo = 2) #Checks quadratic relationship

	Overdispersion test

data:  mpois
z = 6.5297, p-value = 3.295e-11 #It is quadratic
alternative hypothesis: true alpha is greater than 0
sample estimates:
    alpha
0.5091216 #NB is more sutable
```

```
> fm_qpois <- glm(art ~ fem + mar + kid5 + phd + ment, data = bioChemists, family = quasipoisson)
> fm_nb <- MASS::glm.nb(art ~ fem + mar + kid5 + phd + ment, data = bioChemists)
> library(texreg)
> screenreg(list(mpois,fm_qpois,fm_nb))
=======================================================
                 Model 1       Model 2      Model 3
-------------------------------------------------------
(Intercept)         0.30 **       0.30 *       0.26
                   (0.10)        (0.14)       (0.14)
femWomen           -0.22 ***     -0.22 **     -0.22 **
                   (0.05)        (0.07)       (0.07)
marMarried          0.16 *        0.16         0.15
                   (0.06)        (0.08)       (0.08)
kid5               -0.18 ***     -0.18 ***    -0.18 ***
                   (0.04)        (0.05)       (0.05)
phd                 0.01          0.01         0.02
                   (0.03)        (0.04)       (0.04)
ment                0.03 ***      0.03 ***     0.03 ***
                   (0.00)        (0.00)       (0.00)
-------------------------------------------------------
AIC              3314.11                    3135.92
BIC              3343.03                    3169.65
Log Likelihood  -1651.06                   -1560.96
Deviance         1634.37       1634.37      1004.28
Num. obs.         915           915          915
=======================================================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

```
> library(vcdExtra)
> zero.test(table(bioChemists$art))
Score test for zero inflation

  Chi-square = 133.91825
  df = 1
  pvalue: < 2.22e-16 #too many zeros, use ZIM
```

```
> fm_zinb1 <- zeroinfl(art ~ fem + mar + kid5 + phd + ment|1, data = bioChemists, dist = "negbin")
> summary(fm_zinb1)

Call:
zeroinfl(formula = art ~ fem + mar + kid5 + phd + ment | 1, data = bioChemists, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max 
-1.2677 -0.8755 -0.2612  0.4984  6.6572 

Count model coefficients (negbin with log link):
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  0.25620    0.13856   1.849 0.064456 .  
femWomen    -0.21642    0.07267  -2.978 0.002901 ** 
marMarried   0.15047    0.08211   1.833 0.066853 .  
kid5        -0.17641    0.05306  -3.325 0.000885 ***
phd          0.01525    0.03604   0.423 0.672122    
ment         0.02908    0.00347   8.381  < 2e-16 ***
Log(theta)   0.81734    0.11994   6.814 9.46e-12 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -12.85      88.15  -0.146    0.884
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Theta = 2.2645 
Number of iterations in BFGS optimization: 35 
Log-likelihood: -1561 on 8 Df
```

```
> fm_zinb2 <- zeroinfl(art ~ fem + mar + kid5 + phd + ment, data = bioChemists, dist = "negbin")
> summary(fm_zinb2)
Call:
zeroinfl(formula = art ~ fem + mar + kid5 + phd + ment, data = bioChemists, dist = "negbin")
Pearson residuals:
    Min      1Q  Median      3Q     Max 
-1.2942 -0.7601 -0.2909  0.4448  6.4155 
Count model coefficients (negbin with log link):
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  0.4167466  0.1435964   2.902  0.00371 ** 
femWomen    -0.1955076  0.0755926  -2.586  0.00970 ** 
marMarried   0.0975826  0.0844520   1.155  0.24789    
kid5        -0.1517321  0.0542061  -2.799  0.00512 ** 
phd         -0.0006998  0.0362697  -0.019  0.98461    
ment         0.0247862  0.0034927   7.097 1.28e-12 ***
Log(theta)   0.9763577  0.1354696   7.207 5.71e-13 ***
Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)   
(Intercept) -0.19161    1.32280  -0.145  0.88483   
femWomen     0.63587    0.84890   0.749  0.45382   
marMarried  -1.49944    0.93866  -1.597  0.11017   
kid5         0.62841    0.44277   1.419  0.15583   
phd         -0.03773    0.30801  -0.123  0.90250   
ment        -0.88227    0.31622  -2.790  0.00527 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
Theta = 2.6548 
Number of iterations in BFGS optimization: 27 
Log-likelihood: -1550 on 13 Df
```

```
> fm_zinb3<- zeroinfl(art ~ fem + mar + kid5 + phd + ment|ment, data = bioChemists, dist = "negbin")
> summary(fm_zinb3)

Call:
zeroinfl(formula = art ~ fem + mar + kid5 + phd + ment | ment, data = bioChemists, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max 
-1.3041 -0.7684 -0.2632  0.4670  6.3764 

Count model coefficients (negbin with log link):
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  0.404026   0.141716   2.851  0.00436 ** 
femWomen    -0.211893   0.071923  -2.946  0.00322 ** 
marMarried   0.139468   0.081193   1.718  0.08584 .  
kid5        -0.167637   0.052458  -3.196  0.00140 ** 
phd          0.001955   0.035587   0.055  0.95618    
ment         0.024393   0.003518   6.934  4.1e-12 ***
Log(theta)   1.002834   0.142824   7.021  2.2e-12 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)  
(Intercept)  -0.8063     0.3531  -2.283   0.0224 *
ment         -0.6098     0.2459  -2.480   0.0132 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
Theta = 2.726 
Number of iterations in BFGS optimization: 20 
Log-likelihood: -1553 on 9 Df
```

```
> exp(coef(fm_zinb3))
count_(Intercept)    count_femWomen  count_marMarried        count_kid5
        1.4978426         0.8090515         1.1496624         0.8456605
        count_phd        count_ment  zero_(Intercept)         zero_ment
        1.0019574         1.0246933         0.4465119         0.5434470
```

```
> pR2(mpois)
fitting null model for pseudo-r2
          llh       llhNull            G2      McFadden          r2ML          r2CU
-1.651056e+03 -1.742573e+03  1.830343e+02  5.251839e-02  1.813000e-01  1.854110e-01

> pR2(fm_zinb3)
fitting null model for pseudo-r2
          llh       llhNull            G2      McFadden          r2ML          r2CU
-1.553271e+03 -1.609937e+03  1.133320e+02  3.519764e-02  1.164966e-01  1.200537e-01
```

INTERPRETATION

For log:
The log of the number of articles produced by women is 0.21 lower compared to men
With every kid under 5 yo the log of the number of articles decreases by 0.17
Every additional article produced by the mentor increases the log of the number of articles by 0.02
Every additional article produced by the mentor increases the log of odds of not producing any articles by 0.61

For exp:
Women have 20% fewer articles compared to men ((0.8-1)*100 ≈ -20%)
Every kid under 5 yo decreases the number of articles by 15% ((0.85-1)*100 ≈ -15%)
Every additional article produced by the mentor increases the number of articles by 2%
Every additional article produced by the mentor decreases the odds of not producing any articles by 1.85

EXPOSURE

Let’s use “Insurance” data from package “MASS” on the numbers of car insurance claims made by the policyholders.

Variables:

- District – factor: district of residence of policyholder (1 to 4): 4 is major cities.
- Group – an ordered factor: group of car with levels <1 litre, 1–1.5 litre, 1.5–2 litre, >2 litre.
- Age – an ordered factor: the age of the insured in 4 groups labelled <25, 25–29, 30–35, >35.
- Holders – numbers of policyholders.
- Claims – numbers of claims

```
> data("Insurance")
> head(Insurance)
  District  Group    Age Holders Claims
1        1    <1l    <25     197     38
2        1    <1l  25-29     264     35
3        1    <1l  30-35     246     20
4        1    <1l    >35    1680    156
5        1 1-1.5l    <25     284     63
6        1 1-1.5l  25-29     536     84
```

```
> mod1 = glm(Claims ~ District + Group + Age, family=poisson, data=Insurance)
> mod2 = glm(Claims ~ District + Group + Age + offset(log(Holders)), family=poisson,data=Insurance)
> screenreg(list(mod1,mod2))
========================================
                 Model 1      Model 2
----------------------------------------
(Intercept)        3.92 ***    -1.81 ***
                  (0.03)       (0.03)
District2         -0.44 ***     0.03
                  (0.04)       (0.04)
District3         -0.92 ***     0.04
                  (0.05)       (0.05)
District4         -1.44 ***     0.23 ***
                  (0.06)       (0.06)
Group.L           -0.51 ***     0.43 ***
                  (0.05)       (0.05)
Group.Q           -1.02 ***     0.00
                  (0.04)       (0.04)
Group.C            0.22 ***    -0.03
                  (0.03)       (0.03)
Age.L              1.50 ***    -0.39 ***
                  (0.05)       (0.05)
Age.Q              0.47 ***    -0.00
                  (0.05)       (0.05)
Age.C              0.41 ***    -0.02
                  (0.05)       (0.05)
----------------------------------------
AIC              458.63       388.74
BIC              480.22       410.33
Log Likelihood  -219.32      -184.37
Deviance         121.31        51.42
Num. obs.         64           64
========================================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

STEPS TO FOLLOW

1. Check if you need any offsets
2. Estimate a poisson model
3. Check for overdisperssion
4. Check for excessive zeroes
5. Choose the appropriate model and estimate it
6. Interpret the results