Poisson Regression

Hynek Pikhart

Learning objectives

- Understand the Poisson distribution
- Understand and use Poisson regression
- Know how to test the appropriateness of Poisson regression
- Understand and use zero-inflated Poisson regression

Dependent variable

- Is a count variable
  - Indicates how many times something has happened
  - a non-negative integer
- For example

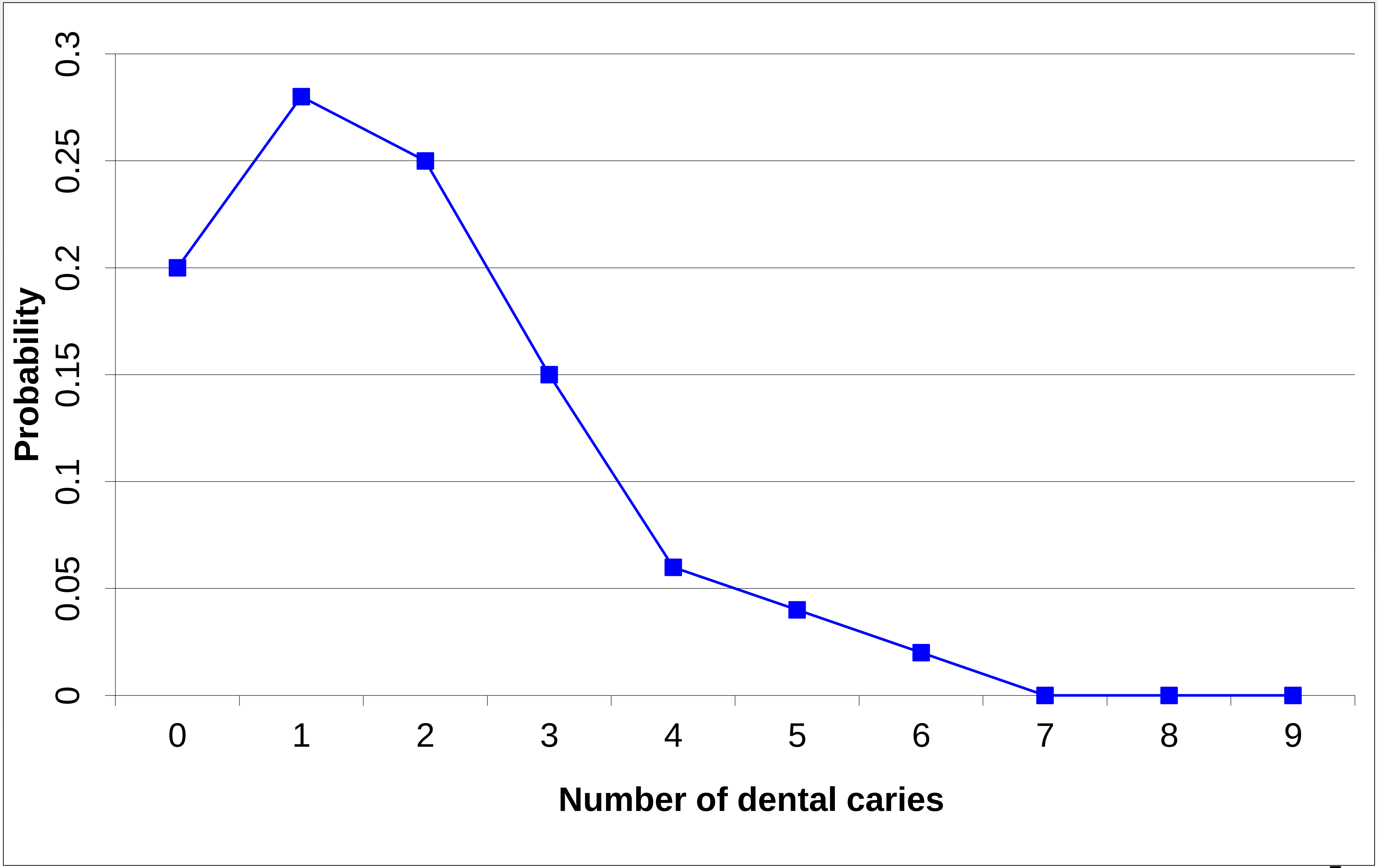
  -Number of patients

  -Number of dental caries

  -Event free days after hospital discharge

An example

caries	Freq.	Percent	Cum.
0	20	20.00	20.00
1	28	28.00	48.00
2	25	25.00	73.00
3	15	15.00	88.00
4	6	6.00	94.00
5	4	4.00	98.00
6	2	2.00	100.00
Total	100	100.00	


Probability
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1
2
3
4
5
6
7
8
9
Number of dental caries

Poisson Distribution

- Let *Y* be a random variable that indicates the number of times (count) a certain event occurs
- Let *μ* be the expected count
  - Often termed "rate" if thinking in terms of a count per unit time or unit of space
- The Poisson distribution specifies the relationship between the expected count *μ* and the probability of observing any observed count *y* as

$$\Pr(Y = y \mid \mu) = \frac{e^{-\mu} \mu^{y}}{y!}$$

Poisson Distribution

- μ is the mean value of the random variable Y

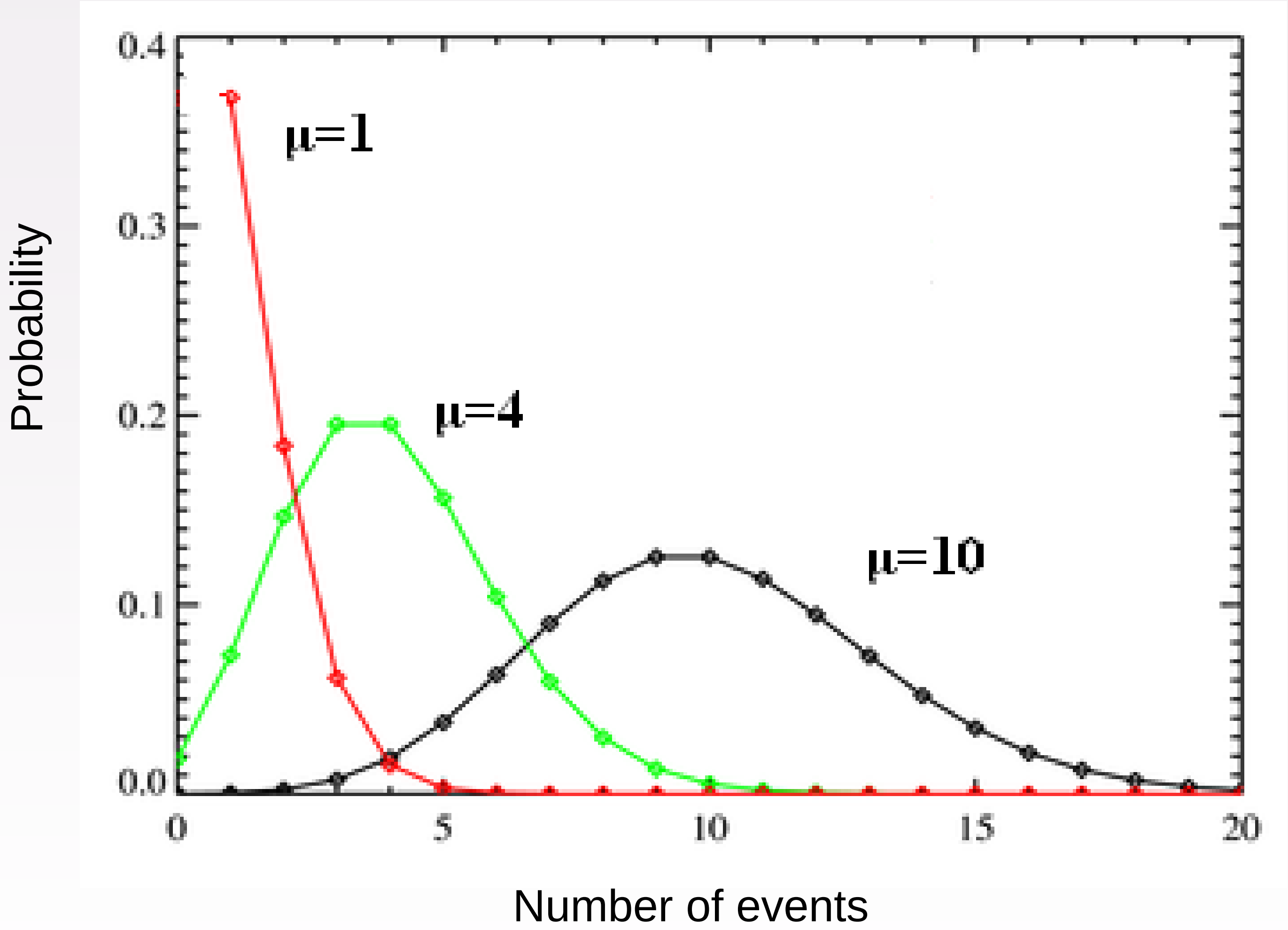
  $E(Y) = \mu$

  μ is also the variance of the distribution, thus

  $Var(Y) = \mu$

  - This is known as *“equidispersion”*

- The mean entirely fixes the shape of the distribution
  - when the mean is small the most common count is predicted to be zero
- Larger values of μ will produce greater probabilities of non-zeros values
  - therefore the probability of a zero count decreases
- For large μ the Poisson distribution is well approximated by the normal distribution



For μ = 1 the prob of obtaining a count = 1 is 0.36

For μ = 4 the prob of obtaining a count = 4 is 0.20

For μ = 10 the prob of obtaining a count = 10 is 0.12

Poisson Regression

- A regression model can be obtained by letting *μ* depend on independent variables **X**
- In order to have a positive mean we set

$$\mu_i = \exp(\beta_0 + \beta_0 x_{i1} + \beta_0 x_{i1} + \cdots + \beta_p x_{ip})$$

i.e. the link function is the logarithm (inverse of exp).

$\beta_0 \ldots \beta_p$ *are* the regression coefficients to be estimated

- The model is called log-linear, since

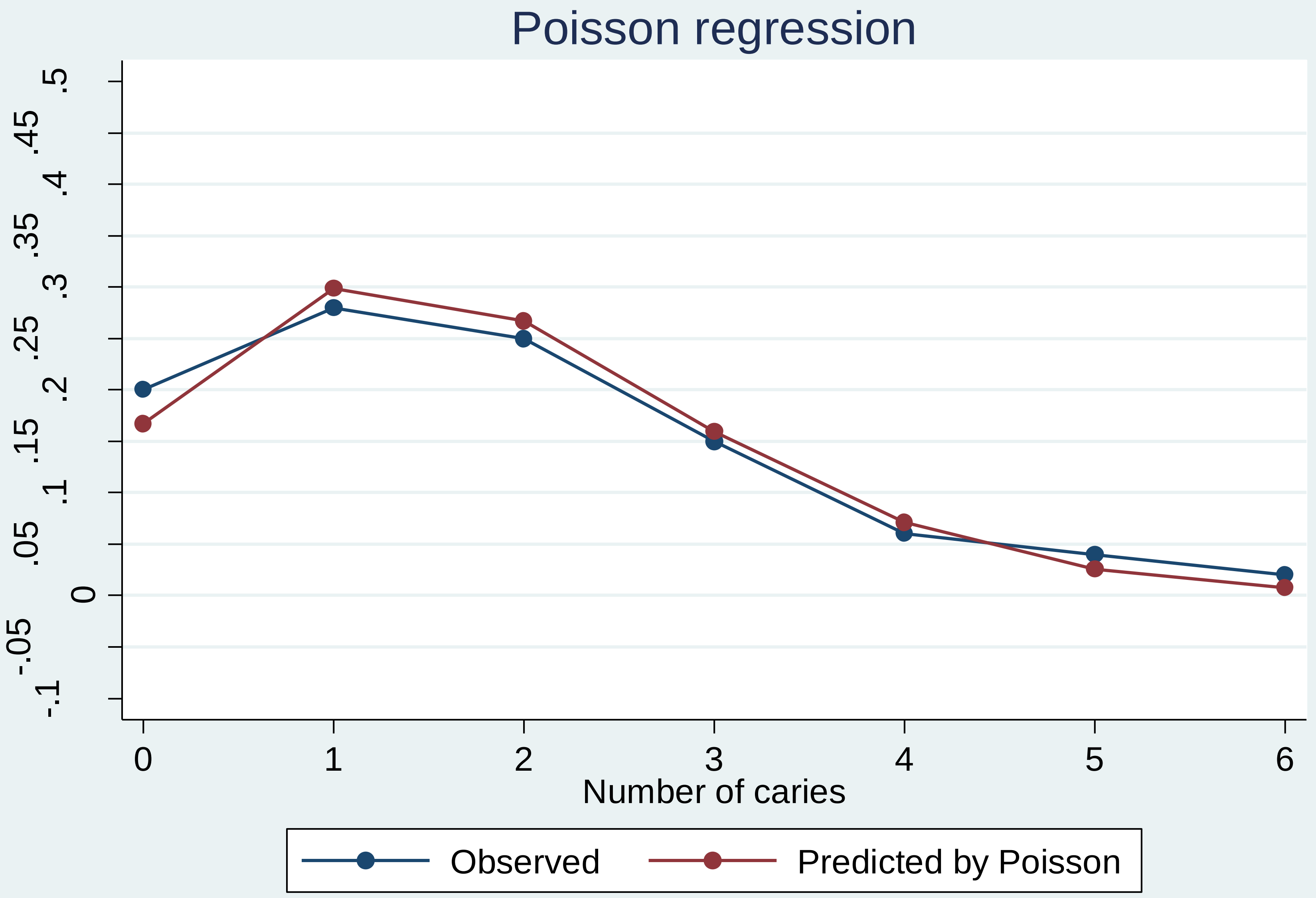
  $\log(\mu_i) = \beta_0 + \beta_1\, x_{i1} + \beta_2\, x_{i2} + \ldots + \beta_p\, x_{ip}$

Assumptions of Poisson regression

- The dependent variable is a count
- The dependent variable is not over-dispersed (equidispersion holds) and does not have an excessive number of zeros
- If thinking in terms of “rates”, each subject has the same unit of time or space
  - if not the Poisson model needs to be adjusted to account for time or size per subject – see later

A graphical comparison of observed vs predicted

- We can compare graphically observed probabilities for each value of the Y variable with predicted probabilities from fitting the Poisson distribution (the empty regression); we do so as a check for whether Poisson could be an adequate regression for our outcome.
  - In Stata do this using **`prcounts`**


Poisson regression
.5
.45
.4
.35
.3
.25
.2
.15
.1
.05
0
-.05
-.1
0
1
2
3
4
5
6
Number of caries
Observed
Predicted by Poisson

Example Poisson regression in Stata

- We look at the relationship between dental caries and age and sex

poisson caries age sex (or **xi: poisson caries age i.sex**)

```
Iteration 0:   log likelihood =  -160.9713
Iteration 1:   log likelihood =  -160.9713

Poisson regression                           Number of obs   =        100
                                             LR chi2(2)      =      18.79
                                             Prob > chi2     =     0.0001
Log likelihood =  -160.9713                  Pseudo R2       =     0.0551

------------------------------------------------------------------------------
      caries |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0810846   .0191089     4.24   0.000     .0436318    .1185374
         sex |   .0277315   .1494894     0.19   0.853    -.2652623    .3207252
       _cons |  -.1129303   .2025551    -0.56   0.577     -.509931    .2840704
------------------------------------------------------------------------------
```

How to interpret the results

- **Iteration Log** - This is a listing of the log likelihood at each iteration. Poisson regression uses ML estimation, and because of the form of the Poisson distribution this has to be done iteratively.
- **Log Likelihood** - This is the log likelihood of the fitted model. It is used in the calculation of the Likelihood Ratio (LR) chi-squared test of whether all independent variables' regression coefficients are simultaneously zero and in tests of nested models.
- **LR chi^2(2)**- This is the LR test statistic for the null hypothesis that the regression coefficients for independent variables are equal to zero. The degrees of freedom (the number in parenthesis) of the LR test statistic is defined by the number of independent variables (2).
- **Prob > chi^2** - This is the probability of obtaining this chi-square test statistic (18.79) if there is in fact the independent variables have no effect. The small p-value from the LR test, $p < 0.001$, leads us to conclude that at least one of the regression coefficients in the model is not equal to zero.

Poisson regression coefficients

- Poisson regression coefficients are interpreted as the difference between the log of expected counts,
- $\beta = \log(\mu_{x+1}) - \log(\mu_x)$

- where β is the regression coefficient corresponding to a variable x, μ is the expected count and the subscripts represent evaluation at the independent variable value of x or x+1 (implying a one unit change in the independent variable x) with all other independent variables being held constant

Poisson regression coefficients

- For a one unit change in the independent variable, the difference in the log of expected count will change by the respective regression coefficient, holding the other variables in the model constant.
- If age increases by one year, the difference in the log of expected count would be 0.081, while holding sex constant.
- The difference in the log of expected count is 0.277 for females compared to males, while holding age constant.

Significance of parameter estimates

- **z** and **P>|z|** - These are the test statistic and p-value, respectively, that test the null hypothesis that an individual regression coefficient is zero given the rest of the coefficients
- The test statistic **z** is the ratio of the **Coef.** to the **Std. Err.** of the respective coefficient.
- In our model **sex** is not significant because its corresponding coefficient has a large (> 0.05) **z** statistic

Using the Incidence Rate Ratio (IRR)

```
poisson caries age sex, irr

Iteration 0:   log likelihood =  -160.9713
Iteration 1:   log likelihood =  -160.9713

Poisson regression                          Number of obs   =        100
                                            LR chi2(2)      =      18.79
                                            Prob > chi2     =     0.0001
Log likelihood =  -160.9713                 Pseudo R2       =     0.0551

------------------------------------------------------------------------------
      caries |      IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  1.084463   .0207229    4.24   0.000     1.044598   1.125849
         sex |   1.02812   .1536929    0.19   0.853     .7670047   1.378127
------------------------------------------------------------------------------
```

How to interpret the IRR

- 1.08 is the estimated rate ratio for a one year increase in age: for each year increase in age, the number of dental caries is expected to increase by a factor of 1.08, while holding sex constant
- 1.02 is the estimated rate ratio comparing females to males: females are expected to have a 1.02 times greater number of dental caries (note: the effect is not significant)

Please Note

- Each subject in our sample was assumed to be followed for one unit of time (a year).
- If this were not the case and we were to neglect the exposure time, our Poisson regression estimates would be biased, since our model assumes all subjects had the same follow up time
  - we want to be sure that we account for the fact that children will have a larger count of dental caries if we observe them for longer

Exposure time

- Different exposure times can be easily incorporated into the model (or exposure space)

- In Stata we use the option **exposure(*varname*)**, where ***varname*** corresponds to the exposure of an individual to adjust the Poisson regression estimates.

poisson caries age sex, exposure(time) irr

```
Iteration 0:   log likelihood = -178.57015
Iteration 1:   log likelihood = -178.57015

Poisson regression                          Number of obs   =       100
                                            LR chi2(2)      =     25.57
                                            Prob > chi2     =    0.0000
Log likelihood = -178.57015                 Pseudo R2       =    0.0668

------------------------------------------------------------------------------
      caries |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.096587   .0206417     4.90   0.000     1.056867    1.137799
         sex |   1.125068   .1682197     0.79   0.431      .839281    1.508169
        time | (exposure)
------------------------------------------------------------------------------
```

Post-estimation

- It is possible to perform a goodness of fit test of the model
- We use the χ^2 test to test the null hypothesis that the data are Poisson distributed, conditional on the independent variables
- If $p<0.05$ we reject the null hypothesis, therefore the Poisson regression model is inappropriate

poisson caries age sex ,irr

```
Iteration 0:   log likelihood =  -160.9713
Iteration 1:   log likelihood =  -160.9713

Poisson regression                                Number of obs   =        100
                                                  LR chi2(2)      =      18.79
                                                  Prob > chi2     =     0.0001
Log likelihood =  -160.9713                       Pseudo R2       =     0.0551

------------------------------------------------------------------------------
      caries |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.084463   .0207229     4.24   0.000     1.044598    1.125849
         sex |    1.02812   .1536929     0.19   0.853     .7670047    1.378127
------------------------------------------------------------------------------
```

estat gof

```
         Goodness-of-fit chi2  =  114.8906
         Prob > chi2(97)       =    0.1038
```

What to do if Poisson is inappropriate

- If we have over-dispersion (the count variable has a variance greater than the mean):

  - use Negative Binomial Regression

- Or if the count variable has a greater number of zero observations than expected from the Poisson model

  - use Zero Inflated Poisson regression...

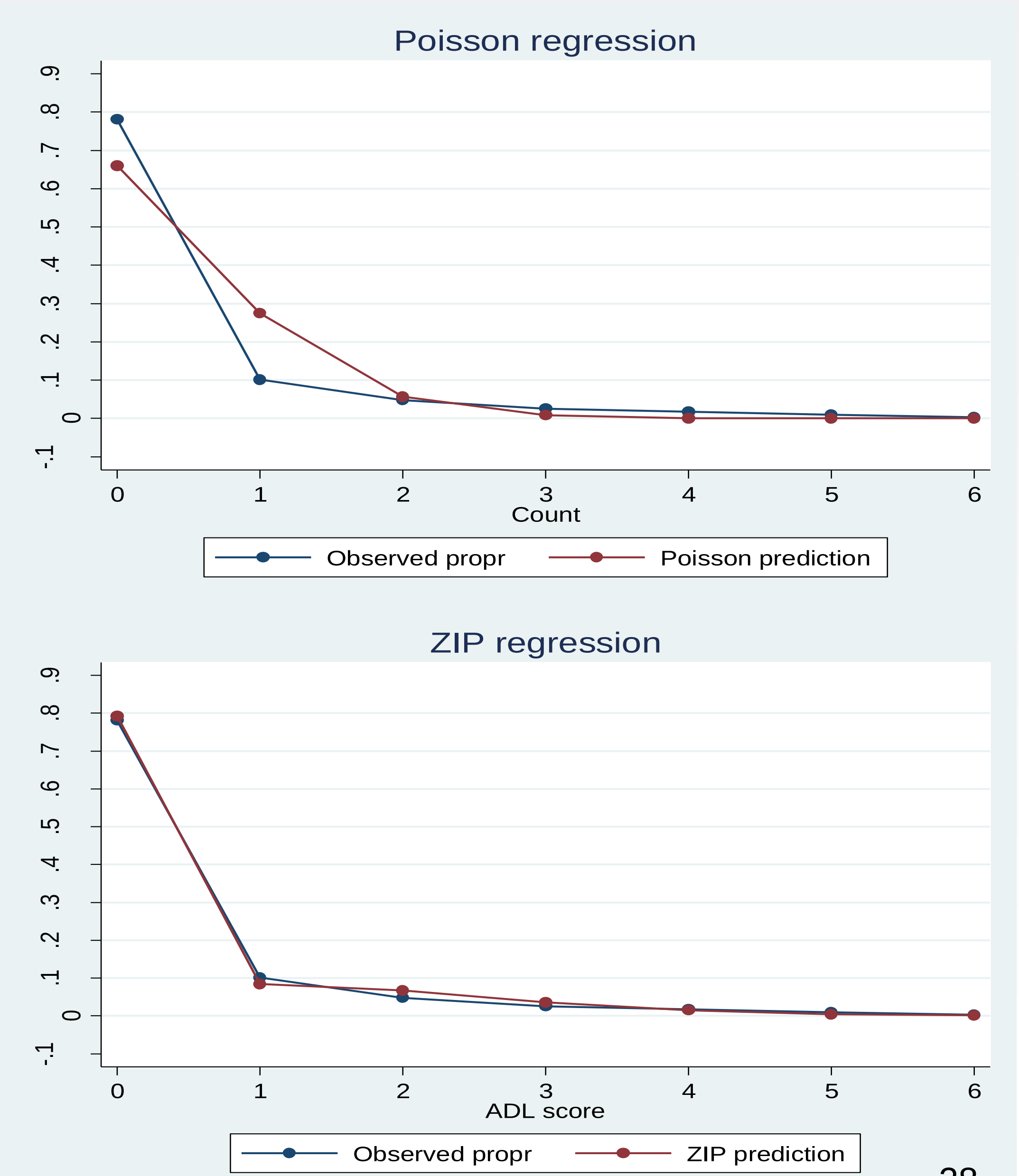
Zero-Inflated Poisson (ZIP)

- ZIP assumes that there are two latent groups: the group of zeros and the group of non-zeros.
- The first group generates only zeros the second is a Poisson

$$\Pr(Y_i = y_i \mid X_i) = \begin{cases} p_i + (1 - p_i)e^{-\mu_i} & y_i = 0 \\ \dfrac{(1 - p_i)e^{-\mu_i}\mu_i^{y_i}}{y_i!} & y_i \geq 1 \end{cases}$$

An example

Number of difficulties with ADL	
0	79.2%
1	10.3%
2	4.9%
3	2.6%
4	1.7%
5	0.9%
6	0.4%



```
. zip adl indager indsex ed2, inf (indager indsex ed2) vuong nolog irr

Zero-inflated Poisson regression                 Number of obs   =     11201
                                                  Nonzero obs    =      2327
                                                    Zero obs     =      8874

Inflation model = logit                           LR chi2(3)     =     11.20
Log likelihood  = -8684.432                       Prob > chi2    =    0.0107

------------------------------------------------------------------------------
         adl |        IRR    Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
adl          |
     indager |   .9966642   .0016994    -1.96   0.050      .993339    1.000001
      indsex |   1.009251   .0388564     0.24   0.811     .9358964    1.088355
         ed2 |   1.126005   .0444662     3.01   0.003     1.042141    1.216619
-------------+----------------------------------------------------------------
inflate      |
     indager |  -.0514357   .0028253   -18.21   0.000    -.0569732   -.0458981
      indsex |   .0283517    .057836     0.49   0.624    -.0850048    .1417081
         ed2 |  -.5104743   .0584694    -8.73   0.000    -.6250722   -.3958765
       _cons |   4.657592   .2038718    22.85   0.000      4.25801    5.057173
------------------------------------------------------------------------------
Vuong test of zip vs. standard Poisson:           z =    21.91  Pr>z = 0.0000
```

Female vs male

Having no formal education vs having formal education

```
. zip adl indager indsex ed2, inf (indager indsex ed2) vuong nolog irr

Zero-inflated Poisson regression              Number of obs    =     11201
                                               Nonzero obs     =      2327
                                                 Zero obs      =      8874

Inflation model = logit                         LR chi2(3)     =     11.20
Log likelihood  = -8684.432                     Prob > chi2    =    0.0107

------------------------------------------------------------------------------
         adl |       IRR    Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
adl          |
     indager |  .9966642   .0016994    -1.96   0.050      .993339    1.000001
      indsex |  1.009251   .0388564     0.24   0.811     .9358964    1.088355
         ed2 |  1.126005   .0444662     3.01   0.003     1.042141    1.216619
-------------+----------------------------------------------------------------
inflate      |
     indager | -.0514357   .0028253   -18.21   0.000    -.0569732   -.0458981
      indsex |  .0283517    .057836     0.49   0.624    -.0850048    .1417081
         ed2 | -.5104743   .0584694    -8.73   0.000    -.6250722   -.3958765
       _cons |  4.657592   .2038718    22.85   0.000      4.25801    5.057173
------------------------------------------------------------------------------
Vuong test of zip vs. standard Poisson:            z =    21.91  Pr>z = 0.0000
```

Poisson part

Logit part

How to interpret the results

- Poisson part:

  -increasing age and female sex are not significantly related to higher number of difficulties with ADL

  -not having an educational qualification is related with a higher number of difficulties with ADL. People with no qualification are expected to have 1.13 times greater number of difficulties with ADL than those with an educational qualification
- Logit part:

  -Increasing age and not having an education qualification are related to lower chances of being in the zero difficulties with ADL by default group. The coefficients can be transformed into odds ratios.
- Vuong test:

  -To test whether the ZIP model is better than the Poisson. As the p-value is <0.0001, the ZIP fits the data better than the Poisson

Suggested reading

- Agresti A. *An introduction to categorical data analysis*. New York; Chichester: Wiley 1996.
- Long JS. *Regression models for categorical and limited dependent variables*. Thousand Oaks; London: Sage 1997.
- Long JS., Freese, J., & Stata Corporation. *Regression models for categorical dependent variables using Stata*. College Station, Texas : Stata Corporation 2003.
- Zaninotto P & Falaschetti E (2010) Comparisons of methods for dealing with a count outcome: an application to Activities of Daily Living (ADL-s). *JECH theory and methods doi:10.1136/jech.2008.079640*