BIOS 312: MODERN REGRESSION ANALYSIS

James C (Chris) Slaughter
Department of Biostatistics
Vanderbilt University School of Medicine
james.c.slaughter@vanderbilt.edu

biostat.mc.vanderbilt.edu/CourseBios312

Updated March 8, 2012

Contents

Chapter 10

Simple Poisson Regression

10.1 Count Data and Event Rates

- Sometimes a random variable measures the number of events occurring over some space and time interval

- Examples include
  - Number of polyps recurring in the three year interval between colonoscopies
  - Number of pulmonary exacerbations experienced by a cystic fibrosis patient in a year
  - Number of reflux events in a 24-hour period

- Count data have (in theory) no upper limit, although very large counts can be highly improbable

- When a response variable measures counts over space and time, we often summarize by considering the event rate

- "Event rate" is the expected number of events per unit of space-time

- The rate is thus a mean count

- In most statistical problems, we know the interval of time and the volume of space sampled
  - Poisson models allow us to take into account the known interval of time/space using an "offset"

10.2 Poisson Model

10.2.1 Poisson distribution

- Often we assume that counts follow a Poisson distribution

- The Poisson distribution can be derived from the following assumptions
  - The expected number of events in an interval is proportional to the size of the interval

  - The probability that two events occur with an infinitesimally small interval of space-time is zero

  - The number of events occurring in disjoint (separate) intervals of space-time are independent

- (Note that the assumption of a constant rate with independence over space-time is pretty strong and rarely holds completely)

- Poisson distribution

- Counts the events occurring at a constant rate λ in a specified time (and space) t
  * Independent intervals of time and space

- Probability distribution has parameter $\lambda > 0$
  * For $k = 0, 1, 2, \ldots$

$$\Pr(Y = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} \tag{10.1}$$

  * Mean: $E[Y] = \lambda t$

  * Var: $V[Y] = \lambda t$

  * (Mean-variance relationship, like binary data)

10.2.2 Regression Model

- When the response variable represent counts of some event, we usually model using the (log) rate with Poisson regression
  - Compares rates of response per space-time (e.g. person-years) across groups

  - "Rate ratio"

- Why not use linear regression? The reasons are primarily statistical
  - The rate is in fact a mean

  - For Poisson Y having event rate λ measured over time t
    * The mean is equal to the variance (both are λt)

  - We want to be able to account for

      * Different areas of space or length of time for measuring counts
      * Mean-variance relationship (if not using robust standard errors)

- In Poisson regression, we tend to use a log link when modeling the event rate
  - As in other models, a log link means that we are assuming a multiplicative modeling
    * Multiplicative model $\rightarrow$ comparisons between groups based on ratios
    * Additive model $\rightarrow$ comparisons between groups based on differences
  - Log link also has the best technical statistical properties
    * Log rate is the "canonical parameter" for the Poisson distribution
    * Being the canonical parameter makes the calculus and mathematical properties easier to derive, and thus easier to understand from a theoretical perspective

- Poisson regression
  - Response variable is count of event over space-time (often person-years)
  - Offset variable specifies amount of space-time
  - Allows continuous or multiple grouping variables
    * But will also work with binary grouping variables

- Simple Poisson Regression
  - Modeling rate of count response Y on predictor X

Distribution	$\Pr(Y_i = k \mid T_i = t_i) = \frac{e^{-\lambda_i t_i}(\lambda_i t_i)^k}{k!}$
Model	$\log\ E[Y_i \mid T_i, X_i] = \log(\lambda_i T_i) = \log(T_i) + \beta_0 + \beta_1 \times X_i$
$X_i = 0$	log $\lambda_i = \beta_0$
$X_i = x$	log $\lambda_i = \beta_0 + \beta_1 \times x$
$X_i = x + 1$	log $\lambda_i = \beta_0 + \beta_1 \times x + \beta_1$

– To interpret as rates, exponentiate the parameters

Distribution	$\Pr(Y_i = k \mid T_i = t_i) = \frac{e^{-\lambda_i t_i}(\lambda_i t_i)^k}{k!}$
Model	$\log\ E[Y_i \mid T_i, X_i] = \log(\lambda_i T_i) = \log(T_i) + \beta_0 + \beta_1 \times X_i$
$X_i = 0$	$\lambda_i = e^{\beta_0}$
$X_i = x$	$\lambda_i = e^{\beta_0 + \beta_1 \times x}$
$X_i = x + 1$	$\lambda_i = e^{\beta_0 + \beta_1 \times x + \beta_1}$

- Interpretation of the model
  - Intercept
    - Rate when the predictor is 0 is found by exponentiation of the intercept from Poisson regression: e^{β_0}
  - Slope
    - Rate ratio between groups differing in the value of the predictor by 1 unit is found by exponentiation of the slope from Poisson regression: e^{β_1}

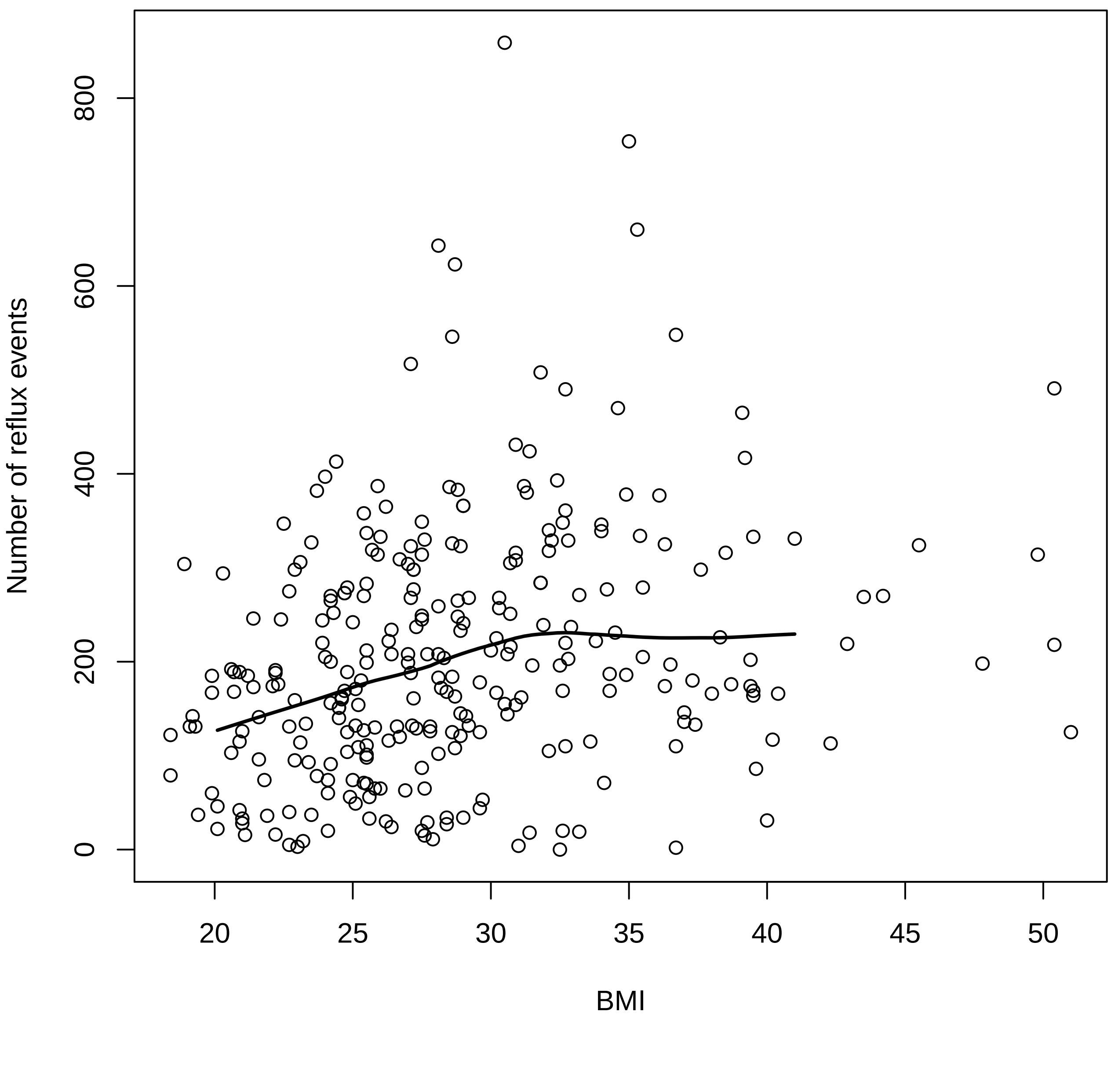
10.3 Example: Acid reflux and BMI

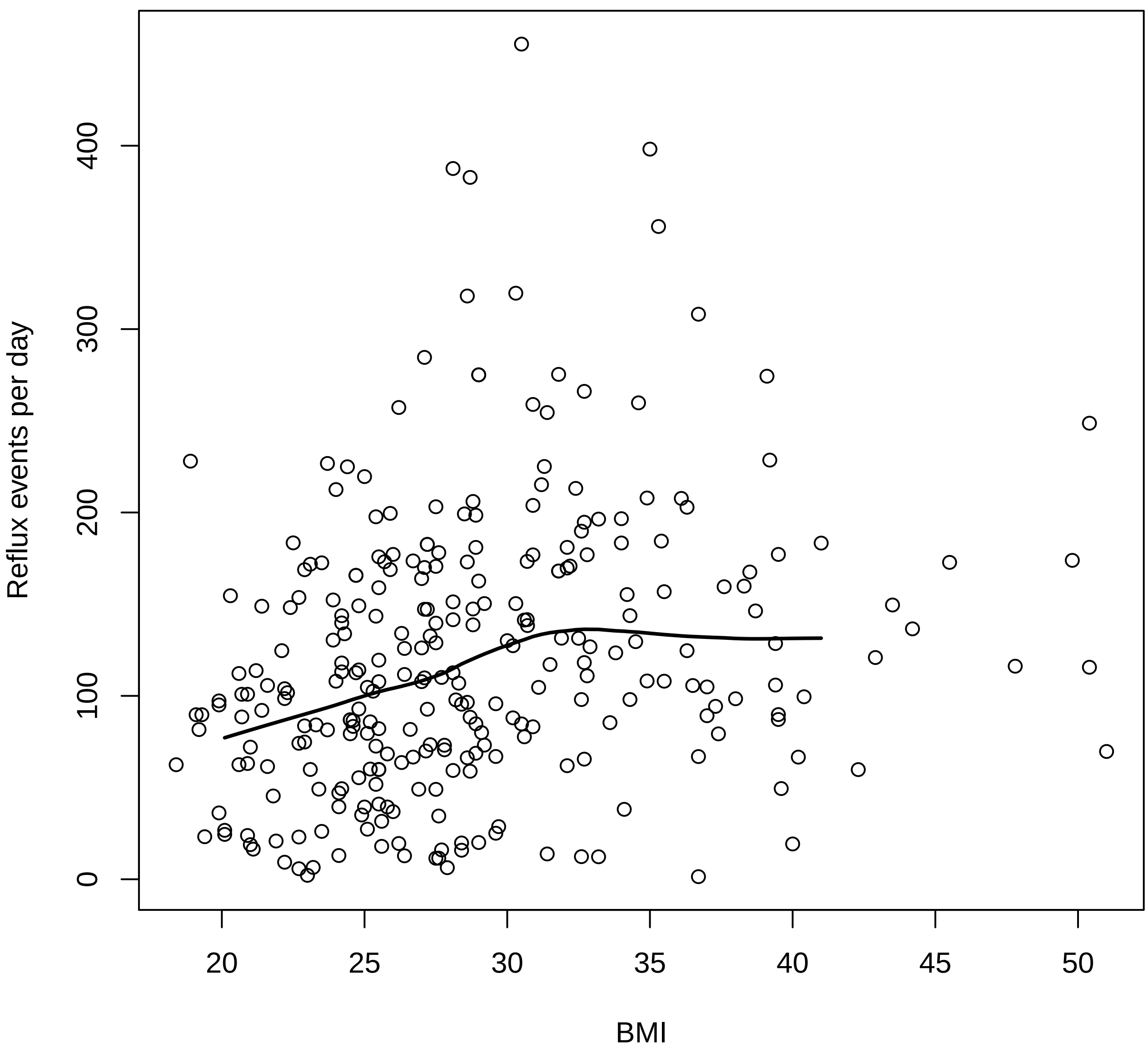
10.3.1 Data description

- Research question: Are the number of acid reflux events in a day related to body mass index (BMI)?

- Each subject pH in the esophagus in monitored continuously for about 24 hours
- Count the number of time pH drop below 4, which is called a “reflux event”
- Analysis (statistical) goals
  - Primary goal: Determine if there is an association between BMI and acid reflux rate
  - Secondary goal: Describe the (mean) trend in reflux rates as a function of BMI
- Variables
  - Response: Number of acid reflux events
  - Offset: Number of minutes subject was monitored
  - Predictor of interest: BMI
  - Other covariates: Presence of esophagitis at baseline

10.3.2 Descriptive Plots


Number of reflux events
0
200
400
600
800
20
25
30
35
40
45
50
BMI



Reflux events per day
0
100
200
300
400
20
25
30
35
40
45
50
BMI

- Characterization of plots
  - Plots are visually similar if we consider the rate (events per day) or the raw number of events
  - First order trend: Event rate increases with increasing BMI
  - Second order trend: Event rate increase until BMI of 32 (or so) and then flattens out
  - Within-group variability
    * Hard to visualize from the plots
    * Model assumes increasing variability with increasing BMI, which looks reasonable

10.3.3 Regression commands

- As before, but need to specify the offset
  - Offset is the log of the exposure time
  - In Stata, can alternatively specify the "exposure" and it will take the log for you
- Stata
  - `poisson respvar predvar, exposure(time) [robust]`
  - `poisson respvar predvar, offset(logtime) [robust]`
- R
  - One method to fit Poisson models

- Uses the `sandwich` and `lmtest` libraries
- Must install the above two libraries using `install.packages("lmtest")` and `install.packages("sandwich")`
- `model.poisson <- glm(response ~ predictors + offset(log(time)), data=data, family="poisson")`
- `coeftest(model.poisson, vcov=sandwich)`

– Another method to fit Poisson models using the `Design` package

- `m1 <- glmD(response ~ predictors + offset(log(time)), data=data, family="poisson", x=TRUE, y=TRUE)`
- `bootcov(m1)` for robust (bootstrap) confidence intervals

– Can also use methods within the `gee` library

10.3.4 Estimation of the regression model

- Regression model for number of reflux events on BMI
  - Answer primary research question: Is there an *association* between BMI and the acid reflux event rate?
  - Estimate the best fitting line to (log) number of reflux events within BMI groups using an offset of log time
    - $\log(\text{Events}|\text{BMI}) = \beta_0 + \beta_1 \times \text{BMI} + \log(\text{time})$
  - An association will exist if the slope β_1 is nonzero

```
. poisson events bmi, offset(logmins) robust

Iteration 0:   log pseudolikelihood =  -11360.89
Iteration 1:   log pseudolikelihood =  -11360.89

Poisson regression                                Number of obs   =        279
                                                  Wald chi2(1)    =      23.42
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood =  -11360.89                 Pseudo R2       =     0.0520

------------------------------------------------------------------------------
             |               Robust
      events |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .0223194   .0046121     4.84   0.000     .0132799    .0313589
       _cons |  -3.119991    .139521   -22.36   0.000    -3.393448   -2.846535
     logmins |   (offset)
------------------------------------------------------------------------------
```

- Interpretation of output
  - $\log \text{rate} = -3.119991 + 0.0223194 \times \text{BMI}$
- Interpretation of intercept
  - Estimated event rate when BMI is 0 is found by exponentiation: $e^{-3.12} = 0.044$
  - This is the rate per 2-minute interval. This unusual time interval is an artifact of the way in pH data is sampled
    * To convert to events per day, multiply by 720 (there are 720 2-minute intervals in a day)
    * $720 \times e^{-3.12} = 31.7 \text{ events per day}$
- Interpretation of slope
  - Estimated ratio of rates for two subjects differing by 1 in their BMI
  - Interpretation by exponentiation of slope
    * A subject with a 1 kg/m^2 higher BMI will have an acid reflux event rate that is 2.3% higher. (calc: $e^{0.0223} = 1.023$)

  * We are 95% confident that the increase in event rate is between 1.3% higher and 3.2% higher

  * There is a significant association between BMI and reflux events $p < 0.001$

10.4 Example: Acid reflux and BMI by esophagitis status

10.4.1 BMI modeled as a linear term

- The following results compare using a Poisson model to a linear regression model

- Both models will control for Esophagitis status, so any interpretation must involve "Holding esophagitis status constant..." ("Among subjects with the same Esophagitis status...")

- Note the different (numerical) estimates for the coefficients and standard errors for BMI and esophagitis, but the similar statistical significance

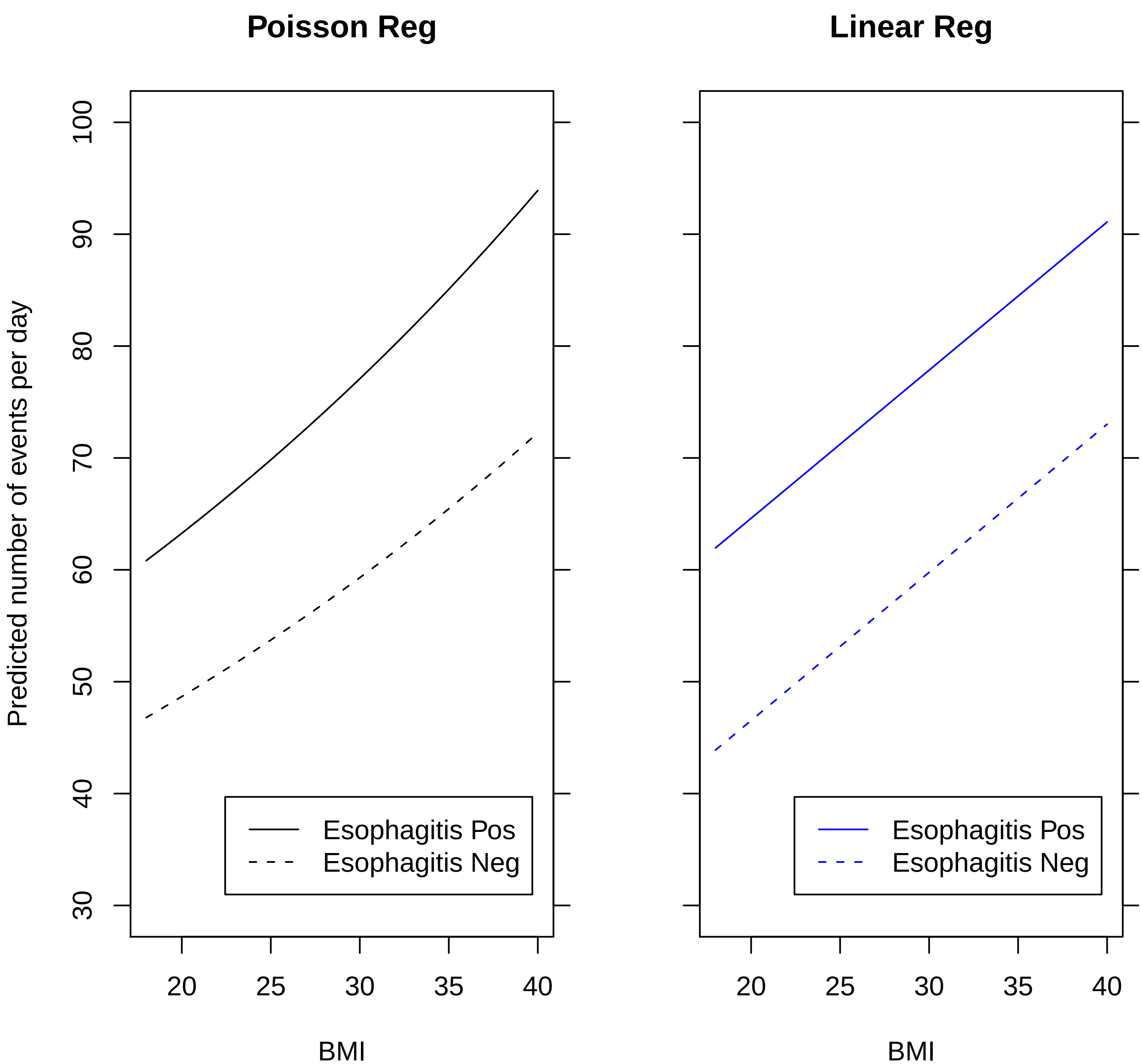
- Also if we plot the predicted number of events per day versus BMI, the results are similar from either model

Stata Output

```
. poisson events bmi esop, offset(logmins) robust

Poisson regression                                Number of obs   =        279
                                                  Wald chi2(2)    =      30.30
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -11072.339                 Pseudo R2       =     0.0761

------------------------------------------------------------------------------
             |               Robust
      events |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .0197465   .0047721     4.14   0.000     .0103934    .0290997
        esop |   .2622171    .083202     3.15   0.002     .0991442      .42529
       _cons |  -3.089033   .1423038   -21.71   0.000    -3.367944   -2.810123
     logmins |   (offset)
------------------------------------------------------------------------------


. gen eventsmins = events / mins
. regress eventsmins bmi esop, robust

Linear regression                                      Number of obs =     279
                                                       F(  2,   276) =   14.16
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0856
                                                       Root MSE      =  .05102

------------------------------------------------------------------------------
             |               Robust
  eventsmins |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |    .001839   .0004618     3.98   0.000     .0009299    .0027482
        esop |    .025104   .0085449     2.94   0.004     .0082826    .0419254
       _cons |   .0278461   .0129053     2.16   0.032     .0024407    .0532515
------------------------------------------------------------------------------
```

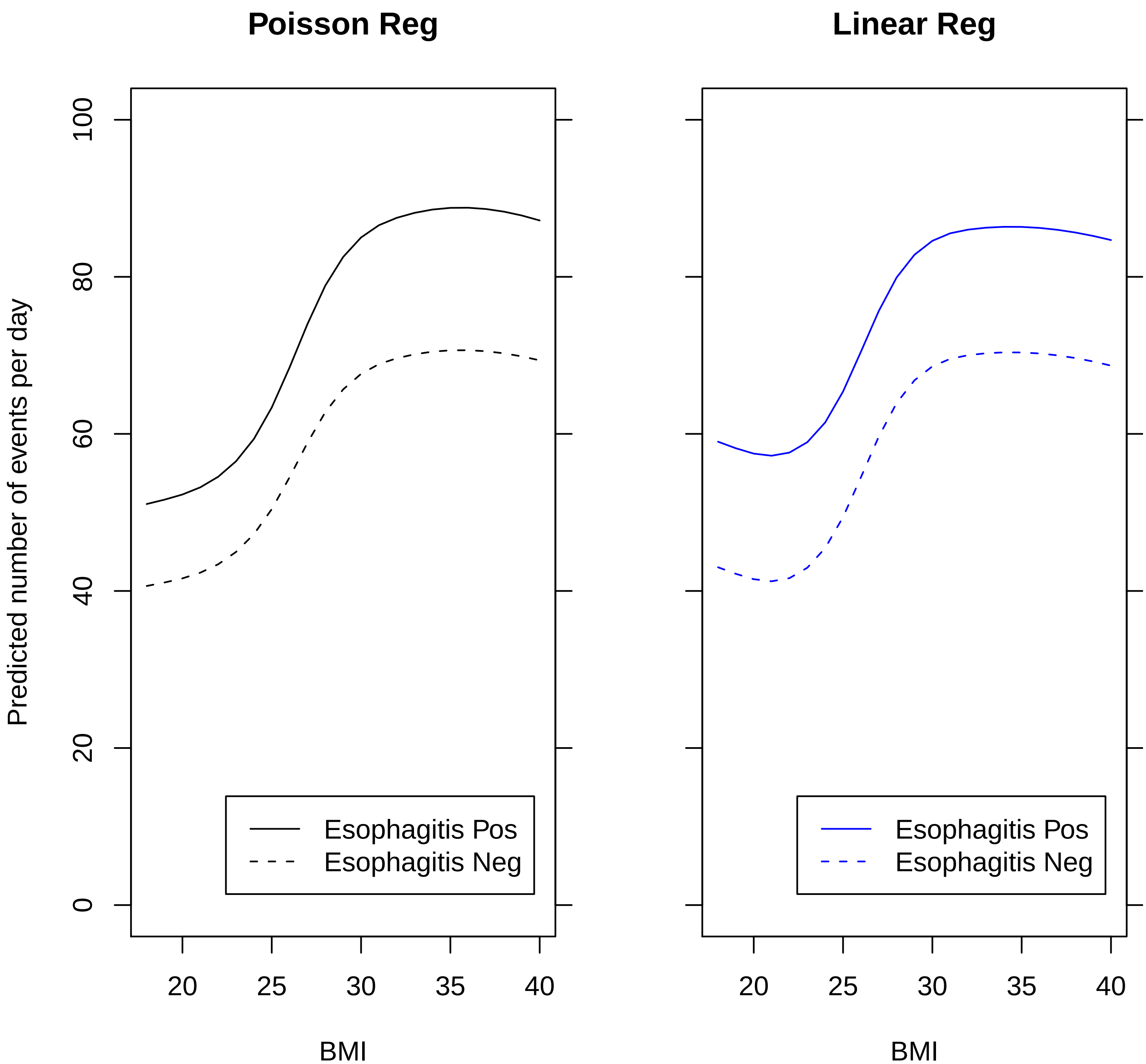
- Example prediction calculations: BMI=30, with esophagitis
  - Linear regression: $0.0278461 + .025104 + .001839 \times 30 = 0.108$
    - Stata: `adjust bmi=30 esop=1`
  - Poisson regression: $e^{-3.089033+0.2622171+.01975465\times 30} = 0.107$
    - Stata: `adjust bmi=30 esop=1, nooffset exp`

– Remember the above rates are for a 2-minute time interval. To convert to daily rates, multiply by 720



10.4.2 BMI modeled using splines

- Regression splines are handled more naturally in R than in Stata
  - `glm(events ~ ns(bmi,4) + esop + offset(log(mins)), data=bmi.data, family="poisson")`
  - `ns(bmi, 4)` specified a natural spline for bmi with 4 degrees of freedom
  - Later, we will discuss regression splines in Stata using `mkspline`
- Note that there is an optical illusion in the following plots
  - For both plots, it appears as if the lines are closer in the middle ranges of BMI
  - For the Poisson regression, the true distance between lines is increasing with increasing with BMI
  - For the Linear regrression, the true distance between lines is constant


Poisson Reg
Linear Reg
Predicted number of events per day
0
20
40
60
80
100
20
25
30
35
40
BMI
Esophagitis Pos
Esophagitis Neg

10.4.3 Comparison of modeling linear BMI to using spline function

- For all regression models, we are more confident modeling associations than predicting means

- When we use a linear term (i.e. a straight line) for the predictor, we are modeling a first-order association
  - Most power to detect this type of association

  - Always need to check that a first-order association answers the scientific question
    - Counter example: Interested in seasonal trends in air pollution. A linear effect of time would only answer if air pollution levels are increasing/decreasing over time, not how they are changing from month to month

- Flexible functions for predictors, including splines, are, in general, more useful if we care about predicting means or individual observations

- Acid reflux example: Which model you choose depends on the scientific goals
  - Primary goal: Is there an association between BMI and the rate of acid reflux?
    - Fitting the linear BMI term answers this question

  - Secondary goal: Describe the (mean) trend in reflux rates as a function of BMI
    - A priori, I would be less inclined to believe a linear function captures the true mean relationship

    - To answer this scientific question, a spline analysis is preferred