Log-linear models

(Poisson regression)

Dr. Maria Tackett

11.13.19

STA 210

[Click for PDF of slides]

Announcements

- HW 06 **due Wed, Nov 20 at 11:59p**
- Project Regression Analysis **due Wed, Nov 20 at 11:59p**
- Looking ahead:
  - Exam 02: Mon, Nov 25 in class
  - Exam review on Nov 20

Poisson response variables

The following are examples of scenarios with Poisson response variables:

- Are the **number of motorcycle deaths** in a given year related to a state's helmet laws?
- Does the **number of employers** conducting on-campus interviews during a year differ for public and private colleges?
- Does the **daily number of asthma-related visits** to an Emergency Room differ depending on air pollution indices?
- Has the **number of deformed fish** in randomly selected Minnesota lakes been affected by changes in trace minerals in the water over the last decade?

STA 210

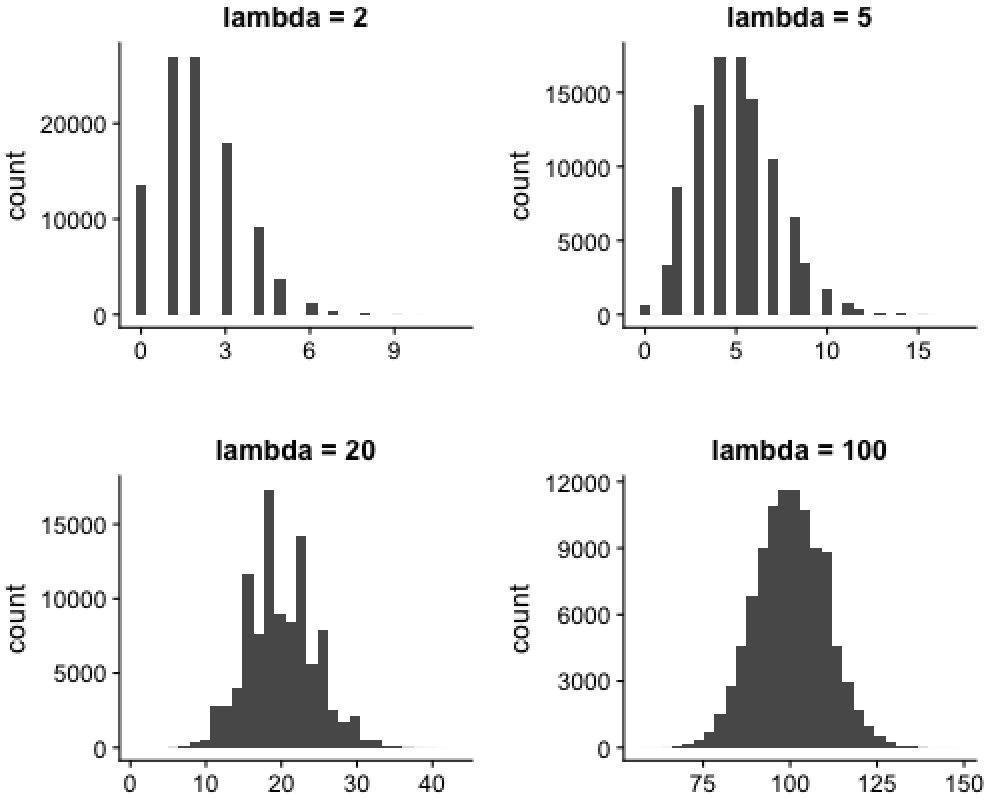
Poisson Distribution

- If Y follows a Poisson distribution, then

$$P(Y = y) = \frac{\exp\{-\lambda\}\lambda^y}{y!} \qquad y = 0, 1, 2, \ldots$$

- Features of the Poisson distribution:
  - Mean and variance are equal (λ)
  - Distribution tends to be skewed right, especially when the mean is small
  - If the mean is larger, it can be approximated by a Normal distribution

Simulated Poisson distributions



Simulated Poisson distributions

	Mean	Variance
lambda=2	2.00740	2.015245
lambda=5	4.99130	4.968734
lambda=20	19.99546	19.836958
lambda=100	100.02276	100.527647

Poisson Regression

- We want λ to be a function of predictor variables $x_1, \ldots, x_p$

Why is a multiple linear regression model not appropriate?

- λ must be greater than or equal to 0 for any combination of predictor variables
- Constant variance assumption will be violated!

STA 210

Multiple linear regression vs. Poisson

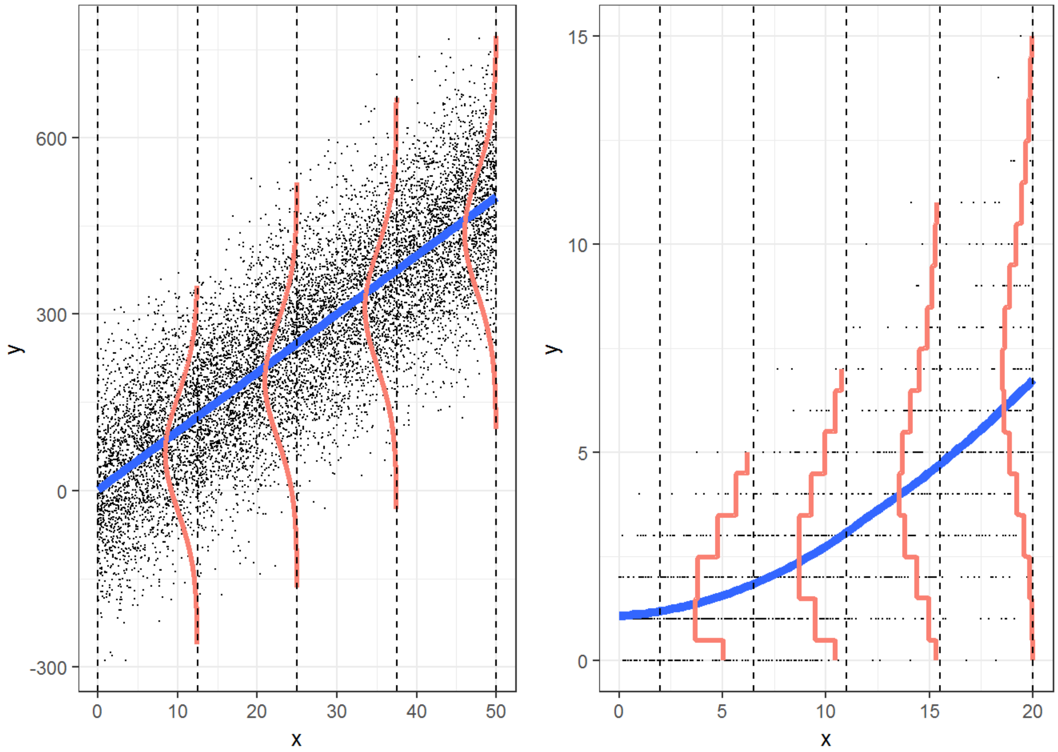


Image from: [*Broadening Your Statistical Horizons*]()

STA 210

Poisson Regression

- If the observed values Y_i are Poisson, then we can model using a **Poisson regression model** of the form

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

- Note that we don't have an error term, since λ determines the mean and variance of the Poisson random variable

STA 210

Interpreting Model Coefficients

- **Slope, β_j:**
  - **Quantitative Predictor**: When x_j increases by one unit, the expected count of y changes by a multiplicative factor of $\exp\{\beta_j\}$, holding all else constant
  - **Categorical Predictor**: The expected count for category k is $\exp\{\beta_j\}$ times the expected count for the baseline category, holding all else constant
- **Intercept, β_0:** When x is 0, the expected count of y is $\exp\{\beta_0\}$

STA 210

Example: Age, Gender, Pulse Rate

- **Goal:** We want to use `age` and `gender` to understand variation in pulse rate
- We will use adults age 20 to 39 in the `NHANES` data set to answer this question
- Reponse
  - **`Pulse`:** Number of heartbeats in 60 seconds
- Explanatory
  - **`Age`:** Age in years. Subjects 80 years or older recorded as 80
    - We will use mean-centered `Age` in the model
  - **`Gender`:** male/female

Example: Age, Gender, Pulse Rate

```r
model1 <- glm(Pulse ~ ageCent + Gender, data = nhanes,
              family = "poisson")
kable(tidy(model1, conf.int = T),format="html")
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.3416799	0.0031800	1365.30794	0.0000000	4.3354407	4.3479061
ageCent	-0.0007360	0.0003933	-1.87118	0.0613201	-0.0015069	0.0000349
Gendermale	-0.0595673	0.0045620	-13.05723	0.0000000	-0.0685091	-0.0506263

1. Write the model equation.
2. Interpret the intercept in the context of the problem.
3. Interpret the coefficient of `ageCent` in the context of the problem.

STA 210

Drop In Deviance Test

- We would like to test if there is a significant interaction between Age and Gender

- Since we have a generalized linear model, we can use the Drop In Deviance Test (similar test to logistic regression)

```
model1 <- glm(Pulse ~ ageCent + Gender, data = nhanes,
              family = "poisson")
model2 <- glm(Pulse ~ ageCent + Gender + ageCent*Gender,
              data = nhanes, family = "poisson")

anova(model1,model2,test="Chisq") %>% kable(format = "markdown")
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
2575	4536.813	NA	NA	NA
2574	4536.345	1	0.4686061	0.4936291

- There is not sufficient evidence that the interaction is significant, so we won't include it in the model.

Model Assumptions

1. **Poisson Response:** Response variable is a count per unit of time or space
2. **Independence:** The observations are independent of one another
3. **Mean = Variance**
4. **Linearity:** $\log(\lambda)$ is a linear function of the predictors

STA 210

Model Diagnostics

- The raw residual for the i^{th} observation, $y_i - \hat{\lambda}_i$, is difficult to interpret since the variance is equal to the mean in the Poisson distribution

- Instead, we can analyze a standardized residual called the **Pearson residual**

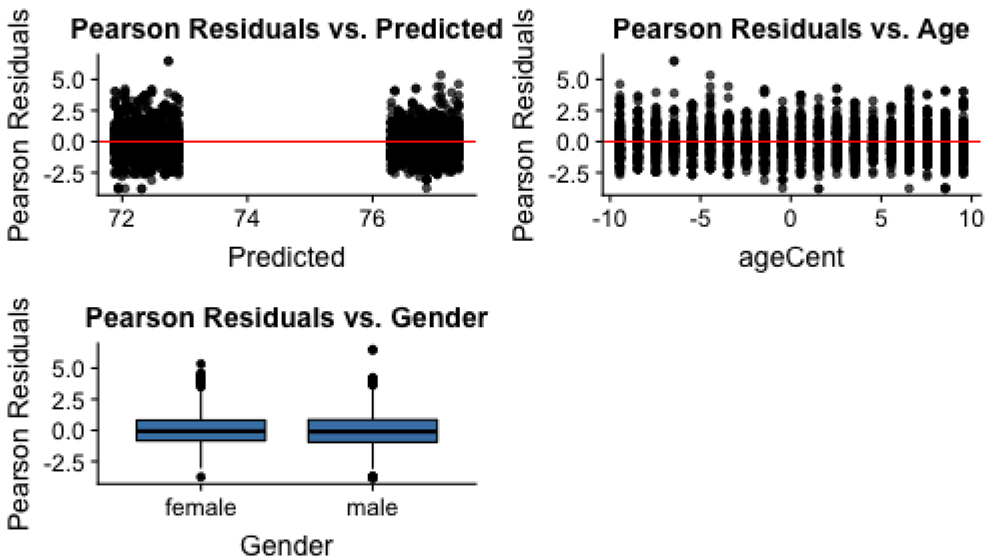
$$r_i = \frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

- Examine a plot of the Pearson residuals versus the predicted values and versus each predictor variable
    - A distinguishable trend in any of the plots indicates that the model is not an appropriate fit for the data

Example: Age, Gender, Pulse Rate

- Let's examine the Pearson residuals for the model that includes the main effects for `Age` and `Gender`

```
nhanes_aug <- augment(model1, type.predict = "response",
                      type.residuals = "pearson")
```



Poisson Regression in R

- Use the **`glm()`** function

```
# poisson regression model
my.model <- glm(Y ~ X, data = my.data, family = poisson)
```

```
# predicted values and Pearson residuals
my.model_aug <- augment(my.model,
                        type.predict = "response",
                        type.residuals = "pearson")
```

Physician Visits

What factors influence the number of times someone visits a physician's office? We will use the variables `chronic`, `health`, and `insurance` to predict `visits`. We will use the `NMES1988` dataset in the AER package.

```
library(AER)
data(NMES1988)
nmes1988 <- NMES1988 %>%
  select(visits, chronic, health, insurance)
glimpse(nmes1988)
```

```
## Observations: 4,406
## Variables: 4
## $ visits    <int> 5, 1, 13, 16, 3, 17, 9, 3, 1, 0, 0, 44, 2, 1, 19, 19, .
## $ chronic   <int> 2, 2, 4, 2, 2, 5, 0, 0, 0, 0, 1, 5, 1, 1, 1, 0, 1, 2, .
## $ health    <fct> average, average, poor, poor, average, poor, average, .
## $ insurance <fct> yes, yes, no, yes, yes, no, yes, yes, yes, yes, yes, y.
```

Physicians Visits

```
visits_model <- glm(visits ~ chronic + health + insurance,
                    data = nmes1988, family = "poisson")
```

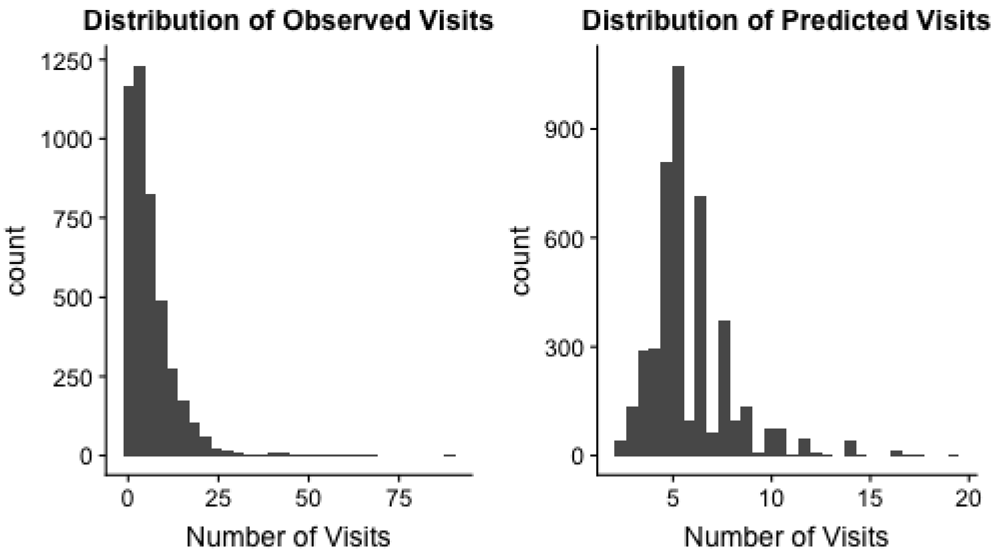
```
tidy(visits_model, conf.int = T) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.217	0.017	71.069	0	1.184	1.251
chronic	0.167	0.004	37.504	0	0.159	0.176
healthpoor	0.290	0.017	16.749	0	0.256	0.324
healthexcellent	-0.360	0.030	-11.889	0	-0.419	-0.301
insuranceyes	0.279	0.016	17.270	0	0.247	0.310

STA 210

Physician Visits

Let's compare the fitted values versus the actual values:



Does the model effectively predict the number of visits? What is the primary difference between the distributions of observed and predicted visits?

Zero-inflated Poisson

- In the original data, there are far more respondents who had zero visits to the physicians office than what's predicted by the Poisson regression model
  - This is called .vocab[zero-inflated data]
- To deal with this, we will fit a model that has 2 parts:
  1. Poisson regression for the number of doctor's visits of those who went to the physician at least one time (parameter = λ)
  2. Logistic regression to find the probability someone goes to the physican at least once (parameter = α)
- We will fit this in R using the `zeroinfl` model in the **pscl** package.

Zero-inflated Poisson Regression

- We will use `chronic`, `health`, and `insurance` for both components of the model
  - Note: We could use different variables for each component of the model.

```
zi_model <- zeroinfl(visits ~ chronic + health + insurance |
                       chronic + health + insurance,
                     dist = "poisson", data = nmes1988)
```

- The first set of coefficients are for the Poisson portion of the model. The second set are for the logistic portion of the model.

STA 210

Zero-inflated Poisson Regression

```
zi_model$coefficients
```

```
## $count
##     (Intercept)         chronic      healthpoor healthexcellent 
##       1.5587860       0.1186671       0.2947644      -0.3019049 
##    insuranceyes 
##       0.1446258 
## 
## $zero
##     (Intercept)         chronic      healthpoor healthexcellent 
##     -0.40531360     -0.55227959      0.02315772      0.23169092 
##    insuranceyes 
##     -0.88637822
```

Let's write the two parts of the model.

Predictions

```
nmes1988 <- nmes1988 %>%
  mutate(pred_count = predict(zi_model, type = "count"),
  pred_zero = predict(zi_model, type = "zero"))
```

```
nmes1988 %>% slice(1:10)
```

```
##    visits chronic  health insurance pred_count  pred_zero
## 1       5       2 average       yes   6.963943 0.08345902
## 2       1       2 average       yes   6.963943 0.08345902
## 3      13       4    poor        no  10.259650 0.06970211
## 4      16       2    poor       yes   9.351253 0.08524762
## 5       3       2 average       yes   6.963943 0.08345902
## 6      17       5    poor        no  11.552315 0.04134603
## 7       9       0 average       yes   5.492655 0.21556659
## 8       3       0 average       yes   5.492655 0.21556659
## 9       1       0 average       yes   5.492655 0.21556659
## 10      0       0 average       yes   5.492655 0.21556659
```

STA 210

References

These slides draw material from [*Broadening Your Statistical Horizons*]