Poisson regression (Chapter 14.13)

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

- Modeling Counts
- Contingency Tables
- Poisson Regression Models

- Number of traffic accidents per day
- Mortality counts in a given neighborhood per week
- Number of customers arriving in a shop daily

We discussed about

- Linear regression: for normally distributed errors
- Logistic regression: for binomial distributed errors

Features of count data

- Counts are not binary (0/1)
- Counts are discrete, not continuous
- Counts typically have a right skewed distribution

So far, the regression strategies we've discussed allow us to model

- Expected values and expected increase in linear regression
- Log odds or log odds ratios in logistic regression
- In modeling counts, we are typically more interested in
 - Incidence rates

• Incidence ratios (when comparing across levels of a risk factor) Poisson regression will provide us with a framework to handle counts properly! • The probability of x occurrence of an event in an interval is:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0, 1, 2, ...$$

where λ is the expected number of occurrences in the interval

•
$$E(X) = Var(X) = \lambda$$

• We can also think of λ as the rate parameter

Poisson Probability



The Poisson distribution can be used to approximate a binomial distribution when

- n is large and p is very small or
- $np = \lambda$ is fixed and n becomes infinitely large

- Yearly cases of esophageal cancer in a large city
- 30 cases observed in 1990

$$P(X=30)=\frac{e^{\lambda}\lambda^{30}}{30!}$$

• $\lambda =$ yearly average number of cases of esophageal cancer

- Men and women are asked whether or not they believed in afterlife (General Social Survey 1991)
- Possible responses were: yes, no or unsure

	Y	N or U	
М	435	147	582
F	375	134	509
Total	810	281	1091

- Question: Is belief in the afterlife independent of gender?
- ${\, \bullet \,}$ We can address this question using a χ^2 test

	Y	N or U	
Μ	435 (432)	147 (150)	582
F	375 (378)	134 (131)	509
Total	810	281	1091

- \bullet We calculated the expected counts to perform the χ^2 test
- Alternatively, we could use a linear model to expression the expected counts systematically

$$Y_{ij} \sim \textit{Poisson}(\lambda_{ij}) \ \lambda_{ij} = \lambda \cdot lpha_{\textit{male}} \cdot \gamma_{\textit{yes}}$$

- λ is the baseline rate, α is the male effect, and γ is the response
- Taking the log of both sides, we have:

$$\log (\lambda_{ij}) = \log (\lambda) + \log (\alpha_{male}) + \log (\gamma_{yes})$$

$$\log (\lambda_{ij}) = \log (\lambda) + \log (\alpha_{male}) + \log (\gamma_{yes})$$

We can also write using $\beta{\rm 's}$

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 I(male) + \beta_2 I(yes)$$

The probabilistic portion of this model enters as:

 $Y_{ij} \sim Poisson(\lambda_{ij})$

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 I(male) + \beta_2 I(yes)$$

- The outcome is the log of the expected cell count
- The baseline β_0 is the log expected cell count for females responding "no"
- β₁ is the increase in log expected cell count for males compared to females
- β₂ is the increase in log expected cell count for the response "yes" compared to "no"

Fitting the afterlife model in R

	Y	N or U	
М	435 (432)	147 (150)	582
F	375 (378)	134 (131)	509
Total	810	281	1091

	\mathtt{count}	male	yes
1	435	1	1
2	147	1	0
3	375	0	1
4	134	0	0

```
(Dispersion parameter for poisson family taken to be 1)
```

Null deviance: 272.685 on 3 degrees of freedom Residual deviance: 0.162 on 1 degrees of freedom AIC: 35.407

Number of Fisher Scoring iterations: 3

So we fit the model:

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 I(male) + \beta_2 I(yes)$$

and out fitted model is:

$$\log(\lambda_{ij}) = 4.88 + 0.13I(male) + 1.06I(yes)$$

Fitting the afterlife model

Using the fitted model:

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 I(male) + \beta_2 I(yes)$$

we can get predicted values for log counts in each of the four cells:For females responding "no":

 $\log E(count | female, no) = 4.88 + 0.134 \cdot 0 + 1.06 \cdot 0 = 4.88$

• For males responding "no":

 $\log E(count | female, no) = 4.88 + 0.134 \cdot 1 + 1.06 \cdot 0 = 5.01$

• Fo female responding "yes":

 $\log E(count|female, no) = 4.88 + 0.134 \cdot 0 + 1.06 \cdot 1 = 5.94$

• For males responding "yes":

 $\log E(count | female, no) = 4.88 + 0.134 \cdot 1 + 1.06 \cdot 1 = 6.07$

Predicting expected cell counts

Using the fitted model

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 I(male) + \beta_2 I(yes)$$

we can get predicted values for counts in each of the four cells:

• For females responding "no":

E(count|female, no) = exp(4.88) = 131

- For males responding "no": exp(5.01) = 150
- For females responding "yes": exp(5.94) = 378
- For males responding "no": exp(6.07) = 432

	Y	N or U	
М	435 (432)	147 (150)	582
F	375 (378)	134 (131)	509
Total	810	281	1091

which are exactly what we got by Poisson regression!

- By fitting the independence model, we force the relative rate of responding "yes" versus "no" to the question of belief in the afterline to be fixed across males and females
- Deviation from the independence model suggests the proportion of those believing in afterlife differs by gender

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 I(male) + \beta_2 I(yes)$$

•
$$\beta_0 = 4.88$$
 is

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 I(male) + \beta_2 I(yes)$$

- $\beta_0=4.88~{\rm is}~$ the log expected count of females responding "no", the baseline group
- $\beta_1=0.134$ is

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 I(male) + \beta_2 I(yes)$$

- $\beta_0 = 4.88$ is the log expected count of females responding "no", the baseline group
- $\beta_1=0.134$ is the difference in log expected counts comparing males to females
- $\beta_2 = 1.05$ is

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 I(male) + \beta_2 I(yes)$$

- $\beta_0 = 4.88$ is the log expected count of females responding "no", the baseline group
- $\beta_1 = 0.134$ is the difference in log expected counts comparing males to females
- $\beta_2 = 1.05$ is the difference in log expected counts for "yes" responses compared to "no" responses

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 I(male) + \beta_2 I(yes),$$

under this independence model:

- $exp(\beta_0) = 131.5$ is the expected count for females responding "no", the baseline group
- $exp(\beta_1) = 1.14$ is the ratio comparing the counts of males to females
- exp(β₂) = 2.85 is the ratio of the number of "yes" responses compared to "no" responses

Outcome Y=number of customers visiting store from region Predictors:

- X_1 : number of housing units in region
- X₂: average household income
- X₃: average housing unit age in region
- X₄: distance to nearest competitor
- X₅: average distance to store in miles

Counts are obtained for 110 regions, so our $n{=}110$

Lumber Company Data

> lumber[1:10,]

	customers	housing	income	age	compet_dist	store_dist
1	9	606	41393	3	3.04	6.32
2	6	641	23635	18	1.95	8.89
3	28	505	55475	27	6.54	2.05
4	11	866	64646	31	1.67	5.81
5	4	599	31972	7	0.72	8.11
6	4	520	41755	23	2.24	6.81
7	0	354	46014	26	0.77	9.27
8	14	483	34626	1	3.51	7.92
9	16	1034	85207	13	4.23	4.40
10	13	456	33021	32	3.07	6.03

Examples for Multinomial Logistic Regression



Histogram of Customer

- The distribution of customer counts is clearly not normally distributed
- Linear regression would not work well here
- Log-linear regression will work just fine

```
> summary(lumber.glm <- glm(customers ~ housing +
   income +age + compet_dist +store_dist, family=poisson()) )
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.942e+00 2.072e-01 14.198 < 2e-16 ***
housing
         6.058e-04 1.421e-04 4.262 2.02e-05 ***
income -1.169e-05 2.112e-06 -5.534 3.13e-08 ***
    -3.726e-03 1.782e-03 -2.091 0.0365 *
age
compet_dist 1.684e-01 2.577e-02 6.534 6.39e-11 ***
store_dist -1.288e-01 1.620e-02 -7.948 1.89e-15 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
   Null deviance: 422.22 on 109 degrees of freedom
Residual deviance: 114.99 on 104 degrees of freedom
ATC: 571.02
```

Number of Fisher Scoring iterations: 4

- We interpret $\hat{\beta}_0$ as baseline log expected count, or log rate, in the group with all covariates (housing, income, age, distance to nearest competitor, and average distance to store) set to zero
- $exp(\hat{\beta}_0) = exp(2.94) = 18.9$ is the expected count of customers in the baseline group
- This baseline value does not quite make sense
- It may be helpful to center our covariates, but this is not a big deal if we don't care about baseline because our primary inference is about the increase with respect to covariates

• We interpret
$$\widehat{eta}_1=6.05 imes10^{-4}$$
 as

- We interpret $\hat{\beta}_1 = 6.05 \times 10^{-4}$ as the increase in log expected count, or the log rate ratio comparing districts whose number of housing units differ by one, adjusting for other covariates
- $exp(\widehat{\beta}_1) = exp(6.05 \times 10^{-4}) = 1.000605$ is

- We interpret $\hat{\beta}_1 = 6.05 \times 10^{-4}$ as the increase in log expected count, or the log rate ratio comparing districts whose number of housing units differ by one, adjusting for other covariates
- exp(β₁) = exp(6.05 × 10⁻⁴) = 1.000605 is the rate ratio comparing districts whose mean housing units differ by one, adjusting for other covariates
- $exp(100 \cdot \widehat{eta}_1) = exp(100 \cdot 6.05 \times 10^{-4}) = 1.062$ is

- We interpret $\hat{\beta}_1 = 6.05 \times 10^{-4}$ as the increase in log expected count, or the log rate ratio comparing districts whose number of housing units differ by one, adjusting for other covariates
- exp(β₁) = exp(6.05 × 10⁻⁴) = 1.000605 is the rate ratio comparing districts whose mean housing units differ by one, adjusting for other covariates
- exp(100 · β₁) = exp(100 · 6.05 × 10⁻⁴) = 1.062 is the rate ratio comparing districts whose mean housing units differ by one hundred → Keeping other factors constant, a 100 unit increase in housing units, would yield an expected 6.2% increase in customer count.
- Question: Based on this model , if we are going to choose a location to build a new store, should we choose areas with higher or lower income? Does it matter?

- Poisson regression gives us a framework in which to build models for count data
- It is a special case of generalized linear models, so it is closely related to linear and logistic regression modelling
- All of the same modelling techniques will carry over from linear regression:
 - Adjustment for confounding
 - Allowing for effect modification by fitting interactions
 - Splines and polynomial terms