Poisson Regression Inference

STA courses next semester

Some classes to consider after STA 214:

- STA 247 Design and Sampling
- STA 279 Statistical Computing
- STA 310 Probability (requires calc II)
- STA 363 Intro to Statistical Learning (requires linear algebra)

Other cool courses to consider:

- STA 311 Statistical Inference (requires STA 310)
- STA 312 Linear Models (requires STA 310 and linear algebra)
- STA 362 Multivariate Statistics (requires linear algebra)
- STA 365 Applied Bayesian Statistics (requires STA 310)
- STA 368 Time Series and Forecasting (requires STA 310)

Last time

A concerned parent asks us to investigate crime rates on college campuses. We have access to data on 81 different colleges and universities in the US, including the following variables:

- `type`: college (C) or university (U)
- `nv`: the number of crimes for that institution in the given year
- `enroll1000`: the number of enrolled students, in thousands
- `region`: region of the US C = Central, MW = Midwest, NE = Northeast, SE = Southeast, SW = Southwest, and W = West)

$$Crimes_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 MW_i + \beta_2 NE_i + \beta_3 SE_i + \beta_4 SW_i + \beta_5 W_i$$

Goodness of fit

Goodness of fit test: If the model is a good fit for the data, then the residual deviance follows a χ^2 distribution with the same degrees of freedom as the residual deviance

Residual deviance = 621.24, df = 75

```
pchisq(621.24, df=75, lower.tail=F)
```

```
## [1] 5.844298e-87
```

So our model might not be a very good fit to the data.

> Why might our model not be a good fit?

Potential issues with our model

- The Poisson distribution might not be a good choice
- There may be additional factors related to the number of crimes which we are not including in the model

> Which other factors might be related to the number of crimes?

Offsets

We will account for school size by including an **offset** in the model:

$$\log(\lambda_i) = \beta_0 + \beta_1 MW_i + \beta_2 NE_i + \beta_3 SE_i + \beta_4 SW_i + \beta_5 W_i + \log(Enrollment_i)$$

Motivation for offsets

We can rewrite our regression model with the offset:

$$\log(\lambda_i) = \beta_0 + \beta_1 MW_i + \beta_2 NE_i + \beta_3 SE_i + \beta_4 SW_i + \beta_5 W_i + \log(Enrollment_i)$$

Fitting a model with an offset

```
m2 <- glm(nv ~ region, offset = log(enroll1000),
          data = crimes, family = poisson)
summary(m2)
```

```
...
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.30445    0.12403 -10.517  < 2e-16 ***
## regionMW     0.09754    0.17752   0.549  0.58270
## regionNE     0.76268    0.15292   4.987 6.12e-07 ***
## regionSE     0.87237    0.15313   5.697 1.22e-08 ***
## regionSW     0.50708    0.18507   2.740  0.00615 **
## regionW      0.20934    0.18605   1.125  0.26053
...
```

- The offset doesn't show up in the output (because we're not estimating a coefficient for it)

Fitting a model with an offset

$$\log(\widehat{\lambda}_i) = -1.30 + 0.10MW_i + 0.76NE_i + 0.87SE_i + 0.51SW_i + 0.21W_i + \log(Enrollment_i)$$

How would I interpret the intercept -1.30?

When to use offsets

Offsets are useful in Poisson regression when our counts come from groups of very different sizes (e.g., different numbers of students on a college campus). The offset lets us interpret model coefficients in terms of rates instead of raw counts.

> With your neighbor, brainstorm some other data scenarios where our response is a count variable, and an offset would be useful. What would our offset be?

Inference

$$\log(\lambda_i) = \beta_0 + \beta_1 MW_i + \beta_2 NE_i + \beta_3 SE_i + \beta_4 SW_i + \beta_5 W_i + \log(Enrollment_i)$$

Our concerned parent wants to know whether the crime rate on campuses is different in different regions.

> What hypotheses would we test to answer this question?

Likelihood ratio test

Full model:

$$\log(\lambda_i) = \beta_0 + \beta_1 MW_i + \beta_2 NE_i + \beta_3 SE_i + \beta_4 SW_i + \beta_5 W_i + \log(Enrollment_i)$$

Reduced model:

$$\log(\lambda_i) = \beta_0 + \log(Enrollment_i)$$

Likelihood ratio test

```
m2 <- glm(nv ~ region, offset = log(enroll1000),
          data = crimes, family = poisson)
summary(m2)
```

```
...
##     Null deviance: 491.00  on 80  degrees of freedom
## Residual deviance: 433.14  on 75  degrees of freedom
...
```

> What is my test statistic?

Likelihood ratio test

```
m2 <- glm(nv ~ region, offset = log(enroll1000),
          data = crimes, family = poisson)
summary(m2)
```

```
...
##     Null deviance: 491.00  on 80  degrees of freedom
## Residual deviance: 433.14  on 75  degrees of freedom
...
```

$$G = 491 - 433.14 = 57.86$$

```
pchisq(57.86, df=5, lower.tail=F)
```

```
## [1] 3.361742e-11
```

Inference

$$\log(\lambda_i) = \beta_0 + \beta_1 MW_i + \beta_2 NE_i + \beta_3 SE_i + \beta_4 SW_i + \beta_5 W_i + \log(Enrollment_i)$$

Now our concerned parent wants to know about the difference between Western and Central schools. They would like a "reasonable range" of values for the difference between the regions.

> How would we construct a "reasonable range" of values for this difference?

Confidence intervals

$$\log(\lambda_i) = \beta_0 + \beta_1 MW_i + \beta_2 NE_i + \beta_3 SE_i + \beta_4 SW_i + \beta_5 W_i + \log(Enrollment_i)$$

```
...
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.30445    0.12403 -10.517  < 2e-16 ***
## regionMW     0.09754    0.17752   0.549  0.58270
## regionNE     0.76268    0.15292   4.987 6.12e-07 ***
## regionSE     0.87237    0.15313   5.697 1.22e-08 ***
## regionSW     0.50708    0.18507   2.740  0.00615 **
## regionW      0.20934    0.18605   1.125  0.26053
...
```

95% confidence interval for β_5:

Class activity

https://sta214-s23.github.io/class_activities/ca_lecture_20.html

Class activity

$$Articles_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 Female_i + \beta_2 Married_i + \beta_3 Kids_i + \beta_4 Prestige_i + \beta_5 Mentor_i$$

> Do I need an offset for this model?

Class activity

$$Articles_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 Female_i + \beta_2 Married_i + \beta_3 Kids_i + \beta_4 Prestige_i + \beta_5 Mentor_i$$

> We are interested in the relationship between prestige and the number of articles published, after accounting for other factors. What confidence interval should we make?

Class activity

```
...
##                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 ***
...
```

> How do I construct a confidence interval for $\exp\{\beta_4\}$?