Poisson Regression

Data

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 violent 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)

Data

We want to know whether there are regional differences in the number of violent crimes on college campuses.

> What would be a reasonable model to investigate this question?

Model

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

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

> What assumptions is this model making?

Model

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

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

> What assumptions is this model making?

- Poisson distribution
- Independence
- Not making a shape assumption, because I just have a categorical predictor (so no notion of linearity)

> How do I assess these assumptions?

Checking assumptions

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

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

- Poisson assumption:
  - Check that response is a count
  - Check that the distribution of the response looks like it could be Poisson
  - Check that the mean and variance of the response are similar
- Independence: think about data

Class activity

https://sta214-f22.github.io/class_activities/ca_lecture_18.html

Class activity

$$\log(\widehat{\lambda}_i) = 1.34 + 0.48\ MW_i + 0.44\ NE_i + 0.77\ SE_i + 0.33\ SW_i + 0.53\ W_i$$

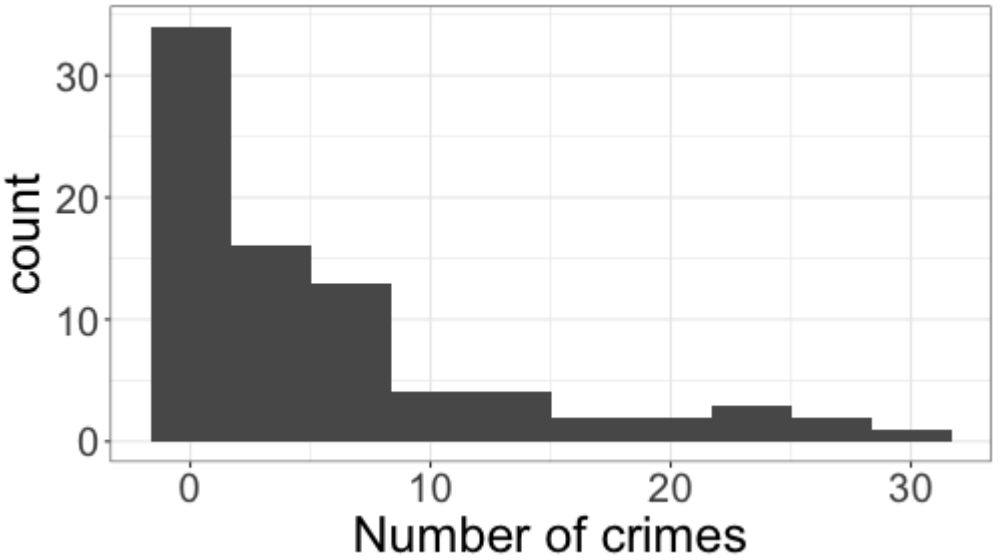
How do I interpret the estimated intercept 1.34?

Class activity

$$\log(\widehat{\lambda}_i) = 1.34 + 0.48\ MW_i + 0.44\ NE_i + 0.77\ SE_i + 0.33\ SW_i + 0.53\ W_i$$

How do I interpret the estimated coefficient 0.48?

Class activity



Does it look reasonable to assume a Poisson distribution for the response?

Class activity

```
mean(crimes$nv)
```

```
## [1] 5.938272
```

```
var(crimes$nv)
```

```
## [1] 54.83364
```

> Does the Poisson distribution still seem reasonable?

Class activity

```
mean(crimes$nv)
```

```
## [1] 5.938272
```

```
var(crimes$nv)
```

```
## [1] 54.83364
```

> Does the Poisson distribution still seem reasonable?

Not necessarily -- the mean is much lower than the variance. We will see a couple options for handling this issue later.

Goodness of fit

Another way to assess whether our model is reasonable is with a *goodness of fit* test.

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

```
...
##     Null deviance: 649.34  on 80  degrees of freedom
## Residual deviance: 621.24  on 75  degrees of freedom
...
```

Residual deviance = 621.24, df = 75

> How likely is a residual deviance of 621.24 if our model is correct?

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 violent crimes which we are not including in the model

> Which other factors might be related to the number of violent 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
```