Inference and overdispersion

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)

Model

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

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

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

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

Confidence interval for β_5:

Computing z^*

Example: for a 95% confidence interval, $z^* = 1.96$

```
qnorm(0.025, lower.tail=F)
```

```
## [1] 1.959964
```

Example: for a 99% confidence interval, $z^* = 2.58$:

```
qnorm(0.005, lower.tail=F)
```

```
## [1] 2.575829
```

Confidence intervals

```
...
##                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-f22.github.io/class_activities/ca_lecture_19.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

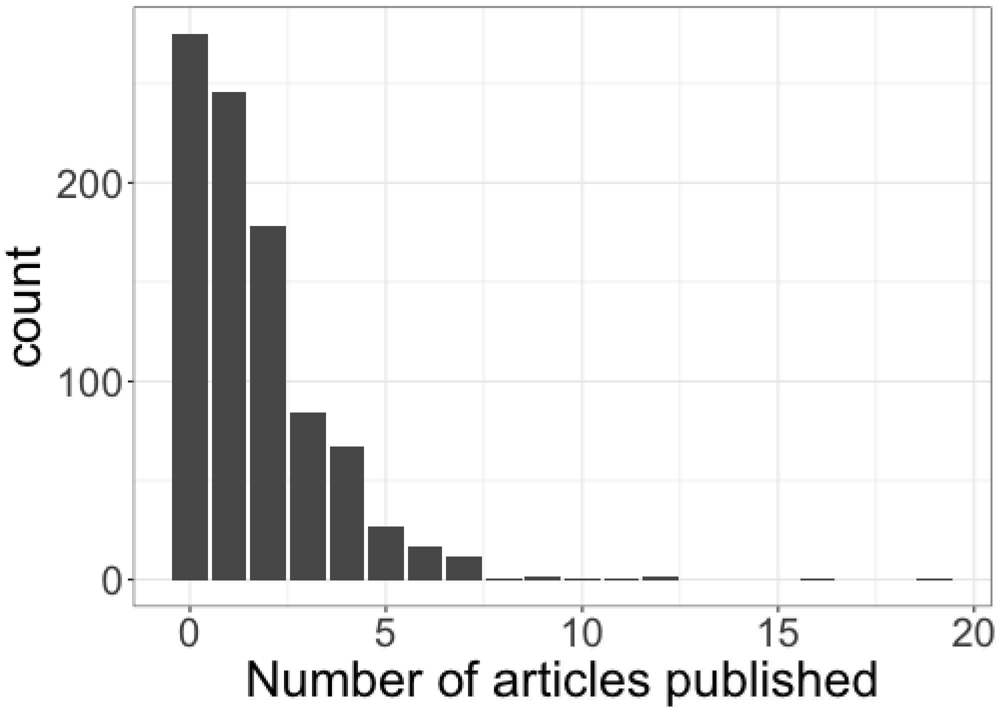
```
...
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.304617   0.102981   2.958   0.0031 **
## femWomen     -0.224594   0.054613  -4.112 3.92e-05 ***
## marMarried    0.155243   0.061374   2.529   0.0114 *
## kid5         -0.184883   0.040127  -4.607 4.08e-06 ***
## phd           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 $\exp\{\beta_4\}$?

Checking assumptions

- But, we haven't checked assumptions yet!
- Let's check the Poisson assumption

Checking assumptions



Does a Poisson distribution seem reasonable, given this plot?

Checking assumptions

Checking the mean/variance condition:

```
mean(articles$art)
```

```
## [1] 1.692896
```

```
var(articles$art)
```

```
## [1] 3.709742
```

> Does it look like the mean and variance could be the same?

Overdispersion

Overdispersion occurs when the response Y has higher variance than we would expect if Y followed a Poisson distribution.

> What problems do you think it causes to assume the mean and variance are the same, when they are not?

Formal checks for overdispersion

First, we need a formal measure of dispersion (relation between mean and variance):

$$\phi = \frac{\text{Variance}}{\text{Mean}}$$

> What should ϕ be if there is no overdispersion?

Hypothesis test for overdispersion

$$\phi = \frac{\text{Variance}}{\text{Mean}}$$

$H_0 : \phi = 1$ (no overdispersion)

$H_A : \phi > 1$ (overdispersion)

Now we need to estimate ϕ...

Pearson residuals and estimated dispersion

The *Pearson residual* for observation i is defined as

$$e_{(P)i} = \frac{Y_i - \widehat{\lambda}_i}{\sqrt{\widehat{\lambda}_i}}$$

$$\widehat{\phi} = \frac{\sum_{i=1}^{n} e_{(P)i}^2}{n - p}$$

- $p =$ number of parameters in model

Example: estimating dispersion parameter in R

```r
# fit the model
m1 <- glm(art ~ ., data = articles,
          family = poisson)

# get Pearson residuals
pearson_resids <- resid(m1, "pearson")

# estimate dispersion parameter
phihat <- sum(pearson_resids^2)/(915 - 6)
phihat
```

```
## [1] 1.828984
```

Back to the hypothesis test

$$\phi = \frac{\text{Variance}}{\text{Mean}}$$

$H_0 : \phi = 1$ (no overdispersion)

$H_A : \phi > 1$ (overdispersion)

$\widehat{\phi} = 1.829$

> Now what?

Calculating a p-value

```
library(AER)
dispersiontest(m1)
```

```
##
##      Overdispersion test
##
## data:  m1
## z = 5.7825, p-value = 3.681e-09
## alternative hypothesis: true dispersion is greater than
## sample estimates:
## dispersion
##    1.82454
```

So there is strong evidence for overdispersion in the data.

Handling overdispersion

Overdispersion is a problem because our standard errors (for confidence intervals and hypothesis tests) are too low.

> If we think there is overdispersion, what should we do?

Adjusting the standard error

- In our data, $\widehat{\phi} = 1.829$
- This means our variance is 1.829 times bigger than it should be
- So our standard error is $\sqrt{1.829} = 1.352$ times bigger than it should be

New confidence interval for β_4:

$$0.0128 \pm 1.96 \cdot \sqrt{1.829} \cdot 0.0264 = (-0.0572,\ 0.0828)$$

Adjusting the standard error in R

```
m2 <- glm(art ~ ., data = articles,
          family = quasipoisson)
```

```
...
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.304617   0.139273   2.187 0.028983 *
## femWomen    -0.224594   0.073860  -3.041 0.002427 **
## marMarried   0.155243   0.083003   1.870 0.061759 .
## kid5        -0.184883   0.054268  -3.407 0.000686 ***
## phd          0.012823   0.035700   0.359 0.719544
## ment         0.025543   0.002713   9.415  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.829006)
...
```

- Allowing ϕ to be different from 1 means we are using a *quasi-likelihood* (in this case, a *quasi-Poisson*)

Adjusting the standard error in R

Poisson:

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

Quasi-Poisson:

```
...
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.304617   0.139273   2.187 0.028983 *
## femWomen    -0.224594   0.073860  -3.041 0.002427 **
## marMarried   0.155243   0.083003   1.870 0.061759 .
## kid5        -0.184883   0.054268  -3.407 0.000686 ***
...
```