

# 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)

## Offsets

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

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

$$\begin{aligned} \log(\lambda_i) = & \beta_0 + \beta_1 MW_i + \beta_2 NE_i + \beta_3 SE_i + \beta_4 SW_i + \beta_5 W_i \\ & + \log(Enrollment_i) \end{aligned}$$

## 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

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

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?

# Goodness of fit

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

```
...  
## (Dispersion parameter for poisson family taken to be 1)  
##  
##      Null deviance: 491.00  on 80  degrees of freedom  
## Residual deviance: 433.14  on 75  degrees of freedom  
...
```

```
pchisq(433.14, df=75, lower.tail=F)
```

```
## [1] 8.33082e-52
```

# Overdispersion

**Overdispersion** occurs when the response  $Y$  has higher variance than we would expect from the specified EDM

Why is it a problem if  $Y$  has more variance than we account for in our model?



# Estimating $\phi$

## Using $\hat{\phi}$

```
pearson_resids <- residuals(m2, type="pearson")
sum(pearson_resids^2)/df.residual(m2)
```

```
## [1] 7.58542
```

```
...
```

##		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	

```
...
```