Overdispersion

- Reminders:
    - No class on Friday
        - I have posted a Poisson regression activity on the course website
    - Exam 1 re-submission due Monday 10/24

Last time: unit deviance for Normal

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{y^2}{2\sigma^2}\right\} \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2}\right\}$$

$$t(y,\mu) = y\mu - \frac{\mu^2}{2}$$

$$\Rightarrow t(y,y) = \frac{y^2}{2}$$

$$2(t(y,y) - t(y,\mu)) = 2\left(\frac{y^2}{2} - y\mu + \frac{\mu^2}{2}\right)$$

$$= 2\left(\frac{1}{2}(y-\mu)^2\right)$$

$$= (y-\mu)^2$$

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)

$Y_i \sim EDM(\mu_i, \phi)$

$g(\mu_i) = \beta^T X_i + O_i$

Offsets

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

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

$$\log(\lambda_i) = \beta_0 + \beta_1 MW_i + \beta_2 NE_i + \beta_3 SE_i + \beta_4 SW_i + \beta_5 W_i + \underbrace{\log(Enrollment_i)}_{\text{offset}}$$

$\Rightarrow \log\left(\frac{\lambda_i}{Enrollment_i}\right) = \beta_0 + \beta_1 MW_i + \cdots$

- let's us βs in terms of rates $\left(\frac{\lambda_i}{Enrollment_i}\right)$
- response is still Crimes
- still assume Poisson distribution for crimes

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?

The estimated average crime rate for central colleges is $e^{-1.3} = 0.273$ crimes per 1000 students

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)
```

Poisson: $\phi = 1$

```
...
## (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
```

≈ 0

Perhaps Poisson is wrong...

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?

Overdispersion

$$Y_i \sim EDM(\mu_i, \phi)$$

$$g(\mu_i) = \beta^T X_i + o_i$$

g = canonical link

$$\text{var}(Y_i) = \phi\, v(\mu_i) \qquad v = \frac{\partial \mu}{\partial \theta}$$

$$\mathcal{I}(\beta) = \frac{X^T V X}{\phi} \qquad V = \text{diag}(v(\mu_i))$$

if $\phi = 1$: $\mathcal{I}(\beta) = X^T V X$

$\phi > 1$, $\mathcal{I}(\beta) = \frac{X^T V X}{\phi}$

$$\Rightarrow \text{Var}(\hat\beta \mid \phi > 1) = \phi \cdot \text{Var}(\hat\beta \mid \phi = 1)$$

$\Rightarrow$ CIs are too narrow

$$u(\beta) = \frac{X^T(Y - \mu)}{\phi}$$

$\hat\beta$ solves $u(\beta) = 0$

$\Rightarrow \hat\beta$ solves $X^T(Y-\mu) = 0$

Analogy to $N(\mu, \sigma^2)$

$$\hat\mu = \bar X$$

$$\hat\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar X)^2$$

Estimating ϕ

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
...
```