EDMs and goodness of fit

Recap: EDMs and GLMs

$$f(y;\theta,\phi) = a(y,\phi)\exp\left\{\frac{y\theta - \kappa(\theta)}{\phi}\right\}$$

$$\theta = g(\mu) \qquad \mu = \mathbb{E}[Y]$$

GLM:

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

$$h(\mu_i) = \beta^T X_i$$

canonical link: $h(\mu_i) = g(\mu_i) = \theta_i$

Why is the canonical link function nice?

The canonical link has nice mathematical properties

<u>Example</u>: observe $(X_1, Y_1), \dots, (X_n, Y_n)$

want to estimate β

$$L(\beta) = \prod_{i=1}^{n} a(Y_i, \phi) \exp\left\{ \frac{Y_i\theta - K(\theta)}{\phi} \right\}$$

$$\Rightarrow \ell(\beta) = \sum_{i=1}^{n} \log(a(Y_i, \phi)) + \frac{1}{\phi} \sum_{i=1}^{n} (Y_i\theta - K(\theta))$$

$$\Rightarrow \text{(HW)} \begin{cases} u(\beta) = \dfrac{1}{\phi} \displaystyle\sum_{i=1}^{n} (Y_i X_i - \mu_i X_i) = \dfrac{X^T(Y-\mu)}{\phi} \\ \mathcal{I}(\beta) = \dfrac{1}{\phi^2} \displaystyle\sum_{i=1}^{n} Var(Y_i) X_i X_i^T = \dfrac{X^T W X}{\phi^2}, \quad W = diag(Var(Y_i)) \end{cases}$$

$$= \frac{X^T V X}{\phi} \qquad V = diag(V(\mu_i)) \qquad V(\mu_i) = \frac{Var(Y_i)}{\phi}$$

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

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

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

Fitted model:

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

> How would I interpret the intercept 1.34?

The estimated average # of crimes in the central region is $e^{1.34} = 3.82$

The average # of crimes for a school in NE is $e^{0.44}$ times higher than the average # of crimes in central region

Model

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

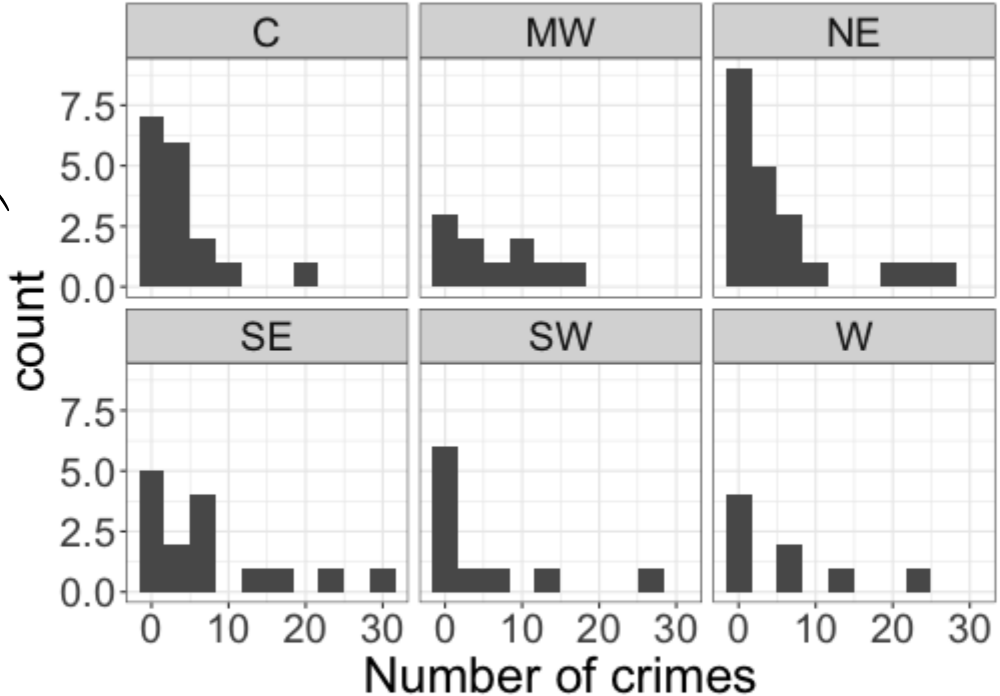
$$\log(\lambda_i) = \beta_0 + \beta_1 MW_i + \beta_2 NE_i + \beta_3 SE_i + \beta_4 SW_i + \beta_5 W_i$$

> What assumptions is this model making?

- Not really a shape assumption, b/c we have a single categorical explanatory variable
- # Crimes is a count variable ✓
- Poisson distribution ←
  - mean ≈ variance
  - unimodal & right skewed
- Independence

Exploratory data analysis

· count look right skewed and unimodal



Exploratory data analysis

Mean and variance for number of crimes by region:

variance >> mean

(generally concerned when variance > 2 x mean)

region	mean	variance
C	3.82	24.28
MW	6.20	37.07
NE	5.95	59.05
SE	8.27	84.35
SW	5.30	75.34
W	6.50	65.71

Goodness of fit

H_0: model is a "good fit"

H_A: model is not a good fit

Test statistic: residual deviance $\approx \chi^2_{n-(k+1)}$

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 ← $n-(k+1)$

$81-6$

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

$E[\chi^2_\nu] = \nu$

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
```

≈ 0

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

Sometimes see $\frac{\text{residual deviance}}{\text{df}} >> 1$

EDMs and deviance

$$f(y; \theta, \phi) = a(y, \phi) \exp\left\{ \frac{y\theta - \kappa(\theta)}{\phi} \right\}$$

Let $t(y, \mu) = y\theta - \kappa(\theta)$ ($\theta = g(\mu)$, so $\exists$ some t for which this is true)

<u>Claim</u>: $t(y, \mu)$ has a maximum at $\mu = y$

<u>Pf</u>: $\frac{\partial}{\partial \theta} t(y, \mu) = y - \frac{\partial \kappa(\theta)}{\partial \theta} = y - \mu$
$= 0$ when $\mu = y$

$$\frac{\partial^2}{\partial \theta^2} t(y, \mu) = -\frac{\partial^2 \kappa(\theta)}{\partial \theta^2} = -\frac{\partial \mu}{\partial \theta} = -V(\mu) < 0 \quad //$$

Let $d(y, \mu) = 2\{t(y, y) - t(y, \mu)\}$ <u>unit (unscaled) deviance</u>

$d(y, \mu) = 0$ when $\mu = y$

$d(y, \mu) > 0$ when $\mu \neq y$

$\Rightarrow d(y, \mu)$ measures <u>distance</u> from y to μ

$$f(y;\theta,\phi) = a(y,\phi)\exp\left\{\frac{y\theta-\kappa(\theta)}{\phi}\right\}$$

$$= a(y,\phi)\exp\left\{\frac{t(y,\mu)-t(y,y)+t(y,y)}{\phi}\right\}$$

$$= \underbrace{a(y,\phi)\exp\left\{\frac{t(y,y)}{\phi}\right\}}\exp\left\{\frac{t(y,\mu)-t(y,y)}{\phi}\right\}$$

$$f(y;\mu,\phi) = b(y,\phi)\exp\left\{-\frac{d(y,\mu)}{2\phi}\right\}$$ dispersion model form

$$Y_i \sim \mathrm{EDM}(\mu_i,\phi)$$

Total (unscaled) deviance: $D(y,\mu) = \sum_{i=1}^{n} d(Y_i,\mu_i)$

Total scaled deviance: $D^*(y,\mu) = \dfrac{D(y,\mu)}{\phi}$

Residual deviance: $D(y,\hat{\mu}) = \sum_{i=1}^{n} d(Y_i,\hat{\mu}_i)$

Scaled residual deviance: $D^*(y,\hat{\mu}) = \dfrac{D(y,\hat{\mu})}{\phi} = 2\left(\ell(\text{saturated}) - \ell(\hat{\beta})\right)$

Saddlepoint approximation

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?

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?