Goodness of fit and overdispersion

- No class on Friday (out of town)
  - I will post an activity or a pre-recorded lecture
- HW 5 released

Recap: EDMs and deviance

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

Let $t(y,\mu) = y\theta - \kappa(\theta)$

$d(y,\mu) = 2\{t(y,y) - t(y,\mu)\}$ unit (unscaled) deviance

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

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

Scaled residual deviance : $D^*(y,\hat{\mu}) = \frac{D(y,\hat{\mu})}{\phi}$

$$= 2\left(\ell(\text{saturated}) - \ell(\hat{\beta})\right)$$

$$L(\hat{\beta}) = \prod_{i=1}^{n} b(y_i, \phi) \cdot \exp\left\{ - \frac{d(y_i, \hat{\mu}_i)}{2\phi} \right\}$$

$$2\ell(\hat{\beta}) = 2\sum_{i=1}^{n} \log b(y_i, \phi) \underbrace{- \sum_{i=1}^{n} \frac{d(y_i, \hat{\mu}_i)}{\phi}}_{-D^*(y, \hat{\mu})}$$

$g(\mu_i) = \beta^T x_i$

$\hat{\mu}_i = g^{-1}(\hat{\beta}^T x_i)$

$$2\ell(\text{saturated}) = 2\sum_{i=1}^{n} \log b(y_i, \phi) - \underbrace{\cancel{\sum_{i=1}^{n} \frac{d(y_i, y_i)}{\phi}}}_{0}$$

$$2\left(\ell(\text{saturated}) - \ell(\hat{\beta})\right) = D^*(y, \hat{\mu})$$

Examples :

1) Poisson :

$$t(y,\mu) = y\log\mu - \mu$$

$$t(y,y) = y\log y - y$$

convention: $0\log 0 = 0$

$$\Rightarrow d(y,\mu) = 2\left(y\log\left(\frac{y}{\mu}\right) - (y-\mu)\right)$$

2) Bernoulli :

$$t(y,\mu) = y\log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)$$

$$= y\log\mu + (1-y)\log(1-\mu)$$

$$t(y,y) = y\log y + (1-y)\log(1-y)$$

$$\Rightarrow d(y,\mu) = 2\left(y\log\left(\frac{y}{\mu}\right) + (1-y)\log\left(\frac{1-y}{1-\mu}\right)\right)$$

3) Normal :

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

$$t(y,\mu) = -\frac{1}{2}(y-\mu)^2 \Rightarrow t(y,y) = -\frac{1}{2}(y-y)^2 = 0$$

$$\Rightarrow d(y,\mu) = (y-\mu)^2$$

$v(\mu) = \frac{\partial \mu}{\partial \theta}$ $Var(Y) = \phi \cdot v(\mu)$ Normal: $v(\mu)=1$ Poisson: $v(\mu)=\mu$

Saddlepoint approximation

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

want $D^*(y,\hat{\mu}) \approx \chi^2_{n-(k+1)}$ need distribution of $d(Y_i,\mu_i)$

Dispersion model form: $b(y,\phi) \exp\left\{ -\frac{d(y,\mu)}{2\phi} \right\}$

Saddlepoint approximation: $b(y,\phi) \approx \frac{1}{\sqrt{2\pi\phi v(y)}}$

$\Rightarrow f(y;\mu,\phi) \approx \frac{1}{\sqrt{2\pi\phi v(y)}} \exp\left\{ -\frac{d(y,\mu)}{2\phi} \right\}$

If saddlepoint approximation holds, $\frac{d(Y_i,\mu_i)}{\phi} \approx \chi^2_1$ (5.4.3 in book)

$\Rightarrow D^*(y,\mu) = \frac{1}{\phi} \sum_{i=1} d(Y_i,\mu_i) \approx \chi^2_n$

$\Rightarrow D^*(y,\hat{\mu}) = \frac{1}{\phi} \sum_{i=1} d(Y_i,\hat{\mu}_i) \approx \chi^2_{n-(k+1)}$

Intuition: lose 1 df for each estimated parameter (Cochran's theorem)

Saddlepoint approximation

- Normal : "approximation" is exact
- Poisson : want $\min\{Y_i\} \geq 3$
- Bernoulli : approximation doesn't hold

$\Rightarrow$ no goodness of fit test for binary response

For Poisson data, if $\min\{Y_i\} < 3$ the GOF test tends to be conservative, so a small p-value still indicates lack of fit

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)

Goodness of fit

Scaled

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

$\phi = 1$

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?

- Need to add explanatory variables
- Poisson distribution is a bad choice

Offsets

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

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

note: no β!

Motivation for offsets

We can rewrite our regression model with the offset:

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

$\Rightarrow \log(\lambda_i) - \log(\text{Enroll}_i) = \beta_0 + \beta_1 MW_i + \dots$

$\Rightarrow \log\left(\frac{\lambda_i}{\text{Enrollment}_i}\right) = \beta_0 + \beta_1 MW_i + \dots$

$\frac{\lambda_i}{\text{Enrollment}_i}$ = expected <u>rate</u> of 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?

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 ϕ

Using $\hat{\phi}$

```r
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
...
```