Overdispersion

Recap: Overdispersion

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

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

Poisson regression: $\phi = 1$ (we use this to estimate $\text{Var}(\hat{\beta})$)

$\Rightarrow$ overdispersion means estimated variances & SEs are too small

Estimating ϕ

dispersion

Intuition: $Y_i \sim N(\mu_i, \sigma^2)$

$$\hat{\sigma}^2 = \frac{SSE}{n-p} = \frac{SSE}{n-(k+1)} = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

How to generalize this?

Option: Recall $d(y, \mu) = (y-\mu)^2$

$$\Rightarrow D(y, \hat{\mu}) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = SSE$$

$$\text{So } \hat{\sigma}^2 = \frac{D(y, \hat{\mu})}{n-(k+1)}$$

Generalize this:

$$\hat{\phi} = \frac{D(y, \hat{\mu})}{n-(k+1)}$$

(mean deviance estimate)

Saddlepoint approximation

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

Log likelihood: $\ell(\mu,\phi) \approx \sum_{i=1}^{n} \left\{ -\frac{1}{2}\log(2\pi\phi V(y_i)) - \frac{1}{2\phi} d(y_i,\mu_i) \right\}$

$$\frac{\partial \ell(\mu,\phi)}{\partial \phi} = \sum_{i=1}^{n} \left\{ -\frac{1}{2}\cdot\frac{1}{\phi} + \frac{1}{2\phi^2} d(y_i,\mu_i) \right\} \overset{\text{set}}{=} 0$$

$$\Rightarrow \sum_{i=1}^{n} -\frac{1}{2}\cdot\frac{1}{\phi} + \frac{1}{2\phi^2}\sum_{i=1}^{n} d(y_i,\mu_i) = 0$$

$$\Rightarrow -\frac{n}{2}\cdot\frac{1}{\phi} + \frac{1}{2\phi^2}\sum_i d(y_i,\mu_i) = 0 \Rightarrow \frac{n}{\phi} = \frac{1}{\phi^2}\sum_{i=1}^{n} d(y_i,\mu_i)$$

$$\Rightarrow \hat{\phi}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} d(y_i,\mu_i) = \frac{D(y,\mu)}{n}$$

we don't know μ

$$\Rightarrow \hat{\phi}_{MLE} = \frac{D(y,\hat{\mu})}{n} \quad \text{(plug in } \hat{\mu}\text{)}$$

But, this is biased

As usual with variance estimates:

$$\boxed{\hat{\phi} = \frac{D(y,\hat{\mu})}{n-(M+1)}}$$

mean deviance estimate

$$\hat{\sigma}^2 = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad V(\mu_i) = 1$$

Option 2: $$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{1} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

Pearson residuals: $\dfrac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$

intuition: account for mean-variance relationship in residuals

$$\Rightarrow \hat{\phi} = \frac{1}{n-(k+1)} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$ ← sum of squared pearson residuals

↑ Pearson estimate

Ex: Poisson regression ($V(\mu_i) = \mu_i$)

Pearson estimate $= \frac{1}{n-(k+1)} \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i}$

Motivation: $\text{var}(y_i) = \phi V(\mu_i)$

Suppose μ_i is known $\Rightarrow \hat{\phi} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)}$

$$E[\hat{\phi}] = E\left[\frac{(y_i - \mu_i)^2}{V(\mu_i)}\right] = \frac{\text{var}(y_i)}{V(\mu_i)} = \phi$$

which estimate $\hat{\phi}$ should we use?

- the mean deviance estimate depends on the saddlepoint approximation
- the Pearson estimate is approximately unbiased if μ & $V(\mu)$ are correct — we just need the first two moments

=> Pearson estimate is more "robust"

- R typically uses the Pearson estimate

Using $\hat{\phi}$

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

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

$\hat{\phi} = 7.59$ (Pearson estimate)

$>> 1$

```
...
##              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
...
```

$\hat{\beta}_S$

CI for β_S: $0.209 \pm z_{1-\frac{\alpha}{2}} (\sqrt{7.59})(0.186)$

$\hat{\beta}_S$ $\qquad$ $\sqrt{\hat{\phi}}$

Quasi-Poisson regression

A model for overdispersed Poisson-like counts, using an estimated dispersion parameter $\widehat{\phi}$, is called a *quasi-Poisson* model.

```
m3 <- glm(nv ~ region, offset = log(enroll1000),
          data = crimes, family = quasipoisson)
summary(m3)
```

already include $\sqrt{\hat{\phi}}$

```
...
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.30445    0.34161  -3.818 0.000274 ***
## regionMW     0.09754    0.48893   0.199 0.842417
## regionNE     0.76268    0.42117   1.811 0.074167 .
## regionSE     0.87237    0.42175   2.068 0.042044 *
## regionSW     0.50708    0.50973   0.995 0.323027
...
```

$\hat{\phi} \simeq \chi^2$

using canonical link:

$U(\beta) = \frac{X^T(Y-\mu)}{\phi} \overset{\text{set}}{=} 0$

Solving $X^T(Y-\mu) = 0$

Poisson vs. quasi-Poisson

Wald test (assume $\phi=1$)

Poisson:

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

same $\hat{\beta}$s

$SE(\text{quasiPoisson}) = \sqrt{\hat{\phi}}\, SE(\text{Poisson})$

e.g., $0.421 = \sqrt{7.59}\,(0.153)$

Quasi-Poisson:

```
...
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.30445    0.34161  -3.818 0.000274 ***
## regionMW     0.09754    0.48893   0.199 0.842417
## regionNE     0.76268    0.42117   1.811 0.074167 .
...
```

$\frac{\hat{\beta}-0}{SE(\hat{\beta})}$

increases p-values

Quasi-likelihood models

what is the Quasi-Poisson model?

- $v(\mu_i) = \mu_i$ $\quad \text{Var}(Y_i) = \phi \mu_i$
- mean-variance relationship uniquely determines EDM

  $\Rightarrow$ the only EDM with $v(\mu_i) = \mu_i$ is a Poisson

Poisson: $\phi = 1$

$\Rightarrow$ Quasi-Poisson doesn't correspond to an EDM