Quasi-Poisson models

Recap: Quasi-Poisson regression

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

```
m1 <- glm(nv ~ region, offset = log(enroll1000),
          data = crimes, family = quasipoisson)
summary(m1)
```

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

Recap: Poisson vs. quasi-Poisson

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 ***
...
```

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 .
...
```

Quasi-likelihood models

Pros and cons of quasi-Poisson

Pros:

- Estimated coefficients are the same as the Poisson model
- Just need to get μ and $V(\mu)$ correct
- Easy to use and interpret estimated dispersion $\widehat{\phi}$

Cons: Uses a quasi-likelihood, not a full likelihood. So we don't get

- AIC or BIC (these require log-likelihood)
- Quantile residuals (these require a defined CDF)

Inference with quasi-Poisson models

```
m1 <- glm(nv ~ region, offset = log(enroll1000),
          data = crimes, family = quasipoisson)
summary(m1)
```

```
...
##             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
## regionW      0.20934    0.51242   0.409 0.684055
...
```

> How can we test whether there is a difference between crime rates for Western and Central schools?

t-tests for single coefficients

Inference with quasi-Poisson models

```
m1 <- glm(nv ~ region, offset = log(enroll1000),
          data = crimes, family = quasipoisson)
summary(m1)
```

```
...
##             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
## regionW      0.20934    0.51242   0.409 0.684055
...
```

> How can we test whether there is any relationship between Region and crime rates?

F-tests for multiple coefficients

F-test example

F-test example

```
m1 <- glm(nv ~ region, offset = log(enroll1000),
          data = crimes, family = quasipoisson)
m0 <- glm(nv ~ 1, offset = log(enroll1000),
          data = crimes, family = quasipoisson)

deviance_change <- m0$deviance - m1$deviance
df_numerator <- m0$df.residual - m1$df.residual
numerator <- deviance_change/df_numerator
denominator <- m1$deviance/m1$df.residual

numerator/denominator
```

```
## [1] 2.003533
```

```
pf(numerator/denominator,  df_numerator,
   m1$df.residual, lower.tail=F)
```

```
## [1] 0.0878041
```

An alternative to quasi-Poisson

Poisson:

- Mean = λ_i
- Variance = λ_i

quasi-Poisson:

- Mean = λ_i
- Variance = $\phi\lambda_i$
- Variance is a linear function of the mean

> What if we want variance to depend on the mean in a different way?

The negative binomial distribution

If $Y_i \sim NB(\theta, p)$, then Y_i takes values $y = 0, 1, 2, 3, \ldots$ with probabilities

$$P(Y_i = y) = \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)}(1-p)^\theta p^y$$

- $\theta > 0, \quad p \in [0, 1]$
- $\mathbb{E}[Y_i] = \dfrac{p\theta}{1-p} = \mu$
- $Var(Y_i) = \dfrac{p\theta}{(1-p)^2} = \mu + \dfrac{\mu^2}{\theta}$
- Variance is a *quadratic* function of the mean

Mean and variance for a negative binomial variable

If $Y_i \sim NB(\theta, p)$, then

- $\mathbb{E}[Y_i] = \dfrac{p\theta}{1-p} = \mu$
- $Var(Y_i) = \dfrac{p\theta}{(1-p)^2} = \mu + \dfrac{\mu^2}{\theta}$

> How is θ related to overdispersion?

Negative binomial regression

$$Y_i \sim NB(\theta,\ p_i)$$

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

- $\mu_i = \dfrac{p_i \theta}{1 - p_i}$
- Note that θ is the same for all i
- Note that just like in Poisson regression, we model the average count
  - Interpretation of βs is the same as in Poisson regression

In R

```r
library(MASS)
m3 <- glm.nb(nv ~ region + offset(log(enroll1000)),
             data = crimes)
```

```
...
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.33404    0.28137  -4.741 2.12e-06 ***
## regionMW     0.14230    0.44824   0.317  0.75089
## regionNE     0.94567    0.36641   2.581  0.00985 **
## regionSE     1.18534    0.39736   2.983  0.00285 **
## regionSW     0.33449    0.45666   0.732  0.46387
## regionW      0.06466    0.47628   0.136  0.89201
##
## (Dispersion parameter for Negative Binomial(1.0662) fami
...
```

$\hat{\theta} = 1.066$