Negative binomial regression

- Reminders:
    - No class on Friday
        - I have posted Lab 6 on the course website

- Today:
    - Negative binomial regression
    - Some time to work on lab

Last time: handling overdispersion

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?

Introducing the negative binomial

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

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

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

Mean and variance for a negative binomial variable

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

- Mean = $\dfrac{p\theta}{1-p} = \mu$
- Variance = $\dfrac{p\theta}{(1-p)^2} = \mu + \dfrac{\mu^2}{\theta}$

> How is θ related to overdispersion?

Large θ: mean $\approx$ variance

Small θ: variance $\gg$ mean

Negative binomial regression

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

$$\log(\mu_i) = \beta_0 + \beta_1 X_i$$

- $\mu_i = \dfrac{p_i \theta}{1 - p_i}$
- 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

Comparing Poisson, quasi-Poisson, negative binomial

Poisson:

- Mean = λ_i
- Variance = λ_i

quasi-Poisson:

- Mean = λ_i
- Variance = $\phi\lambda_i$

negative binomial:

- Mean = μ_i
- Variance = $\mu_i + \frac{\mu_i^2}{\theta}$

In R

library (MASS)

```
m3 <- glm.nb(art ~ ., data = articles)
```

```
...
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.256144   0.137348   1.865 0.062191 .
## femWomen    -0.216418   0.072636  -2.979 0.002887 **
## marMarried   0.150489   0.082097   1.833 0.066791 .
## kid5        -0.176415   0.052813  -3.340 0.000837 ***
## phd          0.015271   0.035873   0.426 0.670326
## ment         0.029082   0.003214   9.048  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## (Dispersion parameter for Negative Binomial(2.2644) fami
...
```

$\hat{\theta} = 2.264$

In R

```
...
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.256144   0.137348   1.865 0.062191 .
## femWomen     -0.216418   0.072636  -2.979 0.002887 **
## marMarried    0.150489   0.082097   1.833 0.066791 .
## kid5         -0.176415   0.052813  -3.340 0.000837 ***
## phd           0.015271   0.035873   0.426 0.670326
## ment          0.029082   0.003214   9.048  < 2e-16 ***
...
```

> How do I interpret the estimated coefficient -0.176?

One additional kid under age 6 is associated with a change in the average # of articles published by a factor of $e^{-0.176} = 0.84$, holding other variables constant

quasi-Poisson vs. negative binomial

quasi-Poisson:

- linear relationship between mean and variance
- easy to interpret $\widehat{\phi}$
- same as Poisson regression when $\phi = 1$
- simple adjustment to estimated standard errors
- estimated coefficients same as in Poisson regression

negative binomial:

- quadratic relationship between mean and variance
- we get to use a likelihood, rather than a quasi-likelihood
- Same as Poisson regression when θ is very large and p is very small

> For this class, either is appropriate to handle overdispersion.