Negative binomial regression

Warm up: class activity

https://sta712-f22.github.io/class_activities/ca_lecture_26.html

H_0: model is a good fit

H_A: model is not a good fit

Under H_0: $D^*(y, \hat{\mu}) \approx \chi^2_{n-(k+1)}$ $\quad n-(k+1) = 2004$

$D^*(y, \hat{\mu}) = 11540$

pchisq(11540, df = 2004, lower.tail=F) ≈ 0

$$\hat{\phi} = \frac{D(y, \hat{\mu})}{n-(k+1)} = 5.76$$

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

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

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

Interpretation: Y = # of successes until r "failures", for independent trials w/ probability p of success

If r is large, $var(Y_i) \approx \mathbb{E}[Y_i]$

If r is small, $var(Y_i) >> \mathbb{E}[Y_i]$

Negative binomial regression

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

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

↑ not the canonical link function

- $\mu_i = \dfrac{p_i r}{1 - p_i}$
- Note that r 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

If r is known, then NB is an EDM

If r is unknown, NB is not an EDM

In R

```
library(MASS)
m2 <- glm.nb(cigsPerDay ~ male + age + education +
               diabetes + BMI, data = smokers)
```

Don't need to specify family

```
...
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.877771   0.123477  23.306  < 2e-16 ***
## male         0.459148   0.027641  16.611  < 2e-16 ***
## age         -0.007010   0.001731  -4.050 5.12e-05 ***
## education2   0.024518   0.032534   0.754    0.451
## education3   0.009252   0.040802   0.227    0.821
## education4  -0.027732   0.044825  -0.619    0.536
##
## (Dispersion parameter for Negative Binomial(3.2981) fami
...
```

$\hat{r} = 3.3$

Negative binomial as an EDM

$$f(y;r,p) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} (1-p)^r p^y$$

$$= \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \exp\{ r\log(1-p) + y\log p \}$$

$$= \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \exp\{ y\log p - (-r\log(1-p)) \}$$

$$\theta = \log p = \log\left(\frac{\mu}{\mu+r}\right)$$

$$\frac{pr}{1-p} = \mu$$

$$\Rightarrow p = \frac{\mu}{\mu+r}$$

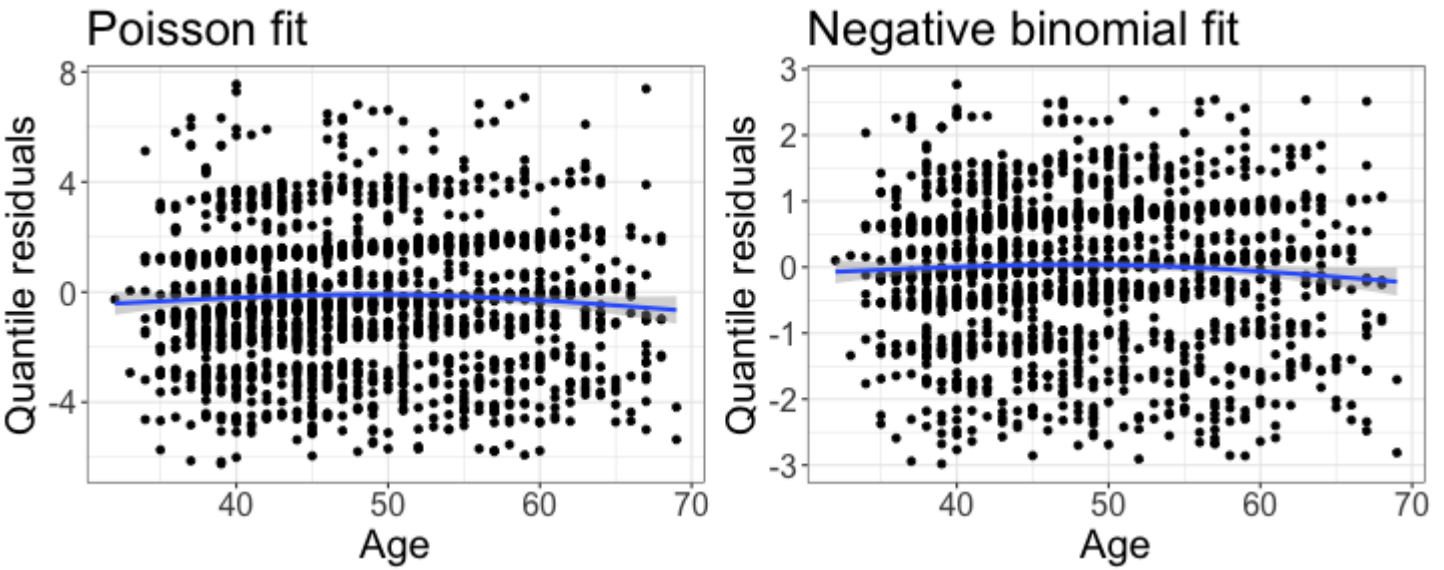
$$\kappa(\theta) = -r\log(1-p)$$

$$= -r\log(1-e^{\theta})$$

$$\phi = 1 \qquad a(y,\phi) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)}$$

So, NB is an EDM for known r

Poisson vs. negative binomial fits



Inference with negative binomial models

```
...
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.877771   0.123477  23.306  < 2e-16 ***
## male         0.459148   0.027641  16.611  < 2e-16 ***
## age         -0.007010   0.001731  -4.050 5.12e-05 ***
## education2   0.024518   0.032534   0.754    0.451
## education3   0.009252   0.040802   0.227    0.821
## education4  -0.027732   0.044825  -0.619    0.536
## diabetes    -0.010124   0.099126  -0.102    0.919
## BMI          0.003693   0.003573   1.033    0.301
...
```

> How would I test whether there is a relationship between age and the number of cigarettes smoked, after accounting for other variables?

Inference with negative binomial models

```
...
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.877771   0.123477  23.306  < 2e-16 ***
## male         0.459148   0.027641  16.611  < 2e-16 ***
## age         -0.007010   0.001731  -4.050 5.12e-05 ***
## education2   0.024518   0.032534   0.754    0.451
## education3   0.009252   0.040802   0.227    0.821
## education4  -0.027732   0.044825  -0.619    0.536
## diabetes    -0.010124   0.099126  -0.102    0.919
## BMI          0.003693   0.003573   1.033    0.301
...
```

> How would I test whether there is a relationship between education and the number of cigarettes smoked, after accounting for other variables?

Likelihood ratio test

```
m2 <- glm.nb(cigsPerDay ~ male + age + education +
               diabetes + BMI, data = smokers)
m3 <- glm.nb(cigsPerDay ~ male + age +
               diabetes + BMI, data = smokers)
m2$twologlik - m3$twologlik
```

```
## [1] 1.423055
```

```
pchisq(1.423, df=3, lower.tail=F)
```

```
## [1] 0.7001524
```