Quasi-Poisson and negative binomial regression

Warm-up: Class activity, Part I

https://sta214-s23.github.io/class_activities/ca_lecture_23.html

Class activity

```
...
##     Null deviance: 13471  on 2011  degrees of freedom
## Residual deviance: 11540  on 2004  degrees of freedom
...
```

Goodness-of-fit test:

Class activity

```
...
## (Dispersion parameter for quasipoisson family taken to be 5.519388)
##
##     Null deviance: 13471  on 2011  degrees of freedom
## Residual deviance: 11540  on 2004  degrees of freedom
## AIC: NA
...
```

What is the estimated dispersion parameter $\widehat{\phi}$?

Nested tests with quasi-Poisson models

```
...
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.837192   0.115099  24.650  < 2e-16 ***
## male          0.456170   0.026095  17.481  < 2e-16 ***
## age          -0.006806   0.001595  -4.267 2.07e-05 ***
## education2    0.016463   0.029675   0.555    0.579
## education3    0.016427   0.037706   0.436    0.663
## education4   -0.015033   0.040477  -0.371    0.710
## diabetes     -0.025398   0.092606  -0.274    0.784
## BMI           0.005001   0.003304   1.513    0.130
...
```

Research question: Is there a relationship between education level and the number of cigarettes smoked per day, after accounting for sex, age, diabetes, and BMI?

> What are my null and alternative hypotheses for this research question?

Nested tests with quasi-Poisson models

Research question: Is there a relationship between education level and the number of cigarettes smoked per day, after accounting for sex, age, diabetes, and BMI?

- In Poisson regression, we would use a likelihood ratio test
- However, the quasi-Poisson model includes the estimated dispersion, $\widehat{\phi}$. We need to use an **F-test** instead

Nested tests with quasi-Poisson models

```
m1 <- glm(cigsPerDay ~ male + age + education +
            diabetes + BMI,
          data = smokers, family = quasipoisson)
m2 <- glm(cigsPerDay ~ male + age + diabetes + BMI,
          data = smokers, family = quasipoisson)

m1$deviance
```

```
## [1] 11539.56
```

```
m2$deviance
```

```
## [1] 11543.84
```

```
m1$df.residual
```

```
## [1] 2004
```

```
m2$df.residual
```

```
## [1] 2007
```

Nested tests with quasi-Poisson models

```
m1$deviance
```

```
## [1] 11539.56
```

```
m2$deviance
```

```
## [1] 11543.84
```

```
pf(0.258, 3, 2004, lower.tail=F)
```

```
## [1] 0.8556648
```

Alternative to quasi-Poisson: 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

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

In R

```r
library(MASS)
m3 <- glm.nb(cigsPerDay ~ male + age + education +
               diabetes + BMI, data = smokers)
```

```
...
## Coefficients:
##              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
##
## (Dispersion parameter for Negative Binomial(3.2981) family taken to be 1)
...
```

Fitted model

```
...
## Coefficients:
##              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    
## 
## (Dispersion parameter for Negative Binomial(3.2981) family taken to be 1)
...
```

How do I interpret the estimated coefficient -0.007?

Hypothesis testing

```
...
## Coefficients:
##              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
##
## (Dispersion parameter for Negative Binomial(3.2981) family taken to be 1)
...
```

Research question: Is there a relationship between education level and the number of cigarettes smoked per day, after accounting for sex, age, diabetes, and BMI?

Likelihood ratio test

```
m3 <- glm.nb(cigsPerDay ~ male + age + education +
                diabetes + BMI, data = smokers)
m4 <- glm.nb(cigsPerDay ~ male + age +
                diabetes + BMI, data = smokers)
m3$twologlik - m4$twologlik
```

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

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

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

Comparing Poisson, quasi-Poisson, negative binomial

Poisson:

- Mean = λ_i
- Variance = λ_i

quasi-Poisson:

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

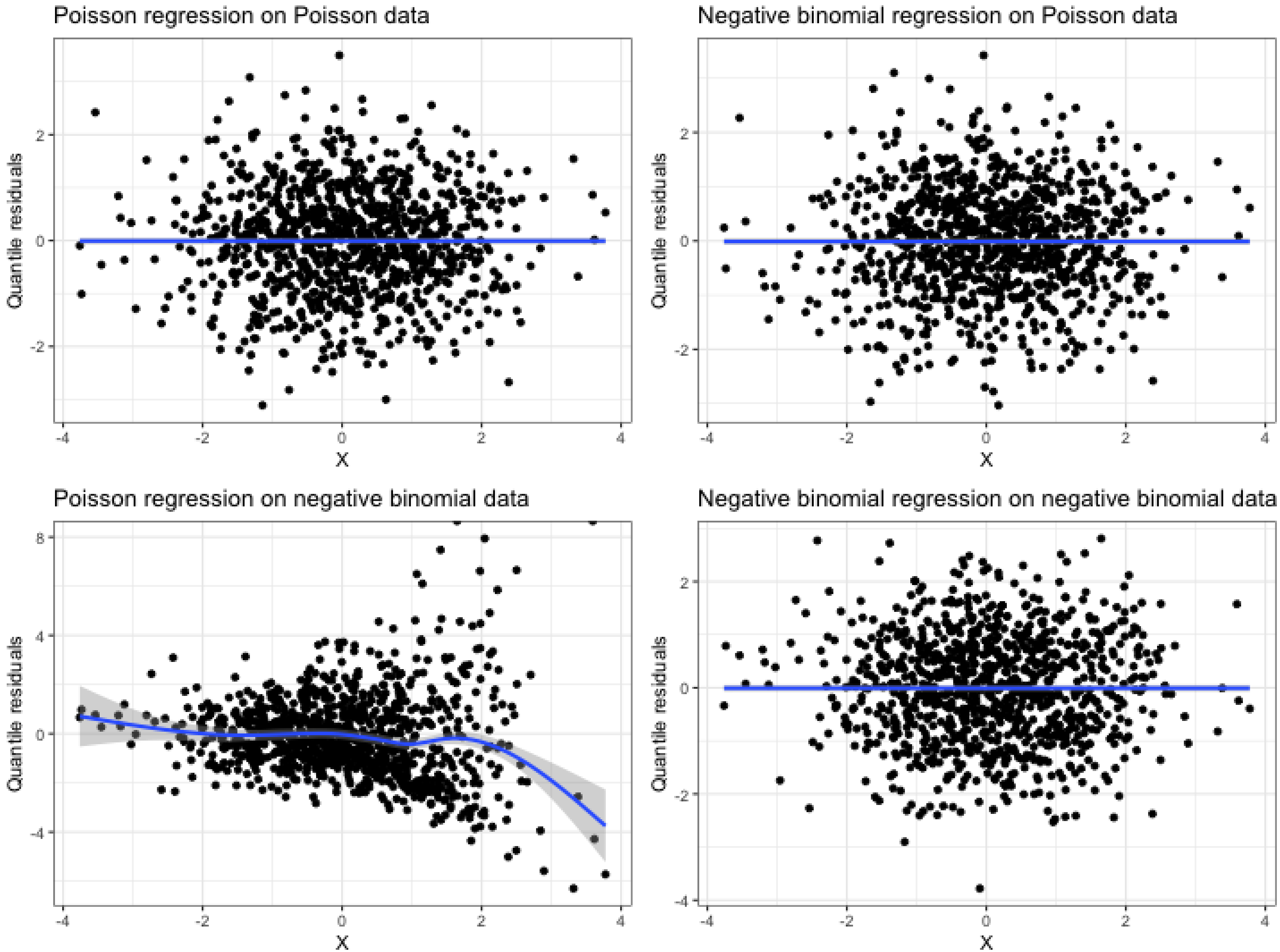
negative binomial:

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

Class activity, Part II

https://sta214-s23.github.io/class_activities/ca_lecture_23.html

Class activity



Choosing a count model with quantile residual plots

- If the residuals have **constant variance** and mostly fall between **-2 and 2**: Poisson is reasonable
- If the residuals have **constant variance** but many residuals are > 2 or < -2: use either quasi-Poisson or negative binomial
- If the residuals have **non-constant variance**: use negative binomial

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
- t-tests and F-tests

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
- Wald tests and likelihood ratio tests