

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  
...  
n-p
```

Goodness-of-fit test:

H_0 : the model is a good fit to the data

H_A : the model is not a good fit to the data

Test statistic: (residual) deviance

under H_0 : deviance $\sim \chi^2_{n-p}$

11540 χ^2_{2004}

$n = \# \text{ obs}$

$p = \# \text{ parameters}$

p-value ≈ 0

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 $\hat{\phi}$?

$$\hat{\phi} = 5.52$$

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?

$$H_0: \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_A: \text{at least one of } \beta_3, \beta_4, \beta_5 \neq 0$$

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

linear regression: $F = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}}) / (\text{df}_{\text{red.}} - \text{df}_{\text{full}})}{SSE_{\text{full}} / \text{df}_{\text{full}}}$

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, $\hat{\phi}$. We need to use an **F-test** instead

Intuition: estimating ϕ changes the distribution of our test statistic

LRT: $\text{deviance}_{\text{reduced}} - \text{deviance}_{\text{full}} \sim \chi^2_{\# \text{parameters tested in model}}$

F-test: $\frac{(\text{deviance}_{\text{reduced}} - \text{deviance}_{\text{full}}) / (\text{df}_{\text{red.}} - \text{df}_{\text{full}})}{\hat{\phi}_{\text{full}}}$

Under H_0 , $F \sim F_{\text{df}_{\text{reduced}} - \text{df}_{\text{full}}, \text{df}_{\text{full}}}$

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)
```

\leftarrow full model
 \leftarrow reduced model

(m1\$deviance)

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

(m2\$deviance)

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

m1\$df.residual

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

m2\$df.residual

$$F = \frac{(11543.84 - 11539.56) / (2007 - 2004)}{5.52} = 0.258$$

$\uparrow \hat{\phi}_{full}$

Under H_0 , $F \sim F_{3, 2004}$

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

Nested tests with quasi-Poisson models

```
m1$deviance
```

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

```
m2$deviance
```

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

test stat. ↓ denominator df (= df for full model)
pf(0.258, 3, 2004, lower.tail=F)
 ↑ numerator df (= # parameters tested)

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

quasi-Poisson

- t - tests for 1 coefficient
- t distribution for CIs
- F - tests for multiple coefficients

Alternative to quasi-Poisson: negative binomial

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

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

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

Negative binomial regression

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

(random component)
(different to Poisson)

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

↑ mean of γ_i

(systematic component)
(Same as Poisson)

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

In R

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

(don't have to specify family)

...

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)

...

$$\Rightarrow \text{Variance} = \hat{\mu} + \frac{\hat{\mu}^2}{\hat{\theta}} \quad \hat{\theta} = 3.3 \quad \emptyset = 1$$

$$\log(\text{mean}) = \beta_0 + \beta_1 X_1 + \dots$$

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?

A one-year increase in age is associated with a decrease in the average # of cigarettes smoked per day by a factor of $e^{-0.007}$

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?

Don't use deviance for LRT with NB

Likelihood ratio test

full model

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

reduced model

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

$$\underbrace{m3\$twologlik}_{Z \log L_{\text{full}}} - \underbrace{m4\$twologlik}_{Z \log L_{\text{reduced}}} = LRT$$

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

$$LRT : G = Z \log L_{\text{full}} - Z \log L_{\text{reduced}}$$

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

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

↑ # parameters tested

Comparing Poisson, quasi-Poisson, negative binomial

Poisson:

- + Mean = λ_i
- + Variance = λ_i

quasi-Poisson:

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

negative binomial:

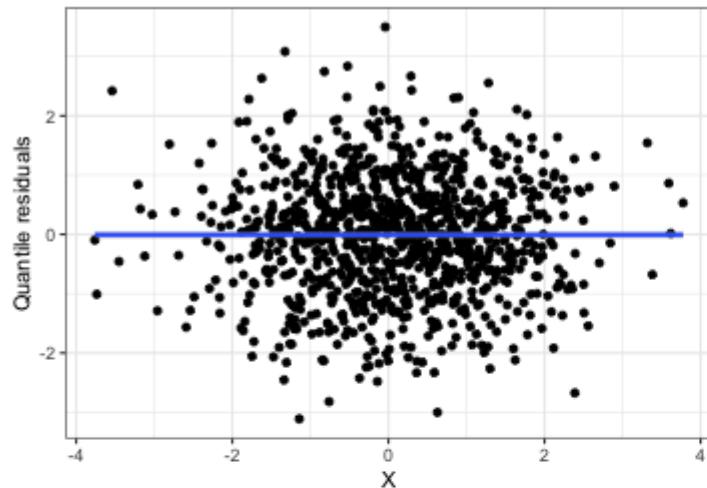
- + Mean = μ_i
- + Variance = $\mu_i + \frac{\mu_i^2}{\theta}$

Class activity, Part II

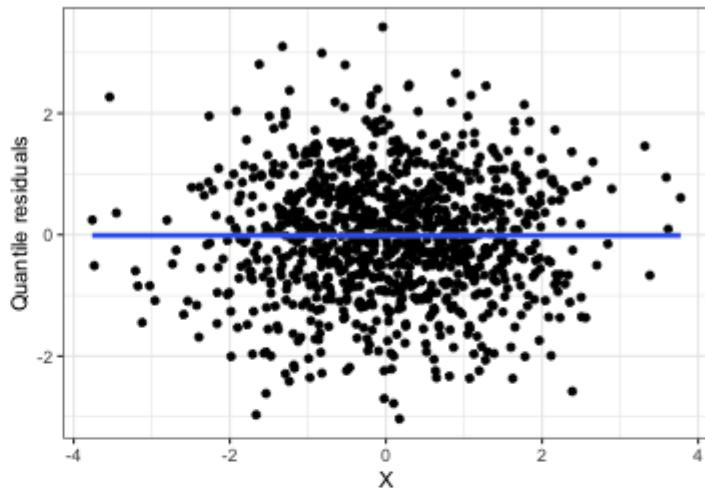
https://sta214-s23.github.io/class_activities/ca_lecture_23.html

Class activity

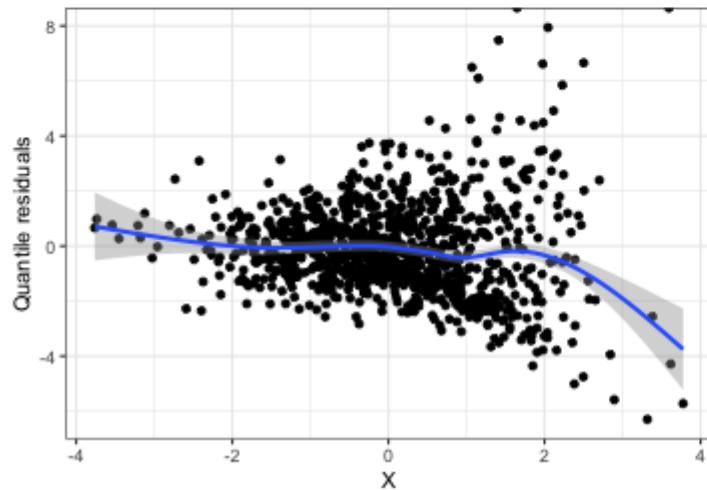
Poisson regression on Poisson data



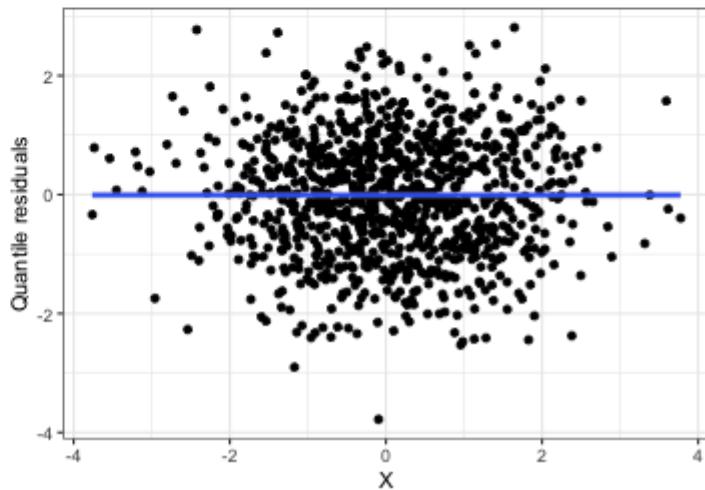
Negative binomial regression on Poisson data



Poisson regression on negative binomial data



Negative binomial regression on negative binomial data



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 $\hat{\phi}$
- + 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