Lecture 17

Inference with quasi-Poisson models

$$(n-p)\frac{\hat{\phi}}{\phi} \approx \chi^2_{n-p}$$

$$F_{d_1,d_2} = \frac{V_1/d_1}{V_2/d_2} \qquad V_1 \sim \chi^2_{d_1} \quad V_2 \sim \chi^2_{d_2} \quad V_1 \perp\!\!\!\perp V_2$$

$$\hat{Var}(\hat{\beta})_{QP} = \hat{\phi}\, Var(\hat{\beta})_{Poisson}$$

$$T \sim t_{n-p} \Rightarrow T^2 \sim F_{1,n-p}$$

Test: $H_0: \beta_j = 0$ vs. $H_A: \beta_j \neq 0$

Test stat: $\frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \sim N(0,1)$ or $\frac{(\hat{\beta}_j - 0)^2}{Var(\hat{\beta}_j)} \sim \chi^2_1$

Test for QP:

$$\frac{(\hat{\beta}_j - 0)^2}{\hat{\phi}\,\hat{Var}(\hat{\beta}_j)_{Poisson}} = \underbrace{\frac{(\hat{\beta}_j - 0)^2}{\phi\,\hat{Var}(\hat{\beta}_j)_{Poisson}}}_{\chi^2_1} \cdot \underbrace{\frac{\phi}{\hat{\phi}}}_{\frac{1}{(\chi^2_{n-p})/(n-p)}} \approx F_{1,n-p}$$

$$\Rightarrow \frac{\hat{\beta}_j - 0}{\sqrt{\hat{\phi}}\, SE(\hat{\beta}_j)_{Poisson}} \approx t_{n-p}$$

LRT: $\dfrac{D(Y, \hat{\mu}_{red}) - D(Y, \hat{\mu}_{full})}{\phi} \approx \chi^2_q$

q = # parameters tested $= df_{red} - df_{full}$

$$\frac{\left(D(Y, \hat{\mu}_{red}) - D(Y, \hat{\mu}_{full})\right) / q}{\hat{\phi}_{full}}$$

$$= \underbrace{\frac{\left(D(Y, \hat{\mu}_{red}) - D(Y, \hat{\mu}_{full})\right)/q}{\phi}}_{\chi^2_q \,/ q} \cdot \underbrace{\frac{\phi}{\hat{\phi}}}_{\frac{1}{\chi^2_{n-p_{full}}/(n-p_{full})}}$$

$\approx F_{q,\, n-p_{full}}$

(intuition: nested F test for linear regression

$$F = \frac{(SSE_{red} - SSE_{full})/q}{(SSE_{full})/(n - p_{full})} \approx F_{q,\, n-p_{full}}$$
)

An alternative to quasi-Poisson

Poisson:

- Mean = λ_i
- Variance = λ_i

quasi-Poisson:

- Mean = λ_i

$\phi \lambda_i$

- Variance = $\varphi\lambda_i$
- Variance is a linear function of the mean

Question: 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

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

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

For a fixed μ, $Var(Y_i) \to \mu$ as $r \to \infty$

Key:
- Count variable ($Y_i = 0, 1, 2, \ldots$)
- A lot more flexibility in the variance

Negative binomial regression

Poisson:

$Y_i \sim \text{Poisson}(\mu_i)$

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

canonical link function

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

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

<u>not</u> 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

In R

If r is known, then the NB is an EDM

if r is unknown, then the NB is not an EDM

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

(Don't need to specify family)

```
...
(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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(3.2981) family taken to be
1)
```

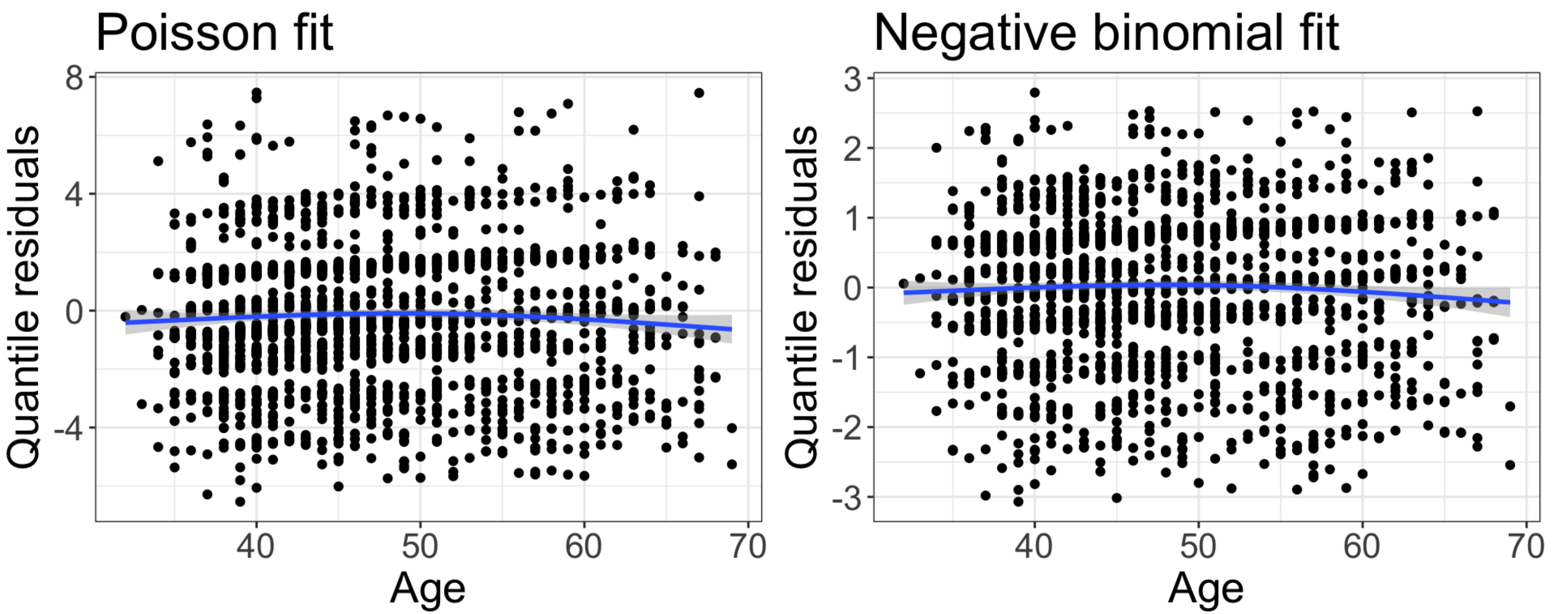
$\hat{r}$

$\hat{r} = 3.3$

$r \neq \phi$

For NB, $\phi = 1$

Poisson vs. negative binomial fits



Inference with negative binomial models

z $\quad$ $Pr(>|z|)$

```
...
(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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

...
```

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

$H_0: \underset{\uparrow \text{age}}{\beta_2} = 0$ $\qquad$ $H_A: \beta_2 \neq 0$ $\qquad$ p-value $\approx 5 \times 10^{-5}$

Inference with negative binomial models

```
...
(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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

...
```

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

LRT : reduced model has sex, age, diabetes, BMI
(no education)

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
```

LRT: $2(\log L_{full} - \log L_{reduced}) \approx \chi^2_q$

Class activity

https://sta712-f23.github.io/class_activities/ca_lecture_17.html