# Negative binomial regression

- Project 2 released this Friday, due before Thanksgiving break
- Choose either two challenges and one project, or one challenge and two projects
- Challenge 7 released on course website
- Exam 2 released next Friday (Nov 11)
  - covers Poisson regression, EDMs, overdispersion, quasi-Poisson, negative binomial, and ZIP models

# Recap: negative binomial regression

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

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

↑ not the canonical link

- $\mu_i = \dfrac{p_i r}{1 - p_i}$
- Note that $r$ is the same for all $i$
- Note that just like in Poisson regression, we model the average count
  - Interpretation of $\beta$s is the same as in Poisson regression

Poisson regression: $V(\mu) = \mu$

NB regression: $V(\mu) = \mu + \dfrac{\mu^2}{r}$

# In R

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

```
...
##                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
...
```

$\hat{r} = 3.3$

the average # of cigarettes someone with an advanced degree is $e^{-0.0277}$ times the average # of cigarettes for someone w/ a HS degree, holding other variables fixed

# Class activity

https://sta712-f22.github.io/class_activities/ca_lecture_27.html

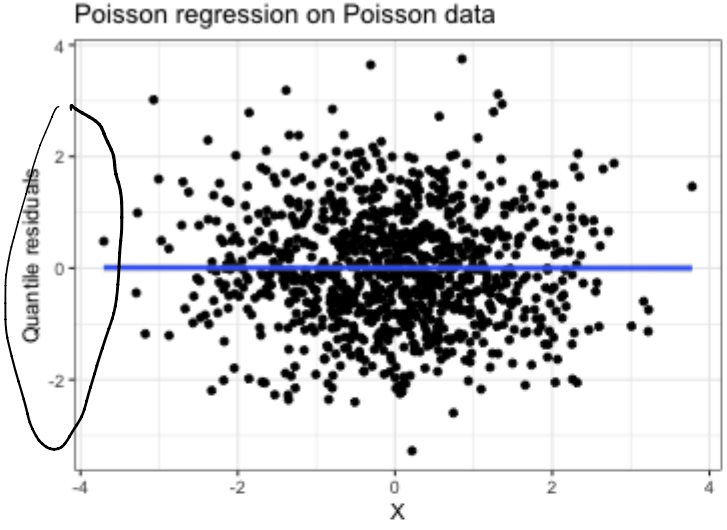
Poisson = NB( $r = \infty$

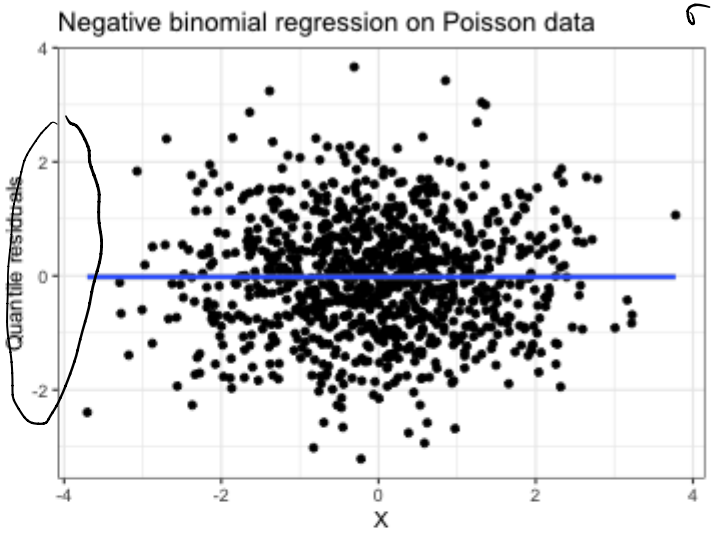
Looking for:
1) residuals $\in (-3, 3)$ (roughly)
2) constant variance
3) random scatter around 0

$\hat{r} \approx 7000$

# Class activity

residuals $\sim N(0, 1)$

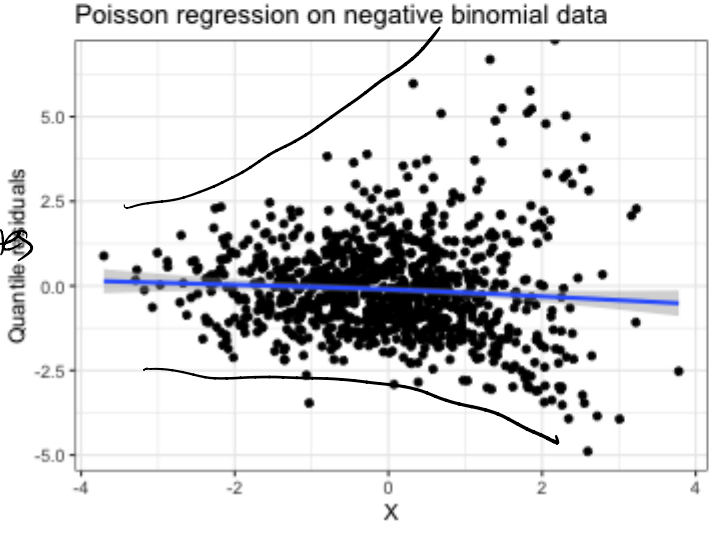


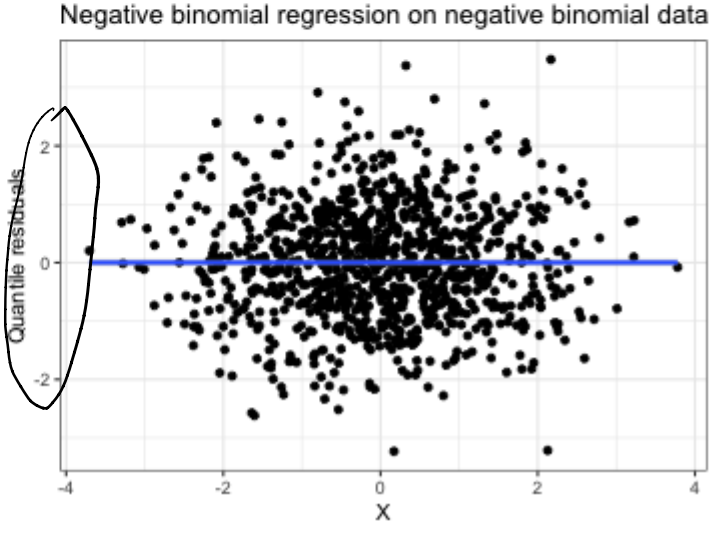




$\mu + \frac{\mu^2}{r}$

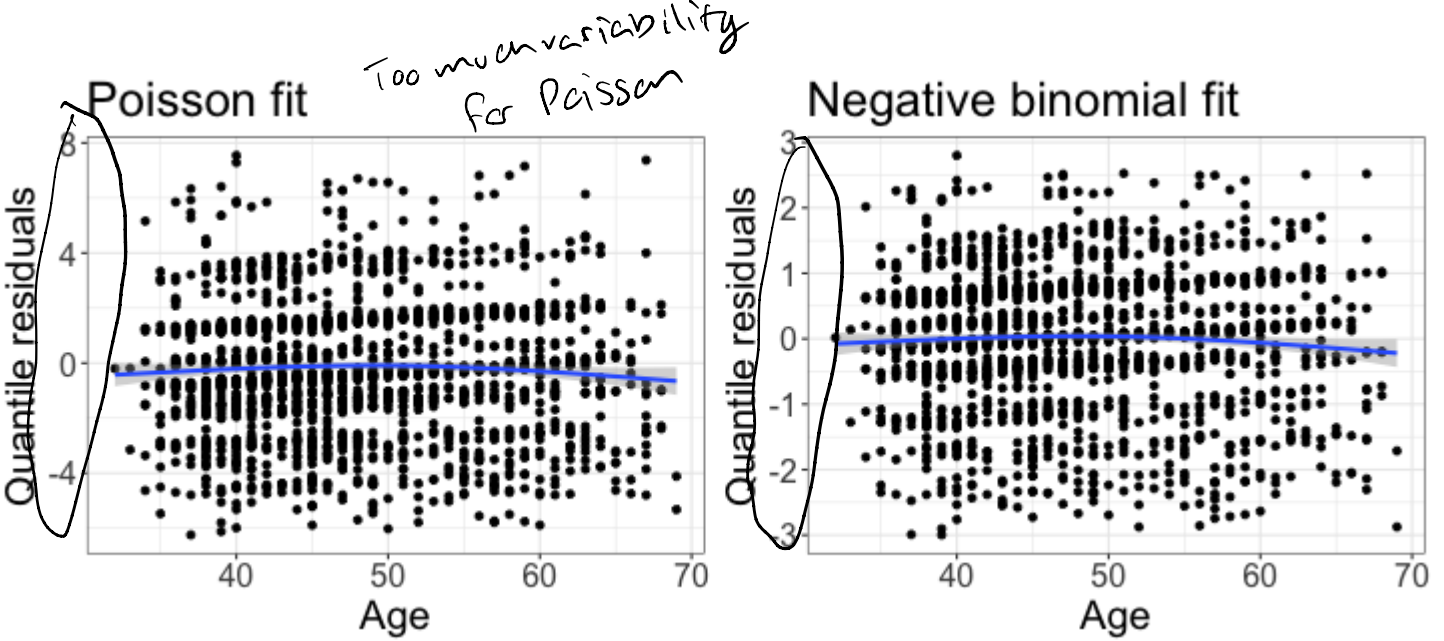
$x \uparrow$
$\Rightarrow \mu \uparrow$

$\Rightarrow$ Poisson underestimates Variance



# Poisson vs. negative binomial fits




- residual variance is too high for Poisson to be a good fit
- variance ≈ constant, so we could use either quasi-Poisson or NB

# Inference with negative binomial models

$-4.05 = \frac{-0.007}{0.00173}$

```
...
##                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
...
```

$H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ $Z = -4.05$

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

Wald test.

$Y|\lambda \sim \text{Poisson}(\lambda)$

$\lambda \sim \text{Gamma}\left(\frac{1}{\psi}, \mu\psi\right)$ $\quad (\Rightarrow \mathbb{E}[\lambda] = \mu)$

$\Rightarrow Y \sim NB\left(\frac{1}{\psi}, \frac{\mu}{\mu + \frac{1}{\psi}}\right)$

$\mathbb{E}[Y] = \mu$

$\frac{1}{\psi} \to \infty$

Point mass for $\lambda$ at $\mu$

$\Rightarrow Y|\lambda \sim \text{Poisson}(\mu)$

$$P(Y=y) = \int P(Y=y|\lambda)\, P(\lambda)\, d\lambda$$

# 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
```