Choosing count models

Class activity

https://sta214-s23.github.io/class_activities/ca_lecture_24.html

Goal: How can we use quantile residual plots to determine whether to fit a Poisson, quasi-Poisson, or negative binomial model?
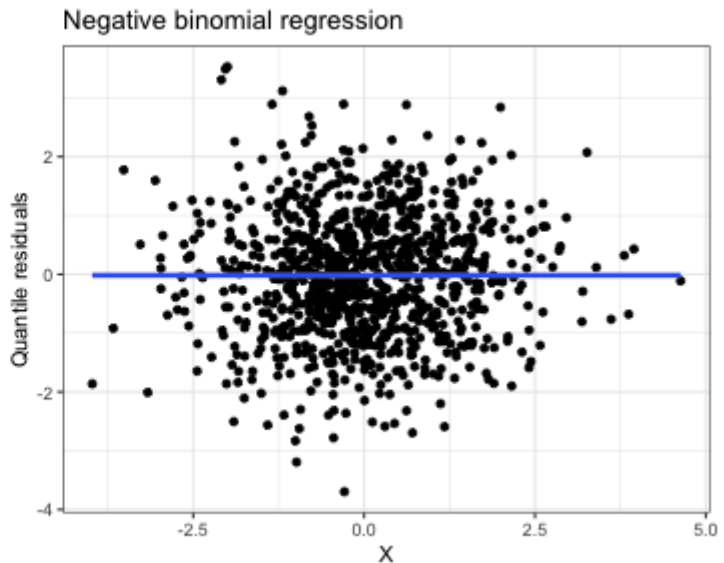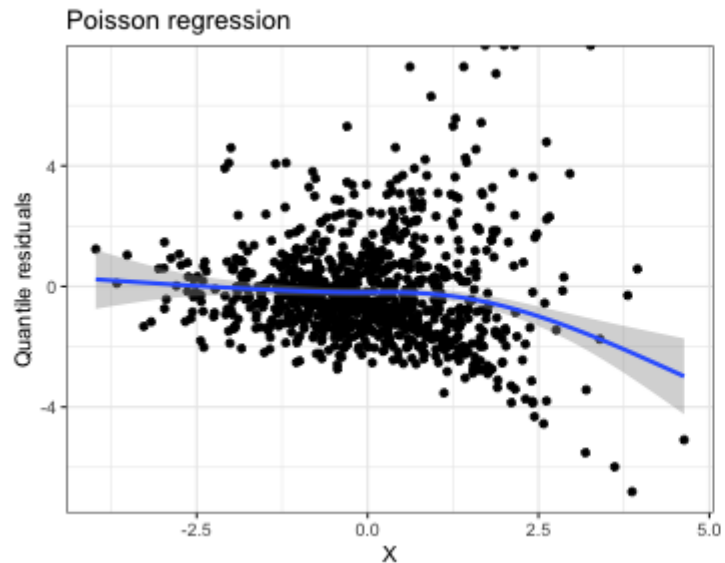
Instructions:

✚ With a neighbor, work throught the class activity

✚ You and your neighbor will then discuss the activity with another pair of students from the class

✚ Finally, we will discuss the activity as a group

Proposed guidelines

Class activity

Quantile residual plots for negative binomial data ($\mu_i = \exp(X_i), \theta = 0.5$):

Class activity

Data: negative binomial with $\mu_i = \exp(X_i), \theta = 0.5$

> How does CI coverage compare for Poisson, quasi-Poisson, and negative binomial models?

Class activity

Data: negative binomial with $\mu_i = \exp(X_i), \theta = 0.5$

Poisson coverage:

```r
n <- 1000
nsim <- 500
theta <- 0.5
contains_beta_poisson <- rep(0, nsim)
for(i in 1:nsim){
  x <- rnorm(n, sd = 1.2)
  y <- rnbinom(n, size=theta, mu=exp(x))

  m1 <- glm(y ~ x, family = poisson)

  upper <- summary(m1)$coefficients[2,1] +
    1.96*summary(m1)$coefficients[2,2]
  lower <- summary(m1)$coefficients[2,1] -
    1.96*summary(m1)$coefficients[2,2]

  contains_beta_poisson[i] <- upper > 1 && lower < 1
}
mean(contains_beta_poisson)
```

```
## [1] 0.278
```

Class activity

Data: negative binomial with $\mu_i = \exp(X_i), \theta = 0.5$

quasi-Poisson coverage:

```
contains_beta_quasipoisson <- rep(0, nsim)
for(i in 1:nsim){
  x <- rnorm(n, sd = 1.2)
  y <- rnbinom(n, size=theta, mu=exp(x))

  m2 <- glm(y ~ x, family = quasipoisson)

  upper <- summary(m2)$coefficients[2,1] +
    qt(0.025, n-2, lower.tail=F)*summary(m2)$coefficients[2,2]
  lower <- summary(m2)$coefficients[2,1] -
    qt(0.025, n-2, lower.tail=F)*summary(m2)$coefficients[2,2]

  contains_beta_quasipoisson[i] <- upper > 1 && lower < 1
}
mean(contains_beta_quasipoisson)
```

```
## [1] 0.586
```

Class activity

Data: negative binomial with $\mu_i = \exp(X_i), \theta = 0.5$

negative binomial coverage:

```r
contains_beta_nb <- rep(0, nsim)
for(i in 1:nsim){
  x <- rnorm(n, sd = 1.2)
  y <- rnbinom(n, size=theta, mu=exp(x))

  m3 <- glm.nb(y ~ x)

  upper <- summary(m3)$coefficients[2,1] +
    1.96*summary(m3)$coefficients[2,2]
  lower <- summary(m3)$coefficients[2,1] -
    1.96*summary(m3)$coefficients[2,2]

  contains_beta_nb[i] <- upper > 1 && lower < 1
}
mean(contains_beta_nb)
```
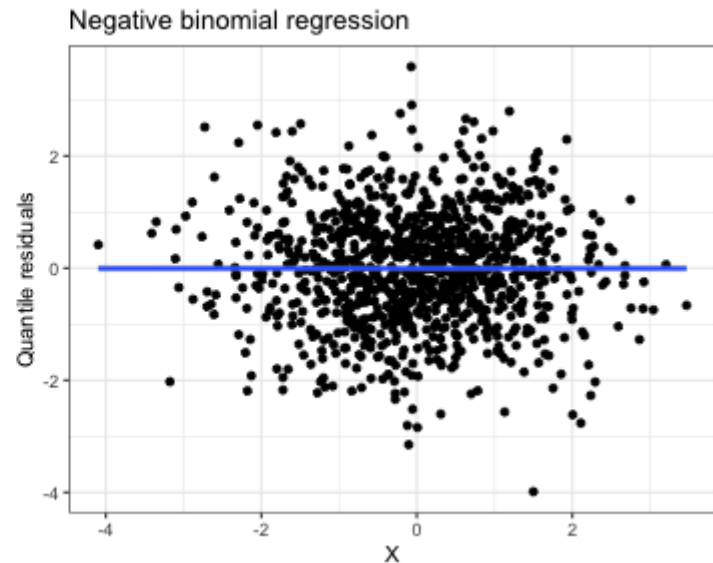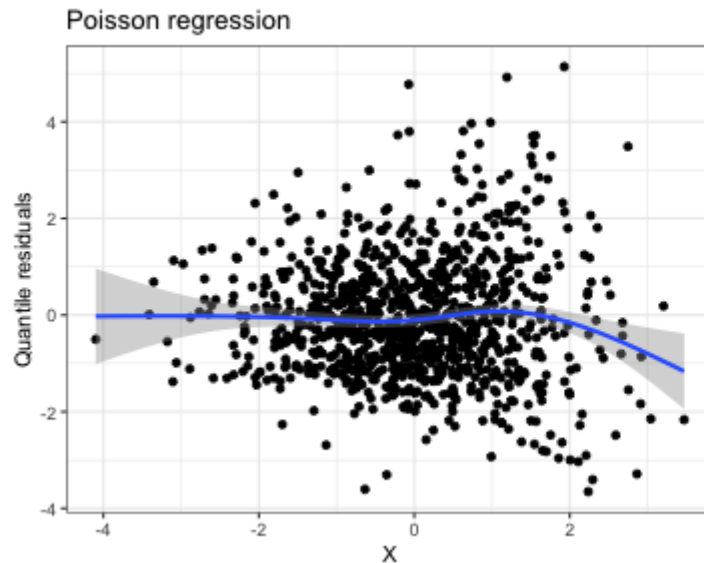
```
## [1] 0.95
```

Class activity

What happens to quantile residual plots as I increase θ? What about coverage?

Class activity

Quantile residual plots for negative binomial data ($\mu_i = \exp(X_i), \theta = 2$):

Class activity

Data: negative binomial with $\mu_i = \exp(X_i), \theta = 2$

Poisson coverage:

```
n <- 1000
nsim <- 500
theta <- 2
contains_beta_poisson <- rep(0, nsim)
for(i in 1:nsim){
  x <- rnorm(n, sd = 1.2)
  y <- rnbinom(n, size=theta, mu=exp(x))

  m1 <- glm(y ~ x, family = poisson)

  upper <- summary(m1)$coefficients[2,1] +
    1.96*summary(m1)$coefficients[2,2]
  lower <- summary(m1)$coefficients[2,1] -
    1.96*summary(m1)$coefficients[2,2]

  contains_beta_poisson[i] <- upper > 1 && lower < 1
}
mean(contains_beta_poisson)
```

```
## [1] 0.508
```

Class activity

Data: negative binomial with $\mu_i = \exp(X_i), \theta = 2$

quasi-Poisson coverage:

```
contains_beta_quasipoisson <- rep(0, nsim)
for(i in 1:nsim){
  x <- rnorm(n, sd = 1.2)
  y <- rnbinom(n, size=theta, mu=exp(x))

  m2 <- glm(y ~ x, family = quasipoisson)

  upper <- summary(m2)$coefficients[2,1] +
    qt(0.025, n-2, lower.tail=F)*summary(m2)$coefficients[2,2]
  lower <- summary(m2)$coefficients[2,1] -
    qt(0.025, n-2, lower.tail=F)*summary(m2)$coefficients[2,2]

  contains_beta_quasipoisson[i] <- upper > 1 && lower < 1
}
mean(contains_beta_quasipoisson)
```

```
## [1] 0.644
```

Class activity

Data: negative binomial with $\mu_i = \exp(X_i), \theta = 2$

negative binomial coverage:

```
contains_beta_nb <- rep(0, nsim)
for(i in 1:nsim){
  x <- rnorm(n, sd = 1.2)
  y <- rnbinom(n, size=theta, mu=exp(x))

  m3 <- glm.nb(y ~ x)

  upper <- summary(m3)$coefficients[2,1] +
    1.96*summary(m3)$coefficients[2,2]
  lower <- summary(m3)$coefficients[2,1] -
    1.96*summary(m3)$coefficients[2,2]

  contains_beta_nb[i] <- upper > 1 && lower < 1
}
mean(contains_beta_nb)
```
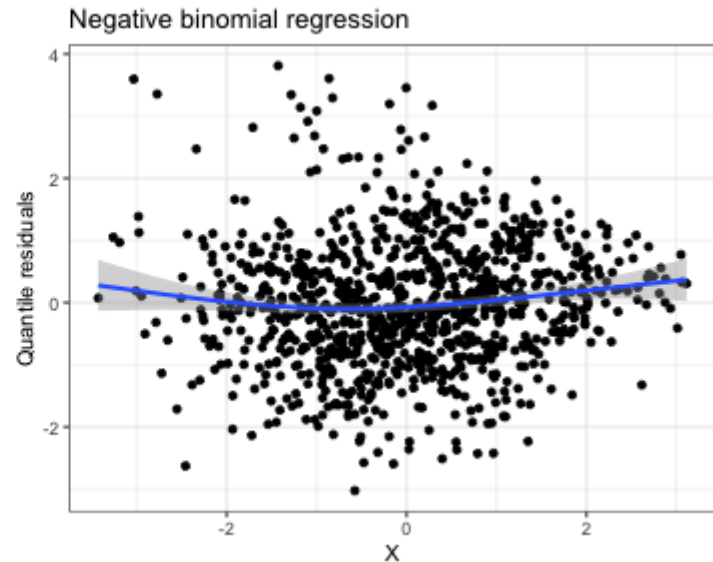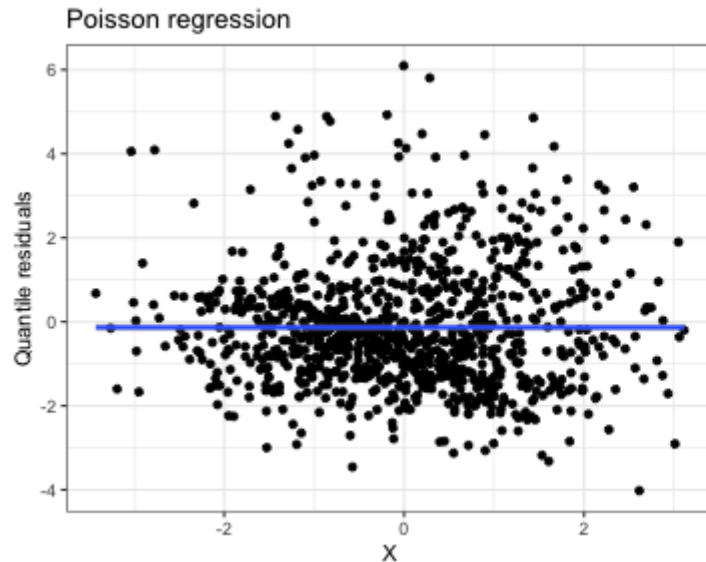
```
## [1] 0.946
```

Class activity

Now let's simulate quasi-Poisson data with $\phi = 3$:

Class activity

Data: quasi-Poisson with $\phi = 3$

Poisson coverage:

```
n <- 1000
nsim <- 500
phi <- 3
contains_beta_poisson <- rep(0, nsim)
for(i in 1:nsim){
  x <- rnorm(n, sd = 1.2)
  y <- rqpois(n, mean = exp(x), dispersion = phi)

  m1 <- glm(y ~ x, family = poisson)

  upper <- summary(m1)$coefficients[2,1] +
    1.96*summary(m1)$coefficients[2,2]
  lower <- summary(m1)$coefficients[2,1] -
    1.96*summary(m1)$coefficients[2,2]

  contains_beta_poisson[i] <- upper > 1 && lower < 1
}
mean(contains_beta_poisson)
```

```
## [1] 0.738
```

Class activity

Data: quasi-Poisson with $\phi = 3$

quasi-Poisson coverage:

```
contains_beta_quasipoisson <- rep(0, nsim)
for(i in 1:nsim){
  x <- rnorm(n, sd = 1.2)
  y <- y <- rqpois(n, mean = exp(x), dispersion = phi)

  m2 <- glm(y ~ x, family = quasipoisson)

  upper <- summary(m2)$coefficients[2,1] +
    qt(0.025, n-2, lower.tail=F)*summary(m2)$coefficients[2,2]
  lower <- summary(m2)$coefficients[2,1] -
    qt(0.025, n-2, lower.tail=F)*summary(m2)$coefficients[2,2]

  contains_beta_quasipoisson[i] <- upper > 1 && lower < 1
}
mean(contains_beta_quasipoisson)
```

```
## [1] 0.92
```

Class activity

Data: quasi-Poisson with $\phi = 3$

negative binomial coverage:

```
contains_beta_nb <- rep(0, nsim)
for(i in 1:nsim){
  x <- rnorm(n, sd = 1.2)
  y <- y <- rqpois(n, mean = exp(x), dispersion = phi)

  m3 <- glm.nb(y ~ x)

  upper <- summary(m3)$coefficients[2,1] +
    1.96*summary(m3)$coefficients[2,2]
  lower <- summary(m3)$coefficients[2,1] -
    1.96*summary(m3)$coefficients[2,2]

  contains_beta_nb[i] <- upper > 1 && lower < 1
}
mean(contains_beta_nb)
```

```
## [1] 0.936
```

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