Negative binomial regression

Recap: 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
...
```

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

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

LRT statistic: $G = 2(\ell(\hat{\mu}_{full}) - \ell(\hat{\mu}_{reduced})) \sim \chi^2_q$

$q = \#$ parameters tested

$(q = 3)$

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
```

$$\underbrace{\texttt{m2\$twologlik}}_{2\ell(\hat{\mu}_{full})} - \underbrace{\texttt{m3\$twologlik}}_{2\ell(\hat{\mu}_{reduced})}$$

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

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

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

Likelihood ratio test

> Why can I use the residual deviance to perform a likelihood ratio test for a Poisson regression model, but not for a negative binomial model?

LRT: $G = 2(\ell(\hat{\mu}_{full}) - \ell(\hat{\mu}_{reduced})) \sim \chi^2_q$

For EDM: $G = D^*(y, \hat{\mu}_{reduced}) - D^*(y, \hat{\mu}_{full})$

But: since we estimate r in NB regression, we can't directly compare full & reduced deviances

EDM in dispersion model form:

$$f(y;\mu,\phi) = b(y,\phi)\exp\left\{-\frac{1}{2\phi}d(y,\mu)\right\}$$

$$2\ell(\hat{\mu}) = 2\sum_{i=1}^{n}\log f(Y_i;\hat{\mu},\phi)$$

$$= 2\left(\sum_{i=1}^{n}\log b(Y_i,\phi) - \frac{1}{2\phi}d(Y_i,\hat{\mu})\right)$$

$$= 2\sum_i \log b(Y_i,\phi) - \underbrace{\frac{\sum_i d(Y_i,\hat{\mu})}{\phi}}_{=\frac{D(y,\hat{\mu})}{\phi} = D^*(y,\hat{\mu})}$$

$$2\left(\ell(\hat{\mu}_{full}) - \ell(\hat{\mu}_{reduced})\right) = D^*(y,\hat{\mu}_{reduced}) - D^*(y,\hat{\mu}_{full})$$

$$f(y;r,p) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)}\exp\left\{y\log p - (-r\log(1-p))\right\}$$

$b(y,\phi)$ includes r

full & reduced models given different $\hat{r}$

$\Rightarrow$ normalizing function doesn't cancel

New data

Survey data from 77 college students on a dry campus (i.e., alcohol is prohibited) in the US. Survey asks students "How many alcoholic drinks did you consume last weekend?"

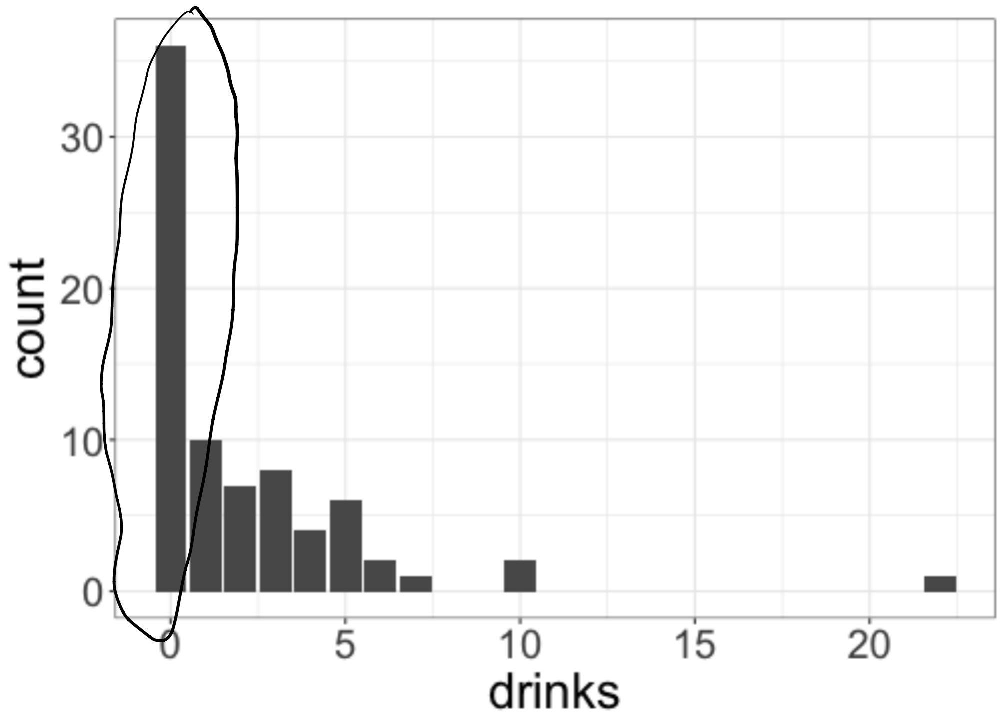
- `drinks`: the number of drinks the student reports consuming
- `sex`: an indicator for whether the student identifies as male
- `OffCampus`: an indicator for whether the student lives off campus
- `FirstYear`: an indicator for whether the student is a first-year student

> Our goal: model the number of drinks students report consuming.

· right-skewed

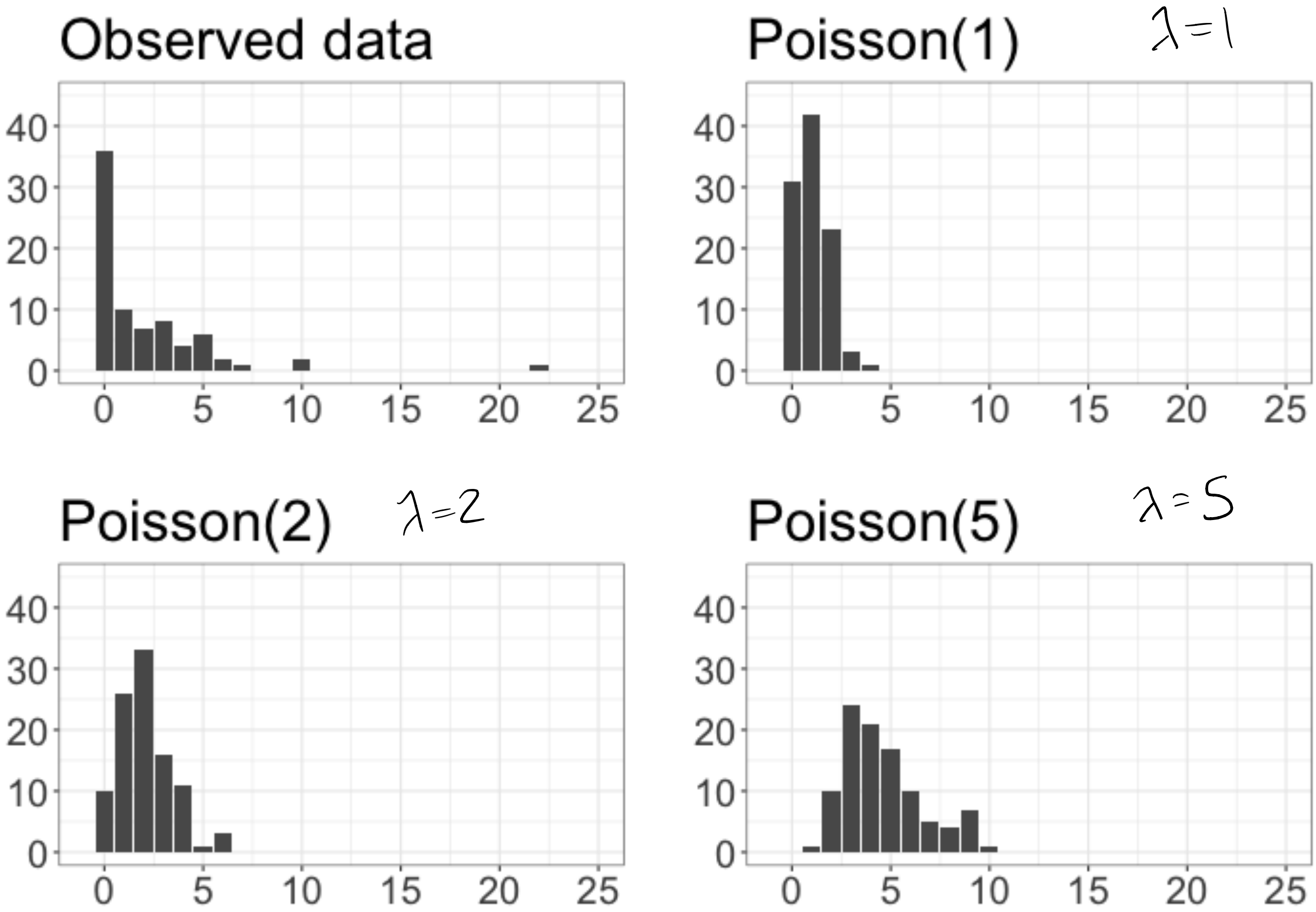
· lots of zeros! (too many zeros)

EDA: drinks



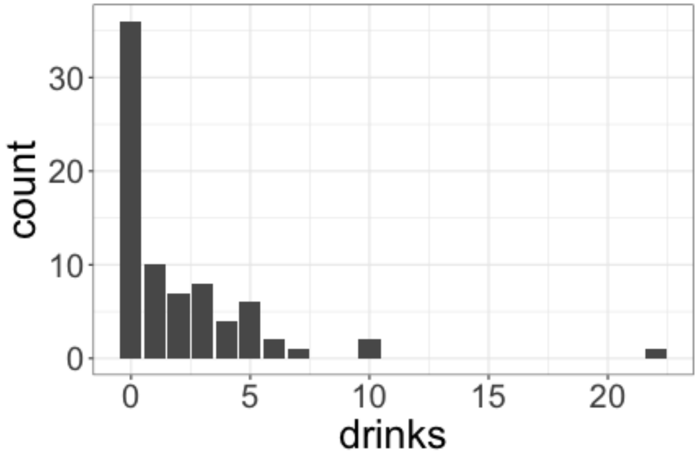
What do you notice about this distribution?

Comparisons with Poisson distributions



Excess zeros

> *Why might there be excess 0s in the data, and why is that a problem for modeling the number of drinks consumed?*



- some students never drink

  ⇒ a Poisson distribution is not a good choice for all students in the data

Modeling

Let $Z_i = \begin{cases} 1 & \text{student } i \text{ never drinks} \\ 0 & \text{student } i \text{ sometimes drinks} \end{cases}$ (latent variable)

Y_i = # drinks consumed

$Y_i \mid (Z_i = 1) \equiv 0$ (point mass at 0)

$Y_i \mid (Z_i = 0) \sim \text{Poisson}(\lambda_i)$ (Poisson distribution)

$Z_i \sim \text{Bernoulli}(\alpha_i)$

$$P(Y_i = y) = \underbrace{P(Y_i = y \mid Z_i = 0)}_{\frac{e^{-\lambda_i}\lambda_i^y}{y!}} \underbrace{P(Z_i = 0)}_{1-\alpha_i} + \underbrace{P(Y_i = y \mid Z_i = 1)}_{= \begin{cases} 1 & y = 0 \\ 0 & y > 0 \end{cases}} \underbrace{P(Z_i = 1)}_{\alpha_i}$$

$$P(Y_i = y) = \begin{cases} e^{-\lambda_i}(1-\alpha_i) + \alpha_i & y = 0 \\ \dfrac{e^{-\lambda_i}\lambda_i^y}{y!}(1-\alpha_i) & y > 0 \end{cases}$$

$$P(Y_i = y) = \begin{cases} e^{-\lambda_i}(1-\alpha_i) + \alpha_i & y = 0 \\ \dfrac{e^{-\lambda_i}\lambda_i^{y}}{y!}(1-\alpha_i) & y > 0 \end{cases}$$

<u>mixture model</u>

(mixture between a point mass at 0, and a Poisson distribution)

<u>Zero-inflated Poisson (ZIP) model</u>

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = \gamma^T X_i$$ ← logistic component

$$\log(\lambda_i) = \beta^T X_i$$ ← Poisson component

Zero-inflated Poisson (ZIP) model

$$P(Y_i = y) = \begin{cases} e^{-\lambda_i}(1-\alpha_i) + \alpha_i & y = 0 \\ \dfrac{e^{-\lambda_i}\lambda_i^y}{y!}(1-\alpha_i) & y > 0 \end{cases}$$

where

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = \gamma_0 + \gamma_1 FirstYear_i + \gamma_2 OffCampus_i + \gamma_3 Male_i$$

$$\log(\lambda_i) = \beta_0 + \beta_1 FirstYear_i + \beta_2 OffCampus_i + \beta_3 Male_i$$

In R

zero inflated model

Poisson

```
library(pscl)
m1 <- zeroinfl(drinks ~ FirstYear + OffCampus + sex |
                 FirstYear + OffCampus + sex,
               data = wdrinks)
summary(m1)
```

logistic component

poisson

logistic

```
...
## Count model coefficients (poisson with log link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.8010     0.1620   4.945 7.60e-07 ***
## FirstYearTRUE   -0.1619     0.3651  -0.444   0.6574
## OffCampusTRUE    0.3724     0.2135   1.744   0.0811 .
## sexm             0.9835     0.1889   5.205 1.94e-07 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.39618    0.39752  -0.997    0.319
## FirstYearTRUE   0.89197    0.65878   1.354    0.176
## OffCampusTRUE  -1.69137    1.47761  -1.145    0.252
## sexm           -0.07079    0.58846  -0.120    0.904
...
```

Interpretation

```
...
## Zero-inflation model coefficients (binomial with logit l
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.39618    0.39752  -0.997    0.319
## FirstYearTRUE    0.89197    0.65878   1.354    0.176
## OffCampusTRUE   -1.69137    1.47761  -1.145    0.252
## sexm            -0.07079    0.58846  -0.120    0.904
...
```

> How would I interpret the estimated coefficient 0.892 in the logistic regression component of the model?

The odds of being a nondrinker are $e^{0.892} = 2.44$ times higher for first year students

Interpretation

```
...
## Count model coefficients (poisson with log link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.8010     0.1620   4.945 7.60e-07 ***
## FirstYearTRUE   -0.1619     0.3651  -0.444   0.6574
## OffCampusTRUE    0.3724     0.2135   1.744   0.0811 .
## sexm             0.9835     0.1889   5.205 1.94e-07 ***
##
...
```

> How would I interpret the estimated coefficient 0.372 in the Poisson regression component of the model?

For drinkers, the average # of drinks ie $e^{0.372} = 1.45$ times higher for off-campus students