Lecture 20

Data from last time

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

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

Recap: Poisson hurdle model

Y_i = # drinks student i reported consuming

$P(Y_i > 0) = p_i$ $\quad \log\left(\frac{p_i}{1-p_i}\right) = \gamma^T X_i$

$Y_i \mid (Y_i > 0) \sim \text{PosPoisson}(\lambda_i)$ $\quad \log(\lambda_i) = \beta^T X_i$

(also a ZTP(λ_i))

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

← involves both γ and β

$$\sum_{y=0}^{\infty} P(Y=y) = P(Y=0) + \sum_{y=1}^{\infty} P(Y=y)$$

$$= 1 - p_i + p_i \underbrace{\sum_{y=1}^{\infty} \frac{e^{-\lambda_i} \lambda_i^y}{y!\,(1-e^{-\lambda_i})}}_{=1 \text{ (PosPoisson)}}$$

Fitting Poisson hurdle models

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

$$Z_i = \mathbb{1}\{Y_i > 0\}$$

$$P(Y_i = y) = (1 - p_i)^{1 - Z_i} \, p_i^{Z_i} \left(\frac{e^{-\lambda_i} \lambda_i^y}{y!(1 - e^{-\lambda_i})} \right)^{Z_i}$$

$$\ell(\beta, \gamma) = \sum_{i=1}^{n} \{ Z_i \log p_i + (1 - Z_i) \log(1 - p_i) \} + \sum_{i=1}^{n} \{ Z_i (Y_i \log \lambda_i - \log(e^{\lambda_i} - 1)) - \log y_i! \}$$

$$= \sum_{i=1}^{n} \{ Z_i \log \left(\frac{p_i}{1 - p_i} \right) + \log(1 - p_i) \} + \sum_{i=1}^{n} \{ Z_i (Y_i \log \lambda_i - \log(e^{\lambda_i} - 1)) + Z_i \log y_i! \}$$

$$\frac{\partial \ell}{\partial \gamma} = X^T (Z - p)$$

X = design matrix

$$Z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}$$

$$p = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix}$$

$$\frac{\partial \ell}{\partial \beta} = X^T \operatorname{diag}(z_i)(Y - \mu)$$

$$\mu_i = \mathbb{E}[Y_i \mid (Y_i > 0)]$$

$$= \frac{\lambda_i}{1 - e^{-\lambda_i}}$$

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

$$\Rightarrow U(\beta, \gamma) = \begin{bmatrix} X^T \operatorname{diag}(z_i)(Y - \mu) \\ X^T (Z - p) \end{bmatrix} \overset{\text{set}}{=} 0$$

Can solve for β, γ separately

Fitting the model in R

$P_i = P(Y_i > 0)$

```
1 library(pscl)
2
3 m1 <- hurdle(drinks ~ sex + FirstYear + OffCampus,
4               dist = "poisson", zero.dist = "binomial",
5               data = wdrinks)
6
7 m1$coefficients
```

```
$count
  (Intercept)          sexm FirstYearTRUE OffCampusTRUE
    0.8132113     0.9706640    -0.2181068     0.3762608

$zero
  (Intercept)          sexm FirstYearTRUE OffCampusTRUE
    0.1230510     0.3377969    -0.8554289     1.5803472
```

Pos Poisson part

logistic regression part

Among students who drink, male students report drinking more (holding First Year & off Campus fixed)

Male students are more likely to report drinking (holding First Year & off Campus fixed)

The odds of any drinking are $e^{0.338}$ times higher for a male student (holding ... fixed)

Model assumptions

```
1 m1 <- hurdle(drinks ~ sex + FirstYear + OffCampus,
2               dist = "poisson", zero.dist = "binomial",
3               data = wdrinks)
4
5 m1$coefficients
```

```
$count
  (Intercept)          sexm FirstYearTRUE OffCampusTRUE
    0.8132113     0.9706640    -0.2181068     0.3762608

$zero
  (Intercept)          sexm FirstYearTRUE OffCampusTRUE
    0.1230510     0.3377969    -0.8554289     1.5803472
```

What assumptions does this model make?

- Distribution: $Y_i | (Y_i > 0) \sim \text{PosPoisson}(\lambda_i)$
- Independence (needed for ML)
- Shape: $\log(\lambda_i) = \beta^T X_i$

  $\log\left(\frac{p_i}{1-p_i}\right) = \gamma^T X_i$

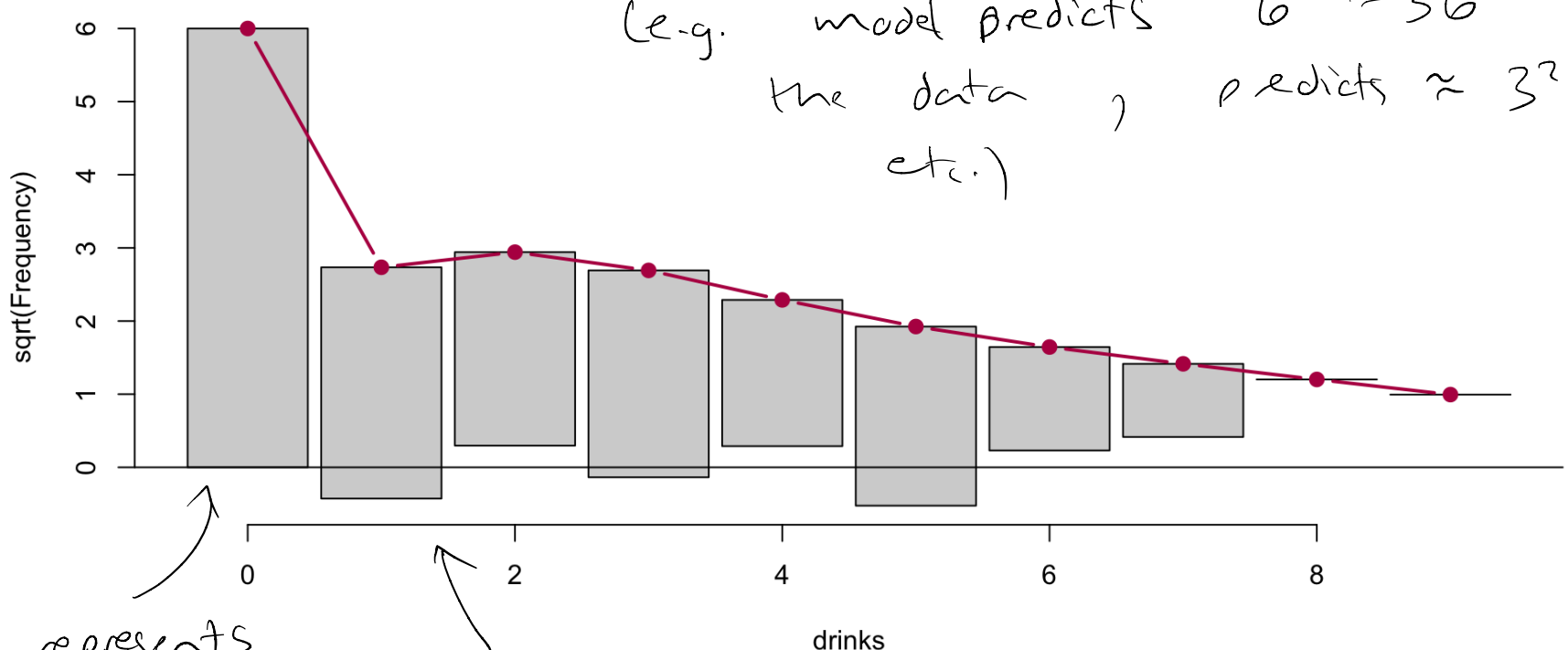
Model diagnostics

- Quantile residual plots
  - For the whole model (distribution of Y_i)
  - For the zero component ($\mathbb{1}\{Y_i > 0\}$)
  - For the cont component ($Y_i | (Y_i > 0)$)

- Rootogram (for assessing distributional assumption)

Rootograms

```
1 library(countreg)
2 rootogram(m1)
```



red: estimated $\sqrt{\#\text{of occurences}}$ from fitted model

(e.g. model predicts $6^2 = 36$ 0s in the data, predicts $\approx 3^2 = 9$ 1s, etc.)

bar represents actual (observed) $\sqrt{\#\text{occurences}}$

bar below 0: $\uparrow$ observed more occurences than we predicted

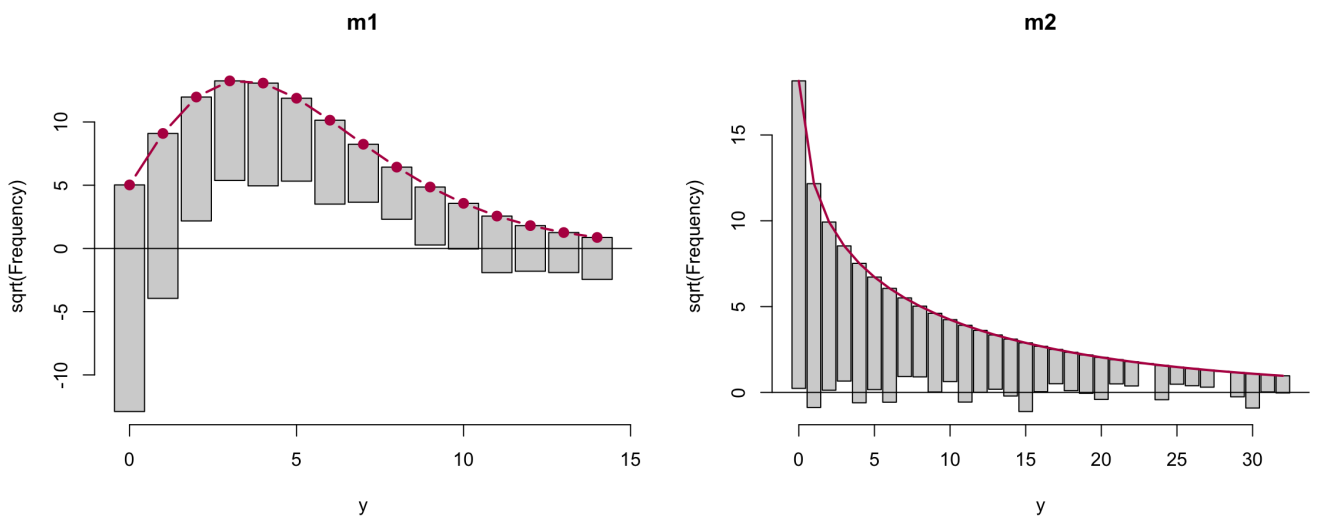
$\downarrow$ or the horizontal

Other examples with rootograms

```
par(mfrow=c(1, 2))

x <- rnorm(1000)
y <- rnbinom(1000, 0.5, mu=exp(1.5 + 0.2*x))
m1 <- glm(y ~ x, family = poisson)
m2 <- glm.nb(y ~ x)

rootogram(m1)
rootogram(m2)
```



suggests overdispersion

Hurdle models for count data

$$P(Y_i > 0) = p_i \quad g_{zero}(p_i) = \gamma^T X_i$$

$$Y_i | (Y_i > 0) \sim ZT(\lambda_i) \quad g_{count}(\lambda_i) = \beta^T X_i$$

Usually: $g_{zero}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$

Usually: $g_{count}(\lambda_i) = \log(\lambda_i)$

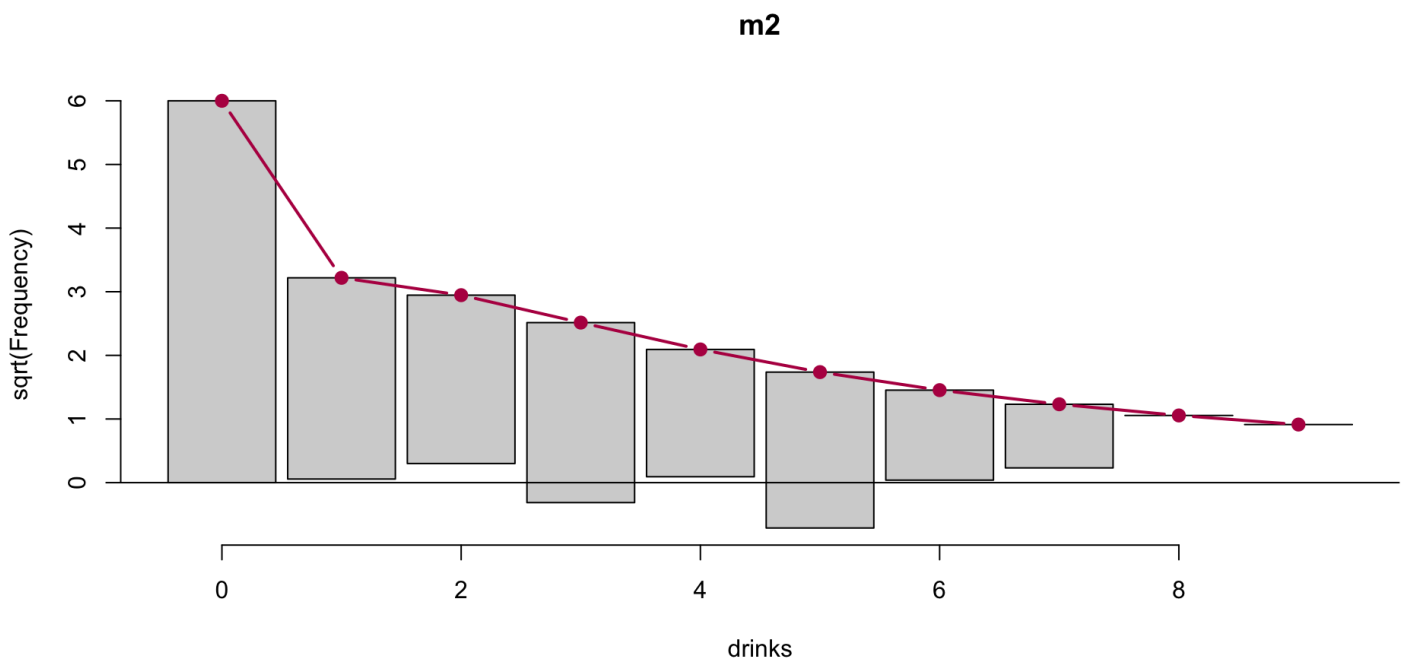
$ZT(\lambda_i)$: Zero-truncated distribution

λ_i = mean of non-truncated dist.

Negative binomial hurdle model

```r
m2 <- hurdle(drinks ~ sex + FirstYear + OffCampus,
             dist = "negbin", zero.dist = "binomial",
             data = wdrinks)

rootogram(m2)
```



Class activity

https://sta712-f23.github.io/class_activities/ca_lecture_20.html