Session 5: loglinear regression part 2

Levi Waldron

CUNY SPH Biostatistics 2

Learning objectives and outline

Session 5:
loglinear
regression
part 2

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion

Zero Inflation

Learning objectives

1. Define and identify over-dispersion in count data
2. Define the negative binomial (NB) distribution and identify applications for it
3. Define zero-inflated count models
4. Fit and interpret Poisson and NB, with and without zero inflation

**Session 5:
loglinear
regression
part 2**

Levi Waldron

**Learning
objectives and
outline**

Review

Over-
dispersion

Zero Inflation

Outline

1. Review of log-linear Poisson glm
2. Review of diagnostics and interpretation of coefficients
3. Over-dispersion
   - Negative Binomial distribution
4. Zero-inflated models

- Vittinghoff section 8.1-8.3

Session 5:
loglinear
regression
part 2

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion

Zero Inflation

Review

**Session 5:
loglinear
regression
part 2**

Levi Waldron

Learning
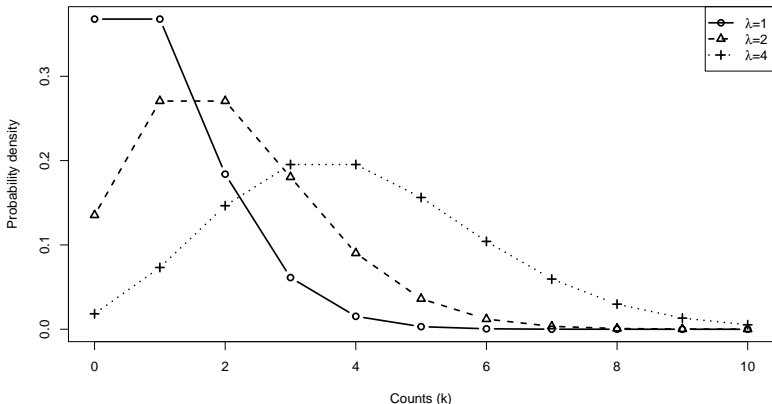objectives and
outline

Review

Over-
dispersion

Zero Inflation

Components of GLM

- **Random component** specifies the conditional distribution for the response variable - it doesn't have to be normal but can be any distribution that belongs to the "exponential" family of distributions
- **Systematic component** specifies linear function of predictors (linear predictor)
- **Link** [denoted by $g(.)$] specifies the relationship between the expected value of the random component and the systematic component, can be linear or nonlinear

Session 5:
loglinear
regression
part 2

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion

Zero Inflation

Motivating example: Choice of Distribution

- Count data are often modeled as Poisson distributed:
  - mean λ is greater than 0
  - variance is also λ
  - Probability density $P(k, \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$

Session 5: loglinear regression part 2

Levi Waldron

Learning objectives and outline

Review

Over-dispersion

Zero Inflation

Poisson model: the GLM

The **systematic part** of the GLM is:

$$log(\lambda_i) = \beta_0 + \beta_1 \mathrm{RACE}_i + \beta_2 \mathrm{TRT}_i + \beta_3 \mathrm{ALCH}_i + \beta_4 \mathrm{DRUG}_i$$

Or alternatively:

$$\lambda_i = exp\left(\beta_0 + \beta_1 \mathrm{RACE}_i + \beta_2 \mathrm{TRT}_i + \beta_3 \mathrm{ALCH}_i + \beta_4 \mathrm{DRUG}_i\right)$$

The **random part** is (Recall the λ_i is both the mean and variance of a Poisson distribution):

$$y_i \sim Poisson(\lambda_i)$$

Session 5: loglinear regression part 2

Levi Waldron

Learning objectives and outline
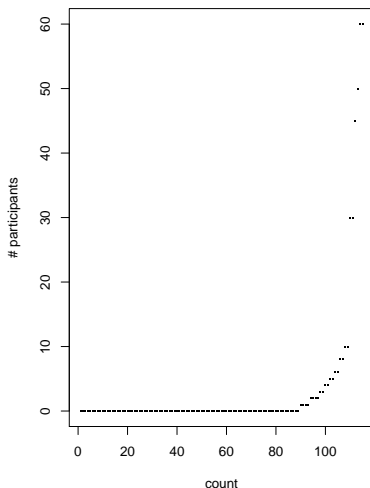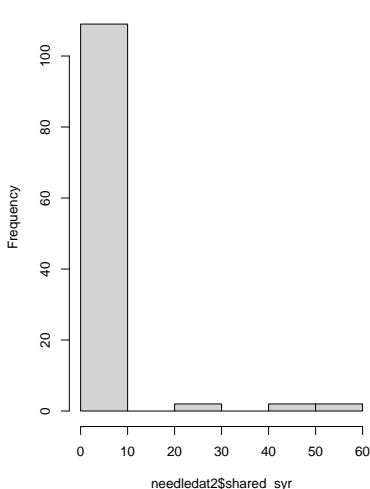
Review

Over-dispersion

Zero Inflation

Example: Risky Drug Use Behavior

- Load the "needle_sharing" dataset
- Outcome is # times the drug user shared a syringe in the past month (shared_syr)
- Predictors: sex, ethn, homeless
  - filtered to only sex "M" or "F", ethn "White", "AA", "Hispanic"

```
##
## -- Column specification --------------------------------
## cols(
##   id = col_double(),
##   sex = col_character(),
##   ethn = col_character(),
##   age = col_double(),
##   dprsn_dx = col_double(),
##   sexabuse = col_double(),
##   shared_syr = col_double(),
##   hivstat = col_double(),
```
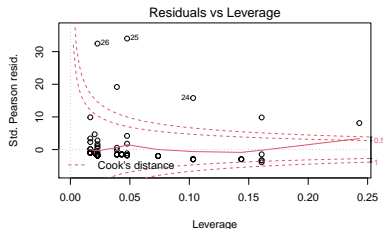
Session 5: loglinear regression part 2

Levi Waldron

Learning objectives and outline

Review

Over-dispersion

Zero Inflation

Example: Risky Drug Use Behavior

Exploratory plots

**Session 5:
loglinear
regression
part 2**

Levi Waldron

Learning
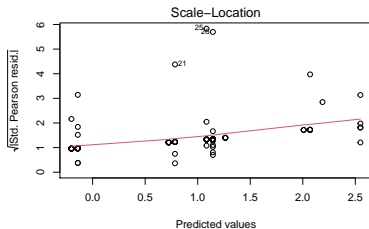objectives and
outline

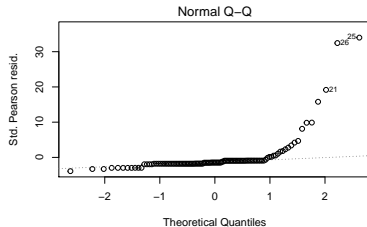**Review**

Over-
dispersion

Zero Inflation
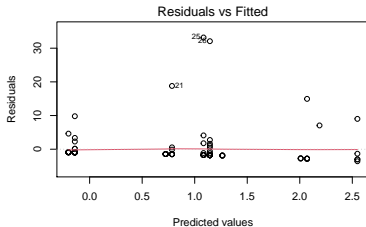
Risky Drug Use Behavior:
fitting a Poisson model

```
fit.pois <- glm(shared_syr ~ sex + ethn + homeless,
                data = needledat2,
                family = poisson(link = "log"))
```

**Session 5:
loglinear
regression
part 2**

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion

Zero Inflation

Risky Drug Use Behavior: residuals plots

* Poisson model is definitely not a good fit.

Session 5:
loglinear
regression
part 2

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion

Zero Inflation

Over-dispersion

**Session 5:
loglinear
regression
part 2**

Levi Waldron

Learning
objectives and
outline

Review

**Over-
dispersion**

Zero Inflation

When the Poisson model doesn't fit

- inference from log-linear models is sensitive to assumptions on the distribution of residuals (e.g. Poisson)
- In the Poisson distribution, the variance is equal to the mean.
- *i.e.* if subjects with a particular pattern of covariates have a mean of 4 visits/yr, then variance is also 4 and the standard deviation is 2 visits / yr.
- The Poisson distribution often fails when the variance exceeds the mean
  - You can *check* this assumption
- Can use alternative random distributions:
  - Negative binomial distribution
- Can introduce zero-inflation

Session 5: loglinear regression part 2

Levi Waldron

Learning objectives and outline

Review

Over-dispersion

Zero Inflation

Negative binomial distribution

- The binomial distribution is the number of successes in n trials:
  - Roll a die ten times, how many times do you see a 6?
- The negative binomial distribution is the number of successes it takes to observe r failures:
  - How many times do you have to roll the die to see a 6 ten times?
  - Note that the number of rolls is no longer fixed.
  - In this example, p=5/6 and a 6 is a "failure"

Session 5:
loglinear
regression
part 2

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion

Zero Inflation

Negative binomial GLM

One way to parametrize a NB model is with a **systematic part** equivalent to the Poisson model:

$$log(\lambda_i) = \beta_0 + \beta_1 \text{RACE}_i + \beta_2 \text{TRT}_i + \beta_3 \text{ALCH}_i + \beta_4 \text{DRUG}_i$$
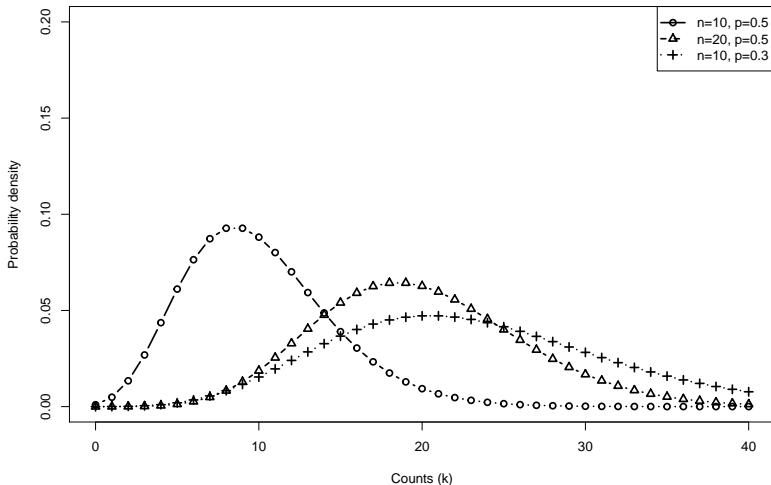
Or:

$$\lambda_i = exp\left(\beta_0 + \beta_1 \text{RACE}_i + \beta_2 \text{TRT}_i + \beta_3 \text{ALCH}_i + \beta_4 \text{DRUG}_i\right)$$

And a **random part**:

$$y_i \sim NB(\lambda_i, \theta)$$

- θ is a **dispersion parameter** that is estimated
- When $\theta = 0$ it is equivalent to Poisson model
- MASS::glm.nb() uses this parametrization, dnbinom() does not
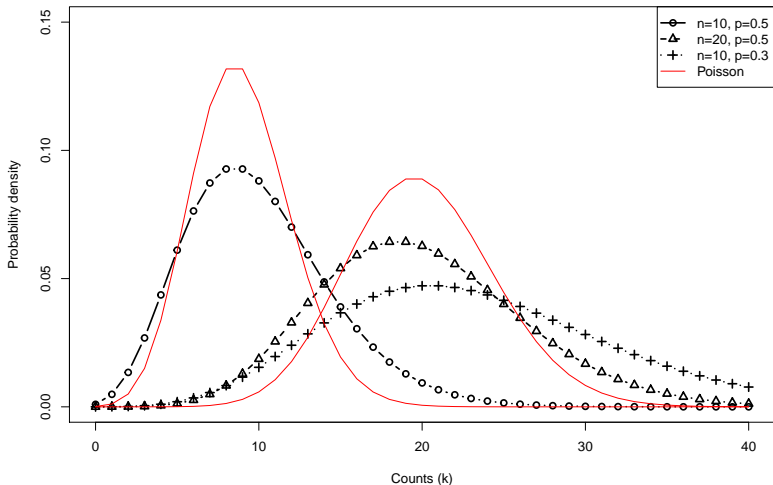- The Poisson model can be considered **nested** within the Negative Binomial model

Session 5: loglinear regression part 2

Levi Waldron

Learning objectives and outline

Review

Over-dispersion

Zero Inflation

Negative Binomial Random Distribution

Session 5:
loglinear
regression
part 2

Levi Waldron

Learning
objectives and
outline

Review

**Over-
dispersion**

Zero Inflation

Compare Poisson vs. Negative
Binomial

Negative Binomial Distribution has two parameters: # of trials
n, and probability of success p

Session 5: loglinear regression part 2

Levi Waldron

Learning objectives and outline

Review

Over-dispersion

Zero Inflation

Risky drug behavior: Negative Binomial Regression

```
library(MASS)
fit.negbin <- glm.nb(shared_syr ~ sex + ethn + homele
                     data = needledat2)
summary(fit.negbin)
```

```
##
## Call:
## glm.nb(formula = shared_syr ~ sex + ethn + homeles
##     init.theta = 0.07743871374, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8801  -0.7787  -0.6895  -0.5748   1.5675
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.4641     0.8550   0.543   0.5876
```

Session 5: loglinear regression part 2

Levi Waldron

Learning objectives and outline

Review

Over-dispersion

Zero Inflation

Likelihood ratio test

Recall the Deviance:

$$\Delta(D) = -2 * \Delta(\log \text{ likelihood})$$

And recall the difference in deviance under H_0 (no improvement in fit) is *chi-square distributed*, with df equal to the difference in df of the two models:

```
(ll.negbin <- logLik(fit.negbin))
```

```
## 'log Lik.' -147.1277 (df=6)
```

```
(ll.pois <- logLik(fit.pois))
```

```
## 'log Lik.' -730.0133 (df=5)
```

```
pchisq(2 * (ll.negbin - ll.pois), df=1, lower.tail=FA
```

```
## 'log Lik.' 1.675949e-255 (df=6)
```

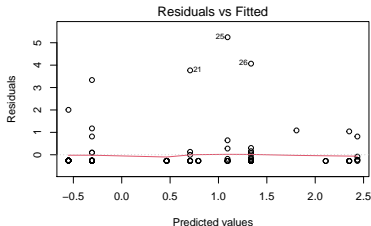**Session 5:
loglinear
regression
part 2**

Levi Waldron

Learning
objectives and
outline

Review

**Over-
dispersion**

Zero Inflation

Risky Drug Use Behavior: NB regression residuals plots

Session 5:
loglinear
regression
part 2

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion

Zero Inflation

Zero Inflation

Session 5:
loglinear
regression
part 2

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion
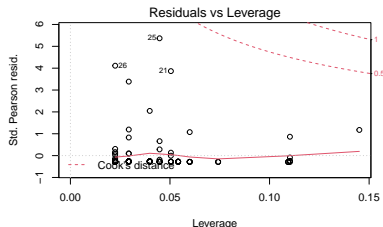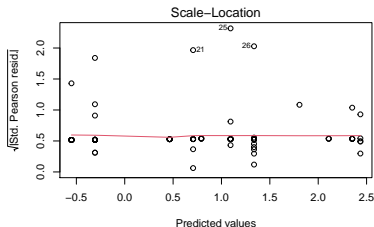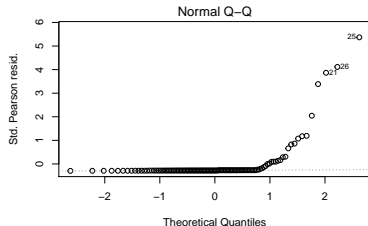
Zero Inflation

Zero inflated "two-step" models

Step 1: logistic model to determine whether count is zero or Poisson/NB

Step 2: Poisson or NB regression distribution for y_i not set to zero by *1*.

**Session 5:
loglinear
regression
part 2**

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion

Zero Inflation

Poisson Distribution with Zero Inflation

**Session 5:
loglinear
regression
part 2**

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion
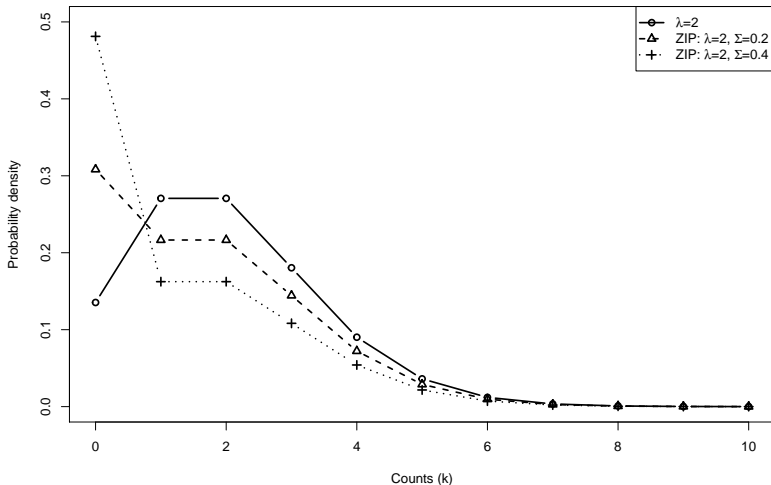
Zero Inflation

Risky drug behavior: Zero-inflated Poisson regression

```
library(pscl)
fit.ZIpois <- zeroinfl(shared_syr ~ sex + ethn + home
                       dist = "poisson",
                       data = needledat2)
```

Session 5: loglinear regression part 2

Levi Waldron

Learning objectives and outline

Review

Over-dispersion

Zero Inflation

Zero-inflated Poisson regression - the model

```
summary(fit.ZIpois)
```

```
##
## Call:
## zeroinfl(formula = shared_syr ~ sex + ethn + homel
##     dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.0761 -0.5784 -0.4030 -0.3341 10.6835
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.2168     0.1796  17.908  < 2e-16
## sexM          -1.4725     0.1442 -10.212  < 2e-16
## ethnHispanic  -0.1524     0.1576  -0.968 0.333244
## ethnWhite      0.5226     0.1464   3.577 0.000348
```

**Session 5:
loglinear
regression
part 2**

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion

Zero Inflation

Risky drug behavior:
Zero-inflated Negative
Binomial regression

```
fit.ZInegbin <- zeroinfl(shared_syr ~ sex + ethn + ho
                         dist = "negbin",
                         data = needledat2)
```

- *NOTE*: zero-inflation model can include any of your
  variables as predictors
- *WARNING* Default in zerinfl() function is to use *all*
  variables as predictors in logistic model

Session 5: loglinear regression part 2

Levi Waldron

Learning objectives and outline

Review

Over-dispersion

Zero Inflation

Zero-inflated Negative Binomial regression model 1

```
summary(fit.ZInegbin)
```

```
##
## Call:
## zeroinfl(formula = shared_syr ~ sex + ethn + homeless, data = needledat2,
##     dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.5402 -0.3255 -0.2714 -0.1926  5.1496
##
## Count model coefficients (negbin with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.8410     1.1845   2.399  0.01646 *
## sexM         -2.2282     0.9351  -2.383  0.01718 *
## ethnHispanic -0.4123     0.9831  -0.419  0.67492
## ethnWhite    -0.4299     0.8648  -0.497  0.61908
## homelessyes   1.9460     0.7103   2.740  0.00615 **
## Log(theta)   -1.1971     0.5159  -2.320  0.02032 *
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.6867     0.8465   1.993   0.0463 *
## sexM          -0.9920     0.8016  -1.238   0.2159
## ethnHispanic -13.1868   281.9134  -0.047   0.9627
## ethnWhite     -0.7455     0.7304  -1.021   0.3074
## homelessyes    0.3554     0.7397   0.480   0.6309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.3021
## Number of iterations in BFGS optimization: 24
```

Session 5: loglinear regression part 2

Levi Waldron

Learning objectives and outline

Review

Over-dispersion

Zero Inflation

Zero-inflated Negative Binomial regression - simplified ZI model

- Model is much more interpretable if the exposure of interest is *not* included in the zero-inflation model.
- E.g. with HIV status as the only predictor in zero-inflation model:

```
fit.ZInb2 <- zeroinfl(shared_syr ~ sex + ethn + homel
                      dist = "negbin",
                      data = needledat2)
```

Session 5:
loglinear
regression
part 2

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion

Zero Inflation

Zero-inflated Negative

Binomial regression model 2

```
summary(fit.ZInb2)
```

```
##
## Call:
## zeroinfl(formula = shared_syr ~ sex + ethn + homeless + hiv | hiv, data = needledat2,
##     dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.4299 -0.3646 -0.3559 -0.3299  6.3053
##
## Count model coefficients (negbin with log link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     3.6685     0.9470   3.874 0.000107 ***
## sexM           -1.7648     0.6205  -2.844 0.004454 **
## ethnHispanic   -1.5807     0.7446  -2.123 0.033769 *
## ethnWhite      -1.1267     0.6924  -1.627 0.103687
## homelessyes     1.0313     0.5693   1.812 0.070028 .
## hivpositive    -1.0820     1.0167  -1.064 0.287235
## hivyes          2.3724     0.7829   3.030 0.002444 **
## Log(theta)      0.1395     0.4647   0.300 0.764009
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.2163     0.2851   4.265   2e-05 ***
## hivpositive   -0.3493     0.9389  -0.372   0.710
## hivyes       -17.9654  3065.6271  -0.006   0.995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 1.1497
## Number of iterations in BFGS optimization: 12
```

Session 5: loglinear regression part 2

Levi Waldron

Learning objectives and outline

Review

Over-dispersion

Zero Inflation

Intercept-only zero-inflation model

```
fit.ZInb3 <- zeroinfl(shared_syr~sex+ethn+homeless|1,
                      dist="negbin", data=needledat2)
summary(fit.ZInb3)
```

```
##
## Call:
## zeroinfl(formula = shared_syr ~ sex + ethn + homeless | 1, data = needledat2,
##     dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.3159 -0.3123 -0.3040 -0.2953  5.2940
##
## Count model coefficients (negbin with log link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.08542    1.42671   1.462   0.1438
## sexM          -1.43809    0.89189  -1.612   0.1069
## ethnHispanic   0.48130    1.16642   0.413   0.6799
## ethnWhite     -0.07418    0.81066  -0.092   0.9271
## homelessyes    1.62076    0.67706   2.394   0.0167 *
## Log(theta)    -1.12538    0.89372  -1.259   0.2080
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5211     0.7600   0.686    0.493
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.3245
## Number of iterations in BFGS optimization: 13
## Log-likelihood: -146.8 on 7 Df
```

Session 5:
loglinear
regression
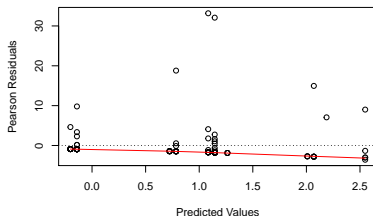part 2

Levi Waldron

Learning
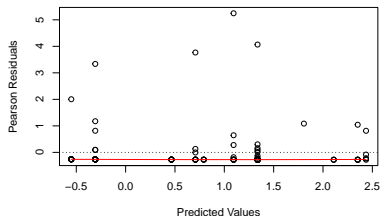objectives and
outline

Review

Over-
dispersion

Zero Inflation

Residuals vs. fitted values

I invisibly define functions plotpanel1 and plotpanel2 that will work for all types of models (see lab). These use Pearson residuals.
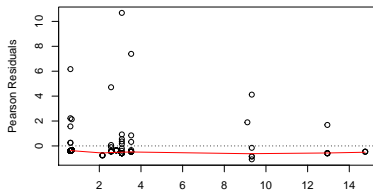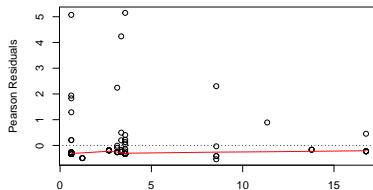
Session 5: loglinear regression part 2

Levi Waldron

Learning objectives and outline

Review

Over-dispersion

Zero Inflation

Quantile-quantile plots for residuals

Normal Q–Q Plot
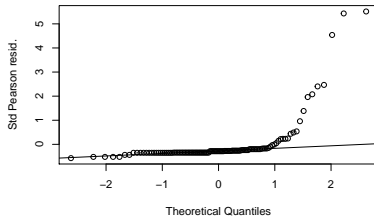**Poisson**
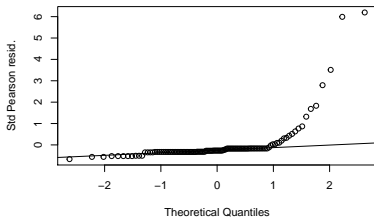
Normal Q–Q Plot
Negative Binomial
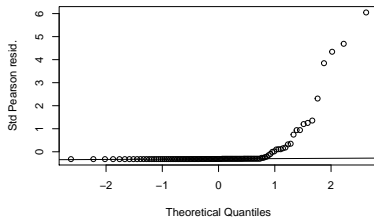
Normal Q–Q Plot
Zero–inflated Poisson
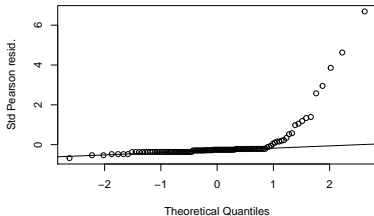
Normal Q–Q Plot
Zero–inflated Negative Binomial

still over-dispersed - ideas?

Session 5:
loglinear
regression
part 2

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion

Zero Inflation

Inference from the different models

Dependent variable:

shared_syr

Poisson

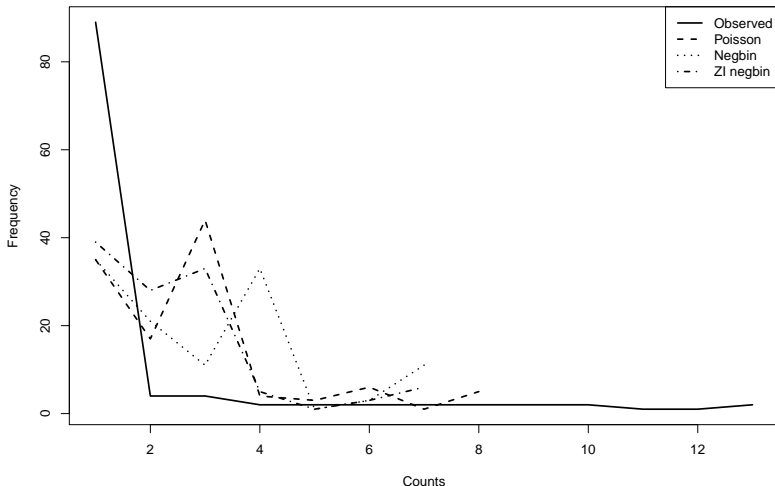negative

zero-inflated

binomial

count data

(1)

(2)

(3)

(4)

(5)

sexM

-0.925***

Session 5: loglinear regression part 2

Levi Waldron

Learning objectives and outline

Review

Over-dispersion

Zero Inflation

Example of plotting observed and predicted counts

**Session 5:
loglinear
regression
part 2**

Levi Waldron

Learning
objectives and
outline

Review

Over-
dispersion

Zero Inflation

Resources for R (and SAS)

- Short, practical tutorials on regression in R and SAS from UCLA at http://www.ats.ucla.edu/stat/:
  - Poisson Regression:
    http://www.ats.ucla.edu/stat/r/dae/poissonreg.htm
  - Negative Binomial:
    http://www.ats.ucla.edu/stat/r/dae/nbreg.htm
  - Zero-inflated Poisson:
    http://www.ats.ucla.edu/stat/r/dae/zipoisson.htm
  - Zero-inflated Negative Binomial:
    http://www.ats.ucla.edu/stat/r/dae/zinbreg.htm