BIOS621 / 821 Session 5

Levi Waldron

loglinear regression part 2

Welcome and outline - session 5

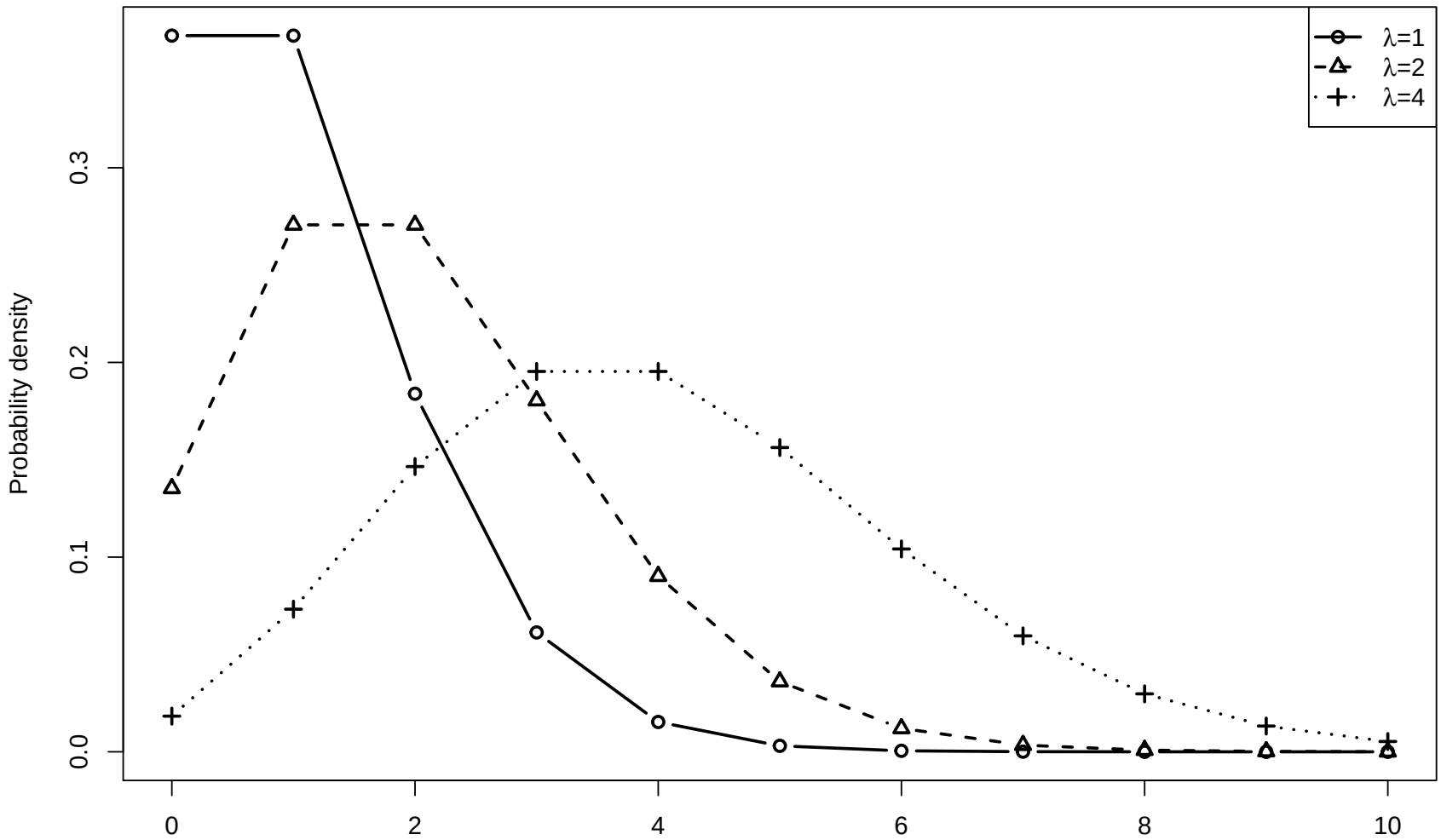
- Review of log-linear Poisson glm
- Review of diagnostics and interpretation of coefficients
- Over-dispersed models:
  - negative binomial distribution
- Zero-inflated models
- Vittinghoff section 8.1-8.3

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

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 $P(k, \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$



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):

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

Example: Risky Drug Use Behavior

- Download the "needle_sharing" dataset in csv format
- Outcome is # times the drug user shared a syringe in the past month (shared_syr)
- Predictors: sex, ethn, homeless

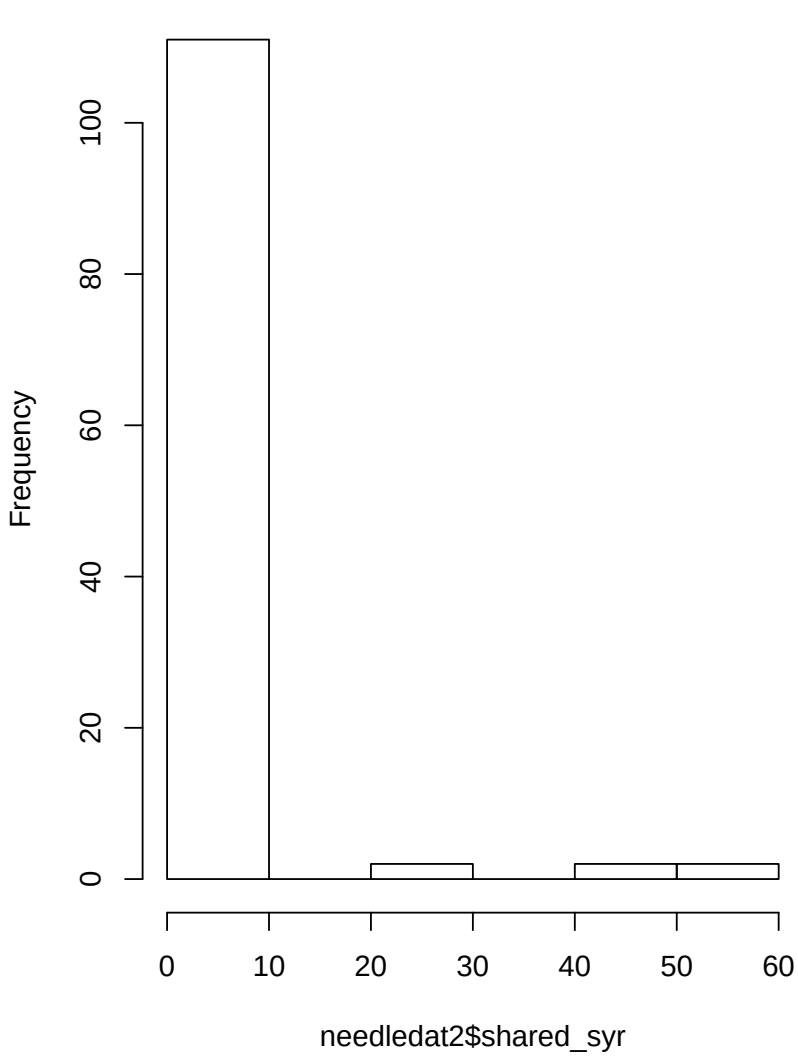
```
needledat = read.csv("needle_sharing.csv")
needledat2 <- needledat[needledat$sex %in% c("M", "F") &
    needledat$ethn %in% c("White", "AA", "Hispanic"), ]
summary(needledat2$shared_syr)
```

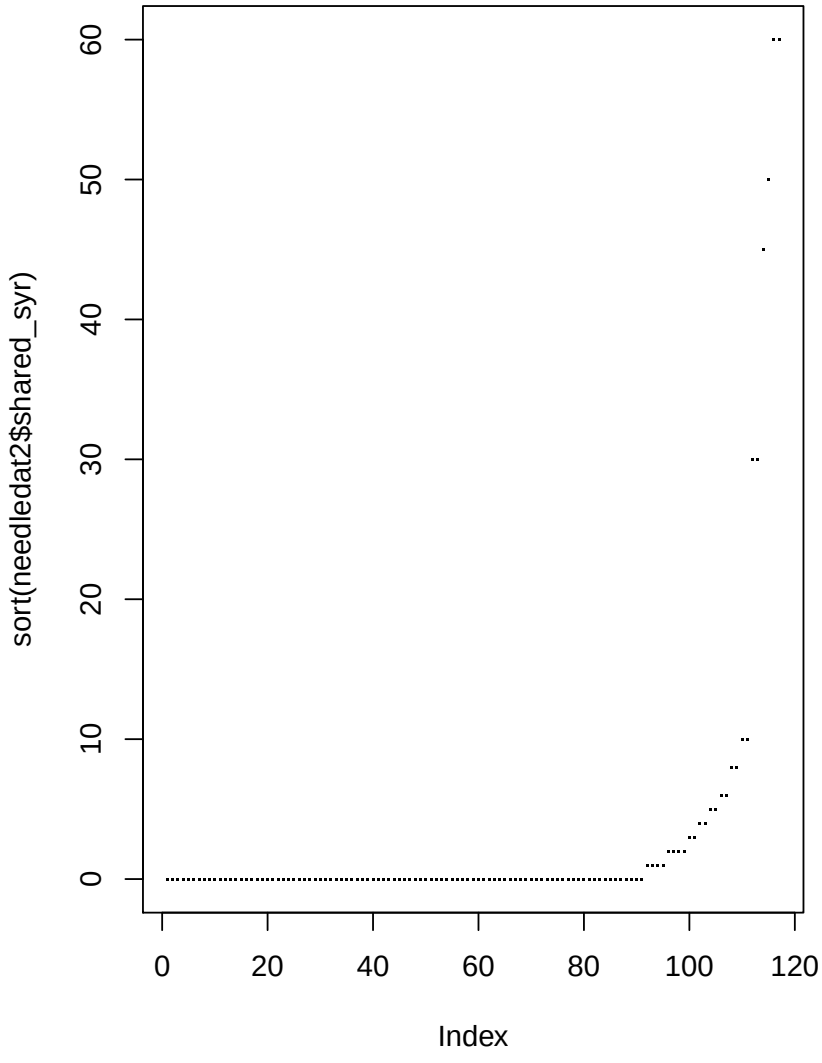
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.000   0.000   0.000   3.068   0.000  60.000       4
```

```
var(needledat2$shared_syr, na.rm=TRUE)
```

```
## [1] 111.5815
```

Example: Risky Drug Use Behavior



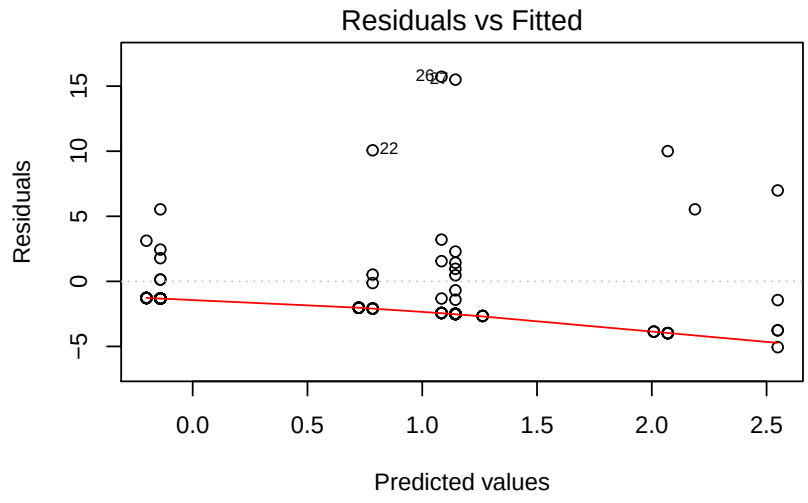


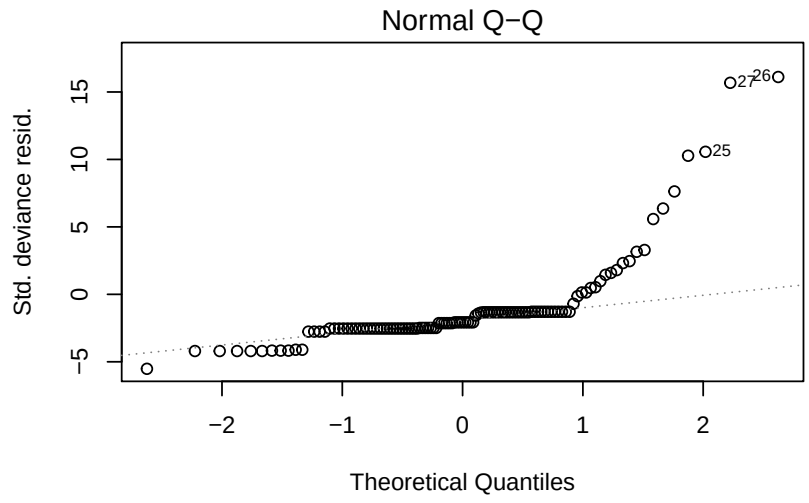
- There are a *lot* of zeros and variance is much greater than mean
  - Poisson model is probably not a good fit

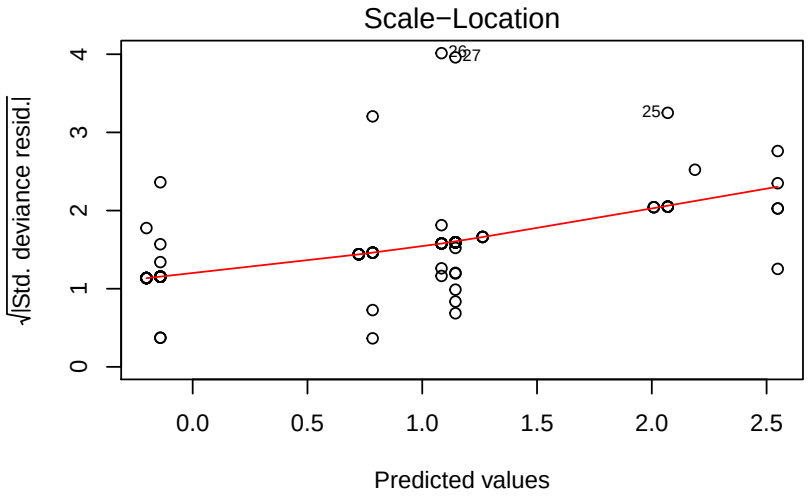
Risky Drug Use Behavior: fitting a Poisson model

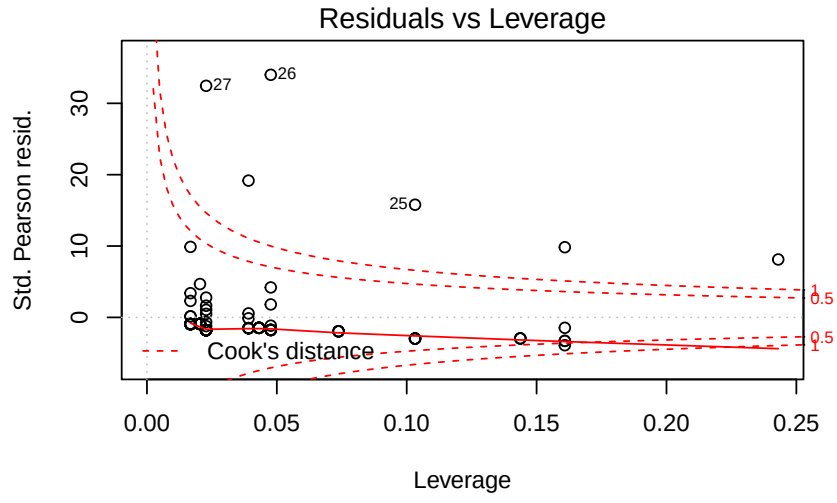
```
fit.pois <- glm(shared_syr ~ sex + ethn + homeless,
                data=needledat2, family=poisson(link="log"))
```

Risky Drug Use Behavior: residuals plots









* Poisson model is definitely not a good fit.

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

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"

Negative binomial GLM

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

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

Or:

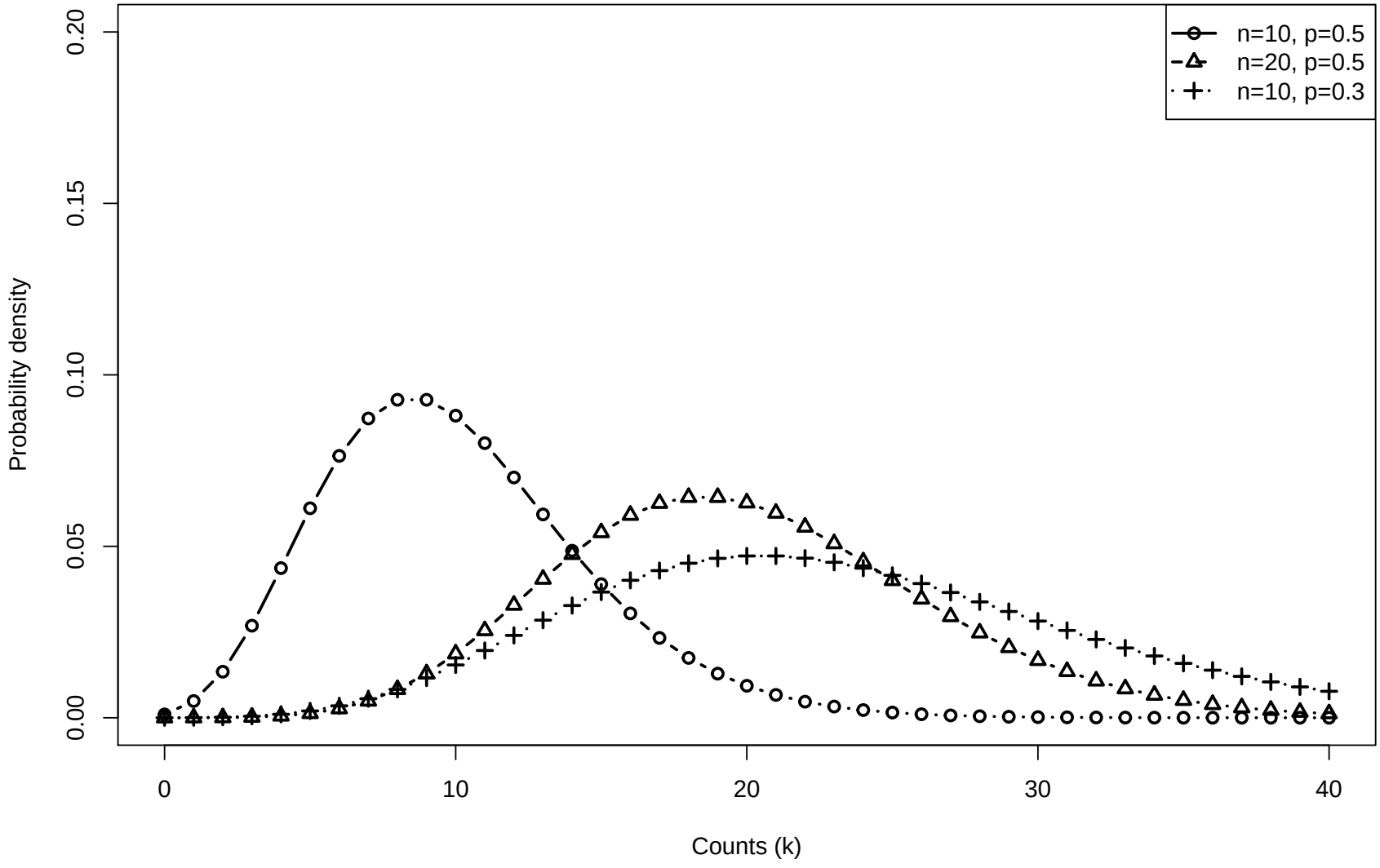
$$\lambda_i = exp\left(\beta_0 + \beta_1 \mathrm{RACE}_i + \beta_2 \mathrm{TRT}_i + \beta_3 \mathrm{ALCH}_i + \beta_4 \mathrm{DRUG}_i\right)$$

And a **random part**:

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

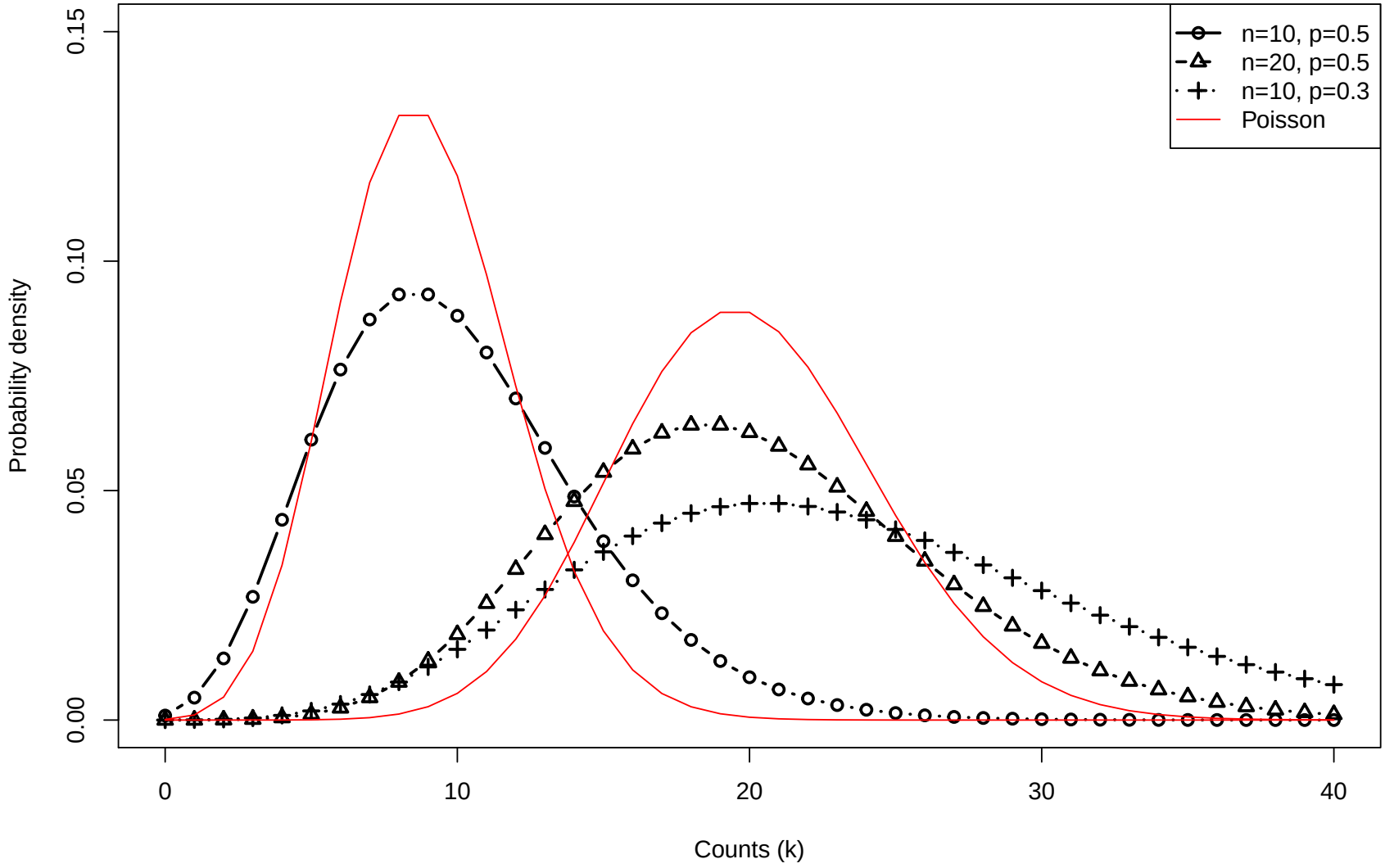
- θ is a **dispersion parameter** that is estimated
- When $\theta = 0$ it is equivalent to Poisson model
- `MASS::glm.nb()` uses this parametrization, `dnbinom()` does not
- The Poisson model can be considered **nested** within the Negative Binomial model

Negative Binomial Random Distribution



Compare Poisson vs. Negative Binomial

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



Risky drug behavior: Negative Binomial Regression

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

```
##
## Call:
## glm.nb(formula = shared_syr ~ sex + ethn + homeless, data = needledat2,
##     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     0.8559   0.542   0.5876
## sexM          -1.0148     0.8294  -1.224   0.2211
## ethnHispanic   1.3424     1.3201   1.017   0.3092
## ethnWhite      0.2429     0.7765   0.313   0.7544
## homelessyes    1.6445     0.7073   2.325   0.0201 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.0774) family taken to be 1)
##
##     Null deviance: 62.365  on 114  degrees of freedom
## Residual deviance: 56.232  on 110  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 306.26
##
## Number of Fisher Scoring iterations: 1
##
##
```

Likelihood ratio test

Recall from class 2 the Deviance:

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

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)
```

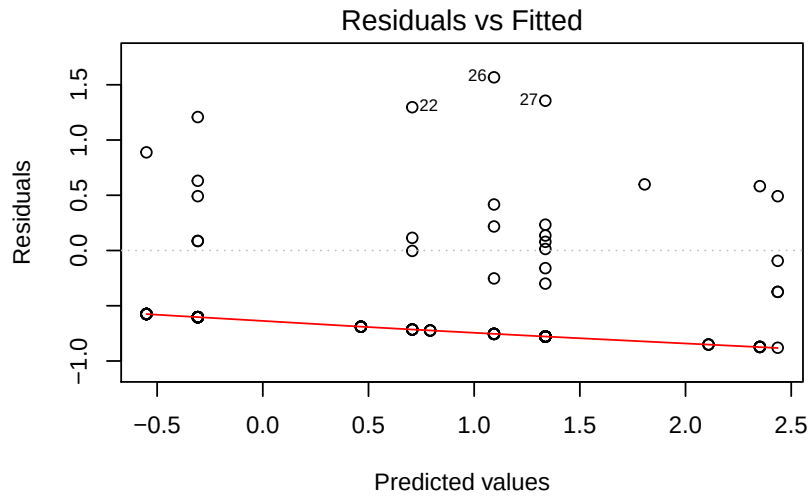
```
(ll.pois <- logLik(fit.pois))
```

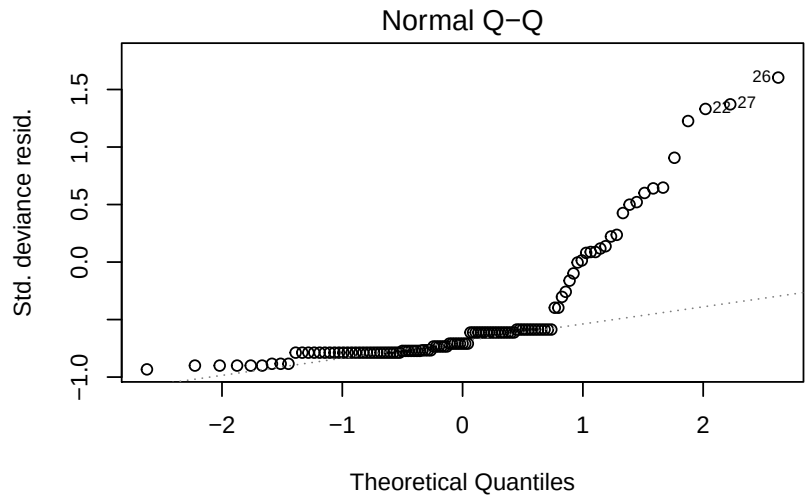
```
## 'log Lik.' -730.0133 (df=5)
```

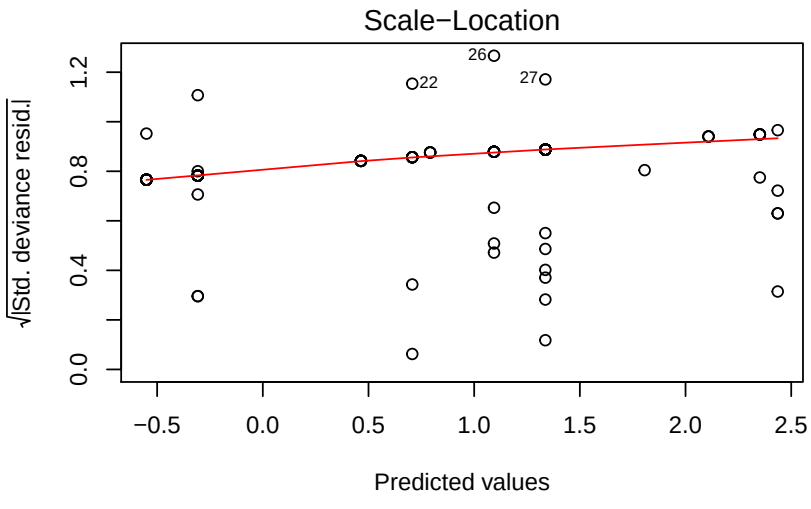
```
pchisq(2 * (ll.negbin - ll.pois), df=1, lower.tail=FALSE)
```

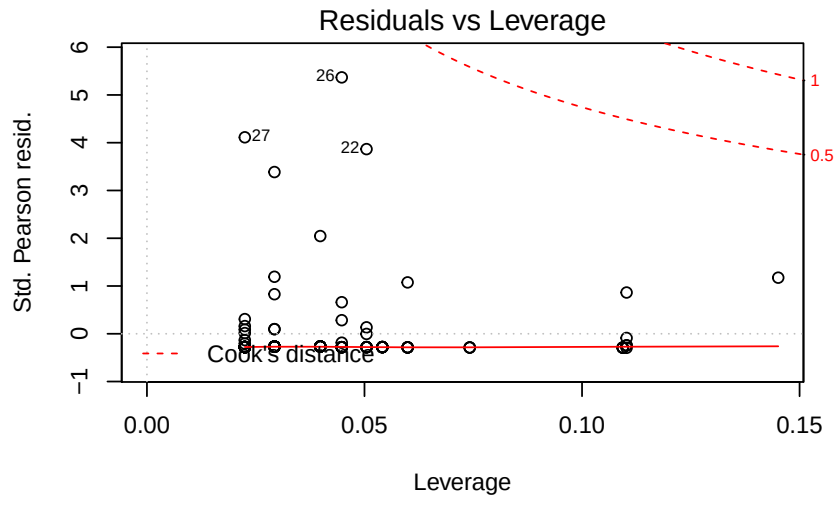
```
## 'log Lik.' 1.675949e-255 (df=6)
```

Risky Drug Use Behavior: NB regression residuals plots





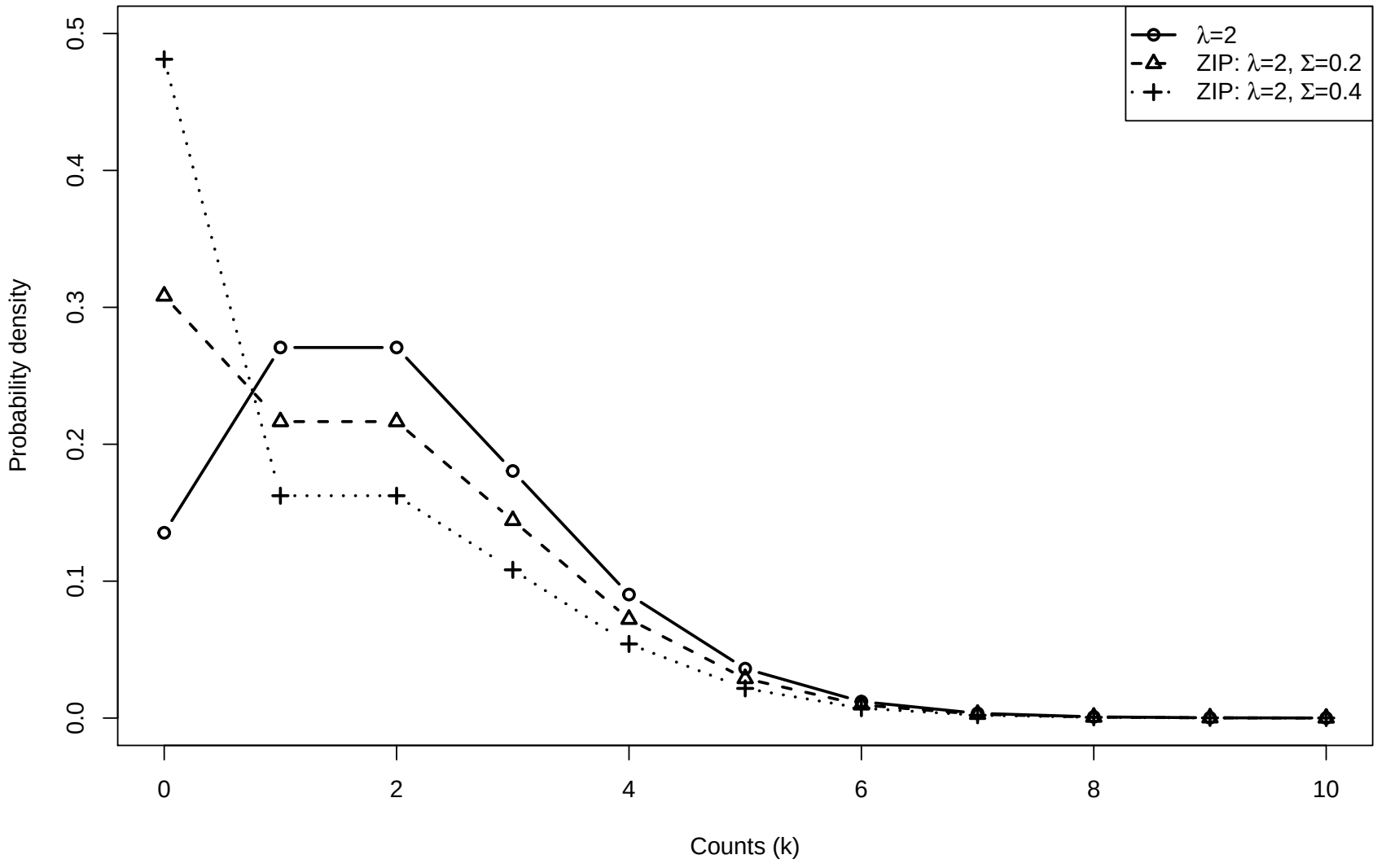




Zero Inflation

- Two-step model:
  1. logistic model to determine whether count is zero or Poisson/NB
  2. Poisson or NB regression distribution for y_i not set to zero by *1.*

Poisson Distribution with Zero Inflation



Risky drug behavior: Zero-inflated Poisson regression

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

Zero-inflated Poisson regression - the model

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

```
##
## Call:
## zeroinfl(formula = shared_syr ~ sex + ethn + homeless, data = needledat2,
##     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)    3.2169     0.1796  17.909  < 2e-16 ***
## sexM          -1.4725     0.1442 -10.212  < 2e-16 ***
## ethnHispanic  -0.1525     0.1576  -0.968 0.333223
## ethnWhite     -0.5236     0.1464  -3.577 0.000347 ***
## homelessyes    1.2034     0.1455   8.268  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.06262    0.65227   3.162  0.00157 **
## sexM         -0.05067    0.58252  -0.087  0.93068
## ethnHispanic -1.76120    0.81177  -2.170  0.03004 *
## ethnWhite    -0.50187    0.56919  -0.882  0.37792
## homelessyes  -0.53013    0.48108  -1.102  0.27048
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 18
## Log-likelihood: -299.8 on 10 Df
```

Risky drug behavior: Zero-inflated Negative Binomial regression

```
fit.ZInegbin <- zeroinfl(shared_syr ~ sex+ethn+homeless,
                         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

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
## -0.5401 -0.3255 -0.2715 -0.1926  5.1489
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.8401     1.1845   2.398  0.01649 *
## sexM          -2.2278     0.9350  -2.382  0.01720 *
## ethnHispanic  -0.4116     0.9832  -0.419  0.67545
## ethnWhite     -0.4294     0.8647  -0.497  0.61949
## homelessyes    1.9461     0.7103   2.740  0.00615 **
## Log(theta)    -1.1972     0.5159  -2.320  0.02032 *
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.6863     0.8466   1.992   0.0464 *
## sexM          -0.9919     0.8016  -1.237   0.2159
## ethnHispanic -11.3556   112.8675  -0.101   0.9199
## ethnWhite     -0.7452     0.7304  -1.020   0.3076
## homelessyes    0.3555     0.7397   0.481   0.6308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.302
## Number of iterations in BFGS optimization: 37
## Log-likelihood: -142.8 on 11 Df
```

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:

```
needledat2$hiv <- factor(ifelse(needledat2$hivstat==0,
                                "negative","positive"))
fit.ZInb2<-zeroinfl(shared_syr~sex+ethn+homeless+hiv|hiv,
                    dist="negbin", data=needledat2)
```

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
## -0.4419 -0.3295 -0.3206 -0.3015  6.0894
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.7521     1.2267   3.059  0.00222 **
## sexM          -1.8727     0.7635  -2.453  0.01418 *
## ethnHispanic  -1.2466     0.9693  -1.286  0.19841
## ethnWhite     -1.2869     0.8436  -1.526  0.12712
## homelessyes    0.9184     0.6822   1.346  0.17827
## hivpositive    1.7342     0.8175   2.121  0.03388 *
## Log(theta)    -0.4561     0.5337  -0.854  0.39287
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0418     0.3515   2.964  0.00304 **
## hivpositive  -0.7252     0.7342  -0.988  0.32327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.6338
## Number of iterations in BFGS optimization: 65
## Log-likelihood: -127.9 on 9 Df
```

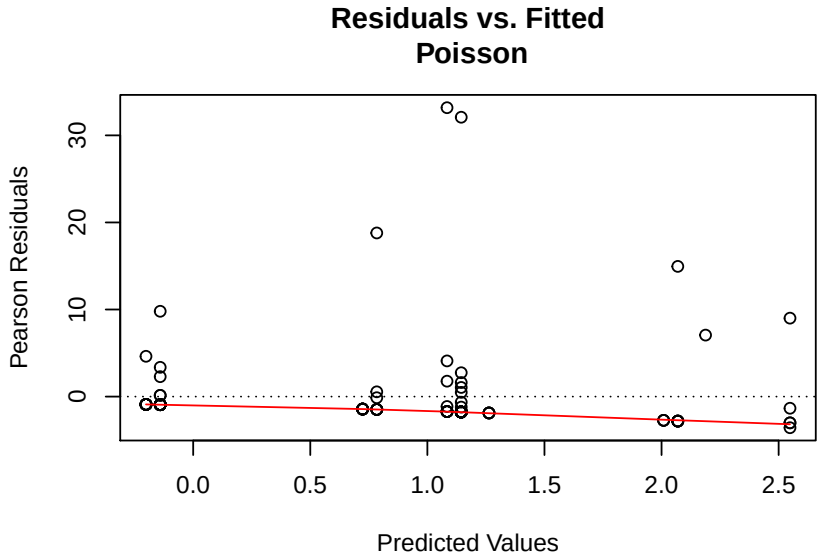
Intercept-only zero-inflation model

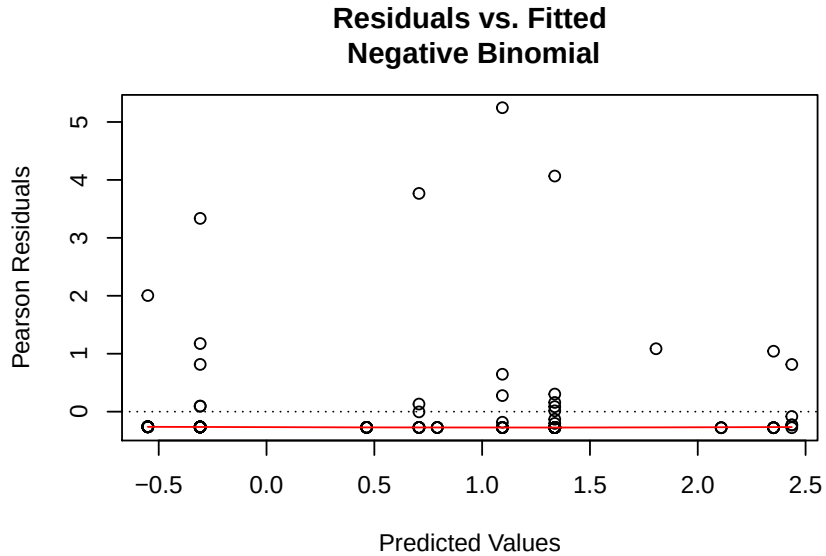
```
fit.ZInb3 <- zeroinfl(shared_syr~sex+ethn+homeless|1,
                      dist="negbin", data=needledat2)
summary(fit.ZInb3)
```

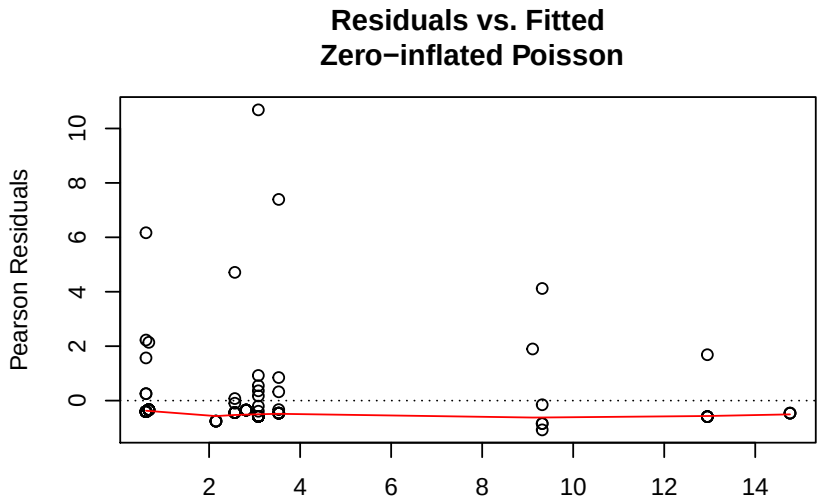
```
##
## Call:
## zeroinfl(formula = shared_syr ~ sex + ethn + homeless | 1, data = needledat2,
##     dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.3159 -0.3123 -0.3040 -0.2953  5.2941
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.08551    1.42665   1.462   0.1438
## sexM         -1.43812    0.89188  -1.612   0.1069
## ethnHispanic  0.48126    1.16639   0.413   0.6799
## ethnWhite    -0.07421    0.81066  -0.092   0.9271
## homelessyes   1.62076    0.67705   2.394   0.0167 *
## Log(theta)   -1.12533    0.89365  -1.259   0.2079
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5211     0.7599   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: 37
## Log-likelihood: -146.8 on 7 Df
```

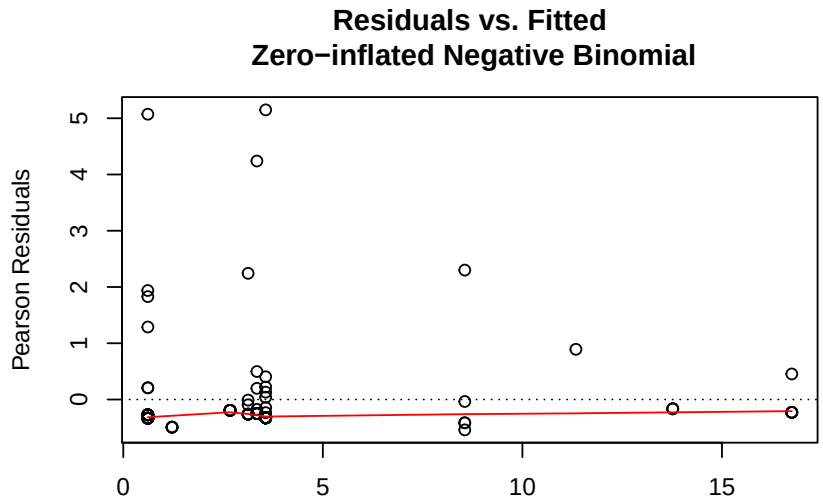
Residuals vs. fitted values

I invisibly define functions plotpanel1 and plotpanel2 that will work for all types of models (see .R or .Rmd file for functions). These use Pearson residuals.

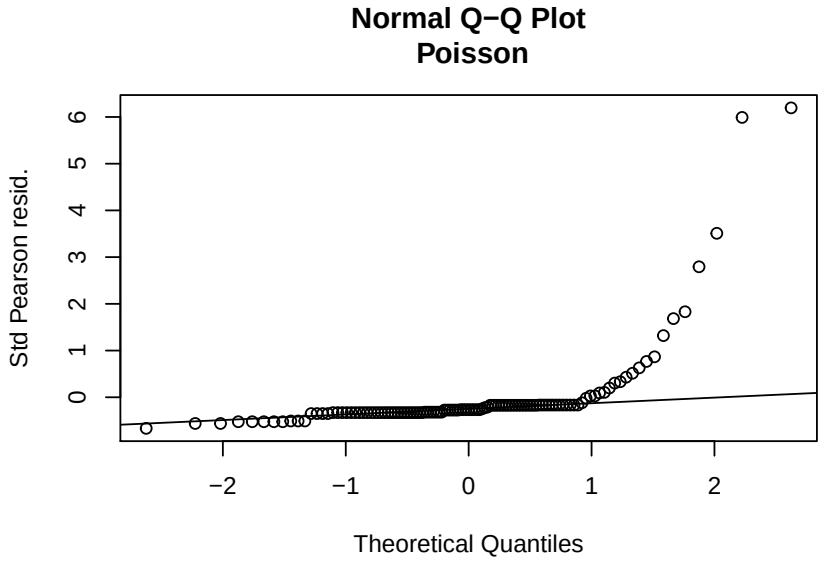


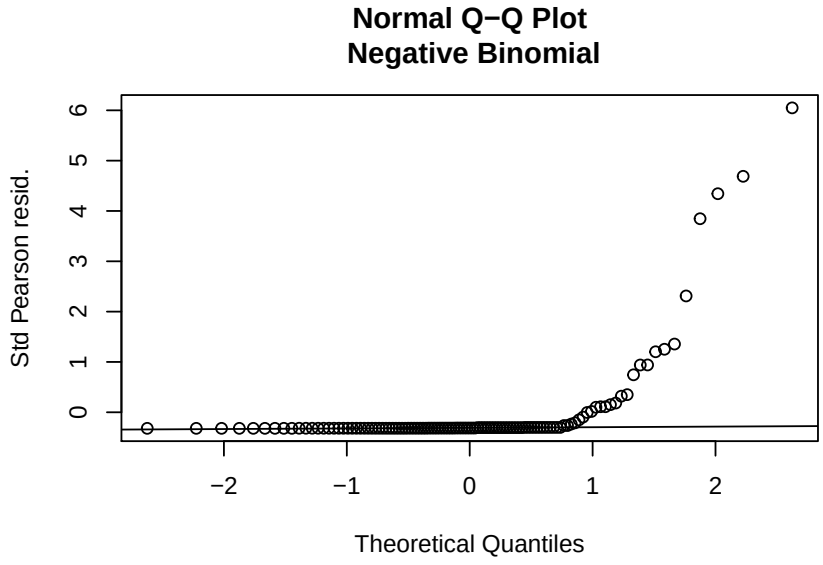


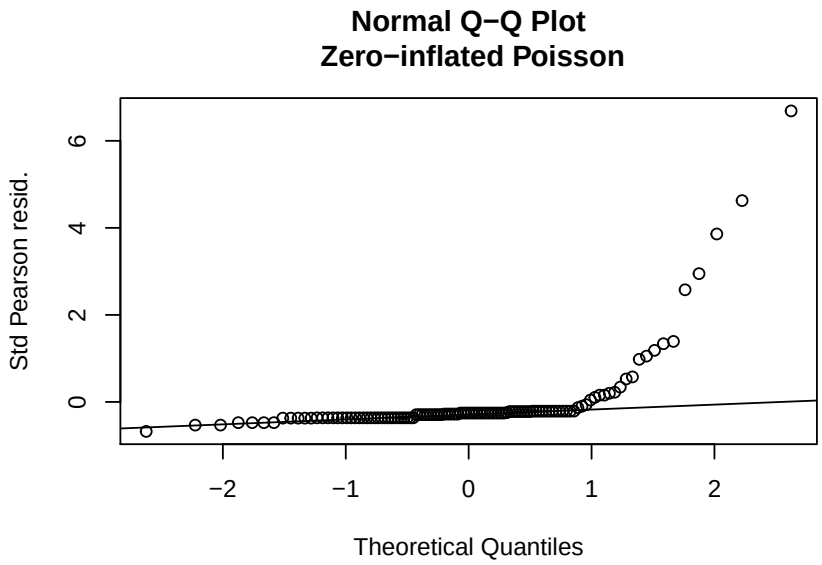


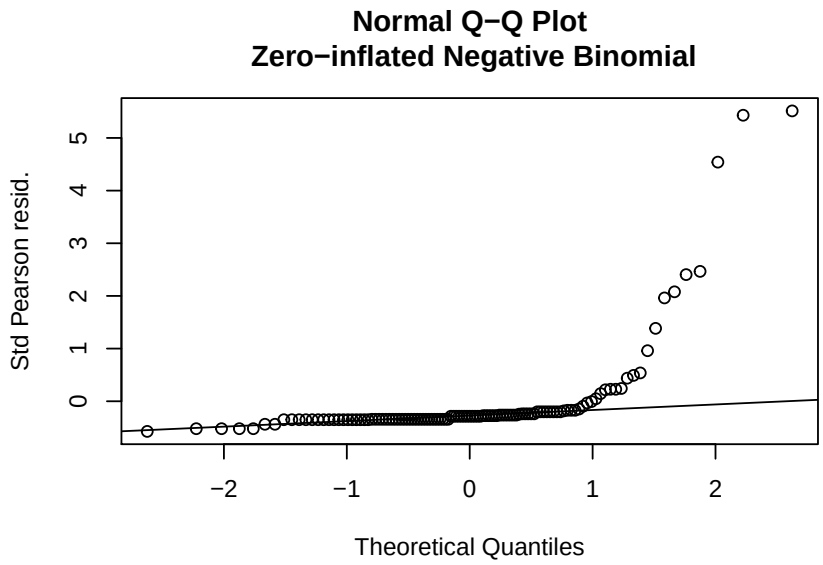


Quantile-quantile plots for residuals









still over-dispersed - ideas?

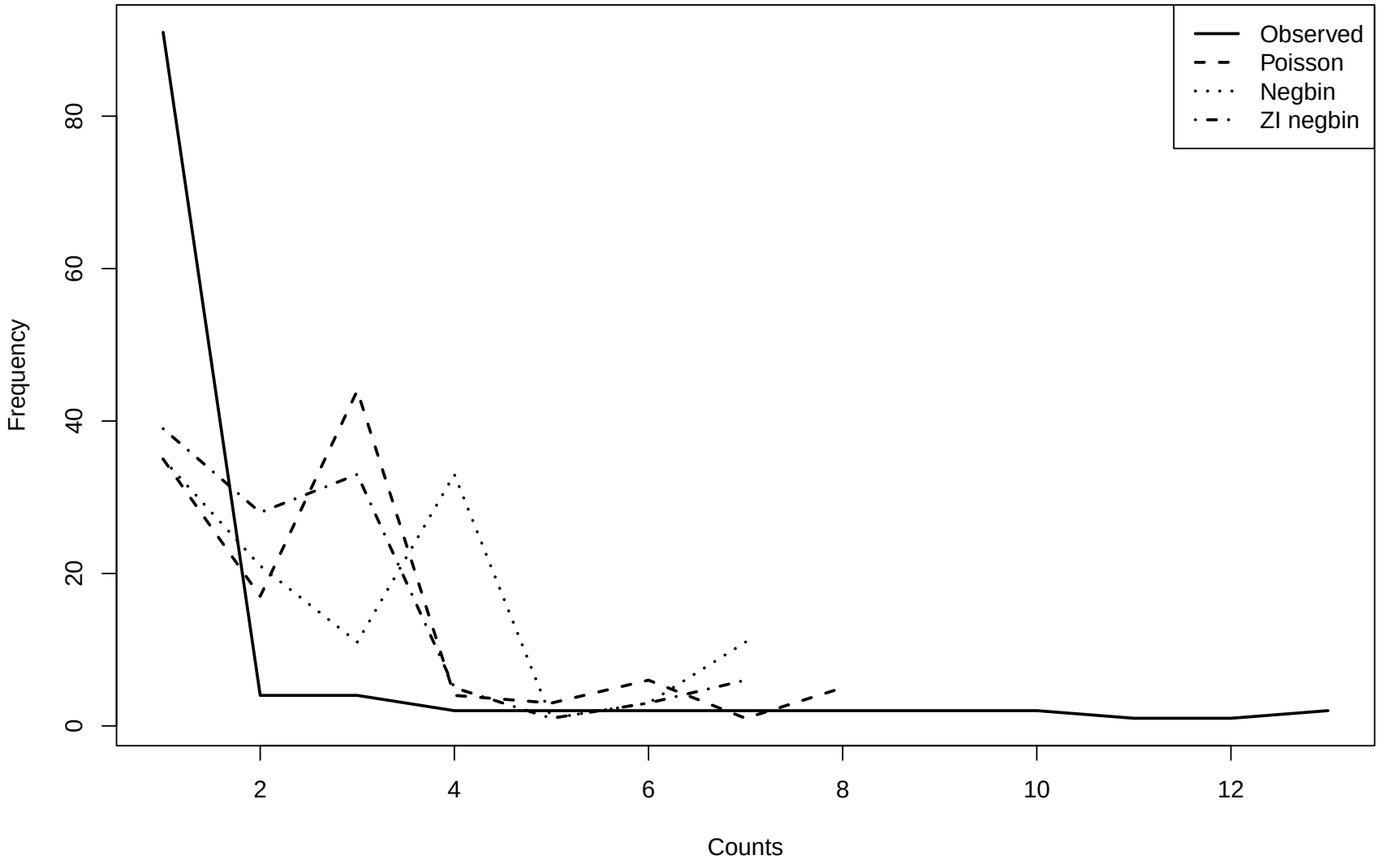
Inference from the different models

Table 1:

	Dependent variable:				
	shared_syr				
	Poisson	*negative binomial*		*zero-inflated count data*	
	(1)	(2)	(3)	(4)	(5)
sexM	−0.925***	−1.015	−1.473***	−2.228**	−1.438
	(0.121)	(0.829)	(0.144)	(0.935)	(0.892)
ethnHispanic	1.465***	1.342	−0.152	−0.412	0.481
	(0.160)	(1.320)	(0.158)	(0.983)	(1.166)
ethnWhite	0.061	0.243	−0.524***	−0.429	−0.074
	(0.133)	(0.776)	(0.146)	(0.865)	(0.811)
homelessyes	1.285***	1.644**	1.203***	1.946***	1.621**
	(0.127)	(0.707)	(0.146)	(0.710)	(0.677)
Constant	0.723***	0.464	3.217***	2.840**	2.086
	(0.145)	(0.856)	(0.180)	(1.184)	(1.427)
Observations	115	115	115	115	115
Log Likelihood	−730.013	−148.128	−299.790	−142.750	−146.768
θ		0.077*** (0.018)			
Akaike Inf. Crit.	1,470.027	306.255			
Note:					*p<0.1; **p<0.05; ***p<0.01

Zero-inflated models are 3) Poisson, 4) Negative Binomial, and 5) Negative Binomial with intercept-only zero inflation model.

Example of plotting observed and predicted counts



Lab exercises

- Perform chi-square nested deviance tests for zero-inflated models
- Try fitting the needle dataset using a zero-inflated gamma count distribution

Resources for R and SAS

- Short, practical tutrorials 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