BIOS621 Session 4 - loglinear regression part 1

Levi Waldron

Welcome and outline - session 4

- brief review of GLMs
- Motivating example for log-linear models
  - Poisson regression
- Checking model assumptions and fit: Residual Analysis
- Note on collinearity

Reading: Vittinghoff textbook chapter 8.1-8.3

Learning Objectives

- Define log-linear models in GLM framework
- Identify situations that motivate use of log-linear models
- Assess model fit of log-linear models
- Define multi-collinearity

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

Linear Regression as GLM

- **The model**:
  $y_i = E[y|x] + \epsilon_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \epsilon_i$
- **Random component** of y_i is normally distributed:
  $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2_\epsilon)$
- **Systematic component** (linear predictor):
  $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$
- **Link function** here is the *identity link*: $g(E(y|x)) = E(y|x)$.
  We are modeling the mean directly, no transformation.

Logistic Regression as GLM

- **The model**:

$$Logit(P(x)) = log\left(\frac{P(x)}{1-P(x)}\right) = \beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+...+\beta_p x_{pi}$$

- **Random component**: y_i follows a Binomial distribution (outcome is a binary variable)
- **Systematic component**: linear predictor

$$\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$$

- **Link function**: *logit* (log of the odds that the event occurs)

$$g(P(x)) = logit(P(x)) = log\left(\frac{P(x)}{1-P(x)}\right)$$

$$P(x) = g^{-1}\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}\right)$$

Additive vs. Multiplicative models

- Linear regression is an *additive* model
  - *e.g.* for two binary variables $\beta_1 = 1.5$, $\beta_2 = 1.5$.
  - If $x_1 = 1$ and $x_2 = 1$, this adds 3.0 to $E(y|x)$
- Logistic regression is a *multiplicative* model
  - If $x_1 = 1$ and $x_2 = 1$, this adds 3.0 to $log(\frac{P}{1-P})$
  - Odds-ratio $\frac{P}{1-P}$ increases 20-fold: $exp(1.5 + 1.5)$ or $exp(1.5) * exp(1.5)$

Motivating example for log-linear models

- Effectiveness of a new case-management program for depression
  - can the new treatment reduce the number of needed visits to the emergency room, compared to standard care?
- *outcome*: # of emergency room visits for each patient in the year following initial treatment
- *predictors*: *race* (white or nonwhite), *treatment* (treated or control), *amount of alcohol consumption* (numerical measure), *drug use* (numerical measure)

Motivating example (cont'd)

- Statistical issues:
  - about $1/3$ of observations are exactly 0 (did not return to the emergency room within the year)
  - highly nonnormal and cannot be transformed to be approximately normal
  - even $log(y_i + 1)$ transformation will have a "lump" at zero
    - over $1/2$ the transformed data would have values of 0 or $log(2)$
  - a linear regression model would give negative predictions for some covariate combinations
  - some subjects die or cannot be followed up on for a whole year

Motivating example (cont'd)

- A *multiplicative* model will allow us to make inference on *ratios* of mean emergency room usage
- Modeling *log* of the *mean* emergency usage ensures positive means, and does not suffer from $log(0)$ problem
- Random component of GLM, or residuals (was $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ for linear regression) may still not be normal, but we can choose from other distributions

Motivating example: proposed model without time

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

Or equivalently:

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

where $E[Y_i]$ is the expected number of emergency room visits for patient i.

- Important note: Modeling $log(E[Y_i])$ is *not* equivalent to modeling $E(log(Y_i))$

Motivating example: accounting for time of follow-up

Instead, model mean count per unit time:

$$log(E[Y_i]/t_i) = \beta_0 + \beta_1 \mathrm{RACE}_i + \beta_2 \mathrm{TRT}_i + \beta_3 \mathrm{ALCH}_i + \beta_4 \mathrm{DRUG}_i$$

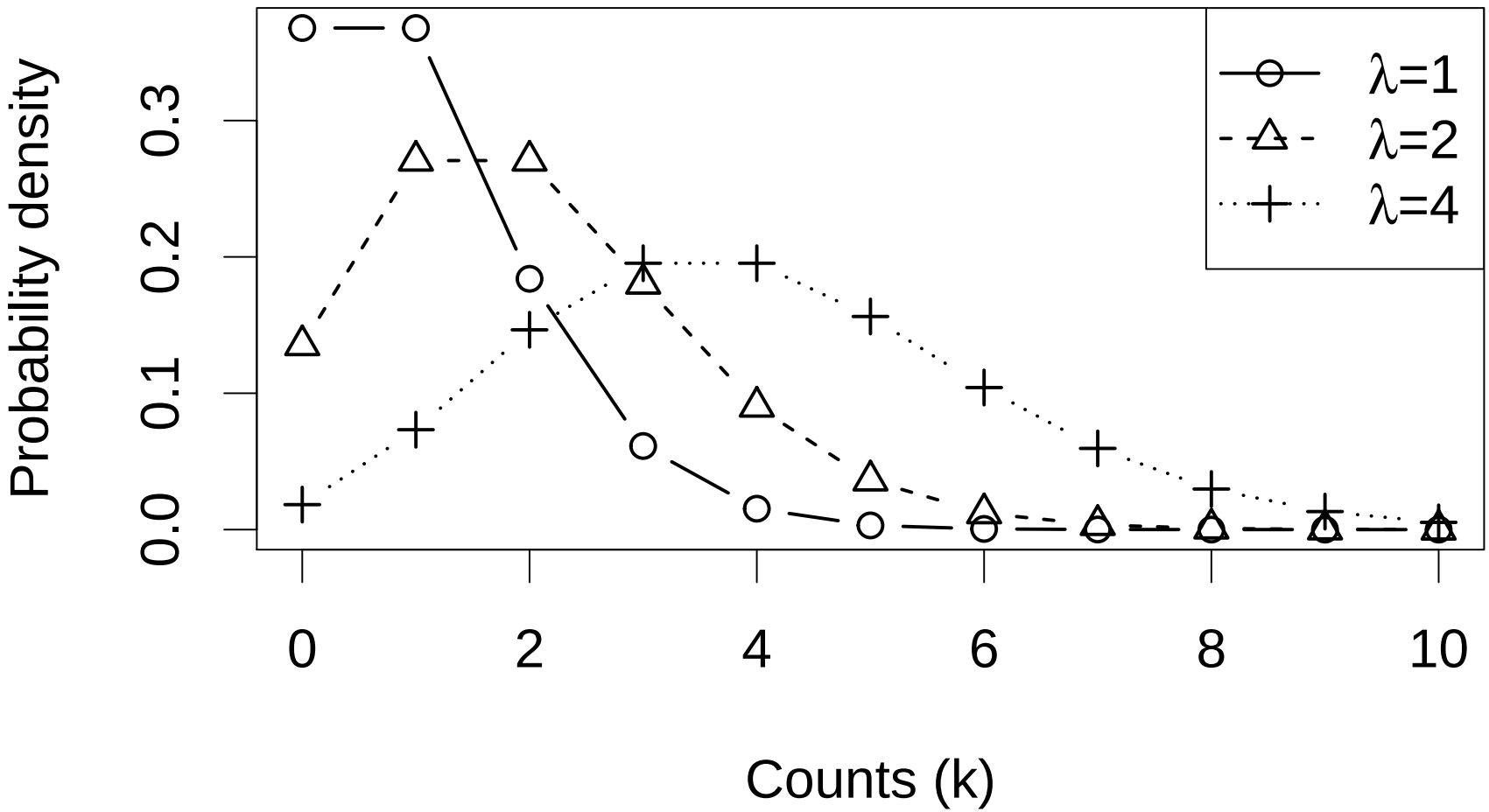
Or equivalently:

$$log(E[Y_i]) = \beta_0 + \beta_1 \mathrm{RACE}_i + \beta_2 \mathrm{TRT}_i + \beta_3 \mathrm{ALCH}_i + \beta_4 \mathrm{DRUG}_i + log(t_i)$$

- $log(t_i)$ is not a covariate, it is called an *offset*

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



Motivating example: the Poisson GLM

- Model the number of counts per unit time as Poisson-distributed
  - so the expected number of counts per time is λ_i

$$E[Y_i]/t_i = \lambda_i$$
$$log(E[Y_i]/t_i) = log(\lambda_i)$$
$$log(E[Y_i]) = log(\lambda_i) + log(t_i)$$

Recalling the log-linear model systematic component:

$$log(E[Y_i]) = \beta_0 + \beta_1 \mathrm{RACE}_i + \beta_2 \mathrm{TRT}_i + \beta_3 \mathrm{ALCH}_i + \beta_4 \mathrm{DRUG}_i + log(t_i)$$

Motivating example: the Poisson GLM

Then 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:

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

Motivating example: interpretation of coefficients

- Suppose that $\hat{\beta}_1 = -0.5$ in the fitted model, where $\mathrm{RACE}_i = 0$ for white and $\mathrm{RACE}_i = 1$ for non-white.
- The mean rate of emergency room visits per unit time for white relative to non-white, all else held equal, is estimated to be:

$$\frac{exp\left(\beta_0 + 0 + \beta_2 \mathrm{TRT}_i + \beta_3 \mathrm{ALCH}_i + \beta_4 \mathrm{DRUG}_i\right)}{exp\left(\beta_0 - 0.5 + \beta_2 \mathrm{TRT}_i + \beta_3 \mathrm{ALCH}_i + \beta_4 \mathrm{DRUG}_i\right)}$$

$$= \frac{e^{\beta_0} e^{0} e^{\beta_2 \mathrm{TRT}_i} e^{\beta_3 \mathrm{ALCH}_i} e^{\beta_4 \mathrm{DRUG}_i}}{e^{\beta_0} e^{-0.5} e^{\beta_2 \mathrm{TRT}_i} e^{\beta_3 \mathrm{ALCH}_i} e^{\beta_4 \mathrm{DRUG}_i}}$$

$$= \frac{e^{0}}{e^{-0.5}}$$

$$= e^{0.5} \approxeq 1.65$$

Motivating example: interpretation of coefficients

- If $\hat{\beta}_1 = -0.5$ with whites as the reference group:
  - after adjustment for treatment group, alcohol and drug usage, whites tend to use the emergency room at a rate 1.65 times higher than non-whites.
  - equivalently, the average rate of usage for whites is 65% higher than that for non-whites
- Multiplicative rules apply for other coefficients as well, because they are exponentiated to estimate the mean rate.

Example by simulation

```r
simdat <- data.frame(race=sample(c("white", "non-white"), size=10000, replace=TRUE))
simdat$race <- factor(simdat$race, levels=c("white", "non-white"))
simdat$y <- rpois(10000, lambda=ifelse(simdat$race=="white", exp(3.5), exp(3)))
fit <- glm(y ~ race, data=simdat, family=poisson("log"))
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ race, family = poisson("log"), data = simdat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5323  -0.7127  -0.0246   0.6616   3.5473
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.500139   0.002446  1431.0   <2e-16 ***
## racenon-white -0.498900   0.004003  -124.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26157  on 9999  degrees of freedom
## Residual deviance: 10111  on 9998  degrees of freedom
## AIC: 60885
##
## Number of Fisher Scoring iterations: 4
```

Inference on deviance residuals 1: compare nested models

- The difference in total deviance between two nested models is χ^2 distributed under H_0 that the more complex model is no better at explaining the response.
  - The difference in deviance residuals is (26157 - 10111) = 16046, with a difference of 1 degrees of freedom.

The critical threshold for rejection at p=0.05 is:

```
qchisq(0.95, df=1)
```

```
## [1] 3.841459
```

So we reject H_0

Inference on deviance residuals 2: test for fit

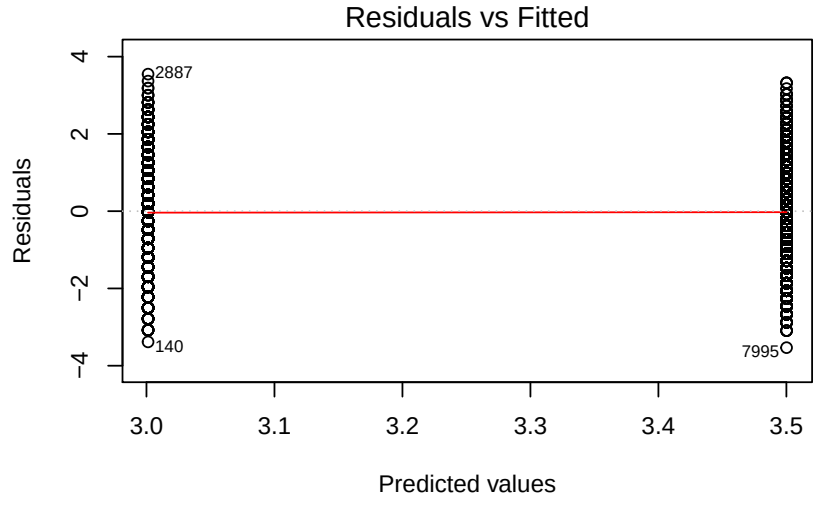
- Total residual deviance is χ^2 distributed if the model is correctly specified
  - What is the critical value for rejecting H_0 at $p < 0.05$ with a χ^2 distribution of 9998 degrees of freedom?

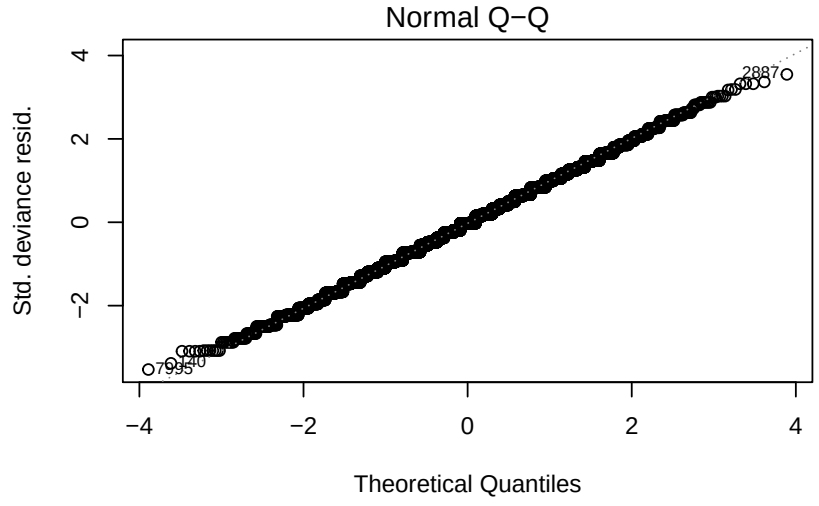
```
qchisq(0.95, df=9998)
```

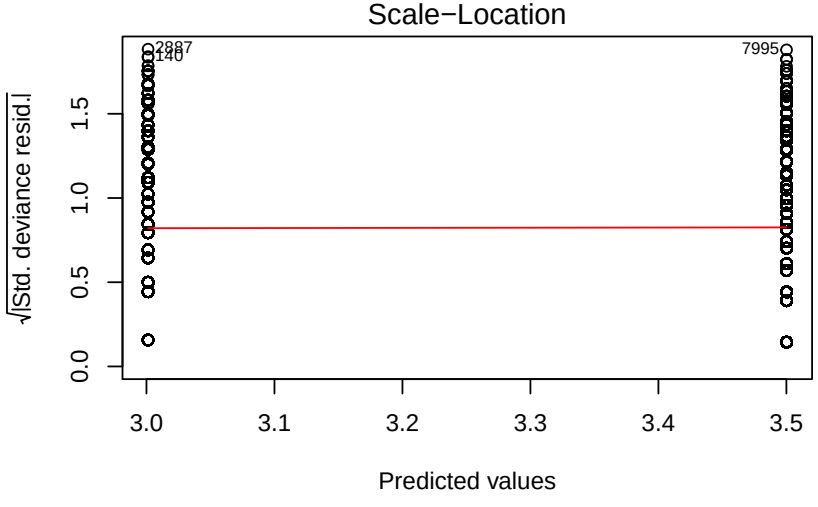
```
## [1] 10231.73
```

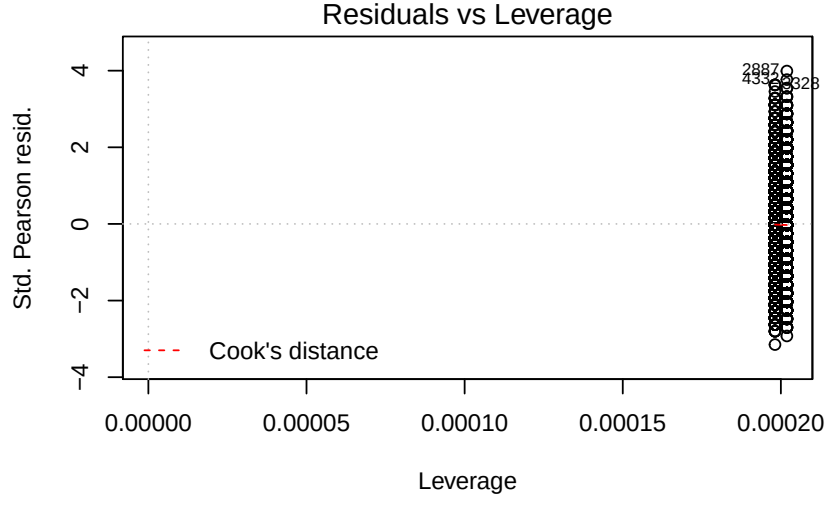
Here total residual deviance is 10111, so we do *not* exceed the threshold and do not reject H_0 that the model is correctly specified.

Example by simulation: Deviance Residuals Plots









Example: Risky Drug Use Behavior

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

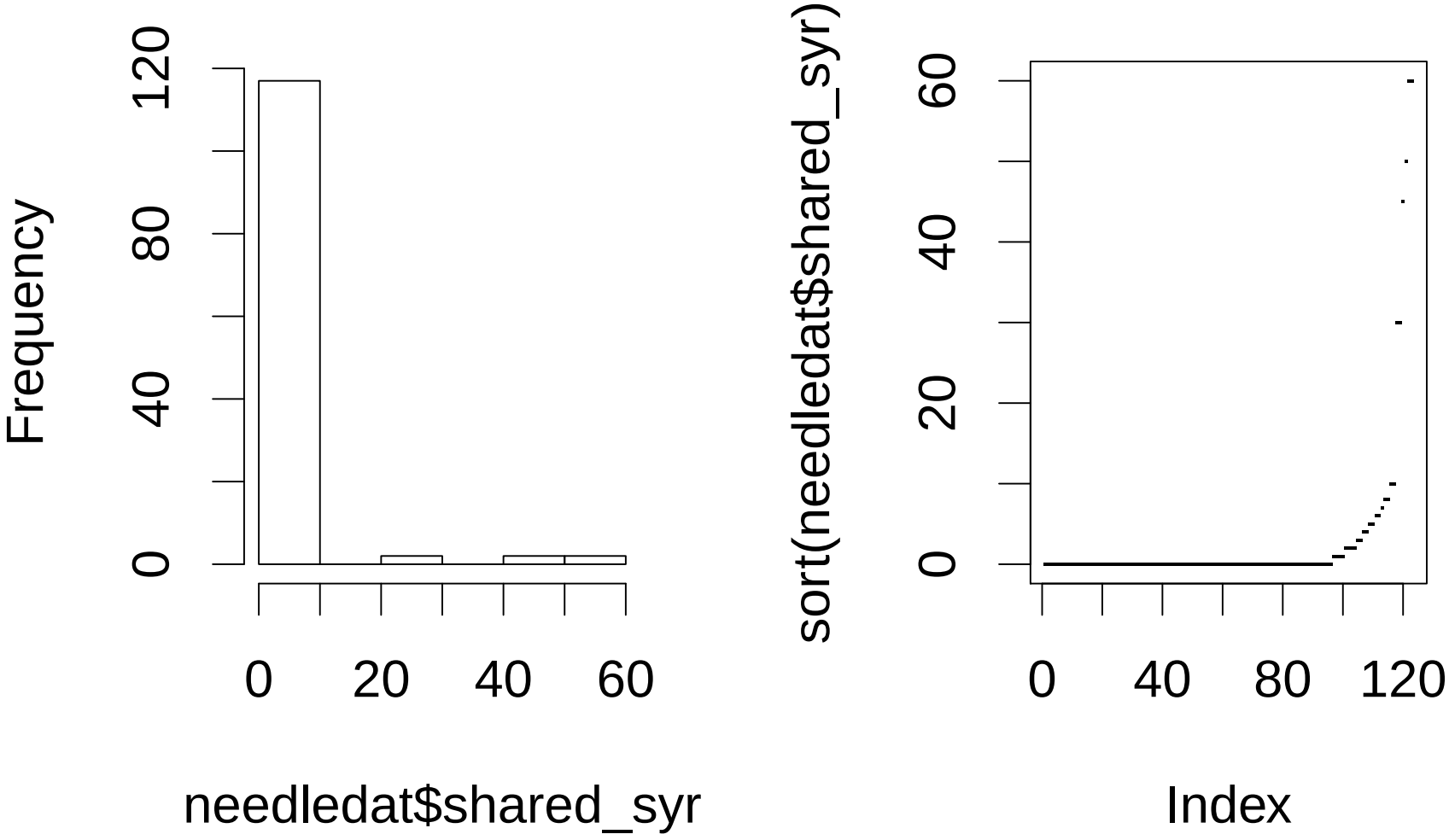
```
needledat = read.csv("needle_sharing.csv")
summary(needledat$shared_syr)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.000   0.000   0.000   2.976   0.000  60.000       5
```

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

```
## [1] 106.5978
```

Example: Risky Drug Use Behavior

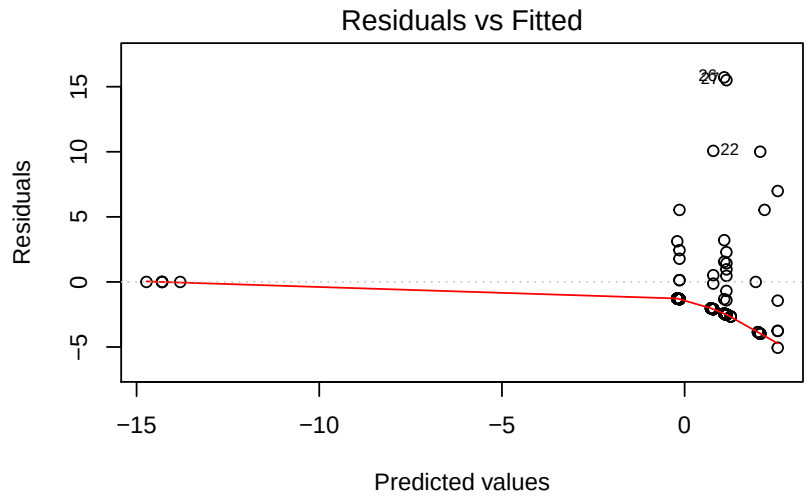


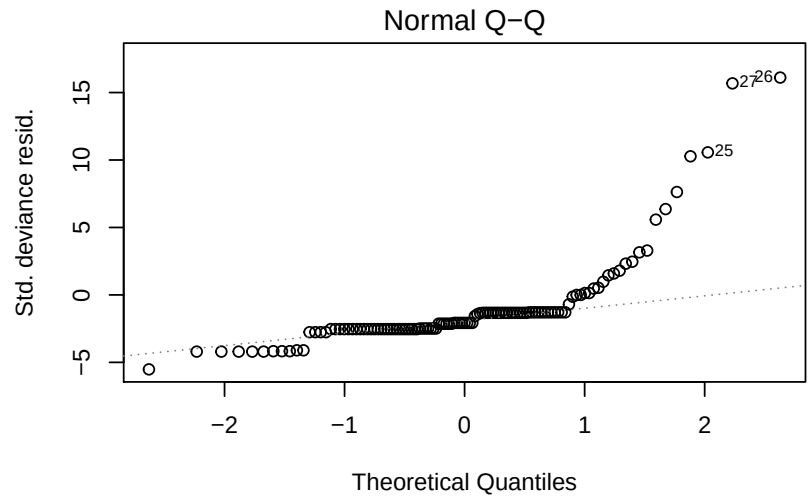
- There are a *lot* of zeros - Poisson model is not a good fit

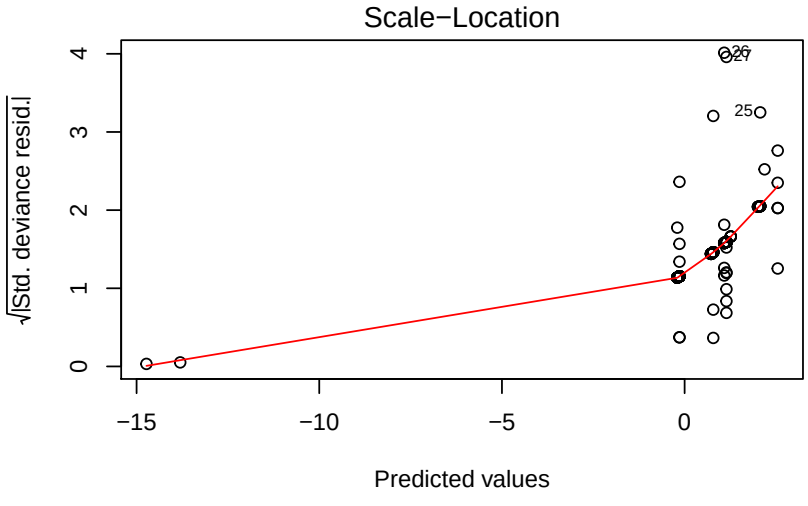
Risky Drug Use Behavior: fitting a Poisson model

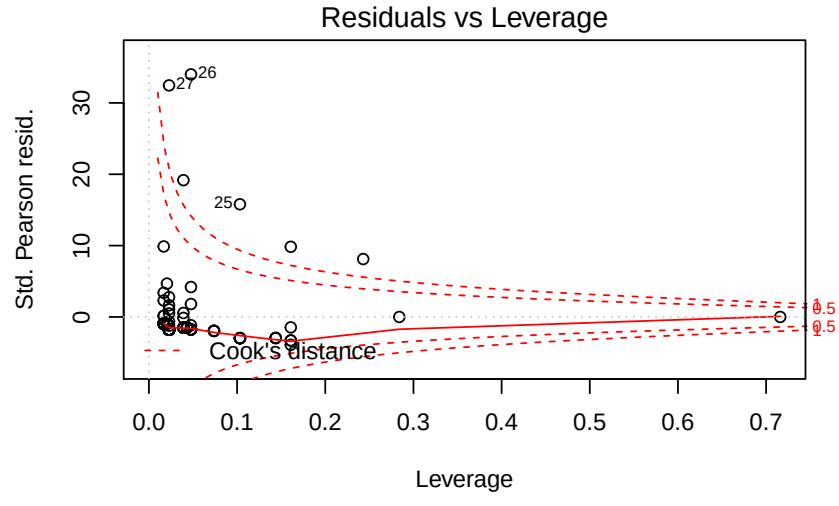
```
##
## Call:
## glm(formula = shared_syr ~ sex + ethn + homeless, family = poisson(link = "log"),
##     data = needledat)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -5.057  -2.506  -2.030  -1.279  15.721
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           0.72332    0.14462   5.002 5.69e-07 ***
## sexM                 -0.92480    0.12133  -7.622 2.50e-14 ***
## sexTrans            -15.08655  773.78384  -0.019   0.9844
## ethnFilipino        -14.52887  510.68253  -0.028   0.9773
## ethnHispanic          1.46454    0.16004   9.151  < 2e-16 ***
## ethnIndian          -14.10111  773.78385  -0.018   0.9855
## ethnIndian & White  -15.02591  773.78384  -0.019   0.9845
## ethnWhite             0.06064    0.13348   0.454   0.6496
## ethnWhite & Hispa     0.86195    0.39872   2.162   0.0306 *
## homelessyes           1.28543    0.12664  10.150  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1621.9  on 120  degrees of freedom
## Residual deviance: 1364.8  on 111  degrees of freedom
##   (7 observations deleted due to missingness)
## AIC: 1483.8
##
## Number of Fisher Scoring iterations: 12
```

Risky Drug Use Behavior: residuals plots









Multicollinearity

- *Multicollinearity* exists when two or more of the independent variables in regression are moderately or highly correlated.
- Multicollinearity implies near-linear relationship among the predictors
- The presence of near-linear dependence dramatically impacts the ability to estimate regression coefficients
- High multicollinearity results in larger standard errors for regression coefficients
  - estimates of such regression coefficients will tend to be less stable over repeated sampling

Concluding notes

- Inference from log-linear models is sensitive to the choice of link function (assumption on distribution of residuals)
- We will cover other options next week for when the Poisson model doesn't fit:
  - Variance proportional to mean, instead of equal
  - Negative Binomial
  - Zero Inflation