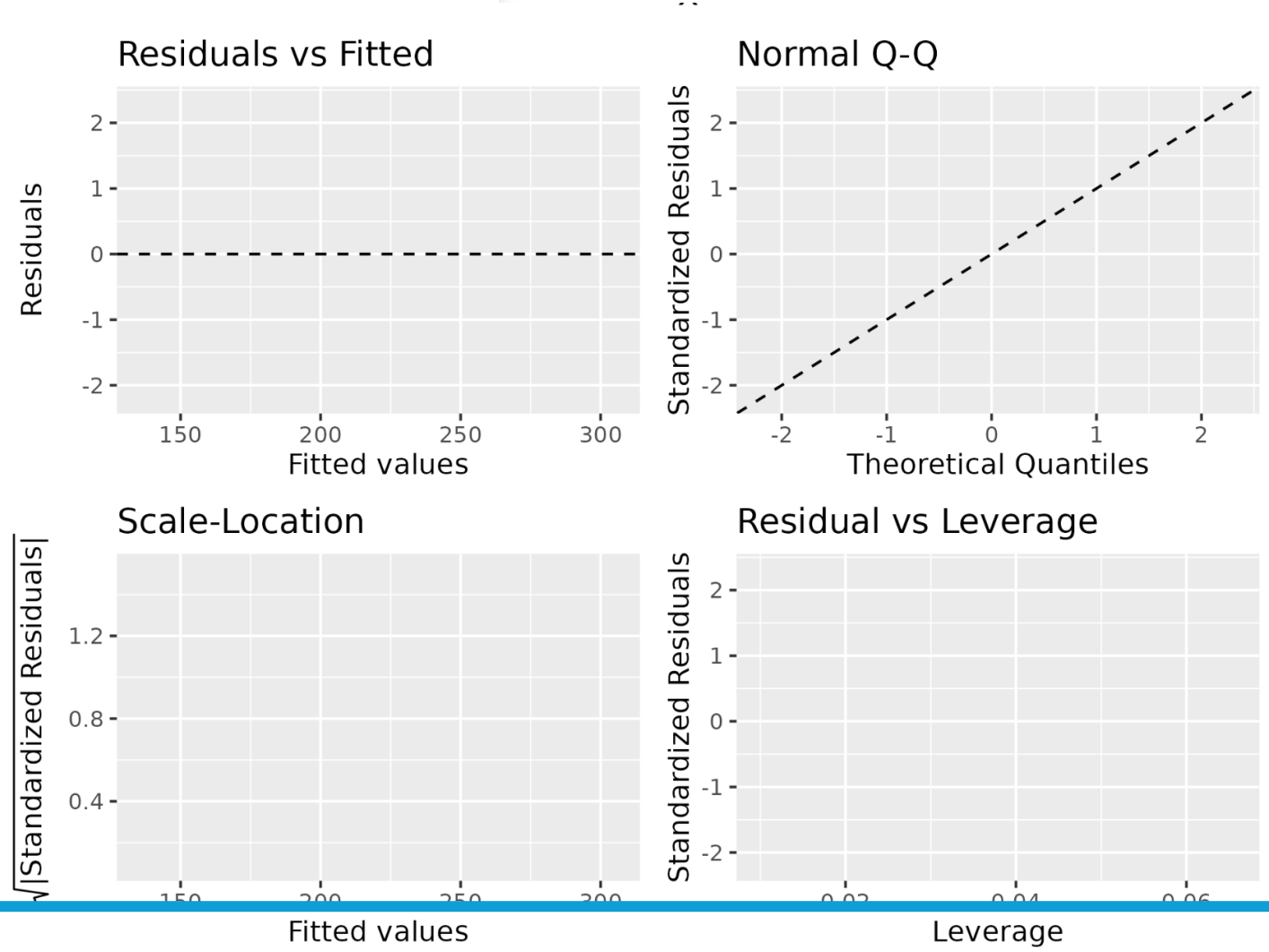
Residuals

Linear Model Residuals



Response Residuals

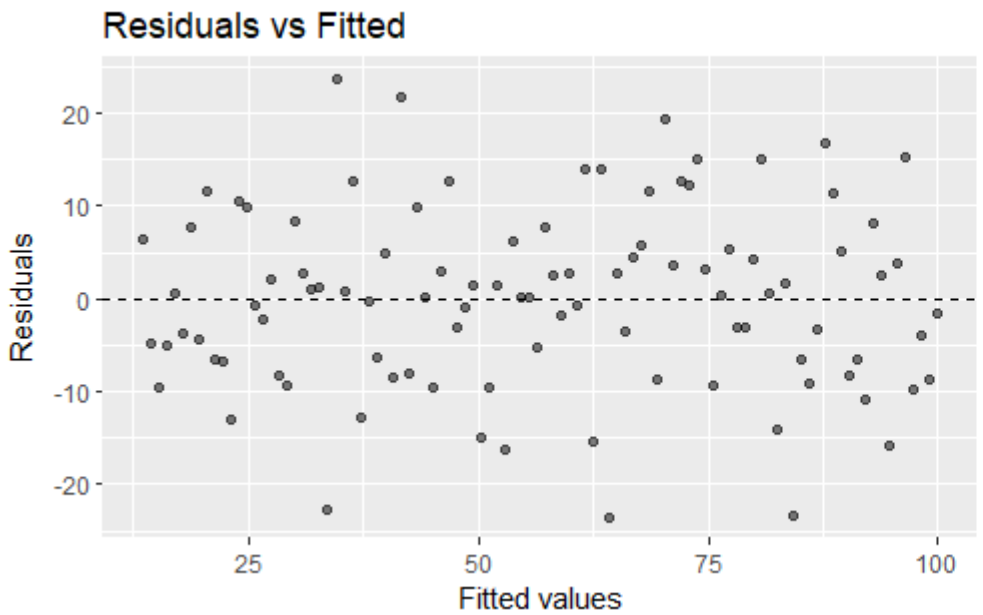
- For a linear model (LM), we looked at response residuals: the observed values minus the predicted values.

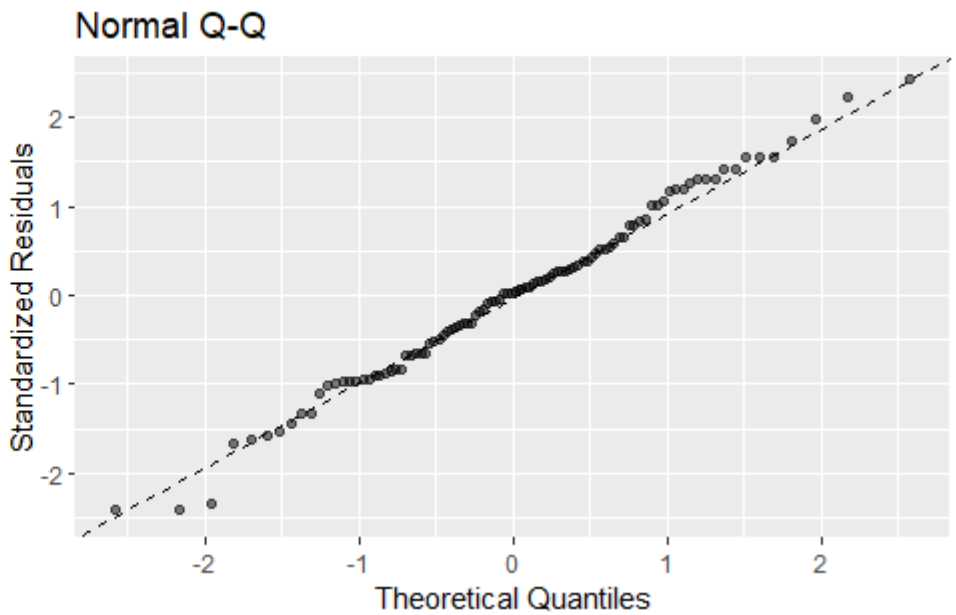
$$r_i = Y_i - \mu_i$$

- Y_i is data point i
- μ_i is the prediction for point i
- r_i is the residual for point i

Response Residuals

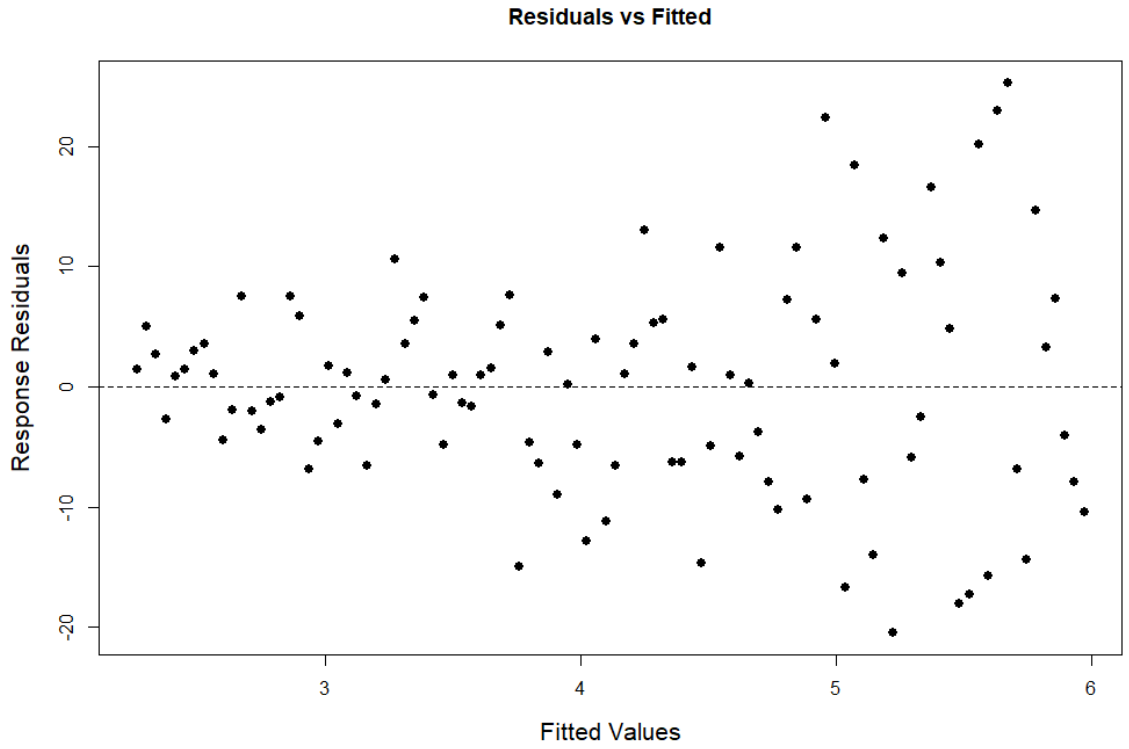
- For a LM, response residuals should be scattered about zero and normally distributed





GLM Response Residuals

- Response Residuals are not helpful for Poisson and logistic GLMs, because the variance is not constant.
- We get non-constant variation and a “fan” or “trumpet” shape in the residuals.



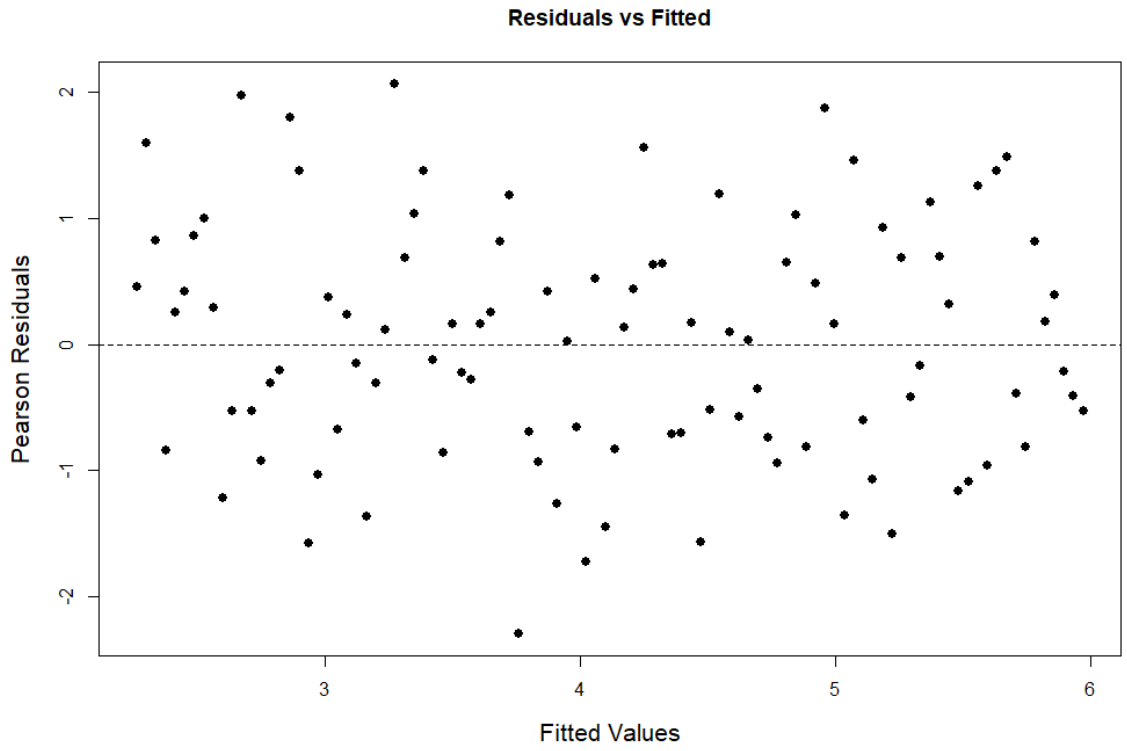
Pearson Residuals

- Pearson Residuals adjust for non-constant variance by dividing response residuals by the modelled standard deviation.
- For a Poisson GLM the formula for a Pearson residual is:

$$\frac{Y_i - \mu_i}{\sqrt{\mu_i}}$$

Pearson Residuals

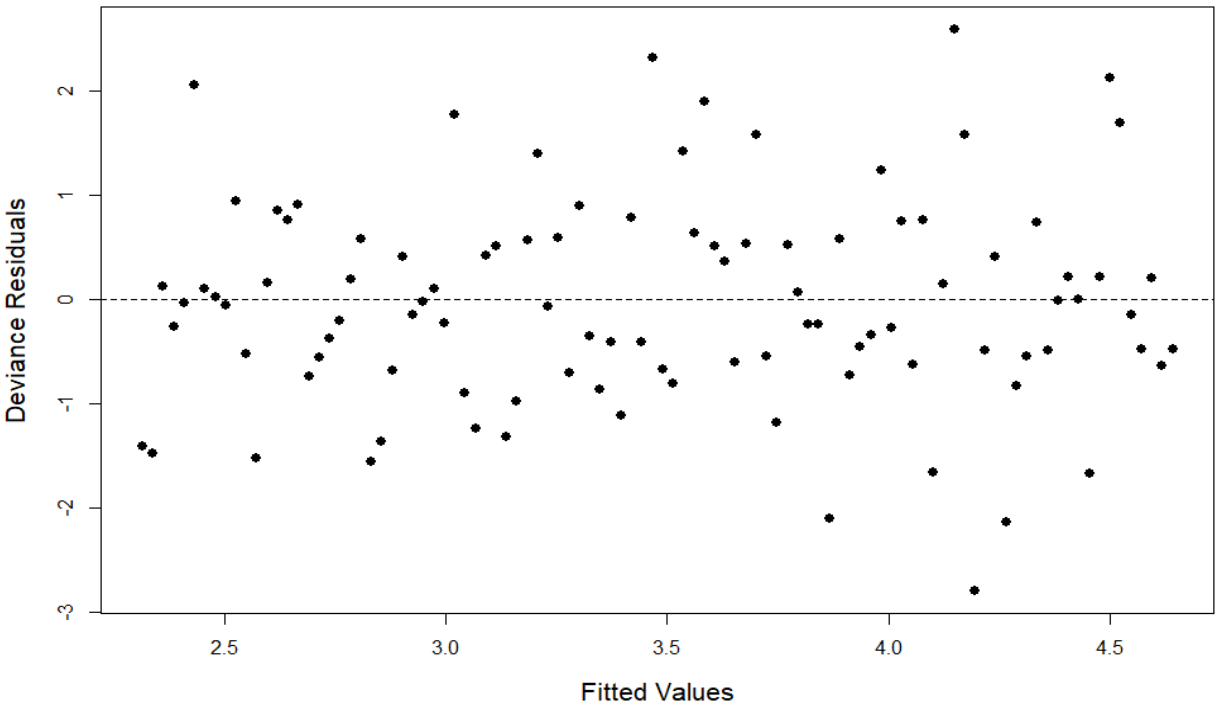
- This plot is for the same model as before. The scatter now looks constant.

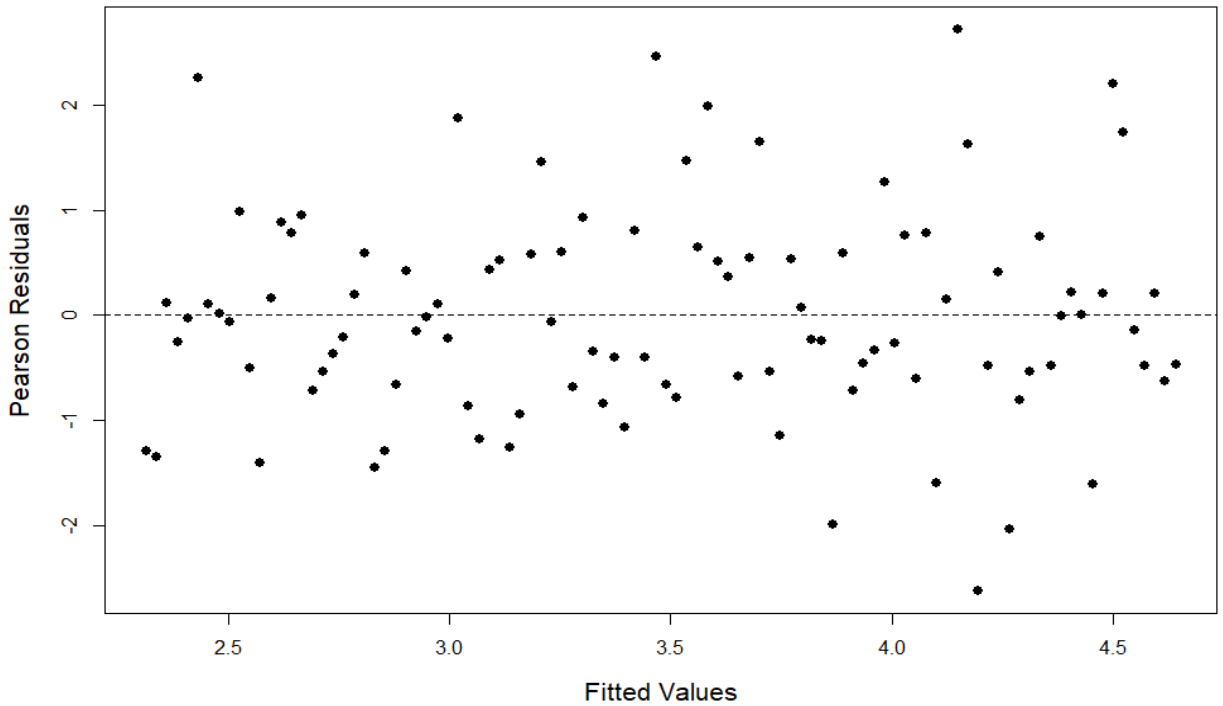


Deviance Residuals

- Deviance Residuals are another type of residuals used with GLMs
- They are related to the concept of deviance which we'll cover shortly
- They are often very similar to Pearson residuals, almost interchangeable.

Pearson v Deviance Residuals



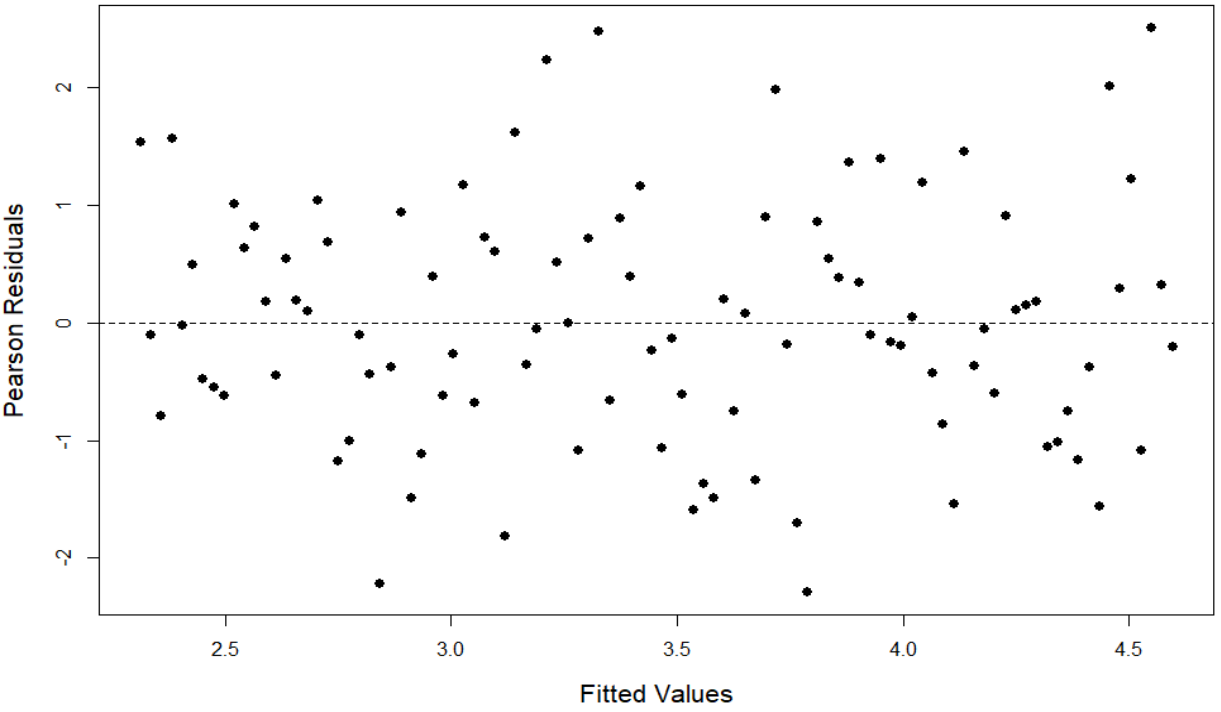


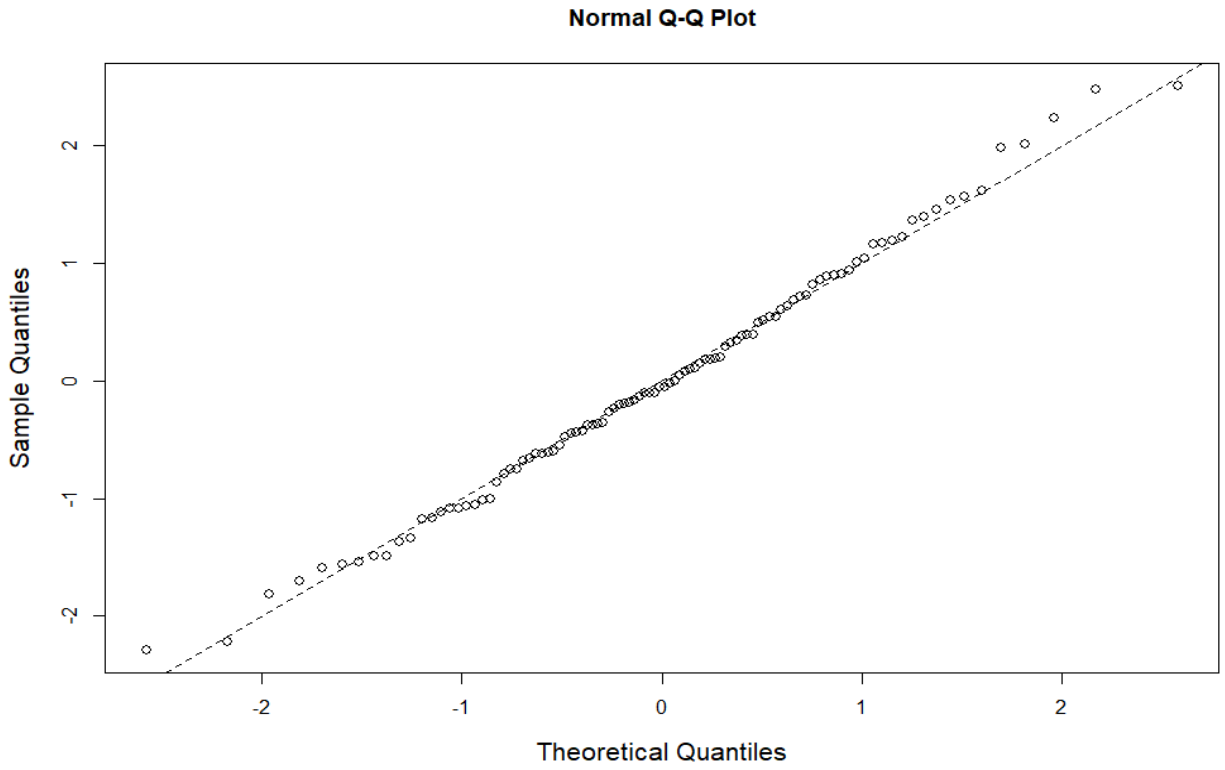
Checking for a Good Fit

- Like response residuals for a LM, Pearson and deviance residuals should:
  - Have an average of zero
  - Be randomly scattered around zero, with no pattern or trend
  - Be normally distributed

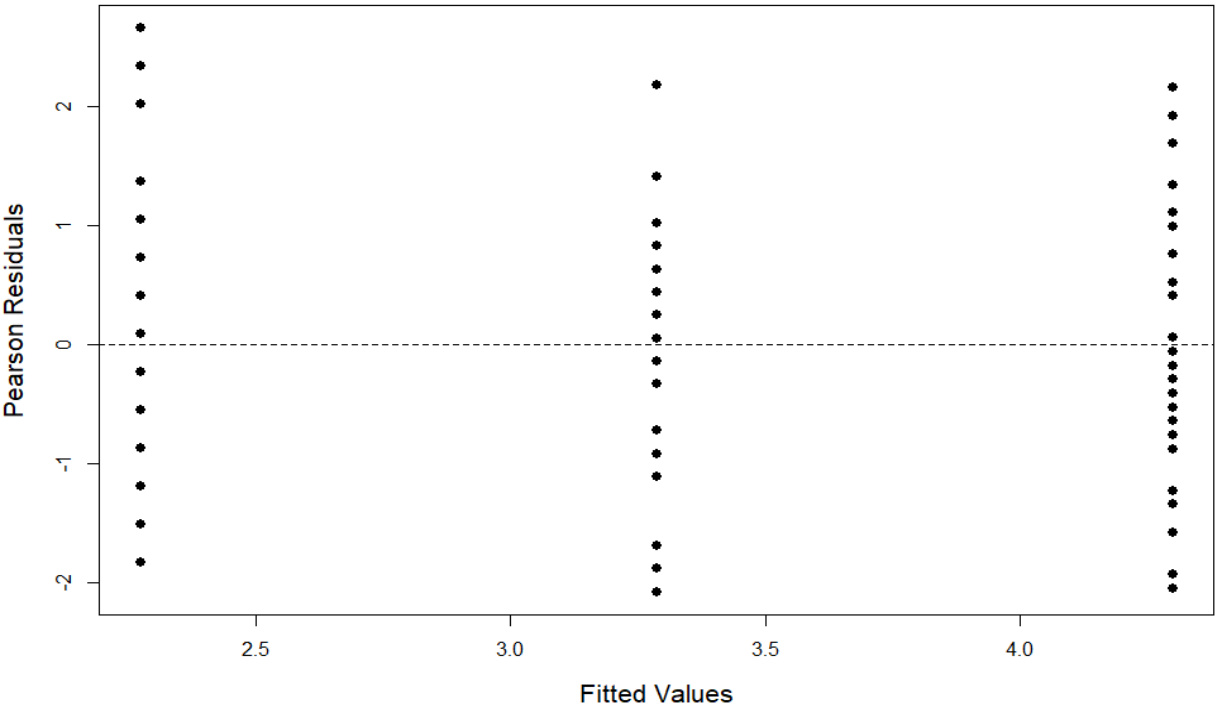
- LM Response residuals can have any variance, but Pearson and deviance residuals should have a variance of 1
  - This means ~95% of them are between -2 and + 2

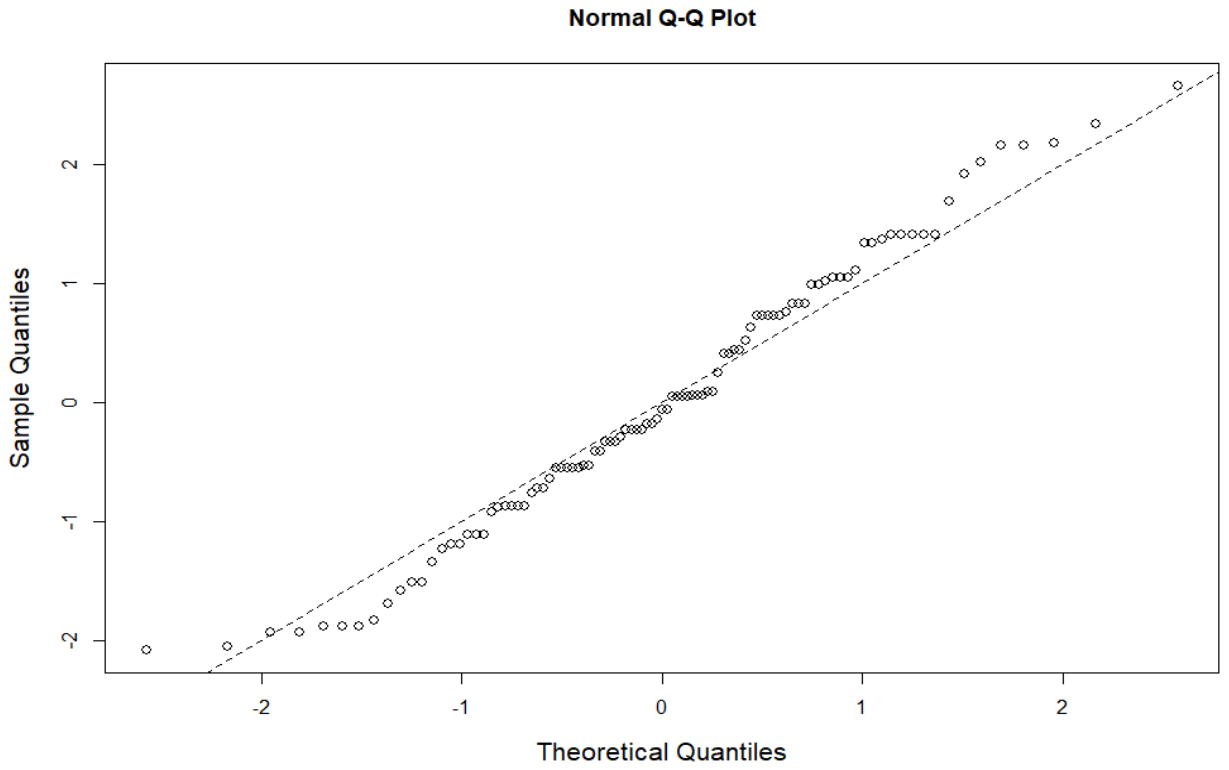
Good Fitting Example I





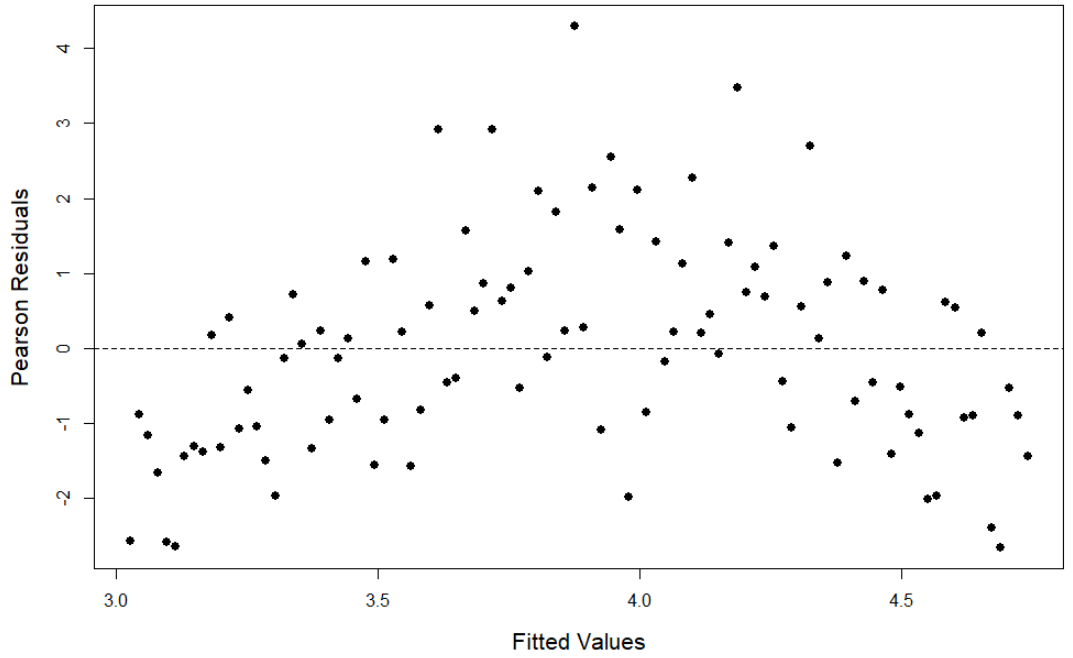
Good Fitting Example II





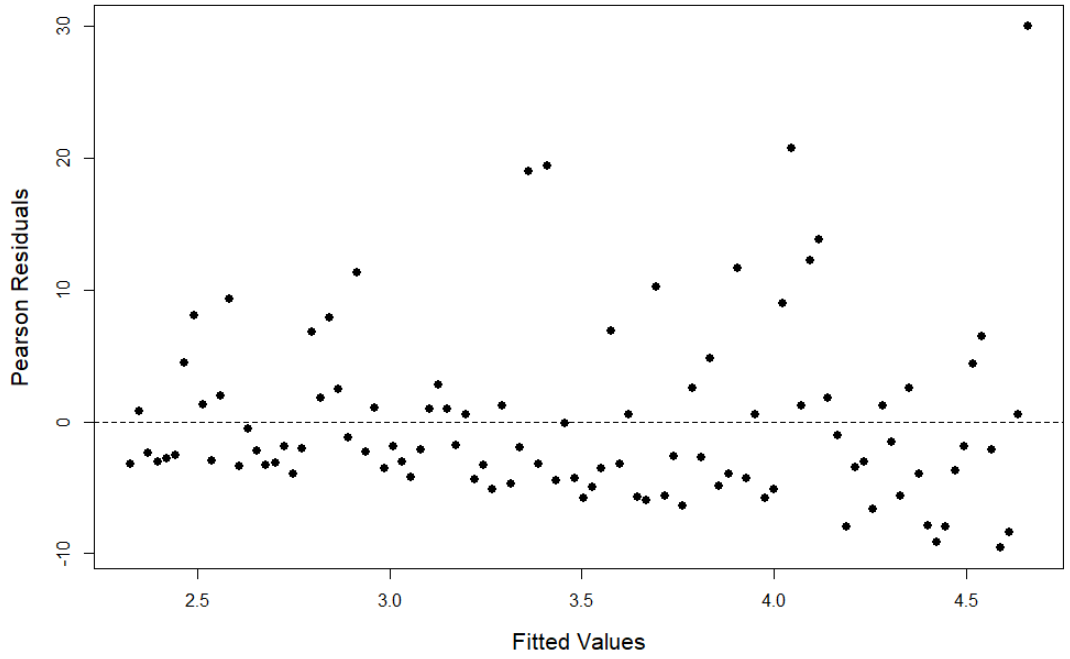
Poor Fit: Pattern in Residuals

- When there is a pattern or trend in the residuals, the assumed linear relationship is probably wrong.
- Try a quadratic model instead, or add additional explanatory variables.



Poor Fit: Residuals too large

- If the residuals have no trend but are too big, the distribution is probably the culprit.
- Try a quasipoisson or negative binomial instead of a Poisson, or a quasibinomial instead of a binomial.



How to do it in R

- The `residuals()` function in `R` calculates all three types of residuals covered
  - `resids <- residuals(mod, type = "response")`
  - `resids <- residuals(mod, type = "pearson")`
  - `resids <- residuals(mod, type = "deviance")`

- A simple way to plot residuals against predicted values is
  - `plot(predict(mod), resids)`

How to do it in R

- The gglm() package plots deviance residuals:
  - `gglm(mod)`

Deviance

Deviance

- Deviance is a way of measuring how well a GLM fits
- A deviance of zero means the model predicts perfectly
- The higher the deviance, the worse the fit
- `deviance()` and `summary()` calculate deviance of a model

```
    Null deviance: 1715.61  on 99  degrees of freedom
Residual deviance:  108.93  on 98  degrees of freedom
AIC: 640.74
```

Chi-Squared Test

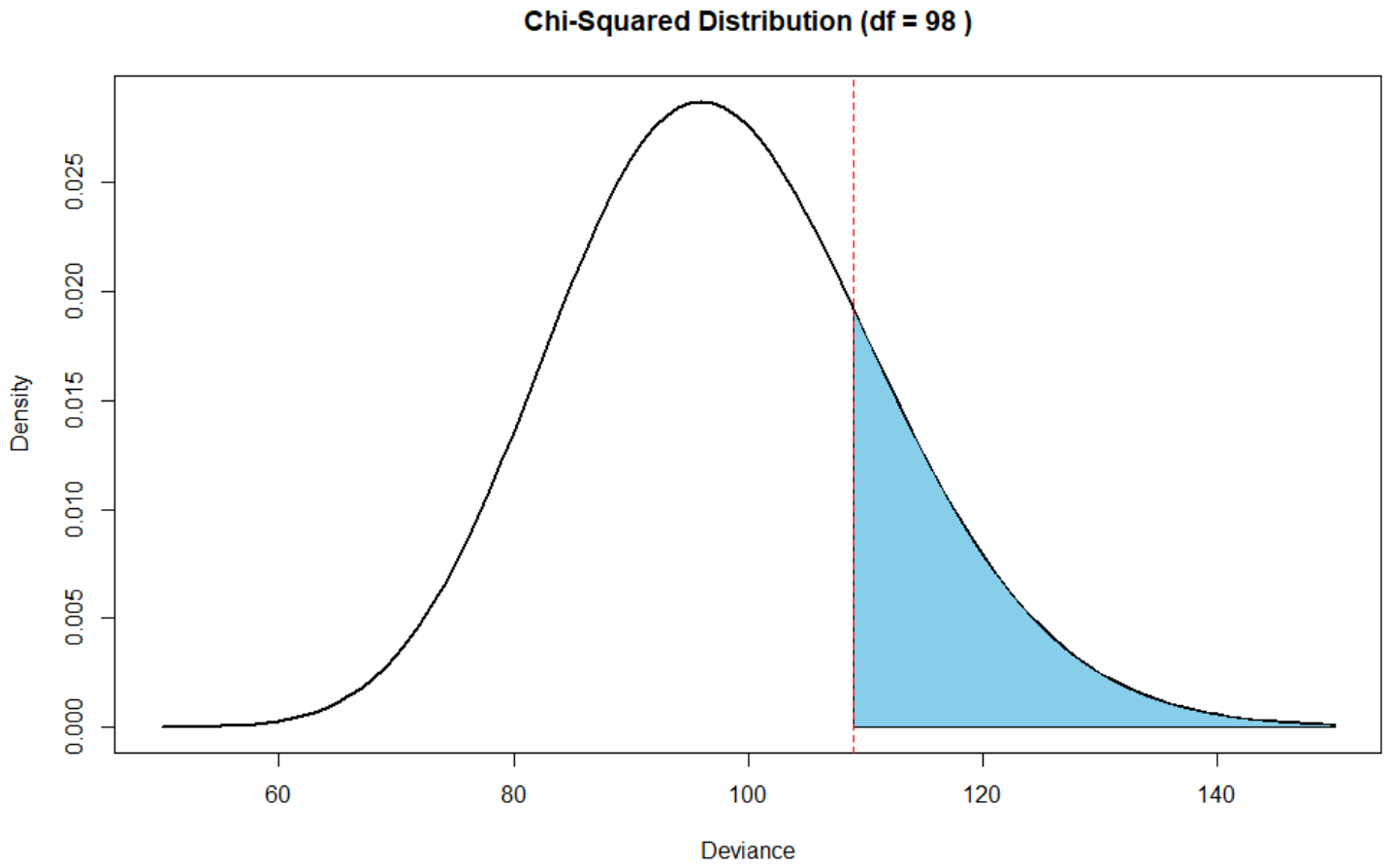
Formally, we can test the null hypothesis that the model is correct by calculating a p-value using

$$p = \Pr(\chi^2_{n-k} > D).$$

- $n - k$ is the degrees of freedom of the model: *n* is the number of data points and *k* the number of parameters

- Higher deviance -> Smaller p-value. If deviance is too high, we reject hypothesis that model is a good fit

Chi-Squared Test



You can get a p-value in `R` for a GLM called `mod`:

```
D = mod$deviance
df = mod$df.residual
1 - pchisq(D, df)
```

Plot generated with assistance from Google AI

Watch Out

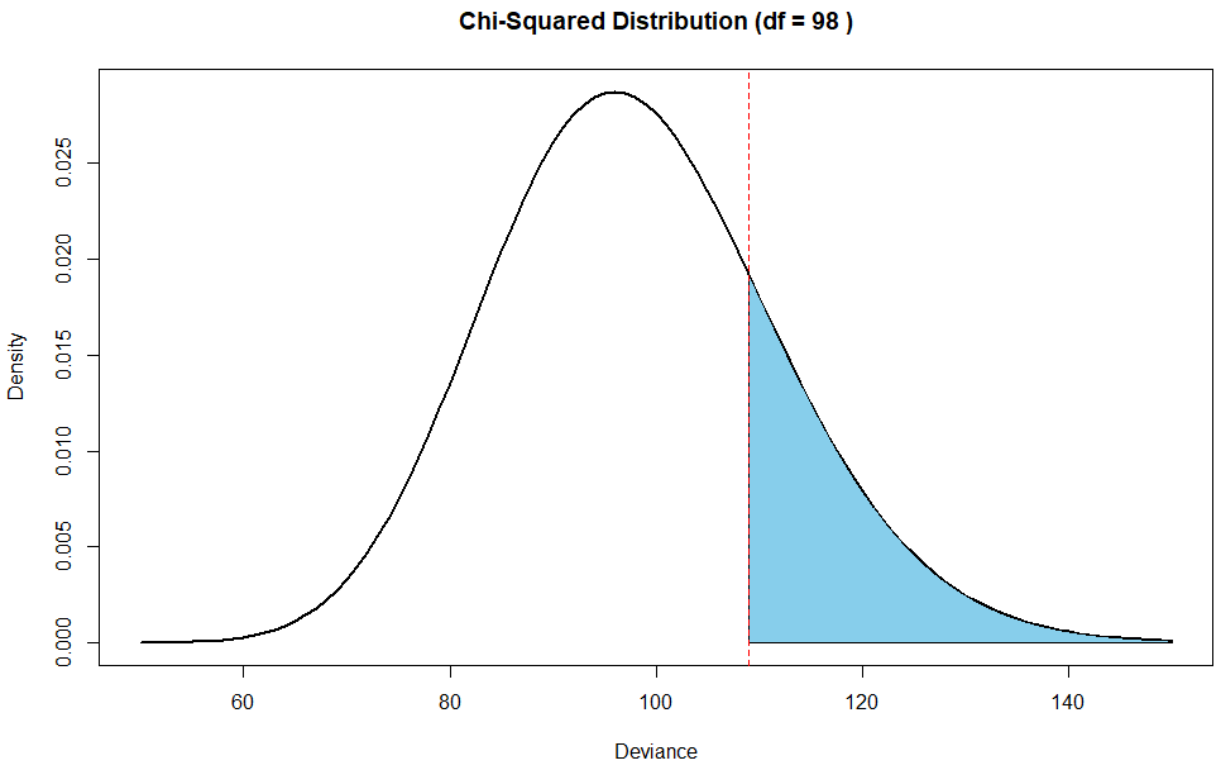
- The concept of deviance doesn't exist for quasipoisson and quasibinomial models

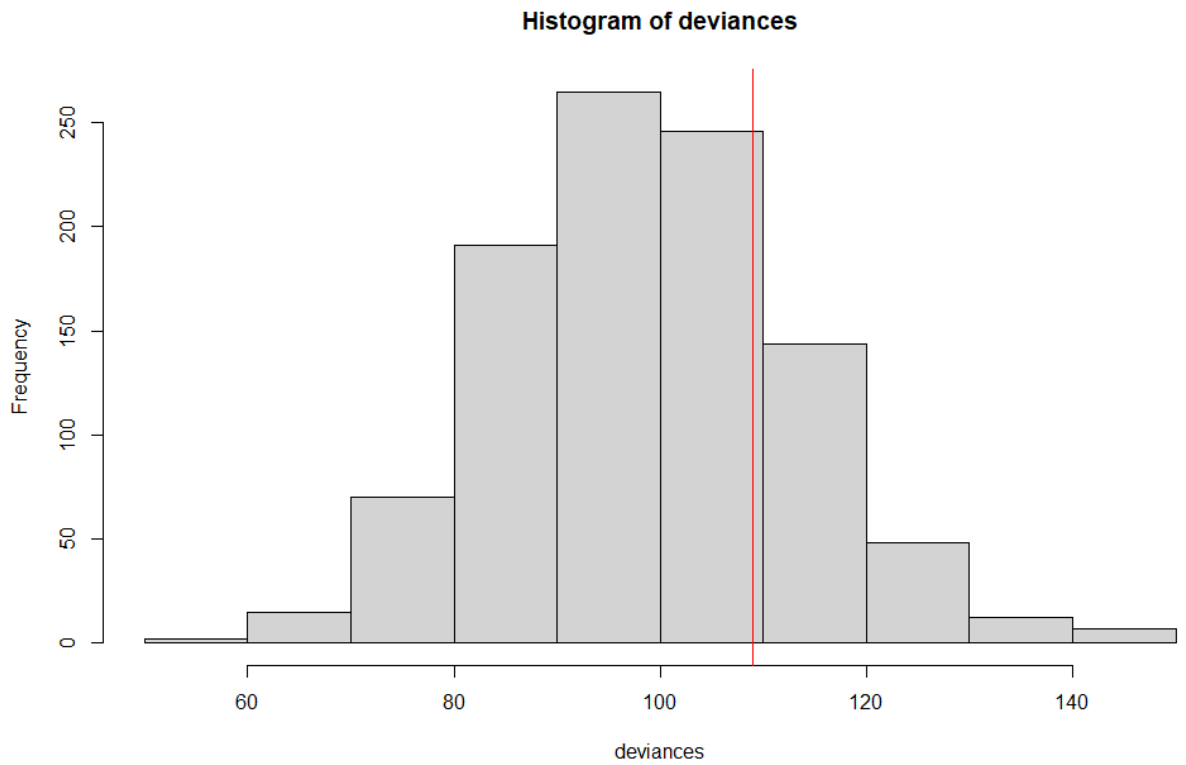
- The Chi-Squared Test doesn't work well for "sparse" counts. As a rule of thumb:
  - For a Poisson GLM, we want $\mu > 5$
  - For a Binomial GLM, we want $n > 5$ at least

Checking Deviance by Simulation

- If we don’t trust the chi-squared test, we can simulate instead.
- We can assume our model is true, simulate data from it, fit our model to the simulated data and calculate a deviance.
- If this is repeated many times, the actual deviance can be compared to the simulated deviances.

Chi-Squared Test versus Simulation





p = 0.212

p = 0.231

No evidence against hypothesis that model fits well; chi-squared approximation is good