

GLMs: Validation

Models behaving badly: Part 2!

Samuel Robinson, Ph.D.

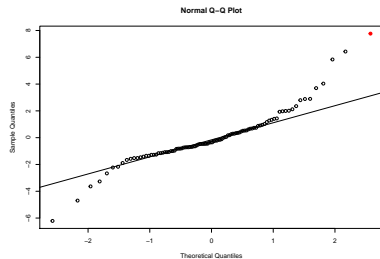
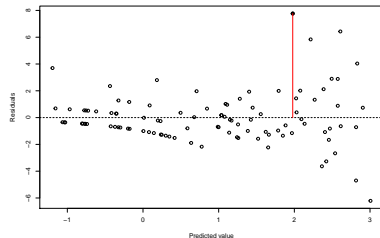
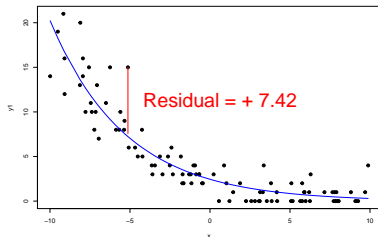
December 3, 2020

Motivation

- Are my model results reliable?
 - Residual checks
 - Overdispersion
 - Zero-inflation
- Model selection - which terms should I use?
 - ML vs REML
 - log-likelihood, χ^2 tests, and AIC
- Other things
 - Binomial GLMs with >1 trial
 - Offsets in count models
 - R^2 for GLMs
- Show-and-tell!

Problem 1: Residual checks

- In LMs, residual checks are used to make sure that:
 - ① Terms are linearly related
 - ② Generating process is valid
 - ③ Variance is constant
- “Regular” residuals don't work this way for GLMs!



There are *many* kinds of residuals!

In addition to *response* (regular) residuals there are:

- Working residuals
- Pearson residuals
- **Deviance residuals**

Deviance residuals use *likelihood*:

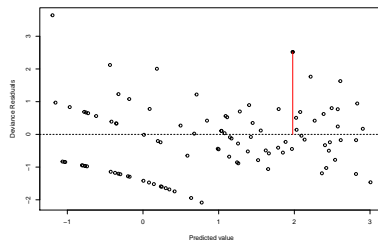
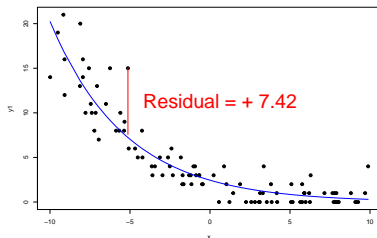
$$r_{dev} = \text{sign}(y - \hat{y}) \sqrt{2(\log(L(y|\theta_s)) - \log(L(y|\theta)))}$$

- This may look scary, but R does this all for you!
- These are analogous to regular residuals in LMs
- For more about the different kinds of residuals, see [here](#)

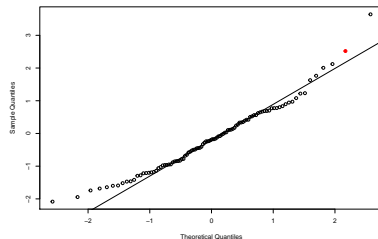
Solution: use deviance residuals for GLMs

Keep in mind:

- Residuals from GLMs will never be as “pretty” as those from LMs
- *Especially* true for:
 - Binomial GLMs
 - Poisson/Negative Binomial GLMs with many zeros

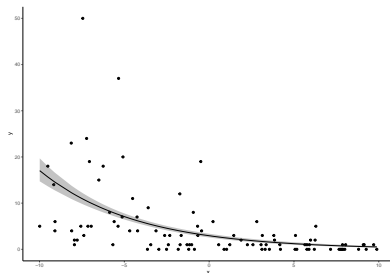


Normal Q-Q Plot



Problem 2: Overdispersion

- Binomial and Poisson families have **no** variance term (e.g. SD).
- Sometimes this assumption doesn't work!
- This is very common for Poisson models



Example: data are much more variable than the predictions from the model

Problem: Overdispersion

```
##
## Call:
## glm(formula = y1 ~ x, family = "poisson", data = d1)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.0843  -0.9460  -0.1897   0.5333   3.6416
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.89455     0.07818   11.44  <2e-16 ***
## x            -0.21145     0.01174  -18.01  <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: 564.27  on 99  degrees of freedom
## Residual deviance: 106.20  on 98  degrees of freedom
## AIC: 362.01
##
## Number of Fisher Scoring iterations: 5
```

- In a regular Poisson or Binomial model, Residual deviance ÷ Degrees of Freedom should be ~ 1
- Residual deviance is the sum of all deviance from the model
- This model looks OK ($106.2 \div 98 = 1.08$)

Problem: Overdispersion

```
##
## Call:
## glm(formula = y2 ~ x, family = "poisson", data = d1)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -4.1009  -1.7543  -0.8805   0.4796   8.6102
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.07897    0.06871   15.70  <2e-16 ***
## x           -0.17581    0.01069  -16.44  <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: 851.96  on 99  degrees of freedom
## Residual deviance: 501.98  on 98  degrees of freedom
## AIC: 735.46
##
## Number of Fisher Scoring iterations: 5
```

- This model does **not** look OK ($501.98 \div 98 = 5.12$)
- Generated using Negative Binomial, but fit to Poisson

Causes

Overdispersion can be caused by different things:

- Using the wrong probability distribution
 - e.g. Poisson, but should be Negative Binomial
- Lots of zeros in count data
 - e.g. Very short observation period
- Leaving out an important term
 - e.g. An important *interaction* term was omitted
- Random effects¹ not accounted for
 - e.g. Data collected at different sites, but ignored

¹Random effects discussed later

Solutions for overdispersion

Try the following (in this order):

- ① Consider terms that may have been left out
 - ① Fixed effects
 - ② Random effects
- ② Try distributions that account for overdispersion
 - ① Negative Binomial, Beta Binomial, Zero-inflated Poisson²
 - ② Quasi-binomial² and quasi-poisson²
 - ③ Transform counts to presence/absence
- ③ Lower your expectations, and use a lower critical p-value (e.g. 0.01 instead of 0.05)
- ④ Design a better study :(

²These can be annoying to deal with, so avoid if possible

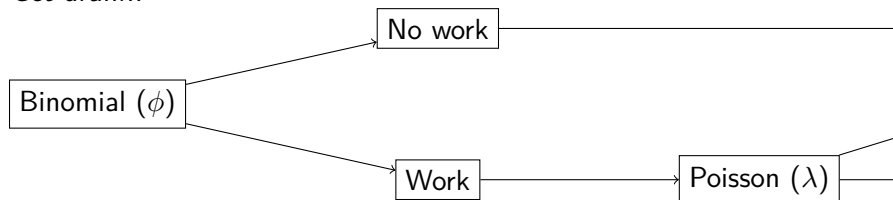
Zero-inflation: drunk monks

An analogy:

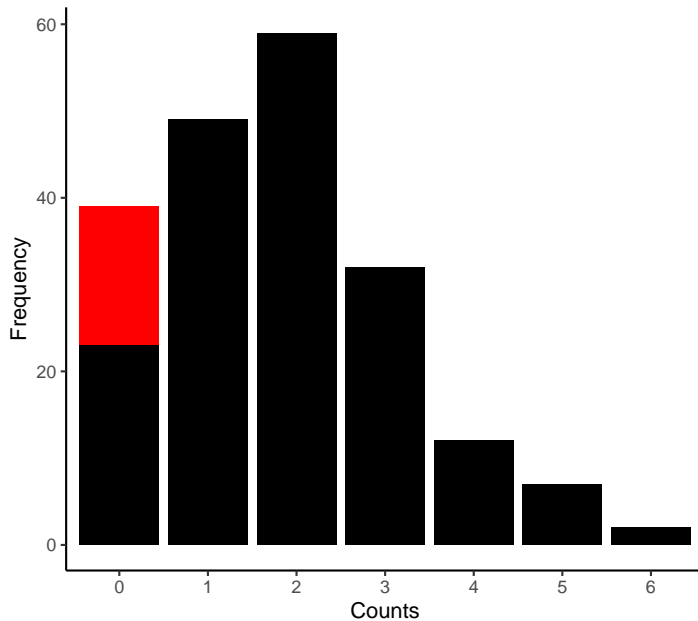
- 1 Monks at a monastery make copies of manuscripts. Most days they make very few (0 or 1), but occasionally they make many (2-5)
- 2 Some days they decide to try out the beer that's been brewing in the cellar! No manuscripts get made on those days.
- 3 The number of manuscripts made (per day) follows a *zero-inflated Poisson distribution*

This is *mixture* of a Poisson and a Binomial:

Get drunk?



Zero-inflation: graphical model



R actually uses *log* Likelihood

- Probabilities multiplied together quickly become *very small*
- Computers can't distinguish between extremely big or small numbers
- Therefore, it uses *log-likelihoods* (also easier to calculate)

