

GLMs: Validation

Models behaving badly: Part 2!

Samuel Robinson, Ph.D.

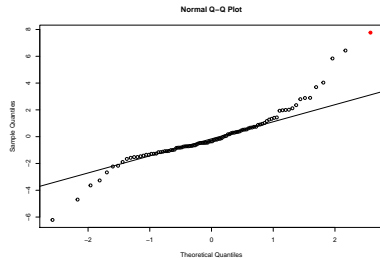
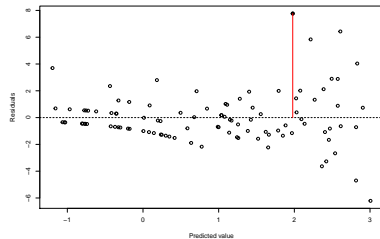
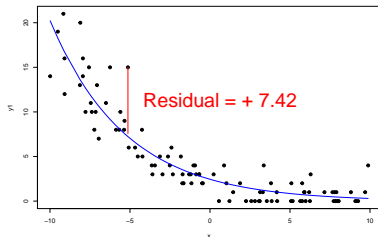
December 3, 2020

Motivation

- Are my model results reliable?
 - Residual checks
 - Overdispersion
 - Zero-inflation
- Model selection - which terms should I use?
 - log-likelihood, χ^2 tests, and AIC
 - ML vs REML
- 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*:

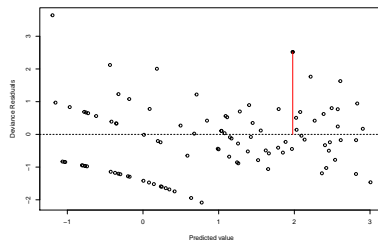
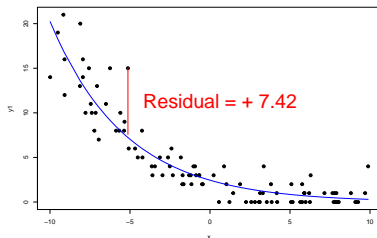
$$r_{dev} = \text{sign}(y - \hat{y}) \sqrt{2(\log(L(y|\theta_s)) - \log(L(y|\theta)))}$$

- 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](#)

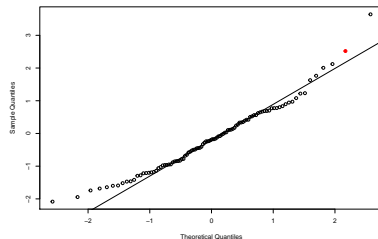
Solution: use deviance residuals for GLMs

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

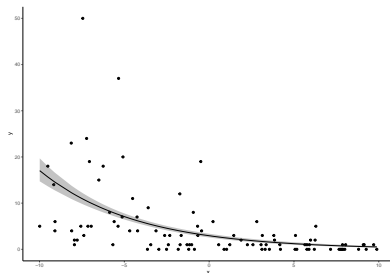


Normal Q-Q Plot



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 :(

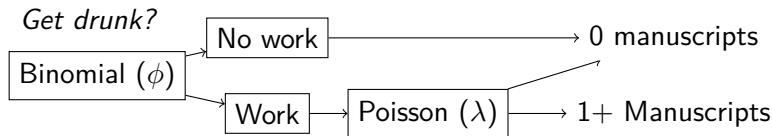
²These can be annoying to deal with, so avoid if possible

Zero-inflation: drunk monks

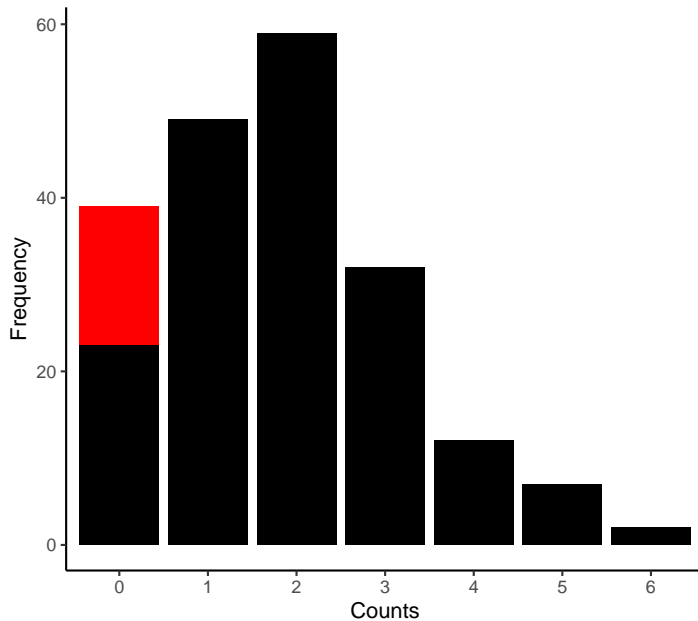
An analogy:

- 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:

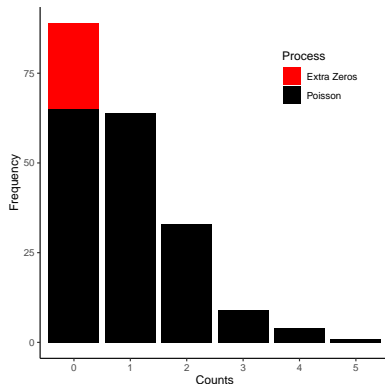
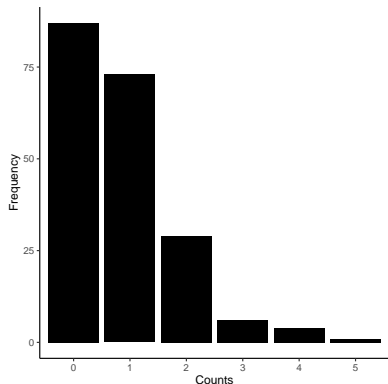


Zero-inflation: graphical model



Problem: hard to fit

- Hard for R to tell the difference between ZIP/ZINB, and a Poisson/NB with a low mean (λ).
- This needs a lot of data in order to work! Consider longer sampling periods in order to reduce zeros



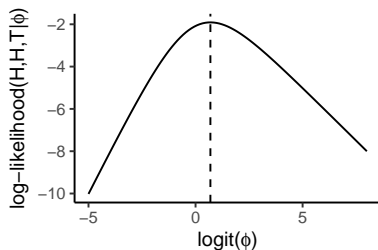
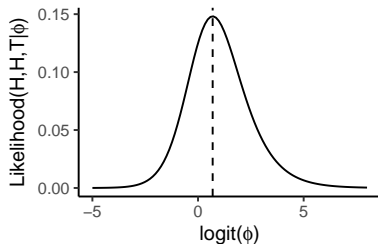
Model selection

How many terms should be in my model?

- Same principle as in regular linear models: what do you think the process is? (see Lecture 4)
- Perhaps some terms should *always* be included on principal?
- To test whether terms are important in predicting your data, use likelihood-ratio tests
 - AIC tests usually say the same thing as LR tests
 - `drop1(model, test='Chisq')`

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)



ML vs REML

- Maximum likelihood (ML) estimates of variance (e.g. SD) are always smaller than the actual variance (biased)
- Restricted maximum likelihood (REML) uses a mathematical trick to get around this, but. . .
- This means that models with different numbers of terms don't have the same REML estimates
- Likelihood between these models technically can't be compared!

Solution = Use ML if comparing between models with different fixed effects, then re-fit with REML once you've decided on a model

Other things!

- Binomial GLMs with >1 trial
- Offsets in count models
- R^2 for GLMs

Binomial GLMs with >1 trial

- If you're measuring single "success/failures", 1s and 0s are used
- If multiple trials occur, R requires counts of successes and failures
- Example: "I counted male and female critters at different sites. How do I tell if temperature affects sex ratios?"

```
#Number of females and males are in 2 separate columns in d1  
glm(cbind(females,males) ~ temperature, family='binomial',data = d1)
```

This will correctly account for different numbers of critters ("trials") at each site

Offsets in count models

- Poisson/NB models assume that counts occur over the same period of time
- Count models use integers only, so you can't just do:
 $\text{counts} \div \text{hours}$
- Solution: use *offsets* to deal with different observation times
 - Predictor with a slope fixed at 1
- Example: “I counted critters for different lengths of time at each site. How do I tell if temperature affect counts?”

```
#hours = observation time at each site, and  
# must be log-transformed before being used in an offset  
#  
glm(counts ~ offset(log(hours)) + temp, family='poisson', data = d1)
```

This will return estimates that have been scaled to a 1-hour observation time

R-squared for GLMs

- Bad news: there isn't really any good way to get R^2 (explained variance) for non-lm models
- OK news: there are many *pseudo- R^2* measures that are *sort of* like R^2 , but nobody really agrees on which one is best
- Good news: ecologists tend to not know or care about this

Solution: pick a single type of R^2 and use that, or omit it completely ³

- See [here](#), [here](#) or [here](#) for more info on R^2
- Try `rsquaredglmm()` from `piecewiseSEM` or `r.squaredGLMM()` from `MuMIn`

³But be prepared to argue with supervisors, committee members, or reviewers! They will want some kind of measure of how well your model predicted your data.

Show-and-tell!

