intermediate generalized linear models

Ben Bolker

packages

```
library(ggplot2)
theme_set(theme_bw())
library(aods3)
```

```
## Loading required package: lme4

## Loading required package: Matrix

## Loading required package: boot
```

overdispersion

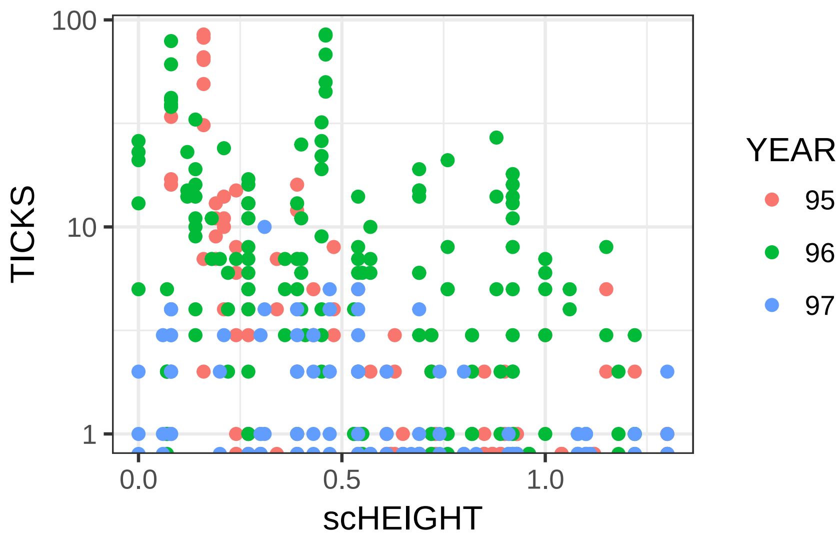
overdispersion

- more variance than expected based on statistical model
- e.g. variance > mean for Poisson
- in general leads to *overconfidence*
  - overly narrow confidence intervals
  - too-small p-values
  - inflated type I error

Tick example

```
ticks <- read.table("../data/Elston2001_tickdata.txt",
    header = TRUE)
ticks <- transform(ticks, YEAR = factor(YEAR),
    scHEIGHT = (HEIGHT - min(HEIGHT))/100)
ggplot(ticks, aes(scHEIGHT, TICKS, colour = YEAR)) +
    geom_point() + scale_y_log10()
```

```
## Warning: Transformation introduced infinite
## values in continuous y-axis
```




```
ticks_glm1 <- glm(TICKS ~ scHEIGHT * YEAR, ticks,
    family = poisson)
aods3::gof(ticks_glm1)
```

```
## D  = 3008.964, df = 397, P(>D) = 0
## X2 = 4496.887, df = 397, P(>X2) = 0
```

methods

- quasi-likelihood models
- compounded distributions
- observation-level random effects

quasi-likelihood

- quantify excess variance
- e.g. ϕ=`sum(residuals(m,type="pearson")^2)/df.residual(m)`
- multiply estimated standard errors by $\sqrt{phi}$
- recompute Z/t statistics, p values
- `family=quasipoisson` or `family=quasibinomial` does this automatically
- no likelihood/AIC available

ticks

```
ticks_QP <- update(ticks_glm1, family = quasipoisson)
summary(ticks_QP)
```

```
##
## Call:
## glm(formula = TICKS ~ scHEIGHT * YEAR, family = quasipoisson,
```

```
##     data = ticks)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.0993  -1.7956  -0.8414   0.6453  14.1356
##
## Coefficients:
##                 Estimate Std. Error t value
## (Intercept)       4.0008     0.2391  16.731
## scHEIGHT         -5.8198     0.8547  -6.809
## YEAR96           -0.9831     0.2729  -3.603
## YEAR97           -2.9448     0.5057  -5.824
## scHEIGHT:YEAR96   4.4693     0.8959   4.988
## scHEIGHT:YEAR97   4.0453     1.2081   3.349
##                 Pr(>|t|)
## (Intercept)      < 2e-16 ***
## scHEIGHT        3.64e-11 ***
## YEAR96          0.000355 ***
## YEAR97          1.19e-08 ***
## scHEIGHT:YEAR96 9.12e-07 ***
## scHEIGHT:YEAR97 0.000890 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 11.3272)
##
##     Null deviance: 5847.5  on 402  degrees of freedom
## Residual deviance: 3009.0  on 397  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

compounded distributions

- instead of Poisson/binomial/etc., use a compounded distribution
- Gamma + Poisson = negative binomial (e.g. `MASS::glmer.nb`)
- Beta + binomial = beta-binomial (e.g. `glmmTMB`, `bbmle::mle2`)

```
ticks_NB <- MASS::glm.nb(TICKS ~ scHEIGHT * YEAR,
    data = ticks)
summary(ticks_NB)
```

```
##
## Call:
```

```
## MASS::glm.nb(formula = TICKS ~ scHEIGHT * YEAR, data = ticks,
##     init.theta = 0.9000852793, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3765  -1.0281  -0.5052   0.2408   3.2440
##
## Coefficients:
##                 Estimate Std. Error z value
## (Intercept)       3.3829     0.2323  14.559
## scHEIGHT         -4.1308     0.4033 -10.242
## YEAR96           -0.2890     0.2829  -1.022
## YEAR97           -2.1926     0.3286  -6.672
## scHEIGHT:YEAR96   2.6132     0.4824   5.418
## scHEIGHT:YEAR97   2.0861     0.5571   3.745
##                 Pr(>|z|)
## (Intercept)      < 2e-16 ***
## scHEIGHT         < 2e-16 ***
## YEAR96          0.307009
## YEAR97          2.52e-11 ***
## scHEIGHT:YEAR96 6.04e-08 ***
## scHEIGHT:YEAR97 0.000181 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.9001) family taken to be 1)
##
##     Null deviance: 840.71  on 402  degrees of freedom
## Residual deviance: 418.82  on 397  degrees of freedom
## AIC: 1912.6
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.9001
##           Std. Err.:  0.0867
##
##  2 x log-likelihood:  -1898.5880
```

observation-level random effects

- use mixed models; add a Normal deviate to each observation (on the link-function/linear predictor scale)

- e.g. logit-Normal-binomial, or log-Normal-Poisson

```
ticks <- transform(ticks, obs = 1:nrow(ticks))
ticks_OR <- glmer(TICKS ~ scHEIGHT * YEAR + (1 |
    obs), data = ticks, family = poisson)
summary(ticks_OR)
```

```
## Generalized linear mixed model fit by
##   maximum likelihood (Laplace  Approximation)
##  [glmerMod]
##  Family: poisson  ( log )
## Formula: TICKS ~ scHEIGHT * YEAR + (1 | obs)
##    Data: ticks
##
##      AIC      BIC   logLik deviance df.resid
##   1903.0   1931.0   -944.5   1889.0      396
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.29773 -0.50197 -0.06591  0.22414  1.91379
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  obs    (Intercept) 1.132    1.064
## Number of obs: 403, groups:  obs, 403
##
## Fixed effects:
##                 Estimate Std. Error z value
## (Intercept)       2.7402     0.2429  11.284
## scHEIGHT         -4.0492     0.4154  -9.746
## YEAR96           -0.2069     0.2958  -0.699
## YEAR97           -1.9407     0.3482  -5.573
## scHEIGHT:YEAR96   2.5381     0.5026   5.050
## scHEIGHT:YEAR97   1.8683     0.5888   3.173
##                 Pr(>|z|)
## (Intercept)      < 2e-16 ***
## scHEIGHT         < 2e-16 ***
## YEAR96           0.48433
## YEAR97          2.50e-08 ***
## scHEIGHT:YEAR96 4.41e-07 ***
## scHEIGHT:YEAR97  0.00151 **
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
```

```
## Correlation of Fixed Effects:
##                (Intr) scHEIGHT YEAR96 YEAR97
## scHEIGHT       -0.830
## YEAR96         -0.818  0.682
## YEAR97         -0.693  0.580    0.568
## sHEIGHT:YEAR96  0.689 -0.826   -0.835 -0.480
## sHEIGHT:YEAR97  0.592 -0.704   -0.485 -0.834
##                sHEIGHT:YEAR96
## scHEIGHT
## YEAR96
## YEAR97
## sHEIGHT:YEAR96
## sHEIGHT:YEAR97  0.583
```

offsets

- account

complete separation

- what happens when a logistic regression model is too good?
- some threshold: all below=0, all above=1
- best slope estimate on logit scale is *infinite*
- Wald approximation breaks down (*Hauck-Donner effect*)
- symptoms: $|\beta| > 10$, crazy SEs and terrible p-values
- strong effects, or slicing data too thin

solutions

- model comparison (`anova()`) still works
- profile CI should get *lower* limit of parameters
- penalization (`brglm`, "Firth's method")
- Bayesian approaches: put a prior on parameters (`blme`, `brms`)

zero-inflation

zero-inflation

- *too many* zeros
- "lots of zeros" can occur just because of low mean
- mode at zero *and* away from zero usually does mean Z-I

zero-inflation models

- *zero-inflation*: mixture of structural and sampling zeroes
  (**not** "true" and "false")

- *hurdle*: zeros plus truncated distribution
- choice depends on meaning of zeros
- Z-I as well as conditional mean may be modeled

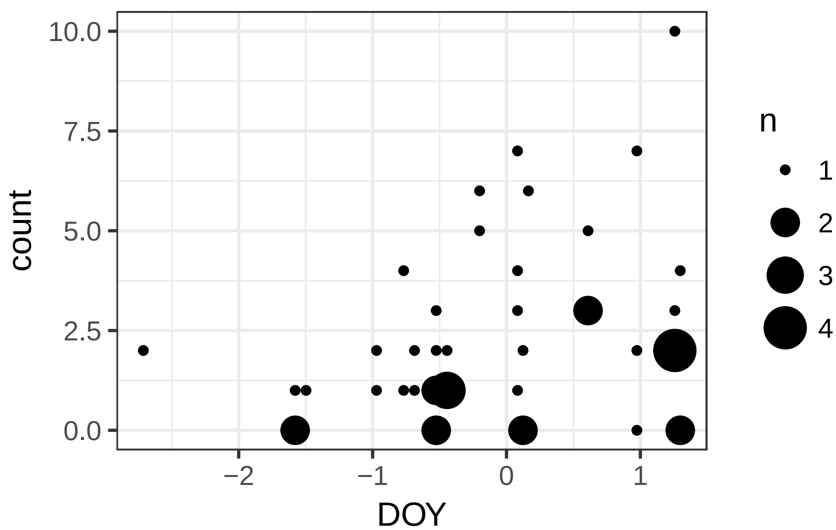
testing for zero-inflation

- a little tricky
- easiest (?) to fit Z-I model and then test whether you needed it or not
- *posterior predictive simulation*

posterior simulation

Use the `simulate()` method, if available

```
data(Salamanders, package = "glmmTMB")
ss <- subset(Salamanders, spp == "GP" & mined ==
    "no")
## fit model
ggplot(ss, aes(DOY, count)) + stat_sum()
```

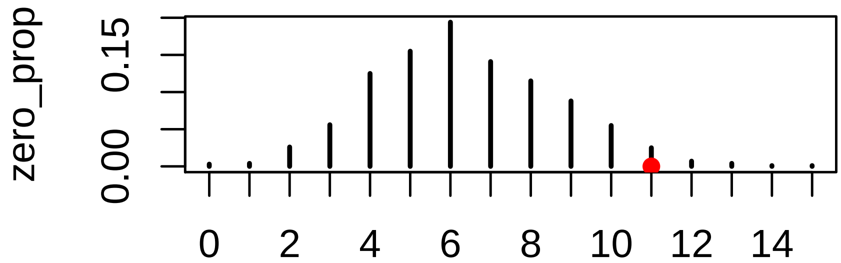



```
salam_1 <- glm(count ~ DOY, ss, family = poisson)
## simulate 1000 realizations from the model
sims <- simulate(salam_1, 1000)
## count proportions of zeros per simulation
zero_prop <- prop.table(table(colSums(sims ==
    0)))
zero_ind <- as.numeric(names(zero_prop))
obs_zeros <- sum(ss$count == 0)
## p-value
sum(zero_prop[zero_ind >= obs_zeros])
```

```
## [1] 0.038
```

zero-inflation plot

```
plot(zero_prop)
points(obs_zeros, 0, col = "red", pch = 16)
```



alternative families and links

Gamma

complementary log-log

beyond the exponential family

beta regression

- GLMs require counts (denominators), e.g. 40% = 4/10
- what if data don't have obvious denominators
- e.g. cover scores, activity budgets
- *Beta distribution*

negative binomial regression