From logistic to binomial & Poisson models

Ben Bolker

October 16, 2018

Version: 2018-10-16 23:20:09

Logistic regression is special in some ways:

- conditional distribution (Bernoulli) is always correct
- model diagnostics especially hard
- no possibility of *overdispersion*

(Aggregated) binomial regression

Binomial with $N > 1$. Basically the same procedures as logistic regression, *except*:

- easier to do exploration, diagnostics (data are already aggregated)
- need to specify response *either* as a two-column matrix: `cbind(num_successes,num_failures)` *or* as a proportion with the additional `weights` variable giving the total number of trials
- need to check for **overdispersion** (see below)

Set up an example to use:

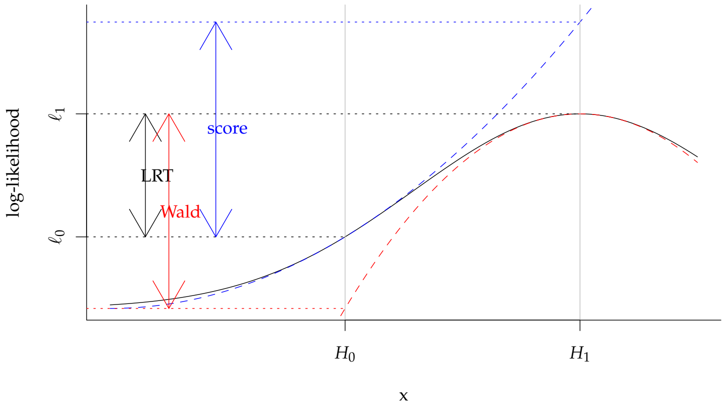
```r
lizards <- read.csv("../data/lizards.csv")
## gfrac (= fraction grahami), N (=grahami+opalinus) already defined
lizards <- transform(lizards,
                     time=factor(time,levels=c("early","midday","late")))
g1 <- glm(gfrac~height+diameter+light+time,
          lizards,family=binomial,weight=N)
## or
g2 <- glm(cbind(grahami,opalinus) ~ height+diameter+light+time,
          lizards, family=binomial)
all.equal(coef(g1),coef(g2))

## [1] TRUE
```

Model diagnostics

Graphical plot computed diagnostic summaries and/or transformations of residuals to highlight particular classes of model deviations

Formal
- compute an overall goodness-of-fit statistic with a known null distribution
- embed the model in a larger parametric family; compare via likelihood ratio test (consider exact or "round" alternative). May use *score test* or single-step update for computational efficiency.



(Fears et al., 1996; Pawitan, 2000)

Residuals

Different types of residuals (`?residuals.glm`, `?rstandard`, `?rstudent`)

Raw $y - \mu$

Deviance $\text{sign}(y-\mu)\sqrt{2w_i\text{dev}_i}$

Pearson $(y-\mu)/(w\sqrt{V(\mu)})$

Standardized $(y-\mu)/(\sqrt{V(\mu)(1-H)}$

Note whether residuals are scaled by (1) variance function, (2) weights, (3) full variance (i.e. including overdispersion factor ϕ), (4) diagonal of *hat matrix* (`hatvalues()`).

(Hat matrix: weighted version of $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$: maps $\mathbf{y}$ to $\hat{\mathbf{y}}$, so h_{ii} is the influence of y_i on $\hat{y}_i$. All hat values are identical for linear models with categorical variables, but not for regression models/GLMs . . .)

Linearity

- (Deviance) residual vs. fitted plot
- (Deviance) residuals vs. individual predictors, or combinations of predictors
- link test [1]; try adding a quadratic term in the linear predictor, see if it fits better
- Adjust by
  - changing link function: `power()`)
  - adding polynomial or spline terms to individual predictors (`poly()`, `splines::ns()`)
  - transforming individual predictors

Variance function

- Scale-location plot: $\sqrt{\mathrm{abs}(\mathrm{residuals})}$ vs. fitted value, or individual parameters, or combinations of parameters. If residuals are scaled and there is no overdispersion (see below) then the center is at 1
- (Banta example?)
- Adjust by
  - fixing some other part of the model
  - tweaking the variance function

Distributional assumptions

The variance function and link function might both be right, but the model distribution can still be wrong (e.g. log-Normal vs Gamma, zero-inflation).

- assessing distributional assumption is hard because it's the *conditional* distribution
- Q-Q plot (examples): good, but only really valid asymptotically (i.e. conditional distribution of *individual samples* $\approx$ Normal: e.g. $\lambda > 5$ for Poisson, $n\min(p, 1-p) > 5$ for Binomial)
- alternatives to Q-Q plot, e.g. (Hoaglin, 1980) (not really practical)
- Improved Q-Q plot: `mgcv::qq.gam()` [2], `DHARMa::simulateResiduals()` [3]
- Adjust by

[1] Pregibon, D. (1980, January). Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* *29*(1), 15–14

[2] Augustin, N. H., E.-A. Sauleau, and S. N. Wood (2012, August). On quantile quantile plots for generalized linear models. *Computational Statistics & Data Analysis* *56*(8), 2404–2409

[3] Hartig, F. (2018). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.2.0

  - alternative distribution (log-Normal/Gamma)
  - ordinal models
  - robust models (`robustbase::glmrob`)

Load data on contagious bovine pleuropneumonia (Lesnoff et al., 2004), taken from `lme4` package:

```
load("cbpp.RData")

## Warning in readChar(con, 5L, useBytes = TRUE): cannot
open compressed file 'cbpp.RData', probable reason 'No such
file or directory'
## Error in readChar(con, 5L, useBytes = TRUE): cannot open
the connection

ggplot(cbpp,aes(period,incidence/size))+
    geom_point(aes(size=size),alpha=0.5)+
    geom_line(aes(group=herd))

## Error in ggplot(cbpp, aes(period, incidence/size)):  object
'cbpp' not found
```

Fit with a Poisson-offset model:

```
m1 <- glm(incidence~herd+offset(log(size)),data=cbpp,
          family=poisson)

## Error in is.data.frame(data):  object 'cbpp' not found
```

```
library(mgcv)
par(mfrow=c(1,2),las=1,bty="l")
plot(m1,which=2)  ## Q-Q plot from base R

## Error in plot(m1, which = 2):  object 'm1' not found

qq.gam(m1,pch=1)  ## improved Q-Q from mgcv

## Error in qq.gam(m1, pch = 1):  object 'm1' not found
```

Influential points

`?influence.measures`

- Cook's distance (overall influence)
- leverage

- Adjust by
  - leaving out influential points to see if it makes a difference
  - robust modeling (`robustbase::glmrob`)

Posterior predictive summaries

Simulate 1000 times; count the number of zeros in each simulation; compute (1-sided) p-value.

```
ss <- simulate(m1,1000,seed=101)

## Error in simulate(m1, 1000, seed = 101):  object 'm1' not
found

zerovec <- colSums(ss==0)

## Error in is.data.frame(x):  object 'ss' not found

zero.obs <- sum(cbpp$incidence==0)

## Error in eval(expr, envir, enclos):  object 'cbpp' not
found

(cbpp.zpval <- mean(zerovec>=zero.obs))

## Error in mean(zerovec >= zero.obs):  object 'zerovec' not
found
```

```
## Error in table(zerovec):  object 'zerovec' not found
## Error in plot.xy(xy.coords(x, y), type = type, ...):
plot.new has not been called yet
## Error in paste0("$\\text{Prob}(z \\geq 22)=", signif(cbpp.zpval,
2), "$"):  object 'cbpp.zpval' not found
```

(this is a 1-sided test)

Overall goodness-of-fit/overdispersion

(`aods3` package)

Detection

- Variance > expected (e.g. assume variance = mean but variance > mean)
- Test: $\sum(\text{Pearson residuals})^2 \approx$ residual df

- More specifically, $\sum r^2 \sim \chi^2_{n-p}$
- `pchisq(sum(residuals(.,type="pearson")^2),rdf,lower.tail=FALSE)`, or `aods3::gof(.)`

Meaning

- May be caused by general lack of fit ...
- *or* may be "intrinsic"

Solutions

- quasi-likelihood $\phi \equiv \sum r^2/(n-p)$: scales all likelihoods by ϕ, all CI by $\sqrt{\phi}$
- compound/conjugate model
  - negative binomial (Gamma-Poisson) (via `MASS::glm.nb`)
  - Beta-Binomial (via `bbmle`?)
- link-Normal model: GLMM with observation-level random effect (Gaussian on linear predictor scale)

References

Augustin, N. H., E.-A. Sauleau, and S. N. Wood (2012, August). On quantile quantile plots for generalized linear models. *Computational Statistics & Data Analysis 56*(8), 2404–2409.

Fears, T. R., J. Benichou, and M. H. Gail (1996, August). A reminder of the fallibility of the Wald statistic. *The American Statistician 50*(3), 226–227.

Hartig, F. (2018). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.2.0.

Hoaglin, D. C. (1980). A Poissonness plot. *The American Statistician 34*(3), 146–149.

Lesnoff, M., G. Laval, P. Bonnet, S. Abdicho, A. Workalemahu, D. Kifle, A. Peyraud, R. Lancelot, and F. Thiaucourt (2004, June). Within-herd spread of contagious bovine pleuropneumonia in Ethiopian highlands. *Preventive Veterinary Medicine 64*(1), 27–40.

Pawitan, Y. (2000, February). A reminder of the fallibility of the Wald statistic: Likelihood explanation. *The American Statistician 54*(1), 54–56.

Pregibon, D. (1980, January). Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 29*(1), 15–14.