From logistic to binomial & Poisson models

Ben Bolker

October 17, 2018

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* (*recommended*) as a proportion with the additional `weights` variable giving the total number of trials.
- need to check for **overdispersion** (see below)

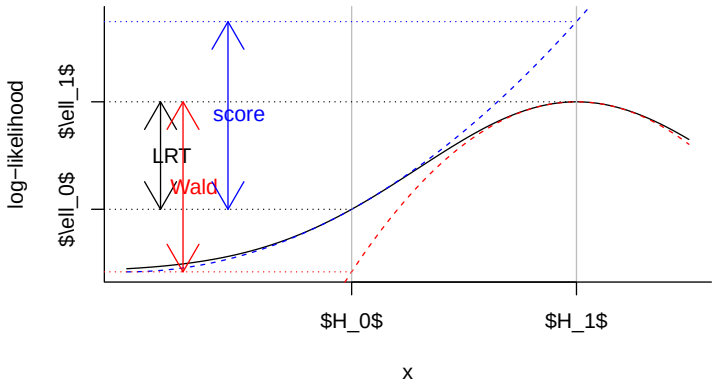
Set up an example to use:

```
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(cbind(grahami,opalinus) ~ height+diameter+light+time,
          lizards, family=binomial)
g2 <- update(g1,  gfrac ~ ., weight=N)
## or

## check answers are the same
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 $\text{sign}(y-\mu)\sqrt{w\text{deviance}}$

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

[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

Variance function

- Scale-location plot: $\sqrt{\text{abs}(\text{residuals})}$ 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
- Adjust by
  - fixing some other part of the model
  - change 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
  - alternative distribution (log-Normal/Gamma)
  - ordinal models
  - robust models (`robustbase::glmrob`)

[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

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`)

Contraception example #2

Contraceptive use data showing the distribution of 1607 currently married and fecund women interviewed in the Fiji Fertility Survey, according to age, education, desire for more children and current use of contraception: downloaded from [http://data.princeton.edu/wws509/datasets/cuse.dat](http://data.princeton.edu/w

```
cuse <- read.table("../data/cuse.dat",header=TRUE)
```

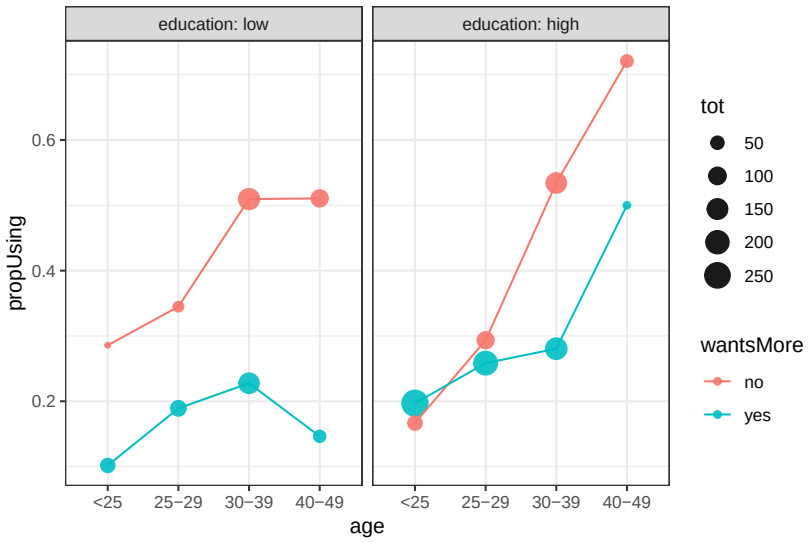
Add convenience variables (proportion and total in each group): change the `education` factor so that "low" rather than "high" is the baseline group:

```
cuse <- transform(cuse,
                  propUsing=using/(using+notUsing),
                  tot=using+notUsing,
                  education=relevel(education,"low"))
```

`ggplot` tricks:

- use `label_both` in the `facet_grid` specification to get the subplots labelled by their factor name as well as the level name
- use `aes(x=as.numeric(age))` to convince ggplot to connect the factor levels on the x axis with lines; use `size=0.5` to make the lines a little skinnier

```
(gg1 <- ggplot(cuse,aes(x=age,y=propUsing,size=tot,colour=wantsMore))+
  facet_grid(.~education,labeller=label_both)+
  geom_point(alpha=0.9)+
  geom_line(aes(x=as.numeric(age)),size=0.5))
```



We could fit the three-way interaction, but it would be a bit silly because there would be as many parameters as observations (this is called a *saturated model*. It would probably be more sensible to include only two-way interactions:

```
fit2 <- glm(propUsing~(age+education+wantsMore)^2,
            weights=tot,
            family=binomial,
            data=cuse)
```

```
plot(fit2)

## Warning in sqrt(crit * p * (1 - hh)/hh):  NaNs produced
## Warning in sqrt(crit * p * (1 - hh)/hh):  NaNs produced
```

```
library(broom)
cuse2 <- augment(fit2,data=cuse)
ggplot(cuse2,aes(.fitted,.resid))+
    geom_point(aes(size=tot))+
    geom_smooth(aes(weight=tot))  ## weight variable

## `geom_smooth()` using method = 'loess' and formula 'y ~
x'

ggplot(cuse2,aes(.fitted,sqrt(abs(.resid))))+
    geom_point(aes(size=tot))+
    geom_smooth(aes(weight=tot))

## `geom_smooth()` using method = 'loess' and formula 'y ~
x'
```

```
p1 <- DHARMa::simulateResiduals(fit2,plot=TRUE)
```

Overdispersion

Detection

- Variance > expected (e.g. assume variance = mean but variance > mean)
- Test: $\sum$(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 poor model ...
- *or* may be "intrinsic"
- **don't worry about overdispersion until other modeling issues are dealt with**
- overdispersion > 2 probably means there is a larger problem with the data: check (again) for outliers, obvious lack of fit
- **only** relevant for families with fixed variance (binomial, Poisson), and **not** for Bernoulli responses

Solutions

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

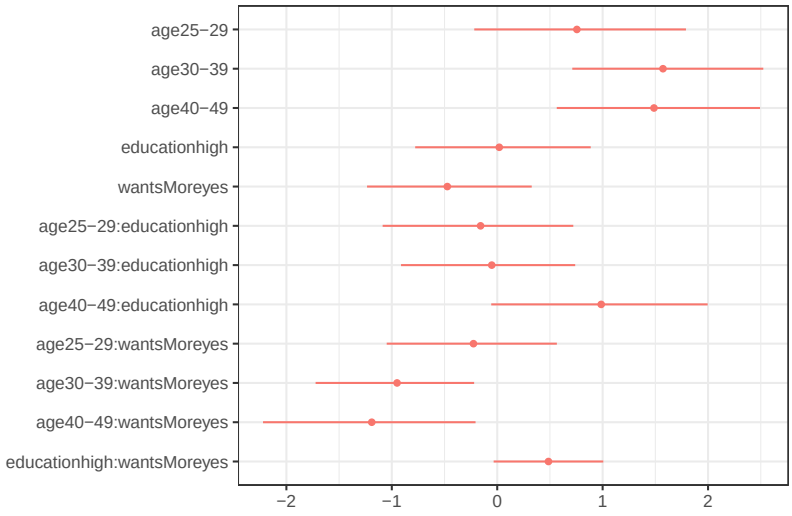
  In this

```
aods3::gof(fit2)

## D  = 2.4415, df = 3, P(>D) = 0.4859584
## X2 = 2.5153, df = 3, P(>X2) = 0.4725266
```

There do indeed seem to be important two-way interactions:

```
drop1(fit2,test="Chisq")

## Single term deletions
##
## Model:
## propUsing ~ (age + education + wantsMore)^2
##                     Df Deviance     AIC     LRT Pr(>Chi)
## <none>                   2.4415  99.949
## age:education        3  10.8240 102.332  8.3826  0.03873 *
## age:wantsMore        3  13.7639 105.272 11.3224  0.01010 *
## education:wantsMore  1   5.7983 101.306  3.3568  0.06693 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(dotwhisker)
dwplot(fit2)
```



Revisiting the AIDS data

```
aids <- read.csv("../data/aids.csv")
aids <- transform(aids,
                  date=year+(quarter-1)/4,
                  index=seq(nrow(aids)))
```

```
g1 <- glm(cases~date, data=aids, family=poisson)
g2 <- update(g1,  . ~ poly(date,2))
```

```
aods3::gof(g1)

## D  = 53.02, df = 18, P(>D) = 2.605035e-05
## X2 = 42.3834, df = 18, P(>X2) = 0.0009773183

aods3::gof(g2)

## D  = 31.992, df = 17, P(>D) = 0.01508225
## X2 = 28.1734, df = 17, P(>X2) = 0.04295199
```

Looks marginal.

```
g3 <- update(g2, family=quasipoisson)
```

```
g4A <- MASS::glm.nb(cases~poly(date,2), data=aids)
g4B <- glmmTMB::glmmTMB(cases~poly(date,2), data=aids, family=nbinom2)
g4C <- glmmTMB::glmmTMB(cases~poly(date,2), data=aids, family=nbinom1)
bbmle::AICtab(poisson=g2,nbinom1=g4C)
```

In this case the fancier model is actually slightly *worse* according to any criteria we measure ...

```
pchisq(-2*logLik(g2)-(-2*logLik(g4C)),lower.tail=FALSE,df=1)
```

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.

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.