Advanced mixed-models workshop: Session 5

Dale Barr

University of Glasgow

Bremen March 2015

Continuous vs. discrete data

Two discrete types of data are common in psychology/linguistics

- categorical (dichotomous/polychotomous)
  - type of linguistic structure produced (X, Y, Z)
  - region looked at in a visual world study (target, other)
  - number of items recalled out of N
  - accurate or inaccurate selection
- counts (no. opportunities ill-defined)
  - no. of speech errors in a corpus
  - no. of turn shifts in a conversation
  - no. words in a utterance

Why not treat discrete data as continuous?

- Proportions range between 0 and 1
- Variance proportional to the mean (expected probability or rate)
- Spurious interactions due to scaling effects (see Jaeger, 2008)

Generalized linear models

- Allows use of regular linear regression by projecting the DV onto an appropriate scale
- Key elements of GLMs:
  - link function
  - variance function

Odds and log odds

Bernoulli trial An event that has a binary outcome, with one outcome typically referred to as "success"

proportion A ratio of successes to the total number of Bernoulli trials, proportion of days of the week that are Wednesday is 1/7 or about .14

odds A ratio of successes to non-successes, i.e., odds of a day being Wednesday are 1 to 6, natural odds= 1/6 = .17

log odds The (natural) log of the odds (turns multiplicative effects into additive effects)

Properties of log odds or "logit"

log odds: $log\left(\frac{p}{1-p}\right)$ or $log\left(\frac{Y}{N-Y}\right)$
where p is a proportion, N is total trials and Y is observed successes

- Scale goes from $-\infty$ to $+\infty$
- Scale is symmetric around zero
- If negative, means that Pr(success)$<$ .5
- If positive, Pr(success)$>$ .5

Logistic regression

DV has 2 categories

model

$\eta = \beta_0 + \beta_1 X$

link function

$\eta = log\left(\frac{p}{1-p}\right)$

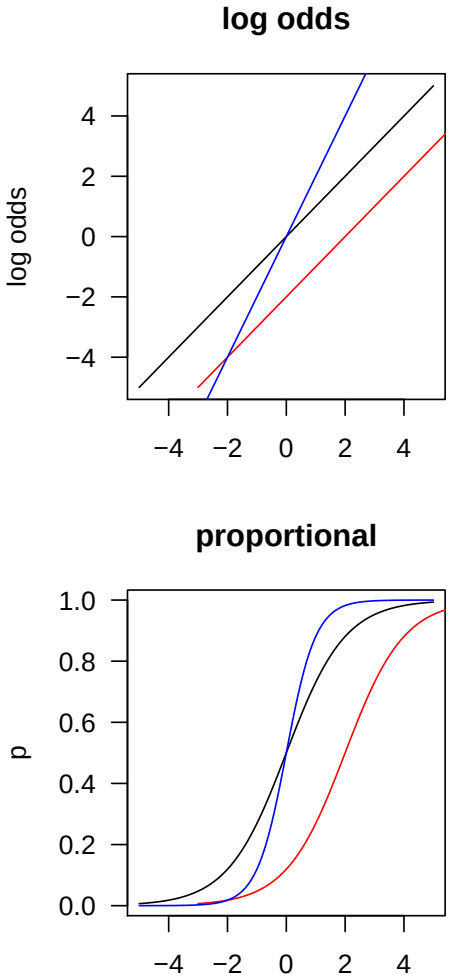
inverse link function

$p = \frac{1}{1+exp(-\eta)}$

getting odds from logit: $\exp(\eta)$

variance function (binomial)

$np(1-p)$



Load in and prepare the data

```
dat <- readRDS("FAN.rds")
dat1 <- subset(dat, Day == 2)

dat1 <- transform(dat1,
                  V1 = (Cond == "same voice") -
                      mean(Cond == "same voice"),
                  V2 = (Cond == "same gender, different voice") -
                      mean(Cond == "same gender, different voice"))
```

Fit a model

- use `glmer()` with optimizer "bobyqa"

```
library("lme4")

mod <- glmer(Accuracy ~ V1 + V2 +
                 (V1 + V2 | SessionID) +
                 (V1 + V2 | ItemID),
             data = dat1, family = binomial(link = logit),
             control = glmerControl(optimizer = "bobyqa"))
```

View results

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: Accuracy ~ V1 + V2 + (V1 + V2 | SessionID) + (V1 + V2 | ItemID)
   Data: dat1
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
  4149.4   4249.3  -2059.7   4119.4     5745

Scaled residuals:
     Min       1Q   Median       3Q      Max
-12.1153   0.1436   0.2594   0.4351   1.3263

Random effects:
 Groups    Name        Variance  Std.Dev. Corr
 ItemID    (Intercept) 0.4185370 0.64694
           V1          0.0202878 0.14244   0.14
           V2          0.0118758 0.10898  -1.00 -0.17
 SessionID (Intercept) 1.3562216 1.16457
           V1          0.0004334 0.02082   1.00
           V2          0.0008073 0.02841  -1.00 -1.00
Number of obs: 5760, groups:  ItemID, 96; SessionID, 20
```

Multiparameter test

```
mod2 <- update(mod, . ~ . - V1 - V2)

anova(mod, mod2)
```

```
Data: dat1
Models:
mod2: Accuracy ~ (V1 + V2 | SessionID) + (V1 + V2 | ItemID)
mod: Accuracy ~ V1 + V2 + (V1 + V2 | SessionID) + (V1 + V2 | ItemID)
     Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
mod2 13 4145.9 4232.4 -2059.9   4119.9
mod  15 4149.4 4249.3 -2059.7   4119.4 0.4632      2     0.7933
```

Conditional versus marginal probabilities

```
pmean <- aggregate(Accuracy ~ Cond, dat, mean)

int <- fixef(mod)
params <- fixef(mod)[-1]

mx <- matrix(c(-1/3,  2/3, -1/3,
               -1/3, -1/3,  2/3), ncol = 2)

df1 <- data.frame(Cond = c("different gender, different voice",
                           "same voice",
                           "same gender, different voice"),
                  logit = as.numeric(mx %*% params + fixef(mod)[1]))
df1$pmod = 1 / (1 + exp(-df1$logit))

merge(df1, pmean)
```

```
                              Cond    logit      pmod  Accuracy
1 different gender, different voice 2.266868 0.9060956 0.8229167
2      same gender, different voice 2.364141 0.9140517 0.8445312
3                       same voice 2.297880 0.9087013 0.8390625
```

Interpreting results: Odds ratios

- use the `exp()` function to get odds ratios

The β associated with `V1` is the change (in logit space) associated with hearing the name in the same voice as in training vs. a different voice
How does that "change the odds" of clicking the right person?

```
c(params["V1"], OR = exp(params["V1"]))
```

```
        V1       OR.V1
0.03101181 1.03149769
```