Advanced Stats: Lec 03 GLMs

Dale J. Barr

University of Glasgow

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
  - hired or not hired
- 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

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(η)

variance function (binomial)

$np(1-p)$

Titanic dataset (kaggle)

```
VARIABLE DESCRIPTIONS:
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1st; 2nd; 3rd)
name            Name
sex             Sex
age             Age
sibsp           N Siblings/Spouses Aboard
parch           N Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg;
                 Q = Queenstown;
                 S = Southampton)
```

```
SPECIAL NOTES:
Pclass is a proxy for socio-economic status (SES)
 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)
 If the Age is Estimated, it is in the form xx.5

With respect to the family relation variables (i.e. sibsp and parch)
some relations were ignored.  The following are the definitions used
for sibsp and parch.

Sibling:  Brother, Sister, Stepbrother, or Stepsister of Passenger
           Aboard Titanic
Spouse:   Husband or Wife of Passenger Aboard Titanic
           (Mistresses and Fiances Ignored)
Parent:   Mother or Father of Passenger Aboard Titanic
Child:    Son, Daughter, Stepson, or Stepdaughter of Passenger
           Aboard Titanic

Other family relatives excluded from this study include cousins,
nephews/nieces, aunts/uncles, and in-laws.  Some children travelled
only with a nanny, therefore parch=0 for them.  As well, some
travelled with very close friends or neighbors in a village, however,
the definitions do not support such relations.
```