0.1 Introduction

Two other kinds of data are common in psychology/linguistics:

- DVs are often 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
  - number of trials in which an error was made
- DVs are sometimes counts (number of opportunities not well-defined)
  - number of speech errors in a corpus
  - number of turn shifts in a conversation

Why not treat categorical data as continuous?

Problems:

- Proportions range between 0 and 1, linear regression range unlimited
- Variance is proportional to the mean (expected probability or rate)
- Application of ANOVA can detect spurious interactions due to scaling effects (see Jaeger, 2008)

Generalized linear models

- A "generalized" approach to linear regression for noncontinuous DVs
- Classical linear models assume an unbounded scale, with uniform variance
- For many cases, scale is bounded, and variance of a DV is proportional to the mean
- A GLM 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., the odds of a day of the week being Wednesday is 1 to 6, natural odds $= 1/6 = .17$

log odds The natural log of the odds (taking the log 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)

$np(1-p)$