Rutgers Linguistics Workshop on Mixed Effects Models

— Mixed effects logistic regression —

09/17/2016

Judith Degen
Stanford University

Logistic regression

- for binary (categorical) instead of continuous outcomes
- instead of predicting the mean of an outcome, we’re predicting the log odds of an event occurring
- also called “logit model”

What kind of data?

- grammaticality (binary)
- syntactic variation (e.g., dative alternation)
- phonological variation (e.g., t-deletion)
- experimental forced choice or eye-tracking data

Why not ANOVA?

- ANOVA over proportion has several problems (cf. Jaeger, 2008 for a summary)
  - Hard to interpret output
  - Violated assumption of homogeneity of variances

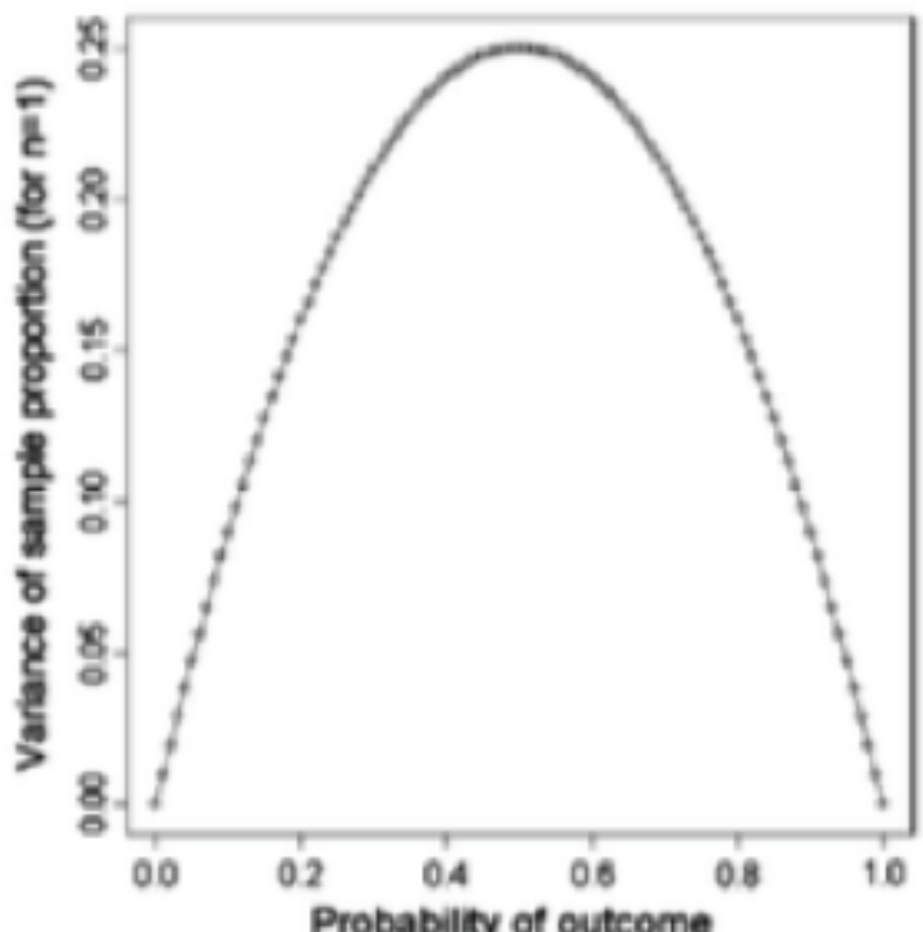


Fig. 1. Variance of sample proportion depending on p (for $n = 1$).

Why can't we use linear regression for categorical outcomes?

The linear model makes impossible predictions (values of $Y > 1$ or $Y < 0$)

The linear model is meaningless if its assumptions are violated

Logistic regression

Recall that **logistic regression** is a kind of **GLM** (with a binomial link function).

- The linear predictor:

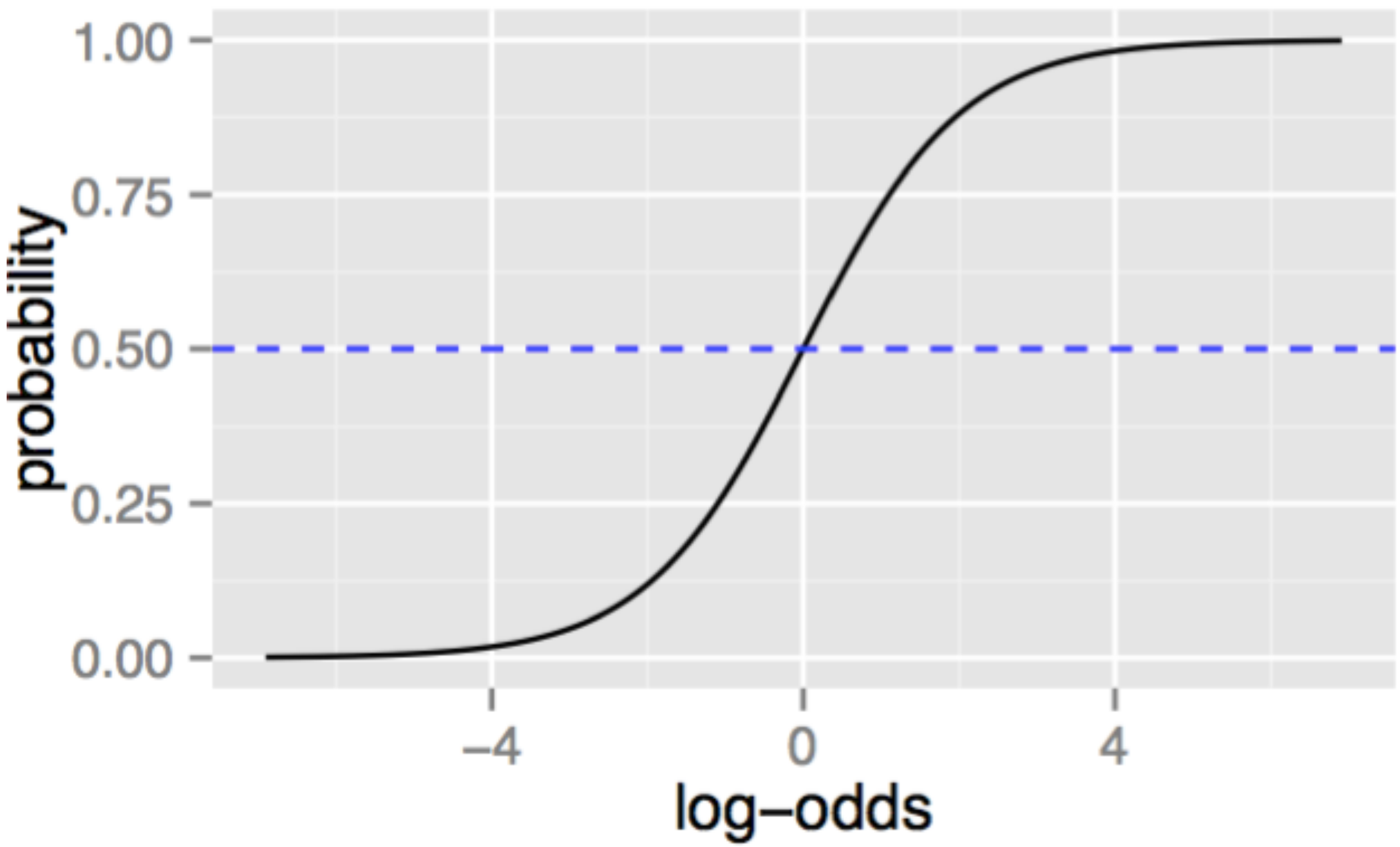
$$\eta = \alpha + \beta_1 \mathbf{x_1} + \cdots + \beta_n \mathbf{x_n}$$

- The link function g is the logit transform:

$$\mathrm{E}(\mathbf{y}) = p = \mathrm{g}^{-1}(\eta) \Leftrightarrow$$
$$\mathrm{g}(p) = \boxed{\ln \frac{p}{1-p}} = \eta = \alpha + \beta_1 \mathbf{x_1} + \cdots + \beta_n \mathbf{x_n}$$

- The distribution around the mean is taken to be binomial.

Log odds and probability



- This relation is particularly clear in the following form of the model:

$$p = \frac{1}{1 + \exp^{-\mathbf{X}\beta}}$$

log odds range from -Inf to +Inf

Mixed effects logistic regression

linear model : mixed linear model ::
logit model : mixed logit model

Assumption: individual differences within a grouping factor are normally distributed in log-odds of event

$$\ln\left(\frac{p}{1-p}\right) = \overbrace{\mathbf{X}\beta}^{\text{Fixed effects}} + \overbrace{\mathbf{Z}b}^{\text{Random effects}} , \overbrace{b_i}^{\sim N(0,\sigma_{b_i})}$$

Back to **`lexdec`**:
Outcome: correct or incorrect response
Inputs: same as in linear model

Let’s translate it into R!