Methods

As an undergraduate thesis, a lot of research into methodology was done. Here I go through some of the essential methodology, while establishing the notation I will use for the rest of this paper.

Logistic Regression

With logistic regression, we seek to fit a model where the response variable is binary. We might consider the response variable, Y, a Bernoulli random variable,

$$Y = \text{Bernoulli}(p),$$

where p is the probability that an observation $y_i = 1$, for any i. (As a binary variable, the support of Y is $\{0, 1\}$, so $y_i = 0$ otherwise.) Thus, in predicting and making inference about a Bernoulli variable, we are concerned with p and how it varies with respect to other quantities.

Logistic regression is, as we will see, one form of regression generalized from linear regression.

Linear regression is the first form of regression most people learn: find the line

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j + \epsilon,$$

based on data with response variable y_i and j predictor variables x_i, coefficients $\beta_0, \ldots, \beta_j$, and error term $\epsilon \sim N(0, \sigma^2)$. We can equivalently write,

$$Y \sim N(\beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j, \sigma^2).$$

Generalized linear regression uses a "link function," g, to modify the regression:

$$g(y_i) = \beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j + \epsilon.$$

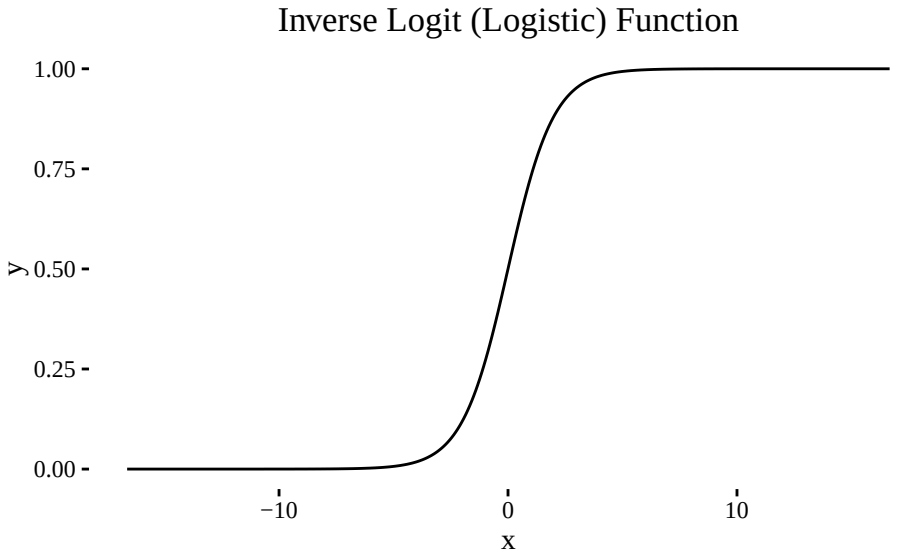
Logistic regression is a form of generalized regression where the 'link' function is the logit function, logit : $[0, 1] \to \mathbb{R}$:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right),$$

also known as the log-odds, odds being $p/1 - p$ for any probability p. So we can model this as a Bernoulli random variable where the probability of a 1 is:

$$\mathbb{P}(y_i = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j).$$

Notice that the inverse logit function maps values from $\mathbb{R}$ to $[0, 1]$. The function provides a convenient way to map linear combinations of other variables onto values that are valid probabilities. Other such functions exist and are also used for regression of binary variables, such as the probit function.



Hierarchical Models and Mixed Effects Models

Many data sets contain nested structures when viewed in some way. For example, a data set of student test scores may contain information about the schools and districts they are in. Or a dataset of soil samples may have multiple samples from each of a set of different sites. In the dataset we examine, rides can be grouped by rider.

Mutlilevel models allow us to address these kinds of relationships in regression models. They provide a number of computational advantages, as we shall describe later.

Description and Notation

These models of course work for other forms of regression, but we will focus on logistic regression, as it is the method we use in this paper. We will be using notation adapted from Gelman's description of multilevel models. Consider a data set composed of

- i observations of a binary response variable y_i,
- m observation level predictors $X_i = x_i^1, \ldots, x_i^m$,
- j groups in which the observations are split into,
- l group level predictors $U_{j[i]} = u_{j[i]}^1, \ldots, u_{j[i]}^l$, where $j[i]$ is the group of the ith observation,.

We could fit a model where the intercept varies by group:

$$\mathbb{P}(y_i = 1) = \text{logit}^{-1}(\alpha_{j[i]} + X_i\beta),$$
$$\alpha_{j[i]} \sim N(\gamma_0 + U_{j[i]}\gamma, \sigma_\alpha^2),$$

where $\alpha_{j[i]}$ is the intercept for the jth group, β are the coefficients for the observation-level predictors, γ_0 are the group-level intercepts, and γ are the coefficients for the group-level predictors. We could also imagine a similar model where there are no group level predictors, such that we simply have different intercepts for each group,

$$\alpha_{j[i]} \sim N(\gamma_0, \sigma_\alpha^2),$$

We can also consider a model that has slopes varying by group. For simplicity, let's consider just one observation level predictor, x_i, that will have varying slopes $\beta_{j[i]}$ as well as one group level predictor. We could specify the model as,

$$\mathbb{P}(y_i = 1) = \text{logit}^{-1}(\alpha_{j[i]} + \beta_{j[i]} x_i),$$

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} = N\left(\begin{pmatrix} \gamma_0^\alpha + \gamma_1^\alpha u_j \\ \gamma_0^\beta + \gamma_1^\beta u_j \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right).$$

Examples and Advantages

Gelman puts forward a framework for thinking about multilevel models as a comprimise between no-pooling and complete pooling. For example, for the school example, one could fit a classical regression ignoring the schoolwide data, with students as the level of observation. (That would be "no pooling".) Alternatively, one could fit a separate regression for each school.