# 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 (i.e. takes on values of 0 or 1, or true or false). 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 $i$ (and probability $1 - p$ for $y_i = 0$). 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 one form of regression generalized from linear regression. Recall that in linear regression we use data with response variable $y_i$ and $j$ predictors $x_{i1}, \ldots, x_{ij}$ to fit the best-fitting linear function

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij} + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$, by estimating $\beta_0, \ldots, \beta_j$. We might can equivalently write,

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

One big problem with this form of regression is that the range does not match the response we want. If we want to measure probabilities as the response variable, we would prefer to have the response limited to [0,1] rather than all of $\mathbb{R}$. Generalized linear regression uses a "link function," $g$, to modify the regression:

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

In logistic regression, the 'link' function is the logit function, $\text{logit} : [0, 1] \to \mathbb{R}$:

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

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. However, logistic regression has better interpretability because it is based on odds.

Interpreting coefficients is slightly different that a linear regression. If we exponentiate the coefficients $\beta_1, \ldots, \beta_j$, we get odds ratios, which tell us the multiplicative effect a one-unit increase in the corresponding predictor has on the odds.

Logistic regression can be thought of both a classification and regression method. We can simply take the estimated probabilities $\hat{p}$ as the response variable, making it a regression method. Alternatively, we can set a

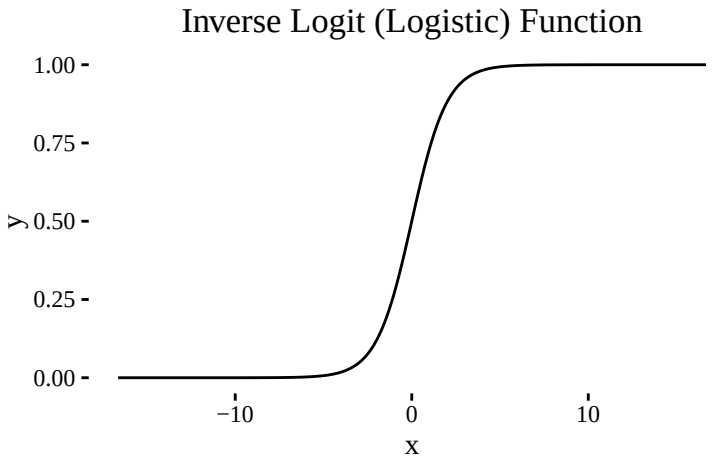


Figure 1: The inverse logit function gives a convenient way to map linear combinations of real numbers to valid probability values.

threshold to classify the observations based on their predicted $\hat{p}$ to get a predicted outcome $\hat{y}$ which we can more easily compare with observed outcomes.

Both can be useful for evaluating models predictive accuracy: simply using $\hat{p}$ avoids having to create an arbitrary decision boundary, but requires us to somehow compute an empirical probabilities from the observed outcomes in order to make a comparison.

## 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 students are in, thus containing the hierarchy of districts and schools. As another example, we might have a dataset of soil samples of which multiple were taken from each of a selection of site, thus containing a hierarchy of sites and samples.

In the data we examine with bicycle rides, we think of rides being grouped by rider. We will talk about different "levels" of variables corresponding to places in this hierarchy. For example, when we refer to ride-level variables, we refer to variables that are specific to a ride, where as we refer to rider-level variables as those specific to the rider, and thus also all the rides that rider takes.

We will also discuss road segment-level variables, which are variables that are specific to road segments in the route of a ride. However, there isn't a clear road segment-ride hierarchy, so in practise this is somewhat different.

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 multilevel models 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 $U_{j[i]} = u^1_{j[i]}, \ldots, u^l_{j[i]}$, where $j[i]$ is the group of the $i$th observation,.

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

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

where $\alpha_{j[i]}$ is the intercept for the $j$th group, $\beta$ are the coefficients for the observation-level predictors, $\gamma_0$ are the group-level intercepts, and $\gamma$ 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,

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

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,

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

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} = N\left(\begin{pmatrix} \gamma^\alpha_0 + \gamma^\alpha_1 u_j \\ \gamma^\beta_0 + \gamma^\beta_1 u_j \end{pmatrix}, \begin{pmatrix} \sigma^2_\alpha & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma^2_\beta \end{pmatrix}\right).$$

**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.

## Tools for evaluating models

There are a couple ways we wish to evaluate our models. Most of the time, we will compare them to some other model.

Predictive accuracy will be one aspect of our model we will want to evaluate. We will use cross-validation to evaluate accuracy, usually with $K$-fold cross-validation. Statistics such as misclassification rate, false-positive rate, and true-negative rate can be calculated for each validation. For a more comprehensive look at predictive accuracy, we use the separation plot.

**The Separation Plot**

The separation plot, created by _____ in their paper _____, is designed to show how well a logistic regression model can distinguish between high and low probability events.

Creating a separation plot first requires a model fit to training data and testing data to evaluate predictive accuracy on. From the testing data, we need a vector $Y$ of observed binary response data and a vector $\hat{Y}$ of predicted probabilities of a 1 for each observation, predicted using our model fitted to training data.

We plot the data ($Y, \hat{Y}$ as a sequence of vertical strips, colored according to observed outcome, $Y$, and ordered from low to high probability based on $\hat{Y}$. A curve is superimposed upon the stripes showing the $\hat{Y}$ as a line graph. And finally, a small triangle is placed indicated the point at which the two colors of lines would meet if all observations $Y = 0$ were placed to the left of all the $Y = 1$ observations; in other words showing where the boundary would be if the two classes were perfectly separated by the model.

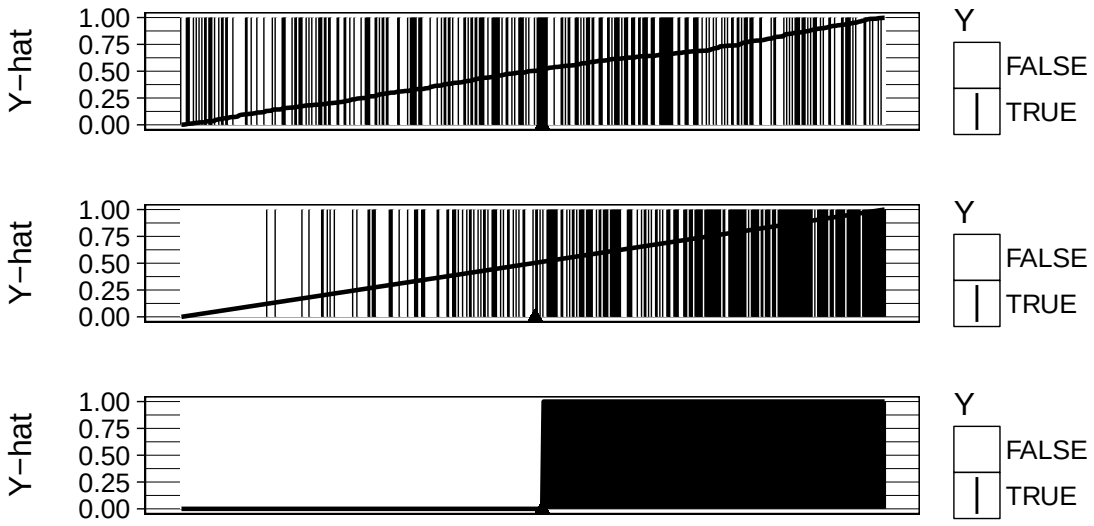


Figure 2: Above we have examples of three separation plots. The first plot shows what it looks like when $Y$ and $\hat{Y}$ are uncorrelated. The second plot