Introduction

Ride Report, created by Knock Software, compiles a crowdsourced data sets of bicycle rides.

The *Ride Report* app combines a simple thumbs-up/thumbs-down rating system with GPS traces of bicycle rides to compile a crowdsourced data set of which routes are stressful for urban bicyclists.

The app privileges reducing barriers to response to increase sample size over ensuring quality and consistent responses. This presents the first problem: how can we analyze ratings when riders are likely rating rides inconsistently?

At the same time, we have another challenge. We have ratings associated with routes, but we would like to know the effect of particular parts of the road network, road segments, for both inference (what effect does this road segment have on the rating?) and prediction (given a route, what do we expect the rating to be?) purposes.

Accounting for Rider Rating Variance

For ratings we are interested in modeling variance between riders (as we might expect different rides to rate differently on average) and within riders (as riders may not rate the same route and conditions the same every time). To model this, we propose using multilevel regression, with random effects from each rider. This approach has been used in similar situations, in one case to model sexual attraction[1].

In a multilevel model, we fit a regression where a slope of intercept term is a random variable whose distribution is unique to all the groupings. For example, if we let r_i be the rating of the ith ride, X_i be the ride-level variables, then we can fit a regression:

$$\mathbb{P}(r_i = 1) = \text{logit}^{-1}\left(\alpha_{j[i]} + \beta \cdot X_i\right),$$

where α_j is the contribution of the j rider:

$$\alpha_j \sim N(\mu_\alpha, \sigma_j^2).$$

We explore multilevels model further in Section 2.1 and multilevel models for riders in Section 4.2.

Addressing Road Segments as a Level

To model road segments, we propose one model that regards segments as a level in the model.

Outlining the terms

We are presented with a interesting data set to examine. (For a through look at the data, read chapter 1.) Ride Report, created by Knock Software, compiles a crowdsourced data sets of bicycle rides. For each bike ride, we have the GPS trace, time, rider, ride length, and a binary (thumbs up / thumbs down) rating of the ride.

Using the time variable, we join in weather and traffic data.

Using some simple map matching, we also join in infrastructure information on street segment level, including road type, and intersection type.

[1] @mackaronis2013

We have a set of data comprised of bicycles rides and binary ratings of those rides. We are seeking to find a model to predict ride rating using predictors that operate on different levels of observation. For example, weather conditions are ride specific, but the presence of a bike lane is specific to the different road segments traveled on during a ride.

Because we have data on different levels of observation—rides, riders, and segments—it will be appropriate to use a multilevel model.

On the ride level, for the ith ride, we have

- r_i, the rating,
- X_i^{rider}, the rider id,
- X_i^{length}, the ride length,
- X_i^{raining}, an indicator variable for rain during the ride,
- $X_i^{\text{rush.hour}}$, an indicator variable for rush hour traffic,
- $X_i^{\text{temperature}}$, the temperature at the time.

On the segment level, for the kth segment, we have

- Y_k^{grade}, road grade,
- $Y_k^{\text{bike.ln}}$, indicator for bike lane,
- $Y_k^{\text{no.car}}$, indicator for no-car road.