Model description

Wesley Brooks

7/31/2020

Overview

This document briefly describes the current state of a Stan model that is used to estimate parameters for bikeshare users and drivers. The model also simulates rides and VMT for unobserved users from the population. In order to do simulation of data from unobserved users, pass the number of users to simulate to the Stan model as the `n_pred` data element. The users' parameters will be simulated, some number of rides will be sampled while drawing each sample, and the mode for which they substitute will be simulated.

Data is returned from the simulation in the samples from the Stan model. The simultated data include:

- `pred_n_rides` Number of simulated bikeshare rides. This is a vector where each element is the count of rides for a specific simulated user.
- `pred_n_rides_carsub` Similar to `pred_n_rides`, but counts how many rides substituted for car travel.
- `pred_substituted_miles` Tallies the number of VMT that were avoided due to bikeshare use by each driver in this pass of the simulation.
- `pred_vmt` Total VMT per user (after removing car-substituting bikeshare miles).
- `pred_rate` Simulated user-level random effect for the (log) number of bikeshare trips. This is the user's Poisson parameter for sampling `pred_n_rides`.
- `pred_log_ranef_ride_dist` Simulated user-level random effect for the log distance of each bikeshare ride by this user.

Rider-level random effects

Sample these random effects from a joint distribution (multivariate log-normal):

$$[\log(r_i), \log(c_i)] \sim N(\boldsymbol{\mu}_{ran.eff}, \Sigma),$$

where $\Sigma = \begin{pmatrix} \sigma_r^2 & \rho_\Sigma \sigma_r \sigma_c \\ \rho_\Sigma \sigma_r \sigma_c & \sigma_c^2 \end{pmatrix}$, $\sigma_r, \sigma_c \sim$ Exponential(1) and $\rho_\Sigma \sim$ lkj_corr(2). The prior for both components of $\boldsymbol{\mu}_{ran.eff}$ is the (independent) student's t-distribution with six degrees of freedom: $\boldsymbol{\mu} \sim t_6(0, 1)$.

Ride frequency

Ride frequency $\boldsymbol{Z}$ (the data are the number of bikeshare rides in a 28-day period, as reported by a user) is assumed to have a Poisson distribution, where the (log) rate parameter is sampled per-user from a population:

$$Z_i \sim \text{Poisson}\{r_i\}.$$

Ride length

The ith user has made Z_i rides in 28 days, and some of those ride lengths are recorded as the vector $\boldsymbol{Y}_i$. The ride lengths for a user are assumed to have a log-normal distribution, where the mean of the log-distance is $\log(c_i)$. The jth ride is distributed as:

$$\log Y_{ij} \sim N\{\log(c_i), \sigma^2_{length}\},$$

where $\sigma_{length} \sim$ Exponential(1).

Mode substitution

The model for mode substitution depends on the length of the ride and some rider-level coefficients. The model says that the jth ride of the ith user substitutes for a trip that would have been by the mode indicated by the random vector $\boldsymbol{D}_{ij}$, which is a vector of six indicators, exactly one of which is indicated. The six indicators, respectively, indicate that the trip would have been by d_1: walking, d_2: ridehail, d_3: car alone, d_4: no trip, d_5: carpool, d_6: transit. The substituted mode follows a multinomial distribution where the relative probabilities of each mode are referred to as the length-six vector $\boldsymbol{\theta}_{ij}$. Here, "relative probabilities" means that the elements of $\boldsymbol{\theta}_{ij}$ need not sum to unity. Instead, the elements are all divided by the sum before being used as probabilities (this is the "softmax" function).

$$D_{ij} \sim \text{multinomial}\left\{\text{softmax}(\boldsymbol{\theta}_{ij})\right\}.$$

The elements of the vector of relative probabilities, $\boldsymbol{\theta}_{ij}$, arise from the model

$$\theta_{ij,k} = a_k + b_{ik} + \beta_k \log\left(\text{distance}_{ij}\right),$$

where k indexes over the alternative modes, a_k is a mode-specific intercept, b_{ik} is a user's mode-specific preference, and β_k is a mode-specific sensitivity of mode substitution to trip distance. The prior distributions for $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{\beta}$ are all independent Student's t-distributions with six degrees of freedom.

$$\boldsymbol{b}_i \sim N(\boldsymbol{\mu}_{mode.sub}, \Sigma_{mode.sub}),$$

where $\boldsymbol{b}_i$ is the length-6 vector of mode-substitution coefficients for the ith user, $\boldsymbol{\mu}_{mode.sub}$ is the length-6 vector of means for those coefficients, and $\Sigma_{mode.sub} =$

$$\begin{pmatrix}
\sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \rho_{1,3}\sigma_1\sigma_3 & \rho_{1,4}\sigma_1\sigma_4 & \rho_{1,5}\sigma_1\sigma_5 & \rho_{1,6}\sigma_1\sigma_6 \\
\rho_{1,2}\sigma_2\sigma_1 & \sigma_2^2 & \rho_{2,3}\sigma_2\sigma_3 & \rho_{2,4}\sigma_2\sigma_4 & \rho_{2,5}\sigma_2\sigma_5 & \rho_{2,6}\sigma_2\sigma_6 \\
\rho_{1,3}\sigma_3\sigma_1 & \rho_{2,3}\sigma_3\sigma_2 & \sigma_3^2 & \rho_{3,4}\sigma_3\sigma_4 & \rho_{3,5}\sigma_3\sigma_5 & \rho_{3,6}\sigma_3\sigma_6 \\
\rho_{1,4}\sigma_4\sigma_1 & \rho_{2,4}\sigma_4\sigma_2 & \rho_{3,4}\sigma_4\sigma_3 & \sigma_4^2 & \rho_{4,5}\sigma_4\sigma_5 & \rho_{4,6}\sigma_4\sigma_6 \\
\rho_{1,5}\sigma_5\sigma_1 & \rho_{2,5}\sigma_5\sigma_2 & \rho_{3,5}\sigma_5\sigma_3 & \rho_{4,5}\sigma_5\sigma_4 & \sigma_5^2 & \rho_{5,6}\sigma_5\sigma_6 \\
\rho_{1,6}\sigma_6\sigma_1 & \rho_{2,6}\sigma_6\sigma_2 & \rho_{3,6}\sigma_6\sigma_3 & \rho_{4,6}\sigma_6\sigma_4 & \rho_{5,6}\sigma_6\sigma_5 & \sigma_6^2
\end{pmatrix}$$

The prior distributions of $(\sigma_1, \ldots, \sigma_6)$ are independent Student's t-distributions with 6 degrees of freedom (restricted to the positive halves of the distributions). The prior distributons of $(\rho_{1,2}, \rho_{1,3}, \ldots \rho_{5,6}$ are all independent lkj_corr(2).