Calculation of By-Stop ORCA Rates

Problem

Given:

- Route X with stops $S_1 \cdots S_n$
- Total ORCA to APC rate $r_X = \frac{o_{\text{total}}}{\text{apc}_{\text{total}}}$ for the route
- Observed ORCA boarding count o_i and census tract population p_i for each $S_i \in X$

We want an estimator $\hat{\mathbf{r}} = \left(\mathbb{E}[r_1] \cdots \mathbb{E}[r_n] \right)$, where r_i represents the ORCA rate $\frac{o_i}{\text{apc}_i}$ for each stop $S_i \in X$, such that $\sum_{i=1}^{n} \frac{o_i}{\text{apc}_i} = r_X$.

Assumptions

Unfortunately, we don't have data on observed APC counts by stop, but we do have the populations of the census tracts within which each stop falls. Let's assume that comparing the observed ORCA count at a stop to the population of the stop's census tract is a good proxy for estimating the ORCA to APC rate. More specifically, the ORCA to population at each stop rate scales linearly with the EV of the ORCA to APC rate. This is a pretty flimsy assumption, but it's the best we can do given the data that we have.

Formal assumption: $\left(\frac{o_1}{p_1} \cdots \frac{o_n}{p_n} \right)$ is linearly dependent with $\left(\mathbb{E}[r_1] \cdots \mathbb{E}[r_n] \right)$. In other words, there is some constant scalar c such that $\forall S_i \in R: \ \mathbb{E}\left[\frac{o_i}{\text{apc}_i}\right] = c\left(\frac{o_i}{p_i}\right)$.

Calculation

Let:

$$\mathbf{o} = \left(o_1 \cdots o_n \right) \qquad \mathbf{p} = \left(p_1 \cdots p_n \right) \qquad \hat{\mathbf{r}}' = \frac{\mathbf{o}}{\mathbf{p}}$$

Under our assumption, $\hat{\mathbf{r}}'$ is linearly dependent with $\hat{\mathbf{r}}$, so there must be some c such that $c\hat{\mathbf{r}}' = \hat{\mathbf{r}}$. Since we are given $r_X = \frac{o_{\text{total}}}{\text{apc}_{\text{total}}}$, we can use the definiton of $\hat{\mathbf{r}}$ to rewrite that r_X as follows:

$$r_X = \mathbb{E}[r_X] = \mathbb{E}\left[\sum_{i=1}^{n} \frac{o_i}{\text{apc}_i}\right] = c\sum_{i=1}^{n}\left(\frac{o_i}{p_i}\right), \text{ where } c\hat{\mathbf{r}}' = \hat{\mathbf{r}}$$

Taking that definition of r_X, we can find $\hat{\mathbf{r}}$ as follows:

$$r_X = c \sum_{i=1}^{n} \left(\frac{o_i}{p_i} \right) \quad \rightarrow \quad c = r_X \sum_{i=1}^{n} \left(\frac{p_i}{o_i} \right)$$

$$\hat{\mathbf{r}} = c\hat{\mathbf{r}}' = \left(r_X \sum_{i=1}^{n} \left(\frac{p_i}{o_i} \right) \right) \hat{\mathbf{r}}' = \frac{(r_X \Sigma \mathbf{p})\, \mathbf{o}}{(\Sigma \mathbf{o})\, \mathbf{p}}$$

Error Analysis

Since the whole point of finding $\hat{\mathbf{r}}$ is to capture the variance of r_X across the route, then we expect lots of variance among its values compared to r_X. Nonetheless, it is still useful to look at the error as a measure of how much variance we have.

In our analysis we used MAE and RMSE. These were calculated as follows:

$$\text{MAE} = \text{average}\,\{(\hat{r}_i - r_X) : S_i \in X, X \in D\}$$
$$\text{RMSE} = \left(\text{average}\,\left\{(\hat{r}_i - r_X)^2 : S_i \in X, X \in D\right\}\right)^{\frac{1}{2}}$$

where D is the dataset containing each route X, each with a unique r_X and $\hat{\mathbf{r}}$.

Using the winter APC counts data with a 15 minute time interval, we calculated an MAE of **1.082** and an RMSE of **4.113**. Approximately **46%** of the predicted by-stop ORCA rate estimates fell within the tolerance interval [0.1, 1]. For the points that fell outside of this interval, we threw them out and replaced them with their corresponding r_D values averaged across each route travelling through the stop and weighted by the routes' APC counts. Note that these error values were calculated before running the tolerance interval.

Although only 46% of our estimates fell within a reasonable interval, we believe that this estimation model is still quite robust given the data to which we have access, and the relatively high MAE and RMSE values indicate that we've captured a good amount of ORCA rate variance within routes.