Problem Statement

Given a dataset of households in need of housing services containing the following information:

1. Household ID
2. Date of Coordinated Entry Enrollment
3. Sub-population identifier (e.g. veteran versus non-veteran)
4. Probability of exiting homelessness if the household received one of the following intervention types: no treatments, rapid rehousing, transitional housing, or permanent supportive housing

And a dataset containing the weekly number of different intervention types available for assignment to a household, finds the optimal assignment to maximize the number of households exiting homelessness that restrains large disproportionalities in assignments across sub-populations.

Set up

Let x_{itj} be a 0/1 indicator for whether or not household *i* during week *t* is assigned treatment *j.* When j=1, this will be the case of "no housing intervention". Let p_{itj} be the probability that household *i* exits homelessness in week *t* if they are assigned treatment j. Let C_{jt} be the number of intervention type j available in week t. The optimization problem from the problem statement becomes:

$$\max_{x_{itj}} \sum_{i}^{N} \sum_{j} p_{itj} x_{itj}$$

Subject to

$$\sum_{i} x_{itj} = C_{jt} \ \forall \mathrm{j} \ \neq 1$$

$$\sum_{j} x_{itj} = 1$$

$$x_{itj} \in \{0, 1\}$$

This is a classical assignment problem from the optimization literature with an added time dimension. Since this optimization will be run each week, it simplifies to the classical assignment problem. In other words, we do not have to optimize over t since we are in a single week.

However, this formulation has no means for dealing with the potential disproportionate treatment assignment across sub-population. To address this, we will introduce a weight to objective function, which will be learned from historical data to help equalize proportionate assignment to housing interventions.

Let α_g be the proportion of group g in the total homeless population taken from historical data.

Let $\gamma_{gt^*} = (\sum_i^{N_g} \sum_{j \neq 1} x_{it^*j})/(\sum_i^{N_g} \sum_j x_{it^*j})$ be the proportion of group g up through time t* assigned to some treatment (other than no treatment).

The proportionality of assignment to treatment is thus $E_{g,t^*} = \gamma_{gt^*}/\alpha_g$

If we take one group to be the base or reference group, we can form risk-ratios as follows:

$$r_{gt^*} = E_{g,t^*} / E^B_{g,t^*}$$

Here E^B_{g,t^*} is the proportionality of assignment of the base or reference group. When the non-reference group has a higher representation than the base group then r_{gt^*} is greater than one, when they are exactly equal then it is one, and when representation is lower than the base group it is less than one.

With this set up we can modify our original optimization problem as follows:

$$\max_{x_{itj}} \sum_i^N ((\sum_j p_{itj} x_{itj}) + (\sum_{j \neq 1} C * (1 - r_{gt^*}) x_{it^*j}))$$

Subject to

$$\sum_i x_{itj} = C_{jt} \ \forall \text{j} \neq 1$$

$$\sum_j x_{itj} = 1$$

$$x_{itj} \in \{0, 1\}$$

The problem introduces a new hyperparameter C which is a weight designed to increase or decrease the likelihood that a household is assigned treatment based on whether or not the group they belong to is over or underrepresented in past assignments to housing interventions (Note the exclusion of j = 1, which is the state of no treatment). For example, if the group is overrepresented then $1 - r_{gt^*}$ is negative, and has the effect of penalizing an assignment to x_{it^*j}.

With this problem set up the problem then becomes finding a value of C that will center the long-term risk ratios around 1 for all groups.

Methods

1. March through each week using the household dataset and number of interventions available that week, and perform the optimization with different values of C.
2. At the end of the dataset, calculate the final risk ratios and compare their values.

Note that this method has a few weaknesses that need to be addressed. It doesn't:

1. Consider the long run implications for choices of C, as we will have about 4 years of data to use.
2. Consider the timing of when households arrive relative to the available set of interventions that week.

Proposed fix to these problems: bootstrapped datasets.

1. In other words, create a new dataset that consists of households first arrival in the Homelessness Management Information System (the database that I am pulling the data from). Households often show up multiple times in the database, so we will want to sample all of a single households data.
2. Estimate the density function for the number of first arrivals in a week.
3. Create a new dataset that samples from number 2 to get the number of arrivals in a week.
4. With the number taken from 3, sample the first arrival dataset with replacement. When a household is sampled here, take all of their enrollment information, including their information after the first arrival.
5. Repeat steps 3-4 for as many weeks as we want in our long run dataset.
6. Now implement steps 1 and 2 from the above methods section.
7. Repeat as many times as computationally feasible.
8. Get the distribution of risk ratios for each C. Pick a final C value.