Thoughts about the subway

Motivation

I really like people watching when I am on public transit. I know the word to describe the realization that the people around you have as vivid lives as yours is called sonder (just because I've looked it up so often) but its interesting to try to work out what you think a person's life may be like just based on the short amount of time that you will have been in the same vicinity as them. This is a very roundabout spiel that explains a little bit about the question that I have now. In my watching of people and being on a lot of public transit I've observed some trends in the way that subway cars are populated.

At first I was deadset on the idea that they were discretely normally distributed, in that the most populated subway car was the one in the center and the density of each subway car would decrease the further away it was from the middle car. Later, however, I became aware of some nuance in this distribution, how time of day or stop at which the subway was affected the variance and realized shape of the distribution.

In general, I find the question pretty interesting as I find myself often moving towards the extremes of the subway train in hopes of being able to more easily find a seat and wanted to build some sort of generalized game-theory behaviour model that would give some insight about the decision making process of an individual when faced with the choice of which subway car to try to enter. Because clearly if there was no cost or necessary "mental capital" associated with thinking to move and to then physically move to subway cars that are likely to be less dense then all subway trains would be uniformly distributed.

In a more academia focussed aspect, however, I find myself to be quite interested in urban and spatial economics and the question of trying to optimize the organization is what tends to ocupy my mind during my time spend on public transit. There could be an argument made that in an ideal world we would place subway stop entrances strategically based on where the density is least in order to minimize costs for an individual and make the density of a subway train evenly distributed.

Initial Notes

There is of course no all encompassing model but I do want to try to get pretty close to including factors that would be involved. For example, although an informed individual may have a decent guess as to how crowded a particular subway train will be in general depending on the time of day and location of the subway stop, there is always a degree of undertainty that we should bake into the model.

Although I've found a couple of data sources that look somewhat promising, just to make sure that this project doesn't take an unreasonable amount of time I won't deal with any of the estimating.

Building the Model

Although people are into all sorts of weird stuff, we will stick to concepts that I personally think most individuals value:

- I think on average, most people would rather sit down on the subway than stand up.
- The above preferences increases the longer the commute an individual has.
- People dislike having to be in a dense subway car.

The shape of the utility curves is something that can get very complicated. Based on my setup and assumptions above, the curve should be inversely proportional to the density of a subway car and the duration of a subway car. But I think the first bullet point speaks to a jump in the utility in terms of the density at which all the seats on a subway car are taken up. This requires a bit more information about the make of subway cars so I'll leave that to later if I happen to have enough time to think about it.

I am also going to assume that the density of a subway car is related to the hour in which the ride takes place (where the peak times are around work start and end times, 9am and 5pm), the stop in which your subway trip begins (where the density is lower the closer to the the ends of the route the subway stops at). Finally, my initial assumption was that population density is approximatly truncated normally distributed (with the max density being 200 and the minimum being 0) with regards to the following variables:

- $x \in [-5, 5] \cap \mathrm{Z}$: the subway car that you choose where 0 is the middle subway car.
- $t \in \mathrm{R}^+$: the length of your subway trip.
- $h \in [0, 24)$: the hour of the day that your trip takes place.
- $s \in [0, 66] \cap \mathrm{Z}$: the stop at which your subway trip begins.

Specifically, let's say that ρ is normally distributed with mean $x_0 e^{-x^2} + h_0 e^{-(h-9)^2-(h-17)^2} + s_0 e^{-(s-33)^2}$ and standard deviation σ. In reality, a truncated normal distribution would be more apt but the derivation for first order condition was so messy so I'll add that later in some appendix type section at the end if I have time. To provide a bit more commentary about why I've chosen this specific distribution; it is mostly because this distribution is relatively easy to handle when taking derivatives (as opposed to something like absolute values) and I wanted to use functions that were symmetric, yield lower values for higher absolute input numbers, and yield higher values for lower absolute input numbers.

And with all this in mind, let's write the utility function as

$$u(\rho, t) = -c|x| - a\ln\left(\frac{t}{b_1}\frac{\rho}{b_2} + 1\right),$$

where an individual's maximization problem is

$$\max_x \mathrm{E}[u(\rho, t) | x, h, s].$$

Optimizing Individual Utility

This problem is nice in the sense that we aren't worrying about any equilibria with respect to multiple agents and rather are just concerning ourselves with one agent's choices (we could say that all the information regarding the choices of other agents is baked into the probability distribution of density). Sparing all the algebra, we can then continue to write

$$\max_x \mathrm{E}[u(\rho, t) | x, h, s] = \max_x \mathrm{E}\left[-c|x| - a\ln\left(\frac{t}{b_1}\frac{\rho}{b_2} + 1\right) \Big| x, h, s\right]$$

$$= \max_x -c|x| - \mathrm{E}\left[a\ln\left(\frac{t}{b_1}\frac{\rho}{b_2} + 1\right) \Big| x, h, s\right]$$

$$= \max_x -c|x| - a\ln\left(\frac{t}{b_1 b_2}\left(x_0 e^{-x^2} + h_0 e^{-(h-9)^2-(h-17)^2} + s_0 e^{-(s-33)^2}\right) + 1\right).$$

And we get an expression that expresses the optimal subway car for one such individual as (given that I haven't made any computational errors)

$$-x^2 + \ln(2ax \pm c) = \ln\left(\pm\frac{c}{x_0}\left(h_0 e^{-(h-9)^2-(h-17)^2} + s_0 e^{-(s-33)^2} + \frac{b_1 b_2}{t}\right)\right)$$

Note the symmetry of choosing an optimal x, the agent doesn't care if they choose x or $-x$ since the way we have formulated the problem is with a lot of baked in symmetry. Of course there are a lot of unknowns in this expression but the general implications of this expression are logically consistent with what we would expect, as out expected density grows throught the $h_0 e^{-(h-9)^2-(h-17)^2} + s_0 e^{-(s-33)^2}$ term, the optimal subway car becomes closer to the center as it makes less sense to commit that much effort for a smaller decrease in density whereas as the cost of walking to further subway cars decreases the first term on the left hand side of the expression increases and thereby makes the optimal car choice further away from the center since it's less costly for the agent to seek out cars with less population density.

Visualizing the Problem

So far we have an equation that expresses the relationship between the optimal subway car choice but it's a bit abstract (at least to me it is) so I'll make a few visualizations to go with it.

In [29]:

```
import numpy as np
import matplotlib.pyplot as plt
from ipywidgets import interact




init_t = 5
init_h = 13
init_s = 33
init_c = 1
init_a = 1
init_b1 = 1
init_b2 = 1
init_x0 = 1
init_h0 = 1
init_s0 = 1

def util_plot(t, c, h, s, a, b1, b2, x0, h0, s0):
    xs = np.linspace(-5, 5, 11)

    mu = (x0*np.exp(-np.power(xs,2))+ h0*np.exp(-np.power(h-9,2)-np.power(h-17,2)) +
          s0*np.exp(-np.power(s-33,2)))

    coeff = t/(b1*b2)
    ys = -c*np.abs(xs) - a*np.log(coeff*mu+1)

    plt.figure(figsize=(15, 5))
    plt.plot(xs, ys)
    plt.xlabel('# Subway Car')
    plt.ylabel('Utility')
    plt.title('Agent Utility')
    plt.grid(True)
    plt.show()


interact(util_plot, t=(0,24,0.5), c=(0,1,0.001), h=(0,24,0.5), s=(0,66,1), a=(0,2,0.5),
                     b1=(0,2,0.5), b2=(0,2,0.5), x0=(0,2,0.5), h0=(0,2,0.5), s0=(0,2,0.5))
```

```
Out[29]: <function __main__.util_plot(t, c, h, s, a, b1, b2, x0, h0, s0)>
```

Future Work

So far, I've only dealt with one agent whose decisions don't have an impact on the "market". I could also expand this scenario to include a finite set of agents who have similar (or dissimilar by changing the cost of walking to futher subway cars if I want to consider a very simple heterogenous agent model) utility preferences and see what the optimal allocation of subway car density then looks like. This may be something I will edit and repost.

Reflections

At work, I keep whatever blog post idea that I'm working on at the time on my whiteboard and people notice it from time to time and talk to me about it. I've gotten a lot of really interesting feedback and insights about this specific problem. The main gist of it is that the distributions of density in a station may not be a very simple normal distribution but rather have multiple modes around where the staircases are placed in a given station. Other suggestions included to think about other motivations for people entering the specific subway car that they enter in such as planning for their future exit or people tending to actually have a preference for dense subway cars at nighttimes because they want to feel safer.

So far I can think of a couple of ways to bake these assumptions into the model (such as making the utility function itself also conditional on time of day - perhaps with some transformation such that by changing the time of day we can also change the shape and modality of the utility function - or making some unique way to classify the number of station on the line of stations in such a way that we can codifiy the conditional expectation of ρ properly).

For now, I left the question as is right now without addressing these issues just because I wanted to post at least something instead of taking an exorbitant amount of time on trying to work around every potential nuance.

Apendix

A. Truncated Normal Distribution

Consider that instead of assuming that population density in subways cars was normally distributed conditional on x, s, h they were truncated normal. That is, we do not allow for negative density or for a car to carry more people than is possible; the pdf function is 0 at all values less than equal to 0 or greater than equal to 200. All this would change is the conditional expectation expression in maximization problem but it involves a bit more algebra for the first order conditions. I won't write out everything explicitly but with a truncated normal distribution we would assume some mean $\mu = x_0 e^{-x^2} + h_0 e^{-(h-9)^2-(h-17)^2} + s_0 e^{-(s-33)^2}$ and standard distribution σ and lower and upper bounds $\alpha = \frac{0-\mu}{\sigma}$ and $\beta = \frac{200-\mu}{\sigma}$ respectively. It would then follow that the mean of this distribution is

$$\mu + \frac{\varphi(\alpha) - \varphi(\beta)}{Z}\sigma$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)$ and $Z = \Phi(\beta) - \Phi(\alpha)$ (Φ is the standard normal CDF).

Processing math: 100%