User Inputs and Scale Economies

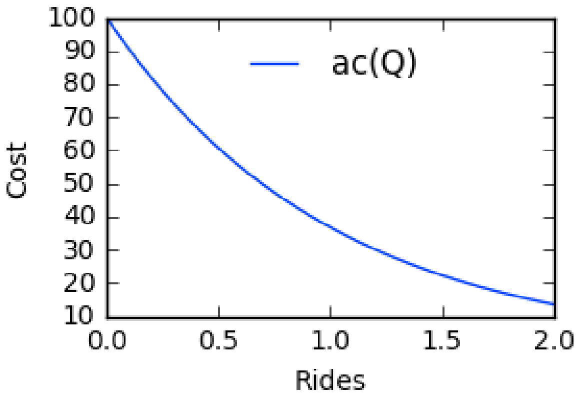
January 28, 2017

```
In [2]: import matplotlib.pyplot as plt
        import matplotlib
        matplotlib.style.use('seaborn-bright')
        import numpy as np
```

1 What are scale economies?

One thing that makes transit systems so impressive is their scale. There are palacial stations hundreds of feet under the earth, grand webs of tunnels and depots full of buses coming and going like huge ants. Nearly everything about transit systems seems to be huge, and it's not an accident: like blast furnaces and oil barges and prisons, they're built large to reap the advantages of *economies of scale.* By "economies of scale," we mean the average cost of a transit trip is cheaper when the system supplies lots of trips. So if Q is the number of rides on our system and $ac(Q)$ is cost per trip, then $ac(Q)$ declines, like this:

```
In [3]: Q = np.linspace(0,2);
        #here we have ac as negative exponential. this is arbitrary.
        #it's just a good example function because it never gets to zero but it alw
        ac = 100*np.exp(-Q);
        plt.close();plt.figure(figsize=(3,2));plt.plot(Q, ac,label='ac(Q)');
        plt.legend(loc='upper center',frameon=False);
        plt.xlabel("Rides");plt.ylabel("Cost");plt.show();
```



2 Sources of scale economies in transit

What are the most obvious sources of economies of scale in transit? Probably those that arise from *sharing* some piece of equipment or infrastructure. When 100 people ride a bus, they can split the cost of compensating the driver over more people than if only 10 rode. But the point of this lecture is that scale economies also arise, less visibly, from a fact particular to urban passenger transportation: users' time is an input to the production of rides—just as much as steel and diesel fuel and drivers' wages are.

Compare transit to another utility: water. Water service has significant *physical* scale economies. Still, everything relevant my transaction with the water company is taken care of on my water bill: I turn on the faucet, water comes out, and I get charged. But to ride a bus, on the other hand, I pay the bus fare *and also*

- walk to the station
- wait for the bus
- and ride on the bus for a while.

Now, among these time costs, *some* have a special property: they get smaller as more service is provided overall. This happens because transit service is not provided continuously across space and time; a bus or train arrives at particular times and at particular places. So when an agency provides more service overall, often it also provides service at new times and places. These new times and places might be closer to *when* I want to travel (meaning less wait time) or *where* I want to travel (meaning less walk time). The upshot is: *the more service is provided, the more convenient the service is.* This idea is called the "Mohring effect" because it was first made explicit by Herbert Mohring, an economist from of University of Minnesota.

Today we are going to look at a series of models that illustrate this idea. For simplicity, these are all models with **exogenous demand**; that is, the number of riders is fixed; it does not depend on quality or price.

3 Model 1: fixed load

Let's say you have a shuttle service from point A to point B with the following parameters:

- F Frequency (buses/hr)
- Q ridership (pax/hr)
- W average wait time at the stop (hrs)
- c_F the money needed to increase frequency by one ($/bus)
- c_w user cost/hour spent waiting. ($/hr)

Given these variables and parameters, the total social cost (TSC) of the service (ignoring time traveling) is

$$TSC = QWc_w + Fc_F$$

where QWc_w is the user's cost and Fc_f is the agency's cost.

The wait time, W, depends on the headway. The headway is $1/F$. If people arrive randomly at the stop, then the average wait time is $W = 1/2F$. Therefore…

$$TSC = Qc_w/(2F) + Fc_F.$$

Now suppose that demand doubles to $2Q$, and that the agency responds by doubling F. Now we have a new total social cost

$$TSC' = 2Qc_w/(4F) + 2Fc_F = Qc_W/(2F) + 2Fc_F$$

Note the first term of this equation (aggregate wait time) is the same as it was before demand doubled! So twice as many people are riding the bus, but all in all they are waiting the same aggregate amount of time. The reason is that headways have been cut in half. Therefore, while demand grew 100%, total cost grew *less than* 100%. Therefore, the cost *per rider* (average social cost) fell.

4 Model 2: variable load

Above, frequency was assumed to rise one-to-one with Q. It doesn't have to. Instead, let's choose F to minimize TSC (given exogenous Q). To do so, we take the the first-order condition with respect to F. We set the derivative with respect to F equal to zero...

$$\frac{\partial TSC}{\partial F} = -\frac{Qc_w}{2F^2} + c_F = 0.$$

Solving this expression for F gives the optimal frequency F^*

$$F^* = \sqrt{\frac{c_w}{2c_F} \cdot Q}.$$

In other words,

$$F^* \sim \sqrt{Q}.$$

Optimal frequency should scale with the square root with ridership, not ridership itself. This result is called a *square root rule*. Likewise, the number of riders per bus (which we'll call L for "load") also scales with the square-root of ridership. Note that $L = Q/F$ and so...

$$L^* = \sqrt{\frac{c_F}{2c_w} \cdot Q}$$

$$L^* \sim \sqrt{Q}.$$

So to handle an increase in ridership, the agency should increase both the number of buses and the number of people per bus.

4.1 Why shouldn't service scale *linearly* with ridership?

4.1.1 or "Why shouldn't we keep a constant load per bus?"

The square-root rule isn't sacrosanct. Below we'll look at some richer models where optimal frequency scales to different powers of ridership than $1/2$. But even this simple model raises a big question: Why shouldn't service scale *linearly* with ridership? If ridership doubles, why shouldn't we run twice as many buses?

Imagine that all the costs of running the buses fall on the riders through their fares. In this case, the fare is equal to

$$P = \frac{c_F F}{Q},$$

which is just the total cost of the service spread over the number of riders. Therefore, when ridership rises by $x\%$, the agency can choose between two extreme options:

- cut P by $x/(1+x)\%$ and hold F constant
- hold P constant but raise F by $x\%$.

The optimal choice is to do something in-between: use some of the extra money to cut fares by putting more people on each bus, and some to run more buses. The reason is that the costs we care about involve *headway*, but headway is proportional to $1/F$, not F. To cut headway by 50%, you need 2x buses; to cut headway by 75%, you have to pay for *4x* as many buses. So there are diminishing returns to adding buses.

Here is a Desmos page I created to illustrate.

However, you don't have to assume that costs always show up as fares to appreciate the tradeoff between having a higher frequency and saving money. If costs aren't paid for via fares, they must be paid for by the government. If that means the transit budget grows, there is less money for education, health care, pensions and security. But if the transit budget is fixed, spending more money on one transit project means cutting spending on other transit projects. These tradeoffs, of course, don't imply you should cut a project's budget down to nothing. The alternative expenditures might not be as worthwhile as the one you're analyzing: some of what doctors, police and teachers do is utterly useless, and some of it is even actively harmful but happens anyway for political reasons.

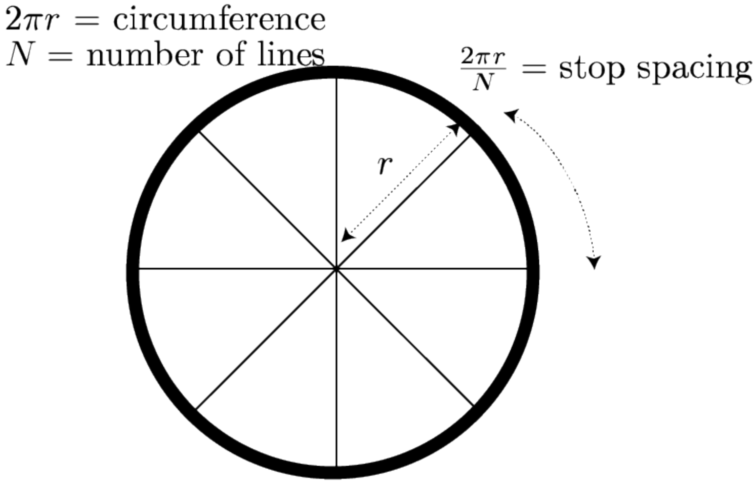
5 Model 3: walk time

To ride the bus (or any other mass transit), passengers don't just wait at the stop; they also walk to the stop. This time is an input into the production of the rides and should be counted in total social cost. Now we'll look at a model. In this model, based on Small (2004) (in bCourses), everyone lives along the periphery of a circular city with radius r. They start their journeys walking from points that are evenly distributed around the circle, walk to the nearest bus line, and then ride into the middle of the city. Here are the parameters:

- N number of bus lines
- Q riders/hr
- M bus-miles/hr
- F frequency (bus/hr)
- S space between lines (miles)
- r radius of the city (miles)
- c_M agency cost of a bus-mile ($/bus-mile)
- c_W cost of users' wait time ($/hr)
- c_K cost of users' walk distance ($/mi) ($c_W$ was already taken)

A figure of the setup:

As before, riders wait on average $1/2F$ at a stop. If riders start their journeys evenly distributed around the circle and walk to the nearest stop, then the average walk distance is $S/4$ (Convince yourself!). Therefore, total social cost is



$$TSC = c_M M + Q \cdot \left[\frac{c_w}{2F} + \frac{c_K S}{4}\right],$$

where the term in brackets is the individual user's wait and walk time, weighted by c_W and c_K. Our task is to choose N and F to minimize TSC. To do so, we'll start by eliminating the variables M and S.

5.0.1 Writing S in terms of N

Since bus lines are evenly spaced, and the circumference of the city is $2\pi r/N$, then the distance between bus lines is $2\pi r/N$. Therefore,

$$S/4 = \pi r/2N$$

is the average walk distance.

5.0.2 Writing M in terms of F and N

Imagine that you were to stand on an infinitesimal slice of bus-route in this city—a slice of length dx miles—and you were to record all the bus-miles traveled in that slice over one hour. How many bus miles would you count? Since frequency is F buses per hour, you would record dx bus-miles exactly F times. And to get the *total* bus-miles per hour, M, you add the bus-miles measured over all such slices. There are Nr miles of bus-routes (look at the picture above), so that sum can be found by the integral

$$M = \int_0^{Nr} F dx,$$

which is just

$$M = FNr.$$

So we have

$$TSC = c_M FNr + Q \cdot \left[\frac{c_w}{2F} + \frac{c_K \pi r}{2N}\right].$$

5.0.3 Solving for F^* and N^*

Now we take the first order condition for F and N.

$$\frac{\partial TSC}{\partial F} = -\frac{Qc_w}{2F^2} + c_M Nr = 0$$

$$\frac{\partial TSC}{\partial N} = -\frac{Qc_K \pi r}{2N^2} + c_M Fr = 0$$

Solving these leads to the expressions for optimal F and N...

$$F^* = \sqrt{\frac{c_W Q}{2c_M Nr}}$$

and

$$N^* = \sqrt{\frac{c_K Q}{2c_M F}}.$$

But this isn't the end of the line, because each expression has the other variable in it. Let's plug the variables into each other's equations and reduce...

$$F^* = \sqrt{\frac{c_W Q}{2c_M r\sqrt{\frac{c_K Q}{2c_M F^*}}}} = \left(\frac{c_W}{rc_K}\right)^{2/3}\left(\frac{Q}{2c_M}\right)^{1/3}$$

$$N^* = \sqrt{\frac{c_K Q}{2c_M \sqrt{\frac{c_W Q}{2c_M r N^*}}}} = \left(\frac{c_K r}{c_W}\right)^{2/3}\left(\frac{Q}{2c_M}\right)^{1/3}$$

Don't get bogged down in all the details here. The important thing is the power laws...

$$F^* \sim Q^{1/3}$$

$$N^* \sim Q^{1/3}$$

.

And since $FNr = M$ is the equation for bus-mileage, this means that

$$M \sim Q^{2/3}.$$

5.1 Takeaways from the model with walk time

Don't focus too much on the coefficients and powers here. The takeaway is this: service should scale more slowly than ridership (because it scales to the 2/3 power rather than linearly), so you get more people per bus. It's just that in this model, M represents service rather than F, and it scales faster than in the previous model. The reason is that increasing M now does two jobs: it reduces wait times *and* walk times. F itself scales more slowly than in the previous model, though, because, for a given M, increasing frequency means incerasing route spacing (and thus increasing walk times).

6 Model 4: boarding time

So far we have only focused on the *benefits* of additional riders. When more people ride, the agency can afford to run more service, which makes riding more convenient. But there are also downsides to additional riders. This model deals with one of hte main downsides: **boarding time**. It is based on Janson (1979) (in bCourses).

Boarding time is the time it takes passengers to get on the bus. For simplicity, we're going to leave route spacing out of this model. Imagine a single bus line that runs between two points, stopping along the way. Here are the parameters.

- Q riders/hr
- M bus-miles/hr
- F frequency (bus/hr)
- D route length (mi)
- l average trip length (mi)
- t_B boarding time (hr)
- c_M agency cost of a bus-mile (\$/bus-mile)
- c_W cost of users' wait time (\$/hr)
- c_IV cost of users' in-vehicle time (\$/hr)
- v bus speed between stops (mi/hr)
- p pace of the bus, counting stops (hr/mi). 'Pace' is a term from engineering that tells how long it takes to travel some unit of distance.

Total social cost is now

$$TSC = c_M M + Q\left[\frac{c_W}{2F} + c_{IV} lp\right],$$

where the product lp is equal to *how long the average passenger spends on the bus*.

6.0.1 Writing p in terms of frequency

Let's calculate p in terms of other variables. To do that, note that if there are M bus-miles traveled per hour, and Q rides per hour, then a single bus must pick up Q/M people. Therefore, for every mile it travels, the bus spends $1/v$ hrs traveling and $t_b Q/M$ sitting still while people board. It follows that

$$p = \left(\frac{1}{v} + t_B\frac{Q}{M}\right).$$

6.0.2 Finding F^* (optimal frequency)

As above, M is the frequency F times the total route-mileage. Above the total route-mileage was Nr; here it's D. So we have

$$TSC = c_M DF + Q\left[\frac{c_W}{2F} + c_{IV} l\left(\frac{1}{v} + t_B\frac{Q}{DF}\right)\right].$$

Pause now to remind yourself what each term means. The first is agency cost. The second is user cost. Within the brackets, the first term is wait cost, and the second is riding time cost. Now take the first-order condition on F...

$$\frac{\partial TSC}{\partial F} = c_M D + Q\left[-\frac{c_W}{2F^2} - c_{IV} t_B l \frac{Q}{DF^2}\right] = c_M D - \frac{Q}{F^2}\left[\frac{c_w}{2} + c_{IV} t_B l \frac{Q}{D}\right] = 0.$$

Solving for F gives

$$F^* = \sqrt{\frac{Q}{c_M D}\left(\frac{c_w}{2} + \frac{t_B c_{IV} l Q}{D}\right)}.$$

6.0.3 Takeway from the model with boarding time

Now that we've included boarding time, F^* scales with Q much faster than it did without boarding time. Mathematically, that's because we have the term $(t_B c_{IV} l Q^2)/(D^2)$ inside the parenthesis. But this should make sense to you intuitively: when passengers take time to get on the bus, you want to run buses more frequently. Higher frequency means that there are fewer people on each bus, so the externality created when each person gets on the bus is smaller; and also each bus picks up fewer people.

7 Takeaway from all the models

We have just looked at four models of simple bus systems. If you're not used to these types of economic/engineering models, you might be tempted to ask, "Which is the most realistic?" Or "The most realistic model would have line spacing *and* boarding time." You might think "These models are worthless because they don't count [insert real-world fact here]."

What I want you to remember, here and throughout the class and even your careers, is that **mathematical models are just stories**. They are really just more logically rigorous versions of Aesop's fables or Zen koans or ancient myths common in every culture. This isn't just true because our models involve people: in physics classes you learn about "free body diagrams" and "frictionless surfaces," but no such things have ever existed nor ever will. In every engineering choice you make, you have to ignore some real facts. The trick is knowing what's relevant and how you can capture it in a pretty way that illuminates more than it obscures. For example, if you're designing an airplane, you can operate in the world of Newtownian physics: forces acting on the plane only accelerate it. But if you're desining a rocket ship that goes incredibly fast, you have to take into account relativistic effects whereby the mass of the vehicle changes as it acquires energy.

So when you look at a model, you have to ask, "What lesson is this story trying to tell?" I think these models have told a few lessons:

- Transit service doesn't just exhibit *physical* scale economies, it also has scale economies that arise from the fact transit service is provided in particular times and places. The total cost of transit service (counting the agency and the users) generally falls as ridership rises. That means anything that raises ridership serves a broader social purpose: it helps all the people who are already riding transit.
- Don't simply increase frequency linearly with ridership, so as to maintain a constant load (number of people per bus). Instead, when ridership rises, put more people on each bus to save money; and run more routes so that people don't have to walk as far.
- Increase frequency more quickly when boarding times are long. That way, there are fewer people on each bus who get slowed down when another person gets on, and each bus doesn't have to pick up as many people.