PPPA 6007: Microeconomics for Public Policy I
Fall 2023

Use Numbers: Assignment 3 of 3
Producer Behavior

Due November 7, 2023

In this assignment we analyze information on producers. Specifically, we study the railroad industry, which provides very detailed data to the federal government.

We are using data the railroads report on revenue received for each shipment.[1] Each individual shipment files a piece of paper (or what used to be a piece of paper) called a "waybill." Railroad company CSX defines a waybill as "A shipping document prepared by a carrier at the point of origin showing the point of origin, destination, route, shipper, consignee, description of shipment, weight, charges and other data necessary to rate, ship and settle."[2] A shipment need not be a whole train.

The data we are using is a random sample of all waybills filed in 2018, 2019, 2020, and 2021 (most recent available) by railroads that shipped goods in the United States. This includes US, Canadian and Mexican railroads (there are very few Mexican railroads, FYI). For each waybill, we observe a variety of things about the shipment, including, but not limited to

- number of cars (rail cars)
- revenue per car-mile (so revenue from the waybill divided by the number of cars times the number of miles)
- the origin location of the shipment
- the destination location of the shipment
- shipment weight
- type of commodity of the shipment

For purposes of this assignment, we will assume that this market is perfectly competitive and that the revenue – and therefore the prices – we observe are linked to costs. In a perfectly competitive market (as we will learn in Lecture 10), firms maximize profits when price equals marginal cost. Thus, if we observe revenue, which is $P * Q$, and we know Q, which we do here, we can infer information about costs, since $P = MC$.

In other words, we observe revenue per car-mile, or P, for each shipment. This is the amount of money the railroad receives per rail car per mile traveled. Thus, if we assume that the market is perfectly competitive, revenue per car mile is equal to marginal cost.

[1] See the data source here.
[2] See https://www.csx.com/index.cfm/about-us/company-overview/railroad-dictionary/?i=W.

Further, we use these data to gain insight into the determinants of cost. Because we observe a measure of cost (price = marginal cost) and other determinants of cost, we can look at the relationship between these two things.

For all questions after the first, you will look in the data for some features that you think should determine costs. You'll make a hypothesis. For example, you might hypothesize that heavy goods cost more to transport than light goods because they require more energy. You test this hypothesis choosing 3 or 4 products and looking in the waybill data.

As with the previous assignment, Google and news archives should be sufficient to answer these questions. This is not a major research paper, so please scale your effort accordingly.

Also as in the previous assignment, you are welcome to discuss parts of this assignment with other students. However, any work you turn in must be your own and written in your own words.

To make graphs, use the software of your choice. Please come see me or the TA if you have questions about how to make graphics in the software of your choice.

Here are datasets of already-prepared waybills data. I have created big and small versions. Unless you have a lot of computing power, I recommend the small version (about 97,000 rows); the large version has almost 1 million rows.

- Small waybills data csv version here, and R dataset version here.
- Big waybills data csv version here, and R dataset version here.

1 Questions

1. First, we want to understand how much the price (revenue per car-mile, by assumption) varies across waybills and by year (2018 to 2020). This is helping us understand whether the railroads charge all customers and products the same, or if they vary price based on the type of shipment.

To understand the extent of variation in price, you will calculate summary statistics using the waybills data linked above. The variable of interest is `rev_per_carmile`, calculated as `freight_revenue / (number_of_carloads * est_short_line_miles)`. Find the mean, median, standard deviation, 25th, and 75th percentiles for each year (in Excel, there is a "percentile" function that helps you do this), as well as the number of observations in each year. The variable for year is `waybill_year`. Report your estimates in a well-labeled table. Let us know if you are using the big or small data.

2. Interpret these statistics

   a Suppose you wanted to transport one railroad car of policy briefs to Los Angeles in 2019, and that the railroad charged you the 75th percentile price. How much would the 3000 mile trip cost?

   b Explain what must be true about the distribution of prices given that the mean is so much larger than the median.

3. Using the same data as the previous question, for 2019 only, choose any three commodities. To find these commodities, you need the name of the variable that identifies commodities, which is `commodity_code`. This variable takes on values of a numeric code. To understand the code, look at this document (I extracted these pages from the full Waybill Reference Guide; see numbered page 112 or pdf page 115 in the full document). For example, raw cotton is code 0112, and popcorn is 01152. All agricultural products have codes that begin with 01.[3]

This list linked above contains all commodities for which the Surface Transportation Board defines a code. There may be codes for which there are no shipments, so be flexible in your choices.

For the three commodities you've chosen, make an initial hypothesis that rank orders the commodities by cost to ship, from largest to smallest (e.g., one might hypothesize that beer costs more to ship than cotton). Remember that the variation in revenue per car-mile is equivalent to the variation in marginal cost, given our assumptions. Then, for each of your three commodities, use the data to find the average across all waybills of

- revenue per carmile
- weight (use `acutal_weight_tons`)
- distance traveled (use `est_short_line_miles`)

Report

- your initial hypothesis, with justification
- a well-labeled table showing the average values requested above for each of your three commodity types
- an evaluation of whether your hypothesis was correct or not (it's absolutely fine if your hypothesis is incorrect; see if you can figure out a reason why)

As best you can, use what we've learned about costs and competition to explain this pattern of revenue by commodity. A few short paragraphs is sufficient to answer this question.

4. For each waybill, the railroad reports the origin of the shipment (`origin_bea_area` and the destination of the shipment (`termination_bea_area`). To describe these areas, the Surface Transportation Board uses three digit codes created by the Bureau of Economic Analysis (BEA). To find what these codes mean, see this file (or starting on page 81 in the full document). If you look for Los Angeles County, you will see that it, along with Orange, Riverside, San Bernardino, and four other counties are all in BEA area 160. San Diego County is alone in BEA area 161.

Choose three origin-destination pairs; these are your first three pairs. Now flip the origin and destination (e.g., make Chicago to LA into LA to Chicago) so that you have three more origin and destination pairs. Make a hypothesis about which pairs should have, on average,

- the most traffic (total number of carloads)

[3] I created a variable for this two-digit code called `commodity_code_2digit` if you are interested.

- the heaviest loads per car
- the highest revenue per carmile

When making your hypothesis, keep in mind that traffic is likely not the same in both directions. In other words, traffic from Chicago to Los Angeles (Chicago is the origin, Los Angeles is the destination) is probably different than from Los Angeles (origin) to Chicago (destination). Ideally you'll have an overarching idea about what drives rail traffic that will apply to all your pairs (e.g., westward travel differs from eastward travel).

Use the data to then find averages across all years (that is, one average) for the variables listed above for each pair in both directions. So, if you chose Chicago and Los Angeles, you'd have two averages of each of the variables below: Chicago to Los Angeles, and Los Angeles to Chicago. In other words, your table should have six rows (three pairs, both directions) and three columns of data (each of the variables listed below).

Report

- your initial hypothesis, with justification
- a well-labled table showing the average values requested above for each of your six pairs
- an evaluation of whether your hypothesis was correct or not (not correct is fine; see if you can figure out a reason why)

As best you can, use what we've learned about costs and competition to explain this pattern of revenue by location. A few short paragraphs is sufficient to answer this question.

5. Now that you've looked at a few cases, let's see what the data have to say when we look across all waybills. I have created two additional datasets based on the waybill data that you've just been using.[4] In the first, I calculated the average revenue per car-mile per ton (so the revenue per car mile we used before divided by the weight of the shipment) for many shipment weights (download here). In the second, I calculated average revenue per car-mile by the distance the shipment goes (download here).

For each of these datasets, make a plot. For the first, plot weight (horizontal axis) versus revenue per car-mile per ton (vertical axis). Then repeat this graph, but using distance instead of weight on the horizontal axis.

Do you see any evidence of economies of scale here? To answer this question, you first need to know what you would expect about price as production increases – here "production increases" means heavier or longer shipments. Then you need to figure out whether the graphs you plot agree with this hypothesis.

Your final answer should be your two graphs, an explanation of what the graphs would look like if there were economies of scale, and then an evaluation of whether you think there are economies

[4] In principle, you could create these yourself; I am doing it to speed things along.

of scale in play. One or two short paragraphs should be sufficient to answer this question.

6. Look at the variables provided in the waybill data and make a hypothesis about another dimension in the data that could influence costs. Remember that you have four years of data, and that some important things may have changed over time in these three years (though you need not focus on changes over time if you'd rather do something else). Write down your hypothesis, and then use the data to test the hypothesis using a few examples as we did in questions 2 and 3 (or if you want to be ambitious and do it for all categories, as we do in question 4, go right ahead). Write a brief paragraph that discusses your hypotheses and what you found how your graph shows this additional element of demand.

2 How to turn it in

Turn this assignment in to your box folder.

Name the assignment "lastname_use_numbers_3". So mine would be "brooks_use_numbers_3."

Turn in one pdf that has all parts together. This makes sure that what you turn in is what we see.

3 Data

Waybill Dataset

The waybill dataset has many many variables. I describe the key ones you'll need in the questions above. If you'd like to look at additional variables, see the user guide to these data here, and look at the listing of variables starting on numbered page 101 (pdf page 104).

Weight and Revenue per Car-mile

I used the waybills data above to calculate the average revenue per car-mile for different carload weights. Variables are as below.

Variable	Definition
`average_weight_per_carload`	Weight for carloads
`average_rev_per_carmile`	freight revenue / (miles traveled * carloads)

Distance Traveled and Revenue per Car-mile

I used the waybills data above to calculate the average revenue per car-mile for different shipment distances. Variables are as below.

Variable	Definition
average_weight_per_carload	Average weight for carloads
average_rev_per_carmile_per_ton	(freight revenue / (miles traveled * carloads))/weight of shipment
average_distance	shipment distance traveled