PPPA 6007: Microeconomics for Public Policy I
Fall 2020

Use Numbers: Assignment 3 of 3
Producer Behavior

Due November 24, 2020

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

We are using data on the revenue railroad firms receive from each individual shipment. 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." [1] A shipment need not be a whole train.

The data we are using is a random sample of all waybills filed in 2018 (the most recent year 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 11), firms maximize profits where price is equal to 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, for these shipments, we observe revenue per car-mile, or P. 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, then revenue per car mile is equal to marginal cost.

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

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.[2]

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, you can use Excel, R (instructions at the end; I suggest this only if you have time to invest) or the software of your choice. We can support technical questions in Excel or R (Brooks and Bayar only).

1 Questions

1. Evaluate how much variation exists in revenue per car mile by calculating summary statistics. Remember that the variation in revenue per car-mile is equivalent to the variation in marginal cost, given our assumptions. Use waybills data (available here, or the smaller version here; either is ok) and the variable `rev_per_carmile`, calculated as `freight_revenue` / (`number_of_carloads` * `est_short_line_miles`). Find the mean, median, standard deviation, 25th, and 75th percentiles (in Excel, there is a "percentile" function that helps you do this). Report your estimates in a well-labeled table. Let us know if you are using the big or small data.

2. Choose five commodities that are shipped in the same data you used for question 1. The variable that identifies commodities is `commodity_code` and is a numeric code. To understand what this code means, 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 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 five commodities you've chosen, make an initial hypothesis, rank ordering the commodities in their cost to ship, from largest to smallest (e.g., beer costs more to ship than cotton). Then, for each of your five 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`)

[2] If you were doing this for an econometrics research paper, you could look at the correlation between weight and revenue per car-mile. If you're feeling ambitious, you can do that here.

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

Report

- your initial hypothesis, with justification
- a well-labeled table showing the average values requested above for each of your five 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.

3. 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)
- 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 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.

4. 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 by the weight of the shipment (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 (vertical axis). Then repeat this graph, but using distance instead of weight.

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

5. Look at the variables provided in the waybills data and make a hypothesis about another dimension in the data that could influence costs. 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 the google folder: for_students → use_numbers → assignment_3

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

Please turn in a pdf (this eases our grading).

3 Data

Waybill Dataset

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

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).

These data are relatively large, so please allow time for download. You can get them from here. If these data are too large for your computer to handle, use the smaller version available here. The smaller version is a random sample of the larger one; I chose 20 percent of the observations in the large dataset.

Weight and Revenue per Car-mile

I used the waybills data above to calculate the average revenue per car-mile for different carload weights. Download here.

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. Download here.

Variable	Definition
`average_weight_per_carload`	Average weight for carloads
`average_distance`	shipment distance traveled

4 R Commands

For this assignment, here are some R commands that may be helpful.

R commands	Description
Question 1: Summary Statistics	
# load packages library(tidverse) # read the data waybills <- read_csv(file = ‘‘waybill_data_20201030.csv’’) # find a variety of summary stats for the revenue per carmile # mean, median, standard deviation, 25th, and 75th percentiles summarys <- summarize(.data = waybills, mean = mean(rev_per_carmile), median = median(rev_per_carmile), sd = sd(rev_per_carmile), p25 = quantile(rev_per_carmile, probs = 0.25), p75 = quantile(rev_per_carmile, probs = 0.75)) summarys	The first command loads the “tidyverse” packages. (You should have already installed them if you did the previous homeworks in R.) The next command reads the waybills csv data. The summarize command creates the relevant summary statistics for the full dataset. The following line prints the summary statistics to the screen.
Q2: R code finding averages across all commodity codes	
# commodities by revenue in car-miles ccsum <- group_by(.data = waybills, commodity_code) ccsum <- summarize(.data = ccsum, mean_rev_per_carmile = mean(rev_per_carmile, na.rm = TRUE), mean_short_miles = mean(est_short_line_miles, na.rm = TRUE), mean_tons_per_car = mean(actual_weight_tons/number_of_carloads, na.rm = TRUE)) head(ccsum)	Here we “group” the data by commodity code and then use summarize to find averages by commodity code. The final line prints the first rows of the output of this summarized data to the screen. In the dataset ccsum, there is one row per commodity code.

R commands	Description
Q3: Summary statistics by origin/destination pair	
od <- group_by(.data = waybills, origin_bea_area, termination_bea_area) od <- summarize(.data = od, tot_number_of_cars = sum(number_of_carloads, na.rm = TRUE), mean_rev_per_carmile = mean(rev_per_carmile, na.rm = TRUE), mean_tons_per_car = mean(actual_weight_tons/number_of_carloads, na.rm = TRUE)) head(od)	Here we group the data by the original and destination codes and find the relevant summary statistics by origin and destination pairs. The final command prints a few row of these data to the screen. The new dataframe od has one row per origin-destination pair.
Q4: Graph relationship between weight and distance and revenue per car mile	
# read the weight data waybills <- read_csv(file = ‘‘weight_vs_rev_20201029.csv’’) # then plot average weight per carload versus revenue per carmile waiter <- ggplot(data = waybill2) + geom_point(mapping = aes(x = average_weight_per_carload, y = average_rev_per_carmile)) + labs(x = "average weight per carload, tons", y = "average revenue per car-mile") # read the distance data waybills <- read_csv(file = ‘‘dist_vs_rev_20201029.csv’’) # then plot average distance versus revenue per carmile waiter <- ggplot(data = waybilld) + geom_point(mapping = aes(x = average_distance, y = average_rev_per_carmile)) + labs(x =	We load the summary data I created and then use ggplot to show these relationships.