Estimating the Demand for Train Tickets

Shunichi Araki[1], Nicholas David Gabriele[2], David Raj[3], Jonathan Simon Wagner[4]

Abstract

> *Teach a parrot the terms 'supply and demand' and you've got an economist.*
>
> –Thomas Carlyle

As the Scottish philosopher Thomas Carlyle astutely observed, the main tools of the economist – and by extension most market participants – are supply and demand. Estimating demand as a market supplier is essential for optimal resource allocation and production decisions. It can also be used for discriminatory pricing strategies to maximize producer surplus. In an environment beset with uncertainty and complicated by technology, market cycles, and competitive forces, accurately estimating demand can mean the difference between success and failure for a firm. This study aims to estimate the demand for train tickets at a specific train station at an undisclosed location. Multiple trains run through the station daily and customers can buy advance or day-of tickets. The results of this demand forecasting exercise can be used to maximize producer surplus by pricing efficiently and enhance operations by optimizing supply.

[1] *A0206507L*, [2] *A0206490J*, [3] *A0056525E*, [4] *A0152784X*

Contents

1. Executive Summary

1.1 Overview of Findings

To estimate the demand for train tickets at a particular station, a two-stage least squares (2SLS) model is selected. As will be discussed, this effectively simulates supply shocks, informing movements in price and quantity along the demand curve. With this model, the proposed demand function is:

$$\ln Q = 2.53 - 0.408 \ln P \tag{1}$$

A fundamental insight of (1) is the price elasticity of demand, which is the coefficient that corresponds to $\ln P$. A price elasticity of 0.408 indicates that the demand for train tickets is relatively price inelastic, and so a change in price results in a disproportionately small change in the quantity demanded.

The primary goal of 2SLS regression is to eliminate endogeneity bias, which arises from a number of factors such as self-selection, omitted variables, measurement error, and simultaneity. The lattermost can be addressed for the purpose of this exercise, but a considerably limited dataset as well as limited statistical power of the overall model suggests that other essential variables may be missing (i.e. omitted variable bias may still persist). The transactions in the dataset are aggregated and crucial information about ticket class, travel time/destination, and number of observations in each cell are missing. This may lead to significantly different conclusions than if a more complete dataset were available.

In constructing the 2SLS model, an instrumental variable is needed to address the endogeneity of price. The model is evaluated on the strength of hypothesis tests and goodness of fit. The following conclusions are made:

- Price is indeed an endogenous variable (i.e. would otherwise be correlated with the error term in an ordinary least squares regression).
- The inventory of train tickets sold to-date is an exogenous variable that is also statistically associated with price.
- A low adjusted R^2 value indicates that there is considerable room for improvement in the goodness of fit. This may be due to the omission of important predictor variables and can be overcome with a more complete dataset.

1.2 Business Implications

Is inelastic demand favorable to the supplier? As mentioned in Section 1.1, the quantity demanded for train tickets is relatively price inelastic. The assumption made in the analysis is that the railroad transportation industry is a natural monopoly or oligopoly with relatively few market participants. With no suitable transportation alternatives available for consumers,[1] this naturally leads to an inelastic demand curve. The findings in this report – that price elasticity is below 1.0 – confirm this hypothesis.

While price insensitivity on the part of consumers may appear to be an ideal situation for a monopolist, profit maximization is only achievable when the demand curve is elastic, according to microeconomic theory. Profits are maximized if an increase in price does not lead to a further increase in profit. In the case of railroad travel, for which demand is consistently inelastic throughout the demand curve, profit maximization is therefore prevented as the strategy of a monopolist would lead to negative marginal revenue. Consequently, marginal revenue is only positive when demand is price elastic (Taylor and Mankiw, 2017). This does not imply that profits cannot be earned, but rather that they cannot be maximized simply by raising ticket prices. While the supplier may be able to initially raise prices without a concomitant decline in quantity demanded, consumers may seek out alternatives in the long-run.[2]

Customer segmentation might present another way for profit maximization. Another way monopolists can achieve profit maximization is via price discrimination. This follows from the notion that the willingness to pay varies among consumers, and relies upon identifying distinct customer segments. Monopolists can exploit this paradigm by offering each customer segment a unique price. In this case, since the transaction data is very limited, identifying narrowly defined segments and quoting a personalized price is not possible. However, groups of customers can be identified by binning 'days to departure' (i.e. the number of days between the booking and departure dates) as a proxy for broad customer types. This is a form of third degree price discrimination, which takes advantage of the common practice that prohibits resale of a ticket, and thus different prices may be assigned to each segment. The findings show that the price elasticities for each of the different groups are significantly different from each other, although with the somewhat counterintuitive suggestion that days to departure is directly linked to price (i.e. the greater the number of days to departure, the higher the price). Hence, the railway company may not be able to maximize profits solely on account of segmenting by booking time. As mentioned previously, this finding may be masked by other subtleties that would only be discoverable in a more complete dataset. If the granularity of the dataset can be increased, this would lead to more effective discriminatory pricing strategies.

2. Assumptions

2.1 Data

During the demand analysis, the following observations are made about the available data in order to simplify the problem formulation and associated calculations:

- $NumT$ is the number of tickets purchased on a specific day for a specific train. The numbers are aggregated on a daily level and therefore do not represent individual transactions.
- $AvgP$ is the average price of a ticket for a specific train on a specific day. The average is taken over individual transactions occurring on that day.
- Dtd is the number of days remaining until the train departs, as measured from the booking date. For instance, a value of zero connotes a same-day purchase, while a value of 14 suggests that the ticket was purchased two weeks in advance.
- Inv is the cumulative number of tickets sold to-date for a specific train. This serves as a count of historical purchases and is floored at zero. If there were five transactions on the first day and three transactions on the second, $Inv = 8$.

Upon initial exploratory analysis, duplicate rows are evident. The assumption made is that this results from the removal of columnar data (e.g. ticket class, destination). For this reason, the rows are not "true" duplicates but rather subaggregated data, which in this case are treated as unique observations. The dataset of train tickets purchased is modified in two ways to make the output more meaningful:

1. The logarithm of $AvgP$ and $NumT$ is taken before proceeding with the 2SLS model. This helps with the interpretability of the coefficients, making the relationships multiplicative (i.e. percentage-based) rather than additive (i.e. unit-based).
2. Outliers containing ticket prices deemed unrealistic (greater than $4000) are removed.

2.2 Market

A general assumption is that train rides are a normal good with a downward sloping demand curve. For this reason, a linear regression is a natural class of model to consider, with logarithmic values chosen for interpretability. Moreover, based on general domain knowledge, the market is presumed to be an oligopoly with the tendency towards a natural monopoly. In many nations, the railway system and train operations are managed by the government or a government-affiliated entity, since private sector companies are often not willing to participate in an arena with high sunk costs. The initial investments and maintenance costs for ongoing operations are especially high, creating barriers to entry that give rise to a

[1] It can be argued that there are competing modes of travel (e.g. air, private car), but in many cases (e.g. commuters) there will not be a comparable substitute good.

[2] This effect could be observed with extended time series data under a scenario where the monopolist steadily increases price.

natural monopoly. In addition, the sovereignty may limit the number of participants by imposing licensing requirements. (Jensen, 1998)

3. Model Formulation

3.1 Dealing with Endogeneity

As mentioned previously, endogeneity may derive from multiple sources, but the main factor considered for this analysis is simultaneity. This arises in the field of econometrics where price and quantity are co-determined, making it difficult to distinguish movement along the demand curve from shifts in the demand curve. In the case of a linear regression model, this causes the independent variable to be correlated with the error term, ε, violating the assumption of normality. The correlated independent variable is referred to as an endogenous variable.

A common way of dealing with endogeneity bias arising from simultaneity is to introduce instrumental variables. The IVs must sufficiently explain the variation in the endogenous term while also being exogenous to the response variable of interest (i.e. uncorrelated with ε).

3.2 Variable Analysis

From the variables presented in Section 2.1, the relationships informing supply[3] and demand can be evaluated and tested as part of an empirical model:

$$Q_S(P) = \beta_{0,S} + \beta_{1,S} P_S(Dtd, Inv) + \varepsilon_S \tag{2}$$

$$Q_D(P) = \beta_{0,D} + \beta_{1,D} P_D(Dtd) + \varepsilon_D \tag{3}$$

In (2), supply is a function of price, which is itself a function of the days to departure and inventory sold to-date. The demand function (3) is assumed to be a function of price, which itself depends on the days to departure. From Section 3.1, it can be inferred that simultaneity is present and a reduced form equation of the endogenous variable (price) must be created. It is also noted that *Inv* is only present in the price function for supply, which speaks to the fact that the information available to producers and consumers is asymmetric (i.e. consumers do not know the inventory sold to-date and it does not affect their behavior). Shifts in the supply function can therefore be simulated to approximate points along the demand curve, making *Inv* the instrumental variable.

$$\ln(NumT) = \beta_0 + \beta_1 \ln(AvgP) + \varepsilon \tag{4}$$

$$\ln(AvgP) = \gamma_0 + \gamma_1 Inv + \nu \tag{5}$$

(4) and (5) represent the structural equation and reduced form of the 2SLS, respectively. The Appendix shows that the association of *Inv* and *AvgP* is indeed meaningful, so the usage of *Inv* as an IV makes sense and γ_1 is shown to be significant.

3.3 Demand Function using 2SLS

As the name suggests, the 2SLS model consists of two steps. The first is to regress the exogenous variable (*Inv*) on the endogenous variable (*AvgP*), typically employing an ordinary least squares (OLS) approach. The predicted values for the endogenous response variable ($\widehat{AvgP}$) are then regressed on the original response variable (*NumT*). In doing so, the exogenous variable should be uncorrelated with the structural equation error term (ε), but sufficiently explain the variation in the endogenous variable (*AvgP*). Using the above model structure, the coefficients using OLS are determined as follows:

$$\ln\widehat{(NumT)} = 2.53 - 0.408 \ln\widehat{(AvgP)} \tag{6}$$

$$\ln\widehat{(AvgP)} = 5.35 - 0.0323 Inv \tag{7}$$

The interpretation of the coefficients in (6) and (7) is straightforward: a unit change *Inv* produces an average 3.2% reduction in *AvgP*, while a percentage unit change in *AvgP* produces an average 0.408% reduction in *NumT*. The demand curve is presented in Figure 1.

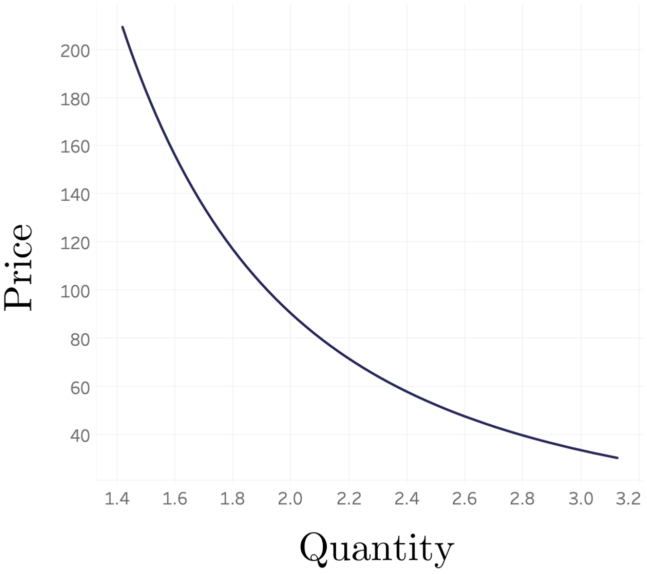

Figure 1. Estimated Demand Curve

Validity of Results

In order to ensure that the modeled results are statistically significant and theoretically sound, several tests can be performed. The first examines whether γ_1 in the reduced form of (5) is statistically significant; that is, whether the exogenous variable *Inv* has a direct association with *AvgP*. The p-value of *Inv* is highly statistically significant (less than 0.001), which leads to the conclusion that *Inv* is indeed a relevant instrument to simulate shifts in the supply of train tickets and therefore estimate the demand curve. This is equivalent to the so-called ‘weak instruments’ test when there is only one instrumental variable involved.

[3] Examining the supply function is important to ensure that the IV is truly exogenous.

The second test, known as the Hausman test, considers whether the purported endogenous variable is truly correlated with the original error term; that is, whether $Cov[AvgP, \varepsilon] = 0$. The associated p-value is highly statistically significant (less than 0.001), confirming the presence of endogeneity in the structural equation.

While the results of the hypothesis tests are statistically significant, the adjusted R^2 value of 4.5% in the reduced form equation suggests a poor fit. That is, *Inv* only explains 4.5% of the variation in *AvgP*. This could be improved upon as part of a future study if the original dataset were expanded in scope (i.e. included other attributes such as ticket class).

3.4 Segmented Demand Function

In Section 3.3, the demand function is estimated using *Inv* as the instrumental variable and the value of β_1 = -0.408 indicates that the consumer demand for train tickets is price inelastic. Although *Dtd* is not used in this model form, it may still have relevance as a segmentation variable. This section aims to investigate whether price elasticity significantly differs for specified ranges of days to departure. The results from the first-stage OLS are classified into three statistically credible[4] bins based on *Dtd*, as shown in Table 1. The hypotheses to be tested using 2SLS are as follows:

Table 1. *Dtd* Segmentation

Description	Bin Range	Observations
Short-term bookings	0-5	1,571
Mid-term bookings	6-40	2,371
Long-term bookings	41+	2,578

$$H_0 : \text{The price elasticities are identical} \quad (8)$$

$$H_1 : \text{The price elasticities are not identical} \quad (9)$$

The above formulation contends that consumers making on-demand bookings (i.e. $Dtd \leq 5$) are potentially less price sensitive than consumers who book tickets far in advance. This can be argued by the fact that consumers making day-of bookings have fewer alternatives available and are less likely to negotiate price. The test is conducted with confidence level $\alpha = 0.05$. The regression equation for each bin can be written as:

$$\ln(NumT_j) = \beta_{0,j} + \beta_{1,j} AvgP_j + \varepsilon_j \quad (10)$$

Note that there are $j = 3$ separate regression models, each representing a separate demand function for each market segment as shown in (11), (12), and (13).

$$\ln(\widehat{NumT_1}) = 4.63 - 0.843 \ln(\widehat{AvgP_1}) \quad (11)$$

$$\ln(\widehat{NumT_2}) = 2.63 - 0.448 \ln(\widehat{AvgP_2}) \quad (12)$$

$$\ln(\widehat{NumT_3}) = 2.93 - 0.460 \ln(\widehat{AvgP_3}) \quad (13)$$

Table 2. Analysis of Variance

	DF	SS	MS	F-Value	p-Value
Bins	2	119	59.42	41.63	< 0.001
Residuals	6,516	9,299	1.43		

ANOVA Results

The F-test performed in Table 2 indicates that the price elasticities of demand are indeed significantly different from one another. The associated p-value is highly statistically significant (less than 0.001), so H_0 can be rejected. However, the category with the highest price elasticity is for consumers who booked zero days in advance, a somewhat counterintuitive result since the argument is that these consumers have fewer alternatives available. As discussed in Section 1.2, this may arise from a lurking variable not present in the original dataset. The distribution of ticket prices can be visualized across customer segments as shown in Figure 2.

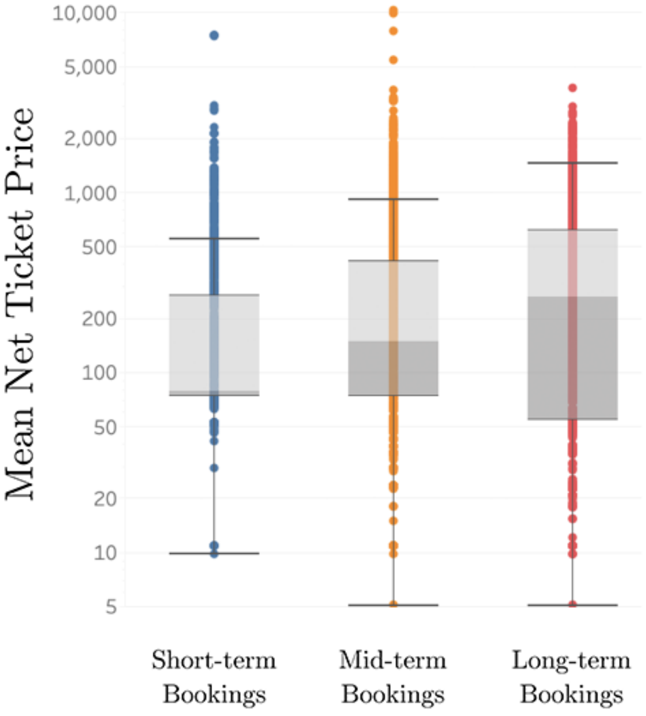

Figure 2. Customer Segments

References

Taylor, M. and Mankiw, N. (2017). *Microeconomics.* Andover: Cengage Learning.

Jensen A. (1998), *Competition in railway monopolies*, Transportation Research Part E: Logistics and Transportation Review, vol. 34, no. 4, pp. 267-287.

[4] According to the definition of Bühlmann credibility.

Appendix

Table 3. Summary of Model Results

Parameters		**Main Model**			**Short-Term Model**		**Mid-Term Model**		**Long-Term Model**	
		Ordinary Least Squares	Reduced Form	Structural Equation	Reduced Form	Structural Equation	Reduced Form	Structural Equation	Reduced Form	Structural Equation
Intercept	β	0.734	5.345	2.531	5.187	4.625	5.408	2.634	5.369	2.926
	SE_β	0.031	0.019	0.176	0.031	0.445	0.032	0.277	0.031	0.440
	p	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
$ln(AvgP)$	β	−0.059		−0.408		−0.843		−0.448		−0.460
	SE_β	0.006		0.034		0.090		0.054		0.084
	p	<0.001		<0.001		<0.001		<0.001		<0.001
Inv	β		−0.032		−0.025		−0.033		−0.044	
	SE_β		0.002		0.002		0.003		0.006	
	p		<0.001		<0.001		<0.001		<0.001	
Adjusted R^2		0.0152	0.0449	−0.5296	0.0752	−0.9961	0.0503	−0.8213	0.0191	−0.9213
Weak Instruments				<0.001		<0.001		<0.001		<0.001
Hausman Test				<0.001		<0.001		<0.001		<0.001