Estimating Demand Function for Train Travel

Team Members (Group 26)

1. Erica Chi Yi Tung (Student ID: A0318639R)
2. Hrithik Kannan Krishnan (Student ID: A0318899B)
3. Om Sanjay Gorakhia (Student ID: A0318038E)
4. Priyanshi Verma (Student ID: A0318672X)
5. Sai Ashwin Kumar Chandramouli (Student ID: A0329294U)

Objective

The purpose of this project is to empirically estimate the demand function for train travel by analyzing detailed ticket sales data collected from a specific train station. This estimation aims to provide insights into consumer behavior, price responsiveness, and travel patterns. To ensure methodological rigor, the analysis employs econometric modeling techniques that account for endogeneity, data structure, and appropriate functional forms in demand estimation.

Data Preparation and Features

We conducted a series data preprocessing and feature engineering procedures to ensure the dataset was suitable for analysis:

Engineered Features/Variables	
estimate_model1	Created to signify our naïve model (OLS).
days_to_departure	Calculated as the difference between 'Dept_Date' and 'Purchase_Date', capturing the booking window.
log_seats	Natural logarithm of 'num_seats_total', applied to stabilize variance and reduce skewness in demand per transaction.
log_price	Natural logarithm of 'mean_net_ticket_price', facilitating elasticity estimation within a log–log demand framework.
trip_type	A categorical variable with three levels: **Oneway**, **Return**, and **Neither**, derived from the 'isOneway' and 'isReturn' indicators. This captures systematic differences in demand patterns across trip types.
Cumulative_sales	Running total of sales for each train departure, proxy for demand pressure.
Ym (used as c(ym))	Month label YYYY-MM from departure/purchase date; absorbs seasonality/holiday/macroeconomic shocks.
Train_Number_All (Train_FE)	Train/service ID; controls for time-invariant route/service traits

Data Cleaning

- Dropped transactions with non-positive prices and negative 'days_to_departure'.
- Winsorized extreme outliers in price and seats.
- Ensured currency/unit consistency for prices before modeling

Literature & Conceptual Framework

In line with demand estimation frameworks commonly applied in transportation economics, our model leverages transaction-level data to capture the price-demand relationship while accounting for heterogeneity in booking behavior, trip type, and customer segments.

Key Variables

- **Outcome (Q)**: 'log_seats' → elasticity-friendly measure of demand.
- **Price (P)**: 'log_price' → allows direct elasticity interpretation.
- **Key Controls**: 'days_to_departure', 'trip_type', 'isNormCabin', 'Customer_Cat', month-year fixed effects, and train fixed effects.

This specification allows us to test the **law of demand** while controlling for temporal, behavioral, and route-specific effects.

Exploratory Demand Analysis (EDA)

Our EDA revealed several key insights into booking behavior and market structure:

- **Booking Window**: Ticket prices tend to rise sharply as the departure date approaches, suggesting that customers booking closer to departure are less price sensitive. This supports the existence of a "last-minute premium," while the significant number of early bookings indicates strong demand for discounted advance-purchase fares (Refer to Fig. A1 titled, "Mean Ticket Price by Booking Window").
- **Bimodal Price Distribution**: Cabin class and demand. Average seats per transaction are higher in Normal cabins than Special cabins, indicating greater group-booking intensity in Normal class (Refer to Fig. A2 in Appendix; Graph titled, "Average Seats by Cabin Type).
- **Customer Segmentation**: Customer A seems to have more diverse booking behavior with a slight preference for more premium traveling options (Special Cabin). They tend to book fewer seats than Customer B with less frequency but can be categorized as a higher-value traveler. Customer B has a higher average in respect to seat booking with stronger preferences for normal cabins (price sensitive) and return trips, indicating that they may represent more routine or business-oriented travel patterns (Refer to Fig. A2 in Appendix for more detailed comparisons).

Econometric Methodology

Our empirical strategy is designed to estimate the causal effect of key independent variables, such as price, on the dependent variable of interest, quantity demanded, while systematically addressing potential concerns related to endogeneity, unobserved heterogeneity, and non-linearities. The analysis adopts a sequential modeling framework, beginning with baseline specifications and progressively incorporating refinements to enhance robustness and interpretability.

The analysis begins using a benchmark model estimated via Ordinary Least Squares (OLS) regression to establish a baseline relationship between the variables. While OLS provides an initial assessment of correlations, it is susceptible to omitted variable bias and measurement error, which may produce inconsistent estimates. To account for potential non-linearities and elasticity-based interpretations, we subsequently implement a log-log transformation, enhancing both interpretability and model fit.

To mitigate potential endogeneity concerns, we employed an **Instrumental Variables (IV) approach** estimated through **Two-Stage Least Squares (2SLS)**. This approach helps correct for biases that may arise from simultaneity between price and demand, yielding more consistent estimates of the demand function. Our methodology follows these steps:

- **First Stage:** Regress the endogenous regressor on chosen instruments and control variables to isolate the exogenous variation.
- **Second Stage:** Use the predicted values from the first stage in the outcome regression to obtain consistent estimates of causal effects.
- We then validate our instrument variable's relevance by checking our first stage **F-statistics** and the **Wu-Hausman test** for endogeneity**.**

First-stage with Year–Month Fixed Effects (YM-FE)

We estimate the price equation using the leave-one-out train×month mean price as an instrument and include Year–Month fixed effects to absorb common shocks (seasonality, holidays, macro). The plot shows actual vs. first stage predicted prices with a strong fit (corr ≈ 0.79 and r-squared ≈ 0.67), indicating a highly relevant instrument after controlling for month effects, which strengthen the causal identification used in the second stage.

Heterogeneity & elasticities: Downstream, we translate the second-stage coefficients into causal price elasticities and can examine subgroup differences (e.g., by cabin or customer segment) via interactions.
Count outcome: Because demand is a count (seats), we also estimate a Poisson IV (2SRI) specification and benchmark models using calibration plots and fit metrics (RMSE/MAE and pseudo-R^2) (Refer to Fig. A3 in Appendix).

Calibration by predicted decile (Poisson 2SRI + YM-FE)
To assess how well the model's score orders demand, we group transactions into ten bins by predicted seats and compare the mean actual vs mean predicted in each bin. The curve rises steadily from D1 to D10, showing strong rank ordering (higher scores → higher realized demand). Calibration is tight in the middle deciles, with slight under-prediction around D6–D7 and mild over-prediction at the very top; the shaded bars indicate balanced sample sizes across deciles, so these patterns aren't driven by tiny cells. Practically, this means the model is dependable for prioritizing inventory/pricing toward top deciles and suggests a modest intercept/scale tweak if perfect calibration is needed at the extremes (Refer to Fig. A4 in Appendix).

Results and Managerial Insights

Our empirical analysis reveals several key findings with both statistical and managerial significance:

1. **Baseline and Elasticities:** The OLS and log-log specifications indicate a negative relationship between price and quantity demanded, consistent with economic theory. Price is highly statistically significant and has an estimated coefficient of –0.0005. For every 1 unit increase in ticket price, the expected ticket sales decrease by 0.0005 seats, holding all else constant; this indicates that there is a tiny marginal effect which means that price is not a strong driver of sales according to the OLS model.

2. **Causal Effects and Endogeneity Correction:** The 2SLS estimates, correcting for potential endogeneity of price, confirm a stronger causal effect compared to OLS, implying that naive regressions may underestimate the true impact. This emphasizes the importance of using instruments when pricing strategies are influenced by unobserved factors like regional competition or demand shocks.

3. **Heterogeneity Across Segments:** The analysis reveals clear heterogeneity in price responsiveness across customer segments. This differentiation carries important implications for pricing strategy as implementations such as dynamic pricing, targeted promotions, and discounts can effectively stimulate demand. This segmentation-based approach allows firms to balance profitability and market penetration.

4. **Fixed Effects Insights:** The elasticity derived using the 2SLS model with Year-Month fixed effects is –1.26, which indicates that the OLS model underestimated the true price sensitivity. By controlling for Year-Month fixed effects, the model isolates exogenous variation in price. The elasticity indicates that demand is elastic and that a 1% increase in price is about a 1.26% decrease in seats sold.

5. **Poisson Model Outcomes:** The Poisson-IV (2SRI) model estimates a price elasticity demand of –0.98. This indicates that price is highly responsive to price changes and that the instrument explains a substantial share of the variation in ticket prices. The relevance of the instrument is confirmed by the first stage r-squared, giving us a value of 0.67. The predictive decile analysis also confirms reasonable calibration across demand levels, highlighting the importance of price sensitivity (Refer to Fig. A4 and A5 in Appendix).

Managerial Implications
Pricing Strategy: Managers should consider segment-specific price elasticity when designing promotions or dynamic pricing. Overpricing sensitive segments could lead to disproportionate demand loss, while stable pricing for less sensitive segments can enhance revenue.
Targeted Marketing: The heterogeneity analysis suggests customizing campaigns to the most responsive segments, optimizing marketing ROI.

Inventory and Planning: Understanding time-based demand fluctuations allows better resource allocation, reducing stockouts or overstocking.
Data-Driven Decisions: Endogeneity-corrected causal estimates provide confidence in making strategic decisions rather than relying solely on historical correlations.

These results collectively form the basis of our proposed **demand function**, which integrates both causal effects and segment-specific insights, providing a practical roadmap for evidence-based managerial decision-making.

Demand Function

Interpretation (Refer to Fig. A5 and A6 in Appendix)

The blue dots represent the actual observed values of seats demanded at different price levels, while the red line represents the predicted demand curve derived from the Poisson IV (2SRI) model. It illustrates a steep decline in demand as ticket prices increase. At low ticket prices (<$200), demand is relatively high with some observed cases exceeding 60 seats. Ticket prices greater than $1,000 render demand negligible, consistent with economic theory and expectations. This framework is appropriate for modeling seat demand using count data which addresses potential endogeneity in ticket prices.

The Poisson IV curve captures the general inverse relationship between price and demand; however, actual demand (represented by the blue dots) show substantial variation around the fitted red line, indicating there are other factors that influence demand (trip type, customer category, timing of purchase, etc.). The sharp contraction in demand at the lower end of the price spectrum reflects significant price sensitivity, whereby incremental increases in ticket prices generate substantial decreases in demand. As prices rise further, the demand curve flattens, indicating to us that remaining consumers are less price sensitive. The accumulation of sales at lower price levels suggests that purchases are largely concentrated in the affordable range with relatively few extreme values, but the tradeoff is that demand is virtually nonexistent at such elevated prices. Demand for train seats is elastic at lower ticket prices but becomes highly inelastic as prices increase beyond a certain range.

Limitations and Reproducibility

Limitations

The analysis does not account for external shocks such as weather events or strikes. The validity of our instruments depends on exclusion restrictions, which cannot be directly tested. The dataset captures only observed transactions, excluding unsold capacity. Although count models better fit the variance structure, they may still understate heterogeneity across observations.

Reproducibility

All feature engineering and regression analyses were implemented in Python using pandas, stats models, and linear models. Plots and tables in the appendix are generated directly from these scripts. The full analysis is reproducible with access to the raw dataset and the accompanying Jupyter notebooks.

Appendix: Tables and Graphs

Fig. A1 — Mean Ticket Price by Booking Window

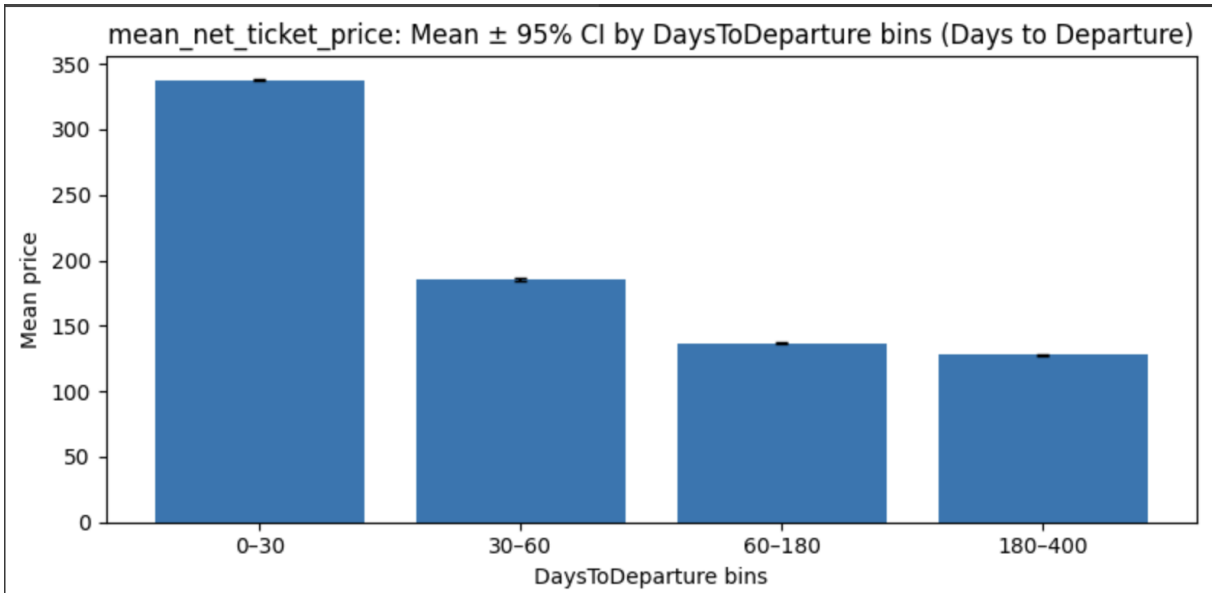

Fig. A2 — Customer Segmentation Analysis

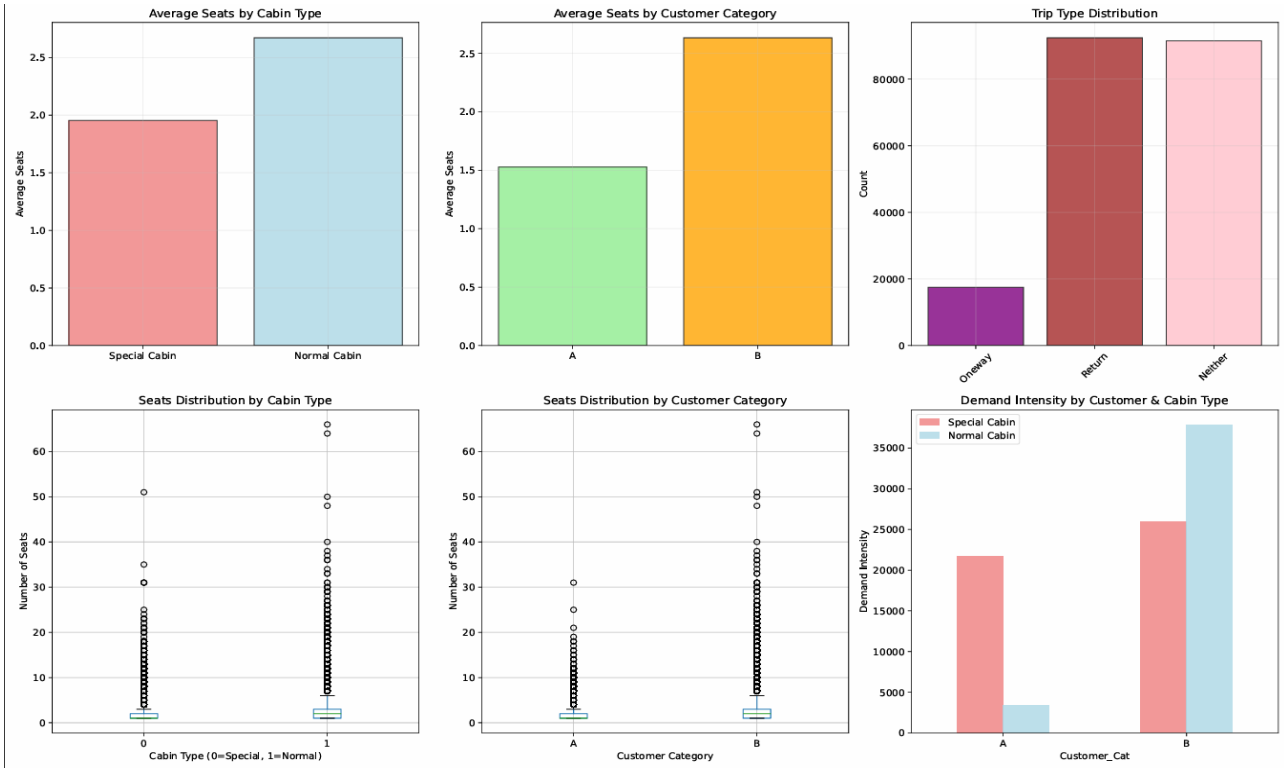

Fig. A3 — Instrumented Price: Actual vs Predicted (YM-FE)

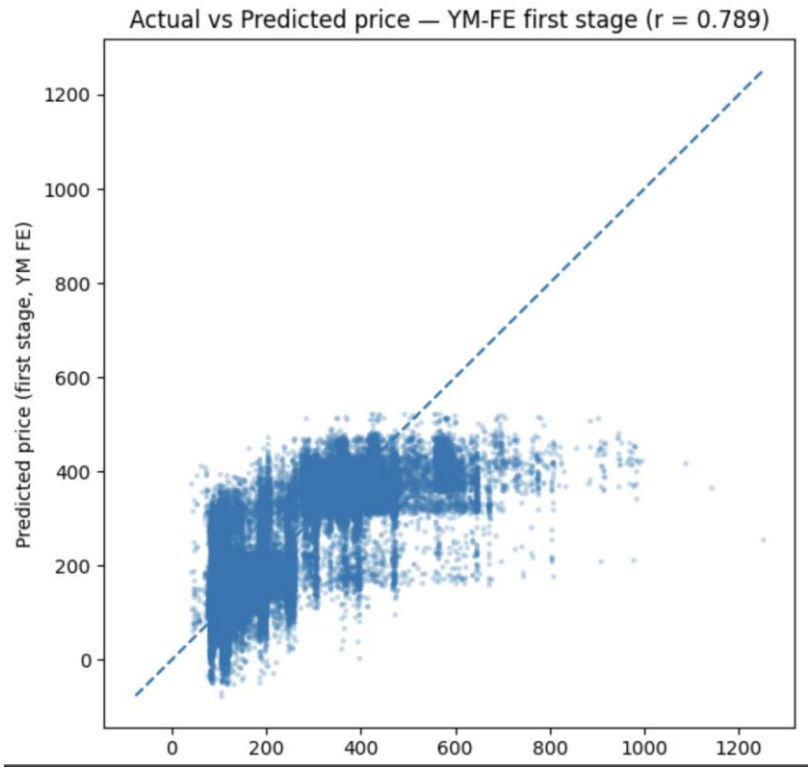

Fig. A4: Calibration: mean Actual vs Predicted by decile (preferred spec)

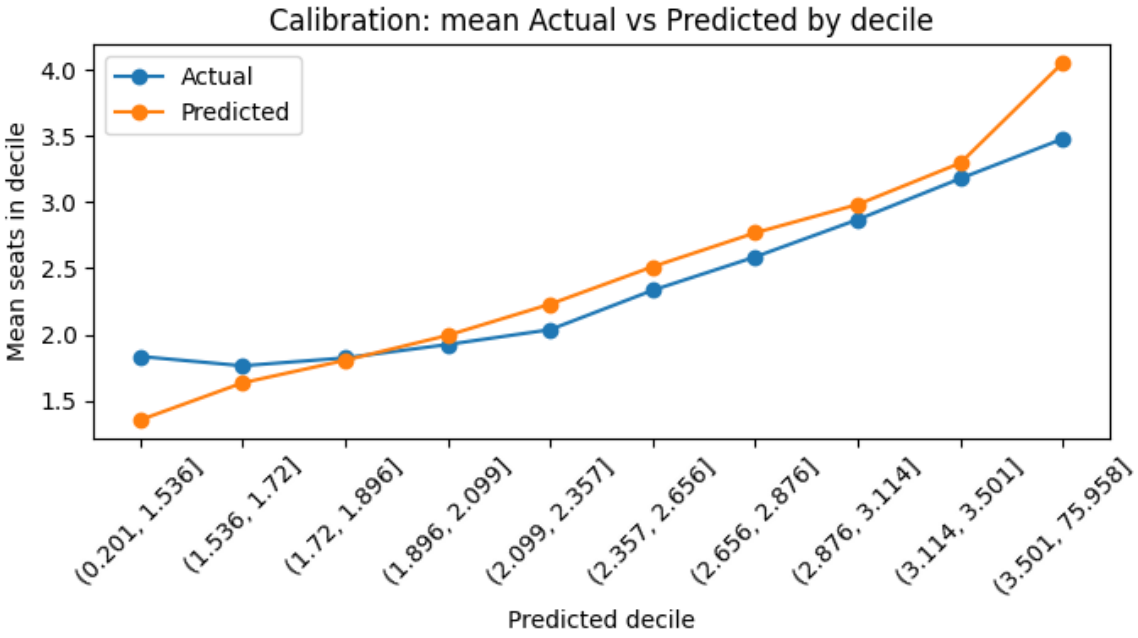

Fig. A5: Demand Curve derived from Poisson IV (2SRI) Model

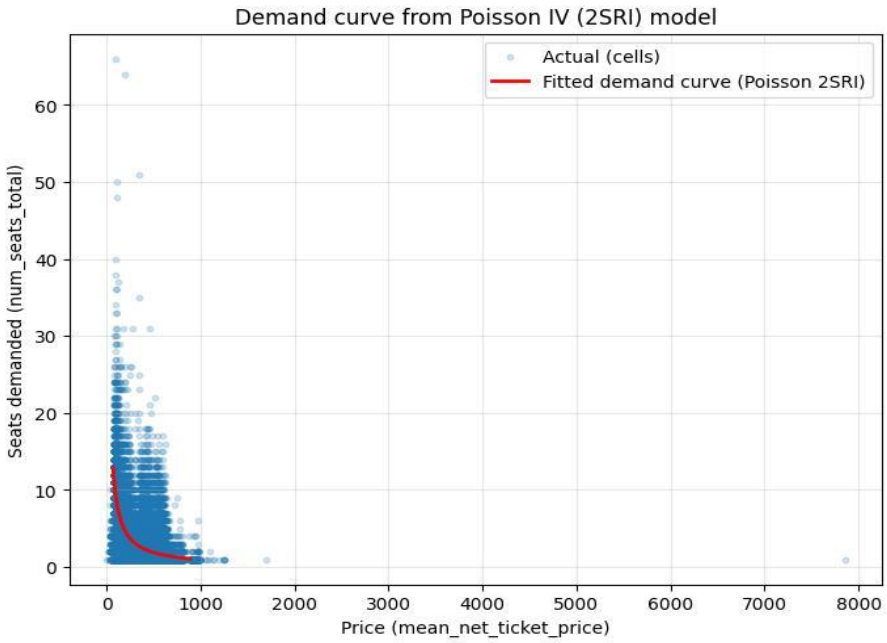

Fig. A6: Summary of all model implementations & results (Generated through the help of ChatGPT)

#	Model (method)	Outcome	Price regressor	Price effect †	SE	95% CI	p-value	Fit metric	N	Clusters	IV
A	OLS (levels)	num_seats_total	mean_net_ticket_price	**−0.0005**	0.000047	[−0.0006, −0.0004]	<0.001	R^2 = 0.085	209,697	—	—
B	OLS (log-log)	log_seats	log_price	**−0.1807**	0.0040	[−0.188, −0.173]	<0.001	R^2 = 0.127	209,697	—	—
C	IV-2SLS (levels)	num_seats_total	mean_net_ticket_price	**−0.0049**	0.0017	[−0.0081, −0.0016]	0.003	R^2 = 0.053	209,690	14	iv_price_loo
D	IV-2SLS (log-log, pooled)	lnQ	log_price	**−0.0095**	0.0059	[−0.0211, 0.0022]	0.111	R^2 = 0.547	209,690	14	lnIVX
E	IV-2SLS (log-log, ym×cabin FE)	lnQ	log_price	**−0.0453**	0.0068	[−0.0587, −0.0319]	<0.001	R^2 = 0.557	209,697	331	lnIV_cabin
F	IV-2SLS (log-log, aggregated market)	lnQ	log_price	**−1.1421**	0.1902	[−1.5149, −0.7692]	<0.001	R^2 = 0.524	15,076	331	lnIV
G	IV-2SLS (log-log, aggregated market, Normal cabin only)	lnQ	log_price	**−1.7471**	0.3283	[−2.3906, −1.1036]	<0.001	R^2 = 0.676	7,788	326	lnIV
H	Poisson IV 2SRI (log-link)	num_seats_total	log_price	**−0.9756**	0.250	[−1.466, −0.485]	<0.001	Pseudo-R^2 = 0.052	≈209,690	Train	lnIV (CF)