FORECASTING TRAIN TRAVEL DEMAND

…a causal treatment of demand forecasting using Two Stage Least Squares Linear Regression (2SLS).

NUS MSBA
DBA 5101 - PROJECT 1

- ANKIT MALHOTRA (A0232322X)
- GINO MARTELLI SY TIU (A0231956Y)
- RACHEL SNG WEI LIN (A0231921N)
- TEERAWAT CHAITEERATH (A0231931M)
- WING KEI TRACY NG (A0231880H)

Table of Contents

1. Problem Statement

Using ticket sales information collated from a particular station, we are tasked with forecasting travel demand using two stage linear regression. The goal is to build the best possible causal model that can be used in BAU settings while minimizing issues caused by endogeneity.

In brief, the focused objective is to "*predict that Y number of tickets will be sold at a certain price, for a certain customer type, a certain train relative to specific month and weekday of departure.*"

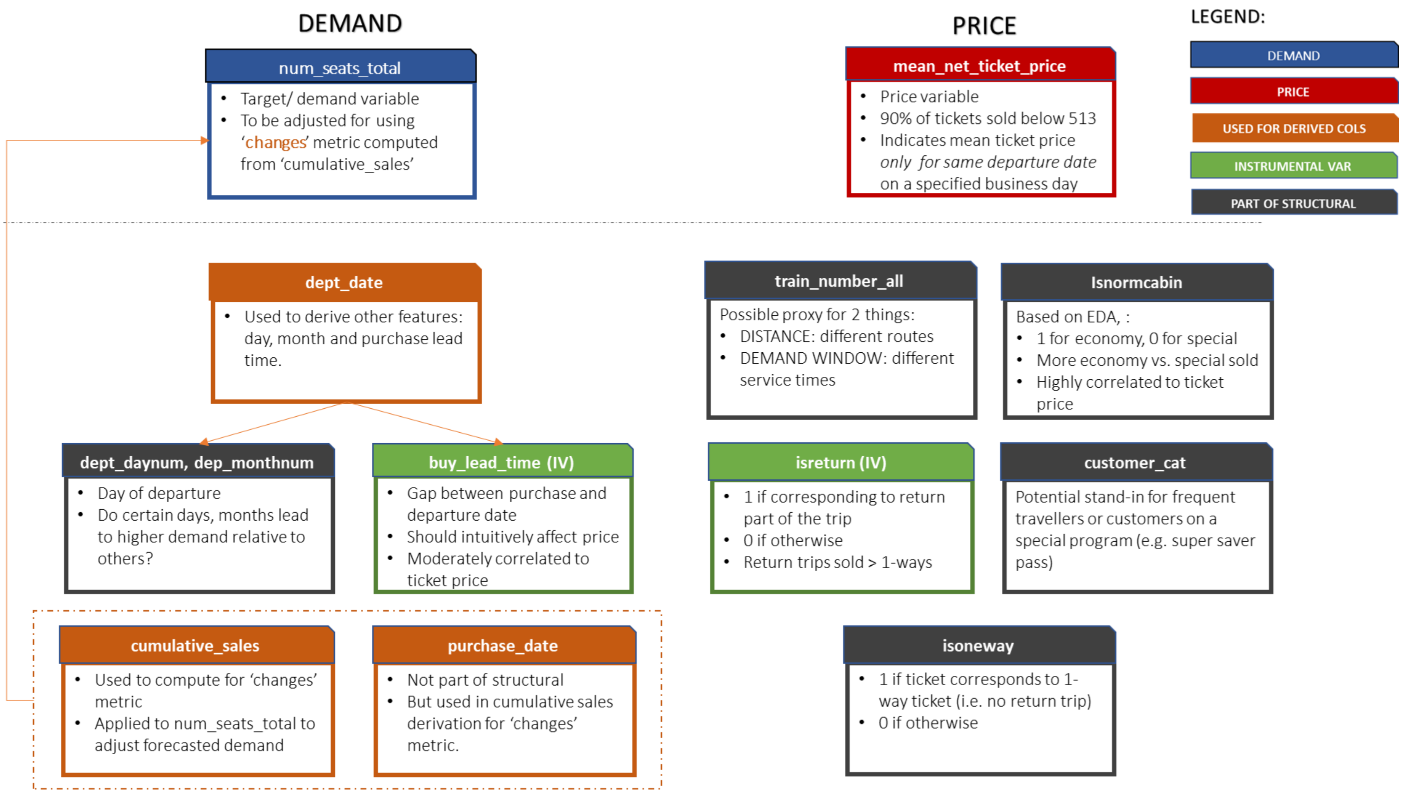
2. Exploratory Data Analysis and Assumptions

EDA and subsequent analysis were carried out to: (a) thresh out assumptions relative to the original dataset and (b) create derived features that may prove useful for the 2SLS implementation. The process undertaken was as follows:

2.1 Initial Data Dimension and Features

2.2 Feature Relationship Analysis

The below diagram elaborates on the relationship between existing variables and newly derived features. Observations based on group-by, pair plot and correlation analysis are included as variable box comments.



2.2.1 Elaborated Variable Assumptions:

- **Train number** is considered a proxy for different service routes and hence has different distances from the train station being studied. An alternative view is trains might service same routes but during different times of the day (peak, non-peak). Train O was dropped since there was only 1 instance of this across the entire dataset.

- **Derived date features** weekday and month are expected to differentiate peak from non-peak periods. For instance, we expect demand to be higher on weekends and months with more holiday travel (e.g., December). Intuitively, '*buy_lead_time'* is also expected to be a significant price influencer.

2.2.2 Commentary on Data Handling:

- **Cumulative Sales Issue**: Analysis shows this is the aggregated seats for the *same train, departure date, customer category and cabin type*. Processed with *num_seats_total*, it is possible to compute changes post-purchase. However, it is impossible to attribute changes back to a specific price/date and cannot be directly used in the demand model.

- **Overstated 2SLS Estimation**: Due to the above limitation, the demand estimate is expected to be overstated. A cancellation factor will be computed and applied to the model estimate to correct this (see section 4b: impact of post-purchase changes).

2.3 Outlier Identification & Management

Significant outliers were noted for *mean_net_ticket_price*, *num_seats_total*, *cumulative_sales* and *buy_lead_time*. As opposed to dropping these rows, outlier capping was used as a strategy to address two goals:

1. Balance the study of causality while making it more generalizable for use

2. Account for *potential measurement and input errors* – e.g., adding an additional 0 – as we expect train ticket prices to be within a certain range. Too high a price would push the consumer to explore other travel modes such as traveling by car or plane.

The upper cap was computed as the 95th percentile and the lower cap was computed at the 10th percentile[1]. Note that the upper cap value is approximately $513 – a more feasible figure vis-à-vis some of the outlier previously identified.

3. Demand Modelling

3a. Structural Equation

Demand for train travel in terms of number of seats sold can be modelled linearly as follows:

$$\textit{num\_seats\_total (seats sold)} = \beta_0 + \beta_1\, \textit{mean\_net\_ticket\_price (price)} + \beta_2\, \textit{isnormcabin} + \beta_3\, \textit{isoneway} + \beta_4\, \textit{train\_number} + \beta_5\, \textit{customer\_cat} + \beta_6\, \textit{dept\_daynum} + \beta_7\, \textit{dep\_monthnum} + u_1\, \textit{(error)}$$

A key assumption of to apply OLS is that the independent variables are exogenous (E(u|X)=0).

However, ***price* is suspected to be endogenous** and correlated with ε since <u>seats sold and price are jointly determined</u>. To elaborate, shocks to *seats sold* may cause *price* to change **(simultaneity).** An example is the case where strong sales cause the train companies to raise prices on remaining seats or vice versa**.** Hence, OLS will not yield an unbiased, consistent estimator (β_1*_hat*) for *price*.

The other independent variables are exogenous, as they are clearly not caused by seats sold – i.e., "an additional unit of *seats sold* cannot change the cabin type, train number, customer type, day of week, month or whether the trip is a round trip".

3b. Reduced Form

To address the endogeneity of price, two instrumental variables (IV) *isreturn* and *buy_lead_time* are introduced. These are reasoned to be **uncorrelated to model error u_1** since the number of seats purchased does not change depending on how far in advance the purchase decision is made or whether the trip is a return one. However, these can be highly influential on price as train companies may give price discounts for earlier purchases or return trips.

---

[1] Refer to code section 3c: Outlier analysis and treatment

The reduced form is thus represented by:

$$\textbf{mean\_net\_ticket\_price (price)} = \pi_0 + \pi_1 \textit{isnormcabin} + \pi_2 \textit{isoneway} + \pi_3 \textit{train\_number} + \pi_4 \textit{customer\_cat} + \pi_5 \textit{dept\_daynum} + \pi_6 \textit{dep\_monthnum} + \boldsymbol{\pi_7} \textbf{\textit{isreturn}} + \boldsymbol{\pi_8} \textbf{\textit{buy\_lead\_time}} + v_1 \textit{(error)}$$

Weak Instrument Test Results:

- F-statistic is large (4.266e+05), with p-value of 0.
- H0: Chosen IVs are weak, slope coefficients are 0 can be **rejected**.
- Hence, we may proceed with using these two IVs as they are correlated with price.

Parameters for both IVs conform to expectations where increase in buy_lead_time and is_return = 1 leads to lower price. In addition, all other independent variables produce results in line with expectations (i.e. cheaper for a normal cabin, more expensive for one-way trip, mixed price premiums for different trains, reduced price paid for customer B, more expensive for peak Friday – Sunday travel and more expensive for year-end November to December travel)[2].

3c. Hausman Test to Verify Endogeneity of Price

To confirm that ***price*** is endogenous, we add in the residuals (v_1*_hat*) from the reduced form model to the structural model and re-run OLS to check the significance of the coefficient of the residual.

Hausman Test Result:

- F-statistic is large (1239.532), with p-value of 0[3].
- H0: All endogenous variables are exogenous can be **rejected** as the coefficient of the error term v_1*_hat* is statistically different from 0.
- Hence, *price* has been correctly identified to be endogenous.

3d. Structural Equation OLS and 2SLS Comparison

We now run 2SLS in the structural equation using values of *price** predicted using the reduced form.

$$\textbf{num\_seats\_total (seats sold)} = \beta_0 + \boldsymbol{\beta_1} \textbf{\textit{mean\_net\_ticket\_price* (price*)}} + \beta_2 \textit{isnormcabin} + \beta_3 \textit{isoneway} + \beta_4 \textit{train\_number} + \beta_5 \textit{customer\_cat} + \beta_6 \textit{dept\_daynum} + \beta_7 \textit{dep\_monthnum} + \varepsilon \textit{ (error)}$$

2SLS Result:

- Coefficient of *price** has materially changed from **–0.0014 to –0.0043** in the 2SLS model.
- This suggests that endogeneity caused the initial β_1*_hat* to be biased and underestimate the impact of *price* on *seats sold*. 2SLS corrects this impact.

3e. Sargan Test for Overidentifying Restrictions

Given there are two IVs, a redundancy check is undertaken. If both IVs are truly exogenous, they should be uncorrelated with the 2SLS residuals. The residuals from 2SLS model are hence regressed against the IVs and test statistic $nR^2 \sim \chi^2_q$ is checked.

Sargan Test Result:

- p-value of the r-square test statistic is 0.899[4].
- H0: All instruments are exogenous cannot be rejected.
- Hence, all instruments are exogenous to the structural model and can be included.

[2] For full results, refer to code section 5b: Reduced Form Model for Price.
[3] For full results, refer to code section 5d: Hausman Test for endogeneity.
[4] For full results, refer to code section 5e: Sargan Test for overidentification.

4. Final Demand Model

4a. Overall Model

The final parameters of the model are summarized below[5]. All coefficients were significant, with p-value of 0.000. These results will enable the train company to better prepare the supply of train capacity to match demand.

Variable	**Parameter (*β*)**	**Interpretation**
isnormcabin	-0.4917	Special cabins are more in demand than normal cabins.
isoneway	-0.0733	One-way trips are less in demand than roundtrips.
train_number	From -0.2883 to +0.4355	Train B has the highest demand, Train N has the lowest. This could mean Train B operates the most popular routes or operates at peak hours.
customer_cat	+0.4722	Customer B purchases more seats than Customer A[6].
dept_daynum	From –0.0826 to –0.0000	Monday is the most popular days for travel while Tuesdays were the least popular, is in line with the travel trends where commuters may surge back into cities for work on Mondays.
dep_monthnum	From –0.1492 to +0.6613	December is the most popular month for train travel and July is the least. This may be due commuters on year-end holidays.
mean_net_ticket_price	-0.0043	Each additional $ increase would lower the number of seats demanded as consumers are less willing to pay and may find other modes of transport.
constant (β_0)	2.8779	Demand for all $x_i = 0$ (i.e. in a normal cabin, round trip, on train A on Monday in January).

4b. Impact of Post-Purchase Changes

For the train company to accurately plan supply of trains, it is important to factor in the average seat changes after ticket sales. These free ticket changes or cancellations can be derived from the running total *culmulative_sales*[7]. Analysis of this data shows that various trains have different change rates from **+11.8% (Train G) to –14.8% (Train L)**[8] that should be factored into capacity planning. The difference by train could be due to factors such as availability of alternative modes of transport for the route or timing serviced by each train type (depending on what the train numbers are proxies for as per section 2.2).

4c. Final Notes

Overall, we remain cognizant of several shortcomings of the model, as indicated by low R^2 (0.0918) and distinctly linear residuals of the final model. This strongly suggests that there are other independent variables that we do not have information on. Various other sources of information that would be useful to build a more robust and accurate model on demand include specific destinations, availability of alternative transportation routes (e.g. taxi, bus, flights), and time of departure among other factors.

[5] Refer to code section 6: Final 2SLS Demand Model

[6] Relatedly, we note that Customer B category is also negative correlated with price (parameter estimate of –42.26 in the reduced form) which may indicate this variable perhaps represents a 'saver' pass of sorts with lower prices and high volume of seats purchased as a result.

[7] To isolate cancellations/transfers, we compute a column of *gross_seats_sold* using a running total for each unique ticket type and compare it against *culmulative_sales* on the last recorded day of sales for that ticket to get the average percentage increase/decrease of seats for a particular train over the entire data set.

[8] For full results, refer to code section 3b: Analysis of post-purchase changes.