1. Executive Summary

As one of the oldest, yet persistent transportation modes, trains remain as one of the go-to public transports adopted worldwide. With established railroads, trains connect large areas and are economical transportation modes. In this project, we will estimate the demand function for trains using train ticket sales data at a particular train station.

The problem is complex by nature as the demand function can never be estimated without complication due to simultaneity of the supply and demand functions. As such, we approached the problem of estimation by adopting the Two-Stage Least Squares (2SLS) regression model. The 2SLS model allows us to eliminate endogeneity bias which would be prevalent in the Ordinary Least Squares (OLS) model of a demand function.

Based on our analysis, the train's 2SLS demand function is:

$$ln(seats) = 1.8183 - 0.2431 * ln(\widehat{price}) + 0.0012 * \ days\_in\_advance \ + \ \epsilon$$

With the instrument variable to predict ln(price):

$$\widehat{ln(price)} = 5.8739 - \ 0.0029 * \ days\_in\_advance - 0.7367 * \ isNormCabin \ + \ \nu$$

The average ticket price was identified as an endogenous variable, estimated using an instrument variable, cabin type. Advanced purchase variable - which estimates the number of days tickets were purchased before departure - was identified as an exogenous variable. We log-transformed the demand and price variables to measure elasticity of demand. The ln(price) coefficient was -0.2431 which means that demand is inelastic. Meanwhile, advanced purchase has a slope of 0.0012 which means that the further away the departure date, the higher the number of seats purchased per transaction.

The intuition of both variables are logical. Train demand is inelastic as other transportation modes, such as cars, planes and ships might not be readily available in certain areas. Similarly, customers are more likely to purchase more tickets further away from the departure date; groups will plan their travel early - wanting to avoid the risk of last minute bookings.

2. Business Insights

There are various degrees of 'sensitivity' of goods demand to the change in price. Suppliers - such as train companies - have to price their goods accordingly. Elasticity represents this sensitivity and products can be deemed to be elastic, unitarily elastic or inelastic. Elastic demand is where there is more than proportionate change in quantity demanded for a corresponding change in price. Inelastic demand would be the inverse - that is a less than proportionate change in quantity demanded for a corresponding change in price.

The demand function for this dataset is judged to be relatively inelastic. Given the vast distances and easy accessibility - as train stations are typically located in key cities - that trains cover, there are few viable substitutes. When train prices rise, there is less than a proportionate drop in demand. Train companies will be able to raise prices without worrying about a significant drop in demand.

Train companies can also capitalise on passengers who urgently need to travel and charge higher prices the closer the purchase date is to departure date. We included such consideration in our analysis by creating a new variable to measure purchase urgency.

3. Data & Features

The datasets used were of train ticket sales data from June 2018 to June 2019. The datasets present 209,697 entries and 14 distinct features. Each entry represents a ticket purchase within the time period.

To maintain consistency throughout the report, we have created a table of all the variables used, and how they would address it moving forward (*See Appendix 1*).

3.1 Exploratory Data Analysis (EDA)

Through our EDA, we have found that the data contained a few notable outliers and inconsistencies, including:

- **Inconsistency of the Cumulative Sales variable**. This variable was supposed to always be increasing until the end of the period but had decreasing values on multiple dates.

- **Train 'O' only had one entry.** This could be caused by data loss or that the purchases for Train 'O' were not recorded properly.

- **Outliers in the ticket price variable.** There were 2 extreme outliers in the ticket variable which cost $7,855.77 and $1,701.56. This is a huge jump from the mean price which was $230.12 (*See Appendix 7*).

3.2 Data Transformation

Natural log was applied to both the ticket price and the number of seats variables, transforming them into ln(price) and ln(seats) respectively. This was done to transform them into variables that could measure demand elasticity.

By applying log, the coefficient is now the ratio of relative change of demand to price. Simply taking the coefficient without log would result in a ratio of absolute change which would not enable us to measure elasticity.

Furthermore, to be able to explore the data further, the train number and customer category variables were transformed into dummy variables with binary input (*See Appendix 7*).

3.3 Feature Engineering

A new variable, advanced purchase, was created by subtracting departure date with purchase date. This was done to explore whether urgency affects customers' demand, as discussed in our business insights.

We also created the weekend factor variable, which indicates whether the purchase was made on a weekend, to better study the effects of the weekend on train ticket sales

3.4 Feature Selection

Based on our EDA and preliminary models, we decided to exclude the following variables:

Cumulative Sales due to its inconsistencies discussed. It comprises of sub-groups which we had no information on, which means its associations with other variables is unknown.

Train Number, Return Trip and One-way Trip variables were not included as intuitively as they should not affect customers' demand for train tickets.

Customers would not base their decision to purchase tickets based on train types, whether it's a one way trip or if it's a return ticket; they will buy a ticket only based on their need to travel.

4. Model

4.1 Assumptions

A few assumptions for the model to hold:

- Supply and demand is in equilibrium.
- No effects of inflation, foreign exchange and cyclical seasonality on any of the variables used.
- Elasticity of demand and price are constant across the time period.
- All train services remain constant across the time period.
- No external shock affecting the price and demand.

4.2 Model Formulation

In this section, we will define the variables that could best explain the demand function.

Independent Variable:
We will use ln(seats) as the total number of seats sold per day estimates the quantity of train tickets demanded.

Dependent Variables:
We chose ln(price) as the first dependent variable that will be adopted in the structural model as price is always a factor in a demand function.

Beside all the excluded variables mentioned in part 3.4, the customer category variable will also be excluded from the structure model. Customer segmentation was done by the train provider and is therefore not relevant to demand. Train riders are most likely not aware of this segmentation.

There are 3 other potential dependent variables: advanced purchase, cabin type and weekend factor. To determine the best variables, we did a backward step model selection to test whether each variable is a good demand estimator by considering its coefficient, p-values and the model's adjusted R^2.

Structural Model 1

To explore the model, we ran OLS with all the candidate dependent variables.

$$ln(seats) = 1.7024 - 0.2236 * \ln(price) + 0.0013 * \ days\_in\_advance + \\ 0.0144 * isNormCabin + 0.0049 * \ isWeekend + \epsilon$$

Weekend factor is insignificant, with a p-value of 0.1083. We further observe that cabin type is also insignificant - when it is removed, the adjusted R^2 value remains at 0.090 (*See Appendix 3*). As such, we can remove both variables.

Intuitively, cabin type cannot be an exogenous variable for demand as it does not affect the number of tickets demanded. Instead, cabin type should affect the number of tickets supplied. The supply of a special cabin is usually less than that of a normal cabin, perhaps because a special cabin takes more space as it is bigger. Furthermore, cabin type is correlated with price as a special cabin is usually priced higher than a normal cabin.

Structural Model 2

After removing the insignificant variables, we find that advanced purchase is a viable exogenous variable. The revised model is indeed suitable with significant variables with p-value < 0.05 and correct slope. Advanced purchase has a positive slope of 0.0013 as travel groups will tend to purchase tickets in advance, while ln(price) has a negative slope of -0.2236 as an increase in price will reduce demand (*See Appendix 3*).

The chosen demand structural function can be written as below:

$$ln(seats) = 1.7541 - 0.2316 * ln(price) + \\ 0.0013 * days\_in\_advance + \epsilon$$

The variables which were not selected in the structural model could be a potential Instrument Variable (IV).

Simultaneity Problem:

The demand curve suffers from a simultaneity problem due to its correlation with the supply curve. As such, the OLS function used to estimate the demand function will suffer from endogeneity bias.

We identify that ln(price) is the endogenous variable in the structural model. Ticket price change could be caused by hidden variables, such as supply shock. When the hidden variables change the ticket price, it will also affect the demand. A model which contains an endogenous term will be inaccurate as it suffers a bias where its error term is not completely random.

Instrument Variable (IV)

There are 3 potential IVs for ln(price): cabin type, customer category and weekend factor. The chosen instrument variable must be correlated with ln(price) but must not correlate with ln(seats). We adopted a forward step model selection to find the best IV model.

We can estimate the reduced form by expressing ln(price) in terms of the chosen IV:

$$\widehat{ln(price)} = \pi_0 + \pi_1 * feature1 + \pi_2 * feature2 + \pi_3 * feature3 + ... + \nu$$

IV Model 1 - Cabin Type

As shown in the structural model, cabin type is not directly linked to ln(seats). This was further validated in the correlation matrix where we found that cabin type was more correlated with ln(price), with coefficient -0.71 than with ln(seats), with coefficient 0.21 (*See Appendix 5*).

Running Model 1:

$$\widehat{ln(price)} = 5.8739 - 0.0029 * days\_in\_advance - 0.7367 * isNormCabin + \nu$$

The cabin type is significant, passed the Hausman test and gives Adjusted R^2 of 0.596, the highest among other variables. Thus, this is the best IV for price (*See Appendix 2*).

IV Model 2 - Customer Category

Customer category was explored as an IV because they might segment leisure vs business travellers. This is akin to cabin type which would have a correlation with price but not directly with seats. The results from the correlation matrix does not lend credence to it however, indicating a similarly weak correlation with ln(price) (-0.49) vs ln(seats) (0.28) (*See Appendix 5*).

Running Model 2 gives :

$$\widehat{ln(price)} = 5.8849 - 0.0041 * days\_in\_advance - 0.4885 * Customer\_Cat + \nu$$

Customer category variable is also significant and passed the Hausman test - albeit, it's Adjusted R^2 is 0.400, lower than that of cabin type (*See Appendix 2*).

IV Model 3 - Weekend Factor

For the weekend factor, the results showed that it has very little correlation not only with ln(price) and ln(seats) but with virtually every other variable (*See Appendix 5*).

Running Model 3 gives:

$$\widehat{ln(price)} = 5.5605 - 0.0052 * days\_in\_advance + 0.0576 * isWeekend + \nu$$

While the weekend variable is significant and passed the Hausman test, the R^2 of this model is the lowest among other models at 0.303. The weekend variable can thus be categorised as a weak instrument variable (*See Appendix 2*).

We also explored models where we combined the potential IVs.

IV Model 4 - Cabin Type & Customer Category:

Running Model 4 gives

$$\widehat{ln(price)} = 5.9836 - 0.0026 * days\_in\_advance - 0.6694 * isNormCabin - 0.2162 * Customer\_Cat + \nu$$

IV Model 5 - Cabin Type & Weekend Factor:

Running Model 5 gives

$$\widehat{ln(price)} = 5.8589 - 0.0029 * days\_in\_advance - 0.7367 * isNormCabin + 0.0583 * isWeekend + \nu$$

Whilst the variables in IV Model 4 and 5 are all significant and passed the Hausman test, both models did not pass the Sargan test as all p-values are < 0.05. Hence, the hypothesis that all IVs are exogenous were rejected in both models (*See Appendix 2*).

Thus, IV Model 1 was chosen. It can be seen from Model 1 reduced form that the F-statistic - 1.544e05 - is large and p-value < 0.05 is significant. Thus we can reject the null hypothesis that cabin type is a weak instrument (*See Appendix 2*).

4.3 Final 2SLS Model

Based on the chosen IV Model, the final 2SLS was run with cabin type as the sole IV

Final Reduced Form Model:

$$ln(\widehat{price}) = 5.8739 - \ 0.0029 * \ days\_in\_advance - 0.7367 * \ isNormCabin \ + \ \nu$$

Reduced Form Model Interpretation

The cabin type coefficient of -0.7367 indicates that price will drop by 73.67% when the cabin type is normal. This supports our understanding that a normal cabin is always cheaper than a special cabin.

On the other hand, the coefficient of advanced purchase is -0.0029, showing that price is higher when a ticket is purchased closer to the departure date. This also aligns with our understanding because train companies usually price rushed tickets higher as they understand that travellers are willing to pay more when there is urgency.

Final 2SLS Model:

$$ln(seats) = 1.8183 - 0.2431 * ln(\widehat{price}) + 0.0012 * \ days\_in\_advance \ + \ \epsilon$$

↓

2SLS Model Interpretation

The ln(price) coefficient of -0.2431 indicates a relatively inelastic demand as a 1% change in price will lead to only a 0.24% change in quantity demanded. This is in line with our understanding from business insights that the travellers will still buy train tickets if price increases given that they need to make the trip.

On the other hand, the coefficient of 0.0012 for advanced purchase is small but significant. It is observed that the customers tend to buy tickets in advance of the departure date. As such, this could have a larger effect on the quantity demanded.

5. Conclusion

By estimating ln(price) with cabin type, we have removed the hidden variables affecting price from this model. The chosen IV model passed all the endogeneity tests and we found that the original OLS model indeed suffered from endogeneity bias.

The coefficients of all the variables used in both the IV and 2SLS models are in line with our understanding and business applications - as described in part 4.3. Thus, the final 2SLS model seems to have solved the endogeneity problem posed by the original OLS model. Nonetheless, the model has a large residual as evident in the low adjusted R^2, and as such, there is room for potential improvements.

5.1 Potential Improvements

- **Better context on datasets and its sampling methods:** since the dataset was given and utilised on an 'as-is' basis, some of the features, such as cumulative sales, were dropped in the feature selection phase due their inconsistency and lack of information on their constituents.

  From a business standpoint, cumulative sales is a variable that would be a good instrumental variable, as train companies' would naturally include it into their pricing strategy. When cumulative sales are high and seats are limited, suppliers would increase prices. If more information were available, this might affect the choice of variables selected which could in turn improve the reliability of the model.

- **Better instrument variables:** The model could also be improved with potentially better candidate instrumental variables that were not available in the dataset. For example, variables like the cost of resources – diesel, electricity, oil, and coal for instance – to power trains would intuitively be included in the supply function and would probably be highly correlated with ticket prices. These variables could further reduce the endogeneity bias.