Factors influencing Scots' satisfaction with public transport

Group 28: Aishwin Tikku, Mengran Li, Steven Kwok, Shaoquan Li, Shuning Li

Introduction

- We **aim** to find factors related to passengers' satisfaction with public transport to help operator to improve its service.
- The **data** are from Scotland's official statistics. The theme of Transport contains seven datasets, Road Transport Expenditure, Public Transport, Road Vehicles, Concessionary Travel Cards, Road Network and Traffic, Travel to Work and Other Purposes. There are 460 observations of 24 variables in this research.

Methods

- Summarize table and density plots are illustrated to detect data patterns. The scatter and correlation plots are proposed to explore the relationship among variables. Potential factors are identified through **EDA**.
- The *Satisfaction* is the response variable, the *DateCode* is the control variable and others are independent variables. A **linear regression** model is applied as Eq.(1):

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \tag{1}$$

- **Model diagnosis** are carried out to check model assumptions. **Stepwise regression** is applied to select variables with **AIC** as the criterion. Compare the selected model with full model on $adj\ R^2$, AIC and BIC.
- The uncertainty of the parameters is determined via **bootstrap** method. The significant variables are verified and their 95% CI are estimated.

Results

Model diagnosis

Fig.1 shows the regression diagnosis results of model1 with all variables.

- Residuals vs Fitted plot (left top) shows that there is no systematic correlation between the residual value and the fitting value.
- Normal Q-Q plot (right top) shows the points on the graph fall on a straight line with an angle of 45 degrees, which indicates the assumption of normality is not violated.
- Scale-location plot (left bottom) displays the points around the horizontal line are randomly distributed. The invariant variance assumption is satisfied.
- Residuals vs leverage (right bottom) figures out special observations.
- According to the correlation plot and VIF, there is obvious **multicollinearity** among variables.

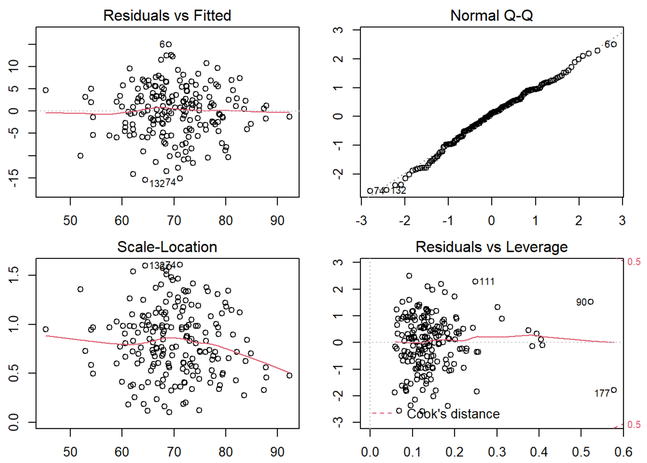

Figure 1: Regression diagnosis results

Stepwise regression

Stepwise regression technique is applied to model selection. The selected model is as Table. 1. Control variable isn't displayed.

Table 1: Model2, selected by Stepwise regression

term	estimate	std.error	statistic	p.value
(Intercept)	49.14	4.04	12.2	0.00
Cards	0.00	0.00	3.3	0.00
Repair	0.24	0.06	3.8	0.00
Work_Bus	0.86	0.10	8.3	0.00
School	-0.12	0.05	-2.6	0.01
Health	0.77	0.51	1.5	0.13
Work_Train	0.81	0.13	6.4	0.00
Train_Stations	-0.14	0.04	-3.3	0.00
Without_Car	0.10	0.07	1.4	0.15
Petrol_Diesel	-0.03	0.01	-2.2	0.03

The variables with high correlation are removed by stepwise regression.

Model comparison

Model2 has higher $adj\ R^2$ and smaller AIC and BIC compared with the model1 as Table. 2.

Table 2: Comparison of model1 with model2 on adj R2, AIC and BIC

model	adj.r.squared	AIC	BIC
model 1	0.55	1271	1366
model 2	0.56	1255	1310

Therefore the **model2** is better.

Bootstrap

To obtain robust results, a bootstrap is developed for the estimation of parameters. This process is repeated **1000** times.

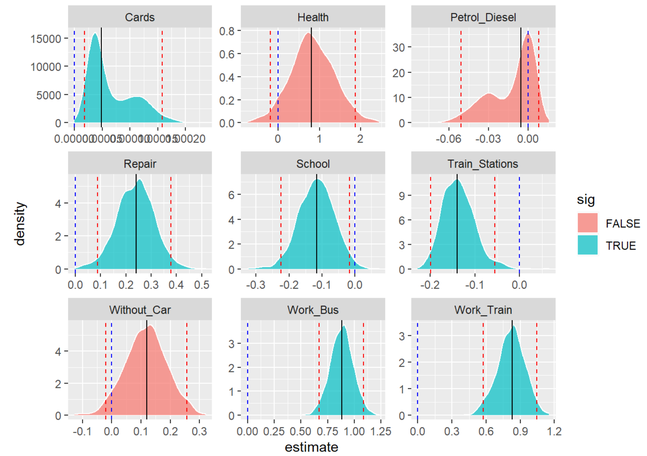

Figure 2: Density plots of parameters via bootstrap

The density plots of parameters are displayed in Fig.2. The variables with orange are not significant while the variables with blue are significant at the $\alpha = 0.05$. The blue dashed lines are zero and the orange dashed lines are 95% CI of parameters.

Conclusion

- The variables **Cards** *(Number of concessionary cards issued to all adults)*, **Repair** *(The Percentage of Roads Needing Repairs)*, **Work_Bus** *(Bus Journeys To Work)* and **Work_Train** *(Train Journeys To Work)* have **positive** influence on satisfaction with public transport. The **School** *(Child Journeys To School By Walking/Cycling)* and **Train_Stations** *(Number of Train Stations)* have **negative** relationship with satisfaction with public transport.
- The reason that more roads needing repairs and less train stations come with higher satisfaction needs to be further explored.

References

- Jim Hester and Hadley Wickham, (2020). *fs: Cross-Platform File System Operations Based on 'libuv.'* R package version 1.5.0
- Kuhn et al., (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*.
- Wickham et al., (2019). *Welcome to the tidyverse. Journal of Open Source Software*, e, 4(43), 1686