Factors influencing Scots' satisfaction with public transport

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

Introduction

Public transport is necessary for most Scottish people. Therefore, its comfort and customer satisfaction are important for operators. To help improve public transport, The purpose is to discover factors having correlation with passengers' satisfaction in this project. The data are from Scotland's official statistics. The theme of Transport contains seven data sets, 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.Find the best results through "Model diagnosis" , "stepwise regression" and comparing the selected model with full model on $adj\ R^2$, AIC and BIC.

Exploratory Data Analysis

The whole data set, combined from seven data sets obtained from website of Scottish government statistics, has 24 factors and 460 observations. The explanatory variables are Road Casualties, Road Transport Expenditure, Public Transport, Road Vehicles, Concessionary Travel Cards, Road Network and Traffic and Travel to work and other purposes.

Firstly, filter and merge data.

```r
#### data wrangling ####
data_dir <- "data"
# read data document name
csv_files <- fs::dir_ls(data_dir, regexp = "\\.csv$")
# dataset name
dataset_name <- c(
  "Casualties", "Expenditure", "Transport",
  "Vehicles", "Cards", "Network", "Purposes"
)
# read data and assign names to dataset
for (i in 1:7) {
  assign(dataset_name[i], readr::read_csv(csv_files[i]))
}
# select total number of casualty outcomes
Casualties <- Casualties %>%
  select(-c(Measurement, Units)) %>%
  filter(Outcome == "Killed Or Seriously Injured") %>%
  spread(Outcome, Value) %>%
  filter(Age == "All" & Gender == "All") %>%
  select(-c(Age, Gender))

# rename variable name
Expenditure <- Expenditure %>%
  select(-c(Measurement, Units)) %>%
  rename(Expenditure = Value)

# remove variables at the whole Scotland level
Transport <- Transport %>%
  select(-c(Measurement, Units)) %>%
  filter(`Indicator (public transport)` %in% c(
    "Number Of Passenger Train Stations",
    "Percentage Of Adults Reporting that they are Very or Fairly Satisfied with Public Transport"
  )) %>%
  spread(`Indicator (public transport)`, Value)

Vehicles <- Vehicles %>%
  select(-c(Measurement, Units)) %>%
  spread(`Indicator (road vehicles)`, Value)

# retain card numbers of all people
Cards <- Cards %>%
  select(-c(Measurement, Units)) %>%
  spread(Age, Value) %>%
  select(-`60 years and over`) %>%
  rename(Cards = All)

Network <- Network %>%
  select(-c(Measurement, Units)) %>%
  spread(`Indicator (road network traffic)`, Value)

Purposes <- Purposes %>%
  select(-c(Measurement, Units)) %>%
  spread(`Indicator (travel to work)`, Value)

# merge all dataset into one complete dataset by 'FeatureCode' and 'DateCode'
Data <- Expenditure %>%
  left_join(Casualties,
    by = c(
      "FeatureCode" = "FeatureCode",
      "DateCode" = "DateCode"
    )
  ) %>%
  left_join(Cards,
    by = c(
      "FeatureCode" = "FeatureCode",
      "DateCode" = "DateCode"
    )
  ) %>%
  left_join(Network,
    by = c(
      "FeatureCode" = "FeatureCode",
      "DateCode" = "DateCode"
    )
  ) %>%
  left_join(Purposes,
    by = c(
      "FeatureCode" = "FeatureCode",
      "DateCode" = "DateCode"
    )
  ) %>%
  left_join(Transport,
    by = c(
      "FeatureCode" = "FeatureCode",
      "DateCode" = "DateCode"
    )
  ) %>%
  left_join(Vehicles,
    by = c(
      "FeatureCode" = "FeatureCode",
      "DateCode" = "DateCode"
    )
  )

# rename variables
names(Data) <- c(
  "FeatureCode", "DateCode", "Expenditure", "Killed_Injured",
  "Cards", "Congestion", "Repair", "Mileage", "Work_Bus",
  "Business", "School", "Commuting", "Work_Cycling", "Education",
  "Health", "Shopping", "Work_Train", "Work_Walking", "Train_Stations",
  "Satisfaction", "One_Car", "More_Car", "Without_Car", "Petrol_Diesel"
)

# Transform the DateCode as a factor
Data$DateCode <- as.factor(Data$DateCode)
```

We get a complete data set called Data, and summarize histogram plots (Fig. 1) are illustrated to detect data patterns.

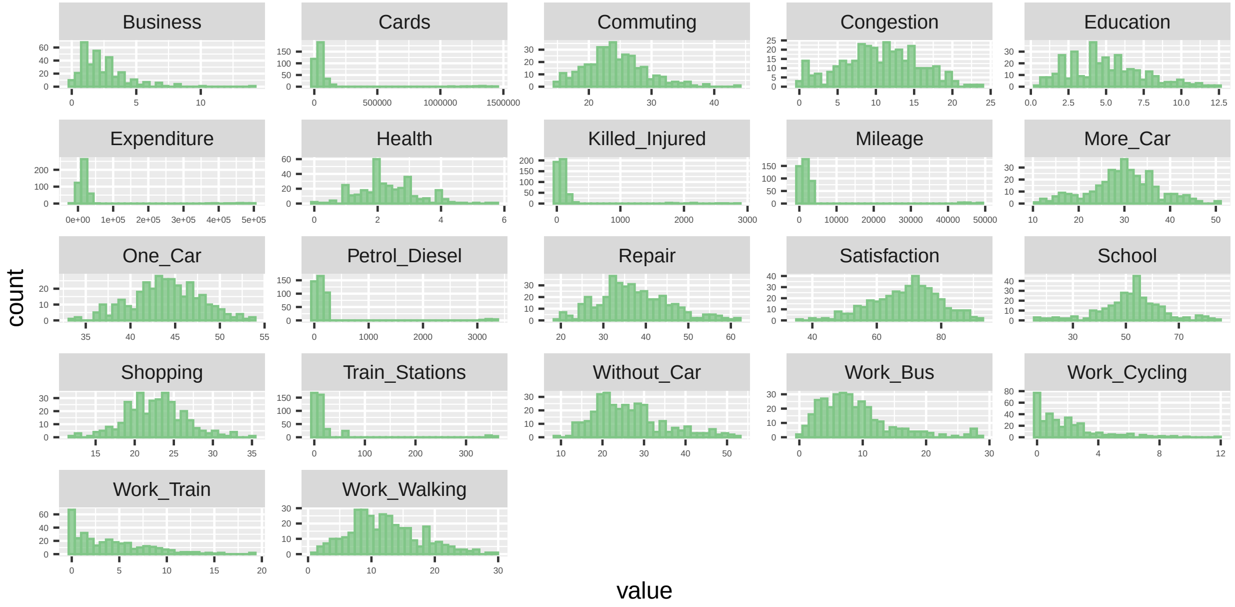


Figure 1: Histogram plot of variables

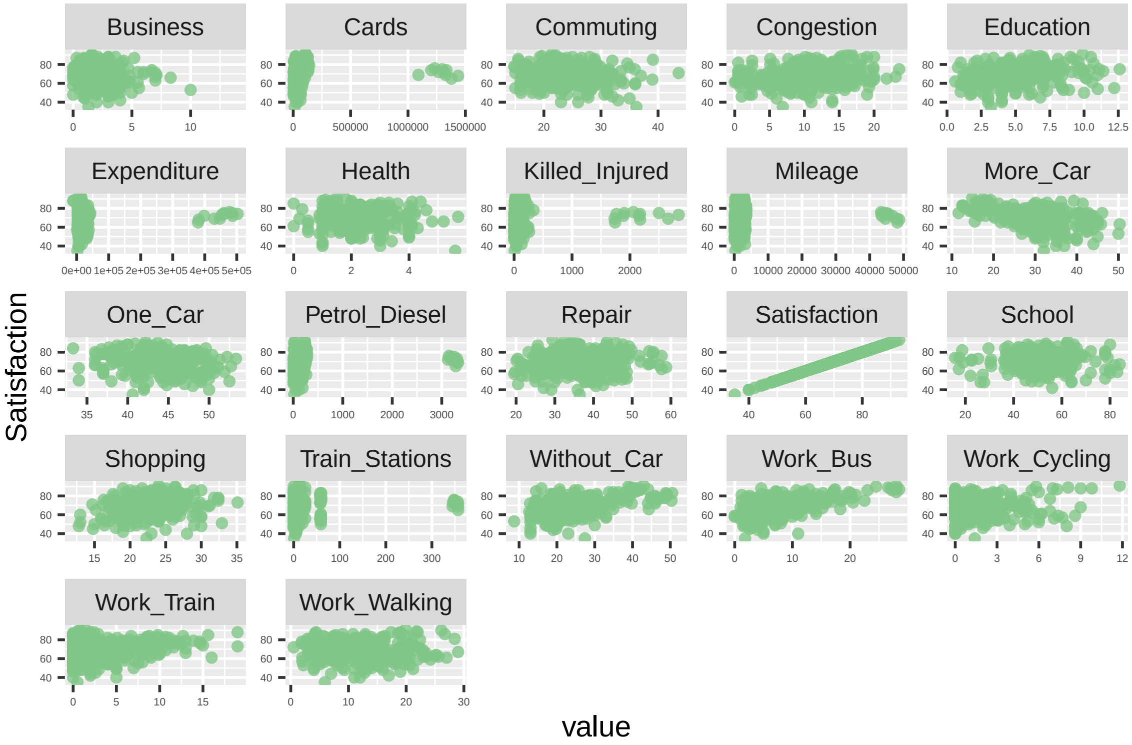
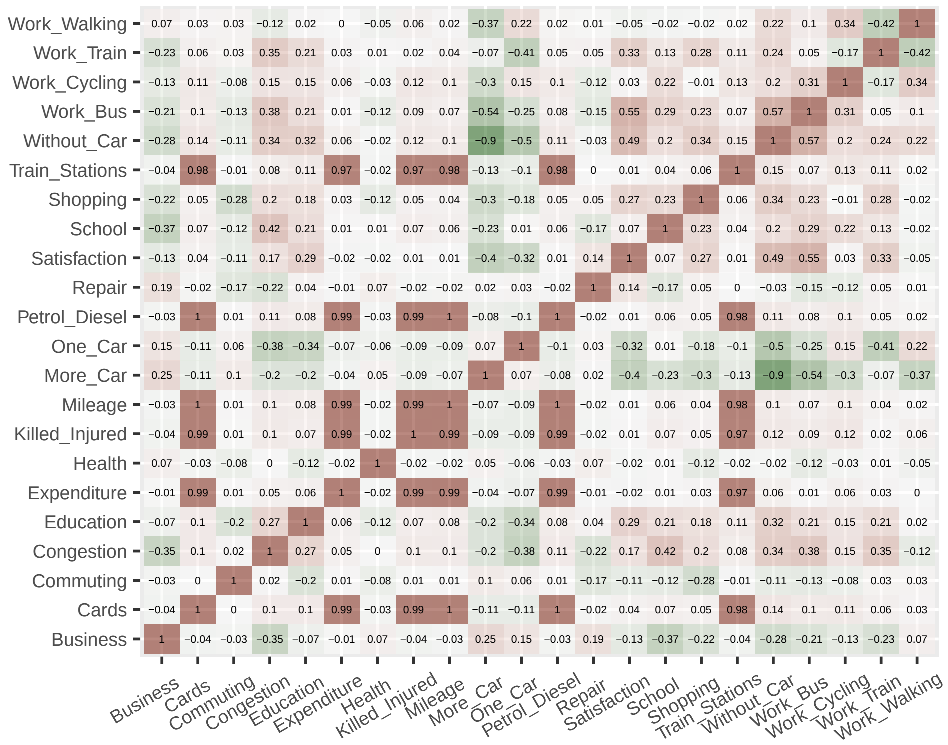


Figure 2: Scatter plot of dependent variabls vs response variable



	Business	Cards	Commuting	Congestion	Education	Expenditure	Health	Killed_Injured	Mileage	More_Car	One_Car	Petrol_Diesel	Repair	Satisfaction	School	Shopping	Train_Stations	Without_Car	Work_Bus	Work_Cycling	Work_Train	Work_Walking
Work_Walking	0.07	0.03	0.03	−0.12	0.02	0	−0.05	0.06	0.02	−0.37	0.22	0.02	0.01	−0.05	−0.02	−0.02	0.02	0.22	0.1	0.34	−0.42	1
Work_Train	−0.23	0.06	0.03	0.35	0.21	0.03	0.01	0.02	0.04	−0.07	−0.41	0.05	0.05	0.33	0.13	0.28	0.11	0.24	0.05	−0.17	1	−0.42
Work_Cycling	−0.13	0.11	−0.08	0.15	0.15	0.06	−0.03	0.12	0.1	−0.3	0.15	0.1	−0.12	0.03	0.22	−0.01	0.13	0.2	0.31	1	−0.17	0.34
Work_Bus	−0.21	0.1	−0.13	0.38	0.21	0.01	−0.12	0.09	0.07	−0.54	−0.25	0.08	−0.15	0.55	0.29	0.23	0.07	0.57	1	0.31	0.05	0.1
Without_Car	−0.28	0.14	−0.11	0.34	0.32	0.06	−0.02	0.12	0.1	−0.9	−0.5	0.11	−0.03	0.49	0.2	0.34	0.15	1	0.57	0.2	0.24	0.22
Train_Stations	−0.04	0.98	−0.01	0.08	0.11	0.97	−0.02	0.97	0.98	−0.13	−0.1	0.98	0	0.01	0.04	0.06	1	0.15	0.07	0.13	0.11	0.02
Shopping	−0.22	0.05	−0.28	0.2	0.18	0.03	−0.12	0.05	0.04	−0.3	−0.18	0.05	0.05	0.27	0.23	1	0.06	0.34	0.23	−0.01	0.28	−0.02
School	−0.37	0.07	−0.12	0.42	0.21	0.01	0.01	0.07	0.06	−0.23	0.01	0.06	−0.17	0.07	1	0.23	0.04	0.2	0.29	0.22	0.13	−0.02
Satisfaction	−0.13	0.04	−0.11	0.17	0.29	−0.02	−0.02	0.01	0.01	−0.4	−0.32	0.01	0.14	1	0.07	0.27	0.01	0.49	0.55	0.03	0.33	−0.05
Repair	0.19	−0.02	−0.17	−0.22	0.04	−0.01	0.07	−0.02	−0.02	0.02	0.03	−0.02	1	0.14	−0.17	0.05	0	−0.03	−0.15	−0.12	0.05	0.01
Petrol_Diesel	−0.03	1	0.01	0.11	0.08	0.99	−0.03	0.99	1	−0.08	−0.1	1	−0.02	0.01	0.06	0.05	0.98	0.11	0.08	0.1	0.05	0.02
One_Car	0.15	−0.11	0.06	−0.38	−0.34	−0.07	−0.06	−0.09	−0.09	0.07	1	−0.1	0.03	−0.32	0.01	−0.18	−0.1	−0.5	−0.25	0.15	−0.41	0.22
More_Car	0.25	−0.11	0.1	−0.2	−0.2	−0.04	0.05	−0.09	−0.07	1	0.07	−0.08	0.02	−0.4	−0.23	−0.3	−0.13	−0.9	−0.54	−0.3	−0.07	−0.37
Mileage	−0.03	1	0.01	0.1	0.08	0.99	−0.02	0.99	1	−0.07	−0.09	1	−0.02	0.01	0.06	0.04	0.98	0.1	0.07	0.1	0.04	0.02
Killed_Injured	−0.04	0.99	0.01	0.1	0.07	0.99	−0.02	1	0.99	−0.09	−0.09	0.99	−0.02	0.01	0.07	0.05	0.97	0.12	0.09	0.12	0.02	0.06
Health	0.07	−0.03	−0.08	0	−0.12	−0.02	1	−0.02	−0.02	0.05	−0.06	−0.03	0.07	−0.02	0.01	−0.12	−0.02	−0.02	−0.12	−0.03	0.01	−0.05
Expenditure	−0.01	0.99	0.01	0.05	0.06	1	−0.02	0.99	0.99	−0.04	−0.07	0.99	−0.01	−0.02	0.01	0.03	0.97	0.06	0.01	0.06	0.03	0
Education	−0.07	0.1	−0.2	0.27	1	0.06	−0.12	0.07	0.08	−0.2	−0.34	0.08	0.04	0.29	0.21	0.18	0.11	0.32	0.21	0.15	0.21	0.02
Congestion	−0.35	0.1	0.02	1	0.27	0.05	0	0.1	0.1	−0.2	−0.38	0.11	−0.22	0.17	0.42	0.2	0.08	0.34	0.38	0.15	0.35	−0.12
Commuting	−0.03	0	1	0.02	−0.2	0.01	−0.08	0.01	0.01	0.1	0.06	0.01	−0.17	−0.11	−0.12	−0.28	−0.01	−0.11	−0.13	−0.08	0.03	0.03
Cards	−0.04	1	0	0.1	0.1	0.99	−0.03	0.99	1	−0.11	−0.11	1	−0.02	0.04	0.07	0.05	0.98	0.14	0.1	0.11	0.06	0.03
Business	1	−0.04	−0.03	−0.35	−0.07	−0.01	0.07	−0.04	−0.03	0.25	0.15	−0.03	0.19	−0.13	−0.37	−0.22	−0.04	−0.28	−0.21	−0.13	−0.23	0.07

Figure 3: Correlation plot of variabls

Through the correlation plot (Fig. 3), there is a strong linear relationship between "Petrol_Diesel" "Train_Stations", "Mileage", "Expenditure", "Killed_Injured" and "Cards". Then we analyse the regression diagnosis results (Fig. 4) of model1 with all variables.

```
#### linear regression ####
# full model
fit <- lm(Satisfaction ~ ., data = na.omit(Data[, -1]))
```

Plot the Fitted values against Residuals and Q-Q plot to assess our assumptions.

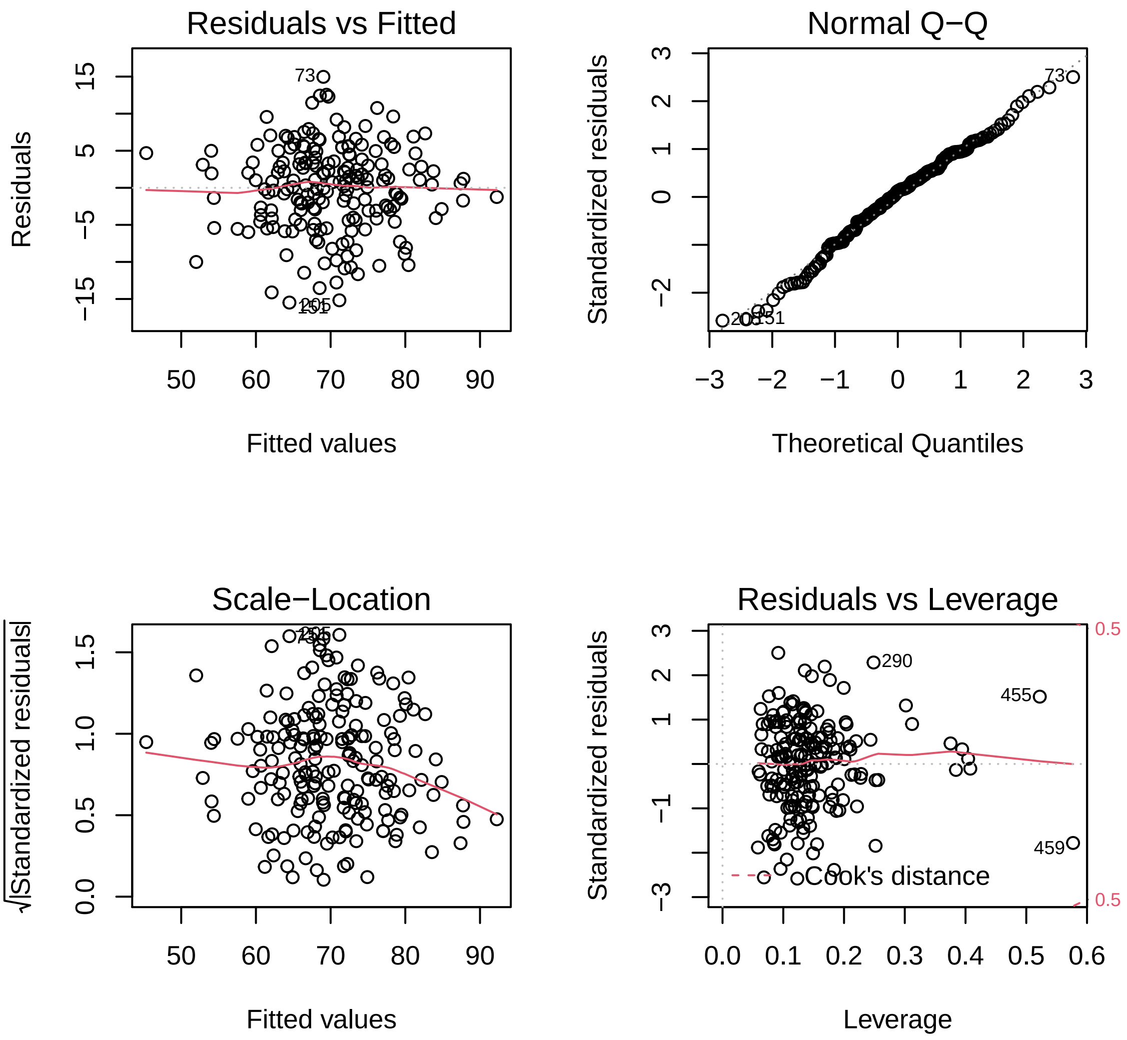


Figure 4: Regression diagnosis results

- 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.

Table 1: Model2, selected by Stepwise regression

term	estimate	std.error	statistic	p.value
(Intercept)	52.9775625	3.9466689	13.4233613	0.0000000
Cards	0.0001289	0.0000352	3.6559946	0.0003356
Repair	0.2494980	0.0665578	3.7485909	0.0002391
Work_Bus	0.7333901	0.1074918	6.8227543	0.0000000
School	-0.1207559	0.0485715	-2.4861470	0.0138198
Health	0.2824925	0.5313962	0.5316043	0.5956519
Work_Train	0.7066629	0.1314720	5.3750078	0.0000002
Train_Stations	-0.1537078	0.0468778	-3.2789066	0.0012497
Without_Car	0.1676106	0.0740901	2.2622533	0.0248705
Petrol_Diesel	-0.0367942	0.0146720	-2.5077725	0.0130293

The invariant variance assumption is satisfied.

- Residuals vs leverage (right bottom) figures out special observations. The strong influence points do not deviate from the regression estimation seriously.
- According to the correlation plot and VIF, there is obvious multicollinearity among variables.

Formal Data Analysis

Linear regression model, a model for examining and discovering relations between the response variable and explanatory variable(s), is applied in this project. The Satisfaction is the response variable, the DateCode is the control variable and others are independent variables. A linear regression model is applied as:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

where y is the response variable, x_1 to x_p is the number of columns selected from 1 to p, β_0 is the intercept of data, β_1 to β_p is the coefficients of corresponding columns of X from 1 to p. The ε is the error terms of the estimations.

```
#library(MASS)
library(MASS)
stepAIC(fit, na.rm = TRUE)

# 3 4 5 8 24
# cards, killed_injured, expenditure, petrol_diesel, mileage are highly correlated.
# model 2 without
names(Data)[c(3, 4, 5, 8, 24)]

fit2 <- lm(Satisfaction ~ DateCode + Cards + Repair + Work_Bus +
             School + Health + Work_Train + Train_Stations + Without_Car +
             Petrol_Diesel, data = na.omit(Data[, -1]))
summary(fit2)

#### Model selection ####

glance(fit)
glance(fit2)
```

A total of four regressions were performed. First, all the independent variables were regressed, and then the highly correlated factors were sequentially removed, and finally, four models were obtained. The variables

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

model	adj.r.squared	AIC	BIC
model 1	0.5478475	1271.288	1365.604
model 2	0.4998739	1274.555	1310.330

with high correlation are removed by stepwise regression. Step AIC method is applied as a criterion for model selection. The model 2 is the best choice. Stepwise regression technique is applied to model selection. The selected model is as Table. 1. Control variable "DateCode" isn't displayed.

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

Therefore the model2 is better.

To obtain robust results, using Bootstrap to select significant variables at 95% level and obtain its confidence interval of parameter estimation and their observations.This process should be repeated 1000 times.

```r
boot_models <- bootstraps(Data[, -1], times = 1000, apparent = TRUE) %>%
  mutate(
    model = map(splits, ~ lm(Satisfaction ~ DateCode + Cards + Repair + Work_Bus +
                               School + Health + Work_Train + Train_Stations + Without_Car +
                               Petrol_Diesel, data = .)),
    coef_info = map(model, tidy)
  )

boot_coefs <- boot_models %>%
  unnest(coef_info)

ci <- int_pctl(boot_models, coef_info)
# significant at 0.05
sig <- ci[sign(ci[, 2]) == sign(ci[, 4]), ] %>% pull(term)

boot_coefs <-  boot_coefs %>%
  mutate(sig = ifelse(term %in% sig, "TRUE", "FALSE")) %>%
  filter(term %in% c("Cards", "Health", "Petrol_Diesel","Repair", "School", "Train_Stations", "Without_

boot_coefs2 <- boot_coefs %>%
  group_by(term) %>%
  summarise(est = median(estimate), lower = quantile(estimate, 0.025), upper = quantile(estimate, 0.975

boot_coefs %>%
  ggplot() +
  geom_density(aes(estimate, fill = sig), alpha = 0.7,  color = "white") +
  geom_vline(data = boot_coefs2, mapping = aes(xintercept = est))+
  geom_vline(data = boot_coefs2, mapping = aes(xintercept = lower), color="Red", lty = 2)+
  geom_vline(data = boot_coefs2, mapping = aes(xintercept = upper), color="Red", lty = 2)+
  geom_vline(data = boot_coefs2, mapping = aes(xintercept = 0), color="Blue", lty = 2)+
  facet_wrap(~term, scales = "free")+
  theme(axis.text.x= element_text(size = 6),
        axis.text.y= element_text(size = 6))
```

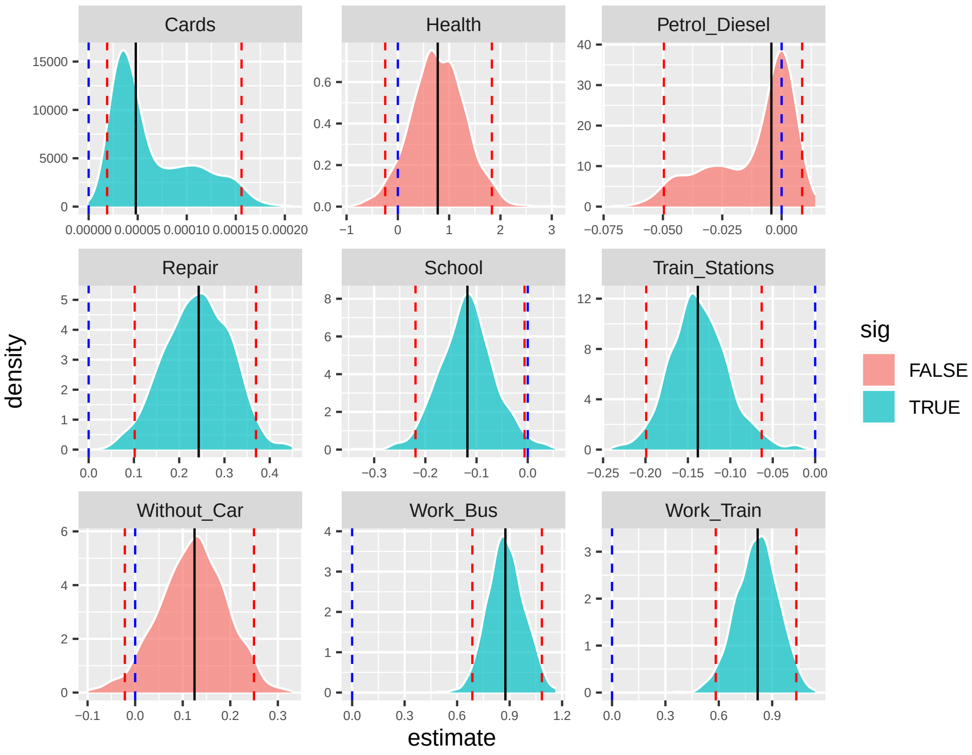


Figure 5: Density plots of parameters via bootstrap

The density plots (Fig. 5) of parameters are displayed. 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. The black lines are estimations of parameters.

Conclusions

- 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