Third Assignment

Christopher Gandrud, Hertie School of Governance, Spring 2016

Emilia Sicari & Rafael Lopez V.

April 19, 2016

Contents

1 Research question and project description

Research question: How the rise in inequality, economic growth and usage of public transportation influences the purchase of cars (as an example of luxury good) in Singapore, from 1995 to 2014.

We collected data on economic growth, income inequality, usage of public transportation and number of privately and publicly owned vehicles, between 1995 and 2014. As suggested by our research question, economic growth, income inequality and usage of public transportations are the explanatory varibales, while purchase of cars is the dependent variable. The reason why we chose cars, is thar they are an example of luxury goods with which wqealthy people can show their social status; in fact, in Singapore purchasing cars is particularly expensive, due to a certificate of car entitlement which may cost even more than 70.000 dollars (Authority 2014). For more details about the research proposal and case justification see ResearchProposal.

2 Processing data

2.1 Data sources and data gathering

This document and particularly data processing was made using: R (2016), Quandl (2015), Corrplot (2013), Ggplot(2015), Pander (2015), Repmis (2016) and Rio (2016).

The data for our empirical analysis were retrieved from:

- IMF Cross Country Macroeconomic Statistics open data available on Quandl. From this source we downloaded data showing the trend in Singopore's GDP per capita meassured in singaporean dollars from 1981 to 2021 (forecasted from 2015 onwards). The data was provided in csv format and imported on R using the URL of the website.
- World Top Incomes Database available on Knoema, provides access to data on the distribution of top incomes in more than twenty five countries across the globe. From this source we downloaded data on the top 10% average income and bottom 90% average income in Singapore from 1947 until 2009, meassured in singaporean dollars. Since it was not possible to directly import the database to R, we requested and received the data via e-mail in csv format. This data set is available in the repository.[1]
- Singapore's open data portal offered two data bases:
  - The Annual Motor Vehicle Population, provides the number of public and private vehicles from 1960 to 2015, including: mortorbikes, rental cars, buses, taxis and other type of vehicles. While mortorbikes, rental cars and cars are private means of transportation, buses and taxis are to be considered public since in Singapore even the taxis are provided by the state.
  - Public transport utilization. This data is expressed as the daily average of thounsand commuters using public trasport by year. It covers the span from 1995 to 2014 and includes the following modes of transportation: MRT (underground), LRT (a localised rail systems acting as feeder services to the Mass Rapid Transit network), taxis (publicly run) and buses.

[1] We did not gather data from the database Clio Infra as initially stated in our ResearchProposal , since it did not provide sufficient data for the time span we are considering.

The following table summarizes the variables taken into consideration for the analysis.

Table 1: Summary of variables

Variable	Description	Time.frame
gdp per capita	measured in singaporean dollars at current prices	1980-2021
inequality	top 10% and bottom 90% singaporean's average income measured in singaporean dollars	1947-2009
anual motor vehicle	number of: cars, rental cars, buses, taxis, buses, motorbikes	1960-2015
public transport utilization	average commuters using daily: MRT, LRT, Buses, Taxis	1995-2014

2.2 Cleaning, processing and merging data sets

- After importing data we used the "date" variable (year) as a unique identifier for all four datasets, in order to merge them afterwards.
- Since time frames of the data were different, we selected a common span of time: 1995-2014. In the case of bottom 90% and top 10% average income, we had to make a linear regression to forecast missing values (from 2009 until 2014). The results, available in a new dataframe, were later on bounded with the orginal one, in order to have the entire time series. As for LRT, values from 1995 until 1998 were missing since the service started to be provided from 1999 (Infopedia 2005); therefore, we completed the dataframe giving the value "0" for the first 4 years of the time span taken into consideration.
- Cleaning the data was limited to changing column names, eliminating the unnecesary ones and organizing the various data frames so to merge them more easily afterwards, using the year as common denominator. Only in the case of the dataframe cointaining the number of private cars in Singapore from 1995 until 2014 (car.pop.1) we had to change the format of the data from characters to integers, due to an incorrect import.
- In order to have an indicator showing the trend in inequality in Singapore between 1995 and 2014, we created a new variable - named "inequality" - by divididing the top 10% avereage income by the bottom 90% average income for each year: the coefficient of the division shows how many times Singaporeans earning the top 10% average income are reacher than the bottom 90% earners of the population.
- As for the number of cars, we simply divided them into the categories provided in the data original set: cars, buses, etc. Originally, they were in one column so we separate them in several ones to have the year as a unique identifier.
- Finally, we merged all the single dataframes into the new one, containing all the variables that we used to perfom descriptive and inferential statistical analyses.

3 Descriptive and inferential statistics: preliminary results

3.1 Descriptive statistics and central tendency

The table below shows the basic decriptive statistics for our variables.

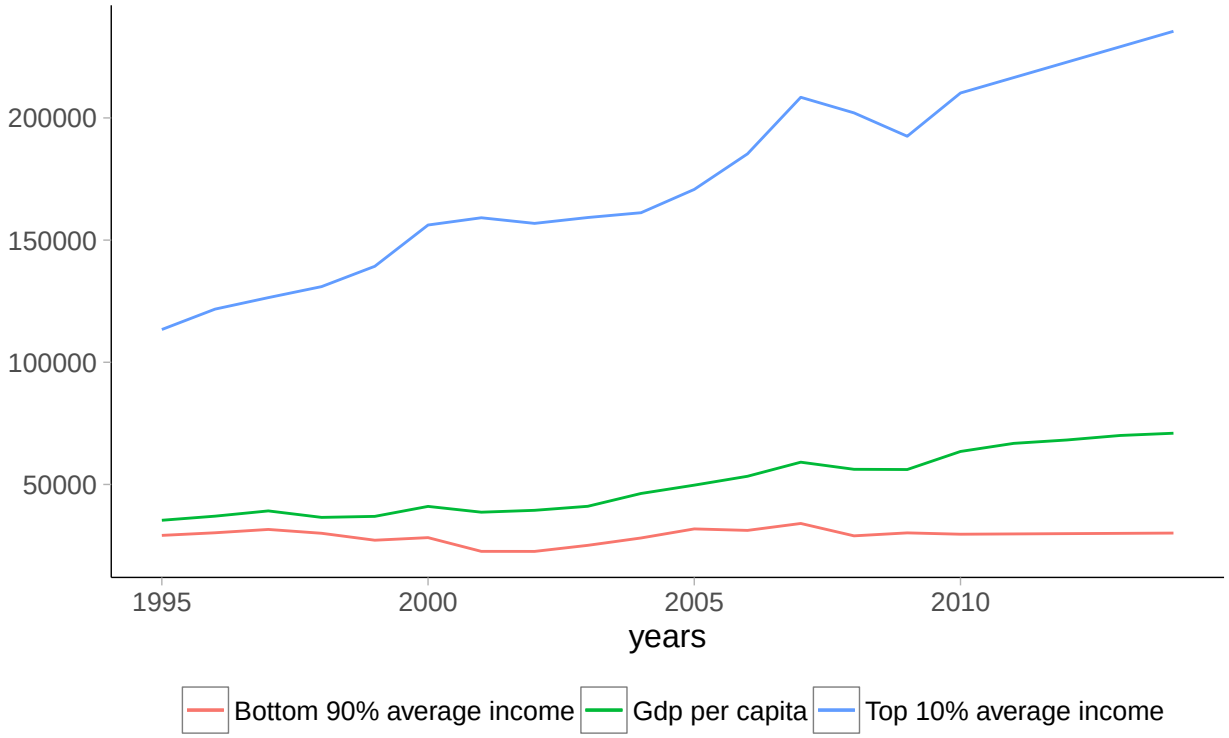
Table 2: General data summary

Statistic	N	Mean	St. Dev.	Min	Max
gdp.per.capita	20	50,277.2	12,717.7	35,345.5	70,966.9
inequality	20	6.1	1.3	3.9	7.8
top	20	174,881.0	38,176.0	113,402.5	235,450.0
bottom	20	29,022.2	2,851.4	22,602.4	34,043.3
cars	20	466,148.3	97,290.5	342,245	607,292
rentalcars	20	10,097.0	3,881.3	5,144	18,847
taxis	20	21,958.7	4,201.0	16,517	28,736
buses	20	13,993.9	2,302.3	10,723	17,554
motorbikes	20	138,985.9	6,435.2	129,587	148,160
other	20	146,548.7	10,180.8	134,756	161,698
bus.u	20	3,159.9	254.5	2,779	3,751
mrt.u	20	1,504.2	635.1	740	2,762
lrt.u	20	62.8	45.0	0	137

3.2 Trends in gdp per capita and inequality

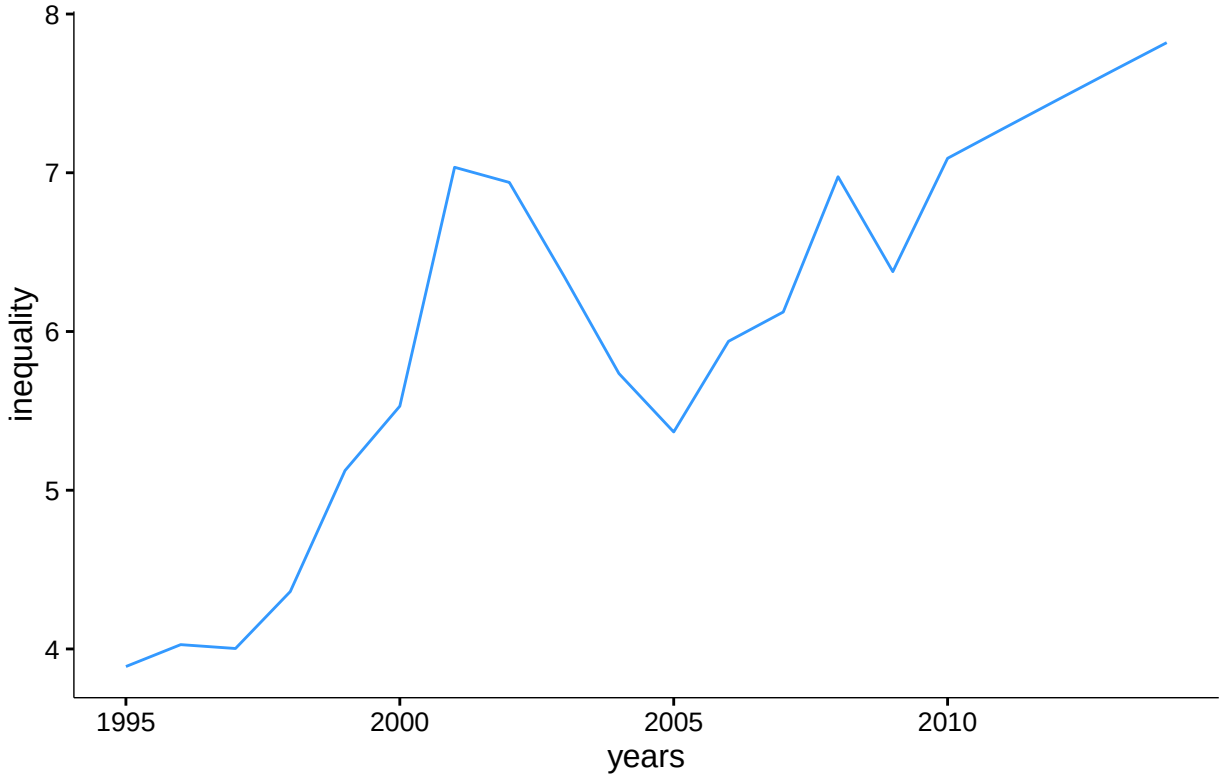
The following graph shows the trend of the explanatory variables in Singapore between 1995 and 2014. As we can see, although slowly, the gdp per capita has risen thourghout the whole period, despite a slight decline between 2002 and 2005 and a more serious reduction in the years of the financial crisis, between 2008 and 2010. The top 10% average income shows the same trend: a steady increase throughout the whole period (in 2014 its value was more than 100% higher than the initial one), with a slight decline between 2002 and 2005, and a more serious reduction in the years of the financial crisis. However, the value of the bottom 90% average income has barely changed, enlarging the difference between the top and bottom populations.

Figure 1 − Gdp per capita, top 10% and bottom 90% average income
in Singapore measured in national currency at current prices (1995−2014)



The growing difference between the top and bottom earners is clarified by the following graph, showing trend in inequality in Singapore, measured in number of times by which the top 10% earners are richer than those earning the bottom 90% average income. The graph confirms what already highlighted above: the difference between the rich and the poor has been increasing all the time, and the trend only reversed between 2002 and 2005 and between 2008 and 2010. The average ratio between both groups is 6.1 and has reaached a maximun value of 7.8.

Figure 2 − Inequality in Singapore measured by the distance between the top 10% and bottom 90% average income in number of times (1995−2014)

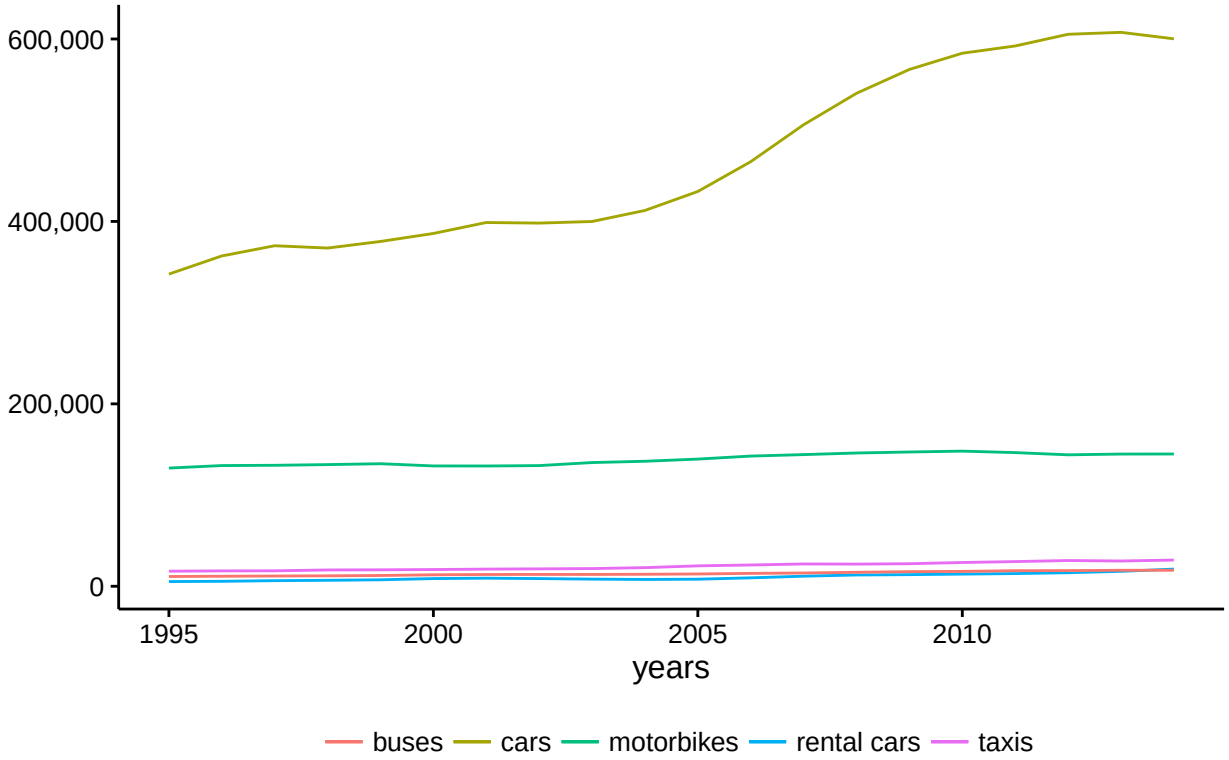


3.3 Trends in ownership and usage of vehicles

The following graph shows the number of public and private vehicles present in Singapore between 1995 and 2015. What is striking, is the continuous and rapid increase in the number of privately owned cars (especially since 2006). This trend supports the hypothesis which links high the economic growth, high the inequality and increase of cars' purchase.[2] A further assumption to be investigated is that such increase might be linked to the likewise rise in the top 10% average income: as the rich become richer, the purchase of luxury goods, such as cars, increases as well.

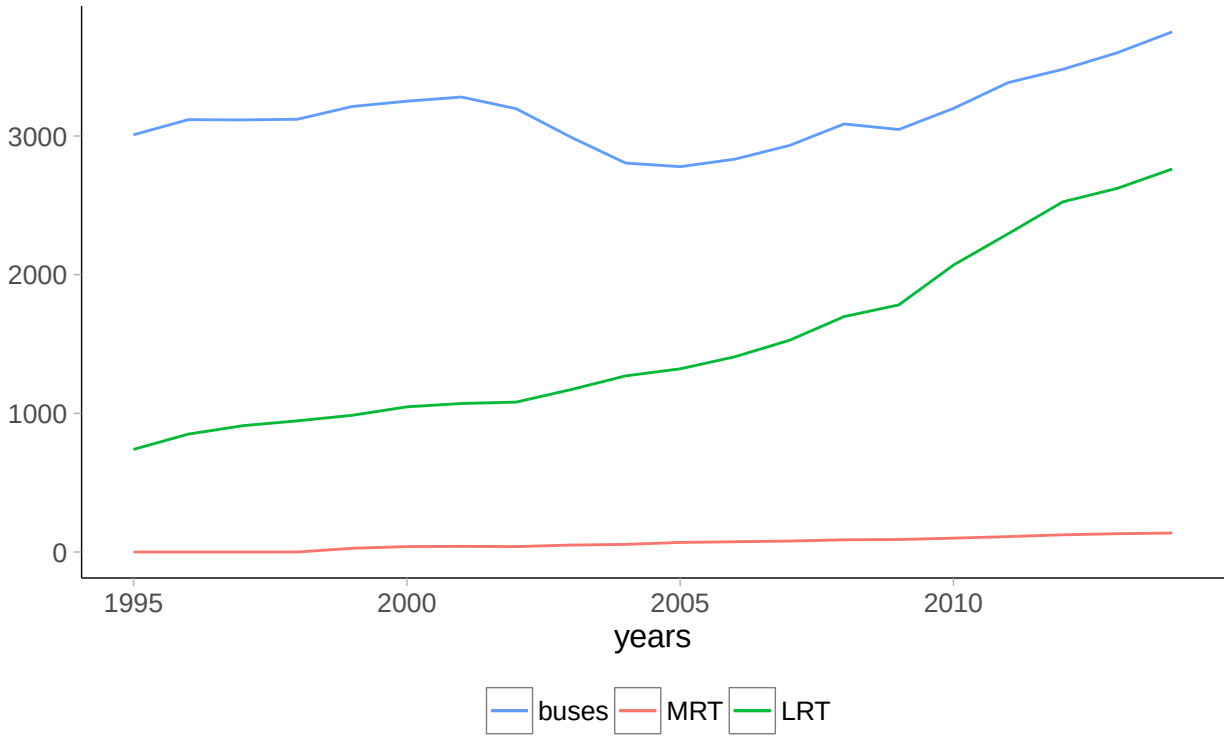
[2]For more details about the hypotheses see https://github.com/EmiliaSicari/ResearchProposal

Figure 3 − Number of public and private vehicles in Singapore (1995−2014)



At the same time, the number of passengers in the main public transportation (MRT and buses) has increased consistenlty over time. Despite that, the publicly owned buses have not significantly changed in number. Consequently, this also supports the hypothesis that the usage of public transport is not entirely linked with the purchase of cars: in fact, usage of public transport has either increased (in the case of MRT and buses) or stayed the same (in the case of LRT), while the number of private cars has grown consistently. Even in this case a further assumption to be investigated is that those using public transports are lower earners.

Figure 4 − Average daily passengers using public transport in Singapore
in thousands (1995−2014)

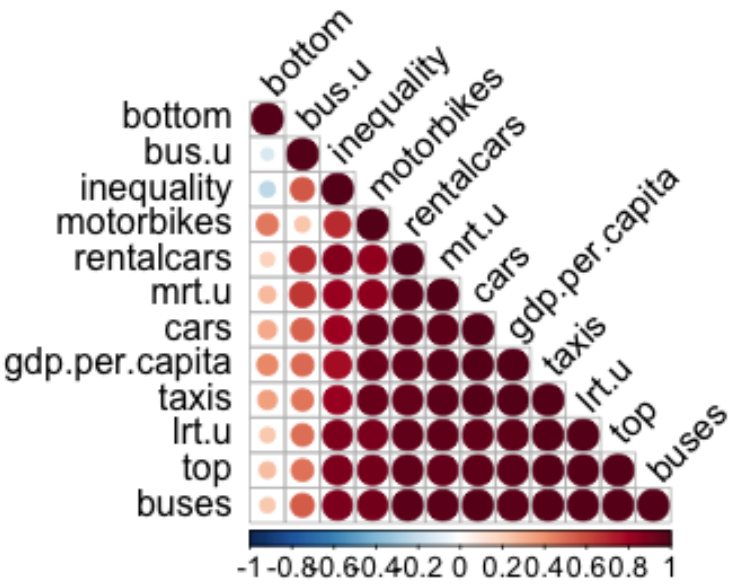


3.4 Correlation analysis

The graph below[3] shows the correlation among the varibles considered in our analysis: the darker the colour, the stronger the correlation. Likewise, the size of the circles shows the magnitude of the correlation. While blue indicates negative correlation, red is associated with positive correlation.

What clearly emerges from the plot, is that the variables are in almost all of the cases highly and positively correlated to each other. Bottom 90% average income and buse utilization are less correlated to the other variables, and bottom 90% average income is also negatively correlated with both inequality and buses utilization (which weakens the assumption that the poorest are those who use more public transportations). However, high correlation among explanatory variables might create problems due to multicollinearity and may show a biased in the variables in general , which will be assesed in the next stage of the research.

[3] bus.u, lrt.u and mrt.u stand for the utilization of these modes of transportation. We also eliminated the "other" variable since it not included in the anlysis



The following table shows the correlation coefficients.[4]

Table 3: Correlation matrix

	gdp	ineq.	top	bottom	cars	rental cars	taxis	buses	m.bikes	bus.u	mrt.u	lrt.u
gdp	1.00	0.77	0.97	0.42	0.98	0.94	0.99	0.97	0.92	0.50	0.97	0.96
ineq.	0.77	1.00	0.88	-0.23	0.80	0.86	0.81	0.88	0.68	0.55	0.81	0.88
top	0.97	0.88	1.00	0.26	0.96	0.95	0.98	0.98	0.90	0.48	0.94	0.98
bottom	0.42	-0.23	0.26	1.00	0.32	0.19	0.35	0.22	0.46	-0.14	0.27	0.22
cars	0.98	0.80	0.96	0.32	1.00	0.95	0.98	0.98	0.94	0.52	0.96	0.95
rental cars	0.94	0.86	0.95	0.19	0.95	1.00	0.94	0.97	0.83	0.69	0.97	0.95
taxis	0.99	0.81	0.98	0.35	0.98	0.94	1.00	0.98	0.93	0.47	0.97	0.98
buses	0.97	0.88	0.98	0.22	0.98	0.97	0.98	1.00	0.90	0.54	0.97	0.98
m.bikes	0.92	0.68	0.90	0.46	0.94	0.83	0.93	0.90	1.00	0.23	0.84	0.89
bus.u	0.50	0.55	0.48	-0.14	0.52	0.69	0.47	0.54	0.23	1.00	0.65	0.49
mrt.u	0.97	0.81	0.94	0.27	0.96	0.97	0.97	0.97	0.84	0.65	1.00	0.96
lrt.u	0.96	0.88	0.98	0.22	0.95	0.95	0.98	0.98	0.89	0.49	0.96	1.00

3.5 Multiple regression analysis

To provide preliminary results about our research question, we used a multiple linear regression model.

Probably due to the high correlation between explanatory variables, the model has a low statistical significance, just one variable has significance at 95% of confidence. Despite the**gdp percapita** is the only significant variable, its explanatory power seem to be biased, since the coefficient suggests that a 5 dollar increase result in the purchase of 1 extra car. Likewise, the high value of the R2 suggests that the model needs to be reviewed, givinen the low statistical significance of the whole model. Therefore, further modelling and data transformation is needed to reach higher standards of significance.

[4]Considerations made in the previous footnote also apply for the correlation matrix.

Table 4: Multiple regression model

	Dependent variable:
	cars
gdp per capita	5.72**
	(2.65)
inequality	16,894.68
	(12,113.70)
buses utilization	−37.76
	(49.54)
MRT utilization	69.34
	(70.49)
LRT utilization	−745.14
	(862.04)
(Intercept)	138,101.80
	(161,762.70)
Observations	20
R^2	0.96
Adjusted R^2	0.95
Residual Std. Error	22,092.91 (df = 14)
F Statistic	70.89*** (df = 5; 14)
Note:	*p<0.1; **p<0.05; ***p<0.01

4 Further steps

- Building at least two alternative models:
  - A probabilistic model, showing the likelihood of an increase in the population of cars given the independant variables.
  - A second model using logartims in all the variables or some of them, in order to show the variable changes in percent terms.
- Assesing multicollinearity, and accordingly correcting the model.

- Analyzing possible bias by homoscedasticity.
- Controlling for other variables such us governmental policies that may have affected the model.
- Making analysis of influential residuals.
- Improving the focus in the literature review.
- Preparing the final report and presentation.

References

Authority, Land Transport. 2014. "Certificate of Entitlement." http://www.lta.gov.sg/content/ltaweb/en/roads-and-motoring/owning-a-vehicle/vehicle-quota-system/certificate-of-entitlement-coe.html.

Chan, Chung-hong, Geoffrey CH Chan, Thomas J. Leeper, Christopher Gandrud, and Ista Zahn. 2016. *Rio: A Swiss-Army Knife for Data I/O*. https://CRAN.R-project.org/package=rio.

Dar'oczi, Gergely, and Roman Tsegelskyi. 2015. *Pander: An R Pandoc Writer*. https://CRAN.R-project.org/package=pander.

Gandrud, Christopher. 2016. *Repmis: Miscellaneous Tools for Reproducible Research*. https://CRAN.R-project.org/package=repmis.

Infopedia, Singapore. 2005. "Light Rail Transit." http://eresources.nlb.gov.sg/infopedia/articles/SIP_538_2005-01-05.html.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Raymond McTaggart, Gergely Daroczi, and Clement Leung. 2015. *Quandl: API Wrapper for Quandl.com*. https://CRAN.R-project.org/package=Quandl.

Wei, Taiyun. 2013. *Corrplot: Visualization of a Correlation Matrix*. https://CRAN.R-project.org/package=corrplot.

Wickham, Hadley, and Winston Chang. 2015. *Ggplot2: An Implementation of the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.