國立政治大學應用經濟與社會發展

英語碩士學位學程

International Master’s Program of Applied Economics

and Social Development

College of Social Sciences

National Chengchi University

碩士論文

Master’s Thesis

利用 Google 關鍵字與機器學習預測日本汽車銷量
Predicting Japanese Car Sales with Google Trends and Machine Learning

Student: Mariia Morozova

Advisor: Kuang-Ta Lo, Tzu-Ting Yang

中華民國 107 年 7 月

July 2018

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

利用 Google 關鍵字與機器學習預測日本汽車銷量

Predicting Japanese Car Sales with Google Trends and Machine Learning

研究生： 莫柔娜　　Student: Mariia Morozova

指導教授： 羅光達　　Advisor: Kuang-Ta Lo

楊子霆　　Tzu-Ting Yang

國立政治大學

應用經濟與社會發展英語碩士學位學程

碩士論文

A Thesis

Submitted to International Master's Program of Applied

Economics and Social Development

National Chengchi University

中華民國 107 年 7 月

July 2018

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

Abstract

Computers and the Internet has been significantly changing our lives over the past few decades and bringing both a lot of opportunities and challenges to our lives. Internet, on the 1 hand, possess a lot of free and important information. For example, information about consumers' moods and preferences that can be extracted from the Web using Google Trends search index data which is undoubtedly precious for market research and forecast. While computers and their computation abilities using machine learning make it feasible to improve to improve task performance, particularly forecasting and planning.

The aim of this research is to utilize both tools – Google Trends data and Least Absolute Shrinkage and Selection Operator (LASSO, a machine learning method) in forecasting Japanese car sales. This paper pursues two main goals: to compare the machine learning method performance with conventional and human-created models and to identify if Google Trend data helps to improve forecasting model for Japanese car sales.

From the results of this research it can be concluded that machine learning methods definitely have some positive implications for forecasting. LASSO definitely outperform human-judgment. Generally, LASSO models with optimal penalty size are very comparable in their out of sample prediction accuracy to autoregressive models. LASSO with optimal lambda also creates models that include a limited number which is undoubtedly easier to interpret.

Google Trends data should be treated with care. It is, in generally, advised to run LASSO-regression when working with Google data as LASSO is able to identify the right lags for the Google search indexes that is of a critical importance due to the fact that different brands might have different characteristics and different consumers.

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

Table of Contents

1.1 Background 1
1.2 Problem statement 2
1.3 Research goal 3
2. Literature review 5
2.1 Forecasting with Google Trends 5
2.2 Forecasting with Google Trends 6
2.3 Forecasting with LASSO Overview 9
3. Data Collection 11
3.1 Japan new cars monthly sales data 11
3.2 Macroeconomic Indicators Data 13
3.3 Google Trends Data 15
4. Methodology 18
4.1 Human Judgement Model Construction 18
4.2 Machine Learning Model 20
4.3 Model Prediction Accuracy Measurements 21
5. Results 24
5.1 Choosing Model by Human Judgement 24
5.2 Choosing Model by Machine Learning Method 28
5.3 Models Comparison: Machine and Human Models 32
5.4 Further Models Comparison 36
6. Discussion of Results 44
References 47
Appendix 1 49
Appendix 2 50

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

Appendix 3 51

Appendix 4 52

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

1. Introduction

1.1 Background

For the past few decades, computers and the Internet have been extensively utilized for various purposes, not only by individuals, but also by investors, businesses, economists, and multiple institutions. On the one hand, the Internet collects and provides a significant amount of data. Machine learning, however, as a new and very promising tool, can analyze those scopes of data and perform tasks connected with this data.

The Internet possesses quickly acquired, current, free information on consumers' metadata, tastes, preferences, and opinions that were previously only obtainable through time and money-consuming surveys. One of the instruments of the Internet that has gained popularity significantly throughout the past few years is Google Trends or «Google Trends Data Goldmine» (Spiegel, 2015). Google Trends data has been excessively involved in forecasting and various marketing purposes by companies in many industries as well as scholars for research purposes. Many believe that Google Trends serves as a tool to understand consumer preferences that result in specific consumer behavior. For this reason, many use this for prediction purposes.

Ever since the paper by Choi and Varian was published in 2009, it has widely been attempted by contemporary scholars to include Google Trends data in the forecasting process in search of better results. Many believe that Google Trends, being a search query index, provides insight into consumers` attention to brands or other matters. Google Trends helps to understand consumers` intention and attention and allow a large-scale collection of free, current information. However, this scope of literature provides different insights. Some of the papers manage to prove the positive contribution from Google Trends data, while others conclude that the old time series method is enough to produce an accurate forecast.

Forecasting methods have been issues of increased interest recently. Researchers have been trying to improve prediction models, not only by including new variables into the model or by modifying existing models, but by using more modern data science methods such as machine learning for predictions.

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

For this reason, machine learning has also been of high interest the past few decades. Machine learning uses statistical techniques and gives computers the ability to learn with data without being programmed to do so, while also being able to improve the performance of a task. Machine learning allows computers to analyze big data and make predictions. Many consider it to be revolutionary and an esteemed innovation.

1.2 Problem statement

Sales forecasting, a company`s attempt at estimating future sales, plays a central role in a company`s planning, cost minimization, and overall efficiency and success. It is essential to obtain the most accurate forecasting method to make informed business decisions. Due to this, there is enormous attention and interest to the newest techniques that help to improve sales forecast accuracies. These techniques include machine learning methods of forecasting and traditional forecasting methods that incorporate Google search.

One market that incorporates these techniques is the automotive market. The automotive market is one of the biggest industries in many countries. Automobiles are an expensive, durable goods that has a high-involvement decision-making process. For this reason, they possess a high risk for consumers concerning the amount of money spent on the purchase, as well as post-purchase service. That is why consumers are highly motivated to do their research before the purchase. The most convenient way to do this research is turning to the search engine that will provide necessary information in just a few seconds. By doing so, the intention to buy the specific brand can be captured in Google Trends.

Traditionally, the methods for quantitative predictions methods are time series and causal forecasting. Time series sales forecast is based on previous sales data, it reveals a trend and assumes that the pattern will be present in the future. Another type is causal forecasting that establishes causal relationships between matters. Thus, in case of sales, it includes data on economic variables that represent how well or poor a country`s economy is performing, predicting the capability of people to obtain certain things. In specific predictions model, researchers combine those methods to provide better forecasting accuracy. Economic variables are recently observed not always to be able to capture some sudden or structural changes in the economy or consumers` opinions. Those traditional

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

methods may ignore a significant part of the consumer`s demand definition – the desire to obtain this or that thing; a consumer’s preference. Since Google Trends data is believed to capture such information, Google Trends data recently has been heavily utilized in forecasting.

Theoretically, it might sound compelling that, by capturing a consumer’s preference, we might improve forecasting results. However, in practice, the performance of different predicting models will depend on the good itself, and numerous characteristics of the country, economy, consumers, and investigated market.

Another aspect of this research is machine learning. Machine learning as a modern data science method, tool for forecasting that has become very popular within the last few years. It provides a selection of sufficient variables and their lags, so it performs model selection. It is also believed to give researchers a more accurate prediction and prevent overfitting as well as no omitted variable bias.

1.3 Research goal

Regarding usage of Google Trends data, the existing scope of literature does not produce consistent results concerning different markets. That may be because different matters bear different characteristics. Alike, different countries with their economies, mentalities and purchasing behavior and preferences are also not identical. Extrapolating results from previous studies on consumers` population of other nations might be complicated. For this reason, there is a necessity to study every case individually to drive more accurate results and to choose the forecasting model correctly.

This paper aims to analyze the Japanese car market and a few forecasting models regarding this market. Provided in this paper is a model that performs best in this research regarding the automotive market of Japan in total sales (of all brands) and for the sales of three leading car brands in Japan. The research includes both, creating forecasts using time series model with Google Trends and macroeconomic variables, as well as the machine learning method that uses data to select the variables for forecasting.

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

The underlying idea behind this research is that the best performing model will be a modern model that includes Google search data. However, the researchers of this paper do not deny the idea that the contrary might be proved. Additionally, the researchers investigate whether machine or human created models perform better. Therefore, the two research questions are as follows:

RQ1: Is the machine learning method outperforming human-judgment and conventional models when forecasting Japanese car sales?

RQ2: Does the Google Trends index help to improve forecasting models for Japanese car sales?

This research will not only produce the forecasting method best suitable for the Japanese car market but also enlarge the knowledge on the matter. From this, one draw other valuable conclusions concerning the Japanese car market.

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

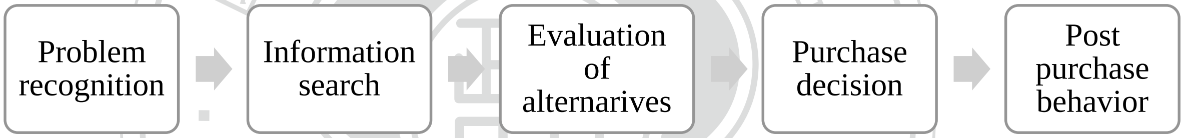
2. Literature review

This section will provide the theoretical knowledge related to the topic. It will illustrate how Google Trends might be an important indicator of one`s purchase intention, especially in the case of durable goods, will explain the theory behind LASSO – one of machine learning techniques, while also covering previous research findings on the issue.

2.1 Forecasting with Google Trends

To illustrate the relevance of Google Trends in sales forecasting, the 5-stage model of the decision-making process will be used. This model is one of the core concepts in marketing, business, and consumer consumption theory. This model describes the stages of consumer consumption.

Figure 2.1.1 5-stage model of the decision-making process. Adapted from: Kotler (2012)



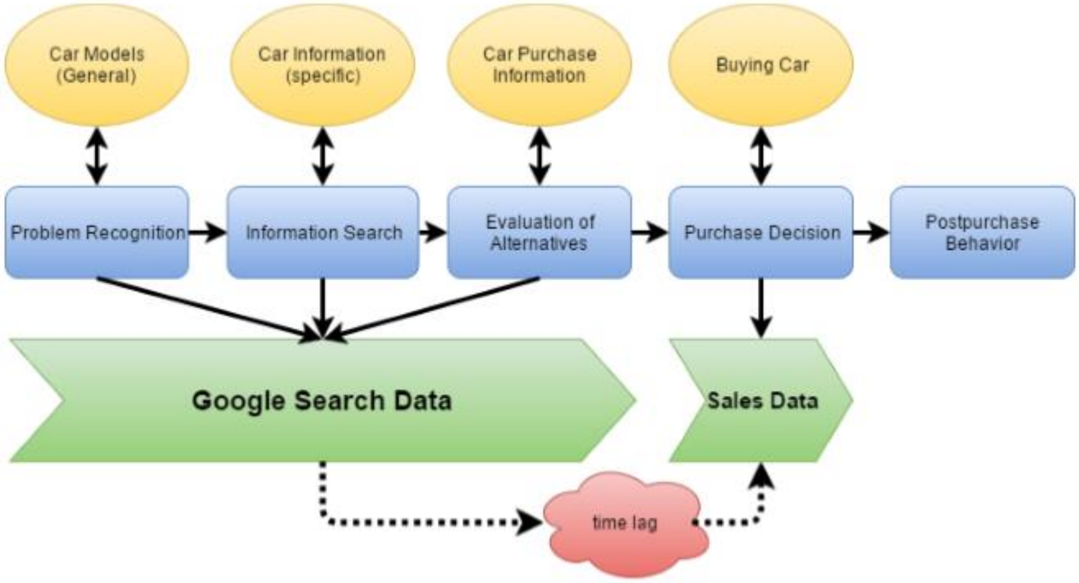
The method of consumer decision making consists of five stages: problem recognition, information search, evaluation of alternatives, purchase decision, and post-purchase behavior. At the first stage, the person recognizes that there is a difference between the desired and the real state, so the consumer identifies a need or a problem. In this case, a need to buy a car. The second stage is the stage of investigating the matter and searching for more information on the issue. The third stage involves a person processing alternative information to make a final decision. This stage is then followed by stages four and five; the actual purchase and post-purchasing behavior (Kotler, 2012).

For this research, we should look more closely at the second stage. At this stage, consumers are likely to look for an individual brand they prefer. Since an automobile is a product that is associated with complex buying behavior, involving a high-involvement decision-making process. In other words, a car purchase is essential to the consumer and it consists of some risk to a consumer. For this reason, people are more likely to spend more time doing the active research, such as browsing the Internet, talking with friends, and visiting stores to learn more about the product. Consumers will be looking for more information on the car's characteristics and features.

Since researching information, of course, will involve some time, the actual sale will come with a lag regarding the Google search. Therefore, it is necessary to include lag terms into the proposed model. At the third stage, people are more likely to look for car purchase information: particular places to buy the car, dealers, and possible special deals.

Thus, as proposed by Muehlen, the above-mentioned model of decision making by Kotler can be modified (Muehlen, 2017) and can take the following form below.

Figure 2.1.1 Modified 5-stage model of decision making. Reprinted from «Improved sales forecasting with consumer behavior» Muehlen, M., 2017



2.2 Forecasting with Google Trends

There was a significant number of papers published that tried to predict consumer behavior or future sales using web search (Google Trends data). The subjects of those papers were ranging from predicting house sales and stock market prices to different types of retail sales. A clear majority of studies are forecasting sales of durable goods because durable goods, as explained in part 2.1, involve a complicated process of evaluation of product price and qualities as well as its potential alternatives. Google Trends also started

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

to be frequently used in prediction of movie theatre tickets and tourism destinations, as those are almost exclusively searched online before purchase.

The original paper on forecasting using Google Trends was published in 2009 by Choi and Varian. The authors decided to present Google Trends to the audience and illustrate how Google data may be utilized in forecasting techniques. The analysis is done by the dataset from the USA and focuses on such industries as retail, home, automotive sales, and travel decisions. The authors attempted to prove that the inclusion of query search indexes might help to reduce the prediction errors, thus, improving the forecasting. It was found that, for different categories of goods, models including Google Trends variables tend to significantly outperform those models that exclude these predictors.

The methodology used was a seasonal autoregressive model or seasonal AR model, for short. Initially, the sales were predicted using just historical data (simple autoregressive model), and this model only included lagged 12 and 1-month sales variables. Furthermore, the second to the last week of the previous month, the Google Trends variable was incorporated into the model accordingly. This inclusion led to smaller absolute errors and smaller mean absolute errors (MAE) for the predicted values. The second to the last week of the previous month Google Trends variable is positively correlated with the current sales. For the automotive sales, it was found that, for both models with and without Google Trends, the proper values are not perfect. However, the error for the model including query search index is smaller. The mean absolute error (MAE) improved by 10,6% of 6,34% with simple AR model to 5,66% when Google trends were included. Choi and Varian highly encouraged researchers to try and carry out their research, as to enlarge the existing knowledge on the topic (Choi and Varian, 2012).

Since this paper was released, inspired by such a promising and exciting new topic, the other studies on the similar topic answer did not take very long to be produced. Since the automotive industry, one of the major industries in several countries, requires accurate forecasts of future sales, much attention was driven to this industry and attempts to improve sales forecasting methods.

Fantazzini and Toktamysova (2015) used the data on new car registration in Germany to build a set of multivariate models, including both economic variables and

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

Google Trends data, to compute out-of-sample predictions for both seasonally adjusted and raw data. There are a few significant findings to be learned from this paper. First, in this case, no significant differences between large, medium-sized and small sellers were found as well as no differences between foreign and German manufacturers. Secondly, rather complex Bayesian VAR models performed well for all car brands and all types of forecasts, while parsimonious bivariate models just built, using car sales and Google Trends data, showed the best predictive power concerning long-term projections. Moreover, Google Trends-based models performed well only in the case of seasonally adjusted data, while in the case of raw data models without Google Trends performed better. In general, prediction models using Google Trends data performed better than others especially in the case of the long-term forecast (Fantazzini and Toktamysova, 2015).

Muehlen, (2017) investigated Mercedes-Benz sales in the Thailand market, found that the forecast accuracy is higher if the Google Trends data is included in the model. The method used in this paper is a simple autoregressive model (ADL) that can consist of lagged dependent and independent variables. Three models were built: the model including macroeconomic variables and seasonal dummies; the forecasting model just with Google trends; the model with both macroeconomic variables and Google trends. The 1-step-ahead, 2-step-ahead, 3-step-ahead, 4-step-ahead and 6-step-ahead forecasts were executed for all three models as well as estimation errors computed. The highest r^2 was obtained for the macro & Google Trends model; the smallest errors (MAPE and MSPE) are found for the macroeconomic model and macroeconomic with Google Trends model. Therefore, based on all evaluation measurements, it was concluded that including Google data into the prediction model increases its accuracy (Muehlen, 2017).

Muehlen (2017) as not the only one to add seasonal dummies into the forecasting model; the same was done by Hand and Judge in their research on predicting cinema admissions in the United Kingdom. The study showed that, in the case of cinema admission prediction, the model performs better if fixed seasonal dummies are included (Hand and Judge, 2010).

Although the literature suggests that there are some positive inputs from Google Trends regarding forecasting, it appears from the experience that every case should be

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

studied individually. Combes and Bortoli (2015) investigated different categories of household consumption in France by using a Bayesian approach as a baseline approach. It is derived from that paper that Google Trends has «limited implications and does not improve sales forecasting except in specifically targeted cases». This is since there is a significant amount of variability in different products. Combes and Bortoli (2015) also claim that, even though Google Trends might have a chance for improving forecasting, the models and their results must be checked regularly.

2.3 Forecasting with LASSO Overview

Since LASSO has been introduced to the public in 1990, it was highly utilized in various spheres of big data application. Although the topic is fascinating, there is limited empirical findings on the matter in sphere of business and marketing. In addition to business-related spheres, machine learning methods are also used in many other studies such as bioinformatics and epidemiology. On average, it proves to create accurate forecasts in all spheres. In order to illustrate this, a few papers will be discussed.

One of the papers on the somewhat similar topic to be mentioned is a paper by Sagaert et al. that utilized the LASSO model to improve predictions for sales of tires in USA and Europe. The prediction accuracy measure used in this paper was mean absolute percent error (MAPE). The main conclusion of the paper is that LASSO outperforms other forecasts (linear regression and company benchmark), while at the same time the LASSO-created model provides transparency on selected macroeconomic indicators that gives additional market understanding (Sagaert et al., 2017).

The paper by Shi et al. (2016) used the data from the Singapore Ministry of Health and LASSO techniques to predict the weekly occurrence of dengue over a 3-month time horizon. The paper also compared the model estimated by LASSO to seasonal ARIMA models using mean absolute percentage error (MAPE). In this research, it was found that although LASSO-models have their limitations, such as their lack of ability to explain the cause of dengue outbreaks, they provide high prediction accuracy. The outcome of this paper has been put into practice and is now used in Singapore’s dengue control program that indicates the reliability of it (Shi et al., 2016).

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

Another paper by Li and Chen used LASSO-based approaches to forecast twenty macroeconomic variables. In this paper, the goal was to investigate LASSO-based approaches and compare them to a dynamic factor model that is proved to be the best techniques for such tasks as macroeconomic variables forecasting. It was found that, in general, LASSO performs better in terms of out-of-sample than dynamic factor models. It was also concluded that forecast combination can be regarded as another way to enhance dynamic factor models using shrinkage estimation (Li and Chen, 2014).

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

3. Data Collection

The data for this research comes from various sources and includes information on new car sales by brand in Japan, using macroeconomic variables as well as Google Trends data. New car sales in Japan is the variable of interest. The explanatory variables include both macroeconomic variables and Google search data. This research argues that only by including both macroeconomic and Google Trends data, the best forecasting might be obtained. Macroeconomic variables give researchers a better understanding of the country`s economic situation and, therefore, indicates consumers` ability to buy a product. Google Trends data, however, is providing insights on consumers` desire to purchase, as it provides information on preparatory steps of a purchase.

3.1 Japan new cars monthly sales data

The dependent variable in this research is monthly consumer sales of automobiles in Japan. The data on new car sales in Japan is obtained from 日本自動車販売協会連合会 (自販連) which in English is referred to as Japan Automobile Dealer Association (JADA). This includes monthly observations from the year of 2014 to the most current data of March 2018. Initial data contains all the brands sold in Japan, both local and foreign, and in total includes 2196 observations.

Table 3.1.1 Japan new car sales data

Dependent variable	Sample period	Frequency	Source
New car monthly sales	January 2014 – March, 2018	Monthly	Japan Automobile Dealer Association (JADA) http://www.jada.or.jp

The market shares for all the brands were calculated to identify the key players in the Japanese automotive market. As it was determined, 80% of the market has consistently over the years belonged to the following brands: Toyota, Honda, Nissan, Suzuki, and Daihatsu. Toyota is the number one brand, occupying around a third of the market (30%), followed by Honda and Nissan with around 15% of the market, and then Suzuki and Daihatsu both with approximately 11% of the market. This research will focus mainly on the top three players in the Japanese market – Toyota, Honda, and Nissan. For each brand, the total number of observations is 51. Since some of the macroeconomic variable data is missing, three observations were dropped during the models' estimation process, making a total of 48 observations for each brand.

Japanese monthly car sales data was graphed to do a preliminary analysis and possibly identify some outliers or seasonality. The monthly sales of all brands were graphed as well as individual sales of the top three brands in Japan.

Figure 3.1.2 All brands monthly car sales (in 1000s)

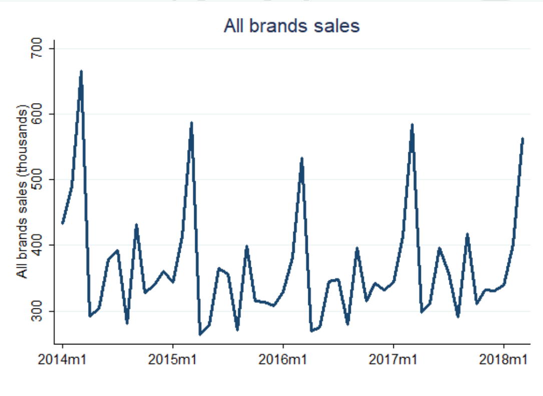


Figure 3.1.2 Toyota monthly car sales (in 1000s)

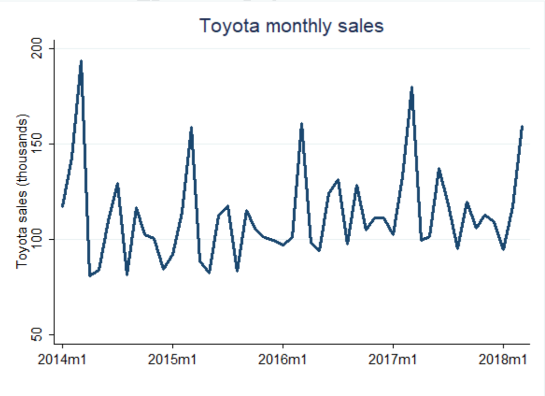


Figure 3.1.3 Nissan monthly car sales (in 1000s)

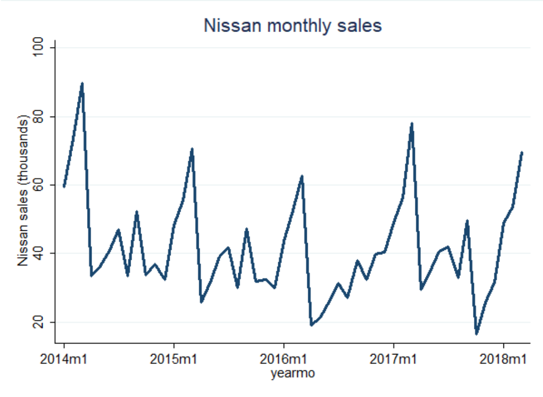
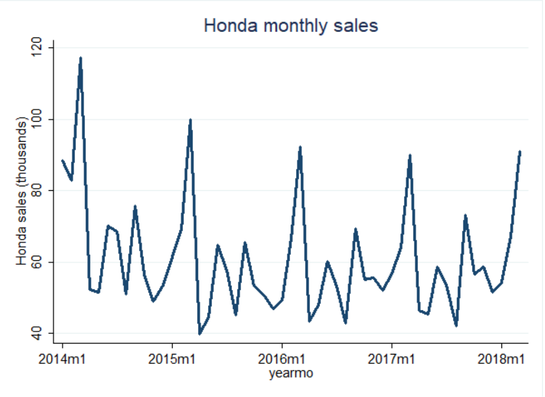


Figure 3.1.4 Honda monthly car sales (in 1000s)



From these figures, there are a few points to be learned. Seasonal behavior of car sales can be observed. It is clear from the figures that the peak of sales for all brands in Japan happens around February-March each year. From the graphs, it appears that data is stationary as there are no «random walks» which was later checked using a statistical test. The Dickey-Fuller test for unit root was carried out for all the brands, and the hypothesis that unit root is present in an autoregressive model was rejected at 1% critical value (See Appendix 1-Appendix 4).

From the figures above, one can also observe certain trends in data; for the aggregate sales of all brands, the trend is downward. Toyota shows a slight upward trend, whereas Honda and Suzuki possess a downward trend.

3.2 Macroeconomic Indicators Data

Data on various macroeconomic indicators were collected. Following such papers by Fantazzini and Toktamysova as well as Muehlen, the following macroeconomic variables were considered as relevant to the car sales forecasting: gross domestic product (GDP), consumer price index for all items in Japan(CPI), working population size (15-64 years old), working for unemployment population, economic uncertainty index, Japan / U.S. foreign exchange rate, Nikkei stock market price, and imported oil price.

Many of the economic variables such as GDP, CPI, working population size and unemployment, exchange rate, stock market price, and oil price are common in various types of economic analysis and do not require detailed explanation. However, it is essential to discuss the economic uncertainty index.

Recently, a significant number of scholars commented on depressing effects of police-related economic uncertainty on economic activity. Uncertainty is believed to play a part in the slow growth or slow recovery from recessions that is because businesses and individuals delay spending and investment until they feel more secured about the future course of policy. The index of economic uncertainty is built from three types of components: newspaper coverage of policy-related economic uncertainty, the number of federal tax code provisions set to expire in future years, as well as disagreement among economic forecasters as a proxy for uncertainty (Baker, Bloom, and Davis, 2011).

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

The data collected for the research can be summarized into the following table.

Table 3.2.1 Macroeconomic variables

Variable	Sample period	Frequency	Standard deviation	Mean	Source
GDP (in billion yen)	January 2014 – December 2017	Q	4626.034	133174.8	E-Stat (Japan government statistics portal site) https://www.e-stat.go.jp
CPI (2010=base year)	January 2014 – February 2018	M	0.8262	103.5172	Organization for Economic Co-operation and Development (OECD) Retrieved from: https://fred.stlouisfed.org/series/JPNCPIALLMINMEI
Working population size (millions)	January 2014 – February 2018	M	0.8883	76.9272	Organization for Economic Co-operation and Development (OECD) Retrieved from: https://fred.stlouisfed.org/series/LFWA64TTJPM647N
Working population unemployment(%)	January 2014 – February 2018	M	0.3752	3.3373	Organization for Economic Co-operation and Development (OECD)

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

					Retrieved from: https://fred.stlouisfed.org/series/LRUN64TTJPA156S
Economic uncertainty index	January 2014 – March 2018	M	32.1403	105.8876	Economic Uncertainty Index http://www.policyuncertainty.com/
Exchange rate	January 2014 – March 2018	M	6.9519	111.7278	Board of Governors of the Federal Reserve System (US) Retrieved from: https://fred.stlouisfed.org/series/DEXJPUS
Stock market price	January 2014 – March 2018	M	2309.754	18284.06	Nikkei Index Website https://indexes.nikkei.co.jp/nkave
Average imported oil price for KL (in thousand yen)	January 2014 – March 2018	M	15.3449	44.5670	Trade Statistics of Japan http://www.customs.go.jp

3.3 Google Trends Data

Google Trends data is publicly available data that is provided by Google Inc., starting from January 1, 2014. Google Trends data are based on Google searches and provides the information on how often a particular search term is entered in different parts of the world in different languages. A researcher can easily collect Google Trends data by clicking the link: https://trends.google.com/trends/. The index for a particular search term can be obtained for a custom period and frequency, as well as for different geographical positions (including areas within the same country). The output is presented in the form of a diagram, and the dataset can be downloaded in CSV format.

It is important to note that Google Trends presents the index that is proportional to the time and location rather than raw data. The process of creating this index is as follows: each data point (search-term) is divided by the total searches of this geographical location and time range it represents, and then this number is scaled on a range from 0 to 100 on a topic's proportion to all searches on all subjects. In total, there are 30 categories at the first level and around 250 categories at the second level (Google Inc., 2018).

Google Trends allows the researcher to look at certain search term popularity over specific periods of time and different geographical locations. It might allow one to reveal trends or sudden changes in consumer preferences. Therefore, Google Trends has recently been recognized as one of the most popular tools in business.

Moreover, relevant Google Trends indexes in the Japanese language were collected in geographical position of Japan starting from January 2014 to March 2018.

Brand	Google search	Sample period	Frequency
All brands	車を買う	January 2014 – March 2018	M
Toyota	トヨタ	January 2014 – March 2018	M
Honda	ホンダ	January 2014 – March 2018	M
Nissan	日産	January 2014 – March 2018	M

Table 3.3.1 Google search terms

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

The Google search data for 3-5 brands and all brands sales were graphed against new car sales data in Japan.

Figure 3.3.2 All brands sales vs. Buy car index

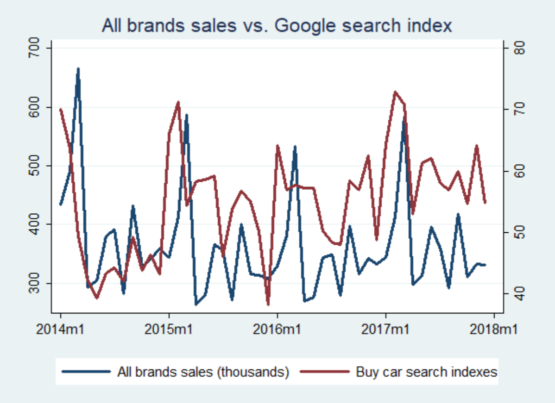


Figure 2.3.3 Toyota sales vs. Toyota search index

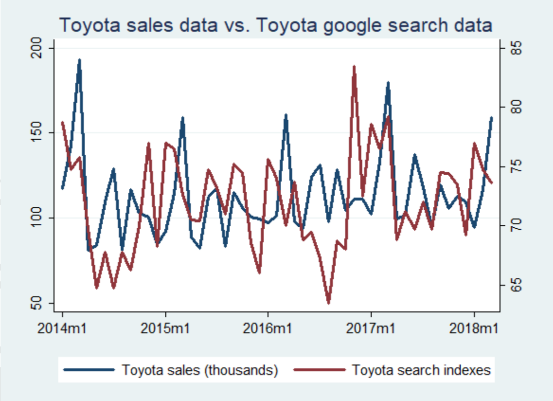


Figure 3.3.4 Nissan sales vs. Nissan search index

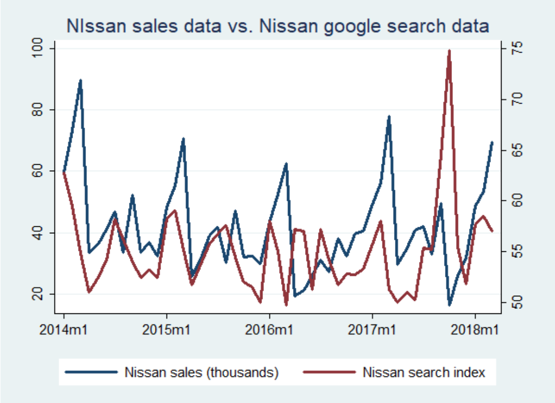
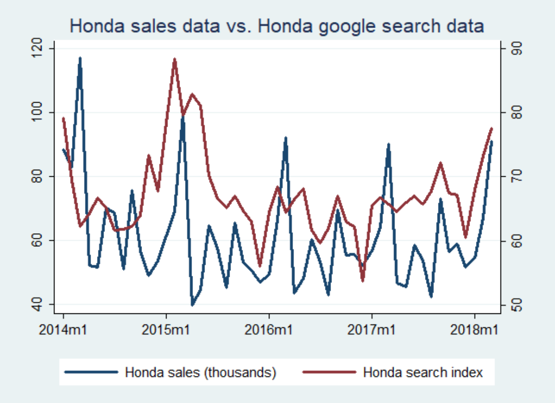


Figure 3.3.5 Honda sales vs. Honda search index



The graphs of the Google Trends data and monthly sales data, for the most part, share the same direction. Google search index appears to be leading, while car sales seem to be lagging in their response to Google Trends data (first people do the research and then complete the purchase), which is in accordance with the literature review.

From the figures, it appears that, for the sales of all brands and Toyota sales, Google data predict the directions of the sales accurately. Nissan sales and Google search index data fit the best among all presented brands.

4. Methodology

This section provides the overview of tools used in this paper for predictions, as well as for prediction accuracy measurements. To fulfill the

4.1 Human Judgement Model Construction

Following Choi and Varian (2012), the superior model (Model 1) for this paper is simple autoregressive model (output variable is determined by its values in the past) or AR model, for short, where sales are predicted using sales data lagged 1 and 12 months accordingly (Choi and Varian, 2012). Following the paper by Hand and Judge on predicting cinema sales admissions, trend variable and seasonal dummies were also included in the model (Hand and Judge, 2012). The model takes the following form:

$y_t = \beta_0 + \beta_1 y_{t-1} + \beta y_{t-12} + \gamma \mathit{TREND} + \sum_{j=0}^{11} \partial_j S_j$ (*Model 1),* where y_t is the dependent variable and y_{t-1} is the 1-month lagged dependent variable, while y_{t-12} is the lagged 12-month dependent variable; *TREND* is trend variable and S_j are month dummy variables. The term c refers to white noise (error term).

In their paper, Choi and Varian (2012) just examined how the inclusion of Google Trends data might affect forecast accuracy without including any macroeconomic variables. The results showed that Google search data has some positive influence on forecast accuracy.

This paper will try to take a step forward and try to improve the forecast, both by including the macroeconomic variable with best predictive power and Google Trends data. Including economic variables might help to identify the economic conditions for buying the car. Google Trends data might help to reveal a consumer's preference. These are the critical components of the purchases.

Therefore, the analysis will be done in two steps. First, individual macroeconomic variables and Google Trends data will be added to the baseline model to analyze which macroeconomic variable will better help the in-sample prediction accuracy. In other words, find which of those macroeconomic variables can be called "the best performing macroeconomic variable". After, the key macroeconomic variable that better helps capture

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

the economic environment will be identified (the best performing economic variable). Further, the forecast can be improved by adding the Google Trends data.

The following models are constructed:

$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12} + \beta_3 x_{t-1} + \beta_4 x_{t-12} + \gamma TREND + \sum_{j=0}^{10} \partial_j S_j$ (*Model 2),* where x is CPI, x_{t-1} and x_{t-12} are lagged 1 and 12 months accordingly CPI.

$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12} + \beta_3 d_{t-1} + \beta_4 d_{t-4} + \gamma TREND + \sum_{j=0}^{10} \partial_j S_j$ (*Model 3),* where d is GDP, d_{t-1} and d_{t-4} are lagged 4 months and 12 months accordingly GDP (GDP data is quarterly data).

$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12} + \beta_3 f_{t-1} + \beta_4 f_{t-12} + \gamma TREND + \sum_{j=0}^{10} \partial_j S_j$ (*Model 4),* where f is the exchange rate, f_{t-1} and f_{t-12} are lagged 1 and 12 months accordingly exchange rate.

$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12} + \beta_3 g_{t-1} + \beta_4 g_{t-12} + \gamma TREND + \sum_{j=0}^{10} \partial_j S_j$ (*Model 5), w*here g is working population size, g_{t-1} and g_{t-12} are lagged 1 and 12 months accordingly working population size.

$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12} + \beta_3 h_{t-1} + \beta_4 h_{t-12} + \gamma TREND + \sum_{j=0}^{10} \partial_j S_j$ (*Model 6), w*here h is unemployment within the working population, h_{t-1} and h_{t-12} are lagged 1 and 12 months accordingly unemployment.

$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12} + \beta_3 k_{t-1} + \beta_4 k_{t-12} + \gamma TREND + \sum_{j=0}^{10} \partial_j S_j$ (*Model 7), w*here k is economic uncertainty level, k_{t-1} and k_{t-12} are lagged 1 and 12 months accordingly economic uncertainty.

$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12} + \beta_3 l_{t-3} + \beta_4 l_{t-12} + \gamma TREND + \sum_{j=0}^{10} \partial_j S_j$ (*Model 8),* where l is Google search index, l_{t-3} and l_{t-12} are lagged 3 and 12 months accordingly Google search index. It was decided to include the lagged 3 months Google Trends variable due to extensive literature that indicates that the decision-making process for cars, on average, can take up to three months (Brooks, 2016; Gevelber, 2016).

$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12} + \beta_3 n_{t-1} + \beta_4 n_{t-12} + \gamma TREND + \sum_{j=0}^{10} \partial_j S_j$ (*Model 9), w*here n is stock market price, n_{-1} and n_{t-12} are lagged 1 and 12 months accordingly stock market price.

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12} + \beta_3 m_{t-1} + \beta_4 m_{t-12} + \gamma \mathit{TREND} + \sum_{j=0}^{10} \partial_j S_j$ (*Model 10),* where m is oil price, m_{t-1} and m_{t-12} are lagged 1 and 12 months accordingly oil price.

Since the aim of this thesis is to test the concept that the inclusion of Google Trends data can further improve forecast for Japanese car sales, the Google Trends data will be added to the best performing model. Therefore, the final model will be as follows:

$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12} + \beta_3 \&_{t-1} + \beta_4 \&_{12} + \beta_5 l_{t-3} + \beta_6 l_{t-12} + \gamma \mathit{TREND} + \sum_{j=0}^{10} \partial_j S_j$ (*Model 11),* where & is the best predicting variable, $\&_{t-1}$ and $\&_{t-12}$ are lagged 1 and 12 months accordingly best predicting variables; l is Google search index, l_{t-1} and l_{t-12} are lagged 1 and 12 months accordingly search index.

The specified above models will be estimated from years 2014 to 2016, leaving out 2017 to further estimation of out-of-sample prediction.

4.2 Machine Learning Model

Machine learning methods that are also frequently addressed as data learning or statistical learning, use data-driven algorithms to perform a task. The primary goal of machine learning is making a prediction.

In this paper, Least Absolute Shrinkage and Selection Operator (LASSO) regression will be used. This is supervised learning since both output and input variables are known. LASSO is a technique that was proposed in the 1990s and has attracted much attention to the problem of “small *n* large *p*” (Tibshirani, 1996).

The idea behind LASSO is that it minimizes the residual sum of squares while at the same time penalizing the model size. Therefore, LASSO will shrink unimportant parameters towards zero. Also, the size of this penalty depends on tuning parameter size – lambda size.

Cross-validation selects optimal lambda and consequently optimal penalty size. The process of cross-validation is performed as follows: the data is separated into k subsets. Therefore, the process is performed k times. For each time, one validation dataset (data is pretended to be unknown) is identified while others become training dataset. The training dataset is used to train the data and estimate model, to make predictions for the validation dataset and estimate the prediction error. Then, for each tuning parameter value, the average of those k experiments is estimated, and the cross-validation error curve can be constructed. The optimal lambda will be the one that minimizes this cross-validation curve (Yang, 2018).

Since LASSO can identify the most essential, variables greatly associated with the output variable, LASSO shrinks unimportant parameters to zero. This means that LASSO can perform variables selection. It is also able to do regularization, which is to prevent overfitting. Therefore, by selecting the critical covariates and selecting a regression model that fits the data best, LASSO is performing the model selection task.

LASSO is usually used when the number of regressors is vast, to select a subset of variables in linear regressors. In this paper, the total number of variables to choose from is 9, and with all lags included – 108 variables. It is considered to be a large number of observations when compared to the number of observations of car sales – which is in total 48. Therefore, LASSO will provide help in picking the best, the most strongly correlated with the output variable covariates and their lags. Moreover, to prevent the problem of subjective approach when a human selects the model.

4.3 Model Prediction Accuracy Measurements

It is necessary to compare the estimated models and their performance. Since there is no consensus on what is the best measure for assessing forecast accuracy, it was decided to extend beyond a single technique. There are a few criteria used to assess forecasting model performance, such as r^2, root mean square error (RMSE), and mean absolute error (MAE) been the most commonly used ones (Diebold, 2017).

The first indicator to compare the created models is r^2 or, as it is also known, the coefficient of determination. The formula of r^2 is presented below:

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

In this model, SS_{res} is the sum of squared residuals while SS_{tot} is the total sum of squares. r^2 shows how close the fitted values are to the actual data. In other words, how well-observed values are replicated by the constructed model. This measure takes a number from 0% to 100%. The higher r^2 is, the better it explains the variability of data. An adjusted r^2 is the coefficient of determination adjusted for degrees of freedom. Adjusted r^2 helps to compare the models with a different number of predictors. In this research r_2 will be utilized in in-sample prediction.

Root Mean Square Error (RMSE) is a frequently used measure of the differences between values that were predicted by the model and the values that were observed. RMSE is a standard deviation of the differences between predicted values and actually observed values (residuals). In other words, it is the square root of the average of squared differences between estimated values and actually observed data. The formula of RMSE is as follows:

$$RMSE = \sqrt{\sum_{j=1}^{n} \frac{(y_j^{\wedge} - y_j)^2}{N}}$$

In this formula, $y_j^{\wedge}$ are predicted values and y_j are actually observed values. Therefore, it is square root of an averaged error.

The perception behind the formula is that RMSE shows how to spread out/close to the actual data the residuals are. RMSE is expressed in the units of the variable of interest, which is, in this case, cars, and can take values from 0 to an endless number. The smaller RMSE means a smaller magnitude of residuals, thus, giving a model that produces a better fit.

Mean absolute error (MAE) is similar to RMSE measure of accuracy forecast; a measure between 2 continuous variables. It is the average of absolute differences between fitted values and actually observed values. Therefore, the formula can be written as follows:

$$MAE = \frac{1}{n}\sum_{j=1}^{n} \left| y_j^{\wedge} - y_j \right|$$

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

In this formula, $y_j^{\wedge}$ are predicted values and y_j are actually observed values.

MAE is also presented in the units of the variable of interest and can take any positive value. This means that the perception behind it is similar to RMSE; the smaller MAE indicates the model that predicts the values better. However, in the case of RMSE, since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors (MAE and RMSE—Which Metric is Better? 2016)

Another technique used is uncertainty coefficient or U Theil's statistic.

$$U=\frac{\left[\frac{1}{n}\sum_{i=1}^{n}(A_i-P_i)^2\right]^{1/2}}{\left[\frac{1}{n}\sum_{i=1}^{n}A_i^2\right]^{1/2}+\left[\frac{1}{n}\sum_{i=1}^{n}P_i^2\right]^{1/2}}$$

If U Theil`s Statistic is equal to 1, it means that the proposed model is as good as the naïve (the forecast that is set to be the value of the last observation) model. If U is greater than 1, there is the estimated forecasting model should not be used, because a naïve method would have better results. One should consider using the constructed model only when U is smaller than 1 and the smaller the number of U statistics is, the better the proposed model is (Small and Wong, 2001).

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

5. Results

Monthly sales of all brands and three individual brands were investigated. Therefore, the results are presented four different times for the human judgment model as well as machine learning model.

5.1 Choosing Model by Human Judgement

For monthly total sales including all brands, the results are the following:

Model	Adjusted R^2	RMSE	MAE	U Theil`s
1	0.9657	8891.2976	7370.1126	0.09974967
2	0.9794	6075.1828	4523.4922	0.06615976
3	0.9783	6236.7172	4817.5938	0.06936932
4	**0.9839**	**5367.0408**	**4452.7109**	**0.05662949**
5	0.9715	7154.5193	6216.7995	0.08396972
6	0.9657	7841.0698	7023.1198	0.08626346
7	0.9630	8147.3719	7526.3333	0.09127866
8	0.9732	6928.8331	6065.446	0.0729507
9	0.9690	7451.1372	6685.9661	0.08299438
10	0.9624	8206.2731	6377.4766	0.08945232
11	**0.9866**	**4149.9483**	**3648.8125**	**0.04486308**

Note: Model 1 to 11 are estimated on regressions described in 4.1 of this paper. Where Model 1 is simple autoregressive model with seasonal dummies and trend variable: $y_t = c + \alpha y_{t-1} + \beta y_{t-12} + \eta x_{t-1} + \delta x_{t-12} + \gamma \text{TREND} + \sum_{j=0}^{10} \partial_j S_j$, Models 2-10 consecutively include different macroeconomic variables and Google Trends variable (8) and the final model (11) includes both Google Trends data and the best performing macroeconomic variable. The sample period of data is from January 2014 to December 2016.

The superior auto-regressive model with seasonal dummies and trend variable (Model 1) in case of all brands in Japan features high r^2 of 0.9657. However, it also has high prediction errors. The further inclusion of macroeconomic variables generally helps to improve prediction errors and to increase r^2, except for the economic uncertainty index

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

(Model 7) and oil prices (Model 10), where improved prediction errors go along with slightly decreased r^2.

The best performing macroeconomic variable, in this case, is believed to be exchange rate, since the model including exchange rate (Model 4) shows the smallest in-sample prediction errors and the highest r^2. In the final model (Model 11), the inclusion of both exchange rate and Google search index as well as their lags as specified in 4.1 of this paper, leads to even further improvement of in-sample prediction based on prediction errors and r^2.

For Toyota, the estimated results are as follows:

Model	Adjusted R^2	RMSE	MAE	U Theil`s
1	0.9358	3120.5059	2444.9355	.11523464
2	0.9352	2764.202	2227.4193	0.10355454
3	0.9236	3002.3457	2470.6263	0.11091421
4	0.9329	2814.7596	2039.1934	0.10571379
5	0.9191	3089.0055	2490.7689	0.11355038
6	0.9294	2886.9892	2322.6686	0.10348007
7	**0.9604**	**2161.6419**	**1704.4303**	**0.07832322**
8	0.9585	2213.906	1726.8359	0.07754637
9	0.9467	2507.9491	1707.8757	0.09792469
10	0.9642	2055.7552	1556.6022	0.08514566
11	**0.9720**	**1537.2156**	**1042.7324**	**0.05451605**

Note: Model 1 to 11 are estimated on regressions described in 4.1 of this paper. Where Model 1 is simple autoregressive model with seasonal dummies and trend variable: $y_t = c + \alpha y_{t-1} + \beta y_{t-12} + \eta x_{t-1} + \delta x_{t-12} + \gamma TREND + \sum_{j=0}^{10} \partial_j S_j$, Models 2-10 consecutively include different macroeconomic variables and Google Trends variable (8) and the final model (11) includes both Google Trends data and the best performing macroeconomic variable. The sample period of data is from January 2014 to December 2016.

In the case of Toyota sales in Japan, the prediction of the superior auto-regressive model (Model 1) with seasonal dummies and trend variable can generally be improved by the inclusion of macroeconomic variables, except for the working population size (Model 5).

The best performing macroeconomic variable, in this case, is economic uncertainty. Even though the model including economic uncertainty (Model 7) does not show the highest r2, prediction errors are the lowest among all the models. In the final model (Model 11), the inclusion of economic uncertainty index and Google search index for Toyota as well as their lags as specified in 4.1 of this paper, leads to even further improvement of in-sample prediction based on prediction errors and r^2.

For Nissan, the estimated results are as follows:

Model	Adjusted R^2	RMSE	MAE	U Theil`s
1	0.9130	2317.5167	1981.8433	0.24178847
2	0.8973	2220.6793	1938.4844	0.21876266
3	0.8944	2251.9049	1969.1309	0.23286921
4	**0.9522**	**1514.4012**	**1079.0927**	**0.14679025**
5	0.9106	2071.8976	1770.8685	0.2136966
6	0.9314	1814.1077	1362.0332	0.18828941
7	0.9088	2092.6813	1763.0156	0.22316873
8	0.9277	1863.4253	1537.5457	0.17487727
9	0.8983	2209.4937	1897.6986	0.2275084
10	0.9196	1964.1563	1696.2917	0.20615934
11	**0.9676**	**1054.3631**	**839.83049**	**0.09474242**

Note: Model 1 to 11 are estimated on regressions described in 4.1 of this paper. Where Model 1 is simple autoregressive model with seasonal dummies and trend variable: $y_t = c + \alpha y_{t-1} + \beta y_{t-12} + \eta x_{t-1} + \delta x_{t-12} + \gamma TREND + \sum_{j=0}^{10} \partial_j S_j$*, Models 2-10 consecutively include different macroeconomic variables and Google Trends variable (8) and the final model (11) includes both Google Trends data and the best performing macroeconomic variable. The sample period of data is from January 2014 to December 2016.*

In the case of Nissan sales in Japan, the inclusion of CPI (Model 2), GDP (Model 3), working population size (Model 5) or stock market prices (Model 9) variables into the superior model (Model 1) leads to a decrease of in-sample prediction errors. Although, r^2

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

decreases as well. In all other models, the inclusion of macroeconomic variables means an increase of in-sample prediction accuracy and higher r^2.

The best performing variable based on the table presented above is exchange rate, which model (Model 4) features both the highest r^2 and smallest prediction errors. When both exchange rate, with its lags, as well as Google index for Nissan, with its lags, included into the superior model (Model 11), the in-sample prediction errors decrease significantly as well as r^2 increases.

For Honda, the estimated results are as follows:

Model	Adjusted R^2	RMSE	MAE	U Theil`s
1	0.9591	1813.0047	1519.4993	0.11105195
2	0.9628	1523.9422	1347.3687	0.09772168
3	0.9479	1804.1269	1495.9661	0.10755163
4	**0.9629**	**1523.333**	**1273.8519**	**0.09386143**
5	0.9516	1739.0762	1439.4424	.010806475
6	0.9510	1749.727	1471.3802	0.1040244
7	0.9562	1653.8004	1458.1816	0.10292646
8	0.9566	1647.2093	1472.2236	0.10143567
9	0.9517	1737.5716	1522.8551	0.10458292
10	0.9589	1602.3587	1417.4053	0.09385221
11	**0.9876**	**743.8478**	**590.88965**	**0.03852638**

Note: Model 1 to 11 are estimated on regressions described in 4.1 of this paper. Where Model 1 is simple autoregressive model with seasonal dummies and trend variable: $y_t = c + \alpha y_{t-1} + \beta y_{t-12} + \eta x_{t-1} + \delta x_{t-12} + \gamma TREND + \sum_{j=0}^{10} \partial_j S_j$*, Models 2-10 consecutively include different macroeconomic variables and Google Trends variable (8) and the final model (11) includes both Google Trends data and the best performing macroeconomic variable. The sample period of data is from January 2014 to December 2016.*

Honda's superior model (Model 1) performance is rather good and comparable to all other models.

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

The best performing economic variable is the same as in the case of Nissan - exchange rate - because the model with the exchange rate (Model 4) shows the highest r2 and the lowest prediction errors. Therefore, the best model (Model 11) is estimated including both exchange rate, with its lags, and Google Trends for Nissan, with its lags. This model boosts the r^2 almost to 99% and makes the in-sample prediction errors smaller than a thousand units.

5.2 Choosing Model by Machine Learning Method

This section will discuss that are selected by machine learning method - LASSO regression with all macroeconomic variables available and Google Trends data, as well as their lags.

The model specified by LASSO for sales of all brands with the highest r^2 and the same size model as in part 5.1 is the following:

$$y_t = \beta_0 + \beta_1 y_{t-12} + \beta_2 x_{t-1} + \beta_3 x_{t-7} + \beta_4 x_{t-8} + \beta_5 x_{t-10} + \beta_6 l_{t-3} + \beta_7 l_{t-5} + \beta_8 l_{t-11} + \beta_9 k_{t-5} + \beta_{10} k_{t-6} + \beta_{11} k_{t-12} + \beta_{12} h_{t-3} + \beta_{13} h_{t-4} + \beta_{14} h_{t-6} + \beta_{15} h_{t-10} + \beta_{16} S_3 + \beta_{17} S_9 + \beta_{18} S_{11}.$$

In this model:

- y is car sales in Japan, y_{t-12} is the sale of cars lagged 12 months;
- x is CPI, x_{t-1}, $x_{t-7,}$ $x_{t-8,}$ x_{t-10}, are lagged 1, 7, 8 and 10 months accordingly for CPI;
- l is Google Trends search index, l_{t-3}, $l_{t-5,}$ l_{t-11} are lagged 3, 5 and 11 months accordingly for Google search index;
- k is economic uncertainty level, $k_{t-5,}$ k_{t-6}, and k_{t-12} are lagged 5, 6 and 12 months accordingly for economic uncertainty;
- h is unemployment within working population, $h_{t-3,}$ h_{t-4}, h_{t-6} and h_{t-10} are lagged 3, 4, 6 and 10 months accordingly for unemployment in working population;
- S_3, S_9 and are S_{11} month dummies;
- β_0 is the constant;

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

The in-sample prediction errors for this model are:

Adjusted R2	RMSE	MAE	U Theil`s
0.9340	6655.4699	5335.0625	0.07773667

Note: Sample period for this model is 2014 to 2016.

The model specified by LASSO for sales of Toyota with the same size model as in part 5.1 is the following:

$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-9} + \beta_3 y_{t-10} + \beta_4 y_{t-12} + \beta_5 n_{t-1} + \beta_6 n_{t-8} + \beta_7 l_{t-1} + \beta_8 l_{t-5} + \beta_9 l_{t-11} + \beta_{10} k_{t-5} + \beta_{11} h_{t-6} + \beta_{12} h_{t-10} + \beta_{13} S_2 + \beta_{14} S_3 + \beta_{15} S_6 + \beta_{16} S_7 + \beta_{17} S_8 + \beta_{18} S_9$.

In this model:

- y is car sales in Japan, y_{t-1}, y_{t-9}, y_{t-10}, y_{t-12} are sales of Toyota cars lagged 1, 9, 10 and 12 months accordingly for sales of Toyota;
- n is stock market price, n_{t-1} and n_{t-8} are lagged 1 and 8 months accordingly for stock market prices;
- l is Google Trends search index for Toyota, l_{t-1}, lt$_{-5}$, and l_{t-11} are lagged 1, 5 and 11 months accordingly for Google Trends search index for Toyota:
- k is economic uncertainty level and k_{t-5},is lagged 5 months for economic uncertainty;
- h is unemployment within working population, h_{t-6} and h_{t-10} are lagged 6 and 10 months accordingly for working population unemployment;
- S_2, S_3, S_6, S_7, S_8, and S_9 are month dummies;
- β_0 is the constant;

The in-sample prediction errors for this model are:

Adjusted R^2	RMSE	MAE	U Theil`s
0.9128	2244.6288	1952.6689	0.08397754

Note: Sample period for this model is 2014 to 2016.

The model specified by LASSO for Nissan with the same size model as in part 5.1 is the following:

$$y_{t} = \beta_0 + \beta_1 y_{t-5} + \beta_2 y_{t-12} + \beta_3 x_{t-5} + \beta_4 x_{t-6} + \beta_5 x_{t-8} + \beta_6 n_{t-4} + \beta_7 l_{t-3} + \beta_8 l_{t-6} + \beta_9 l_{t-7} + \beta_{10} k_{t-1} + \beta_{11} k_{t-2} + \beta_{12} k_{t-3} + \beta_{13} k_{t-6} + \beta_{14} \varphi k_{t-12} + \beta_{15} h_{t-8} + \beta_{16} S_3 + \beta_{17} S_4 + \beta_{18} S_5.$$

In this model:

- y is car sales in Japan, y_{t-5} and y_{t-12} are sales of Nissan cars lagged 5 and 12 months accordingly for sales of Nissan;
- x is CPI, x_{t-5}, $x_{t-6,}$ and x_{t-8} are lagged 5, 6 and 8 months accordingly for CPI;
- n is stock market price, and n_{t-4} is lagged 4 months lagged for stock market prices;
- l is Google Trends search index for Nissan, l_{t-3}, $l_{t-6,}$ and l_{t-7} are lagged 3, 6 and 7 months accordingly for Google Trends search index for Nissan;
- k is economic uncertainty level, $k_{t-1,}$ $k_{t-2,}$ k_{t-3}, kt_{-6}, and k_{t-12} are lagged 1, 2, 3, 6 and 12 months accordingly for economic uncertainty;
- h is unemployment within working population and h_{t-8} is lagged 8 months for working population unemployment;
- S_3, S_4, and S_5 are month dummies;
- β_0 is the constant;

The in-sample prediction errors for this model are:

Adjusted R^2	RMSE	MAE	U Theil`s
0.8926	1388.3437	1126.3446	0.14951881

Note: Sample period for this model is 2014 to 2016.

The model specified by LASSO for the sales of Honda with the highest r^2 and the same size model as in part 5.1 is the following:

$$y_{t} = \beta_0 + \beta_1 y_{t-7} + \beta_2 y_{t-8} + \beta_3 y_{t-10} + \beta_4 y_{t-12} + \beta_5 n_{t-4} + \beta_6 x_{t-4} + \beta_7 x_{t-8} + \beta_8 l_{t-3} + \beta_9 l_{t-12} + \beta_{10} k_{t-3} + \beta_{11} k_{t-7} + \beta_{12} k_{t-11} + \beta_{13} k_{t-12} + \beta_{14} h_{t-3} + \beta_{15} h_{t-6} + \beta_{16} S_3 + \beta_{17} S_4 + \beta_{18} S_8.$$

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

In this model:

- y is car sales in Japan, y_{t-7}, y_{t-8}, $y_{t-10,}$ and y_{t-12} are sales of Honda cars lagged 7, 8, 10 and 12 months accordingly for sales of Honda;
- n is stock market price, and n_{t-4} is lagged 4 months for stock market price;
- x is CPI, x_{t-4}, x_{t-8} are lagged 4 and 8 months accordingly for CPI;
- l is Google Trends search index for Honda, l_{t-3} and l_{t-12} are lagged 3 and 12 months accordingly for Google Trends search index for Honda;
- k is economic uncertainty level, k_{t-3}, kt_{-7}, k_{t-11} and k_{t-12} are lagged 3, 7, 11 and 12 months accordingly for economic uncertainty;
- h is unemployment within working population, h_{t-3} and h_{t-6} are lagged 3 and 6 months accordingly for working population;
- S_3, S_4, S_8 are month dummies;
- β_0 is the constant;

The in-sample prediction errors for this model are:

Adjusted R^2	RMSE	MAE	U Theil`s
0.9647	866.7311	721.18408	0.0548786

Note: Sample period for this model is 2014 to 2016.

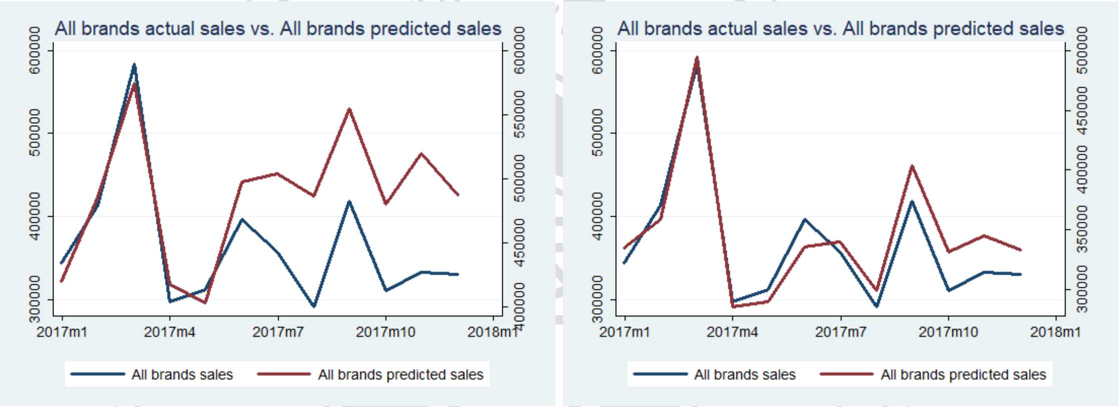
DOI:10.6814/THE.NCCU.IMES.001.2018.F06

5.3 Models Comparison: Machine and Human Models

Out of sample prediction errors for aggregate sales of all brands in Japan:

Model type	RMSE	MAE	U Theil`s
Human-picked	132 520.27	121 831.17	1.4458651
Machine-picked	**37 070.743**	**26 910.693**	**0. 38773194**

Figure 5.3.1 All brands sales comparison: actual and predicted (Human model on the left; Machine model on the right)



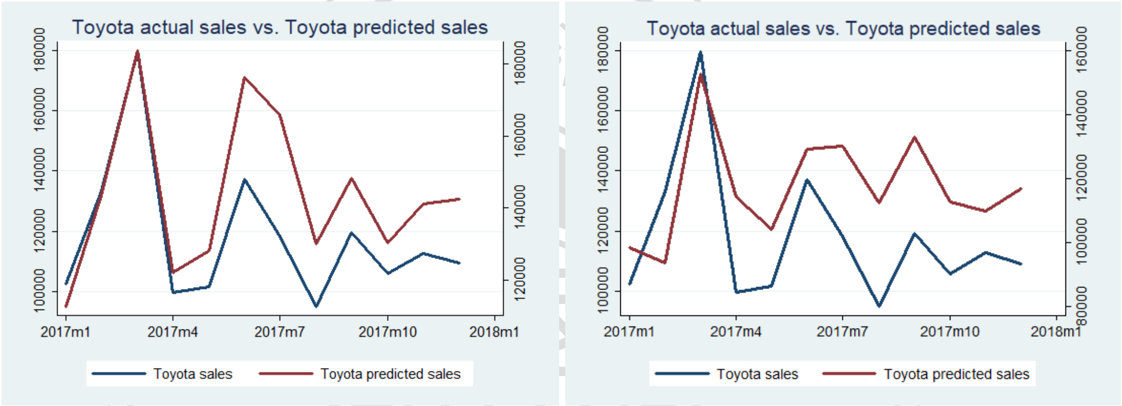
For the sales of all brands in Japan, the machine model (LASSO) performs better as it has lower prediction errors, as well as reproduces the form of the sales better. The human model shows prediction in terms of U Theil`s has a worse performabce than naïve model and has high MAE and RMSE. Fitted values with human model significantly overpredict real sales value

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

Out of sample prediction errors for Toyota:

Model type	RMSE	MAE	U Theil`s
Human-picked	28 604.21	25 905.334	1.0527726
Machine-picked	**16 642.774**	**12 973.738**	**0. 61967838**

Figure 5.3.2 Toyota sales comparison: actual and predicted (Human model on the left; Machine model on the right)



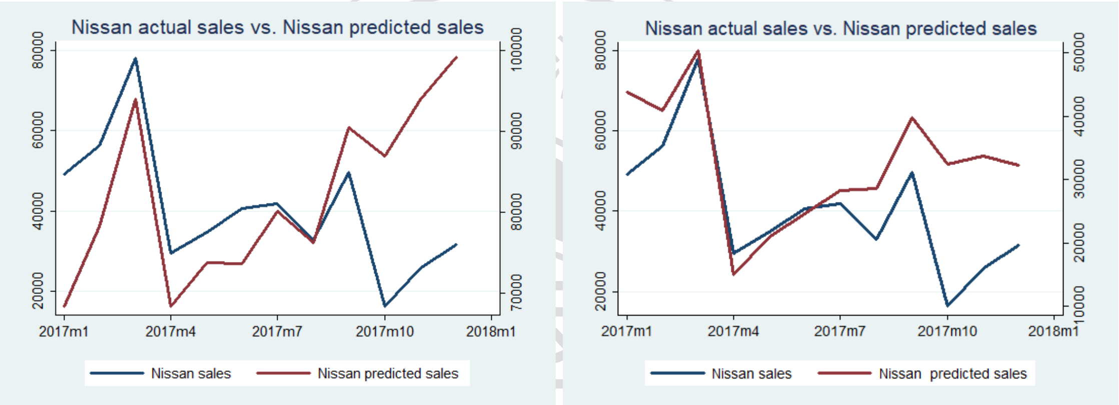
As it can be seen from the results presented in the graph for Toyota, the machine-picked model has smaller out-of-sample prediction errors. From the graph, one can observe that the direction and the form of sales predicted with the human picked model is better, though according to U Theil`s statistics, shows performance that is as good as using last period`s actuals. Generally, both models overpredict sales.

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

Out of sample prediction errors for Nissan:

Model type	RMSE	MAE	U Theil`s
Human-picked	45 093.294	41 388.391	4.2852056
Machine-picked	**13 862.761**	**12 096.133**	**0 .87327995**

Figure 5.3.3 Nissan sales comparison: actual and predicted (Human model on the left; Machine model on the right)



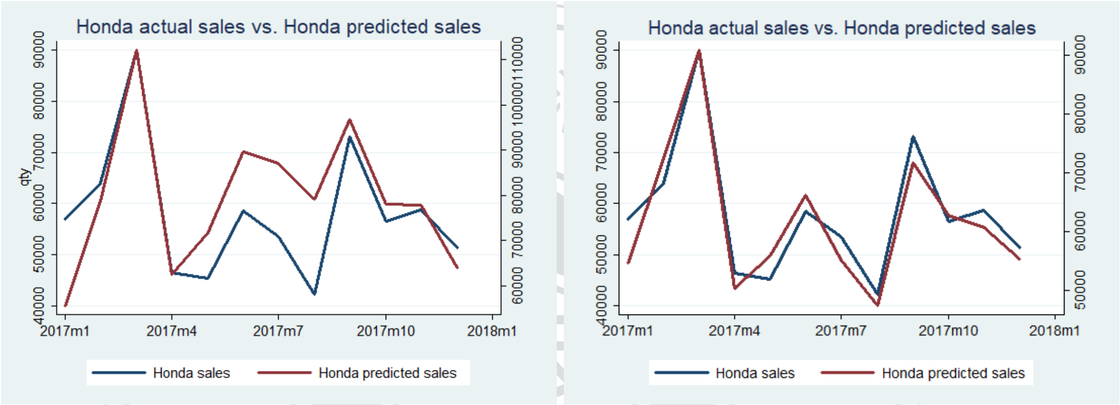
In the case of Nissan, LASSO model has smaller estimation errors. The human constructed model is performing poorly concerning U Theil`s statistics. LASSO model is mispredicting the direction of the sales starting around July 2017, Human model over predict the sales significantly.

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

Out of sample prediction for Honda:

Model type	RMSE	MAE	U Theil`s
Human-picked	23 602.845	21 598.652	1.4271402
Machine-picked	**5433.4728**	**4515.6465**	**0.3346929**

Figure 5.3.4 Honda sales comparison: actual and predicted (Human model on the left; Machine model on the right)



For Honda, judging from both graphs and tables, the machine learning method model (LASSO regression) performs the best, providing the more accurate result regarding prediction errors and the direction of sales.

Overall, based on the comparison above, the models constructed using machine learning method (LASSO regression) perform better than the human created models.

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

5.4 Further Models Comparison

Additionally, to evaluate out of sample prediction, the superior model of this paper was used as well as the model only including best performing macroeconomic variable and LASSO regression with optimal lambda.

As specified in 4.1 of this paper, variables were added gradually to the superior model. That is, added the best performing (regarding in-sample prediction) macroeconomic variable and then include Google Trends data to create the final human picked model. That is Model 11 in 5.1 of this paper, or best performing human-picked model. Here, on the contrary, this best performing human-picked model (Model A) will be degraded. First, the Google trends data will be excluded (Model B), then the best performing macroeconomic variable will be taken out (Model C). Finally, the model will be estimated using the optimal lambda (Model D).

The model estimated by LASSO with optimal lambda (Model D) for sales of all brands in Japan takes the following form:

$y_t = \beta_0 + \beta_1 y_{t-10} + \beta_2 y_{t-12} + \beta_3 x_{t-10} + \beta_4 l_{t-5} + \beta_5 k_{t-5} + \beta_6 k_{t-6} + \beta_7 k_{t-7} + \beta_8 h_{t-3} + \beta_9 h_{t-6} + \beta_{10} h_{t-9} + \beta_{11} h_{t-10} + \beta_{12} S_3 + \beta_{14} S_9$.

In this model, variables, including constant, there are in total 14:

- y is car sales in Japan, y_{t-10} and y_{t-12} are sales of all brands cars lagged 10 and 12 months accordingly for sales of all brands;
- x is CPI, and x_{t-10} is lagged 10 months for CPI;
- l is Google Trends search index, and l_{t-5} is lagged 5 months for Google Trends search index;
- k is economic uncertainty, k_{t-5}, k_{t-6}, and k_{t-7}, are lagged 5, 6, and 7 months accordingly for economic uncertainty;
- h is unemployment within working population, h_{t-3}, h_{t-6}, h_{t-9}, and h_{t-10}, are lagged 3, 6, 9, and 10 months accordingly for unemployment within working population;
- S_3 and S_9 are month dummies;

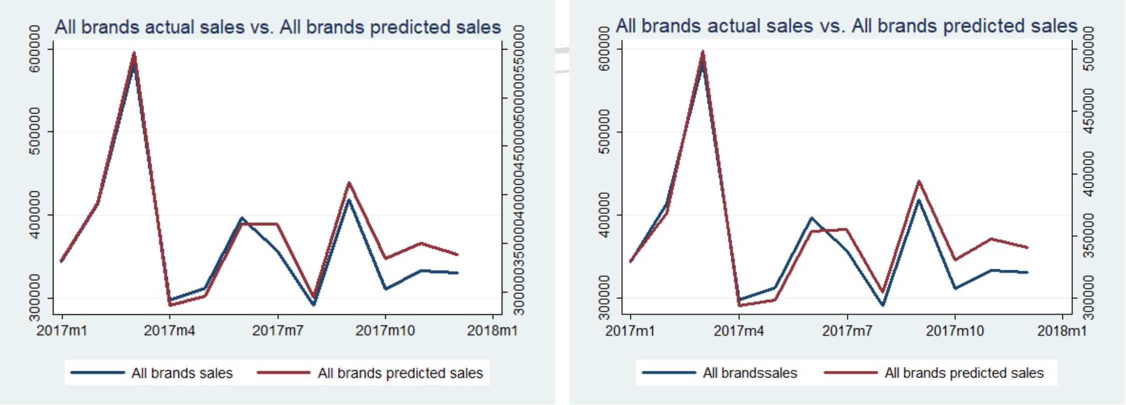
- β_0 is the constant;

Out of sample prediction errors for aggregate sales of all brands:

#	Model type	RMSE	MAE	U Theil`s
A	With Macro Variable and Google Trends	132 520.27	121 831.17	1.4458651
B	With Macro Variable and without Google Trends	41 831.104	36 352.633	0 .41678146
C	**Superior Model (AR)**	**18 686.983**	**16 267.096**	**0.19618075**
D	LASSO with optimal lambda	32 954.319	24 032.323	0.33973458

Based on the table above, in the case of all brands sales, the best model is the autoregressive model with seasonal dummies and trend variable. This being the superior model of this paper (Model C).

Figure 5.4.1 All brands sales comparison: actual and predicted (Superior (AR) model on the left; LASSO with opt. lambda model on the right)



DOI:10.6814/THE.NCCU.IMES.001.2018.F06

However, concerning both out-of-sample prediction errors and direction of sales, LASSO model`s performance is somewhat comparable to the superior model (Model D), though, the superior model perform a little better.

The model estimated by LASSO with optimal lambda for sales of Toyota in Japan takes the following form:

$y_t = \beta_0 + \beta_1 y_{t-9} + \beta_2 y_{t-11} + \beta_3 y_{t-12} + \beta_4 l_{t-5} + \beta_5 m_{t-1} + \beta_6 m_{t-5} + \beta_7 k_{t-1} + \beta_8 h_{t-6} + \beta_9 h_{t-10} + \beta_{10} f_{t-12} + \beta_{11} S_3 + \beta_{12} S_9$.

In this model, including constant, there are 13 variables in total:

- y is car sales in Japan and y_{t-9}, y_{t-11}, and y_{t-12} are sales of Toyota cars lagged 9, 11, and 12 months;
- l is Google Trends search index, l_{t-5}, is lagged 5 months Google Trends search index;
- m is oil price, m_{t-1} and m_{t-5} are lagged 1 and 5 months accordingly for oil prices;
- k is economic uncertainty, k_{t-1}, is lagged 1 month for economic uncertainty;
- h is unemployment within working population, h_{t-6}, and h_{t-10} are lagged 6 and 10 months accordingly for working population unemployment;
- f is exchange rate, f_{t-12}, lagged 12 months for exchange rate;
- S_3 and S_9 are month dummies;
- β_0 is the constant;

Out of sample prediction errors for Toyota:

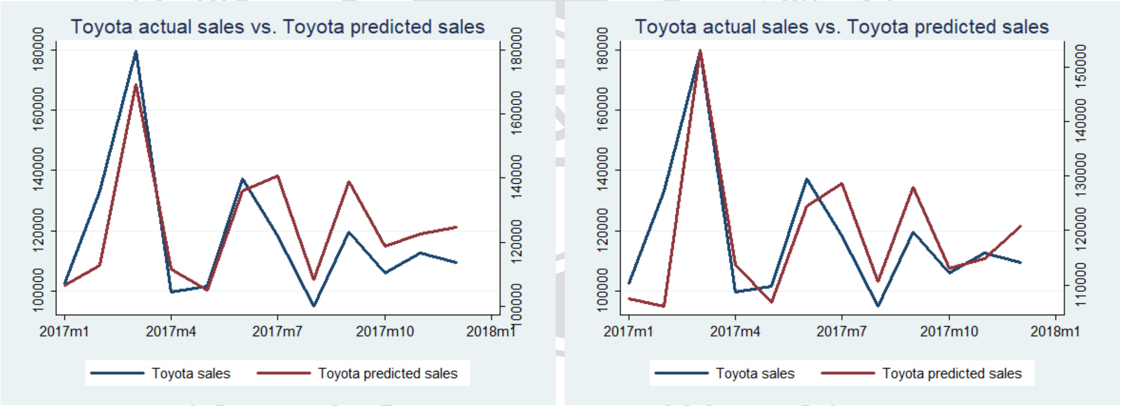
	Model type	RMSE	MAE	U Theil`s
A	With Macro Variable and Google Trends	28 604.21	25 905.334	1.0527726
B	With Macro Variable and Without Google Trends	21 246.29	18 387.173	0.76410689

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

C	**Superior Model (AR)**	**13 725.361**	**12 080.002**	**0.50440673**
D	**LASSO with optimal lambda**	**14 343.443**	**12 198.485**	**0.51652984**

For Toyota, both the superior model (Model C) of this paper and the model specified by LASSO using optimal lambda (Model D) show pretty comparable results. From the graphs below, one can see that, concerning the direction of the sales, the performance is also relatively similar between those two models, while autoregressive model is performing slightly better.

Figure 5.4.2 Toyota sales comparison: actual and predicted (Superior (AR) model on the left; LASSO with opt. lambda model on the right)



When including the economic uncertainty into the superior model (B), the performance is still acceptable regarding U Theil`s Statistics. Although, the out-of-sample prediction errors are higher compared to models C and D. The model with economic uncertainty and Google Trends (Model A) is as good as naïve forecast according to U Theil`s statistics.

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

The model estimated by LASSO with optimal lambda for sales of Nissan in Japan takes the following form:

$$y_t = \beta_0 + \beta_1 y_{t-5} + \beta_2 y_{t-8} + \beta_3 y_{t-12} + \beta_4 x_{t-5} + \beta_5 x_{t-6} + \beta_6 n_{t-2} + \beta_7 n_{t-3} + \beta_8 n_{t-4} + \beta_9 n_{t-6} + \beta_{10} l_{t-2} + \beta_{11} l_{t-3} + \beta_{12} l_{t-4} + \beta_{13} l_{t-6} + \beta_{14} l_{t-7} + \beta_{15} l_{t-11} + \beta_{16} k_{t-1} + \beta_{17} k_{t-2} + \beta_{18} k_{t-3} + \beta_{19} k_{t-6} + \beta_{20} k_{t-7} + \beta_{21} k_{t-11} + \beta_{22} k_{t-12} + \beta_{23} h_{t-6} + \beta_{24} h_{t-8} + \beta_{25} S_3 + \beta_{26} S_4 + \beta_{27} S_5 + \beta_{29} S_8.$$

In this model, there are 29 variables in total:

- y is car sales in Japan, y_{t-5}, y_{t-8} and y_{t-12} are sales of Nissan cars lagged 5, 8 and 12 months accordingly for sales of Nissan;
- x is CPI, x_{t-5}, x_{t-6} are lagged 5 and 6 months accordingly for CPI;
- n is stock market price, n_{t-2}, n_{t-3}, n_{t-4} and n_{t-6} are lagged 2, 3, 4 and 6 months accordingly for stock market prices;
- l is Google Trends search index for Nissan, l_{t-2}, l_{t-3}, l_{t-4}, l_{t-6}, l_{t-7}, and l_{t-11} are lagged 2, 3, 4, 6, 7 and 11 months accordingly for Google Trends search index for Nissan;
- k is economic uncertainty level, k_{t-1}, k_{t-2}, k_{t-3}, k_{t-6}, k_{t-7}, k_{t-11} and k_{t-12} are lagged 1, 2, 3, 6, 7, 11 and 12 months accordingly for economic uncertainty;
- h is unemployment within working population, h_{t-6} and h_{t-8} are lagged 6 and 8 months accordingly for unemployment within working population;
- S_3, S_4, S_5, and S_8 are month dummies;
- β_0 is the constant;

Out of sample prediction errors for Nissan:

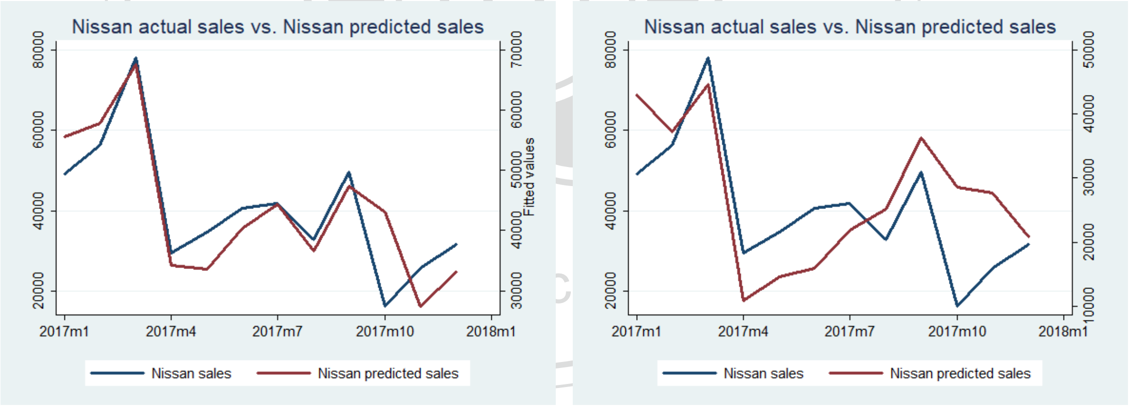
	Model type	RMSE	MAE	U Theil`s
A	With Macro Variable and Google Trends	45 093.294	41 388.391	4.2852056

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

B	With Macro Variable and Without Google Trends	32 634.767	28 377.541	3.3681597
C	**Superior Model (AR)**	**8 721.3859**	**5 210.9806**	**0.46171672**
D	LASSO with optimal lambda	17 766.039	15 682.254	1.1175821

In Nissan`s case, the superior model (Model C) shows better performance, it also much better predicts the direction of sales, which can be seen on the graph below. The models (Model A) and (Model B) are showing a bad performance according to the errors of prediction accuracy.

Figure 5.4.3 Nissan sales comparison: actual and predicted (Superior (AR) model on the left; LASSO with opt. lambda model on the right)



The model estimated by LASSO with optimal lambda for sales of Honda in Japan takes the following form:

$$y_t = \beta_0 + \beta_1 y_{t-7} + \beta_2 y_{t-8} + \beta_3 y_{t-10} + \beta_4 y_{t-12} + \beta_5 n_{t-4} + \beta_6 x_{t-8} + \beta_7 l_{t-3} + \beta_8 k_{t-7} + \beta_9 k_{t-8} + \beta_{10} h_{t-3} + \beta_{11} h_{t-6} + \beta_{12} h_{t-7} + \beta_{13} S_3 + \beta_{14} S_4 + \beta_{15} S_8.$$

In this model, including constant, there are in total 16 variables:

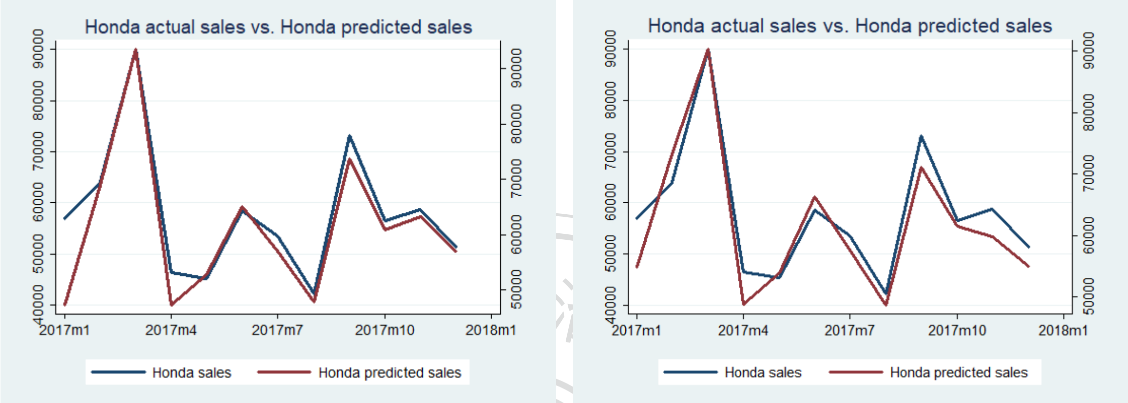
- y is car sales in Japan, y_{t-7}, y_{t-8}, y_{t-10} and y_{t-12} are lagged 10 and 12 months accordingly for the sales of Honda;
- n is stock market price, and n_{t-4} is lagged 4 months for stock market price
- x is CPI, and x_{t-8} is lagged 8 months for CPI;
- l is Google search index for Honda and l_{t-3} is lagged 3 months for Google search index;
- k is economic uncertainty, k_{t-7} and k_{t-8} are lagged 7 and 8 months accordingly for economic uncertainty;
- h is unemployment within working population and h_{t-3}, h_{t-6}, and h_{t-7} are lagged 3, 6, and 7 months accordingly for working population;
- S_3, S_4, and S_8 are month dummies;
- β_0 is the constant;

Out of sample prediction for Honda:

	Model type	RMSE	MAE	U Theil`s
A	With Macro Variable and Google Trends	23602.845	21598.652	1.4271402
B	With Macro Variable and Without Google Trends	12265.909	11863.665	0.73289494
C	**Superior Model (AR)**	**5402.4482**	**4815.276**	**0.33039081**
D	**LASSO with optimal lambda**	**5266.1541**	**4384.2741**	**0.32423967**

For Honda, the model picked with the help of LASSO regression (Model D) and the superior model`s performance (Model C) is comparable regarding prediction errors.

Figure 5.4.4 Honda sales comparison: actual and predicted (Superior (AR) model on the left; LASSO with opt. lambda model on the right)



However, from the graphs, one can observe that the superior model achieves a better fit concerning the direction of sales.

When including the economic uncertainty into the superior model (Model B), the performance is still acceptable regarding U Theil`s Statistics. Although, the out-of-sample prediction errors are higher compared to models C and D. The model with economic uncertainty and Google Trends (Model A) worse than using last period`s sales figures.

Overall, the results presented above show that generally autoregressive model (Model C) is better for Japanese car sales, especially for the direction of sales. Although, the models estimated with LASSO and optimal penalty size regarding their out-of-sample forecast (prediction errors) are relatively close to those of autoregressive model. In fact, in the case of Toyota and Honda, it is hard to decide between those two models. Graphically, those two models comparably fittingly indicate the direction of sales, except for Nissan, where the autoregressive model clearly shows the direction of sales better.

Another item to be learned from this analysis is that the LASSO models (except for Nissan case) feature fewer variables than the autoregressive model, where in total, including constant, 19 variables were added. Those models and all include m3 as a month dummy. This is in accordance with the data definition part, where it was pointed out that there is a pick of sales that happen around February-March each year.

6. Discussion of Results

This paper aims to investigate Google Trends as a tool that can help to create a more accurate forecast of car sales in Japan, as well as compare the human-created forecasting models to the models constructed with the help of machine learning.

There are a few points to be learned from this research and the results of this research.

First, even though models picked by human judgment with Google Trends in combination with macroeconomic variable have a predictive power within the sample, which can be concluded from low estimation errors and high r^2 when evaluating out-of-sample forecast, the predictive accuracy is relatively low, and results are not as good.

The same is true for the macroeconomic variables themselves. The in-sample prediction is improved when macroeconomic variables are included in the superior model. However, out-of-sample prediction errors of those models are much higher than those of the superior model.

Secondly, machine learning has some prospective implications. As it was found, when comparing human-created and computer-created models of the same size (Section 5.3), the computer creates models that have produced more accurate results. Even when the penalty size is determined by human and when optimal lambda is not used, machine estimates (LASSO regression) models are performing better.

Making further comparison of the models (Section 5.4), it was found that, generally, the best forecasting method for Japanese car sales is the autoregressive model with seasonal dummy variables and trend variable, in other words, the superior model of this paper. However, the models constructed with LASSO regression and optimal lambda, in terms of out-of-sample prediction accuracy, are somewhat close to the superior models, especially in case of Toyota and Honda. Concerning the direction of sales, they are also comparable, except for Nissan, where autoregressive model clearly shows the direction of sales better.

Another interesting finding is that the LASSO-models with optimal lambda, in three cases out of four (except for Nissan), include less variables than the superior model.

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

It is also worth noting that LASSO managed to identify the outlier, which is the pick of the sales in Japan that happens around March. LASSO added this month dummy in those models.

All of the models include Google Trends data with Toyota and all brands sales having lagged 5 months Google search index, Honda – 3 months and Nissan having a few lags added.

Therefore, it can be concluded:

RQ1: Is the machine learning method outperforming human-judgment and conventional models when forecasting Japanese car sales?

Models created with machine learning definitely outperform human-picked models when comparing same-size models and their out-of-sample performance.

LASSO models created with optimal lambda create the results that are comparable to the results of the autoregressive model with seasonal dummies and trend variable. However, models picked with LASSO include fewer variables (except for Nissan).

Therefore, it is concluded that machine learning methods definitely has positive implications for forecasting. This is especially the case when researchers are unfamiliar with a specific characteristic of brand or customers since the computer can identify the right variables and their lags.

RQ2: Does the Google Trends index help to improve forecasting models for Japanese car sales?

Combined with macroeconomic variables in the same human-picked model, Google Trends seems to help create more accurate forecast as in-sample prediction accuracy is improved. However, the out-of-sample performance of those models is weak and misleading. That might occur due to the fact that human judgement is biased and misleading.

When models were constructed with the help of machine learning, Google Trends data were included in all of the models. Therefore, in the case of machine-picked models

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

of the same size, Google Trends is helpful in forecasting, but there should be identified right lag terms that is better done by machine learning methods.

Judging from the fact that for Toyota LASSO (with optimal penalty size) picked Google search index data lagged 5 months and for Honda – 3 months, it is concluded that each brand can have unique characteristics and a unique set of consumers and other unobserved aspects. Therefore, Google Trends data might be helpful in forecasting, but in specific cases. As well as the length of the lag might also differ.

It is believed that future research on all brands of cars in Japan will bring more comprehensive results. That is because current research is limited to the scope of available data on Japanese car sales. It is advised to perform a more in-depth research using a longer data period as well as adding more Google search indexes.

Machine learning techniques have positive implications and are expected in the years to come to automate forecasting so that it will be done without the usual time-consuming manual routine.

It is especially important that now scientist and businessmen all over the world are able to utilize Google Trends and it is, as proved in this research, better to use the machine learning techniques when working with Google data as machine learning methods help to identify the right variables for forecast as well as their lags in unbiased manner.

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

References

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593-1636.

Bortoli, C., & Combes, S. Contribution from Google Trends for forecasting the short-term economic outlook in France: limited avenues.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88* (s1), 2-9.

Diebold, F.X. (2017). *Forecasting.* Pennsylvania: Department of Economics, University of Pennsylvania. Retrieved from: http://www.ssc.upenn.edu/~fdiebold/Textbooks.html

Fantazzini, D., & Toktamysova, Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics*, *170*, 97-135.

Gevelber, L. (2016, March). The Car-Buying Process: One Consumer's 900+ Digital Interactions. Retrieved from https://www.thinkwithgoogle.com/consumer-insights/consumer-car-buying-process-reveals-auto-marketing-opportunities/

Google Inc. (2018). How Trends data is adjusted. Retrieved from https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052

Hand, C., & Judge, G. (2012). Searching for the picture: forecasting UK cinema admissions using Google Trends data. *Applied Economics Letters*, *19*(11), 1051-1055.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679-688.

Kotler, P., & Keller, K. L. (2012). *Marketing Management.* Global Edition 14e, London: Pearson Education Limited 2012

Li, J., & Chen, W. (2014). Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models. *International Journal of Forecasting*, *30*(4), 996-1015.

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

MAE and RMSE—Which Metric is Better? (2016, March 23). Retrieved from https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

Muehlen, M. (2017) Improved Sales Forecasting with Consumer Behavior. IMES, National Chengchi University

Sagaert, Y. R., Aghezzaf, E. H., Kourentzes, N., & Desmet, B. (2017). Temporal big data for tire industry tactical sales forecasting. *Interfaces*.

Shi, Y., Liu, X., Kok, S. Y., Rajarethinam, J., Liang, S., Yap, G., ... & Lo, A. (2016). Three-month real-time dengue forecast models: an early warning system for outbreak alerts and policy decision support in Singapore. *Environmental health perspectives*, *124*(9), 1369.

Small, G., & Wong, R. (2002). The validity of forecasting. In *A Paper for Presentation at the Pacific Rim Real Estate Society International Conference, Christchurch, New Zealand* (pp. 1-14).

Spiegelm B. (2015, February 10). The Google Trends Data Goldmine. Retrieved from https://marketingland.com/google-trend-goldmine-117626

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

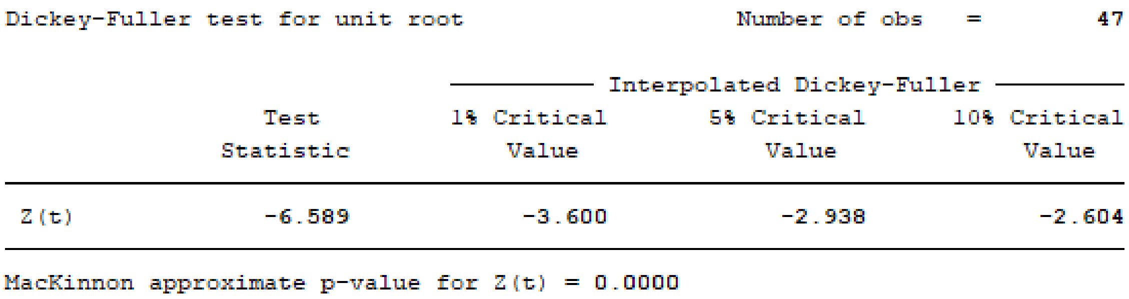
Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), 79-82.

Yang T.T. (2018). *Machine Learning and Casual Inference* [PowerPoint slides]. Retrieved from: https://drive.google.com/file/d/1wUfA6RzcwHkOTId7_dA86-PJ67T6-xgI/view

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

Appendix 1

Dickey-Fuller test results for sales of all brands



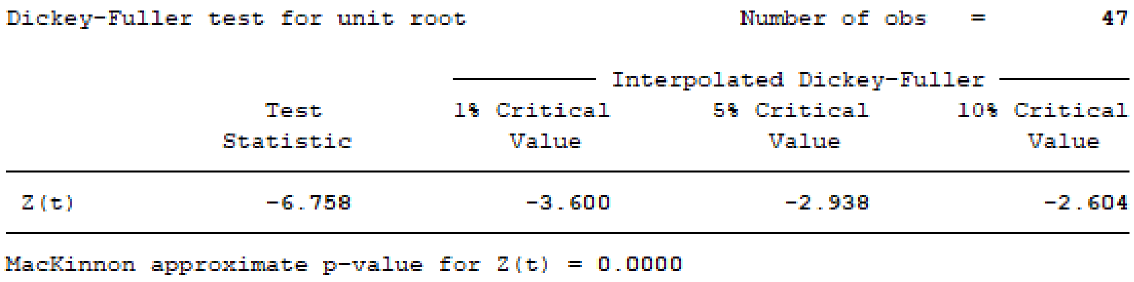
```
Dickey-Fuller test for unit root                     Number of obs   =        47

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -6.589            -3.600            -2.938            -2.604
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

Appendix 2

Dickey-Fuller test results for Toyota:



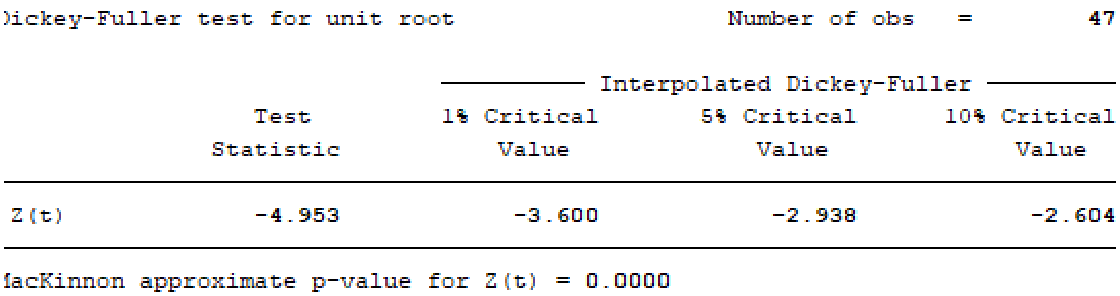
```
Dickey-Fuller test for unit root                   Number of obs   =        47

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -6.758            -3.600            -2.938            -2.604
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

Appendix 3

Dickey-Fuller test results for Nissan:



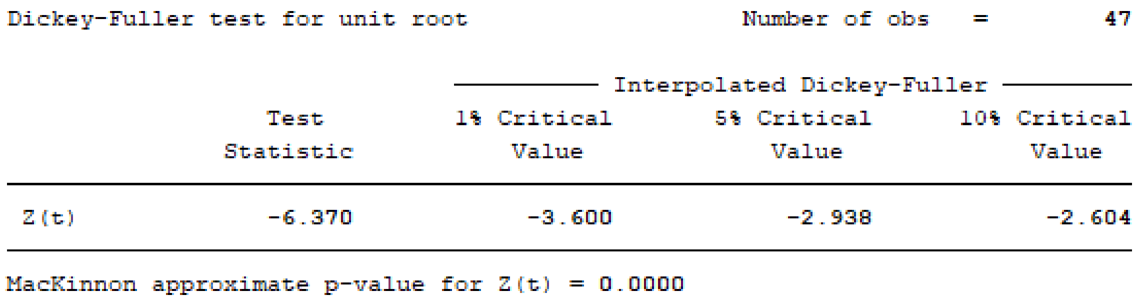
```
)ickey-Fuller test for unit root                    Number of obs   =        47

                              ───────── Interpolated Dickey-Fuller ────────
                  Test          1% Critical        5% Critical      10% Critical
               Statistic           Value              Value             Value
────────────────────────────────────────────────────────────────────────────────
Z(t)             -4.953            -3.600             -2.938            -2.604
────────────────────────────────────────────────────────────────────────────────
lacKinnon approximate p-value for Z(t) = 0.0000
```

DOI:10.6814/THE.NCCU.IMES.001.2018.F06

Appendix 4

Dicky-Fuller test results for Honda:



```
Dickey-Fuller test for unit root                   Number of obs   =        47

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -6.370            -3.600            -2.938            -2.604
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

DOI:10.6814/THE.NCCU.IMES.001.2018.F06