

Auto Car Sales Prediction: A Statistical Study Using Functional Data Analysis and Time Series.

Honors Thesis by Yuchen Lin

Advised by Professor Ed Rothman

University of Michigan

Department of Statistics

Table of Contents

Chapter 1 – Introduction

Chapter 2 – Functional Data Analysis

Chapter 3 – Regression

Chapter 4 – Fitting models to data

Chapter 5 – Results

Chapter 6 – Future Works

Appendix of Code and Graphs

Bibliography

Acknowledgments

I would like to express my appreciation to Professor Rothman for his support and guidance on this project. Additionally, I would like to thank him for encouraging me to do research in economic elements by using statistical models. I would also like to thank Zhenxiang Zhou and Beth Crane for showing me the many applications of functional data analysis. Special thanks to George Fulton for helping me obtain the data for this project.

Chapter 1: Introduction

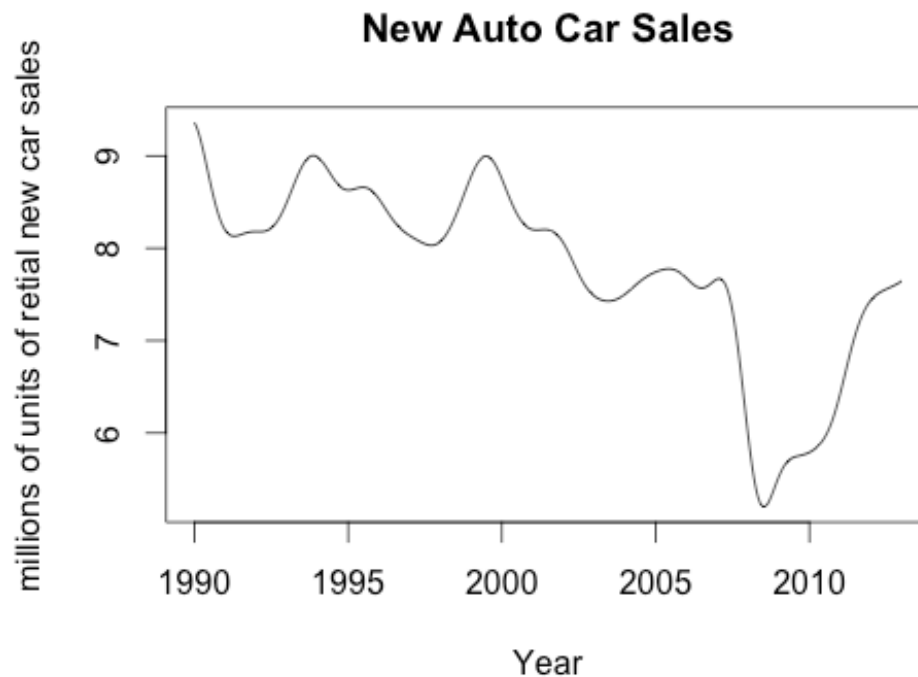
“Autos Sales” is one of the major producers of domestic automobiles report sales monthly. These numbers are seasonally adjusted by the U.S. Department of Commerce and are available to the public one to five business days after the end of each month.¹ This is an important element to the financial markets since it is highly correlated with consumer demands for the market and car sales can express the changes in the economy precisely.

There is an old warning story named “boiled frog.” The frog’s body temperature follows its surroundings. If you put the frog directly in boiling water, it will sense the heat immediately and jump out. However, when you heat the water slowly, the frog keeps adjusting to the rising temperature. When the heat is too much for the frog to take, it is too late. The frog collapses and dies.² This theory can apply into economic as well. For example, in a hamburger restaurant, if the price for one hamburger change from \$5.49 to \$5.99, it would not have many people noticed this change. Therefore, it would not affect their sales a lot based on common sense. Nevertheless, if the price for one hamburger changes from \$5.99 to \$6.99, people will catch this change easily and the sales for the hamburger may decrease due to the change in price. Various differences in auto sales share the same situation here. For predictors which can use for estimating the change in auto sales, when the predictors change a little bit, auto sales may not shift according to this adjustment and it will not be influenced by this. However, if there is a big change in one predictor, then it will lead a change in auto sales just as the “boiled frog” theory. Having a small change in predictors seems like that putting a frog in and heating the water slowly, auto sales will adjust itself and change a little bit to follow the changing trends in predictors. On the other hand, if there is a big change in predictors, it looks like that putting a frog directly in boiling water. In this way, auto sales will follow this big change immediately and then be acted to such big change.

Based on statistics studies, using first derivatives of predictors to measure the level of change in each predictor is a good way to do analysis.

With the new car sales changing a lot in the United States, what affecting units of new car sales has become a topic of great interest to researchers. As we can see from the plot below:

Figure 1.1 Auto Car Sales (With Smoothing)



There is a big downward change in year 2008. Since car sales are an excellent indicator of the financial market, the reason may be financial crisis. However, what specific factors affect units of car sales and how sensitive of each one to units of car sales? This is an important direction to let us explore. If we know what will lead a change in auto car sales, we are able to know the market trend in order to have some accurate predictions. There are many research articles about how to predict auto car sales by using gross domestic product (GDP) to make prediction. Gross domestic product (GDP) is defined by the Organisation for Economic Co-operation and

Development (OCED) as “an aggregate measure of production equal to the sum of the gross values added of all resident institutional units engaged in production (plus any taxes, and minus any subsidies, on products not included in the value of their outputs)”³ Besides GDP plays an important role to predict the auto sales, unemployment rate, price of crude oil and so on are important to predict auto sales too. Those variables are common variables, which are used for measuring the financial market situation.

The purpose of this project was to determine how to use unemployment rate, price of crude oil, S&P 500 index, disposable personal income, consumer price index (CPI) for all items, inflation rate and interest rate on 48-month to predict the number of auto car sales.

Unemployment occurs when people do not have work. The unemployment rate is a measure of the prevalence of unemployment. During periods of recession, an economy usually experiences a relatively high employment rate.⁴ S&P 500 index chosen for market size, liquidity and industry grouping, among other factors. The S&P 500 is designed to be a leading indicator of U.S. equities and is meant to reflect the risk/return characteristics of the large cap universe.

Additionally, it is one of the most commonly used benchmarks for the overall U. S. stock market.

⁵ The consumer price index (CPI) measures the changes in the prices paid by domestic consumers for goods and services.⁶ The inflation rate is a rate at which the general level of prices for goods and services is rising, and, subsequently, purchasing power is falling.⁷ All of these predictors have crucial meanings to the market and they can reflect what the current market is. All activities have shown to impact auto sales, but data for economic elements is limited. The data came from the first quarter of 1990 and the fourth quarter of 2013 and are collected from United States. Analysis was performed on 96 observations. The aim is to offer an objective

analysis on how big change in these predicts will lead an accurate prediction model for auto car sales.

Chapter 2: Functional Data Analysis

Functional data analysis is a method of statistics that doing data analysis in order to provide information about curves over time. In addition, we always use functional data analysis in order to smooth data and then fit these points to a function model. It plays an important role when researchers interpret continuous variables. According to this investigation, using time in quarter to see the changing trend in the auto car sales. Quarter in each year is often reported as a discrete number of years. Functional data analysis will consider separate quarters as continues variables from 1990 to 2013. Treating quarter in year as a continuous variable can be fitted as the proportion of a function of time. In this way, we can obtain the derivatives of the predictor variables and other useful data to do investigation.

In statistics and mathematics field of numerical analysis, interpolation is a common method for constructing new data with the range of a discrete data set. Basic spline method and the Fourier series are normally used in processing data. The reason that using Fourier series is that we can better analyze a new data set in the new domain instead of the original one. Fourier series is an approach that decomposing periodic function into the sum of a set of simple sines and cosines function. A B-spline is an application based on Bézier curves. “The curve construction method that we consider in this section is an alternative to polynomial interpolation and produces what we call *Bézier cureves*.” Basic splines is segment of the polynomial curves and is can be easily be joined smoothly to form various shapes.⁸ The B-spline bases system is defined by the order of the polynomial, which is one more than the value of degree, and the location of the knots. Knots are the points at which the segments points. Additionally, the number of knots determines the number of corresponding derivatives. For example, if there is one know, then the number of corresponding derivatives is two less than the function’s order.⁹

This is the basic idea of function smoothing. Smoothing is a way to make the function and its derivatives to be continuous in order to be provides support to do more analysis.

The smoothing parameter λ provides support to make sure the function is smooth. This parameter λ is chosen to penalize the derivative of the basic spline function. We choose the parameter λ by minimizing the below equation:

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx \quad (1)$$

where y_i is the observed value, $f(x_i)$ is the fitting function to the data, and λ is the smoothing parameter.¹⁰

As we can see from this equation, if λ is close to 0, the fitted model would be smoother in order to let the data fit the function perfectly. If we choose λ in a big value, we cannot receive a smooth curve and we usually obtain a straight curve. Although it seems that smoothing data method is a best way to process data, it has its limitations too. As for the start points and end points of the data, it will be unstable and are usually hard to predict what it will be.

According to this project: analysis of auto sales change among period from 1990 to 2003 in the United States. We choose to use the basic spline to process data. We are going to analyze data based on the changing curve and their first derivatives as a function of time as well. We chose to order our function as degree 6 and had 4 knots to ensure the derivative as continuous. We chose λ to be very small $1e-4$ to ensure a good fit as well as smoothness. The figures below show the effects of different λ (See Figure 2.1) for auto car sales data and how λ influence data point fitness (See Figure 2.2).

Figure 2.1: Large λ makes the function smoother. Small λ makes the function reflecting the real data well.

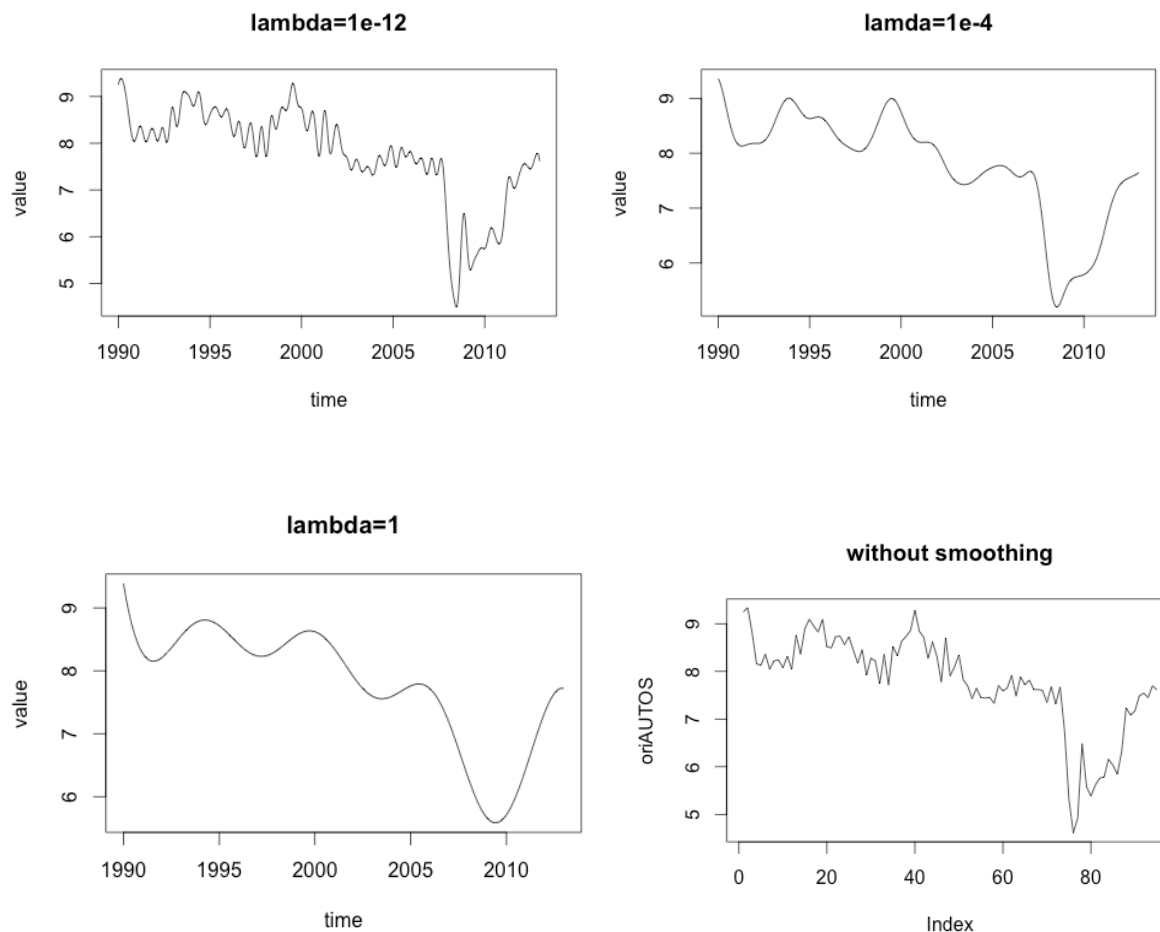
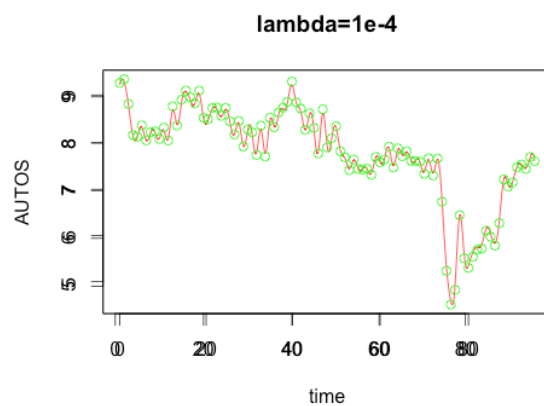


Figure 2.2: As we can see, with smaller λ , the data can fit very well.



Chapter 3: Regression

In statistics, regression analysis is a common method for estimating the relationships between independent variables and dependent variable. In order to figure out which predictor will lead a big change in auto car sales. The most general method is ordinary least squares (OLS) or linear least squares. Ordinary least-squares (OLS) regression is a generalized linear modeling technique that may be used to model a single response variable, which has been recorded on at least an interval scale. The technique may be applied to single or multiple explanatory variables and also categorical explanatory variables that have been appropriately coded.¹¹ In a model, despite the importance of the parameters and the confidence intervals for the estimators. It is important to know how well one model fits the data.

In our model, we did OLS regression with multiple variables. We loaded the data into R and performed various tests to assess the normality of the data and to compare unemployment rate, price of crude oil, S&P 500 index, disposable personal income, consumer price index (CPI) for all items, inflation rate, interest rate on 48-month and the number of auto car sales. Then, simple linear regression was performed with crude oil price, unemployment rate, disposable income, CPI for all items, inflation rate and interest rate on 48-month as the explanatory variables and auto sales as the response variable. We tested for significance between these variables of interest. This process was then altered to include other variables that could act as confounders. We also used R to find the best model via AIC.

For model selection, we felt there could be a better suited model that could give us more accurate information. We can see from the full model that some variables are not considered significant. In order to find a better model compared to the full model, we wished to choose the best model in all combinations, and also delete variables that were not considered useful for the

purpose of our study. In order to accomplish this, we used Akaike Information Criterion (AIC), which is a measure of the relative quality of a statistical model for a specific dataset.

Furthermore, we choose a new model by using Stepwise Regression, specifically using backward forward selection. At first, we started with just the intercept in the model, and the AIC value was 3.92. In step 2 of this procedure, we added CPI for all items as a variable, which lowered the AIC measure to -71.32. In step 3, we added S&P 500 as a variable, which made the AIC value lower at -109.71. In step 4, we added interest rate on 48-month as a variable, which let the AIC equaled -120.2. In step 5, we added crude oil price into this model, and then we get the value of AIC was -132.12. Finally, the consequent AIC value was the lowest out of all possible combinations when we have CPI for all items (PCPI), stock price (SP500), interest rate on 48-month (RVEH48), crude oil price (PBRENT) and unemployment rate (RUM) as explanatory variables. The lowest AIC was -146.38. We can test that this is the best model due to the fact that when we add other variables, the AIC measure is not optimal. As a result, the final model we will use is

$$AUTO\ Sales = \beta_0 + \beta_1(PCPI) + \beta_2(SP500) + \beta_3(RVEH48) + \beta_4(PBRENT) + \beta_5(RUM) + \varepsilon$$

The reason why these variables play an important role to do prediction, not only AIC selection procedures indicates this, but also these variables have the crucial meaning in reality. Economic market are various due to many reasons. In order to derive an accurate model, it is necessary to have crucial variables which can contribute to the results to do estimate the changing in the market. In this way, all these explanatory variables have this feature since all of them are used for expressing the economic market through different prospective. Based on common sense, we can consider personal income, interest rate and CPI are important variables to do prediction. However, after modeling, we find that personal income is not a good predictor since it is not significant variable in the linear model. Therefore, we need to remove this variable

in our model although it may economically meaningful in this case. Interest rate and CPI may directly affect the sales of cars, but stock price, crude oil price and unemployment rate may also reflect the market demand in some ways. Including these variables could make our forecasting model more accurate.

Chapter 4: Fitting Models to data

4.1 Ordinary Least Squares (OLS) Model

We are interested in predicting how big change in predictor variables by using the first derivative of each variable lead a big change in auto car sales. In order to predict the future auto car sales, we need to build a model based on the past data. Following the below formula:

$$y_t = \gamma x_{t-1} + \varepsilon$$

where y_t is the raw data of auto car sales on time t , x_{t-1} is the combination of raw data of predictors on time $t-1$ and γ is the coefficient for each random variable.

The main purpose for this project's idea is to find what cut point for each variable will lead a big change in auto car sales outcome and in order to receive an accurate forecasting. At first, we need to fit the full model by AIC selection:

$$AUTO\ Sales = \beta_0 + \beta_1(PCPI) + \beta_2(SP500) + \beta_3(RVEH48) + \beta_4(PBRENT) + \beta_5(RUM) + \varepsilon_t$$

And then we are going to fit a model with different indicator cut points:

$$AUTO\ Sales = \beta_0 + \beta_1(PCPI)I_1 + \beta_2(SP500)I_2 + \beta_3(RVEH48)I_3 + \beta_4(PBRENT)I_4 + \beta_5(RUM)I_5 + \varepsilon_t$$

where I denotes an indicator. If the first derivative of each variable is larger than one cut point, the indicator I will equal 1. Otherwise, it will equal 0. The indicators for first derivatives of each variable, we used the derivatives of smooth data. But for the dependent and independent variable, we used the raw data instead of smoothing data.

4.2 Time Series Model

Despite fitting Ordinary Least Squares model, it is necessary to fit the time series model since one of the assumptions underlying ordinary least squares (OLS) model is that the error terms are independent. This assumption is easily violated for time series data, such as this one, since it may have pattern in quarters and years. Therefore, we found that the error terms have

autocorrelation. Then, fitting a time series model with lag equaling 1 is necessary. Firstly, we fitted a time series model like this:

$$AUTO\ Sales = \beta_0 + \beta_1(PCPI) + \beta_2(SP500) + \beta_3(RVEH48) + \beta_4(PBRENT) + \beta_5(RUM) + U_t$$

$$U_t = \int U_{t-1} + \varepsilon_t$$

Then adding indicators into this model, it would change to:

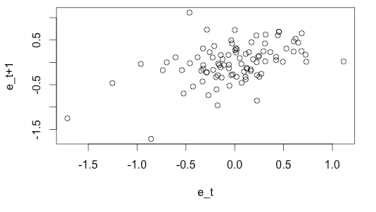
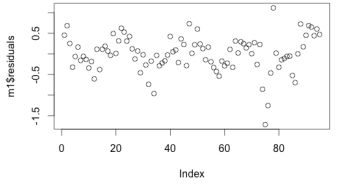
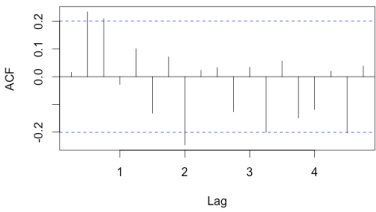
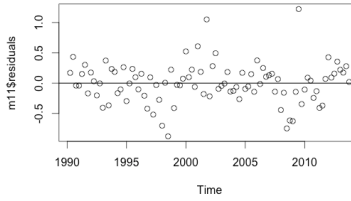
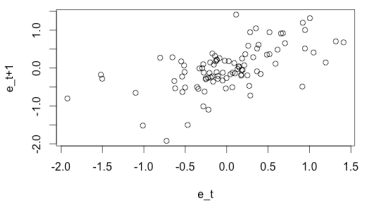
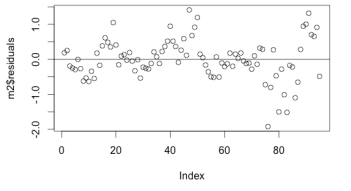
$$AUTO\ Sales = \beta_0 + \beta_1(PCPI)I_1 + \beta_2(SP500)I_2 + \beta_3(RVEH48)I_3 + \beta_4(PBRENT)I_4 + \beta_5(RUM)I_5 + U_t$$

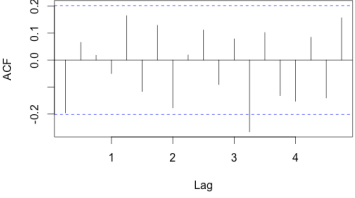
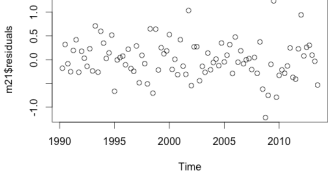
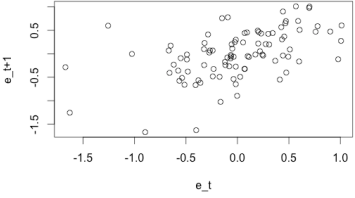
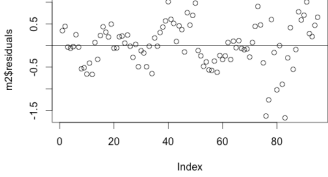
$$U_t = \int U_{t-1} + \varepsilon_t$$

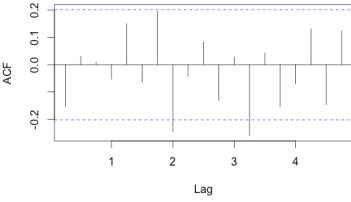
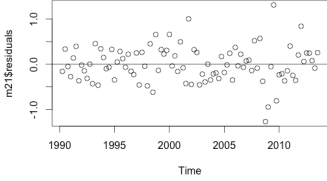
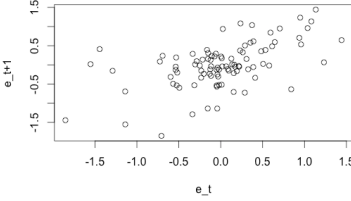
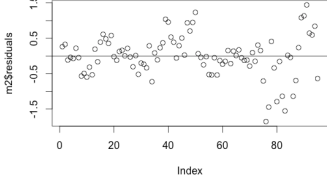
where I denotes an indicator. If the first derivative of each variable is larger than one cut point, the indicator I will equal 1. Otherwise, it will equal 0. The indicators for first derivatives of each variable, we used the derivatives of smooth data. But for the dependent and independent variable, we used the raw data instead of smoothing data.

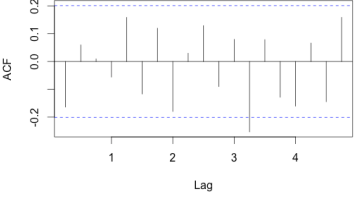
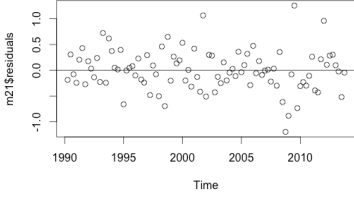
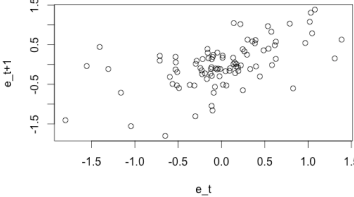
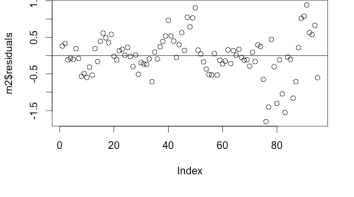
Chapter 5: Results

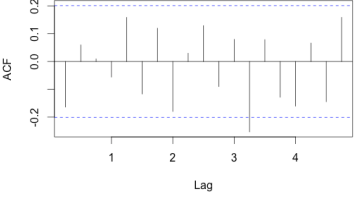
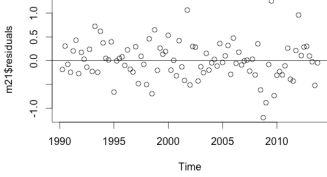
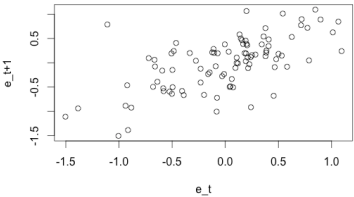
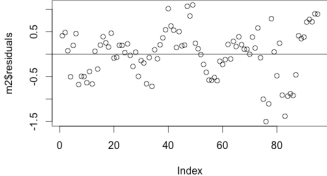
Table 5.1:

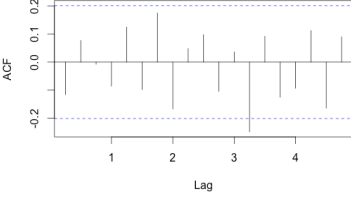
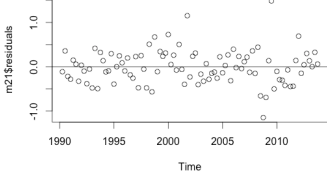
Model	Multiple R Square	Plots	Plot Residuals
<p><i>AUTO Sales</i></p> $= \beta_0 + \beta_1(PCPI) + \beta_2(SP500) + \beta_3(RVEH48) + \beta_4(PBRENT) + \beta_5(RUM) + \varepsilon_t$	<p>0.815</p>	<p>plot e_{t+1} vs. e_t</p> 	<p>Residuals in Full Model</p> 
<p><i>AUTO Sales</i></p> $= \beta_0 + \beta_1(PCPI) + \beta_2(SP500) + \beta_3(RVEH48) + \beta_4(PBRENT) + \beta_5(RUM) + U_t$	<p>0.895</p>	<p>Series m11\$residuals</p> 	<p>residuals in time series model</p> 
<p><i>AUTO Sales</i></p> $= \beta_0 + \beta_1(PCPI)(zscore > .005) + \beta_2(SP500)(zscore > 1) + \beta_3(RVEH48)(zscore > 0.05) + \beta_4(PBRENT)(zscore > 0.0007) + \beta_5(RUM)(zscore > 0.9) + \varepsilon_t$	<p>0.6753</p>	<p>plot e_{t+1} vs. e_t</p> 	<p>Residuals in Full Model</p> 

<p><i>AUTO Sales</i></p> <p>$= \beta_0$</p> <p>$+ \beta_1(PCPI)(zscore$</p> <p>$> .005)$</p> <p>$+ \beta_2(SP500)(zscore$</p> <p>$> 1)$</p> <p>$+ \beta_3(RVEH48)(zscore$</p> <p>$> 0.05)$</p> <p>$+ \beta_4(PBRENT)(zscore$</p> <p>$> 0.0007)$</p> <p>$+ \beta_5(RUM)(zscore$</p> <p>$> 0.9) + U_t$</p>	<p>0.8478</p>		
<p><i>AUTO Sales</i></p> <p>$= \beta_0$</p> <p>$+ \beta_1(PCPI)(zscore$</p> <p>$> 0)$</p> <p>$+ \beta_2(SP500)(zscore$</p> <p>$> 1)$</p> <p>$+ \beta_3(RVEH48)(zscore$</p> <p>$> 0)$</p> <p>$+ \beta_4(PBRENT)(zscore$</p> <p>$> 1)$</p> <p>$+ \beta_5(RUM)(zscore$</p> <p>$> 1) + \varepsilon_t$</p>	<p>0.7391</p>		

<p><i>AUTO Sales</i></p> <p>$= \beta_0$</p> <p>$+ \beta_1(PCPI)(zscore$</p> <p>$> 0)$</p> <p>$+ \beta_2(SP500)(zscore$</p> <p>$> 1)$</p> <p>$+ \beta_3(RVEH48)(zscore$</p> <p>$> 0)$</p> <p>$+ \beta_4(PBRENT)(zscore$</p> <p>$> 1)$</p> <p>$+ \beta_5(RUM)(zscore$</p> <p>$> 1) + U_t$</p>	<p>0.8516</p>		
<p><i>AUTO Sales</i></p> <p>$= \beta_0$</p> <p>$+ \beta_1(PCPI)(zscore$</p> <p>$> .005)$</p> <p>$+ \beta_2(SP500)(zscore$</p> <p>$> 0.95)$</p> <p>$+ \beta_3(RVEH48)(zscore$</p> <p>$> 0.01)$</p> <p>$+ \beta_4(PBRENT)(zscore$</p> <p>$> 0.0007)$</p> <p>$+ \beta_5(RUM)(zscore$</p> <p>$> 1) + \varepsilon_t$</p>	<p>0.6672</p>		

<p><i>AUTO Sales</i></p> <p>$= \beta_0$</p> <p>$+ \beta_1(PCPI)(zscore$</p> <p>$> .005)$</p> <p>$+ \beta_2(SP500)(zscore$</p> <p>$> 0.95)$</p> <p>$+ \beta_3(RVEH48)(zscore$</p> <p>$> 0.01)$</p> <p>$+ \beta_4(PBRENT)(zscore$</p> <p>$> 0.0007)$</p> <p>$+ \beta_5(RUM)(zscore$</p> <p>$> 1) + U_t$</p>	<p>0.8482</p>	<p>Series m2\$residuals</p> 	<p>residuals in time series model</p> 
<p><i>AUTO Sales</i></p> <p>$= \beta_0$</p> <p>$+ \beta_1(PCPI)(zscore$</p> <p>$> .005)$</p> <p>$+ \beta_2(SP500)(zscore$</p> <p>$> 1)$</p> <p>$+ \beta_3(RVEH48)(zscore$</p> <p>$> 0.01)$</p> <p>$+ \beta_4(PBRENT)(zscore$</p> <p>$> 0.0007)$</p> <p>$+ \beta_5(RUM)(zscore$</p> <p>$> 1) + \varepsilon_t$</p>	<p>0.6764</p>	<p>plot e_t+1 vs. e_t</p> 	<p>Residuals in Full Model</p> 

<p><i>AUTO Sales</i></p> <p>$= \beta_0$</p> <p>$+ \beta_1(PCPI)(zscore$</p> <p>$> .005)$</p> <p>$+ \beta_2(SP500)(zscore$</p> <p>$> 1)$</p> <p>$+ \beta_3(RVEH48)(zscore$</p> <p>$> 0.01)$</p> <p>$+ \beta_4(PBRENT)(zscore$</p> <p>$> 0.0007)$</p> <p>$+ \beta_5(RUM)(zscore$</p> <p>$> 1) + U_t$</p>	<p>0.8482</p>	<p>Series m21\$residuals</p> 	<p>residuals in time series model</p> 
<p><i>AUTO Sales</i></p> <p>$= \beta_0$</p> <p>$+ \beta_1(PCPI)(zscore$</p> <p>$> 0)$</p> <p>$+ \beta_2(SP500)(zscore$</p> <p>$> 1.2)$</p> <p>$+ \beta_3(RVEH48)(zscore$</p> <p>$> 0)$</p> <p>$+ \beta_4(PBRENT)(zscore$</p> <p>$> 0.0)$</p> <p>$+ \beta_5(RUM)(zscore$</p> <p>$> 1.2) + \varepsilon_t$</p>	<p>0.7148</p>	<p>plot e_t+1 vs. e_t</p> 	<p>Residuals in Full Model</p> 

<p><i>AUTO Sales</i></p> <p>$= \beta_0$</p> <p>$+ \beta_1(PCPI)(zscore$</p> <p>$> 0)$</p> <p>$+ \beta_2(SP500)(zscore$</p> <p>$> 1.2)$</p> <p>$+ \beta_3(RVEH48)(zscore$</p> <p>$> 0)$</p> <p>$+ \beta_4(PBRENT)(zscore$</p> <p>$> 0.0)$</p> <p>$+ \beta_5(RUM)(zscore$</p> <p>$> 1.2) + U_t$</p>	<p>0.8576</p>		
---	---------------	--	---

The first column in Table 5.1 illustrates the model function we fitted. The second column in Table 5.2 denotes the multiple R square in the fitting model and we used this value to determine how well it fitted. The third column in this table includes two types of graph. For the OLS model fitting, we used the plot about “ ϵ_t vs. ϵ_{t-1} ” to see the relationship between the two error terms. For the time series model fitting, we used the ACF plot to see the residuals. The fourth column in the table provides the residuals plots information in our models.

As we can see in the table, compared two models which share the same cut point condition. The R-square in time series model is larger than the R-square in the OLS model. Therefore, time series model would give us a more accurate forecasting.

Take all the values of first derivatives as absolute values. We can find that consumer price index for all items and interest rate are sensitive to auto car sales. We need to use small cut points approaching to 0 to do prediction. After various trials, we can find using the raw data instead of putting any indicators in this model would give a better result. For other predictors, it is necessary to include when there is a big change in those variables. It does not have any means

to include all variables all the time since it would not give much information for auto car sales prediction. This result follows the supply and demand curve in the economy. For example, stock price is an element, which reflects the changing in the stock market, and it is hard to predict car sales market as an important predictor. However, interest rate is highly correlated with auto car sales. If there is a tiny change in interest, it will lead a big change for the model. When interest goes up, people would prefer to deposit their money in the bank instead of making a loan since the interest rate is high. If they borrow money from bank at a high interest rate, they need to pay more money back. However, if the interest rate goes down, there would be more people to buy a car, because the interest rate is low and people can afford the loan. They are willing to spend more money since making deposits in the bank cannot give them much money. Therefore, for these variables such as price for all items and interest rate, although there is a little bit change in these variables, we need to include these values to do prediction since they are sensitive to auto car sales.

We can conclude from the various R squares that unemployment rate and stock price are not the sensitive variables in the model. Interest rate, crude oil price, and CPI for all items play a meaningful role in predicting auto car sales. It is necessary to include these variables in big first derivative value. Otherwise, it does not have crucial contribution for forecasting auto car sales.

Chapter 6: Future Work

There were several evident limitations in our study. The main limitation was the small sample size and the fact the data that we used in the study is not representative. Doing ordinary least squares model, it is difficult to make any conclusions regarding to future auto car sales. The data that we used to do prediction is not always the past year data. Sometimes, we used future data to do prediction. It cannot give us an accurate forecasting model for predicting future auto car sales. There are various problems that arise from making conclusions from the results of this study, because it does appear to have a selection bias. Deriving this model, we use part of past data and part of future data to do prediction. It may create bias since the main purpose of our model is to forecast future auto car sales. In order to improve this study, we need to have a larger range of past data to do prediction or come up with a method to select the subset of the data.

Although doing time series model can reduce this limitation, we still cannot figure out whether all these years' data are representative. For forecasting one year's future auto car sales, using the past two years' data before this predicting year would be more informative since it can reflect the most recent economic trends. However, in our model, we used whole past years' data to do prediction. It may have bias since the predictors' values in 1990 cannot be representative predictors for forecasting the auto car sales in 2015. There is much change that the economic market changes during these years.

Another issue is self-choosing cut points for each predictor. Those cut points chosen may not be accurate, whether intentional or inadvertent. For instance, certain number we chose for dependent variable cut point is 1.0, may be incorrectly chosen. It may too randomly chosen to derive an accurate model. There is no way to account for this. We can see that measurements be chosen very randomly, only based on the researcher's understanding. This would have to be

remedied for more accurate cut points in the future. It is necessary to have a cut point selection method to choose cut points in order to reduce the bias. Coming up with a selection model for cut points is an efficient way to solve this problem. But we need to think which model is the best one not only solving this problem but also limiting the new bias.

Furthermore, the use of all of these predictors as a proxy for each field in general is questionable, and actually rather problematic. Auto car sale is calculated by simply using these predictors, unemployment rate, CPI for all items, interest rate, crude oil price and stock price. It still has some other phenomena we cannot cover. While, people may be reported unemployed and then start his or her own business. For example, it is popular using Uber as a transportation in the United States, would this be a case that increasing the auto car sales? Therefore, for future modeling, we not only need to include particular predictors, but also need to include the analysis to the every change which related to the auto car sales.

Appendix

R-studio:

```

library(dynlm)
library(fda)
setwd("/Users/YuchenLin/Dropbox/Research/Thesis")
data<-read.csv("data.csv")
result<-read.csv("result.csv")
lambda <- 1e-4
norder <- 6
samples <- seq(1990,2013, length=96)
nbasis <- length(samples) + norder-2
mybasis <- create.bspline.basis(c(1990,2013), nbasis, norder, samples)
myfdPar <- fdPar(mybasis, 4, lambda)
AUTOS<- smooth.basis(samples, data[,2],myfdPar)$fd
RUM<-smooth.basis(samples, data[,3],myfdPar)$fd
RVEH48 <- smooth.basis(samples, data[,4],myfdPar)$fd
PBRENT <- smooth.basis(samples, data[,5],myfdPar)$fd
YD <- smooth.basis(samples, data[,6],myfdPar)$fd
PCPI <- smooth.basis(samples, data[,7],myfdPar)$fd
SP500 <- smooth.basis(samples, data[,8],myfdPar)$fd
RPPERM <- smooth.basis(samples, data[,9],myfdPar)$fd
oriAUTOS<-data[-1,2]
oriRUM<- data[-96,3]
oriRVEH48<-data[-96,4]
oriPBRENT<- data[-96,5]
oriYD <- data[-96,6]
oriPCPI <-data[-96,7]
oriSP500 <- data[-96,8]
oriRPPERM <- data[-96,9]

rawAUTOS<-data[,2]
rawRUM<- data[,3]
rawRVEH48<-data[,4]
rawPBRENT<- data[,5]
rawYD <- data[,6]
rawPCPI <-data[,7]
rawSP500 <- data[,8]
rawRPPERM <- data[,9]
AUTOS.ts=ts(rawAUTOS, start=c(1990,1), freq=4)
RUM.ts=ts(rawRUM,start=c(1990,1),freq=4)
RVEH48.ts<-ts(rawRVEH48,start=c(1990,1),freq=4)
PBRENT.ts<-ts(rawPBRENT,start=c(1990,1),freq=4)
YD.ts <- ts(rawYD,start=c(1990,1),freq=4)
PCPI.ts<-ts(rawPCPI,start=c(1990,1),freq=4)
SP500.ts <-ts(rawSP500,start=c(1990,1),freq=4)
RPPERM.ts<-ts(rawRPPERM,start=c(1990,1),freq=4)

```

```

m1<-lm(oriAUTOS~oriPCPI+oriSP500+oriRVEH48+oriPBRENT+oriRUM)
summary(m1)
a<-m1$residuals
b<-a[-1]
c<-a[-95]
plot(c,b,xlab="e_t",ylab="e_t+1",main="plot e_t+1 vs. e_t")
plot(m1$residuals,main="Residuals in Full Model")
abline(h=0)
m11<-
dynlm(AUTOS.ts~RUM.ts+RVEH48.ts+PBRENT.ts+PCPI.ts+RPPERM.ts+L(AUTOS.ts))
summary(m11)
plot(m11$residuals,type="p",main="residuals in time series model")
abline(h=0)
acf(m11$residuals)

result<-read.csv("result.csv")
RUM<- result[-96,1]
RVEH48<-result[-96,2]
PBRENT<- result[-96,3]
YD <- result[-96,4]
PCPI <-result[-96,5]
SP500 <- result[-96,6]
RPPERM <- result[-96,7]

rrowRUM<- result[,1]
rrowRVEH48<-result[,2]
rrowPBRENT<- result[,3]
rrowYD <- result[,4]
rrowPCPI <-result[,5]
rrowSP500 <- result[,6]
rrowRPPERM <- result[,7]

RUM.ts=ts(rrowRUM,start=c(1990,1),freq=4)
RVEH48.ts<-ts(rrowRVEH48,start=c(1990,1),freq=4)
PBRENT.ts<-ts(rrowPBRENT,start=c(1990,1),freq=4)
YD.ts <- ts(rrowYD,start=c(1990,1),freq=4)
PCPI.ts<-ts(rrowPCPI,start=c(1990,1),freq=4)
SP500.ts <-ts(rrowSP500,start=c(1990,1),freq=4)
RPPERM.ts<-ts(rrowRPPERM,start=c(1990,1),freq=4)
m2<-lm(oriAUTOS~PCPI+SP500+RVEH48+PBRENT+RUM)
summary(m2)
a<-m2$residuals
b<-a[-1]
c<-a[-95]
plot(c,b,xlab="e_t",ylab="e_t+1",main="plot e_t+1 vs. e_t")
plot(m2$residuals,main="Residuals in Full Model")

```

```

abline(h=0)
m21<-
dynlm(AUTOS.ts~RUM.ts+RVEH48.ts+PBRENT.ts+PCPI.ts+RPPERM.ts+L(AUTOS.ts))
summary(m21)
acf(m21$residuals)
plot(m21$residuals,type="p",main="residuals in time series model")
abline(h=0)

```

```

FirstDerivRUM <- deriv.fd(RUM, 1)
FirstDerivPBRENT <- deriv.fd(PBRENT, 1)
FirstDerivAUTOS <- deriv.fd(AUTOS, 1)
FirstDerivSP500<-deriv.fd(SP500,1)
FirstDerivYD<-deriv.fd(YD,1)
FirstDerivPCPI<-deriv.fd(PCPI,1)
FirstDerivRPPERM<-deriv.fd(RPPERM,1)
FirstDerivRVEH48<-deriv.fd(RVEH48,1)

```

```

meanRUM=mean(FirstDerivRUM$coefs)
sdRUM=sd(FirstDerivRUM$coefs)
zscoreRUM=(FirstDerivRUM$coefs-meanRUM)/sdRUM
zscoreRUM=abs(zscoreRUM)

```

```

meanPBRENT=mean(FirstDerivPBRENT$coefs)
sdPBRENT=sd(FirstDerivPBRENT$coefs)
zscorePBRENT=(FirstDerivPBRENT$coefs-meanPBRENT)/sdPBRENT
zscorePBRENT=abs(zscorePBRENT)

```

```

meanAUTOS=mean(FirstDerivAUTOS$coefs)
sdAUTOS=sd(FirstDerivAUTOS$coefs)
zscoreAUTOS=(FirstDerivAUTOS$coefs-meanAUTOS)/sdAUTOS
zscoreAUTOS=abs(zscoreAUTOS)

```

```

meanSP500=mean(FirstDerivSP500$coefs)
sdSP500=sd(FirstDerivSP500$coefs)
zscoreSP500=(FirstDerivSP500$coefs-meanSP500)/sdSP500
zscoreSP500=abs(zscoreSP500)

```

```

meanYD=mean(FirstDerivYD$coefs)
sdYD=sd(FirstDerivYD$coefs)
zscoreYD=(FirstDerivYD$coefs-meanYD)/sdYD
zscoreYD=abs(zscoreYD)

```

```

meanPCPI=mean(FirstDerivPCPI$coefs)
sdPCPI=sd(FirstDerivPCPI$coefs)
zscorePCPI=(FirstDerivPCPI$coefs-meanPCPI)/sdPCPI
zscorePCPI=abs(zscorePCPI)

```

```

meanRPPERM=mean(FirstDerivRPPERM$coefs)
sdRPPERM=sd(FirstDerivRPPERM$coefs)
zscoreRPPERM=(FirstDerivRPPERM$coefs-meanRPPERM)/sdRPPERM
zscoreRPPERM=abs(zscoreRPPERM)

```

```

meanRVEH48=mean(FirstDerivRVEH48$coefs)
sdRVEH48=sd(FirstDerivRVEH48$coefs)
zscoreRVEH48=(FirstDerivRVEH48$coefs-meanRVEH48)/sdRVEH48
zscoreRVEH48=abs(zscoreRVEH48)

```

```

matrix=cbind(zscoreRUM,zscoreRVEH48,zscorePBRENT,zscoreYD,zscorePCPI,zscoreSP500,zscoreRPPERM)

```

Python: (Processing Data)

```

import numpy as np
from numpy import genfromtxt
import csv
def matrix_replacement ():
    np.set_printoptions(suppress=True)
    m = genfromtxt('zscore.csv', delimiter=',')
    q1= float(raw_input("what you want for RUM: "))
    q2= float(raw_input("what you want for RVEH48: "))
    q3= float(raw_input("what you want for PBRENT: "))
    q4= float(raw_input("what you want for YD: "))
    q5= float(raw_input("what you want for PCPI: "))
    q6= float(raw_input("what you want for SP500: "))
    q7= float(raw_input("what you want for RPPERM: "))

    bounds = []
    bounds.append(q1)
    bounds.append(q2)
    bounds.append(q3)
    bounds.append(q4)
    bounds.append(q5)
    bounds.append(q6)
    bounds.append(q7)
    for col in range (7):

        for row in range (100):
            if(m[row, col] < bounds[col]):
                m[row, col] = 0
            else:
                m[row, col] = 1

    m = m[:96, :]

```

```
d = genfromtxt('data1.csv', delimiter=',')
result = np.zeros((96, 7))
for col in range (7):
    for row in range (96):
        result[row,col] = m[row, col] * d[row, col]
with open('result.csv', 'wb') as f:
    writer = csv.writer(f)
    writer.writerow(["RUM", "RVEH48", "PBRENT", "YD", "PCPI", "SP500",
"RPPERM"])
    np.savetxt(f, result ,delimiter=',', fmt = '%10.5f')
matrix_replacement()
```

Bibliography

1. *"Auto Sales Definition | Investopedia."* Investopedia. Investopedia, 19 Nov. 2003. Web. 27 Mar. 2015.
2. *"Snopes.com: Slow Boiled Frog."* Snopes.com: Slow Boiled Frog. Snopes, 12 Jan. 2009. Web. 3 Apr. 2015.
3. *"Gross Domestic Product (GDP) Definition."* OECD Glossary of Statistical Terms - Gross Domestic Product (GDP) Definition. The Organisation for Economic Co-operation and Development, 1 July 2002. Web. 5 Apr. 2015.
4. *"Unemployment Rate"* The Saylor Foundation. 20 Jun. 2012. Web. 29 Mar. 2015
5. *"Standard & Poor's 500 Index (S&P 500) Definition | Investopedia."* Investopedia. Investopedia, 26 Nov. 2003. Web. 15 Apr. 2015.
6. *"CPI News Releases."* U.S. Bureau of Labor Statistics. U.S. Bureau of Labor Statistics, n.d. Web. 29 Apr. 2015.
7. *"Inflation Definition | Investopedia."* Investopedia. Investopedia, 20 Nov. 2003. Web. 13 Apr. 2015.
8. Tom Lyche & Knut Mørken. *Spline Methods Draft*. Department of Informatics. Centre of Mathematics for Applications. University of Oslo. Apr. 5. 2011. Web. 13 Apr. 2015.
9. James Ramsay, Giles Hooker and Spencer Graves. *Functional Data Analysis with R and MATLAB*. (Springer, 2009). 15 Apr. 2015.
10. Ryan Tibashirani. *Smooth Spline*. Department of Statistics and Machine Learning Department. Carnegie Mellon University. May.1.2014. Web. 16 Apr. 2015.

11. Hutcheson, G.D. (2011). Ordinary Least-Squares Regression. In L. Moutinho and G.D. Hucheson, *The SAGE Dictionary of Quantitative Management Research*. Pages 224-228.