Auto Sales Prediction with attention to the Parable of the boiled frog: Functional Data Analysis and Time Series Forecasting.

Undergraduate Honors Thesis by Shuaiji Li

Advised by Professor Ed Rothman

University of Michigan

Department of Statistics

1 Introduction

The parable of the boiled frog is an old warning story describing how frogs react differently to different living environments. Typically, if a frog is put directly into boiling water, it will immediately sense the heat and jump out, but if it is placed into cold water and boiled slowly, it will not sense the danger and will stay in the water until death [1].

This theory has many real-world applications. For instance, global warming has become a hot topic ever since NASA announced that 2009 was tied for the second warmest year in the modern record [2]. Prior to that announcement, the problem with a gradual increase in the average temperature of the earth's atmosphere did not grab enough public attention, mainly because the average variation of land-ocean temperature over a five-year period is normally less than 0.2 degrees Celsius [3]. Not many people would notice this change. However, when the public was told that, in total, average global temperatures have increased by about 0.8 degrees Celsius (1.4 degrees Fahrenheit) since the late nineteenth century [2], which is a significantly dangerous number viewed on a historical timeline, people noticed this change instantly and the rate of increase of the global temperature slowed down due to governments' and the public's immediate efforts.

The same effect may happen to the auto sales market. When considering the prediction of auto sales, a small change in one of the contributing predictors is similar to putting a frog into cold water and heating it gradually: the response variable will react slowly to the little change and balance itself to follow the general trend. Conversely, a significant change in a predictor is more likely to draw public attention, and the auto sales volume will demonstrate an immediate subsequent change to comply with this variation. Previous research has shown that auto sales volumes can be predicted effectively and precisely through Support Vector Regression combined with the Particles Swarm Optimization algorithm (PSO-SVR), which optimizes the

regression parameter with a small mean absolute percentage error [4].

Instead of pursuing precision in estimation, this project focuses on identifying potential predictors that contribute to auto sales predictions, as well as the sensitivity of significant predictors that lead to an accurate and reasonable predictive model. This project uses quarterly data from the first quarter of 1990 to the fourth quarter of 2013 in the United States of America. The data is processed through functional data analysis methods in R and the resulting multiple linear regression model will be validated by an autoregressive integrated moving average model based on different auto sales time series.

2 Functional Data Analysis

Functional data analysis is a technique that can be used to find rates of changes or derivatives of the curve by fitting data points to functional models [5]. A typical way to start is by constructing basis functions with parameters that are easy to estimate and can accommodate curve features properly. A set of functional blocks $\theta_k, k = 1,...K$ within a linear combination are called basis functions. A function $x(t)$ expressed in this way will have a linear basis function expansion:

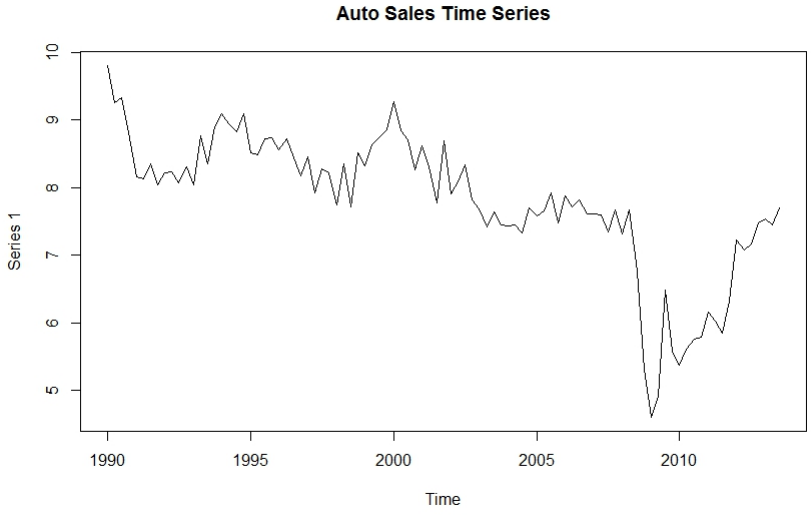
$$x(t) = \sum_{k=1}^{K} c_k \theta_k(t) = c'\theta(t), \tag{1}$$

where parameters c_k are the coefficients of the expansion. Since polynomials consisting of monomial and constant basis systems are less useful when complex functional shapes are required [6], this study considers two other types of basis functions systems that are widely used: Fourier and B-splines basis systems.

Fourier system is typically used for periodic datasets, while Splines and B-splines are commonly used when the data are non-periodic [7]. Splines are piecewise polynomials with domain segmented by knots with matching derivatives to

some order of the knots, and B-splines are convenient basis functions for splines with desirable numerical properties. Particularly for data such as civilian unemployment rates (RUG), consumer sentiment, and S&P 500's index (SP500), which are apparently non-periodically distributed, the B-spline basis system is commonly used for generating spline functions.

To examine the auto sales data, this study converted quarterly observations to time series with frequency four. The curve looks like:



The graph above shows that this curve is a non-stationary time series with many random fluctuations. To better capture the rate of changes of the curve, the raw data needs to be smoothed. A proper smoothing parameter λ that penalizes the derivatives of spline functions should be selected at first. In other words, λ controls the extent of smoothing, which aims to impose a penalty on the roughness of functions so that issues such as overfitting can be avoided. Given λ, the fitting function x(t) is chosen to minimize the following equation:

$$\sum_j [y_j - x(t_j)]^2 + \lambda \int [D^2 x(t)]^2 dt \tag{2}$$

where y_j is the observed data, $x(t_j)$ is the fitting function, and $[D^2 x(t)]^2$ is the curvature at time t [8]. As λ increases from 0, the curvature becomes increasingly penalized, leading to a smoother fit with smaller variance. However, a smoothing

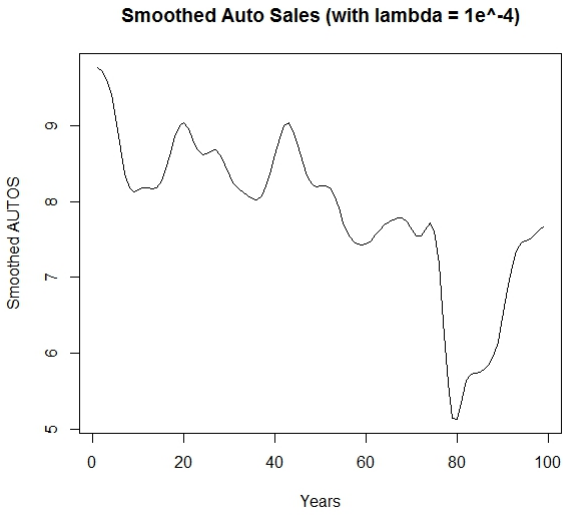
parameter closer to 0 will better keep the initial shape of the raw data and tend to reduce bias.

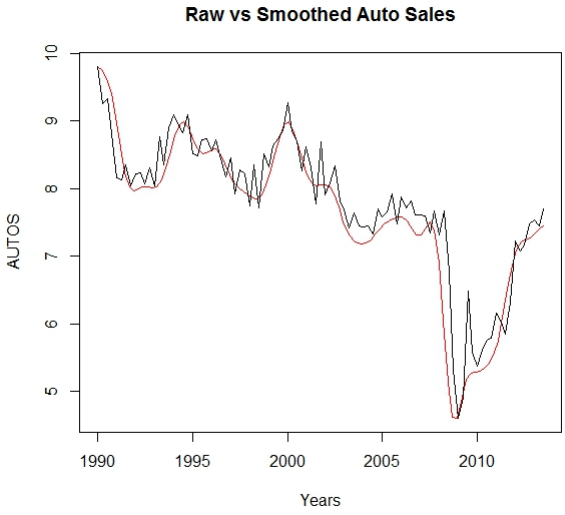
A typical procedure employed to generate a smoothing parameter that gives small mean squared error (MSE) for the target model chooses λ to minimize the generalized cross-validation method (GCV):

$$GCV(\lambda) = (\frac{n}{n - df(\lambda)})(\frac{SSE}{n - df(\lambda)}) \quad (3)$$

Where $df(\lambda) = trace[H(\lambda)]$, $H(\lambda)$ is a smoothing matrix as a function of λ, and SSE is the sum of squared errors predicting each observation from the rest. The designers of this method, Peter Craven and Grace Wahba [9], simulated a Monte Carlo experiment with several smoothed functions to estimate the average squared errors of λ. Their results demonstrated that the minimized value of average squared errors with simulated smoothed functions was close to the minimum of the true error value, showing that the estimator $\hat{\lambda}$ from GCV is a good smoothing parameter.

For this project, six splines were selected to carefully control the curvature of the second derivative. Thus, the derivative of the fourth order polynomial was penalized and the quarterly auto sales data from 1990 to 2013 was smoothed. The GCV returns estimated $\lambda \approx 0.0000316$, which is close to $10^{-}4$. The resulting smoothed curve with a smoothing parameter equals $10^{-}4$, as shown in the graphs below:





The smoothed spline functions fit the data well. Based on the same smoothing parameter, the derivative of functions can be calculated easily to support the future analysis.

3 Regression Analysis

3.1 Model Selection

Ordinary least squares (OLS) is a popular tool used to estimate the unknown parameters in linear models [10]. Considering the complex nature of this project's dataset, a multiple linear regression model predicting a single response variable with a linear function of more than two predictor variables was chosen. The key idea is to determine a linear combination of potential contributing variables that form a regression model that fits well.

There are two major types of variables selection approaches: the testing-based and criterion-based methods. For testing-based approaches such as backward elimination and forward selection, the idea is to test the significance of predictors and eliminate or add them based on individual p-values. The main problem with these methods is that variables not selected can still be correlated with the response, even though they do not improve the fit enough to be included. Smaller models therefore tend to be selected more often than would be desirable for prediction purposes. For

criterion-based approaches, however, models are chosen to optimize a criterion which balances goodness-of-fit and model size, with no p-values involved. To optimize the fit in this project, the second type of variable selection approach was employed and in, particular, the Akaike Information Criterion (AIC) was used with the equation:

$$AIC = nlog(RSS/n) + 2(p+1) \tag{4}$$

where $(2p+1)$ stands for the number of estimated parameters and (RSS/n) is the likelihood function for the model [10]. The combination of contributing variables with minimum AIC value is desired.

Through background research, this project considered the civilian unemployment rate (RUG), two-year interest rates for car loans (RVEH48), the consumer price index for all items (PCPI), the producer price index for finished goods (PPI), the federal funds rate (RFF), disposable yearly income (YD), regular retail gasoline prices (Gas), and lagged auto sales within one quarter (AUTOS_Lagged) as the potential contributing variables. The aim of including lagged auto sales as a predictor into the model was to investigate the application of the parable of the boiled frog on auto sales predictions. A linear model based on these predictors was created. The result shows a model with the smallest AIC equal to -172.54, including the variables lagged auto sales, unemployment rates, two-year interest rates, gasoline prices, and consumer price index:

$$AUTOS = \beta_0 + \beta_1(AUTOS\_Lagged) + \beta_2(RUG) + \beta_3(RVEH48) + \beta_4(PCPI) + \beta_5(Gas) + \epsilon_t$$

To further refine the model, another criterion-based approach, the adjusted R^2 method, was employed to implement the selection again. The definition of adjusted R^2 is:

$$R_a^2 = 1 - \frac{n-1}{n-(p+1)}(1-R^2) \tag{5}$$

where R^2 stands for the coefficient of the determination of regression, which is the proportion of the variance of response variables that can be explained effectively by predictor variables. Equation (5) shows that adjusted R^2 will decrease as more predictors are added if the loss of degrees of freedom covers the increase in model fit. Since maximizing R_a^2 is equivalent to minimizing $RSE\hat{\sigma}$ [11], the model with

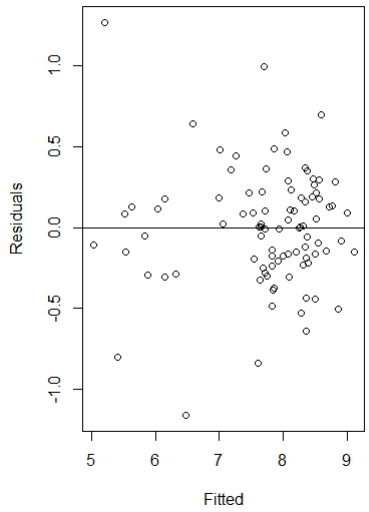
maximized R_a^2 was selected as the prediction model. Building upon the result from AIC method shows:

```
Subset selection object
Call: regsubsets.formula(AUTOS ~ AUTOS_Lagged + RUG + PCPI + RVEH48 +
    Gas, data = data)
5 Variables  (and intercept)
             Forced in Forced out
AUTOS_Lagged     FALSE      FALSE
RUG              FALSE      FALSE
PCPI             FALSE      FALSE
RVEH48           FALSE      FALSE
Gas              FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
         AUTOS_Lagged RUG PCPI RVEH48 Gas
1  ( 1 ) "*"          " " " "  " "    " "
2  ( 1 ) "*"          " " "*"  " "    " "
3  ( 1 ) "*"          " " "*"  "*"    " "
4  ( 1 ) "*"          "*" "*"  "*"    " "
5  ( 1 ) "*"          "*" "*"  "*"    "*"
```

presenting the same model with the largest adjusted R^2 at subset 5. Based on the table, the linear combination containing five variables listed is identified as the combination that best captures the general pattern of the dataset.

3.2 Diagnostics

Diagnostics is a crucial procedure that helps to test the validity of a model. Diagnostics always involves checking the assumption of normality, collinearity of predictors, and correlated errors. The fundamental assumption of ordinary least squares is that errors are independent and identically normal distributed with a mean of zero and variance $\sigma^2 I$. To confirm homoscedasticity (constant variance) of error terms, the residuals versus the fitted value of the model can be plotted:



The resulting graph shows no apparent sign that the residuals tend to converge to or diverge from zero, and there is no indication that points move toward value -1. The concern for heteroscedasticity (non-constant variance) and non-linearity was thus reduced.

The next step was to detect if collinearity existed among predictors. In statistics, collinearity (or multicollinearity) often refers to the situation where two or more predicting variables are correlated in a multiple regression model [12]. Although collinearity will not affect the general predicting power of a regression model, it may lead to an imprecise estimate of regression coefficients β_i, so that significant predictors may be missed due to a non-accurate t-test. One way to test collinearity is by checking the variance inflation factor (VIF) of predictors, which is formulated from the following equation:

$$Var(\widehat{\beta_j}) = \frac{\sigma^2}{1-R_j^2}\frac{1}{(n-1)Var(X_j)} \tag{6}$$

where $Var(\widehat{\beta_j})$ is the estimated variance of each estimated coefficient, n is the sample size, σ is the rooted mean square error, and $1-R_j^2$ is the variance inflation factor. Running function vif from R-package “car” [13], the factor index for each predictor was calculated:

AUTOS_Lagged	RUG	PCPI	RVEH48	Gas
6.236214	3.883123	22.657907	5.176218	9.913060

Of note is that the consumer price index (PCPI) and regular retail gasoline price (Gas) have large VIF values that may be problematic. The correlation table also shows that PCPI and Gas are highly correlated, with covariance 0.9276:

	AUTOS_Lagged	RUG	PCPI	RVEH48	Gas
AUTOS_Lagged	1.0000000	-0.7241805	-0.7427806	0.6546989	-0.6581645
RUG	-0.7241805	1.0000000	0.3871891	-0.4875441	0.4353077
PCPI	-0.7427806	0.3871891	1.0000000	-0.8610233	0.9276005
RVEH48	0.6546989	-0.4875441	-0.8610233	1.0000000	-0.7863504
Gas	-0.6581645	0.4353077	0.9276005	-0.7863504	1.0000000

To deal with this collinearity, this project considered dropping one of the predictors between PCPI and Gas, as well as finalizing the prediction model with the updated variable choice. After several trials of manipulations, the producer price index was found to be more significant than gas price to auto sales in terms of p-values. The individual inflation factor of each coefficient dropped substantially for the new model without Gas. The function VIF from R-package “fmsb” [14] was also employed to evaluate the generalized variance inflation factor of two multiple regression models. The result demonstrates that the new one has a smaller VIF value, which indicates that the overall level of collinearity decreases.

Thus, the refined linear regression model is as follows:

$$AUTOS = \beta_0 + \beta_1(AUTOS\_Lagged) + \beta_2(RUG) + \beta_3(RVEH48) + \beta_4(PCPI) + \epsilon_t$$

with the summary table:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.40597 -0.23435  0.02091  0.22481  1.20504

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.360972   1.575322   4.673 1.05e-05 ***
AUTOS_Lagged  0.634384   0.085785   7.395 7.43e-11 ***
RUG          -0.115240   0.039903  -2.888 0.004868 **
PCPI         -0.013941   0.003423  -4.073 0.000101 ***
RVEH48       -0.164978   0.052497  -3.143 0.002275 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3951 on 89 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8566,    Adjusted R-squared:  0.8502
F-statistic: 132.9 on 4 and 89 DF,  p-value: < 2.2e-16
```

The 0.8566 multiple R-squared indicates a general good fit of the model, with all predictor variables being significant on a 95% confidence interval. These selected variables make sense not only due to model selection and diagnostics, but also because of their influence on car sales in reality. The lagged auto sales is the most significant predictor with the largest coefficient of estimate, which matches with this project’s expectation for the parable of the boiled frog’s application in the car sales market. When people notice that overall car sales increase, they are more likely to be optimistic about the market. The consumer price index is a measure that examines the average price changes associated with the change of the cost of living [15]. This measure is lagged in nature because the government usually adopts future macroeconomic policies based on consumers’ costs and average price levels of past markets. Similar to the two-year interest rates, PCPI has a negative relation to auto sales since people noticing higher interest rates and greater inflation are less likely to purchase new cars for a period of time. However, to explain the sensitivity of unemployment rates on prediction, an indicator variable needs to be set and the cut points where the rate of change has the largest impact on auto sales needs to be found.

3.3 Indicator

The first derivative of a predictor demonstrates the changing rate of the variable to

response. To find the derivative values, one needs to implement smoothed functional data with well-defined roughness. The linear model with the indicator of this rate of change is:

$$AUTOS = \beta_0 + \beta_1(AUTOS_Lagged) + \beta_2(RUG) + \beta_3(RVEH48) + \beta_4(PCPI) + \beta_5 I(ZRUG < C_1) + \beta_6 I(ZRUG > c_2) + \epsilon_t$$

where I denotes the indicator variable. The purpose is to find the appropriate cut points c_1 and c_2 for unemployment rates, which leads to great change in auto sales. This project used z-scores of the first derivative of unemployment rates. If at time t the z-score lies below the specific cut point c_1, then the first indicator is 1. Otherwise, the indicator equals 0. The second indicator is 1 when the z-score exceeds c_2 and 0 otherwise. The errors are assumed to be independent and identically normal distributed here.

Trying with various cut points, this project found 90% (z-score 1.3) and 23% (z-score -0.75) gave the best outcome:

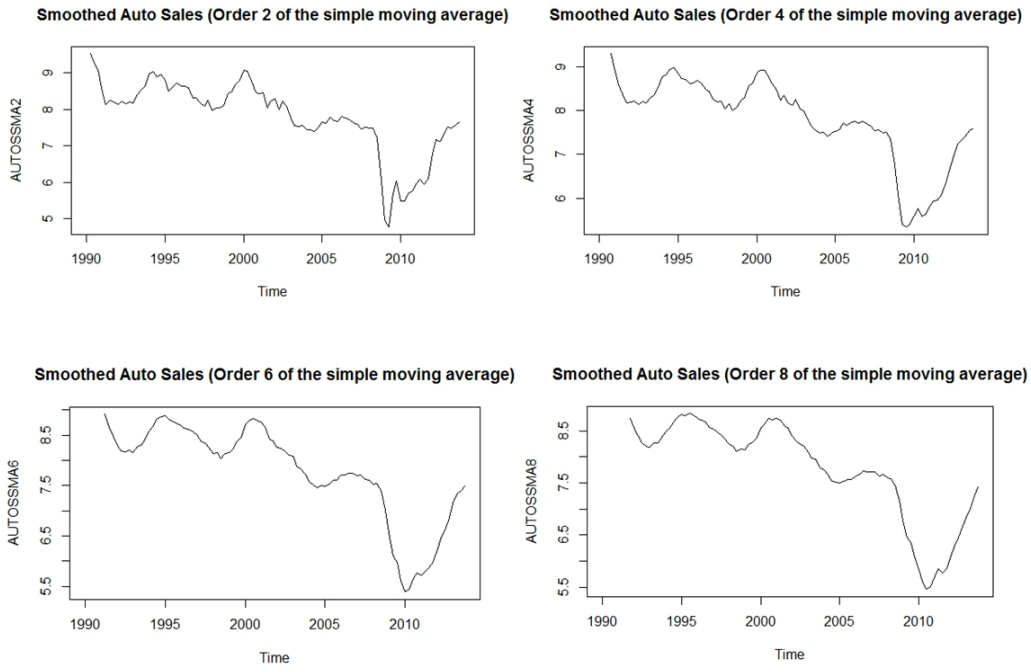
```
Residuals:
     Min       1Q   Median       3Q      Max
-1.15313 -0.20840 -0.01165  0.20803  1.34481

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   7.278860   1.535691   4.740 8.28e-06 ***
AUTOS_Lagged                  0.556782   0.089530   6.219 1.69e-08 ***
RUG                          -0.117348   0.042355  -2.771 0.006841 **
PCPI                         -0.012685   0.003372  -3.762 0.000305 ***
RVEH48                       -0.103674   0.056686  -1.829 0.070837 .
I(zscoreRUG < (-0.75))TRUE    0.173234   0.137572   1.259 0.211317
I(zscoreRUG > (1.3))TRUE     -0.331530   0.151645  -2.186 0.031484 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3846 on 87 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8672,    Adjusted R-squared:  0.8581
F-statistic: 94.72 on 6 and 87 DF,  p-value: < 2.2e-16
```

The model is reasonable on both directions of the derivatives. If the changing rate of unemployment rate grows increasingly, people tend to avoid purchasing new vehicles since they are afraid of losing jobs. In contrast, when the rate of change increases

decreasingly, people are less worried about their positions and are willing to enlarge their budgets for vehicle purchases.

4 Time Series Forecasting

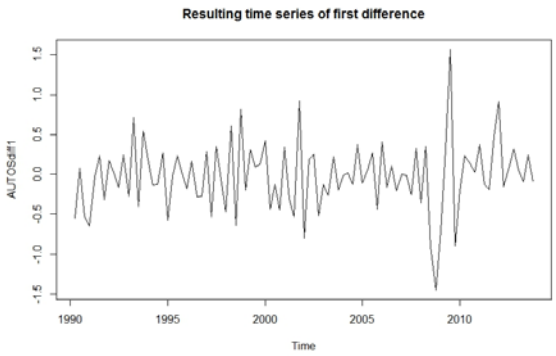
The reason to implement time series forecasting is to find an Autoregressive Integrated Moving Average model (ARIMA) that validates the previous ordinary least squares model in terms of the lagged interval. Auto sales data forms a non-stationary time series with no apparent seasonal trend. To deal with this non-seasonal data sample, this project decomposed the time series to estimate the trend component and irregular component [16]. An order (span) of the simple moving average needed to be specified before smoothing the data. As such, this project tried from order two to eight, with smoothed curves as shown in the graphs below:



One can see that order four of the simple moving average shows the best balance between the extent of roughness and the original shape of the auto sales data. This result resembles to the six spline functions that were selected for smoothing in the functional data analysis section. Because the time series cannot be explained by employing an additive model with always an increasing or decreasing trend, the exponential smoothing method is not useful. Accordingly, the ARIMA model was

chosen to conduct the predictions.

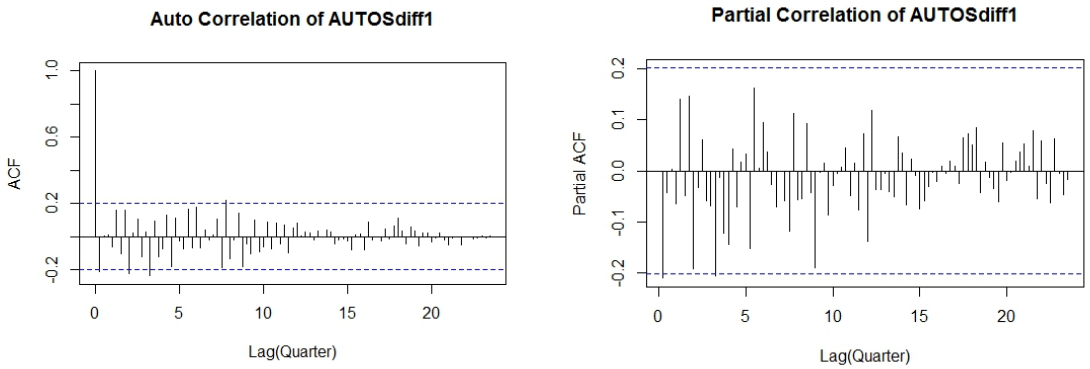
Since the ARIMA model is defined for stationary time series [17], the series must first be differenced until a stationary time series is generated. The order of differencing is parameter d of the ARIMA (p,d,q) model. The result of the first difference is shown below:



The first difference appears to be stationary in mean since the level of the series stays roughly constant over time. Next, parameters p and q must be selected, as these specify the order of autoregressive (AR) and moving average (MA). Here, the AR model is more important because it regresses current values on previous values in the same time series [18], using the following equation:

$$Y_t = c + \sum_{i=1}^{p} \beta_i Y_{t-1} + \epsilon_t \tag{7}$$

where β_i are parameters of the model, ϵ_t is noise, and c is some constant. It is crucial to detect a proper p that defines the number of lag used for the forecasting model. Through the autocorrelation and partial autocorrelation plots shown below:



the following possible ARIMA models can be listed:

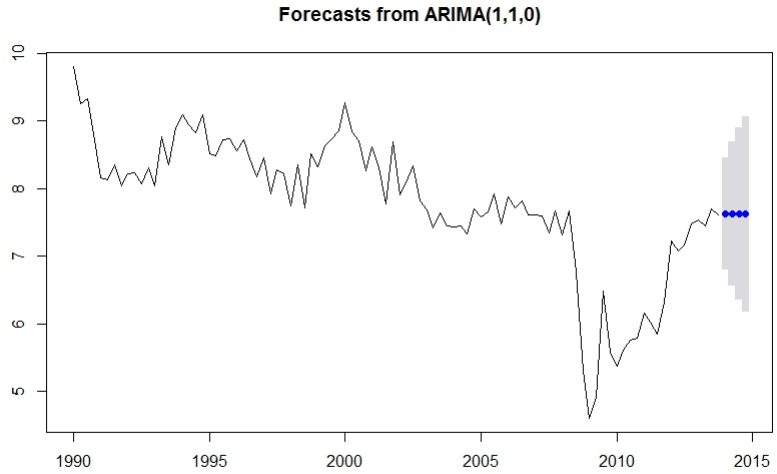
- ARIMA (1,0,0), with an autoregressive parameter p equal to 1, since the partial autocorrelation plot approaches zero after lag one and the autocorrelation plot tails off to zero.
- ARIMA (0,3,0), with a moving average parameter q equal to 3, since the autocorrelation plot approaches zero after lag three and the partial autocorrelation plot tails off to zero.

The law of parsimony is thus employed, which says that the hypothesis with the fewest assumptions should be selected [19]. Consequently, the candidate model with ARIMA (1,1,0) was decided upon. The auto.arima function in the R-package "forecast" validated the project's guess by presenting the same model:

```
Series: AUTOS.ts
ARIMA(1,1,0)

Coefficients:
          ar1
      -0.2086
s.e.   0.1006

sigma^2 estimated as 0.1827:  log likelihood=-53.56
AIC=111.13   AICc=111.26   BIC=116.23
```

The 0.1 standard error is reasonable given the large fluctuation of data from 2007 to 2012. Finally, the actual predictions for auto sales in 2014 with 95% prediction intervals were conducted, as shown below:

	Point Forecast	Lo 95	Hi 95
2014 Q1	7.631732	6.798489	8.464975
2014 Q2	7.628033	6.565430	8.690636
2014 Q3	7.628805	6.358728	8.898882
2014 Q4	7.628644	6.184138	9.073150



5 Conclusion

When significant predicting variables such as unemployment rates lead a big change with large rates (90% and 23%), auto sales in the United States will be substantially influenced by this lag effect. The results of this study demonstrates the validity of the parable of the boiled frog theory in the car sales market, and the number of quarters lagged is justified by finding the parameters of the ARIMA model that have a good fit.

Potential future research may involve studying data over a broader time period. This project conducted regression and time series analyses on 96 observations in 24 years, which somehow limited the accuracy of the project's predictions due to greater variability of the sample mean and a larger estimated variance. With a larger data set, we may employ some resampling methods such as K-fold cross-validation to compare multiple models and validate our choice based on out-of-sample errors. Additionally, the consideration of potential predictors is still narrow, mainly due to the limitation of data. Finally, further research can enlarge the selection scope to multidisciplinary applications and test variables such as consumer sentiments, crime rates and so on, if

the data is supportive enough.

6 Acknowledgments

I would like to thank for Dr. Ed Rothman from the Department of Statistics at the University of Michigan for his valuable comments and generous support. I would also like to thank Dr. Robert Keener from the Department of Statistics, University of Michigan for his careful review and kind suggestions. Thank George Fulton for providing data for this project.

Appendix

R-studio Codes

Functional Data Analysis:

```
library('splines')
library('Matrix')
library('fda')
data = read.csv("C:/Users/admin/Desktop/STATS RESEARCH/carsalesdata.csv",header=TRUE)
AUTOS <- data[,1]
RUG <- data[,2]

AUTOS.ts <- ts(AUTOS, start = c(1990,1),frequency = 4)
plot.ts(AUTOS.ts, main = "Auto Sales Time Series")

argvalsAUTOS = seq(1990,2013.75,len=nAUTOS)
nbasis = 99
AUTOS <-as.matrix(AUTOS)
xAUTOS = AUTOS
xAUTOS.factor <- factor(xAUTOS)
as.numeric(as.character(xAUTOS.factor))
basisobj = create.bspline.basis(c(1990,2013.75),nbasis)
fdParobjAUTOS = smooth.basis(argvalsAUTOS, xAUTOS, basisobj)
loglamvec = seq(0, 0.001, 0.00001)
loglamout = matrix(0,length(loglamvec),4)
m = 0
for (loglambda in loglamvec) {
   m = m + 1
   loglamout[m,1] = loglambda
   fdParobjAUTOS$lambda = 10^(loglambda)
   smoothlist = smooth.basis(argvalsAUTOS, xAUTOS,fdParobjAUTOS）
   xfd = smoothlist$fd
   loglamout[m,2] = smoothlist$df
   loglamout[m,3] = sqrt(mean((eval.fd(argvalsAUTOS, xfd) - xAUTOS)^2))
   loglamout[m,4] = mean(smoothlist$gcv)
}
indx = which.min(loglamout[,4])
(loglamout[index,4])

norder = 6
nbasisAUTOS = nAUTOS + norder -2
basisobjAUTOS = create.bspline.basis(c(1990,2013.75),nbasisAUTOS,norder)
lambdaAUTOS = 10^(-4)
fdParobjAUTOS = fdPar(fdobj=basisobjAUTOS, Lfdobj=4, lambda=lambdaAUTOS)
```

```
smoothlistAUTOS = smooth.basis(argvalsAUTOS, xAUTOS, fdParobjAUTOS)$fd

der.AUTOS <- deriv.fd(smoothlistAUTOS,1)
dercoeffAUTOS <- der.AUTOS$coefs
AUTOScoeff <- smoothlistAUTOS$coefs
plot(AUTOScoeff,lty=1,type = 'l',xlab = 'Years',ylab = 'Smoothed AUTOS',main = "Smoothed Auto Sales (with lambda = 1e^-4)")
plot(AUTOScoeff,lty=1,type = 'l',xlab = 'Years',ylab = 'AUTOS',col = 'red',axes = F)
par(new = T)
plot(AUTOS.ts,lty=1, xlab='', ylab='', main = "Raw vs Smoothed Auto Sales")

RUG_d = read.csv("C:/Users/admin/Desktop/STATS RESEARCH/RUG.csv",header=TRUE)
nRUG = 95
argvalsRUG = seq(1990,2013.75,len=nRUG)
RUG_d <-as.matrix(RUG_d)
xRUG = RUG_d [1,]
xRUG.factor <- factor(xRUG)
as.numeric(as.character(xRUG.factor))
norder = 6
nbasisRUG = nRUG + norder -2
basisobjRUG = create.bspline.basis(c(1990,2013.75),nbasisRUG,norder)
lambdaRUG = 10^(-4)
fdParobjRUG = fdPar(fdobj=basisobjRUG, Lfdobj=4, lambda=lambdaRUG)
smoothlistRUG = smooth.basis(argvalsRUG, xRUG, fdParobjRUG)$fd
der.RUG <- deriv.fd(smoothlistRUG,1)
dercoeffRUG <- der.RUG$coefs
D_RUG1 <- dercoeffRUG

meanD_RUG1 <- mean(D_RUG1)
sdD_RUG1 <- sd(D_RUG1)
zscoreRUG <- (D_RUG1 - meanD_RUG1)/sdD_RUG1
```

Regression Analysis:

```
AUTOS_Lagged <- data[,3]
RVEH48 <- data[,4]
PPI <- data[,5]
PCPI <-data[,6]
Gas <- data[,7]
YD <- data[,8]
RFF <- data[,9]
zscoreRUG <- data[,10]

library('leaps')
```

```
#AIC
lg<- lm(AUTOS ~ AUTOS_Lagged + RUG + PPI + PCPI + RVEH48 +     Gas + RFF + YD)
step(lg)
lg <- lm(AUTOS~AUTOS_Lagged + RUG + PCPI + RVEH48 + Gas)
summary(lg)
#adjusted r-squared
b = regsubsets(AUTOS~AUTOS_Lagged + RUG + PCPI + RVEH48 + Gas, data=data)
summary(b)
rs = summary(b)
which.max(rs$adjr2)

plot(lg$fitted, lg$residual,xlab="Fitted", ylab="Residuals")
abline(h=0)

library('car')
vif(lg)

X <- model.matrix(lg)
cor(X[,2:6])

library('fmsb')
lt <- lm(AUTOS~AUTOS_Lagged + RUG + PCPI + RVEH48)
VIF(lg)
VIF(lt)

fit <- lm(AUTOS~AUTOS_Lagged + RUG + PCPI + RVEH48 + I(zscoreRUG < (-0.75)) + I(zscoreRUG >
(1.3)))
summary(fit)
```

Time Series Forecasting:

```
library("zoo")
library("forecast")
library('timeDate')
library("TTR")
Forecast = read.csv("C:/Users/admin/Desktop/STATS RESEARCH/AUTOS.csv",header=F)
AUTOS <- Forecast[,1]

AUTOSSMA2 <- SMA(AUTOS.ts,n=2)
plot.ts(AUTOSSMA2,main = "Smoothed Auto Sales (Order 2 of the simple moving average)")
AUTOSSMA4 <- SMA(AUTOS.ts,n=4)
plot.ts(AUTOSSMA4,main = "Smoothed Auto Sales (Order 4 of the simple moving average)")
AUTOSSMA6 <- SMA(AUTOS.ts,n=6)
plot.ts(AUTOSSMA6,main = "Smoothed Auto Sales (Order 6 of the simple moving average)")
AUTOSSMA8 <- SMA(AUTOS.ts,n=8)
```

```
plot.ts(AUTOSSMA8,main = "Smoothed Auto Sales (Order 8 of the simple moving average)")

AUTOSdiff1 <- diff(AUTOS.ts,differences = 1)
plot.ts(AUTOSdiff1, main = "Resulting time series of first difference")

plot.acf <- function(ACFobj) {
   rr <- ACFobj$acf[-1]
   kk <- length(rr)
   nn <- ACFobj$n.used
   plot(seq(kk),rr,type="h",lwd=2,yaxs="i",xaxs="i",
        ylim=c(floor(min(rr)),1),xlim=c(0,kk+1),
        xlab="Lag",ylab="Correlation",las=1)
   abline(h=-1/nn+c(-2,2)/sqrt(nn),lty="dashed",col="blue")
   abline(h=0)
}

AUTOSdiff1.acf <- acf(AUTOSdiff1,lag.max = 95, main = "Auto Correlation of AUTOSdiff1", xlab =
"Lag(Quarter)") #plot a correlogram
acf(AUTOSdiff1,lag.max = 95, plot = F)#get the autocorrelation values
plot.acf(AUTOSdiff1.acf)

pacf(AUTOSdiff1,lag.max = 95, main = "Partial Correlation of AUTOSdiff1", xlab = "Lag(Quarter)")
#plot a correlogram
pacf(AUTOSdiff1,lag.max = 95, plot = F)

auto.arima(AUTOS.ts)#ARIMA(1,1,0)

(AUTOSarima <- arima(AUTOS.ts,order = c(1,1,0)))
(AUTOSforecasts <- forecast.Arima(AUTOSarima,h=4,level = c(95)))
plot.forecast(AUTOSforecasts)
```

References

[1] “Boiling frog.” *Wikipedia*. N.p.: Wikimedia Foundation, 18 Aug. 2016. Web. 1 Dec. 2016.

[2] “2009: Second Warmest year on record; end of Warmest decade.” Brian Dunbar, 28 Jan. 2010. Web. 2 Dec. 2016.

[3] “Global Land-Ocean Temperature Index (C).” *NASA*. 2016. Web. 2 Dec. 2016.

[4] X. Lu and X. Geng, "Car Sales Volume Prediction Based on Particle Swarm Optimization Algorithm and Support Vector Regression," *2011 Fourth International Conference on Intelligent Computation Technology and Automation*, Shenzhen, Guangdong, 2011, pp. 71-74.

[5] “Functional data analysis.” *Wikipedia*. N.p.: Wikimedia Foundation, 16 Oct. 2016. Web. 2 Dec. 2016.

[6] Ramsay, J O, Giles Hooker, and Spencer Graves. *Functional Data Analysis with R and MATLAB*. Dordrecht: Springer-Verlag New York, 2009. Print, pp.30

[7] Yuko Araki, Yuko. “Functional Data Analysis and Statistical Modeling.” *Kyushu Univ.* n.d. Web. 2 Dec. 2016.

[8] Ramsay, J O, Giles Hooker, and Spencer Graves. *Functional Data Analysis with R and MATLAB*. Dordrecht: Springer-Verlag New York, 2009. Print, pp.64

[9] Wahba, Grace, and Peter Craven. “Smoothing Noisy Data with Spline Functions.” *Numerische Mathematik* 24.5 (1975): 383–393. Web.

[10] “Akaike information criterion.” *Wikipedia*. N.p.: Wikimedia Foundation, 9 Nov. 2016. Web. 2 Dec. 2016.

[11] Nau, Robert. “Mathematics of simple regression.” *Fuqua School of Business Duke University*. 1 May 2016. Web. 3 Dec. 2016.

[12] “Multicollinearity.” *Wikipedia*. N.p.: Wikimedia Foundation, 18 Nov. 2016. Web. 3 Dec. 2016.

[13] John Fox and Sanford Weisberg (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

[14] Minato Nakazawa (2015). fmsb: Functions for Medical Statistics Book with some Demographic Data. R package versiono.5.2.
https://CRAN.R-project.org/package=fmsb

[15] Investopedia.com. “Consumer Price Index - CPI.” N.p.: Investopedia, 19 Nov. 2003. Web. 3 Dec. 2016.

[16] Avril Coghlan, Avril. *Using R for time series analysis — time series 0.2 documentation*. n.d. Web. 4 Dec. 2016.

[17] Nau, Robert. “Introduction to ARIMA models.” *Statistical forecasting: notes on regression and time series analysis*. 1 May 2016. Web. 4 Dec. 2016.

[18] State, The Pennsylvania. *14.1 - Autoregressive models*. 2016. Web. 4 Dec. 2016.

[19] “Occam’s razor.” *Wikipedia*. N.p.: Wikimedia Foundation, 28 Nov. 2016. Web. 4 Dec. 2016.