The Effects of Economic Factors on Total Vehicles Sold in the USA

Kyle Wu

2023-03-13

Data Application

In the United States, one of the main modes of transportation is the automobile. To the average consumer, it has seemed that the new car has been slowly getting out of reach, with the average price of a new vehicle currently sitting around $49,500 ("No End in Sight: New Vehicle Transaction Prices End 2022 at Record Highs, According to New Data from Kelley Blue Book" 2023). Despite the high prices of vehicles for many Americans a car is not only a luxury, but a necessity, and many citizens find themselves shelling out a large portion of their paychecks for their transportation. Even when individuals opt to purchase used cars, they are still often faced with prices that would have seemed exorbitant not that long ago.

Since this is the case, studying the United States car market over time will allow us to gain useful knowledge that will be of significance not only to the average consumer, but also for economists trying to understand what trends the American auto market may be facing going forwards and what factors influence automotive sales. Past research into the American auto market have been vital to our understanding of the forces driving the auto market. For example, it is well known that the chip shortage that occurred as a result of COVID-19 shutdowns, among other reasons led to a chip shortage that has in many ways created problems for the world economy ("Inflation and the Auto Industry: When Will Car Prices Drop" 2022). Since cars now heavily rely on computers to work, this resulted in many manufacturers around the world decreasing production projections, which decreased vehicle production, and partially led to the rapid rise in vehicle prices and sales patterns that broke from previous market trends. However, if we look at production figures, we can see that the domestic production of cars had been following a decreasing trend since the 90s, so researchers at Federal Reserve Economic Data (FRED) found that it is hard to say if COVID was fully responsible for the decreased production, or if it would have happened regardless ("Long-Term Trends in Car and Light Truck Sales" 2021). Research by FRED also indicated that despite the increase in population since the mid 1970s, the total number of vehicles sold has remained relatively flat aver the past few decades ("What's Been Drivin the Rise in Auto Prices Since COVID" 2022).

Looking at the data offered by the Federal Reserve could allow us to answer even more questions regarding the American auto market. If we take into account other economic factors, such as bank prime rate loans or gas prices, we can then try to measure what economic factors may most affect the sale of motor vehicles. Using the data we obtained and after determining factors that determine automotive sales, we can then create a forecast to determine how each factor relating to the automotive industry will change in the future. For example, we can try to answer the question of whether it is likely new vehicles will continue facing inflation or if it might become stable in the near future. Besides looking at the various factors individually, we can also look at the auto market holistically, asking what the future may be in terms of vehicle purchases in the United States and whether it is likely that vehicle purchases return to pre-COVID levels. For the average consumer, the questions that will be answered will allow them to perhaps better plan for the expenditure that comes with the purchase of a new car.

Furthermore, studying time series data of vehicles can allow us to better understand how or if certain policy changes may change vehicle prices or purchasing behavior. For example, we could potentially find time periods with varying federal funds rates, which influence bank prime loan rates to see if this changed the overall behavior of consumers.

Gathering all this data about the American auto market would then allow us to broadly gain an understanding

not only of factors affecting vehicle sales, but also of the health of the American economy due to the fact that vehicles are often the second most expensive possessions of individuals, second only to homes. Increased purchasing of vehicles would indicate that the American economy has been healthy and following a positive trend, whereas decreased vehicle purchases may indicate that the economy had been following a general undesirable trajectory.

Analysis of Empirical Properties

In all cases, the data I selected came from the Federal Reserve Economic Data database and all variables selected were recorded on a monthly basis and input into a format that was very neat and effective for time series analysis. The variables I have chosen are Total number of vehicles sold, new vehicle consumer price index, domestic auto production, fuel price index, and the bank prime rate. In this study, I will use total number of vehicles sold as the gauge of the american auto market, and the other variables will be used as predictors.

Since we will be creating time series regression models, we first want to get an overview of the data to see how each variable relates to the other, which we can do through a correlation plot.

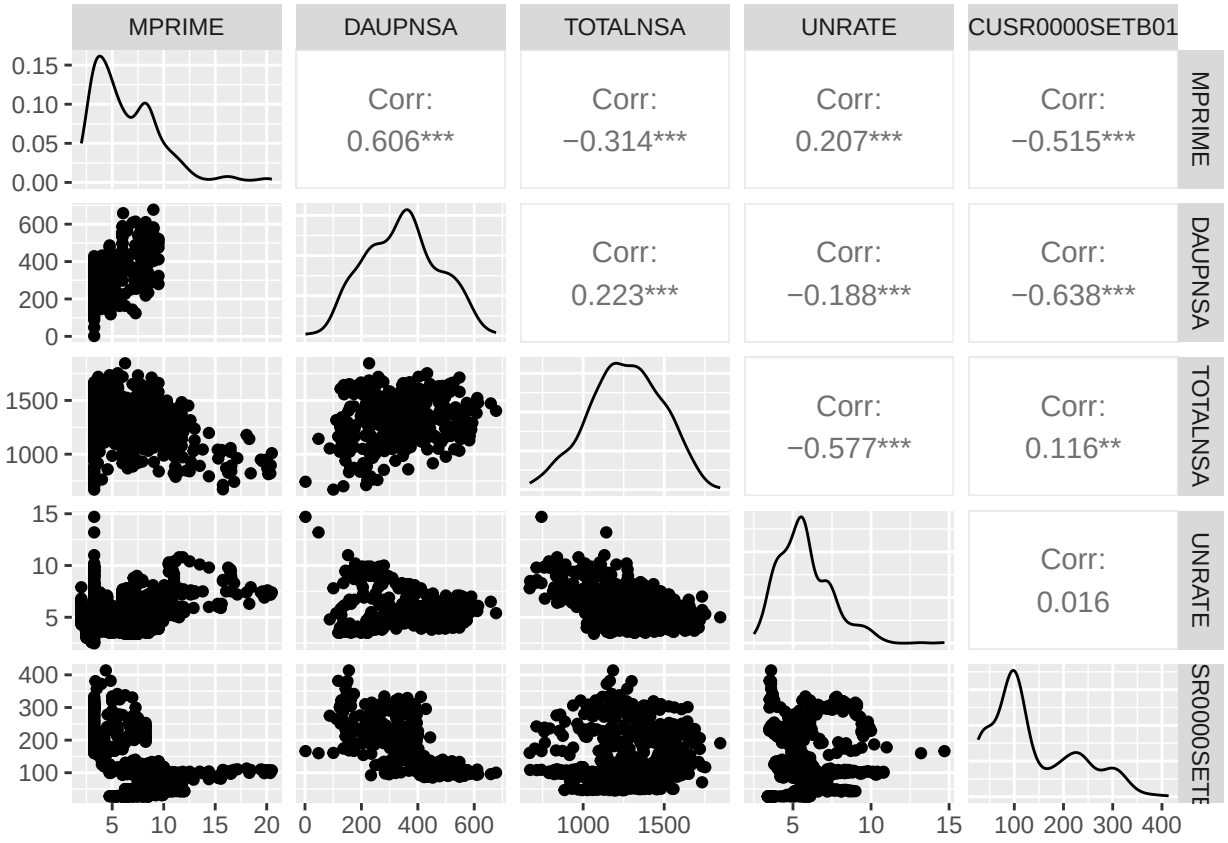


Figure 1: Correlation Plot.

From the correlation plot, we can see that all the variables have some form of statistical significance with the exception of the correlation between the gas price index and the bank prime rate loans. This would make sense because there really shouldn't be a reason why there should be a significant correlation between the variables especially considering that bank prime rate loans are assocaited with larger purchases such as for cars or houses.

We will now analyze the variables individually before talking about all the factors as they may relate to projecting future car sales.

The first variable we will look at is total vehicle sales.

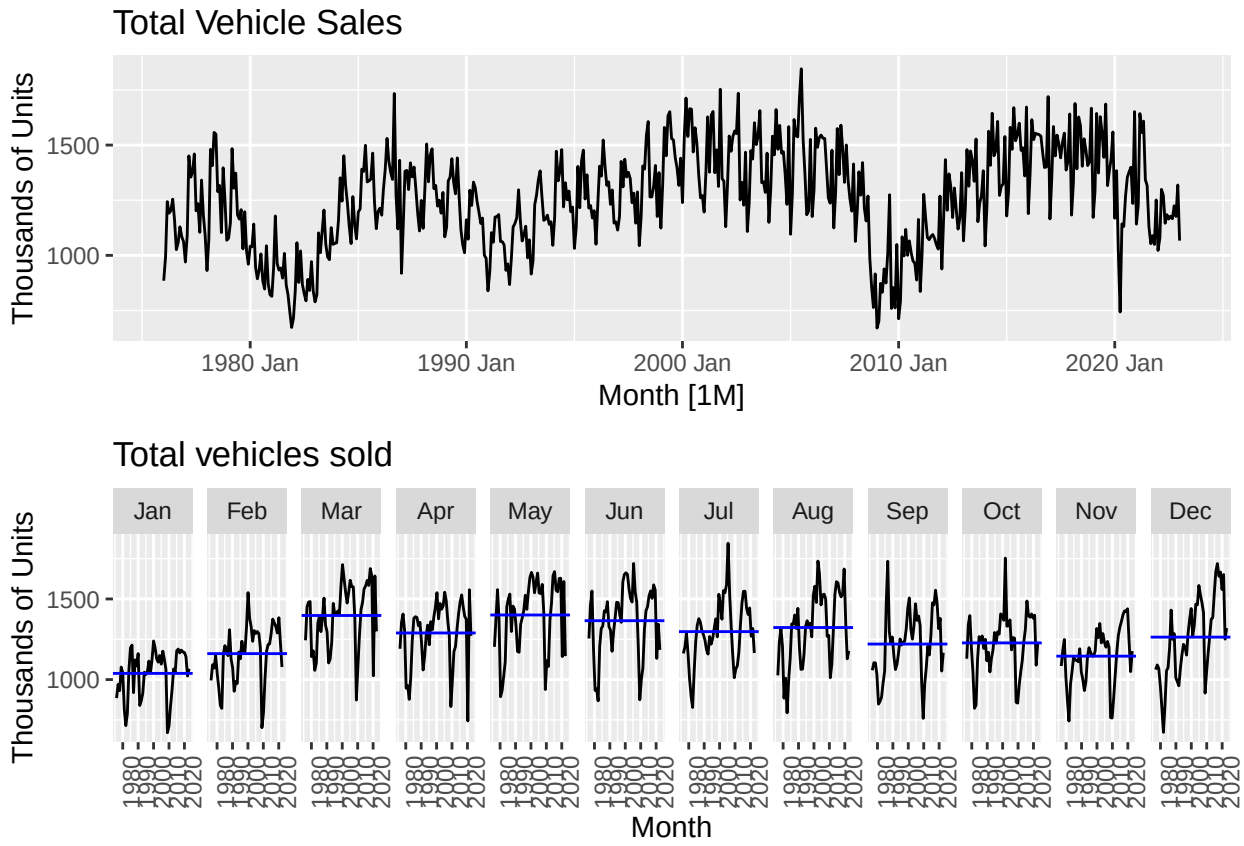


Figure 2: Total New Vehicle Sales time series plot (Top) and Total Vehicle Sales seasonal subseries plot (Bottom).

Our data for vehicles sold starts from January 1976, is recorded monthly, and does not have any missing values.The data was collected from the U.S. Bureau of Economic Analysis, which has formatted the data in a easily usable database. From the subseries plot, we see that the data definitely follows a seasonal pattern, as demonstrated by the fluctuating mean lines based on the month of the year. Additionally, from the scalloped shape of the plot, we can see that the data follows a seasonal trend of peaks and troughs every twelve months. Logically this makes sense, as it is well known that vehicle sales usually increase during the summer and tend to decrease during the winter. From the data we can determine the number of vehicles sold over the last 5 decades and whether there has been a general trend in vehicle sales.

We can now look at the consumer price index of new vehicles in the United States.

Records for the price index of gasoline start from February 1968 and are recorded monthly by the U.S. Bureau of Labor Statistics with no missingness present. From the data, we see that there has been a general increase in the price of gasoline from 1968 until the early 2000s before prices seem to level off, but with high volatility. We also see that following COVID, there was a large spike in gasoline prices, which has since gradually come down. This data will allow us to understand what trends are present in terms of gasoline prices and whether gasoline prices have an effect on the total number of vehicles sold. In other words, do high gasoline prices lead to less consumers buying vehicles?

The data for unemployment rate goes back to January 1948 and is recorded on a monthly basis and is collected by the U.S. Bureau of Labor Statistics and there is no missingess present in this time series. From our data, we see general cyclical fluctiations with periods with varying periods of higher vs lower unemployment. There does not appear to have been a consistent trend of unemployment over the lifecycle of the data. In more recent times, we see that there has been a decrease in the unemployment rate since after the Great

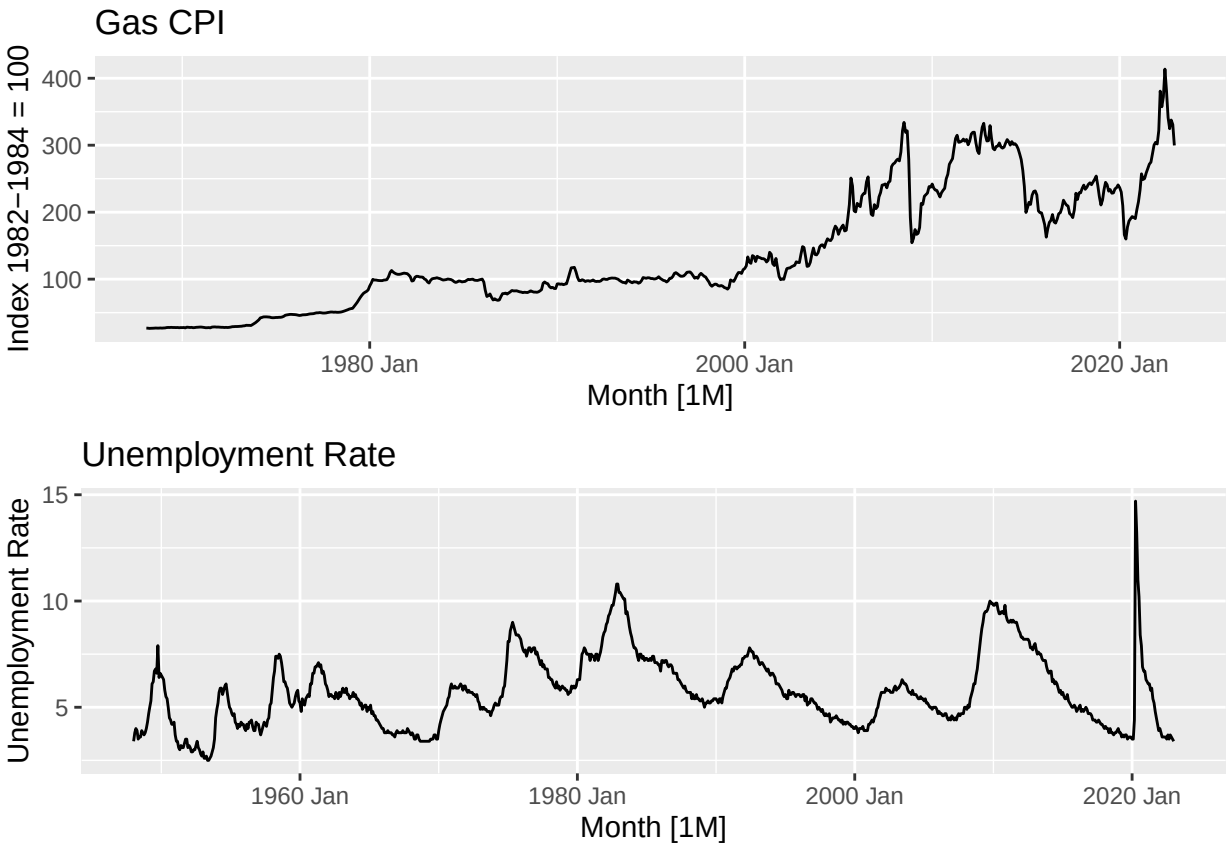


Figure 3: Time series plot of Gasoline CPI (Top) and Unemployment Rate (Bottom).

Recession following 2008 with a large spike in unemplopyment following the beginning stages of the COVID-19 pandemic. We can use the unemployment data to determine whether there is a relationship between rates of unemployment and the number of total vehicles sold in the United States, and if a relationship is present, we can attempt to determine the strength of the relationship.

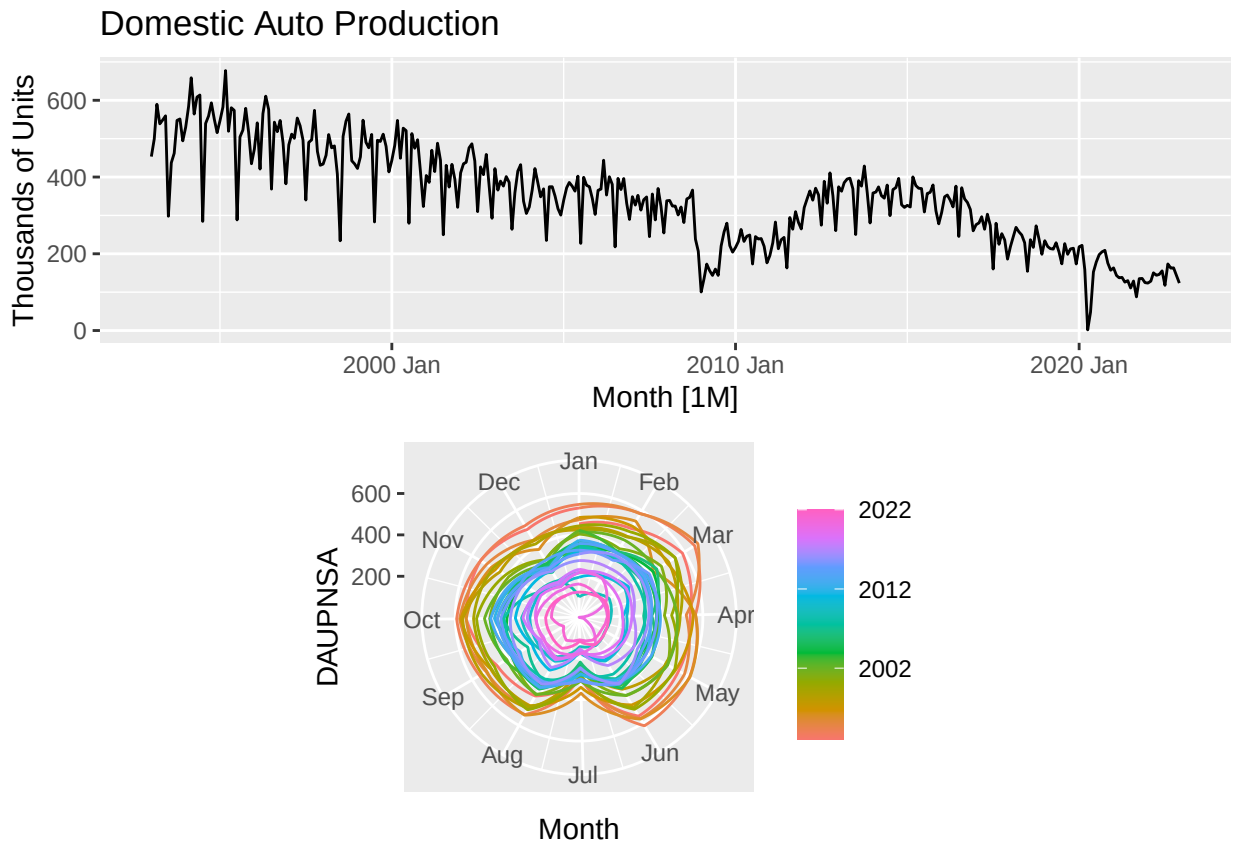


Figure 4: Time series plot of Domestic Auto Production (Top) and Domestic Auto Production Seasonal Plot (Bottom).

For Domestic Auto Production, we have monthly recorded data from the U.S. Bureau of Economic Analysis with no missingess present in the data. Overall, we appear to see volatility month-to-month and pretty consistent seasonal patterns with July being a month with consistently low production numbers in relation to other months and March and October having relatively high production numbers. This data alone can answer a couple questions about the United States auto market. First of all, we can answer if the COVID induced supply shortages drastically impacted U.S. vehicle production numbers or if, as suggested by FRED COVID simply highlighted a long-term trend that would have occurred regardless. Using this data will then also allow us to make forecasts on what the future outlook may be for domestic auto production.

The Bank Prime loan rate is set in relation to the federal funds rate set by the federal reserve and we have data going back to January 1949. The rate is often set based on other determinants of the market so we may not be able to generate too many insights from it individually. However, if we use the bank prime rate in addition to the other factors we discussed earlier, we may be able to create a regression model that will allow us to understand what economic factors influence the car purchasing decisions of Americans. Additionally, studying all the data together will allow us to understand if COVID-19 disruptions altered the purchasing decisions of Americans or if there were vehicle market trends that would have likely occurred regardless of the pandemic.

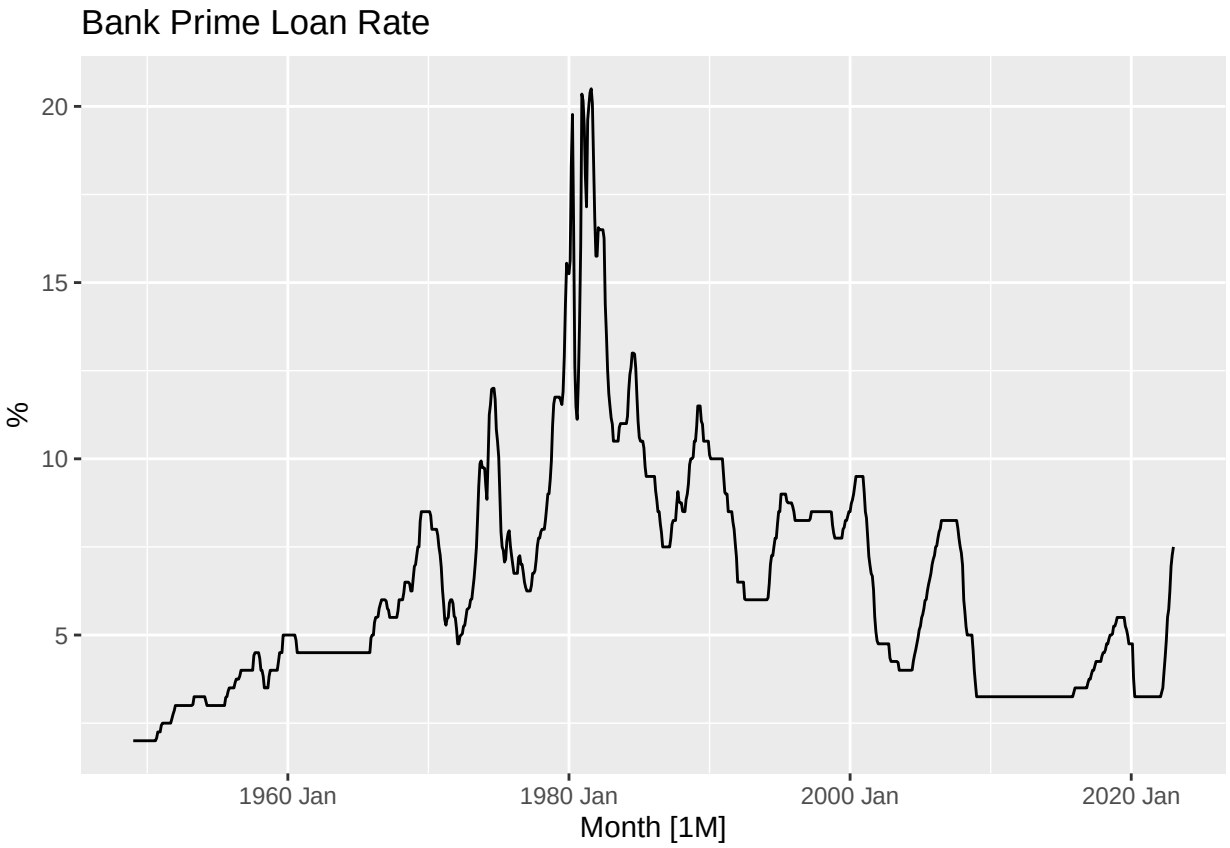


Figure 5: Bank Prime Rate Loan.

Models for Data Fitting

Throughout this case study, we will aim to fit time series regression models that allow us to understand the relationships between the total number of vehicles sold and different economic variables. The selected will be Autoregressive Distributed Lag (ARDL) models, which use a combination of self lagged values, distributed lags, seasonality components, and trend components. Self lagged values means the model takes into account previous values of itself to make future predictions. Distributed lags means that the regression model takes into account previous values of an independent variable to make predictions. Finally, seasonality and trend components are extracted from the data itself. In order to construct an ARDL model, we must fiest examine the data and determine if the data is stationary either at I(0) or I(1). If the data is stationary, we can then begin to fit our models using criteria such as the Akaike Information Criterion (AIC) to select the best model. Finally, we will have to use diagnostic checking where we will examine the residuals from the fitted model to see if they appear sufficient. If we initially appear to fit models that do not fit the data properly, we should then go back and once again attempt to find an appropriate model.

The first step to fitting our model is to see if any of our time series need to be transformed or adjusted, which often allows us to analyze simpler time series. Adjusting the data will allow us to make patterns more consistent across our data, which will allow us to better model the data, and will lead to better forecasts. For the metrics I used, the only ones that may need transformations due to variations in seasonality are those for domestic auto production, total vehicles sold, and the gasoline price index. Since these time series show variations that change with the level of the series, in order to allow the data to remain interpretable, we will apply logarithmic transformations.

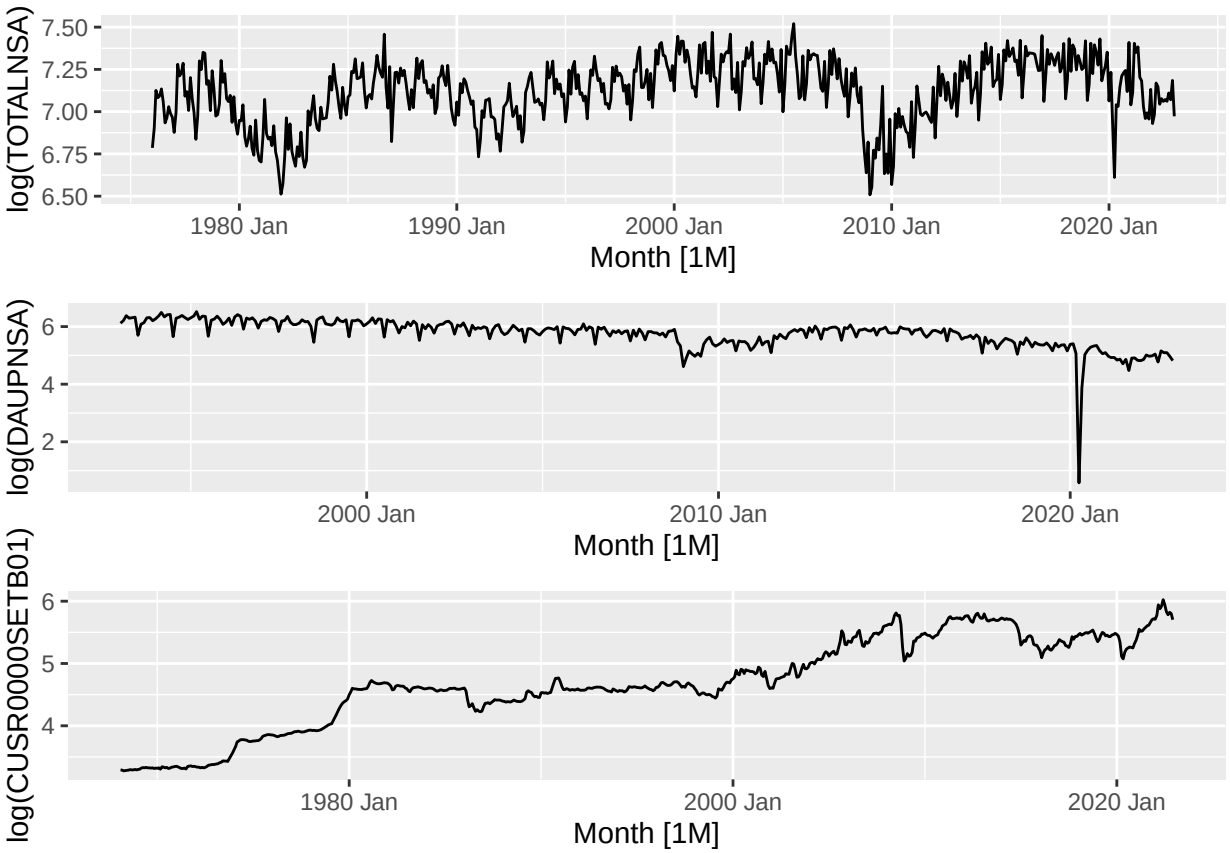


Figure 6: Log adjusted plot for Total New Vehicle Sales (Top), Domestic Auto Production (Middle), and Gasoline CPI (Bottom).

The next step to fitting our model is to decompose our data into the time series' separate components so that we can perform further analysis. Depending on the type of data we have, there are two possible ways

that we can decompose the data. The first possible case of data is a time series with a trend but no seasonal variations, which follows the form $X_t = \alpha + \beta t + \epsilon_t$, where α, β are constants and ϵ_t is a random error term with mean zero. Since our CPI data for gasoline prices and new car prices are both seasonally adjusted, this is the type of decomposition they will undergo.

The second possible case of data is a time series that contains both a trend, and seasonal variation. In this case, there are two possible cases of decomposition we can take. If we assume additive decomposition, then we can use $y_t = S_t + T_t + R_t$, where S_t is the seasonal component, T_t is the trend component, and R_t is the remainder component. The other possible case involves multiplicative decomposition which follows the form $Y_t = S_t \times T_t \times R_t$. We use the multiplicative case if the variation around the trend-cycle appears to be proportional to the level of the time series ("PSTAT 174/274: Time Series Part IV" 2023). The additive case would be if the magnitude of the seasonal fluctuations don't significantly vary over time. Since we used a box-cox transformation to stabilize our variance, we can use the additive case in our decomposition.

In our case, we will use an X-11 decomposition, which is commonly used by the U.S. Census Bureau. X-11 decomposition works by estimating the trend by a moving average, removing the trend to leave the seasonal and irregular components, and estimating the seasonal component using the moving averages to smooth out irregularities. Due to the fact that seasonality cannot be identified without the trend and a trend cannot be estimated without data beign seasonally adjusted, X-11 models use an iterative approach to come up with the best solution. An example of X-11 decomposition is shown below

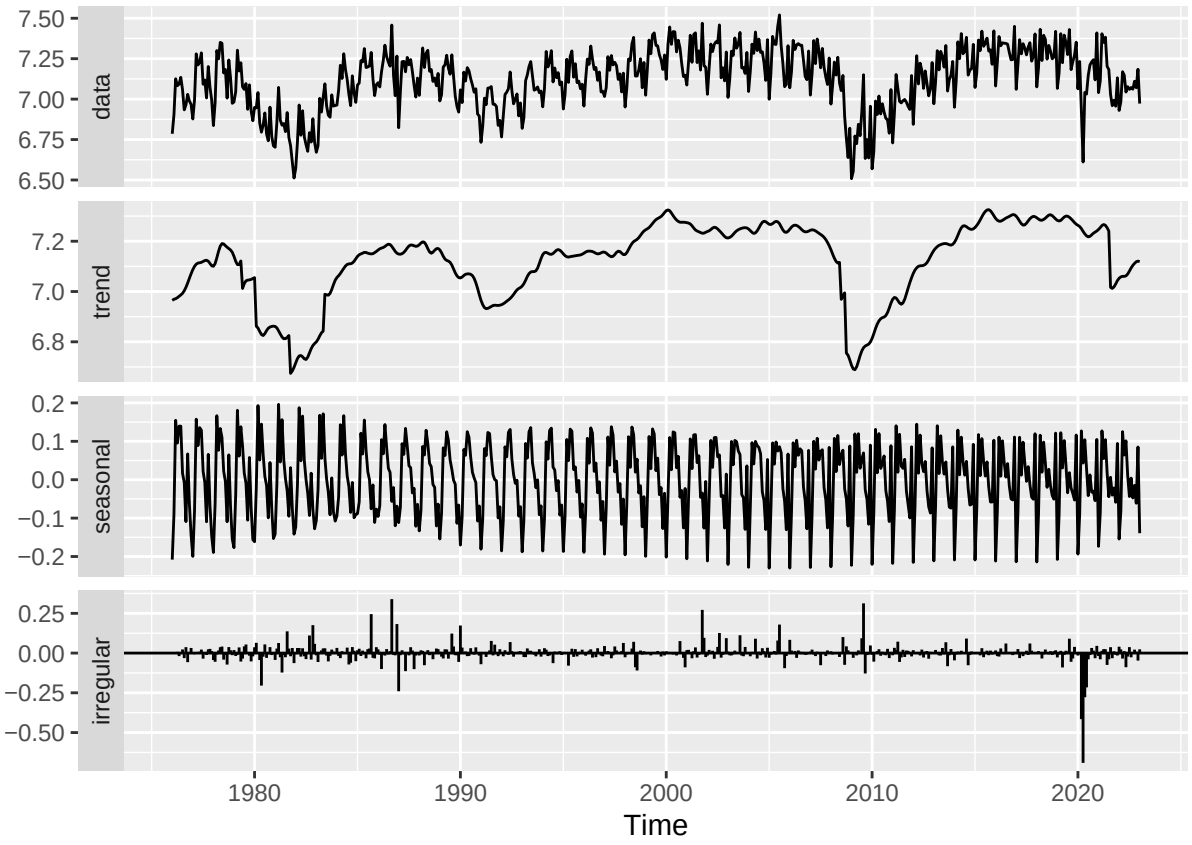


Figure 7: X-11 decomposition of Total New Vehicle Sales

From the decomposition of this time series, along with the other time series, we can get a general idea of the trend and seasonality present within the plot, which will then allow us to better estimate and thus forecast the data moving forwards. Additionally, following the decomposition, we can then see how the trend and seasonally adjusted versions of our data fit with the original data.

For this case study, our goal is to eventually forecast future values of each of our time series data through the

use of ARDL models. The goal of ARDL models is to describe autocorrelations in the data for forecasting (Rob J Hyndman 2021). When attempting to fit ARDL models, we attempt to first remove any trend or seasonality present within the data so that we can create stationary time series, which will allow us to attempt to model the remaining residuals. Stationary time series are time series whose mean and variance is constant over time (Glenn, n.d.).

One way to convert non-stationary time series into stationary time series is the method of differencing, where we compute the differenece between consecutive observations in a time series and can be written as follows

$$y'_t = y_t - y_{t-1}$$

Differencing time series allows us to stabilize our means because it removes or reduces trend and seasonality. It is important to note that for ARDL models, it is important to ensure that all data is stationary at either the I(0) or I(1) level, meaning that either the original data must be stationary or the first difference of the data is stationary. Additionally, for the dependent variable, we cannot interpret long run relationships of the model unless it is I(1).

Before we start differencing our data, we should first test if our data is already stationary, which would then allow us to go straight into further analysis. The test we will use for stationarity is the Kwiatkowski-Phillip-Schmidt_shin (KPSS) test. The KPSS test works by breaking up a time series down into the following decomposition

$$Y_t = r_t + \beta_t + \epsilon_t$$

where r_t is a random walk, β_t is the trend, and ϵ_t is the stationary error. The KPSS test also involves hypothesis testing with

$$H_0 : Y_t \text{ is trend (or level) stationary}$$

$$H_1 : Y_t \text{ is a unit root process}$$

where H_0 and H_1 is the null and alternative hypothesis respectively. Setting our significance level at e can now test each time series to determine whether they will require differencing.

To double check the stationarity of our data, we will also use the Augmented Dickey-Fuller (ADF) test. ADF tests consider the following model

$$Y_t = \mu + \beta t + \alpha Y_{t-1} + \sum_{j=1}^{p} \phi_j \Delta Y_{t-j} + \epsilon_t$$

and they are unit root tests that test the null hypothesis that $\alpha = 1$ with $p = 0$ where α is the coefficient of the first lag on Y and Y_{t-j} is the first lag, with ΔY_{t-j} first difference in series at time $t - j$ ADF tests have a null and alternative hypothesis as follows

$$H_0 : \text{Series is non-stationary or series has a unit root}$$

$$H_A : \text{Series is stationary or series has no unit root.}$$

If out test statistic results in a p-value < 0.05, we will reject the null hypothesis that our time series does not have a unit root, which would mean that it would be stationary and does not hage a time-dependent structure.

The value of the t-statistic for each KPSS test is as follows:

KPSS Test Statistic and Significant Values of Time Series

Time Series	Test-Statistic	P-Values
Total Vehicles	1.3497	0.01
Domestic Production	3.9468	0.01
Bank Prime Rate	2.1478	0.01
Uemployment	1.1680	0.01
Gasoline CPI	8.2357	0.01

ADF Test Statistic and Significant Values of Time Series

Time Series	Test-Statistic	P-Values
Total Vehicles	-2.7392	0.27
Domestic Production	-3.8473	0.02
Bank Prime Rate	-2.8386	0.22
Uemployment	-3.8808	0.01
Gasoline CPI	-2.5347	0.35

From the two tables, we see that we have slightly conflicting information as to which time series are stationary. From the KPSS tests, we see that totalall our variables have p-values less than 0.05, meaning that we will have to difference the data to attempt to make our data stationary. From our ADF test, we see that with the exception of our domestic auto production and unemployment time series, all our other time series are non-staionary. To be safe, data identified as non-stationary by either tests will be differenced and then we will once again conduct the KPSS and ADF tests on them to check for stationarity.

In the case of domestic auto production the following procedure will first apply a differencing function and then plot the function to see if the differencing allowed the data to become stationary.

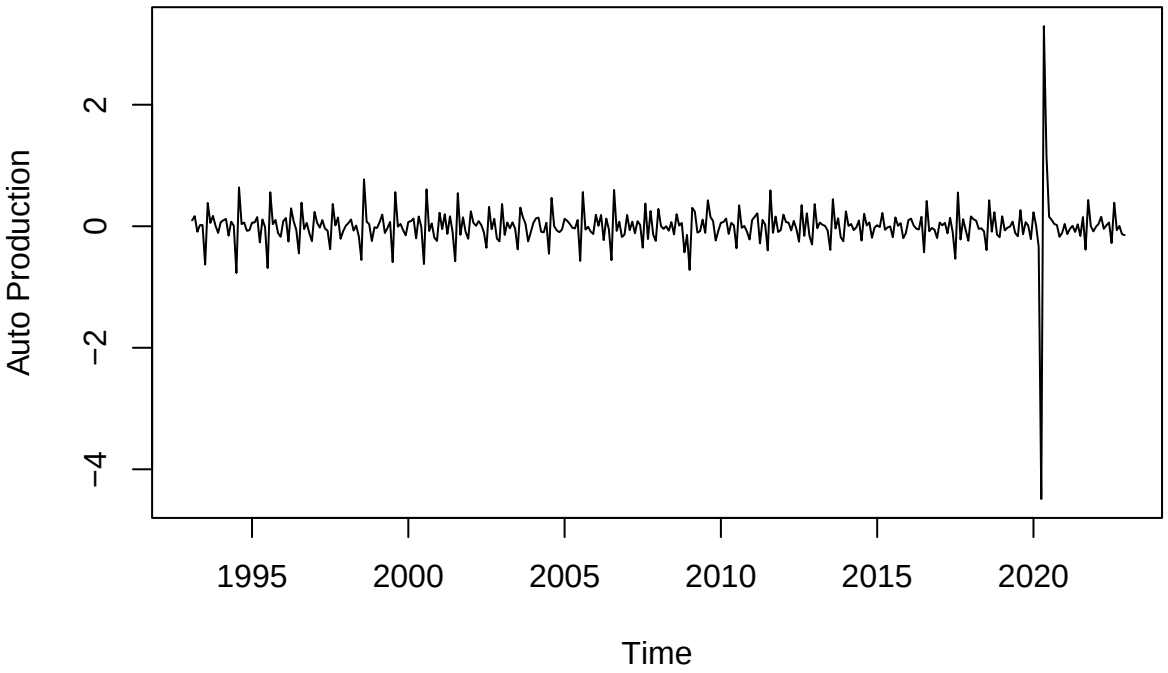


Figure 8: Domestic Auto Production log transformed and differenced.

From the above plot, we see that with the exception of the spikes brought about from COVID-19 production disruptions there does not appear to be any defined pattern that this data follows, so we see that adding one difference allows the time series to become stationary.

KPSS Test Statistic and Significant Values of Differenced Time Series

Time Series	Test-Statistic	P-Values
Total Vehicles	0.0297	0.1
Domestic Production	0.0181	0.1
Bank Prime Rate	0.0715	0.1

Uemployment	0.0527	0.1
Gasoline CPI	0.0614	0.1

ADF Test Statistic and Significant Values of Differenced Time Series

Time Series	Test-Statistic	P-Values
Total Vehicles	-13.5990	0.01
Domestic Production	-10.3160	0.01
Bank Prime Rate	-9.1312	0.01
Uemployment	-9.6830	0.01
Gasoline CPI	-8.9804	0.01

From the tables above, we see that all of our time series are now appropriately differenced and have been made stationary. This will then allow us to complete further analysis of our data.

We can now consider the type of model that we will use for our analysis, the ARDL model. ARDL models are comprised of two main parts the AR, or autoregressive portion of the model and the DL, or distributed lag portion of the model.

We will first discuss autoregressive models. In autoregression, our goal is to forecast the variable of interest using linear combinations of previous values of the variable. This means that in an autoregression, the variable is regressed on itself. An model that utilizes autoregression of order p can be denoted by

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

, where ϵ_t denotes white noise. In such a case, we would say that our model is an AR(p) model. For these types of models, we face certain constraints, mainly:

- For an AR(1) model: $-1 < \phi_1 < 1$
- For an AR(2) model: $-1 < \phi_2 < 1, \phi_1 + \phi_2 < 1, \phi_2 - \phi_1 < 1$.

While autoregression models use past values of itself, distributed lag models use the current and past values of a variable of interest using linear combinations of the previous values of the variable. This means that the variable of interest is regressed on current and previous values of the independent variable. A distributed lag model of order r DL(r) can be written as

$$Y_t = \alpha + \sum_{i=1}^{r} X_{t-i} + \epsilon)t$$

The full ARDL(P,r) model can be expressed as follows:

$$Y_t = \alpha + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \beta_0 X_t + \cdots + \beta_r X_{t-r} + \epsilon_t$$

Now that we have our data properly differenced, we now need to come up with our regression models. Our goal will be to regress each independent varaiable individually on the total number of cars sold to try to understand what relationship, if any is present between our economic factors and the resulting number of vehicles sold.

The first thing we want to consider with our model is that the number of total vehicles sold as shown in our data analysis demonstrated seasonality so in order to account for that in our model, we will want to use dummy coding. After taking seasonality into account, we can use an ANVOA test to determine if the dummy variables results in a significantly different model.

Analyis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
563	10.685423	NA	NA	NA	NA
552	5.183181	11	5.502242	53.27085	1.888952e-79

The small p-value indicates that the seasonal terms play a significant role in the model so we should make sure to take that into account for our models going forwards

Total Vehicles Sold and Domestic Auto Production:

We can now consider the time series model when regressed against each other. The first model we will look at is the ARDL model including total vehicles sold and domestic auto production. When considering ARDL models we must first attempt to calculate the number of lags which are significant for each variable. To do this, I will elect to use Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) and the `auto_ardl` function which uses AIC and BIC for both variables simultaneously to search for the best possible model. Information criterion are of the form $nlog(\frac{RSS}{n}) + kp, p = \#$ of parameters With:

- AIC, k = 2
$$nlog(\frac{RSS}{n}) + 2p$$
- BIC, k = 2
$$nlog(\frac{RSS}{n}) + log(n)p$$

BIC tends to penalize larger models more heavily so if the lags required for AIC models are quite high, we may choose to use the lags chosen through BIC to prevent over fitting.

After Selecting the model that will best fit our data, we can then perform residual analysis to ensure that our data results in the remaining residuals having no autocorrelation present. This can be done both with an inspection of residual plots and with the implmentation of a Breusch-Godfrey test. Ideally, we want to see that there are no patterns present in the residuals vs fitted plot, indicating that our reisudals are evenly distributed that the scale-location plot has a flat line, indicating constant variance, and that the points on the QQ-plot lie close to the line. The Residuals vs. Leverage plot help us identify influential points on the model.

The Breusch-Godfrey test is a statistical test that allow us to determine if autocorrelations are present between the residuals of our time series regression models. The test is conducted first by creating the regression then considering the sample residuals as follows

$$u_t = \rho u_{t-1} + \cdots + \rho_m u_{t-m} + \epsilon_t$$

and where the null hypothesis is

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_m = 0.$$

If the sample size is large enough, then $nR^2 \sim \chi^2(p)$. If we do not reject the null hypothesis, that means that there is no serial correlation of order up to p (Zaiontz 2023).

As a final measure of the effectiveness of our model, we will also apply the Granger causality test to see if including lags of our regressor are actually informative in terms of predicting Y (Christoph Hanck 2021). The null and alternative hypothesis of the Granger Causality Test is as follows

$$H_0 = \text{ Time series X does not cause time series Y to Granger-cause itself.}$$

$$H_A = \text{ Time series X causes time series Y to Granger-cause itself.}$$

It is important to note that for a time series can Granger-cause another variable if it is helpful for forecasting the other variable (Eric 2021).

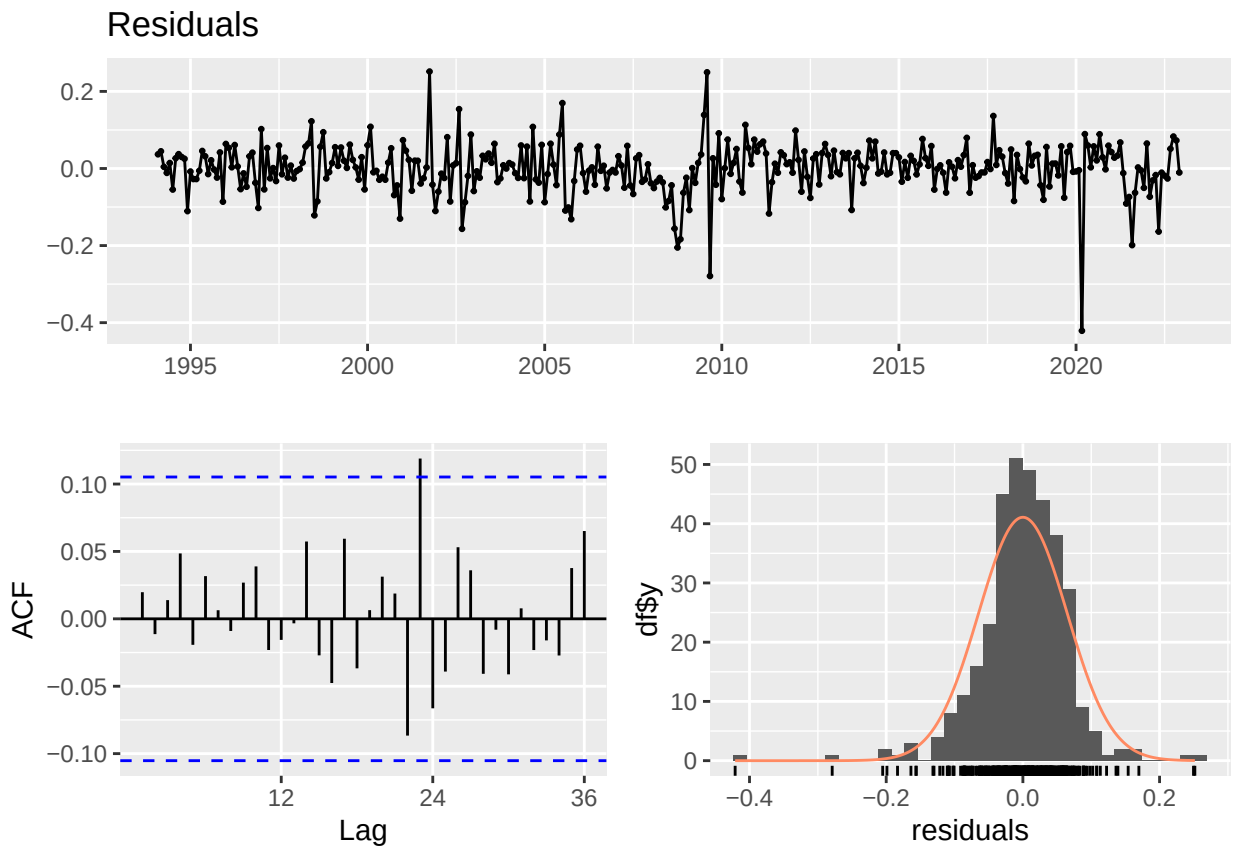


Figure 9: Analysis of ARDL (12,10) on Total New Vehicles Sold with Domestic Auto Production as Exogenuous Variable.

```
##
##  Breusch-Godfrey test for serial correlation of order up to 38
##
## data:  Residuals
## LM test = 50.267, df = 38, p-value = 0.0879
```

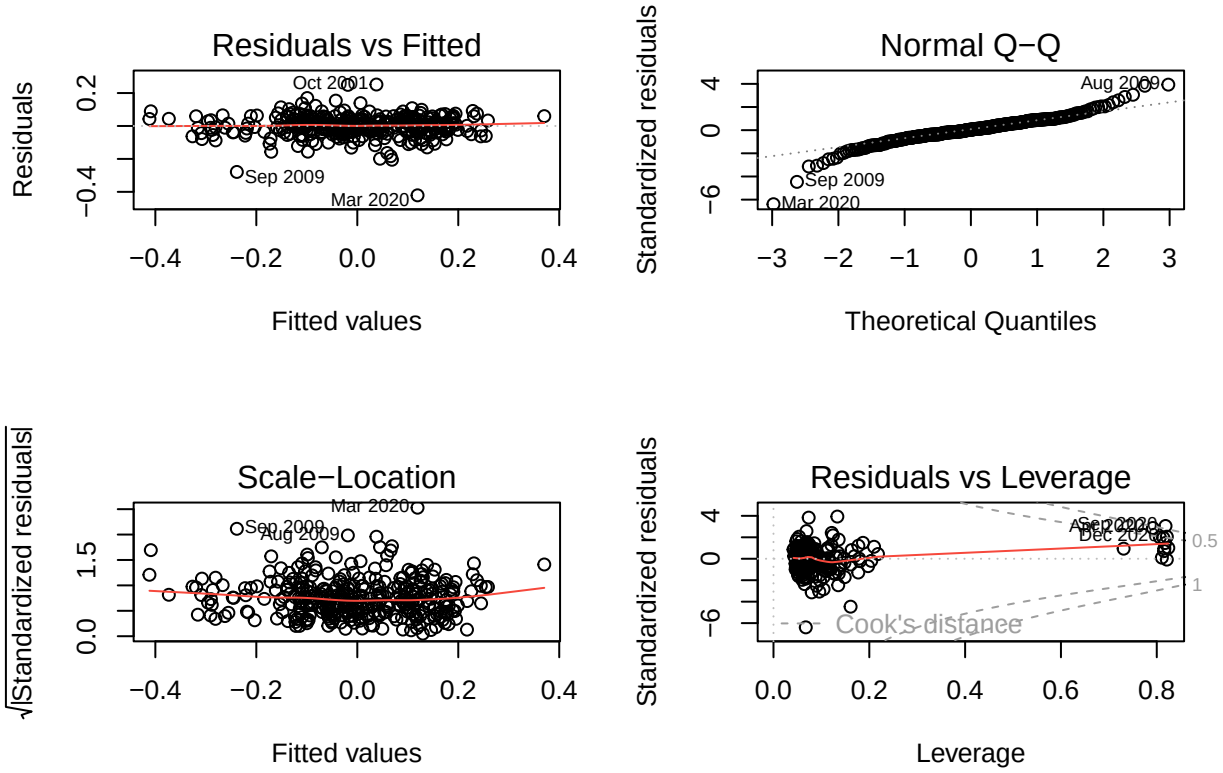


Figure 10: Analysis of ARDL (12,10) on Total New Vehicles Sold with Domestic Auto Production as Exogenuous Variable.

From the outputs, above showing our residual analysis, we can see that the residual values we obtain are consistent with being white noise and that all our values appear normally distributed. Additionally, the p-value we got from the Breusch-Godfrey test was $0.0879 > 0.05$, therefore we can conclude that it is likely that our residuals are not serially correlated. We can then conclude that our current model is appropriate with modeling the true data. We can now test to see if adding the lags for domestic auto production were helpful in predicting the number of vehicles sold.

From the Granger-Causality test conducted above, we get a p-value < 0.05, therefore we can reject the null hypothesis and conclude that it is likely that domestic auto production Granger-causes the number of vehicles sold. For the following time series, the same process will be followed with the output being listed in a table.

Total Vehicles Sold and Unemployment Rate:

Total Vehicle Sales and Unemployment Model Analysis (Breusch-Godfrey Test)

Model	Selection Criteria	P-Values	Model Adequacy	Granger Causality P-Value
ARDL(12,1)	Individually Selected BIC	0.26170	Adequate	< 0.01
ARDL(12, 14)	Auto_ARDL BIC	0.04512	Inadequate	
ARDL(15,14)	Auto_ARDL AIC	0.00733	Inadequate	

From the 3 models that we tried, the best one was the model that was the ARDL(12,1) model, meaning we used 12 lags of total auto sales and 1 lag of the unemployment rate. For the Granger Causality test we also obtained a p-value that was smaller than 0.05 which means we can conclude that it is likely that Unemployment Granger-causes total auto sales.

```
## Analysis of Variance Table
##
## Model 1: diff_total_log ~ L(diff_total_log, 1:12) + L(unemp, 0:1) + season(diff_total_log)
## Model 2: diff_total_log ~ L(diff_total_log, 1:12) + season(total_log)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    526 3.1766
## 2    528 3.5042 -2  -0.32752 27.116 6.195e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Total Vehicles Sold and Gas CPI:

Total Vehicle Sales and Fuel CPI Model Analysis (Breusch-Godfrey Test)

Model	Selection Criteria	P-Values	Model Adequacy	Granger Causality P-Value
ARDL(12,2)	Individually Selected BIC	0.03310	Inadequate	
ARDL(12, 12)	Auto_ARDL BIC	0.12620	Adequate	< 0.01
ARDL(15,12)	Auto_ARDL AIC	0.03477	Inadequate	

From the 3 models that we tried, the best one was the model that was selected by `auto_ardl` using BIC as the selection criterion. The best model is ARDL(12,12) model, meaning we used 12 lags of total auto sales and 12 lag of the gasoline CPI. For the Granger Causality test we also obtained a p-value that was smaller than 0.05 which means we can conclude that it is likely that gasoline prices Granger-causes total auto sales.

Total Vehicles Sold and Bank Prime Rate Loan:

Total Vehicle Sales and Bank Prime Rates Model Analysis (Breusch-Godfrey Test)

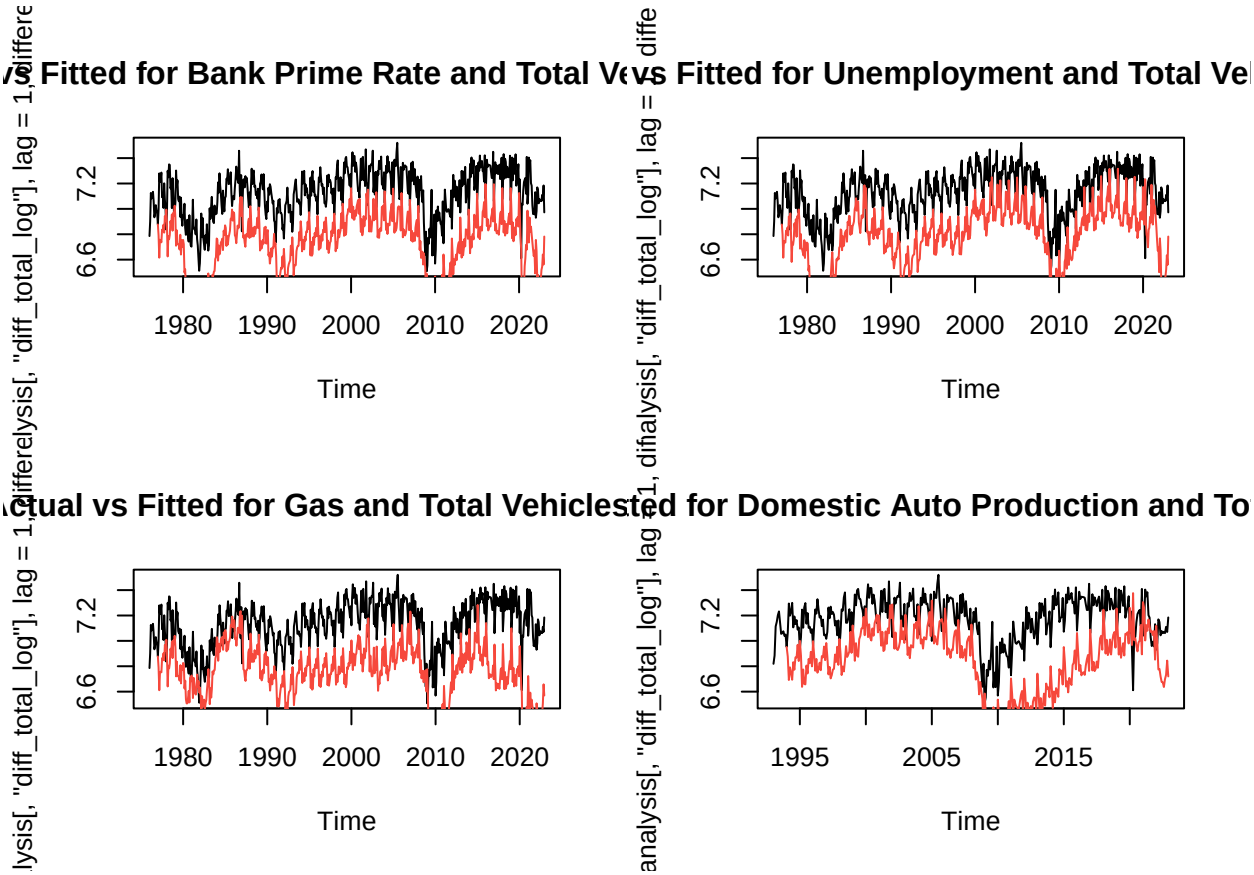
Model	Selection Criteria	P-Values	Model Adequacy	Granger Causality P-Value
ARDL(12,13)	Individually Selected BIC	0.01152	Indequate	
ARDL(12, 0)	Auto_ARDL BIC	0.05509	Adequate	0.4755
ARDL(14,13)	Auto_ARDL AIC	0.00080	Inadequate	

From the 3 models that we tried, the best one was the model that was selected by `auto_ardl` using BIC as the selection criterion. The best model is ARDL(12,0) model, meaning we used 12 lags of total auto sales and the current value of the gasoline CPI. For the Granger Causality test we obtained a p-value that $0.4755 > 0.05$ so we cannot reject the null hypothesis, which means that it is unlikely that bank prime rate loans Granger-cause total auto sales.

```
## integer(0)
```

```
## integer(0)
```

```
## integer(0)
```




```
## integer(0)
```

Forecast Analysis

Now that we have appropriate models, we can analyze the forecasts provided by the models. In order to do this, we will utilize thge method of pseudo out-of-sample forecasting which follows the following steps (Christoph Hanck 2021): 1.) Divide the sample data into $s = T - P$ and P subsequent observations where the P observations are used as pseudo-out-of-sample observations

2.) Estimate the model using the first s observations.

3.) Compute the pseudo-forecast $\tilde{Y}_{s+1|s}$.

4.) Compute the pseudo-forecast-error $\tilde{u}_{s+1} = Y_{s+1|s}$.

Doing this will then allow us to determine how our models forecast and which models appear to have the best forecasting performance when fit to actual data. Since it appears that the bank prime rate loan does not appear to have a significant effect on the total number of cars sold, we will analyze the forecasts of ARDL models including the unemployment rate, gasoline CPI, and domestic auto production.

As a measure of the accuracy of our forecasts, we will use the root mean square error (RMSE) of our forecasts compared to the actual data. The root mean square error allows us to calculate the spread of the residuals and in essence allows us to "measure the average difference between values predicted by a model and the actual values" ("Root Mean Squared Error (RMSE)," n.d.). The formula for RMSE is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

Unemployment Forecast

```
## integer(0)
```

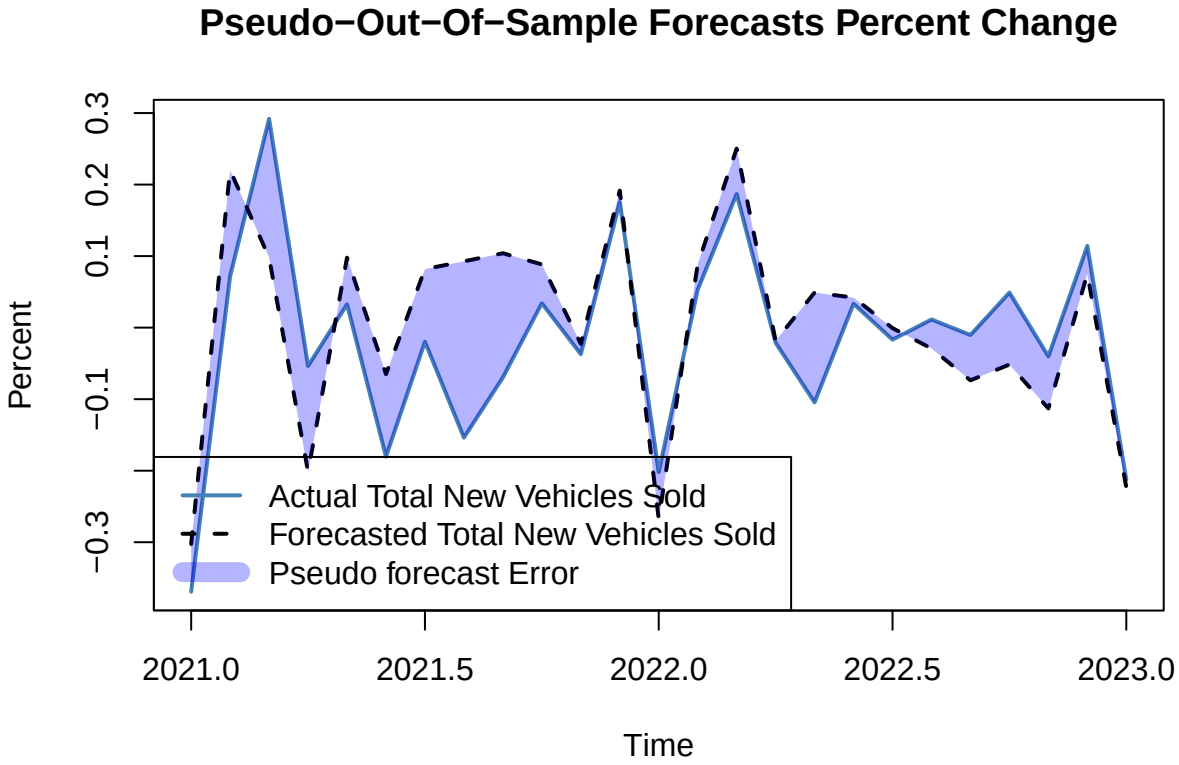


Figure 11: Total New Vehicle Sales, Unemployment ARDL pseudo-out-of-sample residual forecast analysis

Gas Forecast

```
## integer(0)
```

Domestic Auto Production Forecast

```
## integer(0)
```

Forecast RMSE

Model Exogenous Variable	RMSE
Unemployment	0.1028
Gasoline CPI	0.1645
Domestic Auto Production	0.1613

Future Forecasts

Discussion

This case study focused on attempting to model total new vehicle sales through the use of Autoregressive Distributed Lag (ARDL) models. We attempted to fit models that included the unemployment rate, the gasoline CPI, domestic auto production, and the bank prime rate loans.

After fitting our models, we found the different models that would fit best for our different ARDL models.

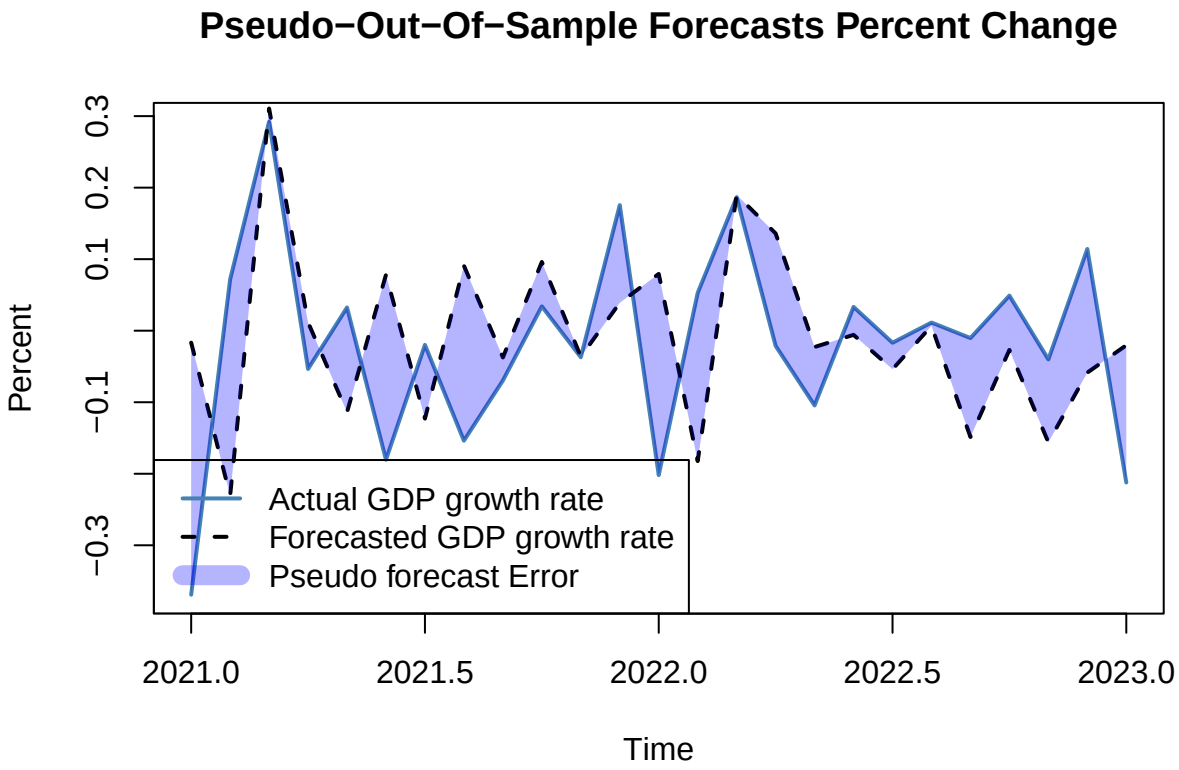


Figure 12: Total New Vehicle Sales, Gasoline CPI ARDL pseudo-out-of-sample residual forecast analysis

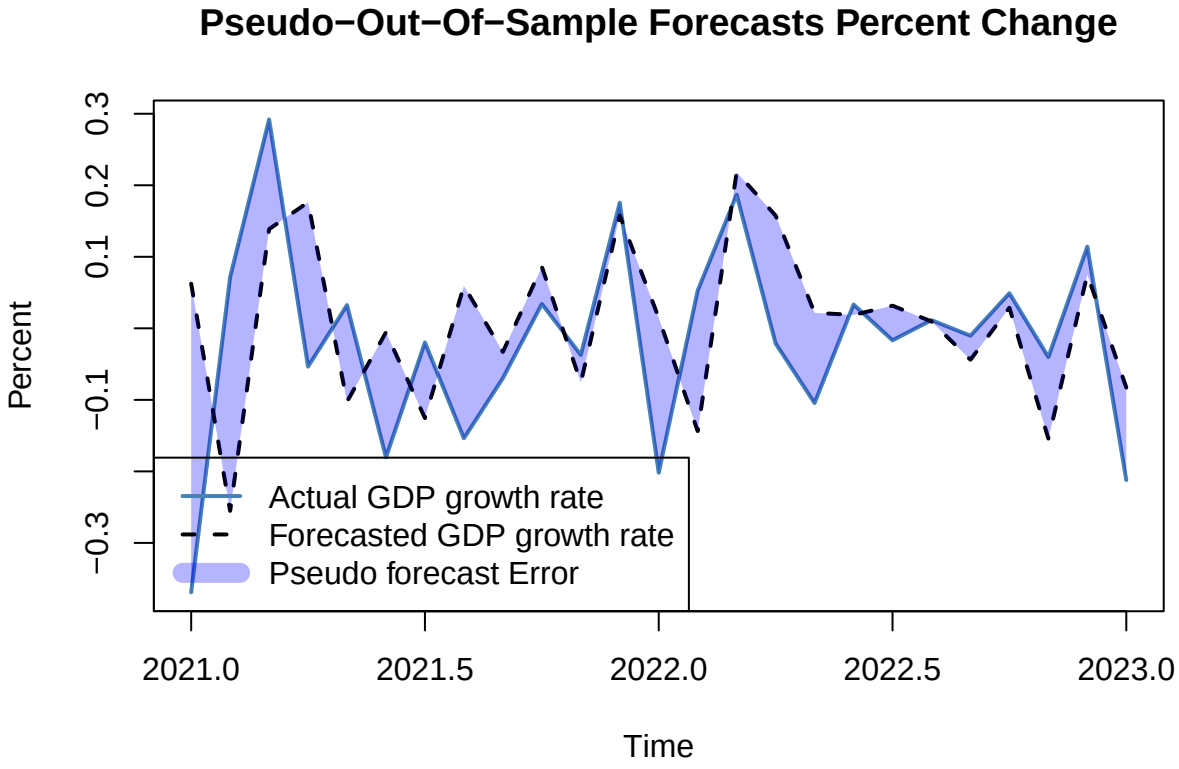


Figure 13: Total New Vehicle Sales, Unemployment ARDL pseudo-out-of-sample residual forecast analysis

We then attempted to find which ARDL models were most important in determining the total number of new vehicles that were sold. From the variables that we used, we then calculated if the exogenous variables that we added as part of the ARDL model allowed us to better estimate the true number of vehicles sold through the use of Granger-causality tests.

It was determined that the bank prime rate loans had no statistical significance in allowing us to better predict the total number of vehicles sold. This could potentially make sense because if we look at how bank prime rate loans have changed over time, we see that they reached a peak in the 80s before slowly decreasing over time; however, if we look at time series of total new vehicles we see that there have been cyclical components but that in general the number of new vehicles sold has remained at a relatively stable level when compared to some of the other time series.

It was therefore determined that the ARDL models where the exogenous variables did make a difference in allowing us to predict total number of new vehicles sold were the unemployment rate, gasoline CPI and domestic auto production. The actual regression model used for the models that were significant is located in the appendix. For unemployment we see that there is a negative relationship, meaning that as unemployment increases, we would expect to predict that the total number of new vehicles sold decreases, which would make sense since when unemployment increases, the population as a whole would likely have less purchasing power, leading to fewer cars being bought.

Interestingly, for gasoline CPI we have 6 lags with positive values, which would seem to indicate that about half the time, an increase in gas price leads to an increase in new vehicles sold. This is not straightforward but may make sense if we consider that in the last few years, even as gas prices have gone up, the purchasing of new car has remained high, which may indicate that cars may be an inelastic good, meaning that despite the ricing prices of gasoline, a complement, people still have a need for purchasing vehicles. In addition, we have recently seen an increase in the sale of hybrid and electric vehicles with 5.5% of light vehicle sales being hybrid and 3.2% of light vehicle sales being electric in 2021 ("Hyrbid-Electric, Plug-in Hybrid-Electric and Electric Vehicle Sales" 2022). Furthermore, gas mileage has improved now even for trucks and SUVs which could make the American consumer less likely to respond to change in gas prices. For further conclusions, this is an area of research that may be interesting to look at.

For domestic auto production the data is as expected, with all values being positive, which would indicate that as more cars are produced domestically, more total new cars would be sold. This relationship could go both ways in that more cars being produced could lead to more total new vehicles being sold and that consumers buying more cars may encourage auto makers to produce more cars.

After analyzing the data, the ARDL models which appeared to have effects on the total new vehicles sold were analyzed through pseudo-out-of-sample forecasts where we found that it appeared that the model that appeared to have the best forecasting performance was the model that used unemployment as its exogenous variable. This intuitively makes sense since it appears logical that as more people became unemployed, there would also be less people purchasing new cars. It is worth mentioning, however, that our other models also appeared to have relatively good forecasting ability.

Conclusion

The automobile is in many ways synonymous with the American lifestyle and this case study aimed to better understand some of the economic forces behind it. We looked at the the unemployment rate, gasoline CPI, domestic auto production, and bank prime rate loans as economic factors that could potentially affect the number of total new vehicles sold. We ultimately found that bank prime rate loans were largely uncorrelated with the number of new vehicles sold but that our other factors demonstrated correlation with total new vehicles sold.

In the process of finding these correlations, we were able to make use if ARDL models to model the relationship between the our exogenous variables and total new vehicles sold, and these relationships can be seen in the appendix. We then utilized pseudo-out-of-sample forecasts to evaluate our models as a way to confirm their efficacy and was able to find that unemployment appeared to be the best of our selected exogenous variables when used in an ARDL model.

Going forwards, this case study provides us with many avenues through which we could continue to better understand the American automobile market. One of the most apparent ways we could further our analysis would be by looking at other economic variables that may affect total number of new vehicles sold such as the amount of disposable income available throughout the country, the general CPI, or by looking at the price of new cars. As mentioned earlier in the discussion, having the same amount of lags of gasoline CPI indicating a positive and negative relationship between gasoline CPI and new vehicles sold may indicate the inelasticity of total vehicle purchases and analyzing the price of vehicles in conjunction with total vehicles sold would allow us to truly test this hypothesis.

Furthermore, since vehicles are often the second most expensive item a family may purchase, we could try to analyze the relationship between vehicles sold and house purchases and see if increases in vehicle sales may preempt a strong housing market.

Ultimately, this case study allowed us to better understand the forces driving Americans to purchase vehicles and its relation to other economic forces.

Apendix

```
##
## Time series regression with "ts" data:
## Start = 1977(2), End = 2023(1)
##
## Call:
## dynlm::dynlm(formula = diff_total_log ~ L(diff_total_log, 1:12) +
##     L(unemp, 0:1) + season(diff_total_log), data = total_unemp_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44799 -0.04068  0.00286  0.03934  0.27965
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.104330   0.016548  -6.305 6.11e-10 ***
## L(diff_total_log, 1:12)1   -0.584703   0.044351 -13.184  < 2e-16 ***
## L(diff_total_log, 1:12)2   -0.402037   0.048614  -8.270 1.10e-15 ***
## L(diff_total_log, 1:12)3   -0.253643   0.047466  -5.344 1.36e-07 ***
## L(diff_total_log, 1:12)4   -0.174381   0.047721  -3.654 0.000284 ***
## L(diff_total_log, 1:12)5   -0.121162   0.047713  -2.539 0.011391 *
## L(diff_total_log, 1:12)6   -0.148613   0.047810  -3.108 0.001983 **
## L(diff_total_log, 1:12)7   -0.109781   0.047991  -2.288 0.022561 *
## L(diff_total_log, 1:12)8   -0.071640   0.047851  -1.497 0.134952
## L(diff_total_log, 1:12)9    0.003180   0.047455   0.067 0.946602
## L(diff_total_log, 1:12)10  -0.133131   0.046642  -2.854 0.004483 **
## L(diff_total_log, 1:12)11  -0.097297   0.044960  -2.164 0.030906 *
## L(diff_total_log, 1:12)12   0.193564   0.040888   4.734 2.83e-06 ***
## L(unemp, 0:1)0             -0.050929   0.007142  -7.131 3.31e-12 ***
## L(unemp, 0:1)1             -0.016568   0.007441  -2.226 0.026404 *
## season(diff_total_log)Feb   0.097045   0.023949   4.052 5.84e-05 ***
## season(diff_total_log)Mar   0.245870   0.022047  11.152  < 2e-16 ***
## season(diff_total_log)Apr   0.160288   0.023085   6.943 1.13e-11 ***
## season(diff_total_log)May   0.181524   0.025774   7.043 5.92e-12 ***
## season(diff_total_log)Jun   0.143522   0.024497   5.859 8.23e-09 ***
## season(diff_total_log)Jul   0.073546   0.025586   2.874 0.004211 **
## season(diff_total_log)Aug   0.103533   0.024362   4.250 2.53e-05 ***
## season(diff_total_log)Sep   0.041330   0.025142   1.644 0.100803
```

```
## season(diff_total_log)Oct  0.082300   0.022847   3.602 0.000345 ***
## season(diff_total_log)Nov  0.011556   0.022394   0.516 0.606034
## season(diff_total_log)Dec  0.111558   0.023763   4.695 3.41e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07771 on 526 degrees of freedom
## Multiple R-squared:  0.6998, Adjusted R-squared:  0.6855
## F-statistic: 49.04 on 25 and 526 DF,  p-value: < 2.2e-16
##
## Time series regression with "ts" data:
## Start = 1977(2), End = 2022(12)
##
## Call:
## dynlm::dynlm(formula = diff_total_log ~ L(diff_total_log, 1:12) +
##     L(gas_cpi, 0:12) + season(diff_total_log), data = total_gas_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45751 -0.04335  0.00303  0.04190  0.29063
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.10300    0.01745  -5.903 6.49e-09 ***
## L(diff_total_log, 1:12)1  -0.48865    0.04253 -11.491  < 2e-16 ***
## L(diff_total_log, 1:12)2  -0.33948    0.04762  -7.130 3.43e-12 ***
## L(diff_total_log, 1:12)3  -0.22603    0.05004  -4.517 7.79e-06 ***
## L(diff_total_log, 1:12)4  -0.18814    0.05102  -3.687 0.000251 ***
## L(diff_total_log, 1:12)5  -0.14734    0.05148  -2.862 0.004379 **
## L(diff_total_log, 1:12)6  -0.15750    0.05106  -3.084 0.002149 **
## L(diff_total_log, 1:12)7  -0.08640    0.05105  -1.692 0.091190 .
## L(diff_total_log, 1:12)8  -0.03543    0.05070  -0.699 0.484989
## L(diff_total_log, 1:12)9   0.05009    0.05024   0.997 0.319260
## L(diff_total_log, 1:12)10 -0.09419    0.04930  -1.911 0.056615 .
## L(diff_total_log, 1:12)11 -0.05778    0.04746  -1.217 0.223977
## L(diff_total_log, 1:12)12  0.25720    0.04220   6.095 2.14e-09 ***
## L(gas_cpi, 0:12)0          0.18568    0.08411   2.208 0.027707 *
## L(gas_cpi, 0:12)1          0.01585    0.09257   0.171 0.864113
## L(gas_cpi, 0:12)2          0.02468    0.09525   0.259 0.795649
## L(gas_cpi, 0:12)3         -0.09144    0.09553  -0.957 0.338973
## L(gas_cpi, 0:12)4          0.10433    0.09566   1.091 0.275938
## L(gas_cpi, 0:12)5         -0.23222    0.09571  -2.426 0.015604 *
## L(gas_cpi, 0:12)6          0.12462    0.09621   1.295 0.195800
## L(gas_cpi, 0:12)7         -0.20945    0.09649  -2.171 0.030412 *
## L(gas_cpi, 0:12)8         -0.18531    0.09708  -1.909 0.056829 .
## L(gas_cpi, 0:12)9         -0.03821    0.09831  -0.389 0.697670
## L(gas_cpi, 0:12)10         0.12818    0.10048   1.276 0.202631
## L(gas_cpi, 0:12)11        -0.14839    0.09787  -1.516 0.130089
## L(gas_cpi, 0:12)12        -0.15743    0.08852  -1.778 0.075923 .
## season(diff_total_log)Feb  0.10001    0.02477   4.037 6.24e-05 ***
## season(diff_total_log)Mar  0.23293    0.02305  10.106  < 2e-16 ***
## season(diff_total_log)Apr  0.13534    0.02391   5.660 2.52e-08 ***
## season(diff_total_log)May  0.17509    0.02696   6.494 1.97e-10 ***
```

```
## season(diff_total_log)Jun  0.14002    0.02583   5.422 9.09e-08 ***
## season(diff_total_log)Jul  0.08110    0.02692   3.013 0.002718 **
## season(diff_total_log)Aug  0.12380    0.02539   4.875 1.45e-06 ***
## season(diff_total_log)Sep  0.05671    0.02633   2.154 0.031719 *
## season(diff_total_log)Oct  0.09232    0.02390   3.862 0.000127 ***
## season(diff_total_log)Nov  0.01705    0.02340   0.729 0.466608
## season(diff_total_log)Dec  0.10808    0.02468   4.378 1.45e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07947 on 514 degrees of freedom
## Multiple R-squared:  0.6919, Adjusted R-squared:  0.6703
## F-statistic: 32.06 on 36 and 514 DF,  p-value: < 2.2e-16
```

```
##
## Time series regression with "ts" data:
## Start = 1994(2), End = 2022(12)
##
## Call:
## dynlm::dynlm(formula = diff_total_log ~ L(diff_total_log, 1:12) +
##     L(auto_prod, 0:10) + season(diff_total_log), data = total_prod_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42106 -0.02910  0.00278  0.03888  0.25162
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.12735    0.02452  -5.194 3.73e-07 ***
## L(diff_total_log, 1:12)1   -0.52768    0.05807  -9.087  < 2e-16 ***
## L(diff_total_log, 1:12)2   -0.39472    0.06617  -5.966 6.61e-09 ***
## L(diff_total_log, 1:12)3   -0.23016    0.07117  -3.234 0.001352 **
## L(diff_total_log, 1:12)4   -0.20640    0.07160  -2.883 0.004214 **
## L(diff_total_log, 1:12)5   -0.15019    0.07158  -2.098 0.036708 *
## L(diff_total_log, 1:12)6   -0.23782    0.07072  -3.363 0.000868 ***
## L(diff_total_log, 1:12)7   -0.19297    0.07033  -2.744 0.006425 **
## L(diff_total_log, 1:12)8   -0.15777    0.06894  -2.289 0.022765 *
## L(diff_total_log, 1:12)9    0.03437    0.06554   0.524 0.600374
## L(diff_total_log, 1:12)10  -0.19547    0.06150  -3.178 0.001629 **
## L(diff_total_log, 1:12)11  -0.05169    0.05494  -0.941 0.347588
## L(diff_total_log, 1:12)12   0.15183    0.05085   2.986 0.003055 **
## L(auto_prod, 0:10)0         0.11740    0.01415   8.296 3.27e-15 ***
## L(auto_prod, 0:10)1         0.09543    0.01809   5.274 2.50e-07 ***
## L(auto_prod, 0:10)2         0.08424    0.02089   4.034 6.92e-05 ***
## L(auto_prod, 0:10)3         0.07797    0.02294   3.400 0.000763 ***
## L(auto_prod, 0:10)4         0.07407    0.02382   3.109 0.002048 **
## L(auto_prod, 0:10)5         0.05695    0.02429   2.345 0.019657 *
## L(auto_prod, 0:10)6         0.05862    0.02406   2.437 0.015388 *
## L(auto_prod, 0:10)7         0.06805    0.02329   2.921 0.003740 **
## L(auto_prod, 0:10)8         0.02180    0.02130   1.024 0.306835
## L(auto_prod, 0:10)9         0.04293    0.01871   2.294 0.022426 *
## L(auto_prod, 0:10)10        0.06520    0.01533   4.252 2.81e-05 ***
## season(diff_total_log)Feb   0.12583    0.03695   3.405 0.000748 ***
## season(diff_total_log)Mar   0.22793    0.03235   7.046 1.18e-11 ***
```

```
## season(diff_total_log)Apr  0.16840    0.03210   5.246 2.87e-07 ***
## season(diff_total_log)May  0.19412    0.03635   5.341 1.79e-07 ***
## season(diff_total_log)Jun  0.15245    0.03533   4.315 2.15e-05 ***
## season(diff_total_log)Jul  0.13505    0.03701   3.649 0.000308 ***
## season(diff_total_log)Aug  0.13846    0.03625   3.820 0.000161 ***
## season(diff_total_log)Sep  0.05225    0.03652   1.431 0.153435
## season(diff_total_log)Oct  0.13960    0.03369   4.143 4.42e-05 ***
## season(diff_total_log)Nov  0.01954    0.03264   0.599 0.549860
## season(diff_total_log)Dec  0.20756    0.03722   5.577 5.30e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06819 on 312 degrees of freedom
## Multiple R-squared:  0.8091, Adjusted R-squared:  0.7883
## F-statistic: 38.89 on 34 and 312 DF,  p-value: < 2.2e-16
```

References

Christoph Hanck, Alexander Gerber, Martin Arnold. 2021. "Introduction to Econometrics with r." University of Duisberg-Essen. https://www.econometrics-with-r.org/index.html.

Eric. 2021. "Introdution to Granger Causality." https://www.statisticshowto.com/kpss-test/.

Glenn, Stephanie. n.d. "KPSS Test: Definition and Interpretation." https://www.statisticshowto.com/kpss-test/.

"Hyrbid-Electric, Plug-in Hybrid-Electric and Electric Vehicle Sales." 2022. https://www.bts.gov/content/gasoline-hybrid-and-electric-vehicle-sales.

"Inflation and the Auto Industry: When Will Car Prices Drop." 2022. https://www.jpmorgan.com/insights/research/when-will-car-prices-drop.

"Long-Term Trends in Car and Light Truck Sales." 2021. https://fredblog.stlouisfed.org/2022/10/whats-been-driving-the-rise-in-auto-prices-since-covid/?utm_source=series_page&utm_medium=related_content&utm_term=related_resources&utm_campaign=fredblog.

"No End in Sight: New Vehicle Transaction Prices End 2022 at Record Highs, According to New Data from Kelley Blue Book." 2023. https://www.coxautoinc.com/market-insights/kbb-atp-december-2022/.

"PSTAT 174/274: Time Series Part IV." 2023. University of California, Santa Barbara.

Rob J Hyndman, George Athanasopoulos. 2021. "Forecasting: Principles and Practice." Monash University, Australia; OTexts.

"Root Mean Squared Error (RMSE)." n.d. https://help.sap.com/docs/SAP_PREDICTIVE_ANALYTICS/41d1a6d4e7574e32b815f1cc87c00f42/5e5198fd4afe4ae5b48fefe0d3161810.html.

"What's Been Drivin the Rise in Auto Prices Since COVID." 2022. https://fredblog.stlouisfed.org/2021/03/long-term-trends-in-car-and-light-truck-sales/?utm_source=series_page&utm_medium=related_content&utm_term=related_resources&utm_campaign=fredblog.

Zaiontz, Charles. 2023. "Breusch-Godfrey Test." https://real-statistics.com/multiple-regression/autocorrelation/breusch-godfrey-test/.