ISDS 526

Forecasting for Analytical Decision Making

Project 4

**(heading)**

Fall 2018 session

**A close up of a sign

Description automatically generated**

**Professor:** Dawit Zerom

**Submitted By:**

Ruchika Narang

Sanchit Singh

Sebastian Stuchetz

Abhinay Sariswal

Table of Contents

[**Executive Summary** 3](#_Toc532595116)

[**The Forecasting Problem** 4](#_Toc532595117)

[**Dynamic regression model building** 5](#_Toc532595118)

[**Step 1: Analyzing regression of Calls vs Rainfall:** 5](#_Toc532595119)

[**Step 2: Analyzing regression of Calls vs Temperature:** 6](#_Toc532595120)

[**Step 3: Analyzing Rainfall vs Temperature:** 7](#_Toc532595121)

[**Step 5: Capturing Further Seasonality** 9](#_Toc532595122)

[**Step 6: Examining and dealing with autocorrelation of errors** 11](#_Toc532595123)

[**Step 7: HEADING IS MISSING** 13](#_Toc532595124)

[Introducing **Rate** 13](#_Toc532595125)

[**Step 8: The final dynamic regression model** 15](#_Toc532595126)

[**Step 9: Forecast Accuracy** 16](#_Toc532595127)

[**Conclusion and Recommendation** 18](#_Toc532595128)

# **Executive Summary**

American Automobile Association (hereafter referred to as AAA) Washington is among the two regional automobile clubs affiliated with AAA which offers its members a variety of automobile and automobile-related services. Within the member benefits, the most popular service is the emergency road side assistance which is the primary reason for people to join AAA. However, being the top service offered by the company, it turned out to be the club’s single largest operating expense with a projected cost of 9.5 million US dollars, which is 37% of the club’s annual operating budget. We have been hired by AAA as forecasting consultants. The objective of our report is to help Mr. DeCoria, the club’s vice president of operations, collect insights into why service calls have surged at particular times with the help of several factors. Among these factors are average daily temperature, the amount of rainfall received in a day, and unemployment rate for Washington State. By analyzing these factors, the collectable insights will assist Mr. DeCoria’s decision making process and allow him to take suitable plan to prevent the additional expenses of excessive roadside service calls.

# **The Forecasting Problem**

We aim to build a dynamic regression model to determine the variation within the emergency service call volume data. Moreover, as forecasting consultants of AAA we need to facilitate Mr. DeCoria in his decision-making process so he can take necessary actions to neutralize the effect of increased emergency service calls. This is done by analyzing the factors such as temperature, rainfall, and rate. However, these factors alone cannot explain the boost in the service calls and hence, we need to utilize correlation and seasonality to data analysis techniques to conclude the effect on service calls and build the final model. In the end, we need to evaluate the accuracy of our forecast model so business decisions can be made to improve the service and manage costs.

A picture containing screenshot

Description automatically generated

**Figure 1: Graphical framework of decision-making process**

# **Dynamic regression model building**

We will now build multiple dynamic regression models to examine the effect of the factors such as rainfall and temperature on the number of calls. We will first analyze the effect of each factor separately followed by both included in a single regression model. The purpose of the model is to explain the variation within the provided service calls volume data by examining the model statistics. By building this model, AAA may explain some of the historical variation which may lead to more accurate forecasts.

## **Step 1: Analyzing regression of Calls vs Rainfall:**

In this step, we will determine whether there is a relationship between the emergency service calls (hereafter, referred to as calls) and amount of rainfall with the help of regression analysis. From Table 1, the coefficient of rainfall indicates that for every 1 unit increase in rainfall there will be 395 (approx.) increase in calls. Now, we will observe the p-value which supports the randomness in the model if it is less than the significance level (0.05). In this model, from the statistics we obtain a p-value of 0 which is less than the significance level(alpha) of 0.05. This establishes a significant relationship between calls and rainfall. Moreover, the R-square (variability covered by the model) is 38%, which means that the predictive power of this model is weak. Therefore, calls and rainfall are not entirely correlated since only 38% of the model’s variability is explained. However, we can say that the increase in the rainfall increases the number of emergency calls and we can also relate naturally that there are more chances of vehicle breakdown during heavy rain and extreme weather conditions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Coefficient** | **Std. Error** | **t-Statistic** | **p-value** |
| Rainfall | 394.8 | 74.21 | 5.321 | 0.000002824 |
| \_CONST | 19,861 | 316.3 | 62.8 | 0 |

**Table 1: Model Statistics in Detail - Dynamic Regression (2 regressors, 0 lagged errors)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample size | 49 | No. parameters | 2 | |
| Mean | 21,206.43 | Std. deviation | | 1,665.74 |
| R-square | 0.38 | Adj. R-square | | 0.36 |
| Durbin-Watson | 1.51 | Ljung-Box(18) | | 27.9 P=0.94 |
| Forecast error | 1,329.86 | BIC | | 1,410.11 |
| MAPE | 5.08% | SMAPE | | 5.05% |
| RMSE | 1,302.44 | MAD | | 1071.74 |
| MAD/Mean Ratio | 0.05 |  | |  |

**Table 2: Within-Sample Statistics**

## **Step 2: Analyzing regression of Calls vs Temperature:**

Here, with the help of regression analysis, we will determine the kind of relationship that exists between calls and temperature. From Table 3, we can infer from the coefficient of temperature that for every 1 unit increase in temperature there will be 126 (approx.) decrease in calls. In this model, the p-value (0.00) is less than the significance level (0.05) which establishes a significant relationship between calls and temperature. The R-square value in Table 4 indicates that the model covers 51% variability, which means that the predictive power of this model is comparatively good, but still needs improvement. Therefore, the calls and rainfall are not entirely correlated. Only 51% of the variation in the data is explained by this correlation. This indicates that there is still 49% of the variation which is unexplained due to some other factors not covered by the model.

**Table 3: Model Statistics in Detail - Dynamic Regression (2 regressors, 0 lagged errors)**

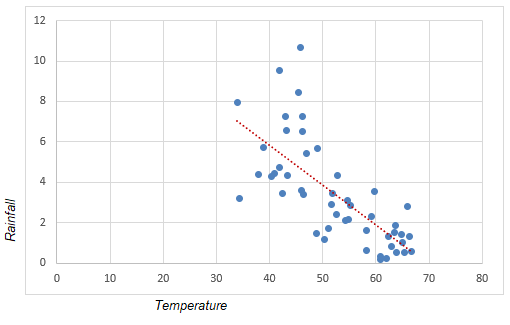
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Coefficient** | **Std. Error** | **t-Statistic** | **p-value** |
| Temperature | -126.2 | 17.86 | -7.062 | 6.572E-09 |
| \_CONST | 27,802 | 948.7 | 29.3 | 0 |

**Table 4: Within-Sample Statistics**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample size** | **49** | No. parameters | 2 |
| Mean | 21,206.43 | Std. deviation | 1,665.74 |
| R-square | **0.51** | Adj. R-square | 0.5 |
| Durbin-Watson | 1.08 | Ljung-Box (18) | **30.3 P=0.97** |
| Forecast error | 1,172.50 | BIC | 1,243.25 |
| MAPE | **4.59%** | SMAPE | 4.55% |
| RMSE | 1,148.32 | MAD | 953.33 |
| MAD/Mean Ratio | 0.04 |  |  |

## **Step 3: Analyzing Rainfall vs Temperature:**

We will now analyze rainfall and temperature to identify their relationship. From Figure 2, we can see that temperature and rainfall are highly correlated. The temperature is low during high rainfall and there is no rainfall when the temperature in high.



**Figure 2: Temperature and rainfall relationship**

We calculated the values of the correlation between the temperature and rainfall. Table 5 represents the correlation matrix which proves that these variables have a high negative correlation. Since they are negatively correlated, for every increase in the temperature there is a decrease in rainfall and vice versa. If there is high correlation, positive or negative, the parameter estimates are unstable and a small change in the data can cause catastrophic changes in the variables estimates. We can counter this problem in the following ways: (1) getting more data, (2) dropping one variable, and (3) combining the variables. We will investigate on these options further in the report.

**Table 5:Pearson Correlation matrix for Temperature vs Rainfall**

|  |  |  |
| --- | --- | --- |
|  | Temperature | Rainfall |
| Temperature | 1 |  |
| Rainfall | **-0.720400255** | 1 |

**Step 4: Analyzing regression of Calls vs (Rainfall and Temperature):**

In this part, we will regress both the independent predicting variables, rainfall and temperature, against the dependent variable to be forecasted that is calls. Then, we will evaluate both the predictors based on parameters such as R-square, Adjusted R-square, and the significance of coefficients. The R-square value for calls vs rainfall and calls vs temperature were 38% and 51% respectively. From Table 7, we can observe that by combining the two predictors (rainfall and temperature), the R-square increased to 53% which is a growth of 4% than Model 2. Moreover, the adjusted R-square for calls vs rainfall and calls vs temperature is .36 and .50. However, after combining rainfall and temperature, the adjusted R-square increased to 0.51, that is, by 2% increase. The improvement is not very significant. Furthermore, in Table 6, the p-value (0.17) for rainfall is greater than the significance level (0.05), which shows that rainfall is an insignificant predictor in this model. On the other hand, the p-value (0.00) for temperature is less than the significance level (0.05). Since temperature outperforms rainfall as a predictor in all the measures, we are eliminating rainfall to get better insights about the model. Hence, calls vs temperature will be hereafter referred to as **MODEL #1**.

**Table 6: Model Statistics in Detail - Dynamic Regression (3 regressors, 0 lagged errors)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Coefficient** | **Std. Error** | **t-Statistic** | **p-value** |
| Rainfall | 128.8 | 93.45 | 1.378 | 0.1748 |
| Temperature | -101 | 25.51 | -3.952 | 0.00026 |
| \_CONST | 26,039 | 1,587 | 16.4 | 0 |

**Table 7: Within-Sample Statistics - Dynamic Regression for Calls vs Temperature and Rainfall**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample size** | **49** | No. parameters | 3 |
| Mean | 21,206.43 | Std. deviation | 1,665.74 |
| R-square | 0.53 | Adj. R-square | 0.51 |
| Durbin-Watson | 1.14 | Ljung-Box(18) | 30.0 P=0.96 |
| Forecast error | 1,161.44 | BIC | 1,267.71 |
| MAPE | 4.47% | SMAPE | 4.43% |
| RMSE | 1,125.33 | MAD | 929.53 |
| MAD/Mean Ratio | 0.04 |  |  |

## **Step 5: Capturing Further Seasonality**

In this section, we will further enhance our MODEL #1 for enhancing its accuracy. MODEL #1 was successful in capturing seasonal patterns due to weather conditions since the calls are seasonal by nature. However, MODEL #1 was unable to explain the variability of calls during times such as long vacations in the summer or national holidays and festivals. Therefore, we harness the Error Autocorrelation Function (hereafter referred to as Error ACF) for calls. By analyzing error ACF in Figure 2, we found that MODEL #1 captures most of the seasonality but failed to capture the seasonality in calls as evident from the spikes at lag 1 and lag 12. This implies that MODEL #1 is itself not enough and should be extended to capture this seasonality. As a result, we will introduce dummy variables or seasonal dummies to help our model have better accuracy.

A screenshot of a social media post

Description automatically generated

**Figure 3: Error Autocorrelation Function or Error ACF with calls vs temperature**

Now to capture the remaining seasonality, we are introducing 11 seasonal dummies with September as the reference month. By examining the plots in figure 3, the reference month is chosen based on the month which has often records the lowest calls for the course of the whole data. Table 8 shows the model details after adding 11 dummy variables in MODEL #1. Since much of the seasonality has already been accounted by weather conditions (as seen in Model #1), not all seasonal dummies will be significant. This can be observed in table 8, where only Apr, Jul, and Aug are significant at 10% level of significance. Therefore, we remove the insignificant seasonal dummies and keep only significant seasonal dummies to build our resulting model. This resulting model will have calls as dependent variable and temperature and significant seasonal dummies as independent variables or predictors and will be hereafter referred to as **MODEL #2**.

**Table 8: Model Statistics in Detail - Dynamic Regression (13 regressors, 0 lagged errors)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Coefficient** | **Std. Error** | **t-Statistic** | **p-value** |
| Temperature | -226.50 | 57.68 | -3.9260 | 0.0003737 |
| \_CONST | 33389.00 | 3531.00 | 9.4550 | 0.000000000027 |
| Jan | -918.00 | 1323.00 | -0.6940 | 0.49220 |
| Feb | -2049.00 | 1284.00 | -1.5950 | 0.11940 |
| Mar | -1194.00 | 1095.00 | -1.0910 | 0.28270 |
| Apr | -1935.00 | 841.20 | -2.3010 | 0.02730 |
| May | -1019.00 | 713.50 | -1.4280 | 0.16190 |
| Jun | 54.19 | 652.90 | 0.0830 | 0.93430 |
| Jul | 1919.00 | 697.70 | 2.7500 | 0.00927 |
| Aug | 1980.00 | 665.70 | 2.9740 | 0.00522 |
| Oct | -229.30 | 824.20 | -0.2783 | 0.78240 |
| Nov | -548.80 | 1075.00 | -0.5107 | 0.61270 |
| Dec | -760.80 | 1372.00 | -0.5545 | 0.58270 |

In Table 9, we can observe that MODEL #2 is a better model than MODEL #1 since the predictive power (R-square) of the resulting model has increased from 53% to 71%. Also, the Adjusted R-square of MODEL #2 has increased from 0.51 to 0.68 which is a good indication for our resulting model.

**Table 9: Within sample statistics - Regression of calls over temperature and significant seasonal dummies**

|  |  |  |  |
| --- | --- | --- | --- |
| Sample size | 49 | No. parameters | 5 |
| Mean | 21,206.43 | Std. deviation | 1,665.74 |
| R-square | 0.71 | Adj. R-square | 0.68 |
| Durbin-Watson | 1.12 | Ljung-Box(18) | 21.8 P=0.76 |
| Forecast error | 939.74 | BIC | 1,086.10 |
| MAPE | 3.07% | SMAPE | 3.04% |
| RMSE | 890.50 | MAD | 637.93 |
| MAD/Mean Ratio | 0.03 |  |  |

## **Step 6: Examining and dealing with autocorrelation of errors**

After building MODEL #2, we plotted error ACF (figure 5) and saw that there is still seasonality left to be captured. The spike at lag 1 resembles that we need to further refine our model to eradicate the seasonality factor from our final model and hence we move forward with our analysis.

A screenshot of a social media post

Description automatically generated

**Figure 4: Error ACF with calls vs temperature and dummy variables**

Another way to test the correlation problem is by using the Ljung-Box (LB) Statistics. This statistic helps us with concrete number to decide whether the error ACF plot is a white noise or not. In LB, we define null hypothesis where we decide whether to reject the null hypothesis or not. According to model diagnostics, it implies that the serial correlation (or autocorrelation) exists with the errors being not white noise. Serial correlation exists successive observations over time are related to one another and it is a huge problem and unless a model is free of this problem, we never proceed with that model.

A possible solution to remove serial correlation is Cochrane-Orcutt procedure. To remove the spike at lag 1 (ref. figure 5), we consider the errors are correlated only at the first lag and hence we add \_AUTO[-1] to our model. After adding, we again plot the error correlation matrix (ref. figure 6) and observe that there are no spikes left and hence we can say that now the errors of the regression can be considered as white noise and the model is free from serial correlation problem. Also, the LB statistics value dropped from 21.8 (MODEL #2) to 17.5 and hence it implies that LB statistics is significant. This model will be hereafter referred to as **MODEL #3**.

A screenshot of a social media post

Description automatically generated

**Figure 5: Error ACF with calls vs temperature and dummy variables and AUTO[-1]**

Now we compare the two models (ref. table 9), MODEL #2 and MODEL #3, based on certain criteria to check if there are any noticeable improvements and below are the observations:

**Table 10: Comparison of MODEL #2 vs MODEL #3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **R-square** | **MAPE** | **Adj. R-square** | **Ljung-Box statistics** |
| **MODEL #2** | 71% | 0.031 | 0.68 | 21.8 |
| **MODEL #3** | 77% | 0.028 | 0.74 | 17.5 |

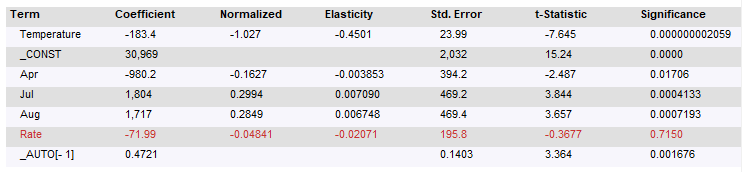
From table 9, we can observe that there is an oveall improvement in the model and now we can futher improve our model from here by taking MODEL #3 as the base for our model analysis.

## **Step 7: HEADING IS MISSING**

Based on Mr. DeCoria observation of cyclic trend, in this section we will try to improve the accuracy of the model by adding the unemployment factor “Rate” in our regression model. This will help us analyzae if the unemployment rate is a good predicfttor of the economic cyclic trend.

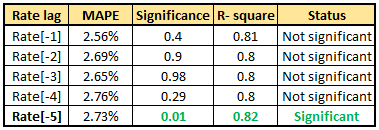
### Introducing **Rate**

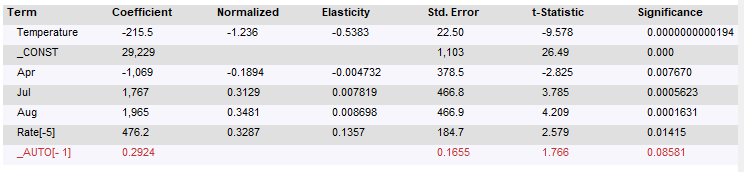
After adding the rate variable, we observed that at a significance level of 10%, the variable is insignificant. This means that volume of emergency calls are not dependent on the rate. The rate does not effect the volume of current calls, however, we know that the economic cycle or the unemployment rate lags behind the cyclic trend of the time series, hence, we will test the model by applying lagged values of the rate of unemployment. This will show us which lagged value of rate gives the best model when added to the model as a predictor.



On adding the lagged values of rate, we observe that at a significance level of 10%, Rate[-5] is significant with significance level of .014. The rate lags less than 5 appear to be insignificant for a 10% significance level.

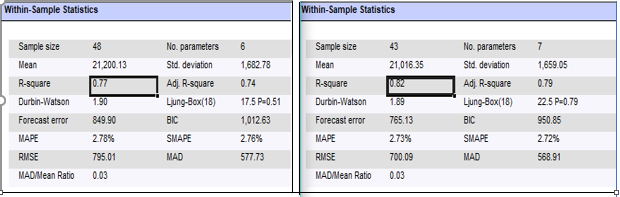
**Table 11: Statistics for model with Rate[-5] included in the predictors.**





When we compare model 3 with model 4, we find that model 3 explains 77% of the variability and model 4 explains 82% of the variability within the model. The values of adjusted R2, and Ljung box are higher for model 4 as compared to model 3. So, we can say that model 4 is better than model 3. We have a lesser Mean absolute percentage error which explains that the model has a better accuracy than the model 3.

**Model 3 Model 4**



## **Step 8: The final dynamic regression model**

As discussed in step 7, MODEL #4 our best model since it is the healthiest among all our resulting models and hence, we are building our final equation based on this model. The final dynamic regression model equation is as follows:

Our equation consists of the following components: temperature, 3 seasonal dummies, a lagged predictor variable (unemployment rate) of order 5, and a lagged error as independent variables.

While interpreting the coefficients estimates, when all the independent variables are kept constant except temperature, 1 unit increase in temperature will decrease the calls by 216 (approx.). Likewise, when all the independent variables are kept constant except 3 seasonal dummies, in comparison to September, there are on average 1069 less calls in April, 1767 more calls made in July, and 1965 more calls in August. Next, when all the independent variables are kept constant except unemployment rate, 1 unit increase in rate will increase the number of calls by 476 (approx.) after a period of 5 months. Also, when all the independent variables are kept constant except lagged error, we can say that 29% of the current month’s errors can be accounted because of last month’s errors.

After interpreting all the coefficients, the model makes perfect business sense. According to the model, if the temperature rises by 1 unit, on an average 216 less calls will be received. The reasons that affect cars in cold temperatures are car’s battery works harder to start and it diminishes quickly; tire pressure can decrease at rest and increase when the car is moving meaning a shortened lifespan for the tire. Now talking about the unemployment rate, when a person is unemployed it effects his lifestyle and services his/her car less often and hence the car breaks more often resulting in more road side assistance calls. Therefore, we can state that the resulting model makes more business sense.

## **Step 9: Forecast Accuracy**

Now we will use holdout analysis to assess the genuine forecast accuracy. This is carried out by testing the model using outside data not used in the model building process which is of the period from Sep 1992 to Jan 1993. We append this external data to our original data and MODEL #4 variable composition is applied. Then, a 5 months forecast is achieved by putting a holdout period of 5 observations. Hence, we create table 15 by using the results of Out-of-Sample Static Evaluation from Forecast Pro and to compare our forecast we have included the actual values. Furthermore, we analyze this table by utilizing Mean Absolute Percentage Error (MAPE) values to evaluate forecast accuracy. In 1993-Jan, we can see that our forecast gives a 3.38% MAPE and since lower MAPE for a forecasting model implies better forecasting accuracy, we can say that the generated forecast is accurate.

**Table 12: Dynamic Regression Out-of-Sample Static Evaluation from Forecast Pro**

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Forecast** | **Actual** | **MAPE** |
| 1992-Sep | 20,119.93 | 20,251.00 | 0.65% |
| 1992-Oct | 21,130.09 | 22,069.00 | 2.45% |
| 1992-Nov | 22,934.36 | 23,268.00 | 2.11% |
| 1992-Dec | 24,469.63 | 26,039.00 | 3.09% |
| 1993-Jan | 24,936.90 | 26,127.00 | 3.38% |

When we used the same data and calculate for Expert Selection, we did not receive better outcome, as shown in table 16, since our forecast gives a higher MAPE of 6.86% and since lower MAPE for a forecasting model implies better forecasting accuracy, we cannot say that the generated forecast is accurate.

**Table 13: Expert Selection Out-of-Sample Static Evaluation from Forecast Pro**

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Forecast** | **Actual** | **MAPE** |
| 1992-Sep | 20,168.86 | 20,251.00 | 0.41% |
| 1992-Oct | 21,899.33 | 22,069.00 | 0.59% |
| 1992-Nov | 22,210.31 | 23,268.00 | 1.91% |
| 1992-Dec | 22,512.70 | 26,039.00 | 4.82% |
| 1993-Jan | 22,203.44 | 26,127.00 | 6.86% |

Moreover, we take Expert Selection method into consideration to compare our regression model. We observe that overall Expert Selection method is not a better model than Dynamic Regression since the predictive power of the resulting model has decreased from 82% to 63%. Also, the Adjusted R-square of the Expert Selection method is .63 which is lower in comparison to 0.79 of Dynamic Regression. Consequently, the MAPE is compared from both model’s Out-of-Sample Static Evaluation and we observe that MAPE has increased from 2.83% to 6.86%. Hence, Dynamic Regression model can be considered a good model with better forecast accuracy.

**Table 14: Comparison of Expert Selection and Dynamic Regression**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **R-square** | **MAPE** | **Adj. R-square** | **Ljung-Box statistics** |
| **Expert Selection** | 63% | 6.86% | 0.63 | 16.4 |
| **Dynamic Regression** | 82% | 3.38% | 0.79 | 22.5 |

Since, Dynamic Regression model is a better selection, we plot the forecast values of both the regression model and that of expert selection with the actual values to see how they compare with each other. From figure 4, it is observed that there is an initial rise in the Expert Selection but after a month it goes down. However, the forecasting for Dynamic regression is low on the first month but it rises to the par with Actual Values making it the best choice for Mr. Decoria’s business.

**Figure 4: Line chart for comparing Regression, Actual Values and Expert Selection**

# **Conclusion and Recommendation**