ISDS 526-01 21352

Forecasting for Analytical Decision Making

**Project 4: Explaining Variation of Service Calls to Counteract Operational Expenses**

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](#_Toc532837309)

[**The Forecasting Problem** 4](#_Toc532837310)

[**Dynamic Regression Model Building** 5](#_Toc532837311)

[**Step 1: Analyzing regression of Calls vs Rainfall:** 5](#_Toc532837312)

[**Step 2: Analyzing regression of Calls vs Temperature:** 6](#_Toc532837313)

[**Step 3: Analyzing Rainfall vs Temperature:** 6](#_Toc532837314)

[**Step 5: Capturing Further Seasonality** 8](#_Toc532837315)

[**Step 6: Examining and dealing with autocorrelation of errors** 10](#_Toc532837316)

[**Step 7: Incorporating the Unemployment Rate Factor in the Model** 12](#_Toc532837317)

[**Step 8: The final Dynamic Regression model** 14](#_Toc532837318)

[**Step 9: Forecast Accuracy** 16](#_Toc532837319)

[**Conclusion** 18](#_Toc532837320)

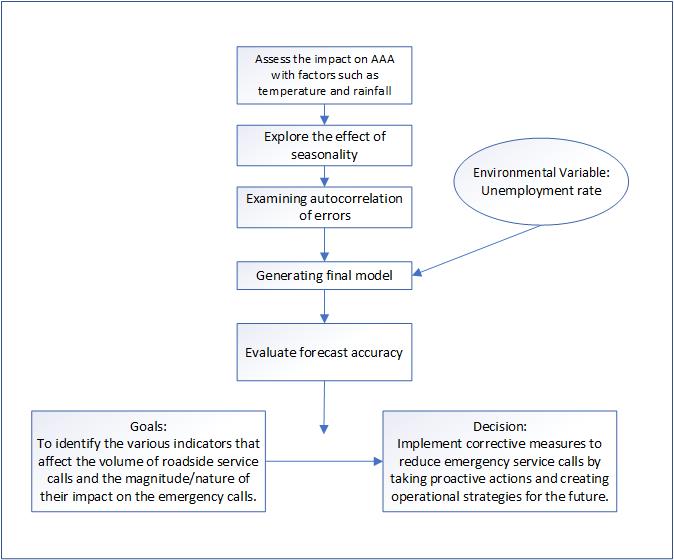
# **Executive Summary**

American Automobile Association (hereafter referred to as **AAA**) Washington is among the two regional automobile clubs affiliated with AAA, offering its members a variety of automobile and automobile-related services. Within the member benefits, the most popular service is emergency roadside assistance – 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 total operating budget.

AAA has hired us 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. We have analyzed the factors such as average temperature, the amount of rainfall, the unemployment rate for Washington State, and the seasonal trend in the time series. By analyzing these factors, we will get to know the impact of these factors on the service calls. We build a forecasting model using the significant factors, and we will also incorporate the behavior of the service calls with respect to the economic cycle. It is also important to check the effect of errors from the previous observations. These collectible insights will assist Mr. DeCoria’s decision-making process and will thereby allow him to take appropriate actions to prevent additional expenses of excessive roadside service calls. Mr. Decoria can use the interpretations from our project to make the decision regarding the company’s operations to counter the operational cost.

# **The Forecasting Problem**

To determine the variation within the emergency service call volume data, we aim to build a dynamic regression model. 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 factors such as temperature, rainfall, and rate. However, these factors alone might not fully explain the fluctuation in service calls. Hence, we need to utilize data analysis techniques such as correlation, autocorrelation and seasonality checks to conclude the effect on service calls. When we are sure to have incorporated all the fluctuations of the time series, we can build the final model. In the end, we need to evaluate the accuracy of our forecast model, so that business decisions can be efficiently made to improve the service and reduce cost.



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

# **Dynamic Regression Model Building**

In this section, we will build multiple dynamic regression models to examine the potential effects of factors such as rainfall, temperature, and the unemployment rate on the number of calls. We will first analyze the effect of each factor separately, followed by both factors included simultaneously 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, we will explain some of the historical variations which may lead to more accurate forecasts to help AAA.

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

In this step, we will determine whether there is a relationship between the emergency service calls (hereafter, referred to as calls) and the 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 in Table 2, 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.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Coefficient** | **Std. Error** | **t-Statistic** | **p-value** |
| Rainfall | 394.8 | 74.21 | 5.321 | 0.000002824 |

**Table 2: Within-Sample Statistics**

|  |  |  |
| --- | --- | --- |
| R-square | Adj R-square | MAPE |
| 38% | 36% | 5.08% |

## **Step 2: Analyzing regression of Calls on 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. The p-value in Table 3 indicates that there is 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 |

***Table 4: Within-Sample Statistics***

|  |  |  |
| --- | --- | --- |
| **R-square** | **Adj R-square** | **MAPE** |
| 51% | 50% | 4.59% |

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

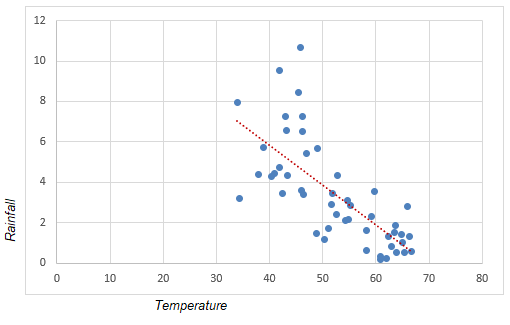
We will now analyze rainfall and temperature to identify their relationship. We found that the temperature and rainfall are highly correlated. The temperature is low during high rainfall, and there is no rainfall when the temperature in high.

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.

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

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

From Figure 2, we can see the time series plot of temperature and rainfall from the observations. Since they are negatively correlated, for every increase in the temperature there is a decrease in rainfall and vice versa.



**Figure 2: Temperature and rainfall relationship**

When there is a 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 these options further in the report.

**Step 4: Analyzing regression of Calls on (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. This will help us to identify the effect of both the factors on service 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 on rainfall and calls on temperature was 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 2% more than our previous model with temperature. The adjusted R-square for calls on rainfall and calls on temperature were 36 % and 50 % respectively. However, after combining rainfall and temperature, the adjusted R-square increased to 51 %, which is a 1% improvement. As we can see, 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 from the model. Hence, the **calls on 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 |

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

|  |  |  |
| --- | --- | --- |
| R-square | Adj R-square | MAPE |
| 53% | 51% | 4.47% |

**Step 5: Capturing Further Seasonality**

In this section, we will further enhance our MODEL #1 for improving 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 3, we found that MODEL #1 captured 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. To address this problem, we will introduce dummy variables as 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 monthly seasonal dummies with September as the reference month. By examining the plot in figure 3, the reference month is chosen based on the month which has often recorded the lowest calls for the course of the whole data. Table 8 shows the model details after adding 11 monthly 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.

**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 |

*\*\*Rows marked in red are insignificant regressors at 10% significance level*

Let us remove the insignificant seasonal dummies highlighted red in Table 8 and keep only significant seasonal dummies to build our resulting model. This resulting model will have calls as dependent variable with temperature and significant seasonal dummies as independent variables or predictors and will be hereafter referred to as **MODEL #2**.

**Table 9: Model 2 Within sample statistics - Regression of calls over temperature and significant seasonal dummies**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coefficient** | **p-value** | **R-square** | **Adj R-square** | **MAPE** |
| Temperature | -178 | 2.957E-12 | 71% | 68% | 3.07% |
| Apr | -1,306 | 0.01183 |  |  |  |
| Jul | 1,900 | 0.001956 |  |  |  |
| Aug | 1,963 | 0.0006781 |  |  |  |

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%, which is very significant. Also, the Adjusted R-square of MODEL #2 has increased from 51 % to 68 % which is a good indication for our resulting model.

## 

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

After building MODEL #2, we plotted error ACF shown in figure 4 and saw that there is still some 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**

We tested the correlation problem by using the Ljung-Box (LB) Statistics. This statistic helps us with concrete numbers to decide whether the error is a white noise or not. According to model diagnostics, the LB indicates that there is no major serial correlation at higher values of lag. However, from the error ACF in Figure 4 we can see there is a serial correlation at lag 1 which needs to be addressed. 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 where we consider the errors encountered in the previous observations. Hence, to remove the error autocorrelation spike at lag 1 in figure 4, we add the error from the previous observation as \_AUTO [-1] to our model. This will incorporate the error from the previous observation. After adding \_AUTO [-1], we again plot the error autocorrelation function shown in figure 5 to finally observe that there are no spikes left and the errors are white noise. Hence, we can now say that the errors of the regression are 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 well 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 accuracy measures to check if there are any noticeable improvements and below are the observations:

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **R-square** | **MAPE** | **Adj. R-square** | **Ljung-Box statistics** |
| **MODEL #2** | 71 % | 3.07 % | 68 % | 21.8 |
| **MODEL #3** | 77 % | 2.78 % | 74 % | 17.5 |

From table 10, it is clear that there is an overall improvement in the model. Now we look to futher improve our model from here by taking MODEL #3 as the base for our analysis.

## **Step 7: Incorporating the Unemployment Rate Factor in the Model**

Based on Mr. DeCoria analysis of cyclic trend, we will try to improve the accuracy of the model by adding the unemployment factor “Rate” in our regression model. This will help us analyze if the unemployment rate helps the model as a good predictor of the economic cyclic trend.

Let us add the “Rate” attribute to our model, in Table 11 we observed that at a significance level of 10%, the rate variable is insignificant. This means that volume of emergency calls are not dependent on the rate of unemployment during that particular month.

**Table 11: Model Statistics in Detail - Dynamic Regression (6 regressors, 1 lagged errors)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Coefficient** | **Std. Error** | **t-Statistic** | **p-value** |
| Temperature | -183.4 | 23.99 | -7.645 | 2.059E-09 |
| \_CONST | 30,969 | 2,032 | 15.24 | 0 |
| Apr | -980.2 | 394.2 | -2.487 | 0.01706 |
| Jul | 1,804 | 469.2 | 3.844 | 0.0004133 |
| Aug | 1,717 | 469.4 | 3.657 | 0.0007193 |
| Rate | -71.99 | 195.8 | -0.3677 | 0.715 |
| \_AUTO [-1] | 0.4721 | 0.1403 | 3.364 | 0.001676 |

*\*\*Rows marked in red are insignificant regressors at 10% significance level*

The Table 11 indicates that the current unemployment rate is not related to 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. Considering the relation, we will test the model by applying lagged values of the rate of unemployment to our model. 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 Rate[-5] is significant with a p-value of .014. The rate lags less than 5 appear to be insignificant for a 10% significance level. The comparison is provided in Table 12.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rate lag** | **MAPE** | **R- square** | **p-value** | **Status** |
| Rate [-1] | 2.56% | 0.81 | 0.403 | Not significant |
| Rate [-2] | 2.69% | 0.8 | 0.913 | Not significant |
| Rate [-3] | 2.65% | 0.8 | 0.979 | Not significant |
| Rate [-4] | 2.76% | 0.8 | 0.293 | Not significant |
| **Rate [-5]** | 2.73% | **0.82** | **0.014** | **Significant** |

**Table 12: Statistics showing the significance of different lagged values of Rate**

We will label this model with Rate [-5] indicator as **MODEL #4**. The models statistics are shown in the Table 13.

**Table 13: Model Statistics in Detail - Dynamic Regression (6 regressors, 0 lagged errors)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Coefficient** | **Std. Error** | **t-Statistic** | **p-value** |
| Temperature | -215.5 | 22.5 | -9.58 | 1.942E-11 |
| Apr | -1069.0 | 378.5 | -2.83 | 0.00767 |
| Jul | 1767.0 | 466.8 | 3.79 | 0.00056 |
| Aug | 1965.0 | 466.9 | 4.21 | 0.00016 |
| Rate [-5] | 476.2 | 184.7 | 2.58 | 0.01415 |
| \_AUTO [-1] | 0.3 | 0.2 | 1.77 | 0.08581 |

We can now compare MODEL #3 with MODEL #4 in Table 14. The R-square value indicates that MODEL #3 explains 77% of the variability and MODEL #4 explains 82% of the variability within the calls. The values of adjusted R2 is 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 indicates better accuracy. The lower value of the Bayesian information criterion (BIC) further indicates that the MODEL #4 is simpler than the MODEL #3 and more accurate.

**Table 14: Comparison of MODEL #3 and MODEL #4**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **R-square** | **Adj. R-square** | **MAPE** | **BIC** |
| **MODEL #3** | 77 % | 0.74 | 2.78 % | 1021.4 |
| **MODEL #4** | 82 % | 0.79 | 2.73 % | 950.85 |

## **Step 8: The final Dynamic Regression model**

MODEL #4 has proved to be our best model since it is the healthiest among all our resulting models. Hence, we are building our final regression 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. Let us interpret the effect of each component.

1. Temperature - When all the independent variables are kept constant except temperature, 1 unit increase in temperature will decrease the calls by approximately 216.
2. Seasonal dummies - When all the independent variables are kept constant except for the seasonal dummies, on and average, there are 1069 less calls in April, 1767 more calls made in July, and 1965 more calls in August as compared to September.
3. Unemployment rate - When all the independent variables are kept constant except unemployment rate, 1 unit increase in rate will increase the number of calls by approximately 476 after a period of 5 months.
4. Lagged error (**et-1**) - As we have incorporated the error from the previous observation, we can say that when all the independent variables are kept constant, we can say that 29% of the current month’s errors can be accounted in the next month’s forecast.

After interpreting all the coefficients, the model makes proper business sense. Some of the reasons that affect cars in cold temperatures are:

1. The car batteries are less effective and lose their charge quicker than in higher temperatures.
2. The tire pressure is generally lower at rest in lower temperatures and increase when the car is moving, leading to a shortened lifespan.
3. When a person becomes unemployed, their lifestyle might change drastically. This lifestyle change also affects the way the person’s car is used. Instead of using the vehicle as a daily driver to work and back home, it might be used less (which leads to the vehicle being parked for longer periods, further possibly leading to damaged tires and stale fuel). A less “planned out” lifestyle might also lead to the vehicle not getting serviced enough.

## **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.

We created table 15 by using the results of Out-of-Sample Static Evaluation from Forecast Pro to compare the Model 4 forecast with the Expert selection model. Furthermore, we will evaluate the accuracy of the models by utilizing the Mean Absolute Percentage Error (MAPE) values. We can see that the forecast by Model 4 gives a 3.38% MAPE which indicates a better forecasting accuracy, and we can say that the generated forecast is quite accurate.

When we used the same data to generate the statistics for Expert Selection, we did not receive a better outcome. As shown in table 15, the expert selection forecast gives a higher MAPE of 6.86% which is almost double the MAPE % of our Model 4.

**Table 15: Comparison of Expert Selection and Dynamic Regression**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **R-square** | **Adj. R-square** | **MAPE** |
| **Expert Selection** | 64 % | 63 % | 6.86 % |
| **Dynamic Regression** | 82% | 79 % | 3.38% |

The Expert Selection method has a significantly lower predictive power (64 %) than the Dynamic Regression model (82 %). Also, the Adjusted R-square of the Expert Selection method is 63 % which is lower in comparison to 79 % of the Dynamic Regression.   
Hence, the Dynamic Regression model can be considered a good model with better forecast accuracy.

Since the Dynamic Regression model is a better performing model, 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 6, 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 6: Line chart for comparing Regression, Actual Values and Expert Selection**

# **Conclusion**

To conclude this project on explaining the variation within the service calls volume data, we can confidently say that we have found the potential predictors to be of actual statistical value, making them decent predictors for the number of incoming roadside service calls. We can explain the variation in the number of roadside service calls and how it is related to different factors in the following way:

The unemployment rate as a general measure of Washington state’s economic state has an effect on how people use their vehicles. Personal vehicles might not be services regularly because unemployed people tend to live a less structured and planned lifestyle. It also means that the vehicles might be used differently (e.g. for longer road trips instead of daily commutes) and could be parked for longer periods of time if the vehicle’s owners have less reason to use the car regularly; leading to prematurely worn-out tires, depleted batteries, and stale vehicle fluids.

A lower temperature has a negative effect on tires when compared to more mild temperatures. This is the case because tires in colder weather (especially when parked) have a lower pressure than warmed-up tires. Every time the vehicle is being started and driven in cold weather, the tires will expand a little, and vice versa when being parked. This leads to the tires being worn out more easily and being at risk of breaking. Lower temperatures also lead to more slippery conditions (even when roads are dry, colder weather leads to tires not having as much grip as they would in warmer temperatures), which can cause more accidents.

The other factor that has to be considered is rainfall, because the more rainfall on average, the higher the number of emergency roadside service calls. As it is true for the lower temperature, more rainfall causes more dangerous driving conditions, which lead to more accidents and therefore more service calls. Hence, Mr. Decoria has to consider the above aspects carefully to establish effective emergency service which helps to reduce the overall operational cost.