Effects on Automobile Fuel Consumption

Edward Moradian

California State University, East Bay

Author Note

Edward Moradian, Department of Statistics, California State University, East Bay.

Correspondence concerning this article should be addressed to Edward Moradian, Department of Statistics, California State University, East Bay, Hayward, CA 94542

E-mail: emoradian@horizon.csueastbay.edu

Abstract

The research question of the study was to estimate the combined miles per gallon of a vehicle using different characteristics of a vehicle as the predictor variables. A best fit regression model was explored using data from the Environmental Protection Agency. A first-order regression model included 11 predictor variables along with the response variable of combined miles per gallon fuel consumption. After diagnostic testing of the first-order regression model, it was decided that a polynomial regression model of the second-order would be the model that would better fit the data. There were a number of reasons for this – the presence of a curvilinear appearance in residuals plots, a failed transformation of the first-order regression model, a statistically significant result from a Tukey’s test for nonadditivity and a very similar r-squared result. For model validation, the polynomial regression model was tested on newer data from the Environmental Protection Agency. A correlation between the predicted values of the polynomial regression model with the actual values of the data returned r = .9897.

Keywords: Automobile Fuel Consumption, Miles Per Gallon, Estimating MPG, Polynomial Linear Regression, Predicting MPG

Effects on Automobile Fuel Consumption

The research question investigated in the current study was how to fit a regression model to estimate the combined miles per gallon of a vehicle. The current study estimated the response variable of combined miles per gallon by different characteristics of a vehicle as the predictor variables. Some of these characteristics were quantitative variables, while others were qualitative variables. The research literature has a myriad number of studies that have statistically modeled vehicle fuel consumption. One recent study modeled vehicle fuel consumption by fitting some of the same predictor variables of the current study into regression models (Ang, Fwa, & Poh, 1991; Essenhigh, Shull, Blackadar, & McKinstry, 1979). The current study attempts to build on the research literature by introducing new predictor variables into the statistical model and by using new data.

Prediction was the primary research goal, but a secondary goal was descriptive. The secondary research goal was to determine which combination of predictor variables were the best fit for the regression model. In other words, which combination of predictor variables combined to make the strongest prediction of combined miles per gallon fuel consumption. A large number of predictor variables was selected intentionally to determine which ones were included and were not included in a final regression model. In this way, combined miles per gallon fuel consumption was described by what characteristics of a vehicle are strong enough predictors to be included in a final regression model.

The practical significance of the study is to be able to estimate the combined miles per gallon of a vehicle through the vehicle’s characteristics. This can provide a prediction of the combined miles per gallon consumption of a vehicle before any actual testing is done on the vehicle to determine its combined miles per gallon. Regression models like the one in the current study are also useful in that they allow for miles per gallon to be estimated before a car is manufactured. The different specifications of a vehicle can be inputted for an estimate on a future vehicle’s miles per gallon fuel consumption.

# Method

## Data Preparation

The data used to model a regression equation was taken from the Environmental Protection Agency’s website on fuel economy – fueleconomy.gov (Environmental Protection Agency, 2017). The data is the result of the vehicle testing by the Environmental Protection Agency and by vehicle manufacturers. The data set was for vehicles of the model year of 2017. Not all types of vehicles were included – plug-in hybrids, electric vehicles and fuel-cell vehicles were not included in the data set.

The data preparation started from deciding which of the vehicle characteristics to include as predictor variables in the preliminary regression model. The original data set included 160 different characteristics of the vehicles listed. Of those 160 different characteristics, only 12 predictor variables were selected for in regression modeling. Of these 12, the following 6 predictor variables were quantitative predictor variables: engine displacement, the number of cylinders, the number of gears, greenhouse gas rating, smog rating and combined carbon dioxide emissions. Of these 12, the following 6 predictor variables were qualitative predictor variables: country of manufacture, transmission type, drive, fuel usage, carline class and oil viscosity. Four interactions between the 12 predictor variables were also chosen for the preliminary regression model. These interactions were included: number of cylinders with engine displacement, greenhouse gas rating with smog rating, smog rating with combined carbon dioxide emissions and greenhouse gas rating with combined carbon dioxide emissions. The reasoning for the first interaction was that if a vehicle had a high number of cylinders with a high volume in engine displacement, then there might be a combined effect of the two, a general high-performance engine, on the combined miles per gallon of a vehicle. The reasoning for the second interaction was that if a vehicle had a high level of emissions of greenhouse gas as well as high emissions of smog, then those two effects might have a combined effect, a general polluting vehicle, on the combined miles per gallon of a vehicle. The same logic followed for the third and fourth interaction pairs.

There was a choice between choosing the response variable as either city miles per gallon fuel consumption, highway miles per gallon fuel consumption or combined city and highway miles per gallon fuel consumption. The choice was made to use combined city and highway miles per gallon fuel consumption as to have an average of fuel consumption as the response variable.

Overall, in the final data set there were 1,085 observations of 13 variables. This meant that there was a total of 1,085 vehicles with 13 different characteristics (i.e., the 12 predictor variables and 1 response variable). The distribution of the response variable combined miles per gallon had a range from 11 to 58 with a mean of 23.37 and a median of 23.00.

### Exploratory Analysis

Correlations were calculated among the response variable and each predictor variable to determine if there was a relationship between the response variable and each predictor variable. None of the correlations between the response variable and predictor variables were close to zero which indicates that there is a relationship between the response variable and each of the predictor variables. This suggested that the predictor variables had predictability towards the response variable.

Correlations were calculated among the predictor variables to determine if any predictor variables were highly correlated with one another and therefore redundant. The number of cylinders was highly positively correlated with engine displacement (r = .92). Consequently, the number of cylinders as a predictor variable was dropped from the preliminary model. In turn, the interaction between the number of cylinders and engine displacement was also dropped from the preliminary model. Greenhouse gas rating was highly negatively correlated with combined carbon dioxide emissions (r = -.97). Consequently, greenhouse gas rating was dropped from the preliminary model. In turn the interaction between greenhouse gas rating with smog rating and with combined carbon dioxide emissions were dropped from the preliminary model. Greenhouse gas rating and number of cylinders were dropped instead of engine displacement and combined carbon dioxide emissions because the latter two had a more precise measure with a larger scale.

### Model

The final regression model was a polynomial regression model of the second-order. It was in the second-order because of having the predictor variable combined carbon dioxide emissions transformed to the second-power. All other predictor variables were to the first-power and not transformed in any way.

There were four main reasons for choosing a polynomial regression model of the second-order than a linear regression model of the first-order. First, in the first regression model, that was in the first-order, the residual plots showed a curved general trend (see Figure 1). This indicated a need for a curvilinear regression function. Second, in the first regression model, from the residual plots, a Tukey’s test for nonaddivity returned a p-value of < 2.2e-16. This supports the need for a curvilinear regression function indirectly by supporting the need for something other than an additive model. Third, a reciprocal transformation was tried on the response variable to keep the model in the first-order. This transformation was given by a box-cox transformation that returned a lambda value of -1. The residual plots for this transformed model showed linearity. A studentized Breusch-Pagan test, however, showed that the variances were not constant in this model (p-value < 2.2e-16). The residual plots also indicated that the variances were not constant (see Figure 2). Lastly, the model with the reciprocal transformation in the first-order had a R-squared value at 99.56%. This was only marginally larger than the R-squared value for the polynomial regression model at 98.29%.

Finding the best fit for a polynomial regression model started with creating a full model with all of the chosen predictor variables except for the ones that were dropped in the exploratory analysis. Because the residuals plots of the first-order linear regression model showed a curvilinear general trend only with the predictor variable combined carbon dioxide emissions when plotted against the residuals, the predictor variable combined carbon dioxide emissions was transformed into the second-order. With a full-model with the transformed predictor variable combined carbon dioxide emissions, a model comparison using the Akaike’s Information Criterion was used against an intercept only model using forward, backward and mixed selection. This yielded the model of best fit with an AIC of – 377.78. The model of best fit was the model with the following predictor variables: combined carbon dioxide emissions (to the second-order), fuel usage, engine displacement, carline class, country of manufacture, number of gears, oil viscosity and drive description. Overall, none of the interactions were included in the model of best fit. Smog rating, number of cylinders, greenhouse gas rating and transmission type were also not included in the model of best fit.

#### Procedure

# Results

The R-squared value for the polynomial regression model was 98.29% indicating that 98.29% of the variation in our response variable is reduced by our predictor variables. An F-test on the model returned an F-value of 2,241 with a p-value of < 2.2e-16. This indicates that the model is a better fit than the intercept-only model. All of the quantitative predictor variables were highly statistically significant at an alpha level of .05. At least one level of each qualitative predictor variable was statistically significant at an alpha level of .05 and therefore all were kept in the polynomial regression model. The y-intercept of the regression model was highly statistically significant at an alpha level of .05.

In the diagnostic plots of the polynomial regression model, it was observed that the residual plots for the residuals vs. numeric predictor variables showed that the residuals generally fall within a horizontal band centered at y=0 (see Figure 3). The same is observed for the residuals vs. fitted values plot. For the residual vs. categorical predictor variable plots, the boxes had a very similar center and spread. In the marginal model plots, the dashed and solid smooth lines are almost matching. This indicates that our model fits the data (see Figure 4).

The residual vs. categorical predictor variable plots showed the presence of potential outliers. This was also the case for these residual plots for the first regression model. When formal testing was done, only the observations 602 and 1056 were decided to be removed. The Bonferroni p-values for outlier testing was statistically significant for both observations. Both observations also had high influence.

For model validation, the polynomial regression model was used to predict the combined miles per gallon of vehicles in a new data set. This new data set came from the same source, the Environmental Protection Agency and was the data set for the vehicles of the model year 2018. A prediction was performed using the polynomial regression model on the vehicles of the data set to predict their combined miles per gallon. The correlation between the predicted values and the actual values was r = .9897 indicating a very good model with strong predictive power (see Figure 5).

# Discussion

External validity was one limitation of the study as the regression model is only generalizable to the class of vehicles that are not plug-in hybrids, electric vehicles or fuel-cell vehicles. This is because the data set used for the regression analysis did not include these class of vehicles. Indeed, Figure 5 shows that the prediction of the polynomial regression model weakens substantially for vehicles with a combined miles per gallon greater than 35. These types of vehicles would typically fall in a combined miles per gallon rating greater than 35. It was observed that the data set consisted of only 30 observations of vehicles with a combined miles per gallon greater than 35. Therefore, the polynomial regression model was mostly trained on vehicles with a combined miles per gallon less than 35. In turn, the polynomial regression model is not a strong predictor for cars with a combined miles per gallon greater than 35.

Another problem with the external validity of the regression model is if the model is generalizable to older vehicle model years. The model year of all of the vehicles in the data set used was 2017. Changes in the build of vehicles have changed in many ways from older generation vehicles. If we are measuring predictor variables based on changing underlying variables of a car coming from its parts or the effects of its parts then a regression model that accounts for a range of model years would be appropriate. However, the miles per gallon fuel consumption of existing vehicles is already well-established. Regression models like the one in the current study have utility in predicting the miles per gallon fuel consumption of future vehicles that have not gone through testing.

Another limitation of the study, in terms of internal validity, was that the study estimated the response variable of the combined miles per gallon only by vehicle characteristics. This limited the scope of the measuring of combined miles per gallon to only by vehicle characteristics and not in general. Other types of predictor variables, other than vehicle characteristics, may have predictive power for estimating miles per gallon. One study investigated whether driver characteristics had an effect on vehicle emissions and fuel consumption. The researchers found statistically significant differences between novice drivers, experienced smooth drivers and aggressive experienced drivers on vehicle fuel consumption and emissions (Zheng, Lib, Zuylena, & Lua, 2017). Incorporating all the possible predictor variables in a regression analysis may yield a better model than the one in the current study with possibly stronger predictor variables.

To help alleviate the problem of generalizability of the regression model to vehicles that are plug-in hybrids, electric vehicles and fuel-cell vehicles new data should be used that contains data about these types of vehicles. With these new types of vehicles, a new regression model can have a new predictor variable such as fuel consumption type to test for the prediction power of vehicles that are conventional, hybrid, electric or fuel-cell. This would all help with the prediction power towards vehicles with a combined miles per gallon greater than 35.

Overall, the polynomial regression model is useful in estimating vehicle combined miles per gallon with a few caveats – for vehicles less than 35 combined miles per gallon, vehicles that are not of the class plug-in hybrids, electric vehicles or fuel-cell vehicles, vehicles yet to be tested and for effectual characteristics only by vehicle characteristics.

References

Ang, B. W., Fwa, T. F., & Poh, C. K. (1991, August). A Statistical Study on Automobile Fuel Consumption. *Energy, 16*(8), 1067-1077. doi:https://doi.org/10.1016/0360-5442(91)90139-D

Environmental Protection Agency. (2017). *Download Fuel Economy Data.* Retrieved from www.fueleconomy.gov: https://www.fueleconomy.gov/feg/epadata/17data.zip

Essenhigh, R. H., Shull, H., Blackadar, T., & McKinstry, H. (1979). Effect of Vehicle Size and Engine Displacement on Automobile Fuel Consumption. *Transportation Research Part A: General, 13*(3), 175-177. doi:https://doi.org/10.1016/0191-2607(79)90068-2

Zheng, F., Lib, J., Zuylena, H. v., & Lua, C. (2017). Influence of Driver Characteristics on Emissions and Fuel Consumption. *Transportation Research Procedia, 27*, 624-631. doi:https://doi.org/10.1016/j.trpro.2017.12.142

Residual Plots for Residuals vs. Combined Carbon Dioxide Emissions and Residuals vs. Predictor Values of the First Regression Model

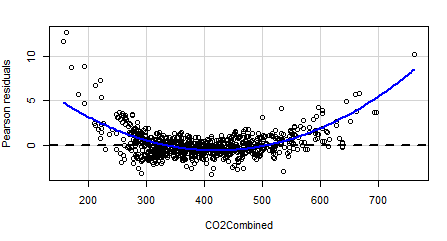
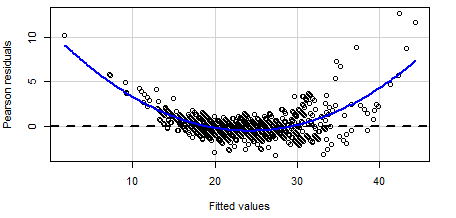
 

Figure 1. In both residual plots of the first regression model that was in the first-order, a clear curvilinear nature is the general trend.

Residual Plots for Residuals vs. Combined Carbon Dioxide Emissions and Residuals vs. Predictor Values of the First Regression Model

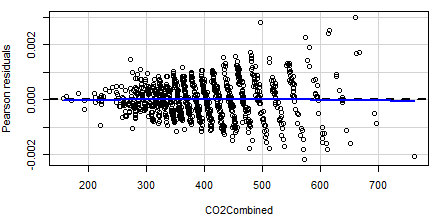
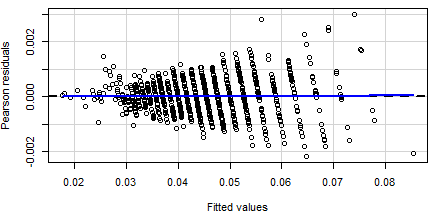
 

Figure 2. In both residual plots of the first regression model, the variances of the residuals increase with increasing values of combined carbon dioxide emissions or the fitted values. This shows heteroskedasticity and therefore that the constant variance assumption was not upheld. Linearity, however, was upheld.

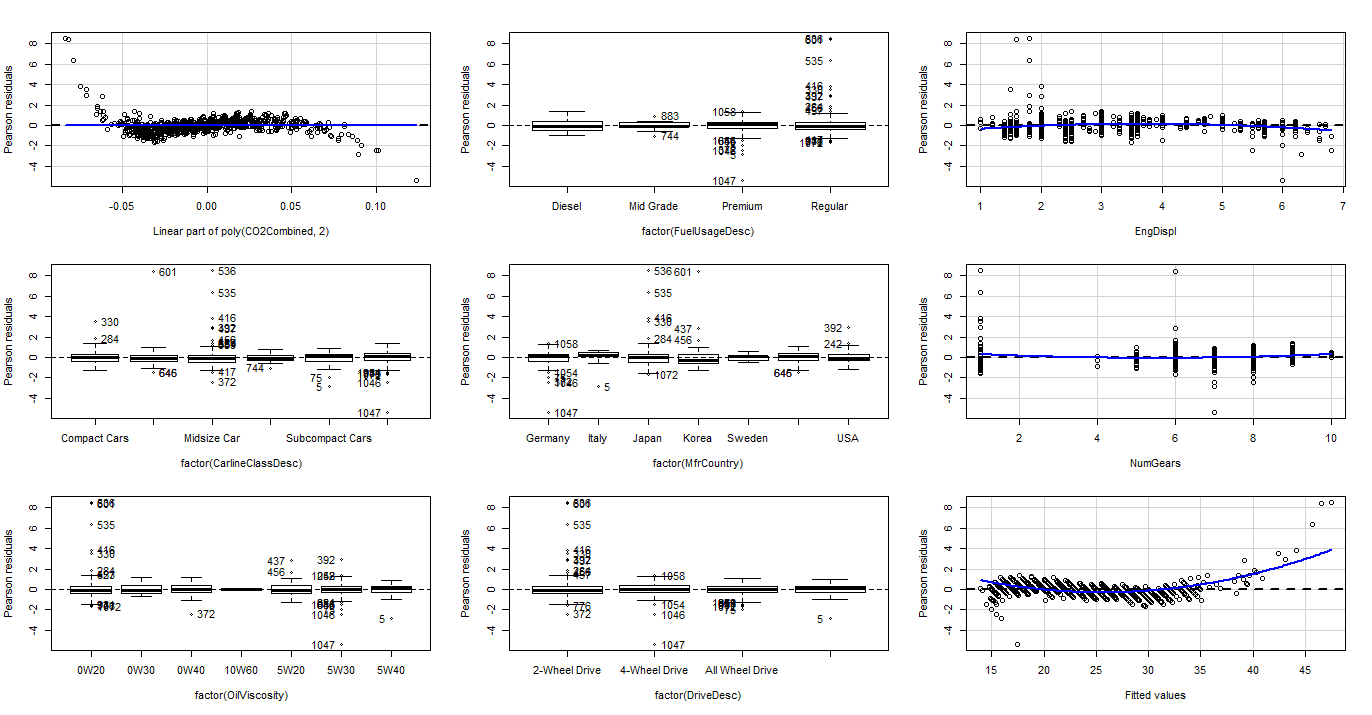
Residual Plots for Residuals vs. Predictor Variables and Residuals vs. Fitted Values of the Polynomial Regression Model 

Figure 3. The residuals plots showed that the constant variance and linearity assumptions were valid for the regression model.

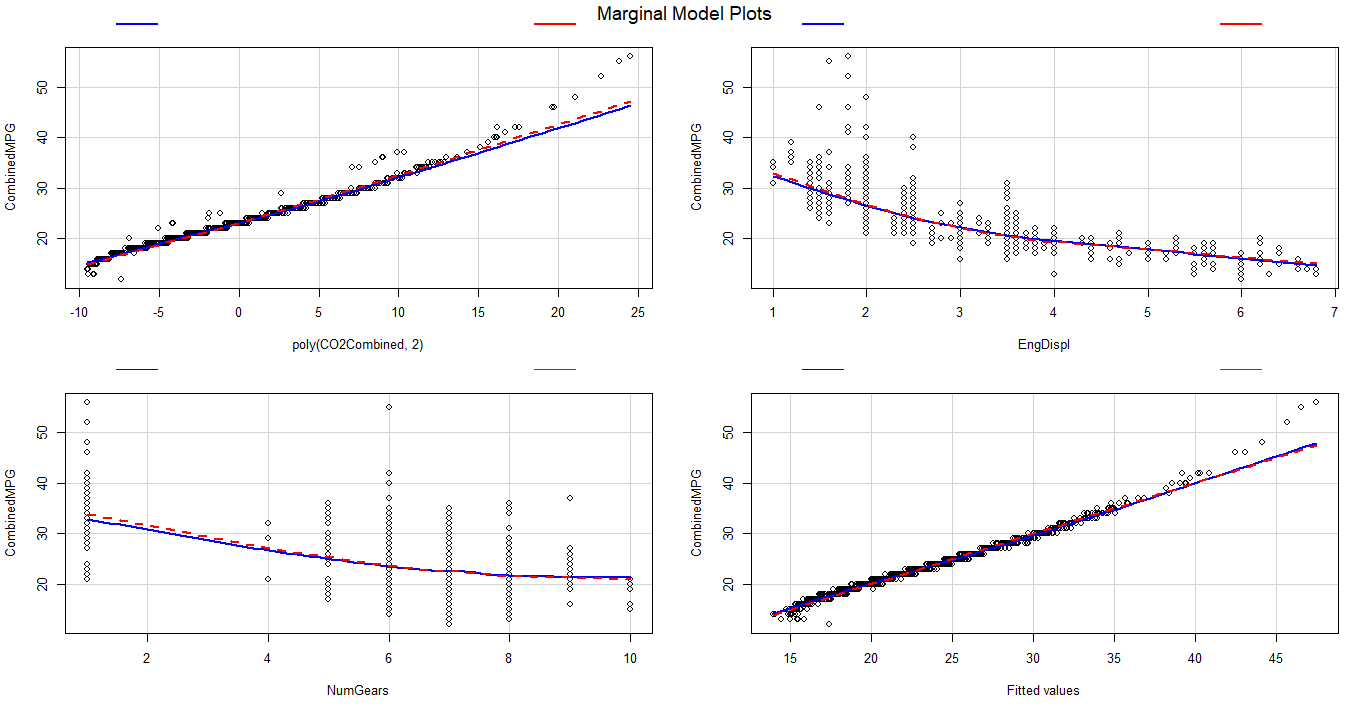
Marginal Model Plots of the Response Variable vs. the Quantitative Predictor Variables in the Polynomial Regression Model 

Figure 4. The solid blue smooth and the dashed red smooth are matching in the marginal model plots of the quantitative predictor variables.

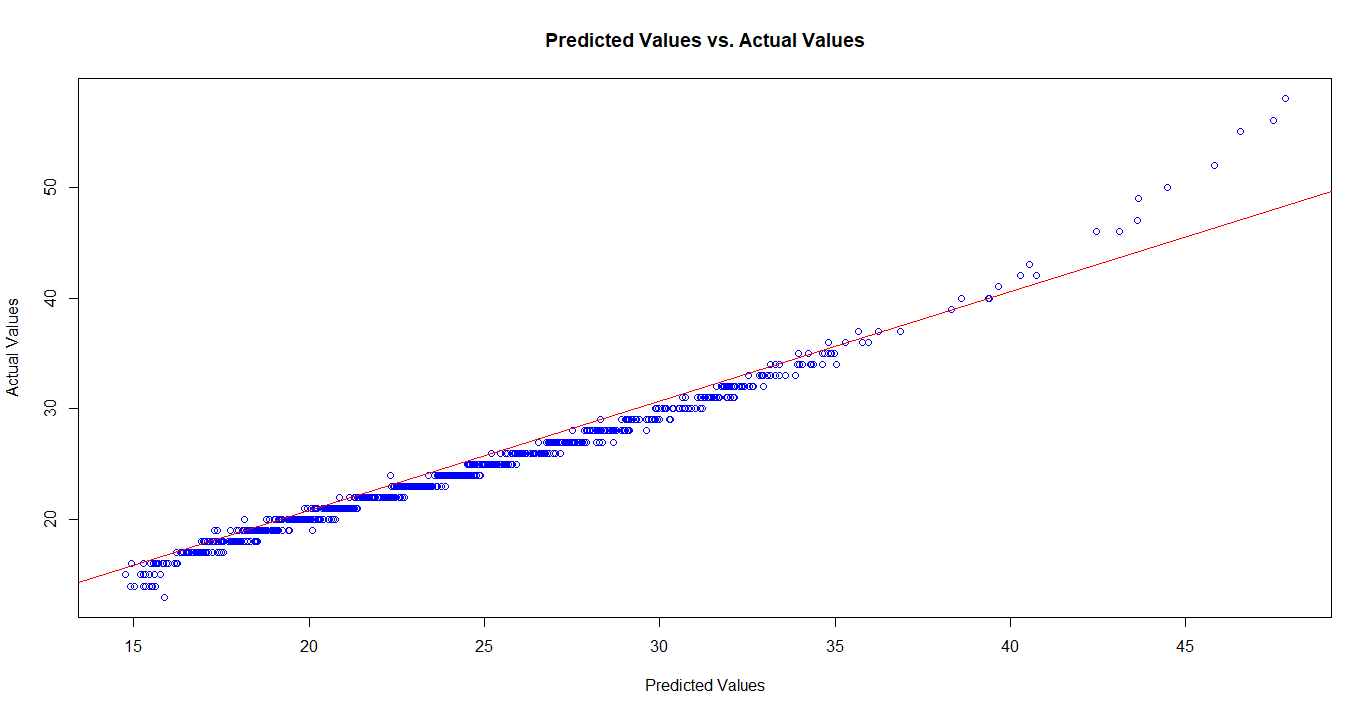
Correlation Graph of the Actual Values vs. the Predicted Values in Model Validation 

Figure 5. A correlation plot shows the predicted values of the polynomial regression model correlated against the actual values of combined miles per gallon combined. Prediction is strong except after 35 combined miles per gallon combined.