Factors Correlated with Value Retention Amongst Motor Vehicles in India

Shan Santhakumar, David Alber, Leo Shi

Contents

List of Figures

List of Tables

Variable	Min	Max	Median	SD	Description
Year	2003	2018	2014	2.89	Year the vehicle was sold.
Selling_Price	$120.31	$42,107.10	$4,331.02	6,114.93	Price vehicle was sold at.
Present_Price	$384.98	$111,403.36	$7,699.58	$10,399.39	Original price of vehicle.
Owner	0	3	0	0.2479	Number of previous owners.
Mileage	310.7	310,700	19,884.8	24,164.3	Number of miles vehicle has on it.
RVRP	0.1054	0.9893	0.6546	0.2024	Percent of value vehicle has retained (as a decimal).

Table 1: Data Summary Statistics

Introduction

Our research question is what factors lead to value retention in motor vehicles. We are interested in this topic because motor vehicles are a big investment and it would be useful to know how one can best retain the value of a motor vehicle, so as not to face too big of a financial loss when purchasing one. Additionally, this information could benefit those looking to get a motor vehicle for a discounted price as they could look at motor vehicles with characteristics that cause them to lose their value and therefore get a good deal on a motor vehicle.

The Data

Our dataset had 301 observations of motor vehicle sales which were scrapped from websites in India. Each observation contained 9 variables.

- Car_Name
- Year
- Selling_Price
- Present_Price
- Kms_Driven
- Fuel_Type
- Seller_Type
- Transmission
- Owner

However, in order to extract more meaningful data, we converted Selling_Price and Present_Price from its original units which were Lakhs (100,000) of Indian Rupees into US dollars. We also converted Kms_Driven into miles driven, so it is easier for us to interpret. Lastly, in order to quantify the response variable we wished to obtain we created a new variable to store the Relative Value Retention Percentage (RVRP) of each motor vehicle, using the formula:

$$RVRP = 1 - \frac{Present\_Price - Selling\_Price}{Present\_Price}$$

EDA

Summary Statistics

From our summary statistics (see Table 1), we discovered that the vehicle that sold for the lowest price was a Bajaj Pulsar 150, a motorcycle sold in 2006 for only $120.31. While the vehicle that sold for the most is a Toyota Land Cruiser, an SUV which was sold in 2010 for $42,107.10. We have year of purchase ranging from

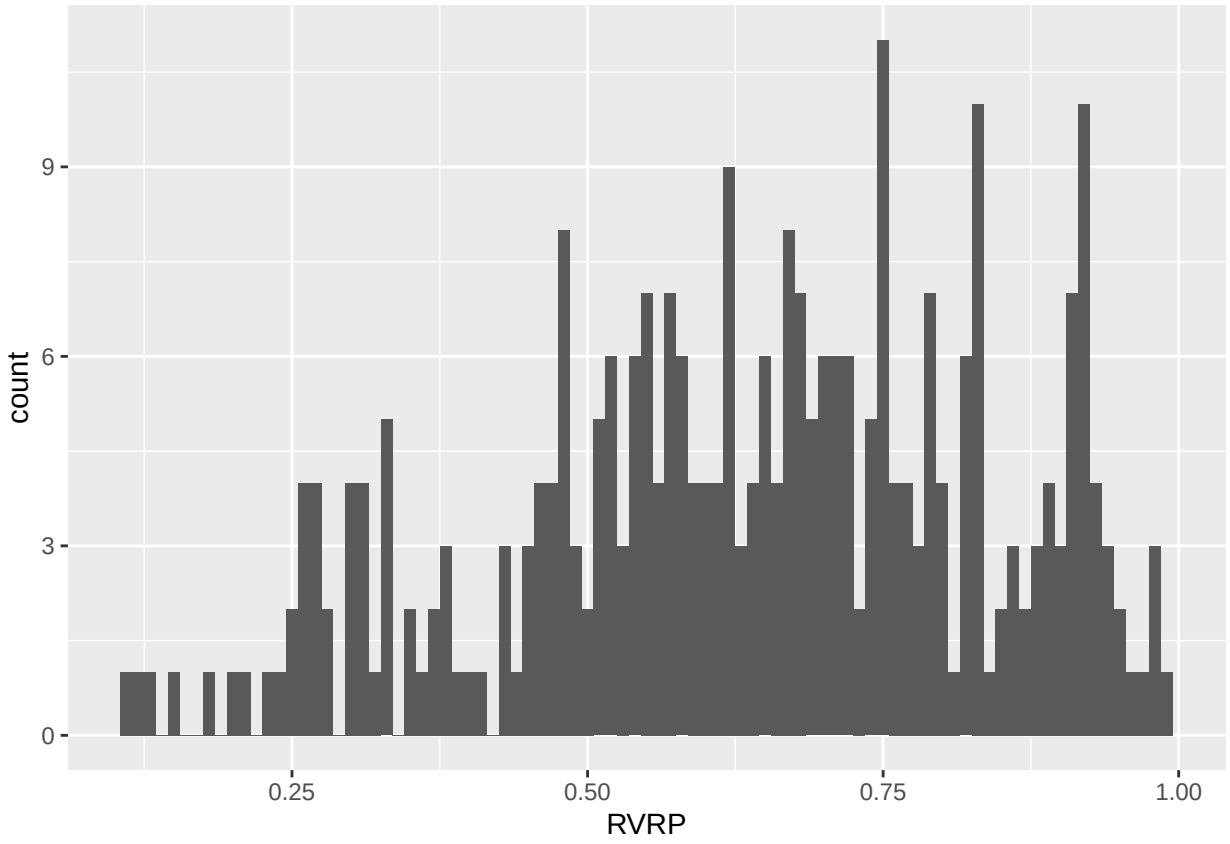


Figure 1: Histogram of Response Variable

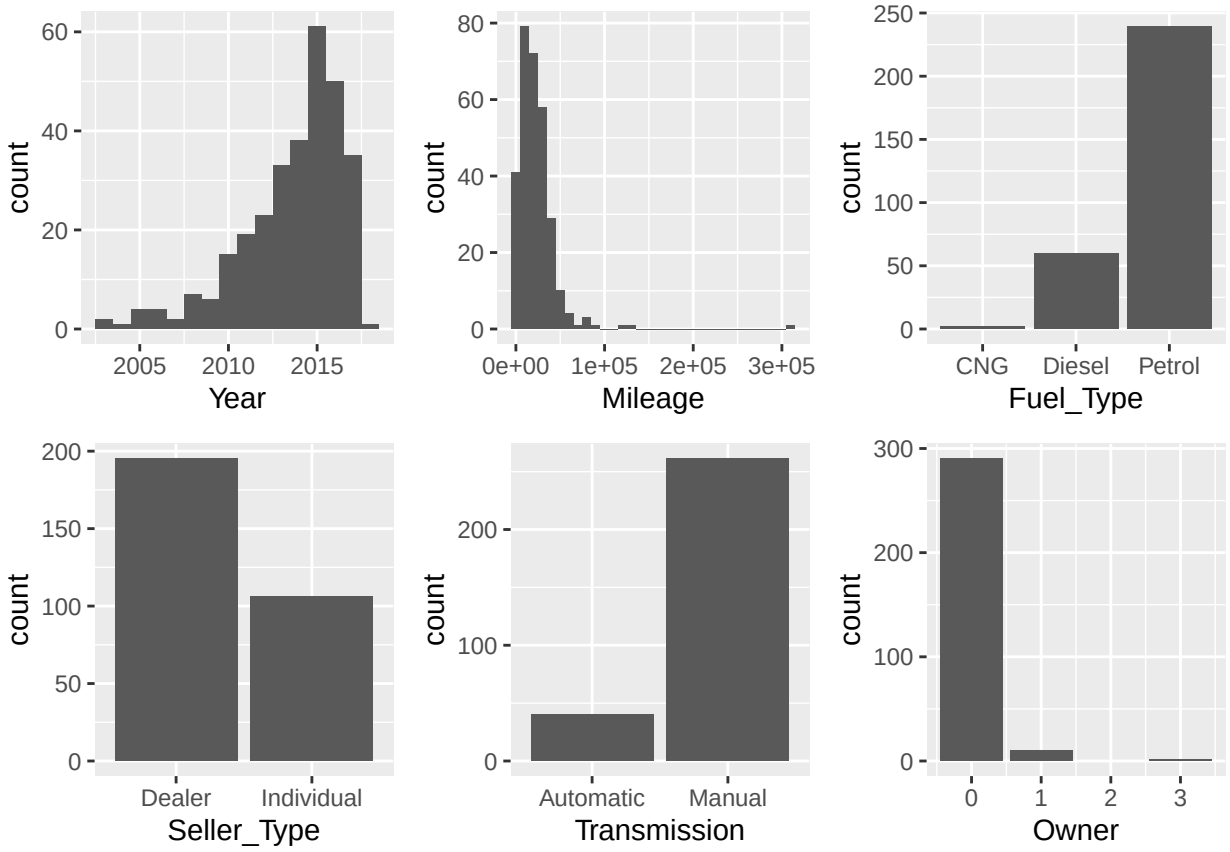


Figure 2: Histogram of Exploratory Variables

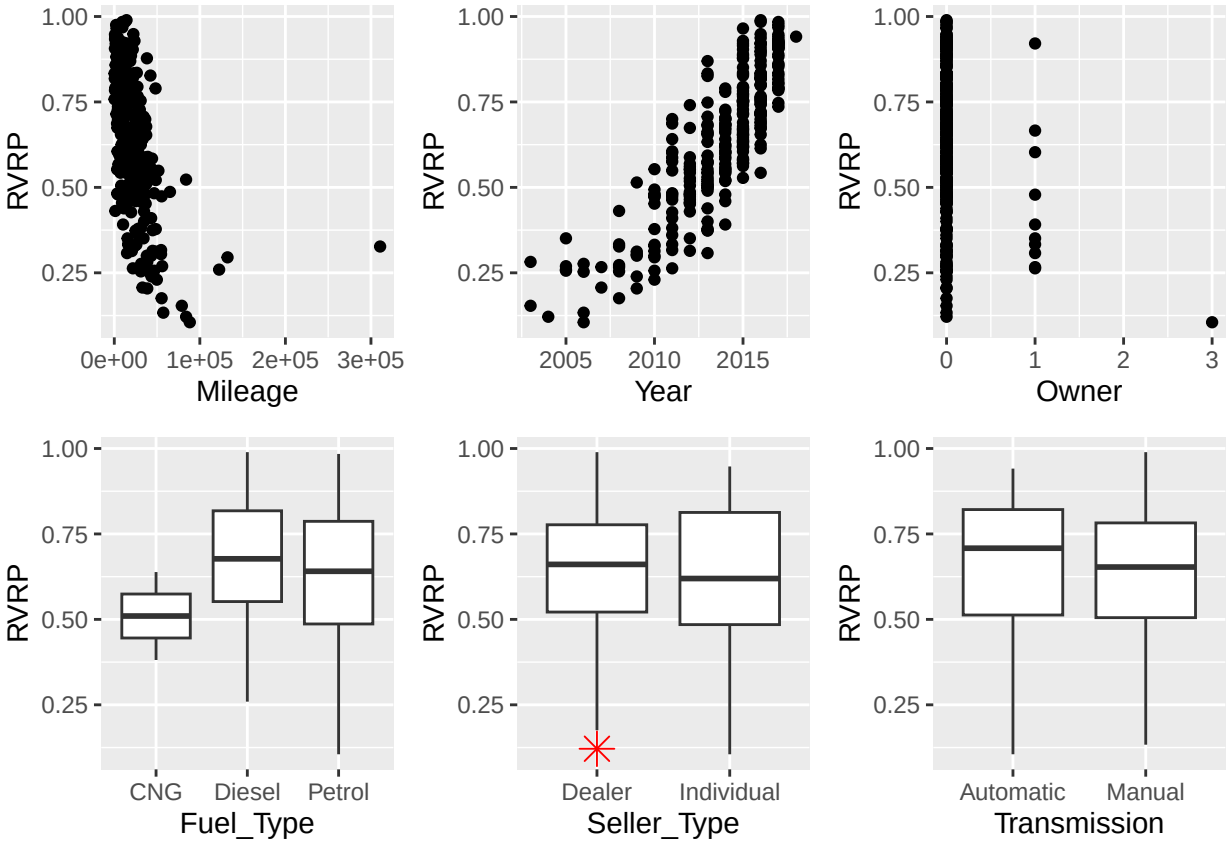


Figure 3: Predictor vs Response Scatterplot/Boxplot

2003 to 2018, number of owners ranging from 0 to 3, and original price of the vehicle ranging from $384.98 up to $111,403.36. Our variable of interest, RVRP, also has quite the range with some vehicles retaining as little as 10.54% of its original value and others retaining up to 98.93% of its original value. From this summary of our data, we get an idea of what range our data covers. When making models and conclusions using this data we can avoid extrapolation by keeping our scope within the bounds of this particular dataset.

Regression Analysis

Final Model

```
##
## Call:
## lm(formula = RVRP ~ Year + I(Year^2) + Owner + Mileage + Year *
##     Mileage, data = carData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24448 -0.07259 -0.00185  0.05663  0.28676
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.386e+03  2.751e+03   2.685  0.00767 **
## Year         -7.402e+00  2.734e+00  -2.708  0.00717 **
## I(Year^2)     1.854e-03  6.790e-04   2.731  0.00670 **
## Owner        -6.773e-02  2.379e-02  -2.847  0.00473 **
## Mileage       4.307e-04  2.067e-04   2.084  0.03801 *
## Year:Mileage -2.147e-07  1.028e-07  -2.088  0.03769 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09993 on 295 degrees of freedom
## Multiple R-squared:  0.7602, Adjusted R-squared:  0.7562
## F-statistic: 187.1 on 5 and 295 DF,  p-value: < 2.2e-16
```

Our final model is:
$RVRP = 7386 - 7.402 * Year + 0.001854 * Year^2 - 0.06773 * Owner + 0.0004307 * Mileage - 2.147 * 10^{-7} * Year * Mileage$

The residuals for our final model satisfies the linearity, constant variance, and normality assumptions, however it does not satisfy the independence assumption (see "Checking Assumptions of Final Model" in Appendix for details). Our model has an R^2 value of 0.7562, meaning 75.62% of the variance in RVRP can be explained by our model. However, since our data does not satisfy the independence assumption, we cannot make conclusions outside of our sample.

Processs of Obtaining the Final Model

When obtaining our final model, we began with our stepwise regression model (see "Building a Stepwise Regression Model" in Appendix for details). We took the variables discovered to be of significant contribution to our response variable, RVRP, from our stepwise model and looked to see if the addition of second order and interaction terms would improve the residuals further and create an even more accurate model. These variables were Year, Mileage, Owner, and Fuel_Type. In order to determine which interaction and second order term would be needed, if any at all, we looked at the residuals for each individual variable one at a

time and found their functional forms by trial and error and moved from one variable to the next. We knew we found a variable's proper functional form when its residuals no longer showed a pattern. After going through this process for each variable from our stepwise regression model, we arrived at our final model.

Note: We did not consider any transformations on our response variable (RVRP) as it was already normally distributed (see figure 1).

Alternate Model Considerations

Models we excluded include the initial step wise regression model we made as well as any models with interaction terms with the categorical variable, Fuel_Type, as we ultimately found Fuel_Type to not have a significant relationship with our response variable (see "Recovering Function Form" in Appendix for details). We also tried models removing some of the variables from the final model we ended up with, however all of the variables in our final model contributed a significant amount to our Y variable (see "Alternate Model Considerations" in Appendix for details).

Conclusion

Based on our final model, we would suggest that for those that want a vehicle to retain as much of its value as possible, to make sure that the vehicle has as few previous owners as possible, has driven the least possible number of miles possible, and is as new as possible. Those that want to find a good deal on a car should do the opposite, find a car with as many previous owner as possible, with the most miles driven as possible, and have the car be as old as possible. However, it is important to keep in mind that this recommendation is only applicable to this sample, and may not be applicable to data outside of this sample.

We came to this conclusion based on the fact that the coefficient of the Year^2 term in our model is positive, which would indicate an upward facing curve between Year and RVRP, therefore newer cars will keep more of their original value. Since Owner has a negative coefficient, more owners decreases RVRP, and since Mileage and the Interaction between Year and Mileage is negative, then as Mileage increases RVRP is expected to decrease.

Limitations

Our model failed to meet the independence of errors assumption, therefore it cannot be said to be representative of data outside of our sample. In order for us to meet this assumption, we would either have to collect additional data or clean up our current data until it satisfies our independence assumption.

Additionally, we only had 301 observations originating from India, and the data collection method was described vaguely. The only information we were given about how the data was collected was that it was scrapped from the web, therefore we were unable to determine an accurate population for this data.

Lastly, our final model assumes that our stepwise regression model selected all of the variables relevant for explaining RVRP. It should be noted that the other two categorical variables not included in the stepwise model, Seller_Type and Transmission, which we did not consider interaction terms for in our final model, may have been able to improve our model. The other variable we did not account for as we felt it would be challenging to incorporate into our model given our time constraints, was the name of the vehicle. The name of a vehicle, and the associated brand recognition that comes with a vehicle's name could very well have an effect on that vehicle's value retention.

Appendix

Check for Collinearity in Variables

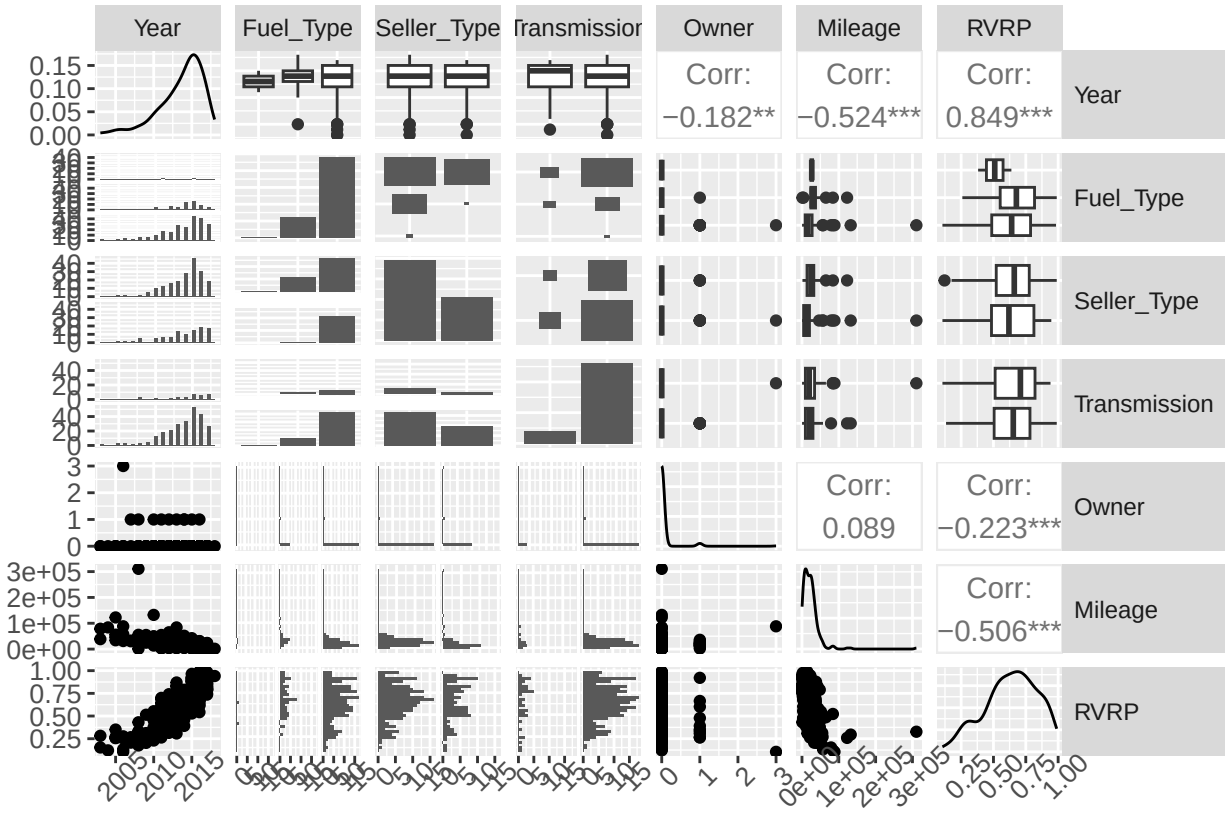


Figure 4: Correlation Matrix

Our predictor variables do not seem to have strong correlations with one another from the correlation matrix, with the highest correlations between predictors being between year and mileage at -0.524 and Owner and Year at -0.182. However, in order to determine if these correlations will cause multicollinearity issues, further testing will be required.

Multicollinearity: Non Significant t-tests for all (or nearly all) independent variables when the overall F-test is significant

```
##
## Call:
## lm(formula = RVRP ~ Year + Fuel_Type + Seller_Type + Transmission +
##     Owner + Mileage, data = carData2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.239918 -0.073147 -0.004528  0.069101  0.274804
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.084e+02  5.128e+00 -21.141  < 2e-16 ***
## Year                   5.413e-02  2.544e-03  21.276  < 2e-16 ***
## Fuel_TypeDiesel        1.181e-01  7.554e-02   1.563  0.11909
## Fuel_TypePetrol        7.769e-02  7.477e-02   1.039  0.29963
## Seller_TypeIndividual  7.935e-03  1.368e-02   0.580  0.56227
## TransmissionManual    -1.391e-02  1.821e-02  -0.763  0.44582
## Owner                 -5.850e-02  2.507e-02  -2.334  0.02028 *
## Mileage               -9.150e-07  3.086e-07  -2.965  0.00328 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1049 on 293 degrees of freedom
## Multiple R-squared:  0.7376, Adjusted R-squared:  0.7314
## F-statistic: 117.7 on 7 and 293 DF,  p-value: < 2.2e-16
```

The F-test for H0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ is highly significant (F=117.7, p-value=2.2e-16). Therefore we can reject H0 for any α greater than 0.0001 and conclude that at least one of the parameters $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ is nonzero.

The t-tests for Seller_Type, Fuel_Type, and Transmission, however, are nonsignificant, thus are less likely to end up in the final model.

Multicollinearity: Opposite signs (from what is expected) in the estimated parameters

From data from a Progressive Car Insurance article we expect value retention, y, to increase when Year sold (X1) increases or when Fuel Type changes from CNG (X2 & X3). Y will decrease when the number of owners (X6) or mileage (X7) increases. Therefore Progressive expects a positive relationship between y and X1, X2 and X3 , and a negative relationship with y and X6 and X7, which supports what is seen by the output.

Multicollinearity: Variance Inflation Factor (VIF) for a beta parameter greater than 10

```
##                   GVIF Df GVIF^(1/(2*Df))
## Year          1.475791  1        1.214821
## Fuel_Type     1.206842  2        1.048124
## Seller_Type   1.167754  1        1.080627
## Transmission 1.045963  1        1.022723
## Owner         1.052987  1        1.026151
## Mileage       1.516368  1        1.231409
```

No VIF values are greater than 10, therefore our explanatory variables do not seem significantly correlated, therefore we will proceed onto our stepwise regression model without dropping any of these variables.

Building a Stepwise Regression Model

```
## Start:  AIC=-960.8
## RVRP ~ 1
##
##                 Df Sum of Sq     RSS      AIC  F value    Pr(>F)
## + Year           1    8.8494  3.4365 -1342.27 769.9655 < 2.2e-16 ***
## + Mileage        1    3.1443  9.1416 -1047.78 102.8443 < 2.2e-16 ***
```

```
## + Owner         1    0.6086 11.6773  -974.09  15.5839 9.845e-05 ***
## + Fuel_Type     2    0.1790 12.1070  -961.21   2.2027    0.1123
## <none>                      12.2859  -960.80
## + Seller_Type   1    0.0173 12.2686  -959.22   0.4213    0.5168
## + Transmission  1    0.0011 12.2849  -958.82   0.0264    0.8711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-1342.27
## RVRP ~ Year
##
##                Df Sum of Sq     RSS     AIC  F value  Pr(>F)
## + Mileage       1    0.0628  3.3737 -1345.8   5.5488 0.01914 *
## + Owner         1    0.0588  3.3777 -1345.5   5.1872 0.02346 *
## + Fuel_Type     2    0.0536  3.3829 -1343.0   2.3529 0.09687 .
## <none>                       3.4365 -1342.3
## + Transmission  1    0.0012  3.4353 -1340.4   0.1008 0.75106
## + Seller_Type   1    0.0002  3.4363 -1340.3   0.0142 0.90510
## - Year          1    8.8494 12.2859  -960.8 769.9655 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-1345.82
## RVRP ~ Year + Mileage
##
##                Df Sum of Sq    RSS     AIC  F value  Pr(>F)
## + Fuel_Type     2    0.0850 3.2887 -1349.5   3.8250 0.02290 *
## + Owner         1    0.0597 3.3140 -1349.2   5.3514 0.02139 *
## <none>                      3.3737 -1345.8
## + Transmission  1    0.0070 3.3667 -1344.4   0.6135 0.43409
## + Seller_Type   1    0.0024 3.3712 -1344.0   0.2147 0.64346
## - Mileage       1    0.0628 3.4365 -1342.3   5.5488 0.01914 *
## - Year          1    5.7679 9.1416 -1047.8 509.4847 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-1349.5
## RVRP ~ Year + Mileage + Fuel_Type
##
##                Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + Owner         1    0.0555 3.2332 -1352.6   5.0601  0.025221 *
## <none>                      3.2887 -1349.5
## + Transmission  1    0.0042 3.2844 -1347.9   0.3816  0.537210
## + Seller_Type   1    0.0010 3.2876 -1347.6   0.0928  0.760861
## - Fuel_Type     2    0.0850 3.3737 -1345.8   3.8250  0.022904 *
## - Mileage       1    0.0942 3.3829 -1343.0   8.4798  0.003864 **
## - Year          1    5.3365 8.6252 -1061.3 480.3160 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-1352.62
## RVRP ~ Year + Mileage + Fuel_Type + Owner
##
##                Df Sum of Sq    RSS     AIC  F value    Pr(>F)
```

```
## <none>                      3.2332 -1352.6
## + Transmission  1    0.0061 3.2271 -1351.2   0.5592  0.455163
## + Seller_Type   1    0.0034 3.2298 -1350.9   0.3121  0.576795
## - Owner         1    0.0555 3.2887 -1349.5   5.0601  0.025221 *
## - Fuel_Type     2    0.0807 3.3140 -1349.2   3.6835  0.026299 *
## - Mileage       1    0.0938 3.3270 -1346.0   8.5539  0.003715 **
## - Year          1    5.0572 8.2904 -1071.2 461.4206 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After running our stepwise model selection using both forward and backward selection, our model chose Year, Mileage, Fuel_Type and Owner as the most significant variables that will give the lowest AIC value and therefore create the best model.

```
##
## Call:
## lm(formula = RVRP ~ Year + Mileage + Fuel_Type + Owner, data = carData2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.239378 -0.076329 -0.007458  0.068116  0.269960
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.087e+02  5.090e+00 -21.355  < 2e-16 ***
## Year             5.427e-02  2.526e-03  21.481  < 2e-16 ***
## Mileage         -8.857e-07  3.028e-07  -2.925  0.00372 **
## Fuel_TypeDiesel  1.207e-01  7.534e-02   1.602  0.11024
## Fuel_TypePetrol  8.277e-02  7.437e-02   1.113  0.26663
## Owner           -5.584e-02  2.482e-02  -2.249  0.02522 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1047 on 295 degrees of freedom
## Multiple R-squared:  0.7368, Adjusted R-squared:  0.7324
## F-statistic: 165.2 on 5 and 295 DF,  p-value: < 2.2e-16
```

From the summary of our stepwise model selection, we can see that this model has an R^2 value of 0.733, so 73.3% of the variance in RVRP can be explained by the variables Year, Mileage, Fuel_Type and Owner. Moreover, there is a positive relationship between Year and Fuel_Type and RVRP and a negative correlation between the variables Mileage and Owner and RVRP.

Checking Assumptions for Stepwise Model

Linearity

The plot of residuals vs. predicted does not show a pattern.
The plot of residuals vs. Year does not show a pattern.
The plot of residuals vs. Mileage does not show a pattern.
The plot of residuals vs. Fuel Type does not show a pattern.
The plot of residuals vs. Number of Owners does not show a pattern.
The linearity condition is satisfied for all significant explanatory variables.

Constant Variance

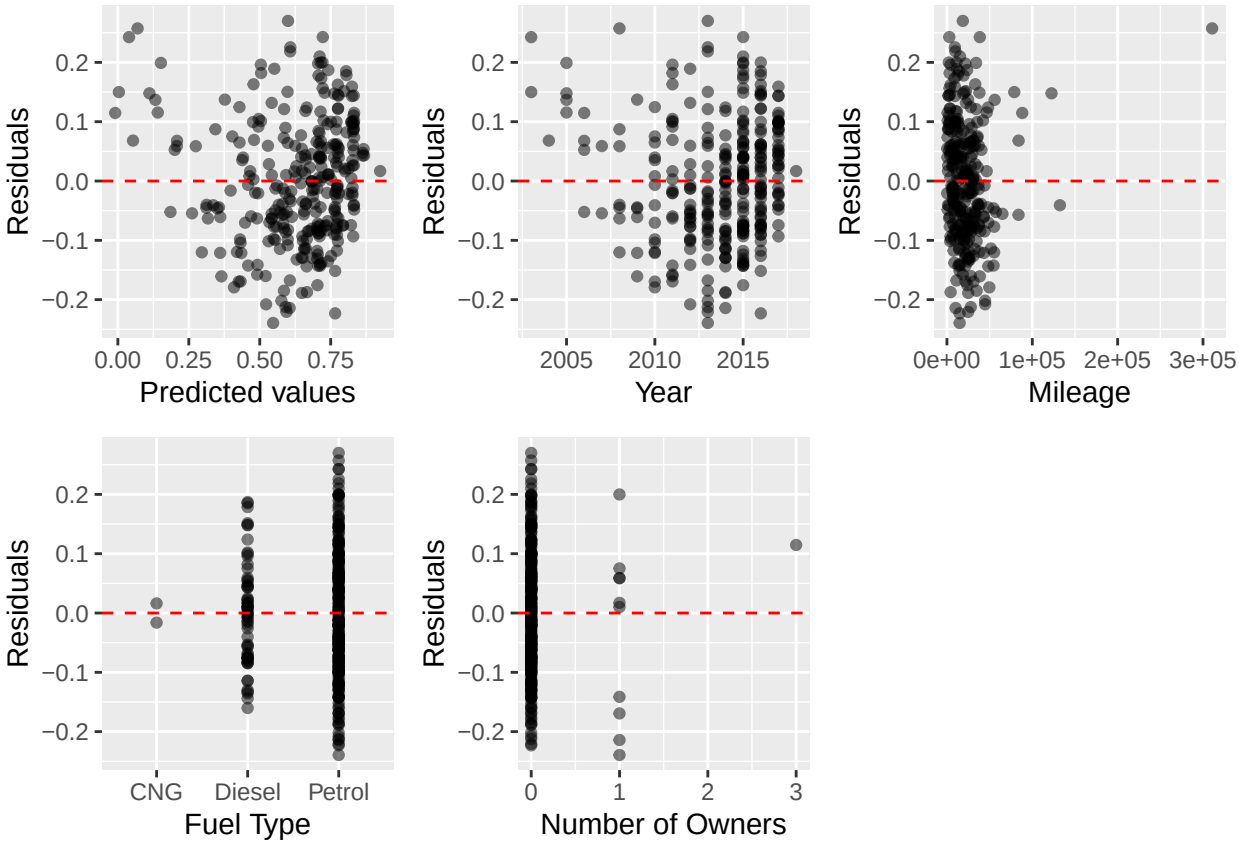
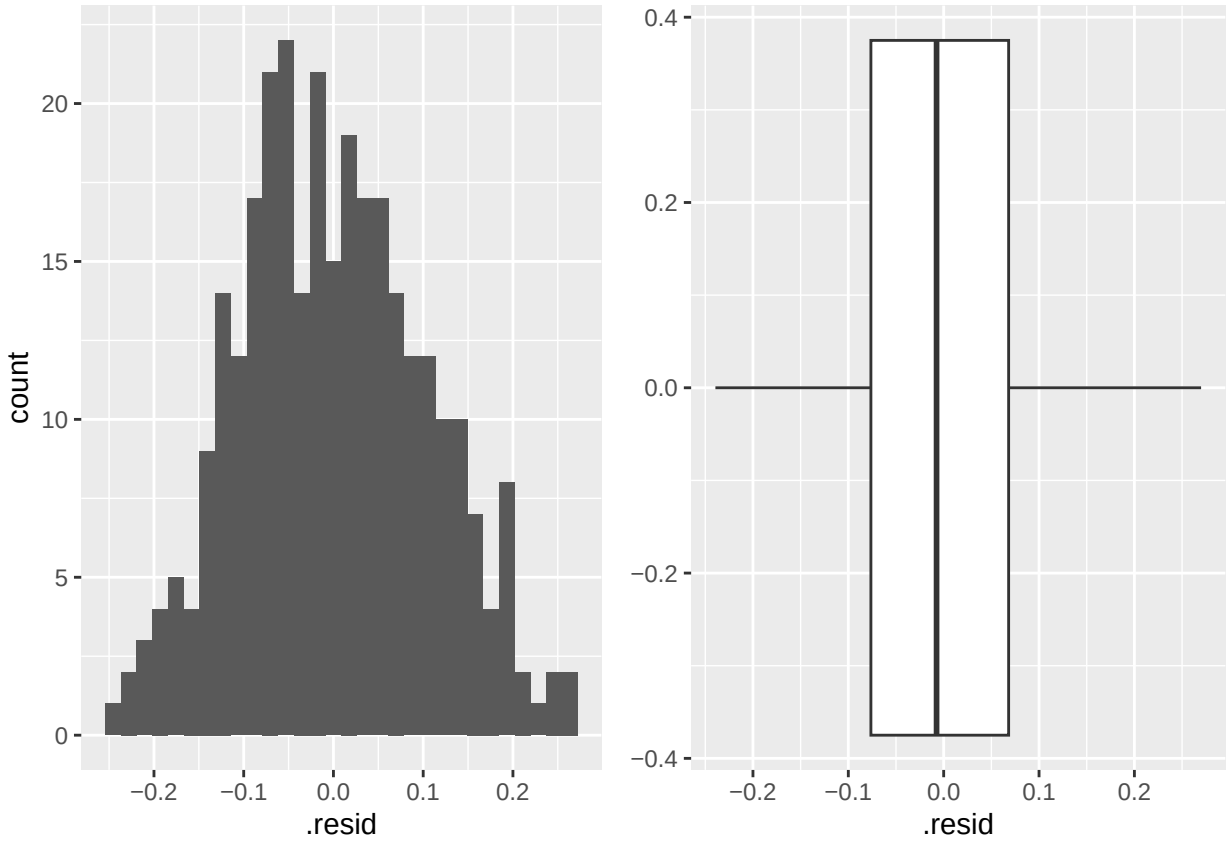
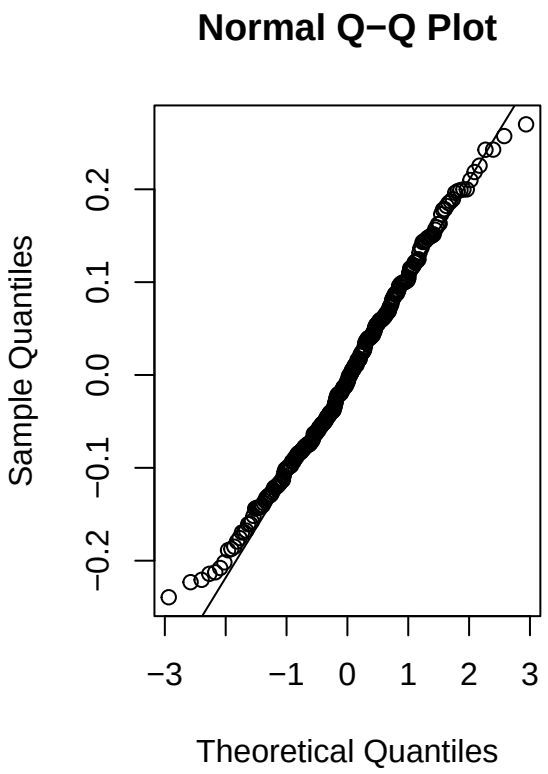


Figure 5: Residuals of Stepwise Model

The vertical spread of the residuals is nearly constant across the plot for the residuals of our model. Therefore, the constant variance condition is satisfied.

Normality






```
##
##  Shapiro-Wilk normality test
##
## data:  resid(step_model1)
## W = 0.99198, p-value = 0.1032
```

Histogram, qqplot and Shapiro-Wilk test (p-value=0.1032) all suggest that residuals appear to be normally distributed, therefore our data passes the normality assumption.

Independence

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1449279      1.701973   0.016
##  Alternative hypothesis: rho != 0
```

Since the p-value for Durbin-Watson test is less than 0.05, we can reject the null hypothesis and conclude that the residuals in this regression model are auto correlated. With the failed independence test for Y, the team concluded that extra data is needed/data needs to be cleaned and that the data is not representative of the population so the data can only be representative of this specific sample.

Recovering Functional Form

Based on the residuals of our stepwise regression model it seems that our model may require further transformation of our x values and the inclusion of some interaction terms in order to create a more accurate model.

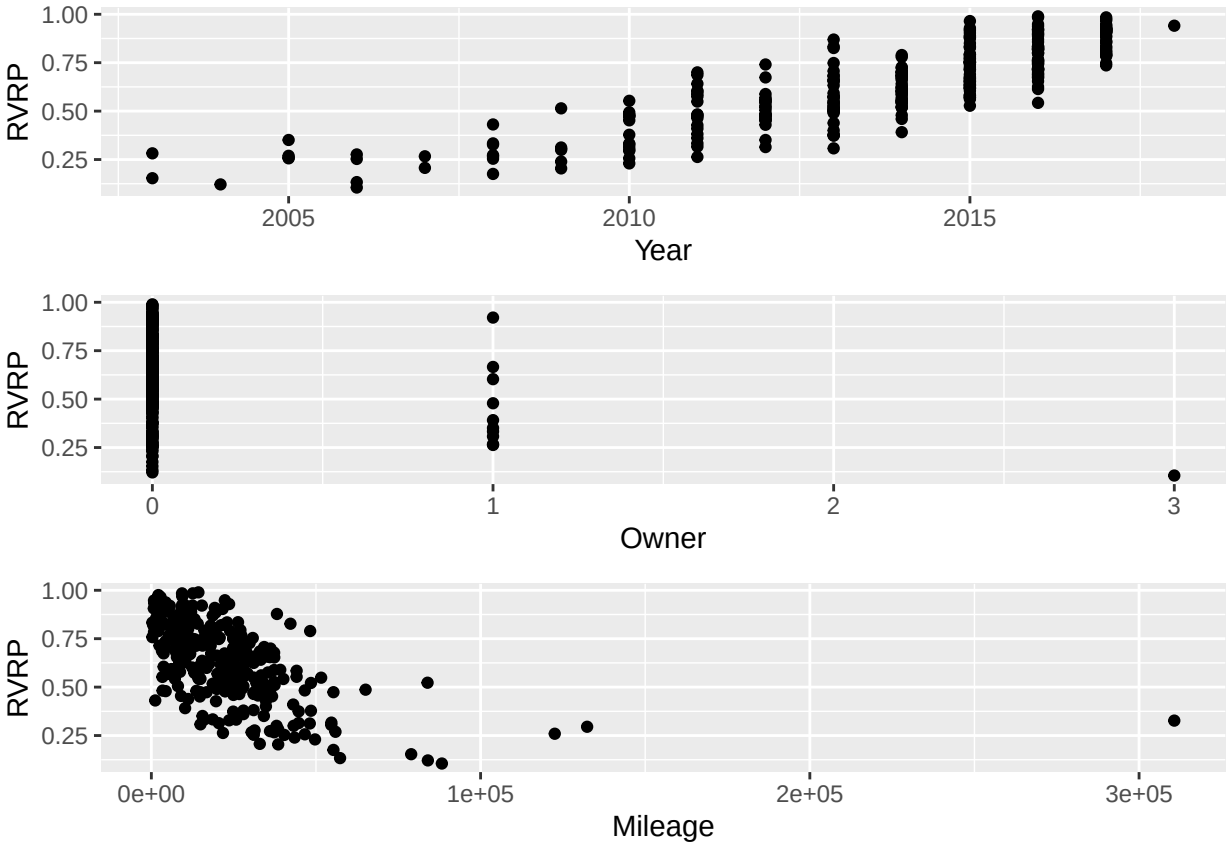


Figure 6: Finding Functional Form: Original Residuals

Year appears to have the most obvious relationship with Y (RVRP). If year is not linear then it appears to have a slight quadratic relationship with RVRP. Therefore, we will try adding a Year^2 term to our model. For building our model, we will refer to RVRP as Y, Year as X1, Owner as X6 and Mileage as X7.

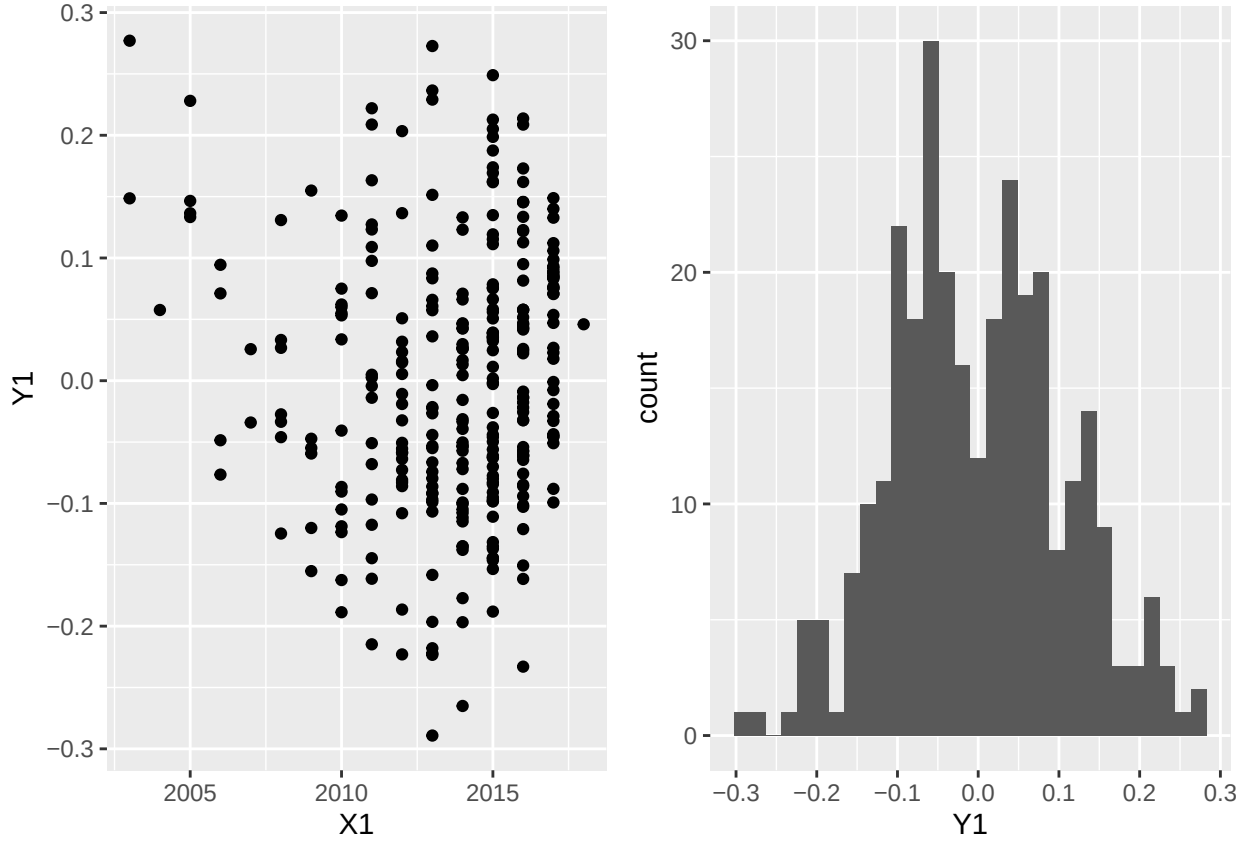


Figure 7: Residuals After Adding Year^2 Term

After adding X1^2 to our model, the distribution of errors now appears slightly bi modal. There still seems to be a slight pattern to the residuals, therefore we will need to add an additional term.

After adding an interaction between X1 and X7 the pattern in the residuals has gone away. Now we found the functional form of X1 and can move on to the other terms. Additionally the error now seems more normally distributed, which is another indication that we are headed in the right direction.

We are just left with X6 and X7, let us see how Y relates to X6 and X7 at different levels of fuel type.

There does not seem to be any interaction term between fuel type and X6 and X7, therefore the categorical variable Fuel_Type is not needed and will be dropped from our model.

We will add just the linear term for X6 as no other relationship would prove to be a better fit, and the residuals do appear more evenly spread out afterwards. The residuals appear to show a decreasing telescoping pattern, however that can likely be attributed to a lack of data points for cars with more owners. Lastly, our distribution of errors appears more normally distributed as well.

We will just keep the linear term for X7 as no other relationship would prove to be a better fit, and the residuals do appear more evenly spread out afterwards. The residuals appear to show a decreasing telescoping pattern, however that can likely be attributed to a lack of data points at higher mileages. The errors now seem to follow a normal distribution, which further affirms we have reached the correct functional form.

Therefore our final model takes the form Y = X1 + X6 + X7 + X1^2 + X1*X7.

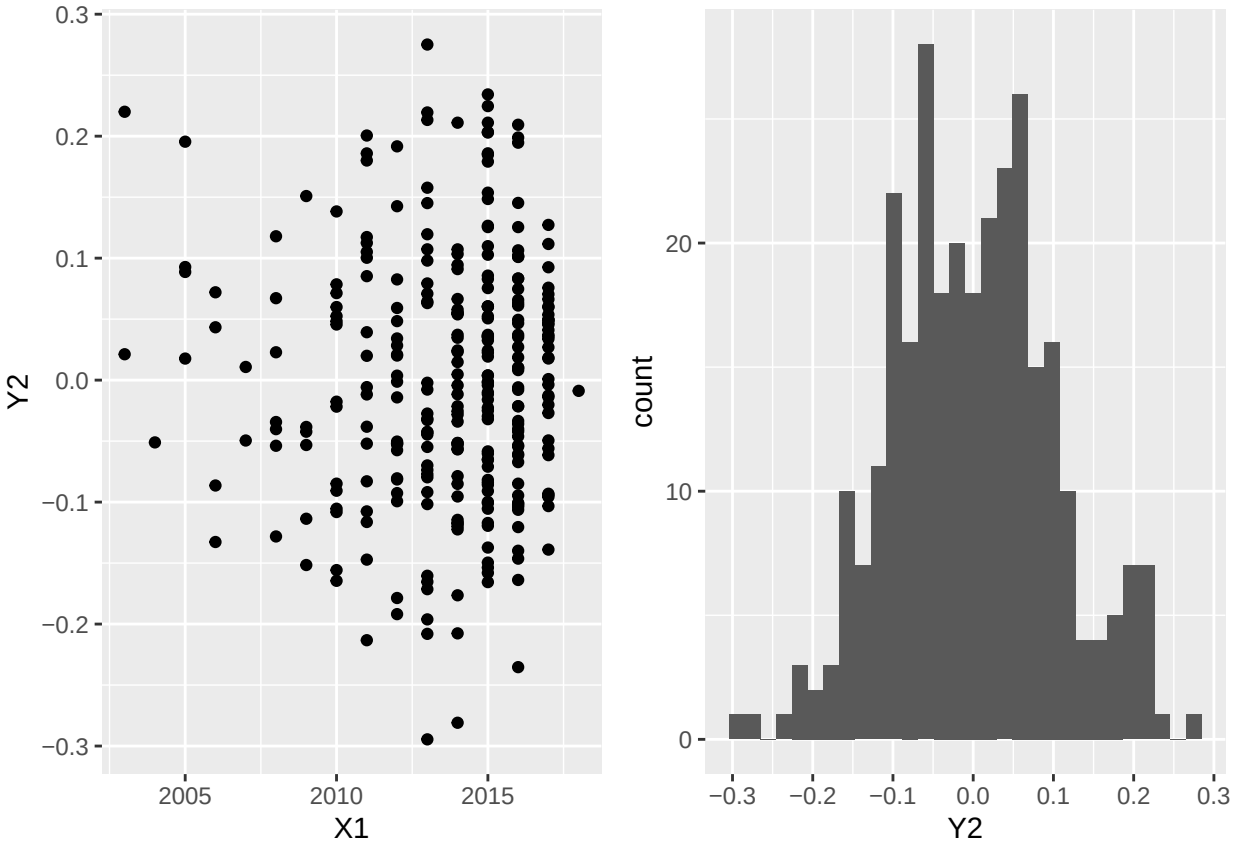


Figure 8: Residuals After Adding Year*Mileage Interaction Term

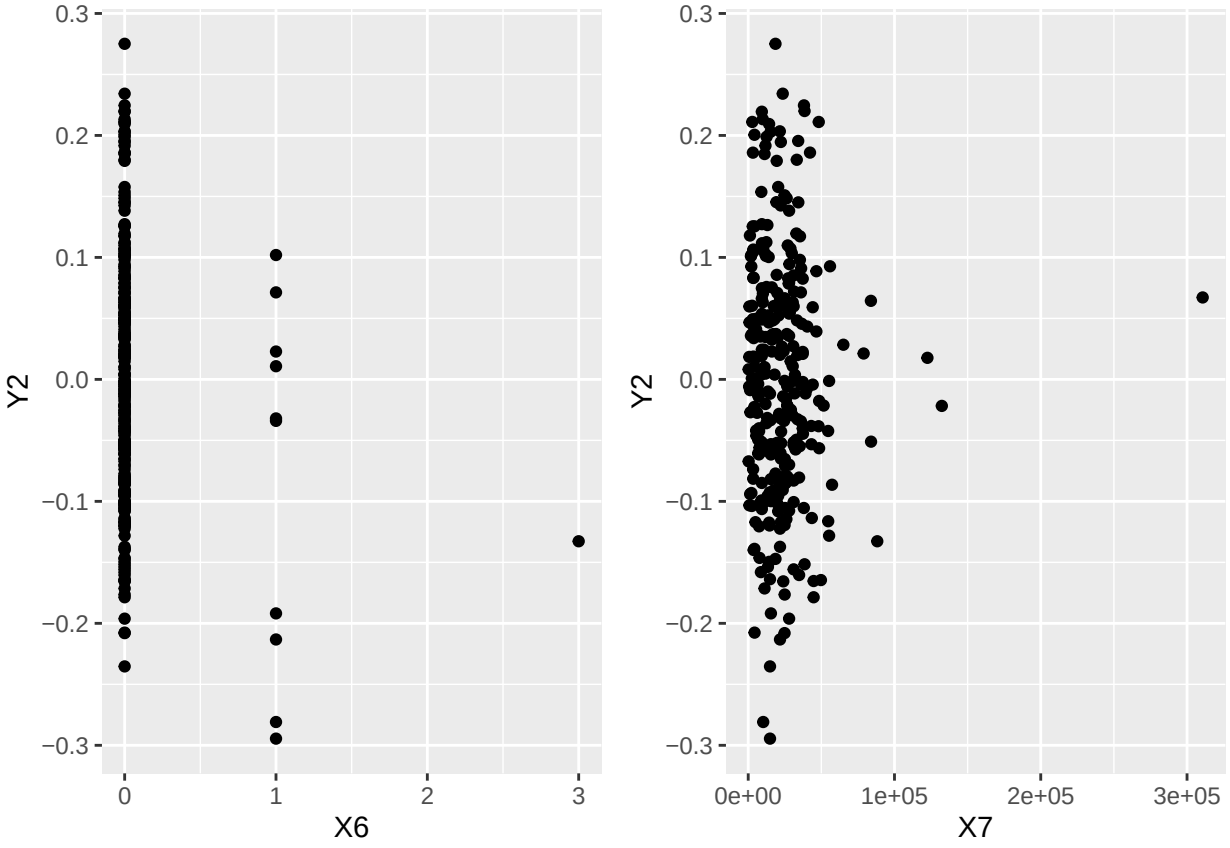


Figure 9: Residuals of Owner (X6) and Mileage (X7)

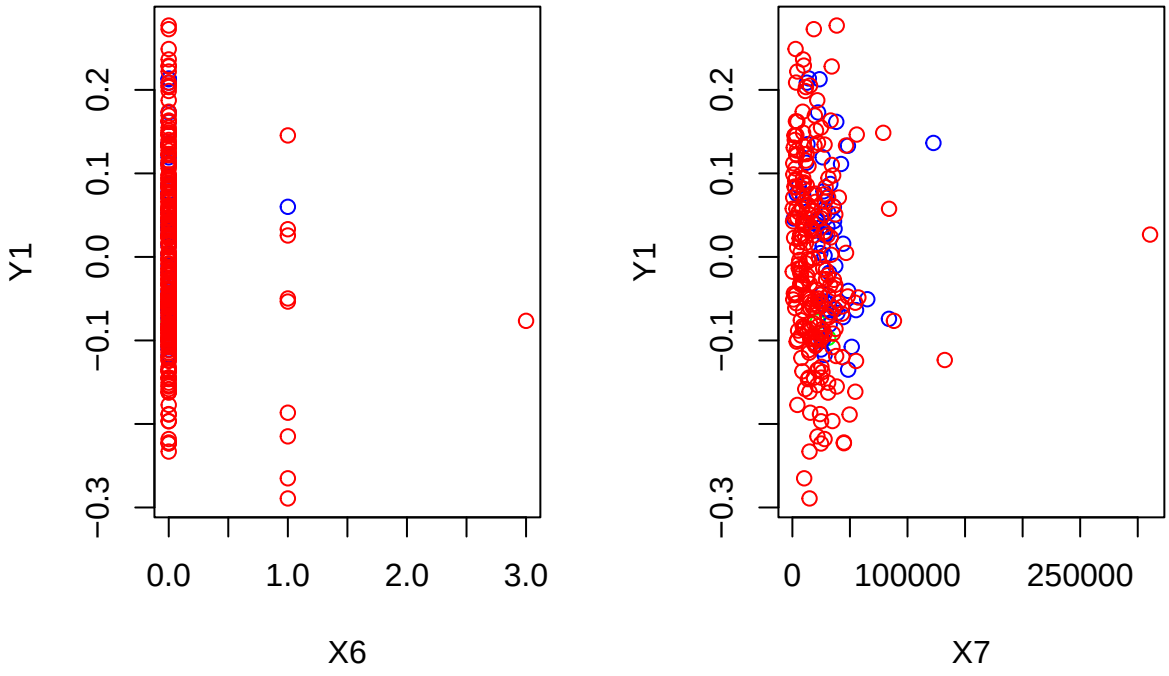


Figure 10: Residuals of Owner and Mileage at Different Levels of Fuel Type

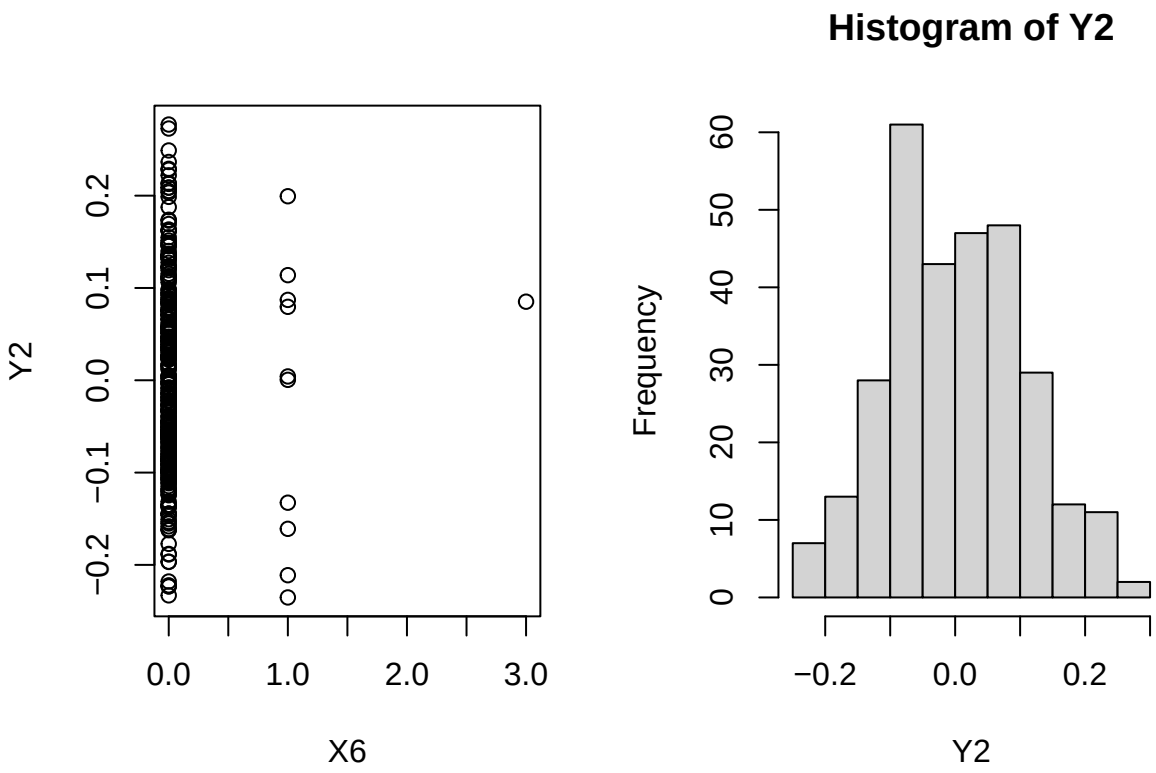


Figure 11: Residuals after Adding First Order Owner Term

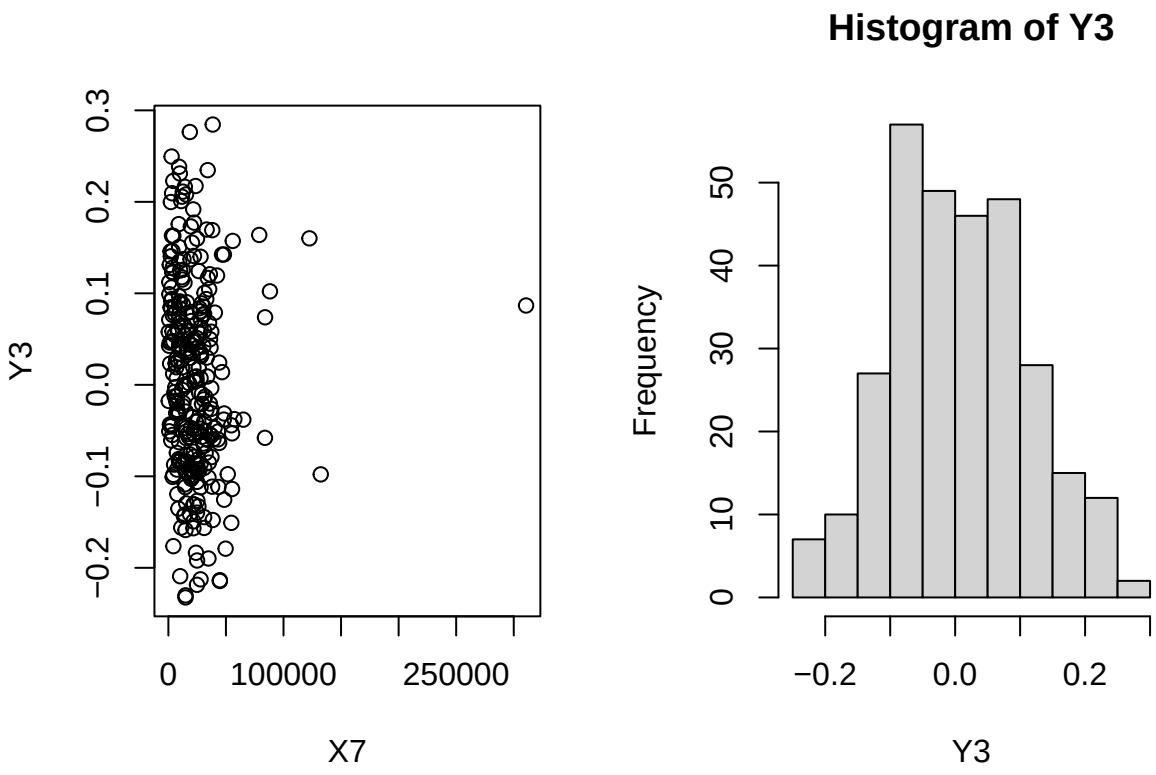


Figure 12: Residuals after Adding First Order Mileage Term

Checking for Multicollinearity in Final Model

Multicollinearity: Non Significant t-tests for all (or nearly all) independent variables when the overall F-test is significant

```
##
## Call:
## lm(formula = RVRP ~ Year + I(Year^2) + Owner + Mileage + Year *
##     Mileage, data = carData2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24448 -0.07259 -0.00185  0.05663  0.28676
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.386e+03  2.751e+03   2.685  0.00767 **
## Year         -7.402e+00  2.734e+00  -2.708  0.00717 **
## I(Year^2)     1.854e-03  6.790e-04   2.731  0.00670 **
## Owner        -6.773e-02  2.379e-02  -2.847  0.00473 **
## Mileage       4.307e-04  2.067e-04   2.084  0.03801 *
## Year:Mileage -2.147e-07  1.028e-07  -2.088  0.03769 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09993 on 295 degrees of freedom
## Multiple R-squared:  0.7602, Adjusted R-squared:  0.7562
## F-statistic: 187.1 on 5 and 295 DF,  p-value: < 2.2e-16
```

The F-test for H0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ is highly significant (F=187.1, p-value=2.2e-16) and all of our terms are also significant, so this does not seem to indicate a multicollinearity issue.

Multicollinearity: Opposite signs (from what is expected) in the estimated parameters

From data from a Progressive Car Insurance article we expect value retention, y, to increase when Year sold (X1) increases. Y will decrease when the number of owners (X6) or mileage (X7) increases. Therefore Progressive expects a positive relationship between y and X1, and a negative relationship with y and X6 and X7, however our model shows a negative relationship with X1 and a positive relationship with X7, while X6 is negative like progressive expects. The cause for this discrepancy may be due to the fact that there are interaction terms and second order terms included in our model. The interaction term between year and mileage (X1 and X7) is negative and the second order term for X1 is positive so these values likely balance out and lead to the correct interpretations when they are all taken into consideration.

Multicollinearity: Variance Inflation Factor (VIF) for a beta parameter greater than 10

```
##         Year    I(Year^2)        Owner      Mileage Year:Mileage
## 1.876976e+06 1.875016e+06 1.045181e+00 7.491926e+05 7.482309e+05
```

Year, Year^2, Mileage, and Year*Mileage all have vif scores greater than 10, however this is expected since they are all either interaction terms or a second order terms and therefore we expect them to be correlated to one another, so this is not an issue.

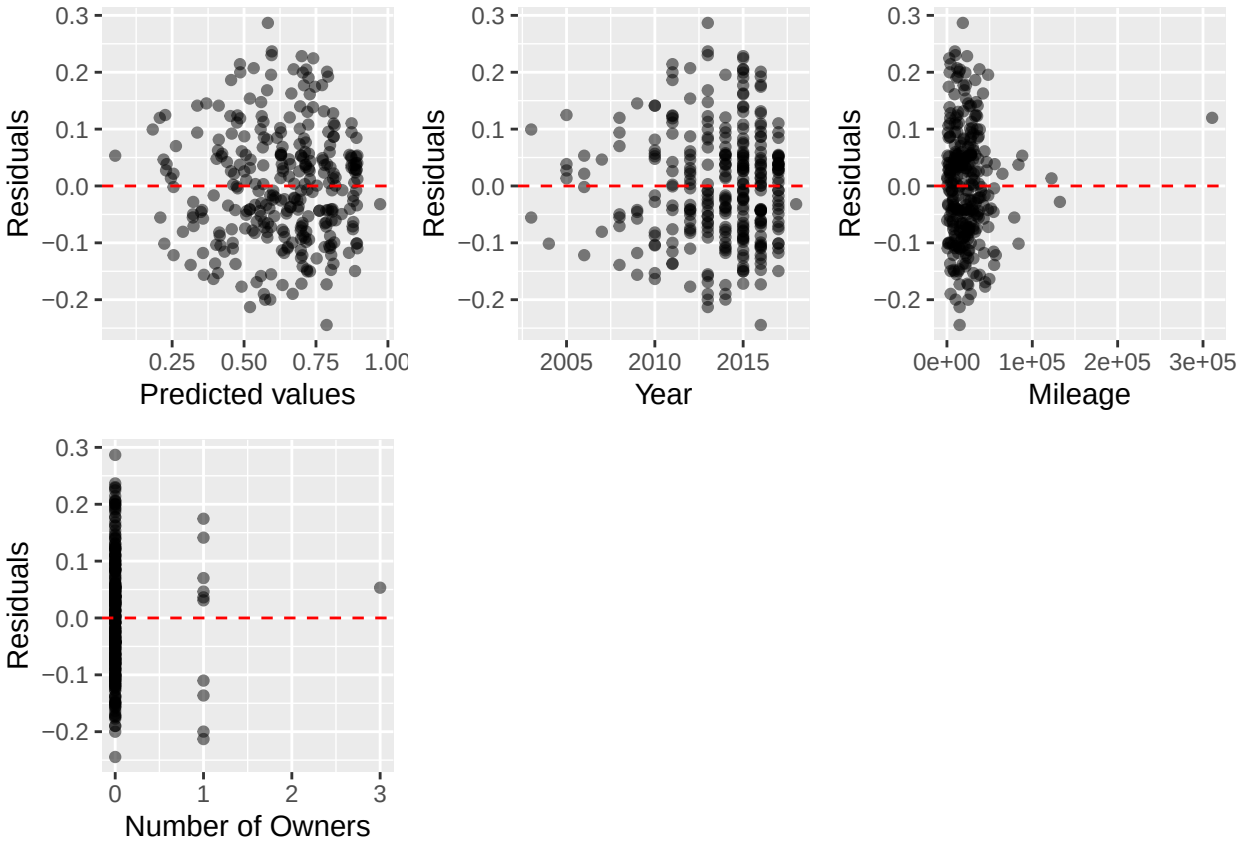


Figure 13: Residual Plots for Final Model

Checking Assumptions of Final Model

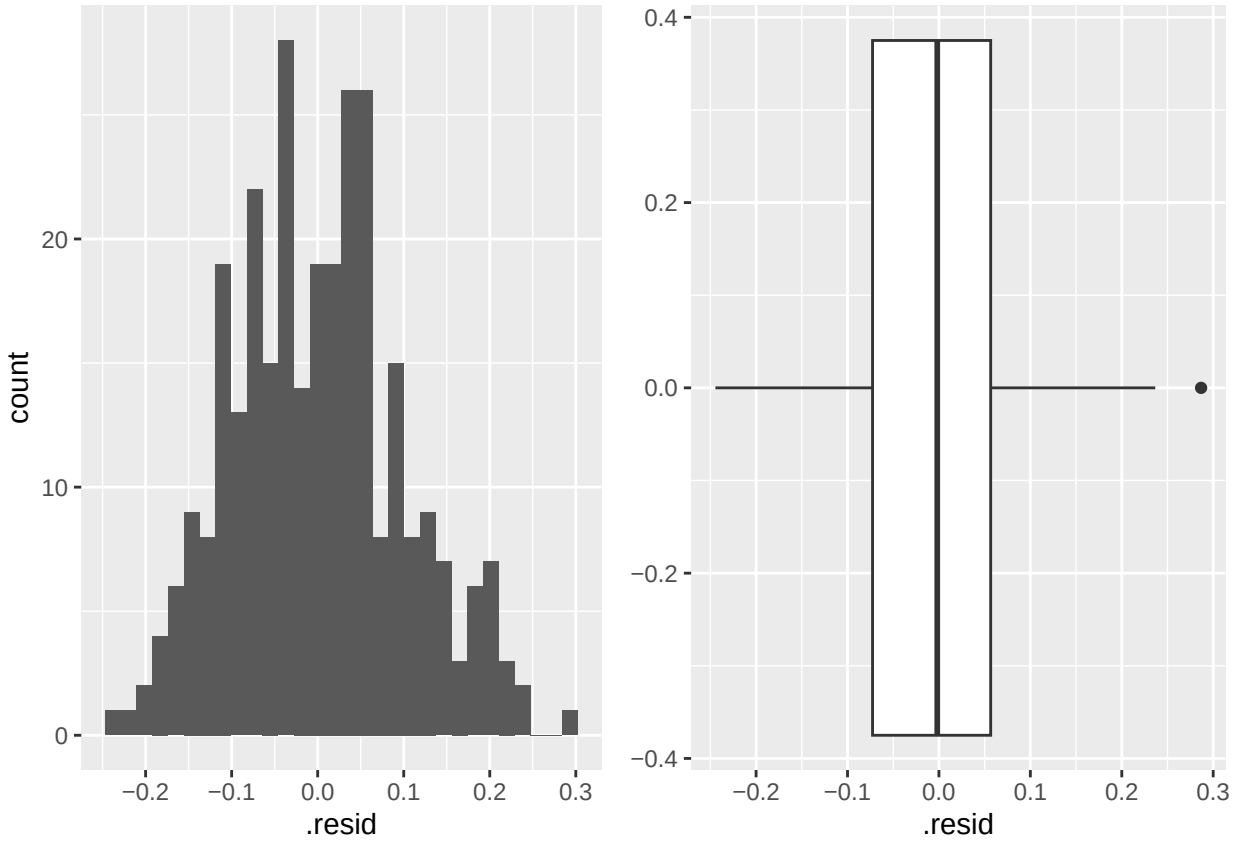
Linearity

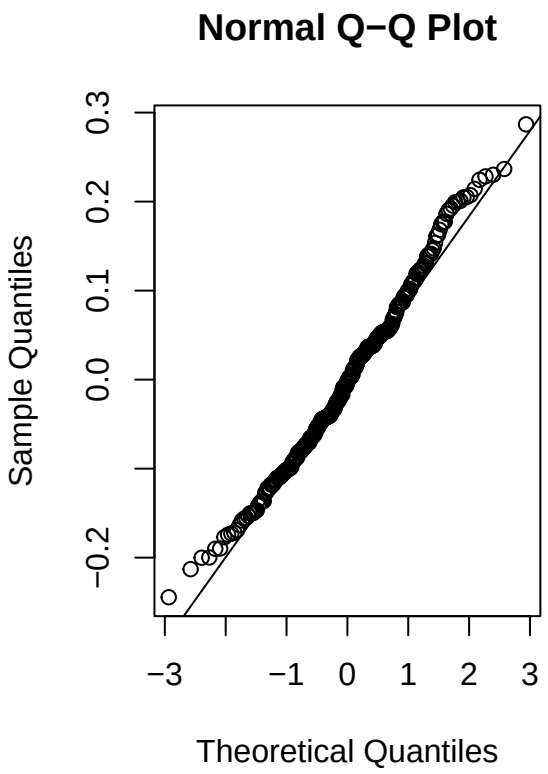
The plot of residuals vs. predicted does not show a pattern.
The plot of residuals vs. Year does not show a pattern.
The plot of residuals vs. Mileage does not show a pattern.
The linearity condition is satisfied for all significant explanatory variables.

Constant Variance

The vertical spread of the residuals is nearly constant across the plot for the residuals of our model. Therefore, the constant variance condition is satisfied.

Normality






```
##
##  Shapiro-Wilk normality test
##
## data:  resid(car.final.mlr)
## W = 0.99125, p-value = 0.07079
```

Histogram, qqplot and Shapiro-Wilk test (p-value=0.07079) all suggest that residuals appear to be normally distributed.

Independence

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1281001      1.738244    0.02
##  Alternative hypothesis: rho != 0
```

Since the p-value for Durbin-Watson test is less than 0.05, we can reject the null hypothesis and conclude that the residuals in this regression model are auto correlated. With the failed independence test for Y, the team concluded that extra data is needed/data needs to be cleaned and that the data is not representative of the population so the data can only be representative of this specific sample.

Alternate Model Considerations

```
##
## Call:
```

```
## lm(formula = RVRP ~ Year + I(Year^2) + Mileage + Year * Mileage,
##     data = carData2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.276461 -0.071245 -0.005743  0.058969  0.290588
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.673e+03  2.782e+03   2.758  0.00618 **
## Year         -7.687e+00  2.764e+00  -2.781  0.00577 **
## I(Year^2)     1.925e-03  6.866e-04   2.804  0.00538 **
## Mileage       3.752e-04  2.082e-04   1.802  0.07254 .
## Year:Mileage -1.871e-07  1.036e-07  -1.806  0.07200 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1011 on 296 degrees of freedom
## Multiple R-squared:  0.7536, Adjusted R-squared:  0.7503
## F-statistic: 226.4 on 4 and 296 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = RVRP ~ Year + I(Year^2) + Mileage, data = carData2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.268941 -0.073060 -0.004537  0.060662  0.295602
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.111e+04  2.035e+03   5.462 9.97e-08 ***
## Year        -1.110e+01  2.023e+00  -5.489 8.65e-08 ***
## I(Year^2)    2.774e-03  5.027e-04   5.517 7.50e-08 ***
## Mileage     -7.161e-07  2.848e-07  -2.514   0.0125 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1015 on 297 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7484
## F-statistic: 298.5 on 3 and 297 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = RVRP ~ Year + I(Year^2), data = carData2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.261314 -0.073190 -0.000901  0.061604  0.300558
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.107e+04  2.053e+03   5.391 1.43e-07 ***
## Year        -1.106e+01  2.041e+00  -5.420 1.23e-07 ***
```

```
## I(Year^2)    2.763e-03  5.072e-04   5.449 1.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1024 on 298 degrees of freedom
## Multiple R-squared:  0.7456, Adjusted R-squared:  0.7439
## F-statistic: 436.8 on 2 and 298 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: RVRP ~ Year + I(Year^2) + Owner + Mileage + Year * Mileage
## Model 2: RVRP ~ Year + I(Year^2) + Mileage + Year * Mileage
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    295 2.9458
## 2    296 3.0267 -1 -0.080921 8.1037 0.004727 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: RVRP ~ Year + I(Year^2) + Owner + Mileage + Year * Mileage
## Model 2: RVRP ~ Year + I(Year^2) + Mileage
##   Res.Df    RSS Df Sum of Sq     F   Pr(>F)
## 1    295 2.9458
## 2    297 3.0600 -2  -0.11426 5.721 0.003651 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: RVRP ~ Year + I(Year^2) + Owner + Mileage + Year * Mileage
## Model 2: RVRP ~ Year + I(Year^2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    295 2.9458
## 2    298 3.1252 -3  -0.17938 5.9878 0.0005664 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All Code for Report

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
library(readxl)
library(ggplot2)
library(gridExtra)
library(dplyr)
library(GGally)

carData = read_excel(file.choose())

# Conversion Indian Lakh to USD
carData$Present_Price = carData$Present_Price*1203.06
```

```
carData$Selling_Price = carData$Selling_Price*1203.06

# Convert KM to Miles
miles = carData$Kms_Driven*0.6214
carData$Mileage = miles

# Create Variable ValueRetentionPercent
Depreciation = carData$Present_Price - carData$Selling_Price
ValueRetentionPercent =  1 - (Depreciation / carData$Present_Price)
carData$RVRP = ValueRetentionPercent
# Plot Histogram of Response Variable
ggplot(carData, aes(x=RVRP)) + geom_histogram(binwidth=0.01)
# Plot eda histograms
h1<-ggplot(carData, aes(x=Year)) + geom_histogram(binwidth=1)
h2<-ggplot(carData, aes(x=Mileage)) + geom_histogram(binwidth=10000)
h3<-ggplot(carData, aes(x=Fuel_Type)) + geom_histogram(stat="count")
h4<-ggplot(carData, aes(x=Seller_Type)) + geom_histogram(stat="count")
h5<-ggplot(carData, aes(x=Transmission)) + geom_histogram(stat="count")
h6<-ggplot(carData, aes(x=Owner)) + geom_histogram(stat="count")
grid.arrange(h1,h2,h3,h4,h5,h6, ncol = 3)
# Plot eda predictor vs response
s1<-ggplot(carData, aes(x=Mileage, y=RVRP)) + geom_point()
s2<-ggplot(carData, aes(x=Year, y=RVRP)) + geom_point()
s3<-ggplot(carData, aes(x=Owner, y=RVRP)) + geom_point()
box1<-ggplot(carData, aes(x=Fuel_Type, y=RVRP)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=4)
box2<-ggplot(carData, aes(x=Seller_Type, y=RVRP)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=4)
box3<-ggplot(carData, aes(x=Transmission, y=RVRP)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=4)
grid.arrange(s1,s2,s3,box1,box2,box3, ncol = 3)


## Summary Statistics
sapply(carData[,c(2,3,4,5,9,10,11)], summary)
sapply(carData[,c(2,3,4,5,9,10,11)], sd)
car.final.mlr = lm(RVRP ~ Year + I(Year^2) + Owner + Mileage + Year*Mileage, data = carData)
summary(car.final.mlr)
# remove unused columns (Car_Name, Selling_Price, Present_Price, Kms_Driven)
carData2 <- carData[-c(1, 3, 4, 5)]

# create correlation matrix
ggpairs(carData2)+ theme(axis.text.y=element_text(size=10),
                         axis.text.x=element_text(angle=45, size=10),
                         strip.text.y=element_text(angle=0, hjust=0))
car.mlr = lm(RVRP ~ Year+Fuel_Type+Seller_Type+Transmission+Owner+Mileage, data = carData2)
summary(car.mlr)
library(car)
vif(car.mlr)
# Specify a null model with no predictors
```

```r
null_model <- lm(RVRP ~ 1, data = carData2)

# Specify the full model using all of the potential predictors
full_model <- lm(RVRP ~ ., data = carData2)

# Use a stepwise algorithm to build a parsimonious model
step_model1 <- step(null_model, scope = list(lower = null_model,
                                             upper = full_model),
                    direction = "both",test="F")
summary(step_model1)
linear_plot1 = ggplot(data =step_model1, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted values", y = "Residuals")

linear_plot2 = ggplot(data =step_model1, aes(x = Year, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Year", y = "Residuals")

linear_plot3 = ggplot(data =step_model1, aes(x = Mileage, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Mileage", y = "Residuals")

linear_plot4 = ggplot(data =step_model1, aes(x = Fuel_Type, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Fuel Type", y = "Residuals")

linear_plot5 = ggplot(data =step_model1, aes(x = Owner, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Number of Owners", y = "Residuals")


grid.arrange(linear_plot1, linear_plot2, linear_plot3, linear_plot4,
             linear_plot5,  ncol = 3)
par(mfrow = c(1, 2))
p1 = ggplot(data =step_model1, aes(x = .resid)) +geom_histogram()
p2 = ggplot(data =step_model1, aes(x = .resid)) +geom_boxplot()
grid.arrange(p1,p2, ncol = 2)
qqnorm(resid(step_model1))
qqline(resid(step_model1))
shapiro.test(resid(step_model1))
library(car)
dwt(step_model1)
# Recovering Functional Form

s1<-ggplot(carData, aes(x=Year, y=RVRP)) + geom_point()
s2<-ggplot(carData, aes(x=Owner, y=RVRP)) + geom_point()
s3<-ggplot(carData, aes(x=Mileage, y=RVRP)) + geom_point()
```

```
grid.arrange(s1,s2,s3, ncol = 1)
Y = carData$RVRP
X1 = carData$Year
X6 = carData$Owner
X7 = carData$Mileage

X1sq<-X1^2
lm1<-lm(Y~0+X1 + X1sq)
Y1<-Y-lm1$fitted.values
p1 = ggplot(carData, aes(x=X1, y=Y1)) + geom_point()
p2 = ggplot(carData, aes(x=Y1)) + geom_histogram()
grid.arrange(p1,p2, ncol = 2)

lm2<-lm(Y~0+X1 + X1sq + X1*X7)
Y2<-Y-lm2$fitted.values
p1 = ggplot(carData, aes(x=X1, y=Y2)) + geom_point()
p2 = ggplot(carData, aes(x=Y2)) + geom_histogram()
grid.arrange(p1,p2, ncol = 2)
p1 = ggplot(carData, aes(x=X6, y=Y2)) + geom_point()
p2 = ggplot(carData, aes(x=X7, y=Y2)) + geom_point()
grid.arrange(p1,p2, ncol = 2)
par(mfrow = c(1, 2))
plot(X6,Y1, type="n")
points(X6[carData$Fuel_Type == "CNG"],Y1[carData$Fuel_Type == "CNG"], col="green")
points(X6[carData$Fuel_Type == "Diesel"],Y1[carData$Fuel_Type == "Diesel"], col="blue")
points(X6[carData$Fuel_Type == "Petrol"],Y1[carData$Fuel_Type == "Petrol"], col="red")

plot(X7,Y1, type="n")
points(X7[carData$Fuel_Type == "CNG"],Y1[carData$Fuel_Type == "CNG"], col="green")
points(X7[carData$Fuel_Type == "Diesel"],Y1[carData$Fuel_Type == "Diesel"], col="blue")
points(X7[carData$Fuel_Type == "Petrol"],Y1[carData$Fuel_Type == "Petrol"], col="red")
par(mfrow = c(1, 2))
lm2<-lm(Y1~0+X6)
Y2<-Y1-lm2$fitted.values
plot(X6,Y2)
hist(Y2)
par(mfrow = c(1, 2))
lm3<-lm(Y2~0+X7)
Y3<-Y2-lm3$fitted.values
plot(X7,Y3)
hist(Y3)
# Check for collinearity issues
car.final.mlr = lm(RVRP ~ Year + I(Year^2) + Owner + Mileage + Year*Mileage, data = carData2)
summary(car.final.mlr)
library(car)
vif(car.final.mlr)
# Check for linearity
linear_plot1 = ggplot(data =car.final.mlr, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted values", y = "Residuals")

linear_plot2 = ggplot(data =car.final.mlr, aes(x = Year, y = .resid)) +
```

```
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Year", y = "Residuals")

linear_plot3 = ggplot(data =car.final.mlr, aes(x = Mileage, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Mileage", y = "Residuals")

linear_plot4 = ggplot(data =car.final.mlr, aes(x = Owner, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Number of Owners", y = "Residuals")

grid.arrange(linear_plot1, linear_plot2, linear_plot3, linear_plot4,  ncol = 3)
# normality testing
par(mfrow = c(1, 2))
p1 = ggplot(data =car.final.mlr, aes(x = .resid)) +geom_histogram()
p2 = ggplot(data =car.final.mlr, aes(x = .resid)) +geom_boxplot()
grid.arrange(p1,p2, ncol = 2)
qqnorm(resid(car.final.mlr))
qqline(resid(car.final.mlr))
shapiro.test(resid(car.final.mlr))
# independence testing
library(car)
dwt(car.final.mlr)
## Checking alternative models
car.final.mlr2 = lm(RVRP ~ Year + I(Year^2) + Mileage + Year*Mileage, data = carData2)
summary(car.final.mlr2)
car.final.mlr3 = lm(RVRP ~ Year + I(Year^2) + Mileage, data = carData2)
summary(car.final.mlr3)
car.final.mlr4 = lm(RVRP ~ Year + I(Year^2), data = carData2)
summary(car.final.mlr4)
anova(car.final.mlr, car.final.mlr2)
anova(car.final.mlr, car.final.mlr3)
anova(car.final.mlr, car.final.mlr4)
```