April 27, 2022

DETERMINING THE BEST CAR VALUE

USING REGRESSION TESTS

Prepared For :

Professor Bazan

MAS 202-B

By

Samara Alsalah

Matteo Constantine

Isabella Gurdian

Iva Iontcheva

Ines Ruiz

Table of Contents

I have actively participated in the preparation of this assignment and I now understand its solution.

Samara Alsalah

Matteo Constantine

Isabella Gurdian

Iva Iontcheva

Ines Ruiz

Cover Memo

This report analyzes the measures taken to statically determine which predictor variables are of utmost importance when valuing a car. The reported data includes 20 sedans, each with a value score determined by four predictor variables: price, cost per mile, road-test score, and predicted reliability. We began by creating a descriptive statistics table via Microsoft Excel, highlighting general statistics vital to understanding the framework of our data. Next, we used this data to create a table that calculates the relationship and extent of differences between two variables. Then, we estimated whether each predictor variable was significant and if it held influence over the value score through a calculated value. Another test was conducted to determine which predictor variables would stay within the model. The tests highlighted that both price and cost per mile have influence over the value score, were proven significant, and should remain in the model used to determine car value.

Project Introduction

Our report focuses on analyzing how to find the value of a car. The data used within our report comes from Consumer Reports, which uses value scores determined by five-year owner costs, cost per mile, overall road-test scores, and predicted reliability ratings. The five-year owner costs encapsulate expenses incurred within the first five years of ownership, including depreciation, fuel, maintenance, and repairs; average cost per mile is used to measure those expenses. The data collected from road-test scores are a collection of a little more than 50 evaluations and tests. Lastly, the predicted-reliability ratings are taken from the Consumer Reports Annual Auto Survey. The variables used within our report consist of 20 sedans with the price, cost per mile, road-test score, predicted reliability, and value score for each. Testing was conducted to determine which predictor variables have influence over the value score of a car. As seen in Exhibit 1, the average price of the cars is $26,886.20. The cheapest car is $21,800.00, and the most expensive car is $32,360.00. The sample variance is $11,385,793.54, while the standard deviation of the price is $3,374.28, which suggests that price values are spread out rather than concentrated around the average price. The average cost per mile is $0.64, the minimum cost per mile is $0.56, and the maximum cost per mile is $0.74. There is a sample variance of $0.00 for the cost per mile variable and a standard deviation of $0.06, which shows that cost-per-mile values are clustered around the average. The average road-test score is 80.450, with the minimum being 52.000 and the maximum being 93.000. The sample variance is 98.050, and the variable's standard deviation is 9.902, which shows that road-test scores values are also concentrated around the average. The predicted reliability variable's average is 3.750. It had a minimum of 3.000 and a maximum of 5.000, along with a sample variance of 0.408 and a standard deviation of 0.639, meaning that most values are close to the average. Lastly, the value score variable has an average of 1.458, a minimum of 1.050, a maximum of 1.750, a standard deviation of 0.197, and a sample variance of 0.039. These variables and measurements aided in the statistical analysis to determine the best car value.

Descriptive Statistics	Price	Cost/Mile	Road-Test Score	Predicted Reliability	Value Score
Mean	$ 26,886.20	$ 0.64	80.450	3.750	1.458
Standard Error	$ 754.51	$ 0.01	2.214	0.143	0.044
Median	$ 28,067.50	$ 0.67	82.000	4.000	1.425
Mode	#N/A	$ 0.67	81.000	4.000	1.730
Standard Deviation	$ 3,374.28	$ 0.06	9.902	0.639	0.197
Sample Variance	$ 11,385,793.54	$ 0.00	98.050	0.408	0.039
Range	$ 10,560.00	$ 0.18	41.000	2.000	0.700
Minimum	$ 21,800.00	$ 0.56	52.000	3.000	1.050
Maximum	$ 32,360.00	$ 0.74	93.000	5.000	1.750
Sum	$ 537,724.00	$ 12.84	1609.000	75.000	29.160

Exhibit 1

Managerial Report

Our report focuses on comparing the predictor variables of Price, Cost per Mile, Road-Test Score, and Predicted Reliability all to the Value Score. Comparing each of these values with the value score allows for a clear path to distinguish which variables were the most and least significant in determining the car value.

Price vs. Value Score

The first predictor variable analyzed was price and how it compared to the value score. To begin our analysis, we used a regression test to create an ANOVA table that would allow our team to see the statistical relationship between price and value score visually. The inputs used were the value score for the Y input range and price for the X input range. The ANOVA table gave us critical information to populate an estimated regression equation using the intercept coefficient (identified slope) and price coefficient (identified y-intercept). The intercept coefficient is 2.3587, and the price coefficient is -0.00003. This information is then used within the formula $y = mx + b$, where m is equal to the slope or intercept coefficient, 2.3587, and b is equal to the y-intercept or price coefficient, -0.00003. This results in the following estimated regression equation of $y = 2.3587 - 0.00003x$, Exhibit 2. This regression equation tells us that there is a negative correlation between price and value score. This means that when one variable increases, the other decreases, and vice versa, as shown by the negative slope of the trendline on the graph, Exhibit 3. To interpret, the value score decreases as price increases, and as price decreases, the value score increases. This regression equation determines the direct relationship between the predictor and outcome variables.

We next determined whether price was significant or not to the relationship. To measure this data, a significance level, or a threshold, was set to 0.05. This number is used to measure the strength of the evidence presented within the testing set, determining the predictor variable's significance. This means that if the number compared to the significance level is less than 0.05, the evidence would be statistically significant. To proceed, our group compared the p-value, 0.0083, from the ANOVA table to the significance level of 0.05 and found the p-value (a measure of statistical significance) to be less than 0.05, meaning the price variable is significant Exhibit 2. Moreover, the relationship between price and value score can not be entirely explained by chance or a random factor. Instead, a real relationship occurs between the predictor variable, price, and value score. When the relationship between price and the value score was compared with the relationships of the other predictor variables to the value score, price proved its relationship held greater significance.

Furthermore, our group looked at how the r-squared (measures relationship strength) value measured the predictor variable of price. The r-square determined that 32.8% of the value score is explained by the total variation of price, a percentage higher than some of the other predictor variables, Exhibit 2. This high percentage measures the strength of the relationship between the price and the value score, signaling a strong relationship compared to the other predictor variables.

To further our analysis of price compared to value score, our group determined whether this predictor variable should stay within the model used to predict a car's value. To perform this analysis, we used a t-test, further explained on page 14 of the report. To conduct a t-test, a t-estimate, -2.962, and t-critical, ± 2.101, is required, Exhibit 2. The t-critical is used to determine the range of which predictor variables are allowed to stay in the model. The t-estimate determines where the predictor variables lie within or outside that range. If the predictor variables land within the range, they will be rejected from the model, but if they land outside of the range, they will stay within the model. The price t-estimate concluded that price would remain within the model, meaning that price has a high level of significance in determining a car's value score.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.572
R Square	**0.328**
Adjusted R S	0.290
Standard Err(	0.166
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.242764665	0.24276467	8.7754135	0.008340605
Residual	18	0.497955335	0.02766419		
Total	19	0.74072			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	**2.3587**	0.306305707	7.70036542	4.2026E-07	
Price ($)	**-0.00003**	1.13084E-05	-2.9623324	**0.0083**	< 0.05 **significant**

Estimated Regression Equation	**T-Estimate**	**T-Critical**
y = 2.3587 - 0.00003x	**-2.962**	**± 2.101**

Exhibit 2

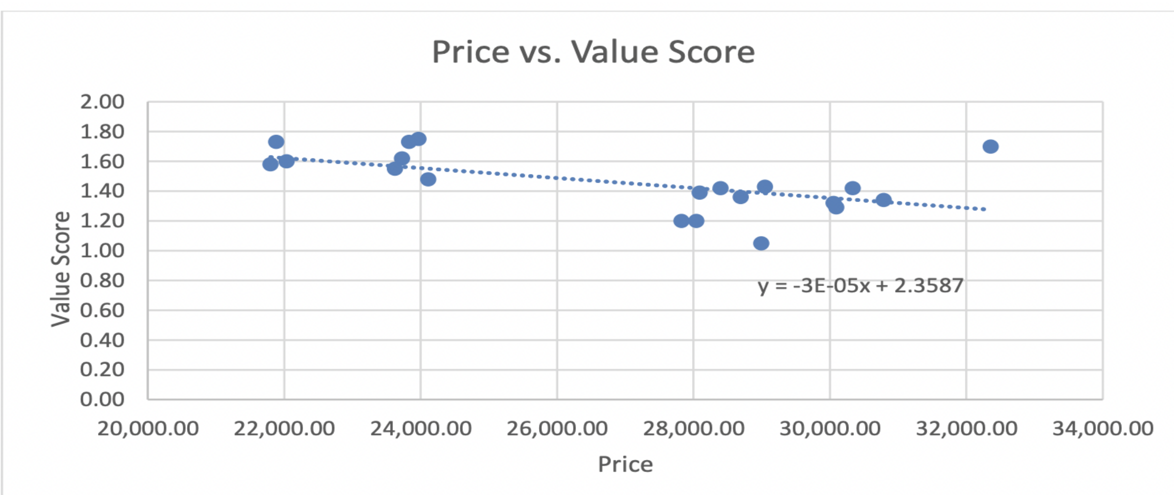

Exhibit 3

Cost per Mile vs. Value Score

The following variable our group analyzed is cost-per-mile and how it compares to the value score. Similarly to price, we used a regression test to find the ANOVA table to assist our analysis. The inputs used were the value score for the Y input range and cost per mile for the X input range. The ANOVA table calculated an intercept coefficient of 2.9422 and a price coefficient of -2.312, creating the regression equation of y = 2.9422x - 2.312, Exhibit 4. The regression equation shows a negative correlation between the value score and the cost-per-mile predictor variable, as shown by the negative slope of the trendline on the graph, Exhibit 5. This means that the higher the cost per mile is, the lower the value score is and vice versa.

Next, we determined whether the cost per mile was significant or not its relationship with the value score. The p-value of cost per mile is 0.0004, meaning that it is less than 0.05, Exhibit 4. The relationship between cost per mile and value score is significant, and the relationship is not determined by a random factor or by chance.

Our group then evaluated how r-squared determined whether the relationship between cost per mile and value score was strong or weak. The r-squared of cost per mile is 51.3%, the highest r-squared out of all the predictor variables. This high percentage indicates a strong relationship between the cost per mile and the value score. Thus, we can conclude that the cost per mile is arguably the most crucial factor in determining the value of a car.

To round out our analysis, we used a t-test to determine whether the cost per mile should stay within the factors used to predict a car's value model. The t-estimate calculated for cost per mile was -4.354, and the t-critical calculated was ± 2.101, Exhibit 4. Because the t-estimate falls outside of the t-critical range, cost per mile will be included in the model. Cost per mile is concluded to be a strong determinant of valuing a car.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.716
R Square	**0.513**
Adjusted R Square	0.486
Standard Error	0.142
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.380117216	0.38011722	18.97409058	0.000380856
Residual	18	0.360602784	0.02003349		
Total	19	0.74072			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	**2.9422**	0.342202258	8.59789414	8.65063E-08	
Cost/Mile	**-2.312**	0.530740722	-4.3559259	**0.0004**	< 0.05 **significant**

Estimated Regression Equation	**T-Estimate**	**T-Critical**
y = 2.9422 - 2.312x	**-4.354**	**± 2.101**

Exhibit 4

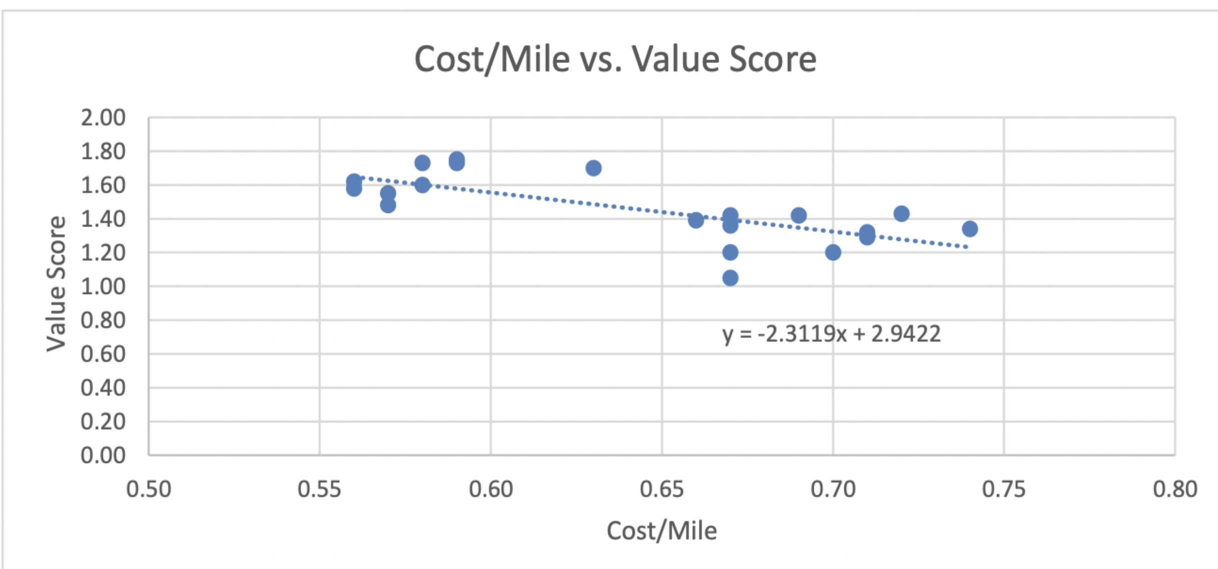

Exhibit 5

Road-Test vs. Value Score

The analysis for the road-test had a different outcome than those of price and cost per mile. A regression test was performed to create an ANOVA table. The Y input range was again the value score, and the X input range was the road-test scores. This gave us the intercept coefficient of 0.7978 and price coefficient of 0.008, resulting in a regression equation of $y = 0.7978 + 0.008x$, Exhibit 6. The regression equations highlight a positive relationship between the road-test score and the value of a car. As the road-test score increases, so will the car's value, as seen by the positive trendline in the graph, Exhibit 7.

Our group then determined his the road-test was significant or not to its relationship with the value score. The p-value of the road-test is 0.0714, meaning that it is greater than 0.05, Exhibit 6. The relationship between the road-test and the value score is not significant, and the relationship could be determined by a random factor or by chance.

Next, our group evaluated the r-square and whether it has a strong or weak relationship between the road-test and the value score. The r-squared of the road-test was 16.9% which was the lowest r-squared out of the price and cost per mile, suggesting that there is a weak relationship between the road-test and the value score, Exhibit 6.

The last determining matric used was the t-test to decide whether the road-test should stay within the car valuing model. The t-estimate calculated for the road-test was 2, and the t-critical calculated was ± 2.101, Exhibit 6. Because the t-estimate falls inside of the t-critical range, the road-test will not be included in the model. The road-test is concluded not to be a strong determinant of valuing a car.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.412
R Square	**0.169**
Adjusted R S	0.123
Standard Err	0.185
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.125458517	0.12545852	3.670396028	0.071409382
Residual	18	0.615261483	0.03418119		
Total	19	0.74072			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	**0.7978**	0.347073898	2.2986459	0.033720422	
Road-Test Sc	**0.008**	0.004283443	1.91582777	**0.0714**	> 0.05 **not significant**

Estimated Regression Equation	**T-Estimate**	**T-Critical**
y = 0.7978 + 0.008x	**2**	**± 2.101**

Exhibit 6

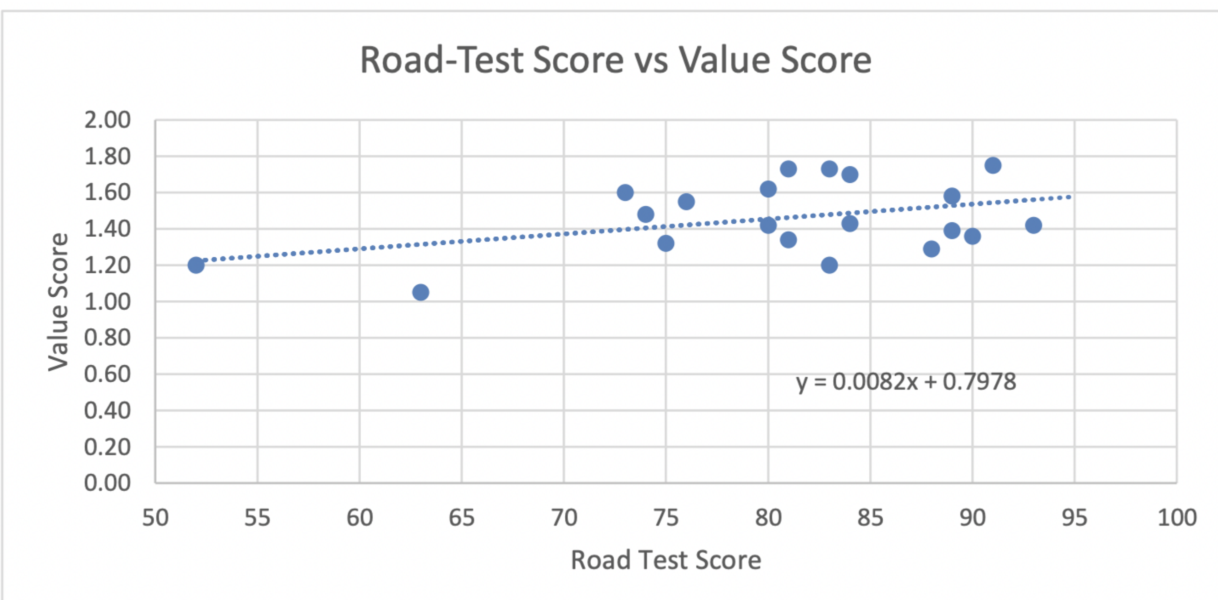

Exhibit 7

<u>Predicted-Reliability Ratings vs. Value Score</u>

The last predictor variable analyzed was predicted-reliability ratings to value score. Again the regression test was performed, which produced an ANOVA table from the Y inputs of the value score and the X inputs of the predicted-reliability ratings. The table showed an intercept coefficient of 1.0515 and a price coefficient of 0.108, creating a regression equation of $y = 1.0515 + 0.108x$, Exhibit 8. The regression equation shows a positive correlation between the predicted-reliability ratings and value score, as shown by the positive slope of the trendline on the graph, Exhibit 9. This means that the higher the predicted-reliability ratings are, the higher the value score is and vice versa.

The predicted-reliability rating was then tested to discover whether it was significant or not in its relationship with the value score. The p-value of the predicted-reliability ratings is 0.1296, which is increasingly more prominent than 0.05, Exhibit 8. The relationship between the predicted-reliability ratings and value score is insignificant, and the relationship could be determined by a random factor or by chance.

The r-squared value was analyzed to determine whether the relationship between predicted-reliability ratings and value score was strong or weak. The r-squared of the predicted-reliability ratings was 12.3%, the lowest r-squared value out of all predictor variables, Exhibit 8. This low percentage suggests a weak relationship between the predicted-reliability ratings and value score.

Lastly, the t-test was used to decide whether the predicted-reliability ratings should stay within the model that values cars. To conduct the t-test, the t-estimate of 1.588 and the t-critical of $\pm$ 2.101 were used, Exhibit 8. Because the t-estimate falls inside of the t-critical range, the predicted-reliability ratings will not be included in the model. The predicted-reliability ratings are not a strong or essential metric for valuing a car.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.351
R Square	**0.123**
Adjusted R Square	0.074
Standard Error	0.190
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.091045161	0.09104516	2.522512503	0.12964188
Residual	18	0.649674839	0.03609305		
Total	19	0.74072			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	**1.0515**	0.259414856	4.05353958	0.000745447	
Predicted Reliability	**0.108**	0.068243441	1.58824195	**0.1296**	>0.05 **not significant**

Estimated Regression Equation	**T-Estimate**	**T-Critical**
y = 1.0515 + 0.108x	**1.588**	**± 2.101**

Exhibit 8

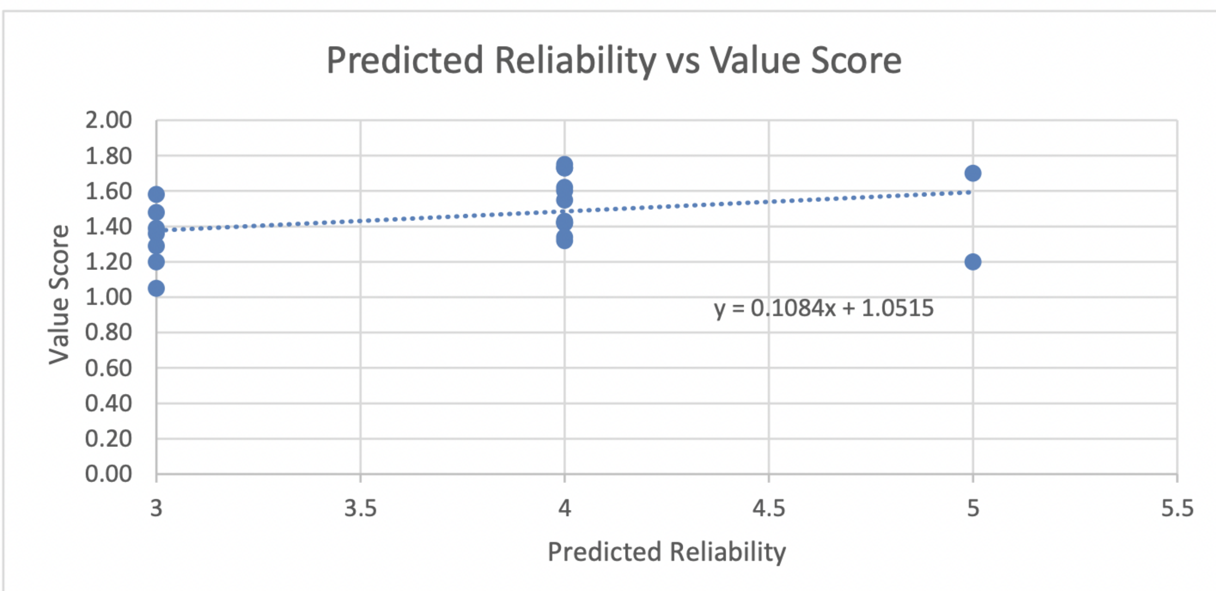

Exhibit 9

T-Test

A t-test was used to distinguish which of the four variables, price, cost per mile, road-test, and predicted reliability, should stay in the model to determine a value score. To conduct this test, we must find what is known as a t-critical value. This requires two things: a threshold value to determine whether or not a statistic is significant; this threshold value is known as alpha (α). We set this value of $\alpha = .05$. However, since this t-test is two-sided, we will be dividing that value by two; thus, $\alpha/2 = .025$. Now that we have a threshold significance value, we need to find the degrees of freedom (d.f.), equal to the number of values that have the freedom to vary in the sample data. In this case, the d.f. is equal to 18. With $\alpha/2 = .025$ and d.f. = 18, we can go to the t-table and see that the t-critical value equals 2.101. Since this test is two-sided, if one of four variables' estimated t-values is less than -2.101 or greater than 2.101, then that variable must stay in the model. But, if one of the four variables' estimated t-values is between -2.101 and 2.101, then that variable must leave the model. In this case, price (-2.962) and cost per mile (-4.354) both had values less than -2.101, which means that they both will stay in the model. On the other hand, the road-test score (2.000) and predicted reliability (1.588) both had values between -2.101 and 2.101, which means that they must leave the model. Staying in the model tells us, statisticians, that the correct regression equations to be used are the ones for price and cost per mile.

Conclusion and Recommendation

We can see from our analysis of the data how the price, cost per mile, road-test score, and predicted reliability scores are related to the value score. Out of those four predictor's, the price and cost per mile were the two that have influence over the value score and the two that must stay in the model. This is due to the alignment of both their P-values being less than 0.05 and their t-estimates being less than -2.101. To provide further evidence for this conclusion, the price and cost per mile both had significantly greater R-squared percentages at 32.8% and 51.3% when compared to the road-test score and predicted reliability at 16.9% and 12.3%. The greater the R-squared percentage, the more the variability in the value score is explained and the more fit the data is for the model. Based on the evidence, our final recommendation is to build a new model that uses both price and cost per mile to predict the value score.