Spring 2020 | BUAN6337.002 | HW 2

Predictive Analytics using SAS

Group2

Adam Butcher

Dong In Kim

Maria Phetteplace

Mahmoud Ali

Q1. '93car' data

Q1-a. Correlation between **horsepower** and **midrange price**

Correlation	0.78822
p-value	<.0001
Conclusion	
There is a significant correlation between **horsepower** and **midrange price.** It is significantly different from zero. Because the p-value is close to 0, we would reject the null hypothesis of the variables not being correlated.	

Q1-b. Regression model on midrange price

Regression Model: (Air_Bag_Standard = 0 is the base)
Midrange_price = 12.98 – 0.28 City_MPG + 3.54 Air_bags_Driver + 6.74 Air_Bag_Drive_Passager + 0.10 horsepower – 2.52 Manual_transmission – 4.87 Domestic

Q1-c. Regression analysis

1. It is a good fit because the R-sq is 0.73 meaning that the variables explain 73% of the variance in Midrange_range
2. R-sq Values

R-sq	Adjusted R-sq	Why Adjusted R-sq
0.7256 0.72 implies that all the 5 explanatory variables explain around 73% of the variance in the dependent variable, midrange_price.	**0.7065** When factoring in a penalty on any variable added to the model that has a very small explanatory power, it went down only slightly.	Adjusted R-sq has been adjusted based on the number of predictors in the model. Since Adjusted R^2 could actually go down if we add more variables to the model, it is more reliable and accurate in determining the efficiency of the model.

3. Significant Variables

Significant Variables	p-value
City_MPG	0.0466 (significant)
Air_Bag_Driver	0.0069 (highly significant)
Air_Bag_Driver_Passager	0.0003 (highly significant)
Horsepower	0.0001 (highly significant)
Domestic	0.0001 highly significant)
Manual_transmission_available	.0714 (only significant at the 10% level)

4. Interpretations of 'Horsepower' and for 'Domestic'

Coefficients	Interpretation
0.09846	For every 1 unit increase in horsepower, there is a $98.46 increase in midrange priced cars.
-4.87087	Midranged priced cars that are domestically made have are priced $4870.87 lower than non-domestically made cars.

5. Importance using STB

Most important	Why?
Horsepower	After running the model with standardized betas, horsepower had an STB of 0.53 which is the highest absolute value of all the variables in the model.

6. Elasticity

Approach	To compute average estimate of price elasticity, multiply the horsepower coefficient with the average horsepower and divide by the average midrange_price.
Computation	0.09846 * 143.828 / 19.510 = 0.7259
Result	The elasticity of midrange price with respect to horsepower is 0.7259.

7. Non-linear Effect

New model	Run a new model with the same variables but add a new variable ‘HP^2’
p-value for hp^2	0.0599
Conclusion	
When checking whether horsepower has a non-linear effect on midrange price, it is NOT significantly different from zero at the 5% level.	

8. Interaction Variables

New model	Run a new model with the same variables but add a new variable ‘HP * Weight’
p-value for hp*weight	0.0358
Conclusion	
Yes, there is an interaction because it significantly different from zero. However, the coefficient is 0.0000 so it doesn’t actually change the price.	

Q1-d. Updated Regression Model

New Regression	Price = 6.95 – 0.37 City_MPG + 3.87 Air_Bag_Driver +6.78 Air_Bag_Driver_Passenger + 0.11 Horsepower – 3.34 Manual_Transmission – 4.19 Domestic + 0.00 Revolutions_per_mile
New R-sq	0.734
Adjusted R-sq	0.712
Significant Variables	p-value
City_MPG	0.0148 (significant)
Air_Bag_Driver	0.0034 (highly significant)
Air_Bag_Driver_Passager	0.0002 (highly significant)
Horsepower	0.0001 (highly significant)
Domestic	0.0016 (highly significant)
Manual_transmission_available	0.0245 (significant)

Engine_Revolutions_per_mile	0.1073 (not significant)
Conclusion	The adjusted R-sq raised by adding Revolutions_per_mile although the variable is not significant according to the p-value.

Q2. ‘Diamond’ data

Q2-1. Relationship between cut and clarity

H_0	There is no relationship between cut and clarity	
H_A	There is a relationship between cut and clarity	
Result	Reject	Conclusion
Chi-sq: .0041	DO NOT Reject the H_0	Although the chi-sq is in the rejection region, there are cells that have a count less than 5. Because of this we should not reject the hypothesis.

Q2-2. Difference of D and E

H_0	There is no significant difference in price between D and E color.	
H_A	There is a significant difference in price between D and E color.	
Result	Reject	Conclusion
t-value: 18.92 p-value: <.0001	Reject the H_0	Since we can reject the result of Equality of Variance, we used the satterwaite unequal coefficient whose t-value is 18.92. Their variances are not equal. There is a significant difference in price between D and E color.

Q2-3. Regression of price (Base for dummy variables are ‘E,’ ‘Very Good,’ and ‘VVS1’)

Regression Model:
Price = -4747.5175 + 16067 Carat + -335.4792 Good – 572.1877 Fair + 266.5103 Ideal - 141.9852 D – 1420.4205 VVS2 – 3781.2142 VS1 – 4240.9402 VS2

Q2-3-a. Significance of Color

Significance Dummy Variable, E	Conclusion
p-value: 0.3206	Because the p-value is above 0.05, there is NO significant difference in price between color 'D' and 'E.'

Q2-3-b. Ideal versus Good

Ideal	A diamond that is an ideal cut would raise the price by $266.51 when compared to a diamond that is not an ideal cut.
Good	A diamond that is a good cut would lower the price by $355.48 when compared to a diamond that is not a good cut.
Conclusion	There is a $621.99 difference between an ideal and good cut diamond.

Q2-3-c. VVS2 versus VS1

VVS2	A diamond that has a clarity rating of 'VVS2' would lower the price by $1,420.42 when compared to a diamond that has a clarity rating that is not 'VVS2'.
VS1	A diamond that is a good cut would lower the price by $3,781.21 when compared to a diamond that is not a good cut.
Conclusion	There is a $2,360.79 difference between a clarity rating of VVS2 and VS1.

Q2-3-d. Variable Significance

Significant Variables	p-value
carat	<0.0001
Fair	0.0002
Good	0.0002
Ideal	0.0100
VVS2	<0.0001
VS1	<0.0001
VS2	<0.0001

Q2-3-e.

R-sq	Adjusted R-sq	Conclusion
.9359	.9338	This is a very good fit. 93% of the variance in price is explained by independent variables. Since we are comparing models with different numbers of predictors, looking at the predicted r-sq is fine.