A black background with blue text

Description automatically generated

**Course Number: MMA 860**

**Course Name: Acquisition and Management of Data**

**Assignment Name: Assignment 2**

**Due Date: July 19, 2024 11:59 pm**

**Team Name: Team Gordon**

|  |  |
| --- | --- |
| **Student Name** | **Student Number** |
| Alisha Sahota | 20497348 |
| Anthony Ramelo | 20499391 |
| Chris Wu | 10182394 |
| Elizabeth Zhang | 20161231 |
| Emily Zhao | 10096273 |
| Sam Hossain | 20466500 |

## Order of files:

|  |  |  |
| --- | --- | --- |
| **Filename** | **Pages** | **Comments and/or Instructions** |
| MMA 860 Assignment 2 | 15 |  |

**Additional Comments:**

|  |
| --- |
|  |

# Question 1:

The p-value for the joint F-test evaluates whether at least one of the variables in the model is statistically significant. If the p-value is greater than 0.05, it suggests that some of the variables in the model may not be significant for prediction. This does not mean we should dismiss all variables in the model. Instead, we should look at each variable to determine the significance of individual predictors.

## b)

Outlier does impact the analysis, but it also provides important information about the data distribution. Simply replacing it with the mean can lead to misleading conclusions. Depending on the type of outliers, different methods of managing outliers should be considered:

* Error in recording: remove
* Out of Scope extreme value: delete
* In Scope extreme value: retain

## c)

The choice of alpha should not be arbitrary, it should be context-specific. Alpha = 0.05 is commonly used but it may not be suitable for all contexts. Choosing alpha should depend on the balance between Type I errors (false positives) and Type II errors (false negatives), as well as the consequences of each in the specific context. In scenarios where the cost of a Type I error is high, a smaller alpha might be appropriate. Another way of choosing alpha is to consider the cost of failure, if the cost of failure is high, alpha should be small.

## d)

Heteroskedasticity refers to the non-constant error term in regression analysis. It makes the estimates less efficient and invalidates standard inferences, but it does not invalidate the results. There are several methods to address the issues without scrapping the entire analysis:

* GLS/Feasible GLS: provides more efficient and unbiased estimates
* HCCME: corrects the inferences and covariances matrix in the regression

# Question 2:

## a)

Multiple imputations are used to handle missing data in a dataset. Instead of guessing a single value to replace each missing value, multiple imputation generates a range of possible values, creating several complete datasets. Each of these datasets is then analyzed and the average is used as if there were no missing data. Finally, the results from these analyses are combined to produce a single result. The process of the multiple imputations is like filling in a blank puzzle piece several times, using clues from the surrounding pieces to guess what might fit. Instead of guessing once, we make several educated guesses, creating different complete puzzles. We then solve each puzzle as if it were complete. In the end, we combine the solutions from all the puzzles to get a final answer that considers different possibilities for the missing pieces.

## b)

Multiple imputation can be used when:

* MAR (Missing at Random)
* Large Sample Size

Example:

An institute is conducting a study on patient outcomes after receiving a treatment, involving thousands of participants from different areas and backgrounds. The dataset includes age, treatment dose, side effects, and recovery time, but recovery time is occasionally missing.

* Recovery time is more likely to be missing for patients who experienced severe side effects. The data can be considered MAR. This allows for the use of multiple imputations. We can impute missing recovery times based on observed severities of side effects along with other variables such as age and treatment dose.
* The study has a very large number of participants, ensuring that there is enough data to accurately model the relationships between the observed variables and the missing recovery times.

## c)

We can substitute the missing value with the Mean of the available recovery times. However, this approach will reduce the variability of the data, skew the distribution and lead to biased estimates.

## d)

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

## e)

X2 does belong in the model.

P-value = 0.0395. (Let’s use alpha = 0.05).

P-value < alpha, therefore X2 is significant.

# Question 3:

## a)

Rating = B0 + B1Price + B2Alcohol + B3Sulphates + B4CountryFrance

Rating = -4.6623 + 1.025Price + 3.2948Alcohol – 12.999Sulphates + 9.0108CountryFrance

A screenshot of a computer

Description automatically generated

Development Steps:

1. Data Exploration: Checked the structure and completeness of the data. There’s no missing data.
2. Preprocessing: Created a dummy variable 'CountryFrance' to handle Country categorical data.
3. Model Building:
   * + Constructed a linear regression model between the predictors and the target variable ('Rating ~ Price + Alcohol + Residual\_Sugar + Sulphates + pH + CountryFrance')
     + T-tests were used for each variable in the model. Since ‘Residual\_Sugar' and 'pH' has high p-value, those two variables does not belong to the model and was removed from the model.
4. Model Training and Assessing:
   * + Used train\_test\_split to divide the data into training and testing sets for model validation.
     + Trained the model on the training data.
     + Used the trained model to predict the ratings for the test dataset, evaluating the model against new data.
     + Validate the model accuracy by calculating R2, RMSE and MAE.

|  |  |  |
| --- | --- | --- |
|  | Training Data Results | Testing Data Results |
| R2 | 0.8199 | 0.8289 |
| RMSE | 7.1145 | 6.0616 |
| MAE | 5.7983 | 4,1374 |

Given the high R2 values for both training and testing data. The variability in wine ratings can be well explained by the model.

Comparing to the training data set, RMSE and MAE are both lower in the testing data, suggesting that the prediction errors are smaller on the testing set, and the model’s predictions are closer to the actual values in the testing set.

The model has no overfitting as the performance on the training and test sets is similar: R2 is not significantly higher on the training data. RMSE and MAE are not significantly lower on the training data.

* + - Residual plots are randomly distributed. P-value in Breusch-Pagan test are higher than 0.05. Both indicates that the model is not heteroskedastic.

1. Conclusion:

The linear regression model developed to predict wine ratings based on price, alcohol content, sulphates, and the country of origin indicates a good fit. Significant portion of the variance in wine ratings is explained by these variables.

Rating = B0 + B1Price + B2Alcohol + B3Sulphates + B4CountryFrance

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generatedA screen shot of a computer screen

Description automatically generated

A computer screen shot of text

Description automatically generated

## b)

The data is not heteroskedastic. As P-value in Breusch-Pagan test are higher than 0.05.

## c)

Wine would be rated 76.83

A computer screen shot of a program code

Description automatically generated

## d)

As mentioned above, the P-value associated with price is very small, indicating that the relationship between price and rating is statistically significant.

To determine if the price of a wine would increase its expert rating, we need to look at the coefficient associated with price in the regression model. The coefficient of Price is 1.025. This means, holding all other variables constant, for every one unit increase in price of the wine, the expert rating is expected to increase by 1.025 units.

# Question 4:

## a)

Sales = B0 + B1Ad\_Budget + B2Price + B3CountryUS

Sales = 435.3492 + 50.2401Ad\_Budget – 78.1967Price + 1010.3254CountryUS

A screenshot of a computer

Description automatically generated

Development Steps:

1. Data Exploration: Checked the structure and completeness of the data. There’s no missing data.
2. Preprocessing: Created a dummy variable 'CountryUS' to handle Country categorical data.
3. Model Building:
   * + Constructed a linear regression model between the predictors and the target variable ('Sales ~ Ad\_Budget + Price + Distance + CountryUS')
     + T-tests were used for each variable in the model. Since ‘Distance' has high p-value, this variable does not belong to the model and was removed from the model.
4. Model Training and Assessing:
   * + Used train\_test\_split to divide the data into training and testing sets for model validation.
     + Trained the model on the training data.
     + Used the trained model to predict the ratings for the test dataset, evaluating the model against new data.
     + Validate the model accuracy by calculating R2, RMSE and MAE:

Given the high R2 values for both training and testing data. The variability in Sales ratings can be well explained by the model.

The model has no overfitting as the performance on the training and testing sets are quite similar.

* + - Residual plots are randomly distributed. P-values in the Breusch-Pagan test are higher than 0.05. Both indicate that the model is not heteroskedastic.

1. Conclusion:

The linear regression model was developed to predict sales based on Ad\_Budget, Price, and the country people reside in. Significant portion of the variance in curry sales is explained by these variables.

Sales = B0 + B1Ad\_Budget + B2Price + B3CountryUS

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screen shot of a computer screen

Description automatically generated

A screen shot of a computer

Description automatically generated

## b)

A screenshot of a computer

Description automatically generated

Sales = B0 + B1Ad\_Budget + B2Price + B3CountryUS + B4Ad\_Budget\_US

The model includes an interaction term between Ad\_Budget and CountryUS to specifically measure different responses to advertising in the US versus Canada.

H0: B4=0

H1: B4 ≠ 0

P-value = 0.3492, failed to reject the null hypothesis, there’s not enough evidence to say the response to the ads in US is different from in Canada.

## c)

A screenshot of a computer code

Description automatically generated

Sales = B0 + B1Ad\_Budget + B2Price + B3CountryUS + B4Ad\_Budget\_US + B5Price\_US

H0: B4=B5=0

H1: At least one of B4 ≠ 0 or B5 ≠ 0

P-value = 5.02 \*10-25, reject the null hypothesis, there is a difference between the two countries, on how pricing and ads budget impact sales.

Combined with the result we got from part b, while ad budgets alone may not impact sales in different countries, price and advertising together do impact sales in different countries. These two factors should be considered together when building the model.