DEAKIN UNIVERSITY

2023

SIG 718
Real World Analytics

Mid Term Assessment

Link to Video : YouTube Video

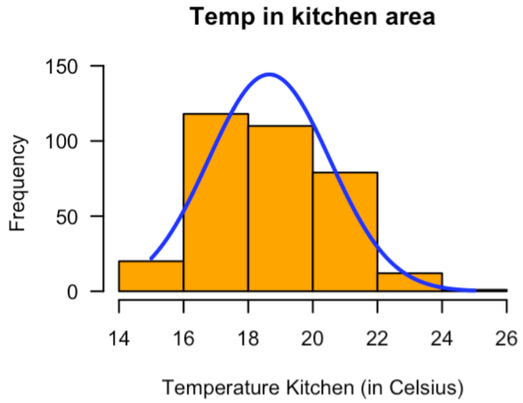
Presented By :
Suraj Mathew Thomas
S223509398

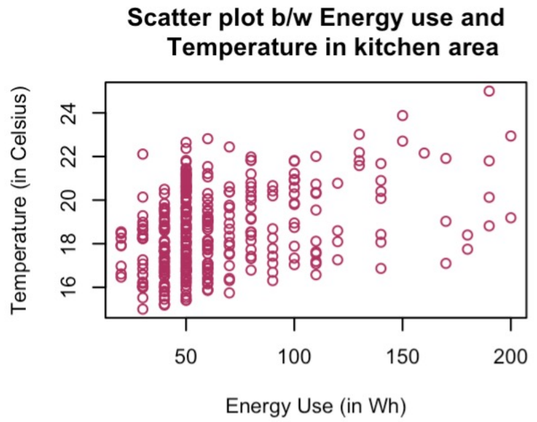
- **Data Set** : Energy Appliance Dataset [Energy Use of Appliances]
- **Input Variables** : 5 (X1, X2, X3, X4, X5)
- **Target Variable** : 1 (Y)
- **Data Definition** :
  - **X1:** Temperature in kitchen area, in Celsius
  - **X2:** Humidity in kitchen area, given as a percentage
  - **X3:** Temperature outside (from weather station), in Celsius
  - **X4:** Humidity outside (from weather station), given as a percentage
  - **X5:** Visibility (from weather station), in km
  - **Y:** Appliances, energy use, in Wh
- **Sample Size – 671 | Sample Used - 340**

I. Identified Data Distributions of all the input variables and their relationship with the variable of interest

Before any Data Transformation

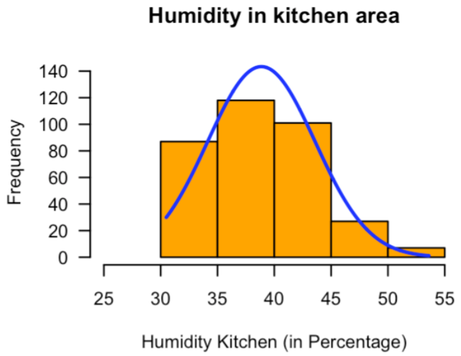
X1

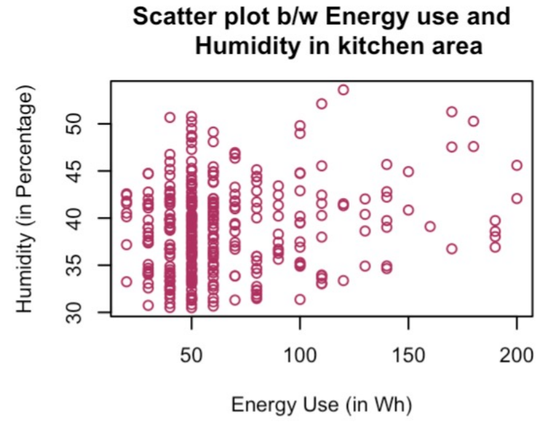




- Distribution of the variable is ~ normal as the histogram graph shows symmetry.
- The K-S test reveals the **p-value was > 0.05**, thereby validating that the distribution of X1 is approximately normal.
- The scatter plot between X1 and Y shows a positive relationship between the input and the target variable.

X2





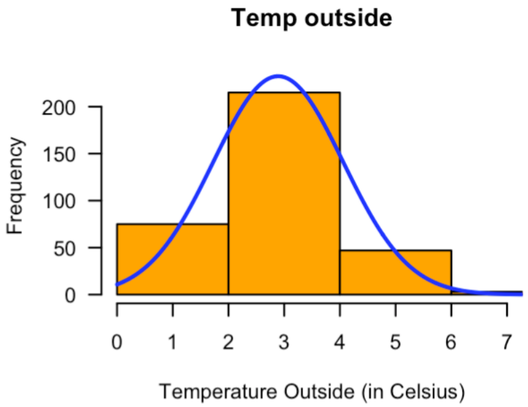
- Distribution of the variable is ~ normal as the histogram graph shows symmetry.
- The K-S test reveals the **p-value was > 0.05**, thereby validating that the distribution of X2 is approximately normal.
- The scatter plot between X2 and Y shows a positive relationship between the input and the target variable.

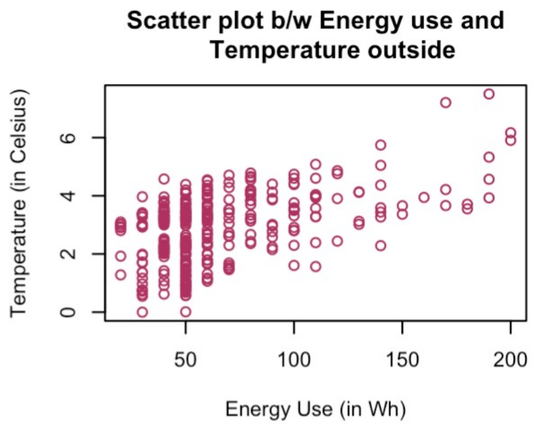
DEAKIN UNIVERSITY

I. Identified Data Distributions of all the input variables and their relationship with the variable of interest

Before any Data Transformation

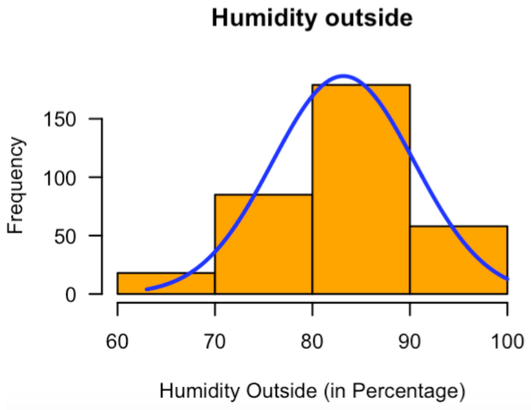
X3

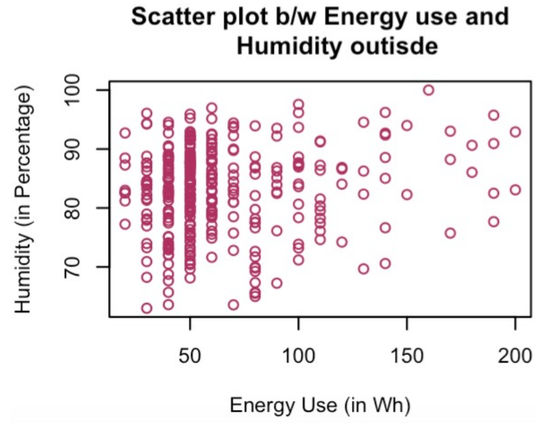




- Distribution of the variable appears normal as per the histogram
- But the K-S test reveals the **p-value was < 0.05**, thereby confirming that there is skewness. When testing the skewness, **it is 0.09 which is less than 0.5 and indicates moderate negative skewness.**
- The scatter plot between X3 and Y shows a positive relationship between the input and the target variable.

X4





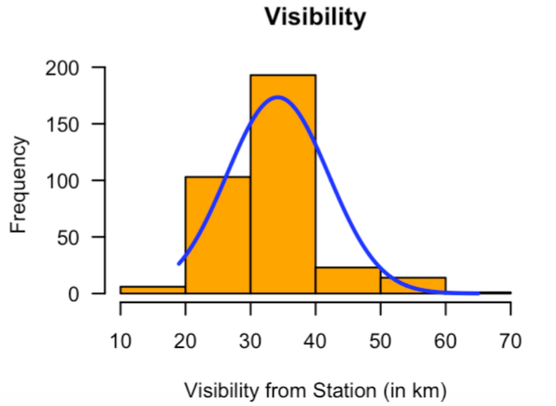
- Distribution of the variable is ~ normal as the histogram graph shows symmetry.
- The K-S test reveals the **p-value was > 0.05**, thereby validating that the distribution of X4 is approximately normal.
- The scatter plot between X4 and Y shows a positive relationship between the input and the target variable.

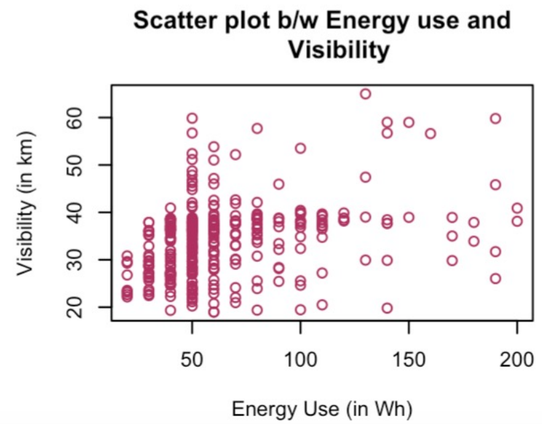
DEAKIN UNIVERSITY

I. Identified Data Distributions of all the input variables and their relationship with the variable of interest

Before any Data Transformation

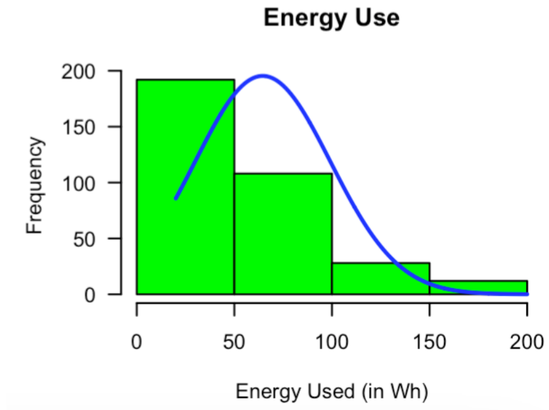
X5





- Distribution of the variable appears normal as per the histogram
- But the K-S test reveals the **p-value was < 0.05**, thereby confirming that there is skewness. When testing the skewness, **it is 0.717 which is greater than 0.5 and indicates moderate positive skewness.**
- The scatter plot between X5 and Y shows a positive relationship between the input and the target variable

Y



- Distribution of the variable appears normal as per the histogram
- But the K-S test reveals the **p-value was < 0.05**, thereby confirming that there is skewness. When testing the skewness, **it is 1.798 which is greater than 0.5 and indicates high positive skewness.**

DEAKIN UNIVERSITY

Normality Test Metrics

P- VALUES [Kolmogorov - Smirnov Test]

Variable	P-Value
X1	0.2185
X2	0.2698
X3	0.01134
X4	0.09223
X5	8.807e-05
Y	1.554e-15

SKEWNESS

Variable	Skewness
X1	0.3572175
X2	0.4270809
X3	-0.4451346
X4	0.09757824
X5	0.7172966
Y	1.798201

INFERENCE

Variable	Inference
X1	Can be approx to normal distribution
X2	Can be approx to normal distribution
X3	Can be approx to normal distribution
X4	Can be approx to normal distribution
X5	positively skewed - We need to apply log transformation
Y	positively skewed - We need to apply log transformation

- X3, X5 and Y are less than 0.05 & X1, X2, X4 are greater than 0.05
- Therefore, since X1, X2 and X4 are greater than 0.05, there is no significant difference between normal distribution and the distribution of X1, X2 and X4.
- Since these are almost normally distributed, (we can cross verify this using the histogram that we generated above) we need not perform any polynomial or log transformation.
- X3, X5 and Y looks a little skewed in the histogram and also the KS test says that they are not normally distributed. Hence let us test the skewness to see what is the transformation we need to apply.

- Skewness = 0, then it is perfect normal distribution
- Skewness is between -0.5 and 0.5, then we approximate it to a normal distribution
- Skewness > 0.5 then positively skewed
- Skewness < 0.5 then negatively skewed

DEAKIN UNIVERSITY

II. Selecting 4 Variables By Performing Correlation Analysis

Variables	Pearson Correlation Test	Spearman Correlation Test
X1: Temp in Kitchen Area	0.3579697	0.32754308
X2: Humidity in Kitchen Area	0.1694812	0.11100195
X3: Temp outside (weather station)	0.4965041	0.42428935
X4: Humidity outside (weather station)	0.1075600	0.07760502
X5: Visibility (weather station)	0.3152205	0.35049866

- We can see in both Pearson and Spearman correlation test the variable X3 (Temp outside (from weather station)) has the strongest correlation with the variable of interest (Energy Consumption) or Y.
- The variable X4(Humidity outside (from the weather station)) has the weakest correlation with the variable of interest (Energy Consumption) or Y
- Therefore, the 4 variables that we have selected are (X1, X2, X3, X5) and the Variable of Interest (Y).
- These will now be used in the transformations.

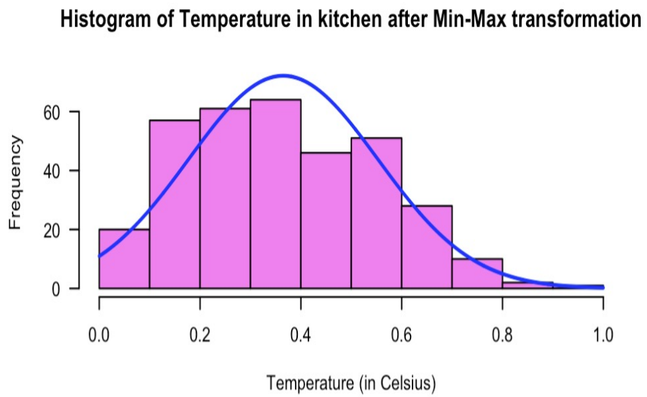
III. Transformation Of The Selected Variables

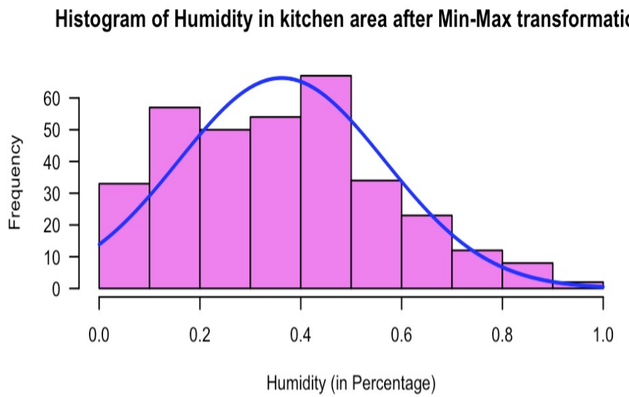
- For the variables X1, X2 and X3, we have seen above that it can be approximated to normal distribution as skewness is between -0.5 and 0.5
- Also, Pearson Correlation test results indicated that all three variables have positive relationship with the variable of interest. Hence, there is no need of negation transformation on these variables.
- Therefore, we apply min-max transformation to X1, X2, X3.
- For the variable X5 and Y, we first apply log transformation and then apply the min-max transformation as it is positively skewed.
- The min-max transformation ensures that the variables that are taking values over different ranges are transformed to the same unit interval [0,1]. This is necessary for aggregation.

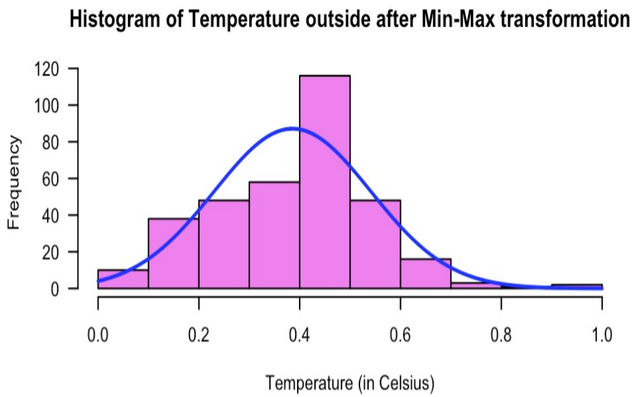
Min-Max Transformation Formula - $x(new) = x - \min(x) \,/\, \max(x) - \min(x)$

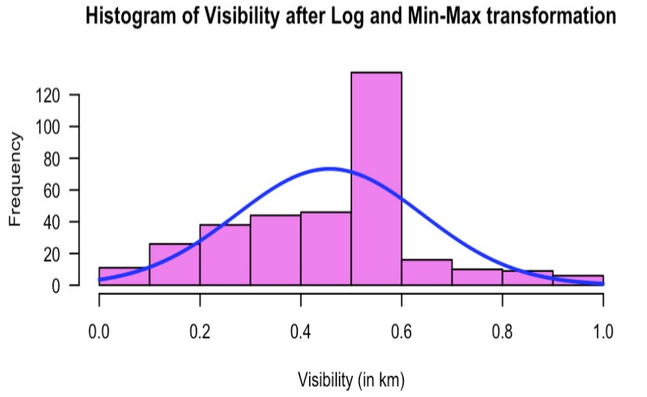
Log Transformation Formula - $x(new) = \log10(x)$

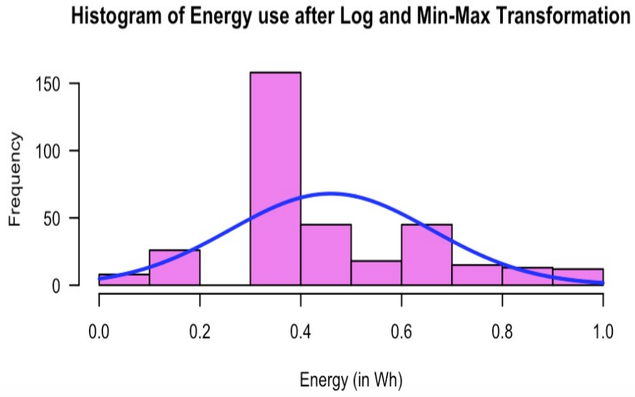
III. Transformation Of The Selected Variables











DEAKIN UNIVERSITY

IV. Model Fitting – Error Measures, Correlation Coefficients, Weights & Parameters

Error Measures & Correlation Coefficients	Models				
	WAM	WPM (p = 0.5)	WPM (p = 2.0)	OWA	Choquet Integral
Root Mean Square Error (RMSE)	0.170118046971246	0.177480177373415	0.162877603623949	0.167443931237899	0.154012557951177
Average Absolute Error	0.130007322359533	0.135443262722266	0.123947764760593	0.130751897605541	0.118444291784058
Pearson Correlation	0.594387535149108	0.56405425418301	0.616687399508369	0.55110631249925	0.640774620517391
Spearman Correlation	0.539680480113747	0.527475112604894	0.550437136544852	0.508871037957623	0.585900862326587

Weights & Parameters	Models				
	WAM	WPM (p = 0.5)	WPM (p = 2.0)	OWA	Choquet Integral
Weight/Shapley (in the case of Choquet) (w1)	0.250823618479399	0.220544620543378	0.247907845911032	0.146960667724284	0.272151349102236
Weight/Shapley (in the case of Choquet) (w2)	0	0	0	0.279742350066015	0.0307166347895854
Weight/Shapley (in the case of Choquet) (w3)	0.444616145291077	0.412229212898623	0.548258621484214	0.134489486934526	0.527691767022004
Weight/Shapley (in the case of Choquet) (w4)	0.304560236229524	0.367226166557999	0.203833532604755	0.438807495275171	0.169440249084761
Orness (only for OWA and Choquet)	-	-	-	0.621714603253527	0.67796428110174

Observations:

- Choquet Integral Model **has the lowest RMSE and Av. Absolute Error**. (meaning the average difference between predicted value and original value is minimal).
- **The Pearson and Spearman Correlation coefficients for Choquet is the highest** when compared to the other models. (meaning the strength of the relationship between the independent and dependent variables is the strongest).
- **Therefore, we choose the Choquet Integral Model over the other models.**

DEAKIN UNIVERSITY

Observations (contd.)

- From the correlation analysis, we find that variable **X3** (Temperature outside (from weather station), in Celsius) **has the strongest relationship** with the target variable (Appliances, energy use, in Wh).

- Therefore, more importance or weight should be assigned to the third variable for the model to perform better.

- In the previous Weights and Parameters table we can see except for OWA model, all the other models have the highest weight against X3 (W3). And amongst all the models, Choquet has the highest weight against X3.

- The reason why W3 in the OWA model is lower could be because the OWA has the highest RMSE values amongst all the models and therefore it could be due to underperformance.

- Also, the orness measure in Choquet is the highest at 0.67796428110174. It indicates in extent to which the averaging function favors high inputs. The orness is > 0.5. Therefore, we can conclude that it favors high inputs.

- Variable X2 has the least importance as we see the weights W2 for all models except OWA is zero.

V. Choosing The Best Fitting Model And Predicting For New Values Using The Model

- The best fitting model → **Choquet Integral Model**

- New Data (22,38,4,34)

- Transformed New Data by applying the same transformations as before → Transformed data (0.7000000 0.3248195 0.5333333 0.4736989). Now we apply this to predict the “Energy Use” using this new transformed data.

- **Measured value of Y is 100 Wh** - given in the question.

- **The predicted value of Y is 82.95919 Wh** - which is a pretty good prediction.

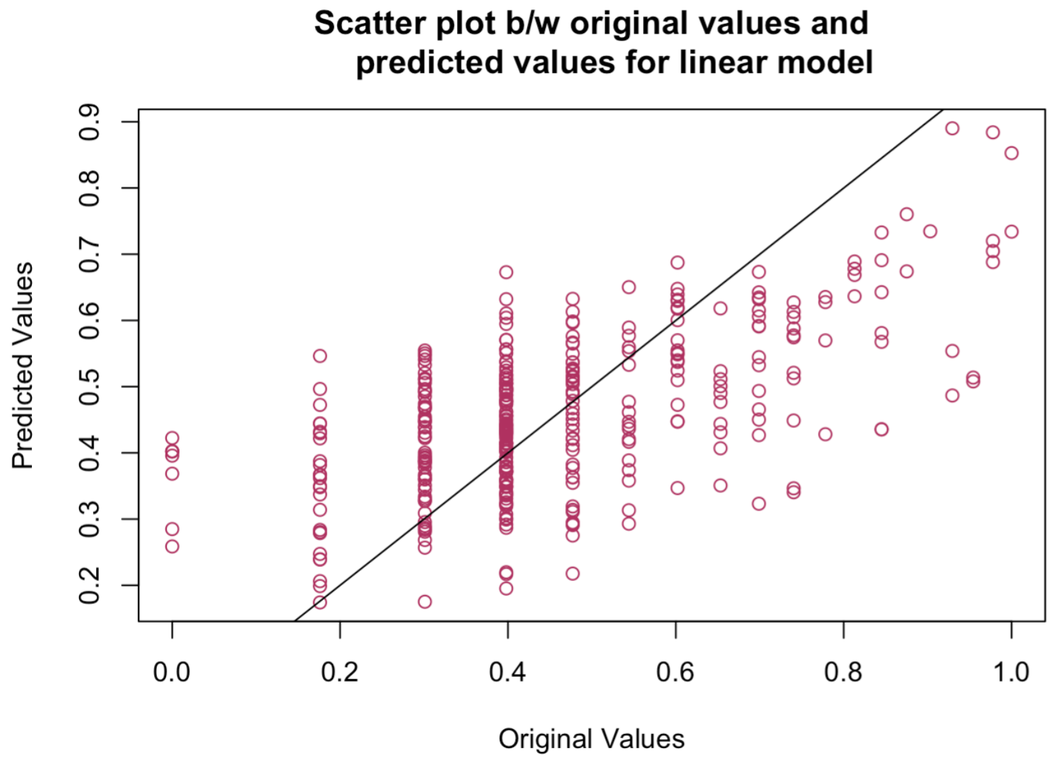
- X2 – has least importance → w2(0.0307166347895854)
- X3 – has highest importance → w3(0.527691767022004)
- X1 & X5 – have moderate importance → w1(0.272151349102236) | w4(0.169440249084761)

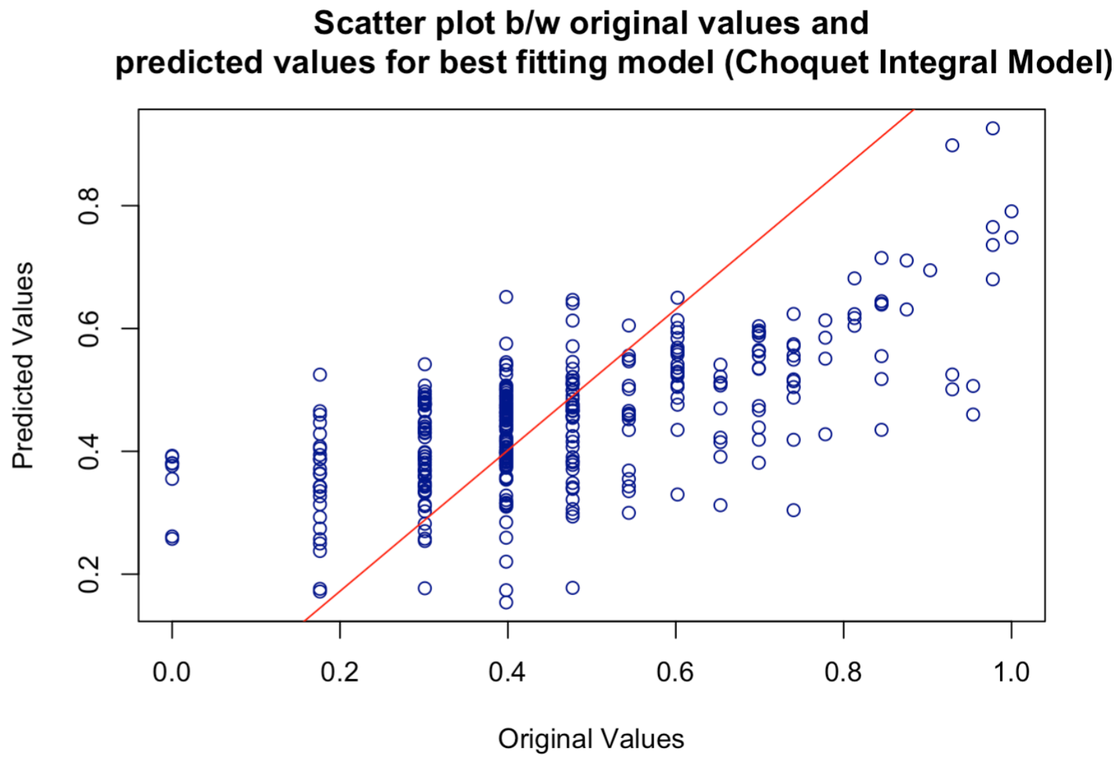
- Therefore if you need to get a low energy for the appliance usage, then input should be the lowest at X3, lower values at X1 and X5
    - Low Temperature outside (from weather station), in Celsius (X3)
    - Low Temperature in kitchen area, in Celsius (X1)
    - Low Visibility (X5)
    - Same humidity approximately (X2)

Yields Low Energy Use of Appliance

DEAKIN UNIVERSITY

VI. Linear Model Vs Choquet Integral Model





Model	Performance Measures	Values
Choquet Integral	RMSE	0.154012557951177
Linear Model	Residual Standard Error	0.1593

DEAKIN UNIVERSITY

VI. Linear Model Vs Choquet Integral Model

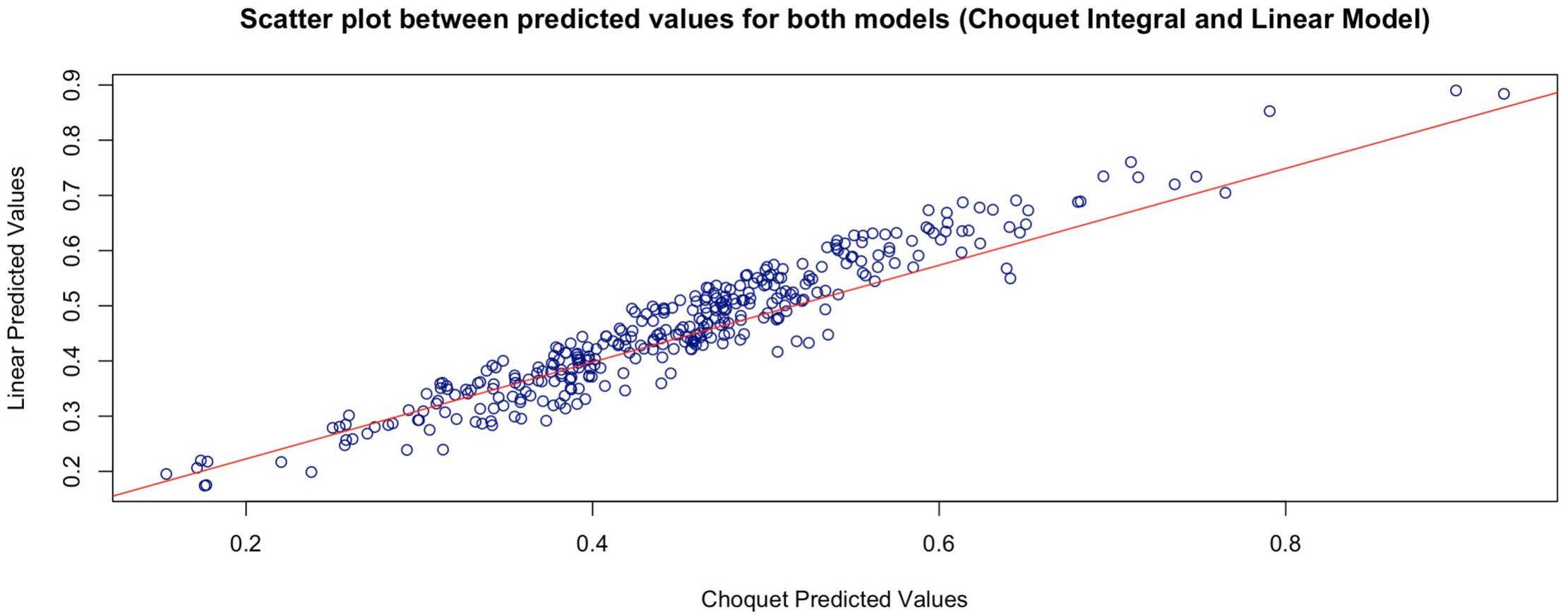
Model	Performance Measures	Values
Linear Model	Residual Standard Error	0.1593
	R- Square (Multiple Regression)	0.369
	Adjusted R – Square (Multiple Regression)	0.3615
	p – value	2.2e-16
	F Statistic	48.98

- The residual standard error is very small (1.593) – This indicates that the model fitment is good.
- Only 36 % of variance in Y (target variable) can be explained by all the independent variables used. This is inferred from the Adjusted R Square value.
- Adjusted R Square is a good measure to evaluate the model performance. Having this measure is an advantage of linear model over the Choquet model.

VI. Linear Model Vs Choquet Integral Model

- The p-vaue is < 0.05 which indicates that the model is statistically significant and has good predictive power
- But the linear model comes with assumptions such as linearity, homoscedasticity, normality and independence of errors etc.
- The relationship between the predicted values of both the linear and choquet models is positive and strong. This can be seen in the chart in the next slide. Therefore, both the models exhibit similar performance.
- Linear Models often tend to exhibit an unbiased estimation of the model performance as compared to the choquet model.

VII. Predicted Values Comparison b/w Choquet and Linear Model



References

Videos

1) Simon James (2020) ‘An Introduction to Choquet Integral, YouTube, accessed 30 November 2023, Link
2) Simon James (2020) ‘Learning the weights of Choquet Integral, YouTube, accessed 30 November 2023, Link

Course Material

1) Deakin Study Material (n.d) ‘Week 1 Aggregation Function’, Real World Analytics SIT718, Deakin University
2) Deakin Study Material (n.d) ‘Week 2 Means and Weights’, Real World Analytics SIT718, Deakin University
3) Deakin Study Material (n.d) ‘Week 3 Error Measures’, Real World Analytics SIT718, Deakin University

Link to Video Presentation: https://www.youtube.com/watch?v=T7IGLSY6jp0

DEAKIN UNIVERSITY

Thank You