**Article Critique**

**MSCI 718**

**Areeb Ullah Ansari (20925657)**

**Submitted to: Oliver Schneider**

**Introduction:**

The article ‘Mutiple linear regression’ reviewed and critiqued is written by Usamsa baig (waterloo grad) as part of his assignment for MSCI 718. The article aims to determine which predictor variables from the Melbourne housing dataset (Containing 13580 observation and 21 variables) are critical for prediction of prices in the city. Usama has deployed Multiple linear regression model to determine them -to further his aim of aiding policy makers and potential buyers make informed and strategic decisions while buying or legislating for the house marker. This critique will explore the weakness and strengths of the article and give potential improvements based on logic, completeness, correctness, and presentation of the said report.

**Process and logic:**

The article follows a logical order from the beginning. Almost every step is instinctive outcome of the previous and builds upon the foundation laid to portray a succinct and cohesive outlook of the situation. For e.g., data description (Conveying level of measurements, no of observations etc..) is followed by data cleaning which is logical as it informs users of what predictor variables exist in the data set followed by what proportion of them are missing and need to be imputed/deleted before determining the final predictor variables as multi linear regression models would not build if there were missing values in any of the predictor variables selected. But although the data wrangling was concrete the article missed to report as to what insights/statistic led to dropping of Longitude and latitude which can dent the validity of the conclusion/prediction.

The logical considerations can also be seen by the fact that the author gives a roadmap to follow in the planning section to make it easier for the readers to follow along. He describes the method for selecting the predictor variables first, then checks if the two models are significantly different using ANNOVA and finally selects the model with higher R square value using summary statistic -showing that all his analysis follows a rational structure.

Text

Description automatically generated

**Image 1: Use of R-squared value by Author to pick best model.**

But even though the article follows a logical footpath it still can’t detract us from some logical jumps which could be scrutinized. For e.g., the method deployed for picking predictor variables is not justified as there is no logical reason given as to why their selection was made. Adequate reasoning (apriori research) should have been made/stated even when using hierarchal approach for feature selection. Also, visual tools like histograms or scatter plots should have been used to show outliers. Also, it should have been combined with data cleaning and not proceed model selection as it is commonly deemed to be a part of preprocessing the data. Overall, the report provides a clear problem statement which is followed by comprehensive analysis to further justify the selection of the predictor variables in the final model. The final findings reported are also justified by the general research as mentioned in the referred link (researchgate.net, 2023) presented in the report.

**Completeness and Correctness:**

The report offers concise, clear, and well-defined analytical methods for comparing models and checking the robustness of the model against the assumptions. For e.g., it assesses all assumptions of the linear model using tests such as Durbin Watson, multicollinearity between predictors using VIF and other visualization tools such as residual vs fitted plot, QQ plot and other scatter plots for assumptions like linearity, normality and zero variance. It dwells into comprehensive discussion about why each assumption was met or not met based on the test statistic. However, the author did not consider the impact of the violations of the model assumptions gravely. For e.g., in case non-normality as reported by QQ-plot a log or square root transform could have been used to transform the data, so that the assumptions would have been met or CLT could be invoked to assume the assumption of normality, or the author could have stated the impact of non-normality on the model. Similarly, the author could have mentioned alternative ways to deal with the violation of homoscedasticity like weighted R square, data transformation, robust regression, or non-parametric regression.

**Text

Description automatically generated**

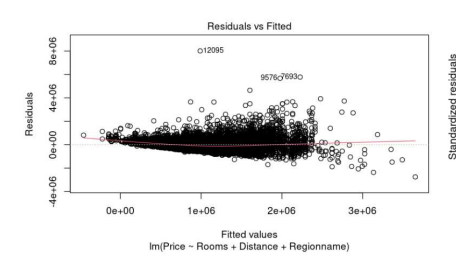
**Image 2: Use of additional resources by the author**

It is also visible by image 2 that the user used additional resources to validate and analyze his result which adds to the degree of correctness of the report. Moreover, the value of test statistic reported while testing the assumptions align with the class slides which further the validity of the report in terms of completeness and correctness.

However, dropping outliers without considering practical real-life research could result in misleading conclusions. That in addition to the lack of intercept and coefficient value of linear regression and their interpretation would mark a dent on the spine of the spine of the conclusion

**Presentation:**

The report has a coherent flow with concise and well-presented graph. The plots have low data-to-ink ratio and are clear with proper titles and label as shown by the exhibit attached below. The graphs used have no distractions, less ink, no chart junk and use an appropriate scale. Although, the aesthetics of graph can be used by using different shapes/color for data points.



**Image 3: Exhibit of plots used in report.**

The report doesn’t use technical jargon, uses simple and clear language. Hasa logical structure with headings and graphs to present information visually. It uses a consistent font, spacing and formatting which makes it look polished. It caters to both beginners and experts as an audience. Finally, it includes relevant information in the appendices, including level of measurement, relevant graphs, and other supplementary information which allows the reader to explore data in further detail if they wish to. All this makes the report a professional document in terms of presentation.

**Conclusion:**

The report demonstrates an excellent process of deducing the most critical predictor variables from existing data. It uses adequate test, models, and visualization techniques to convey the story and methodology all the way from the introduction to the final deduction. However, some follow ups, future recommendations, visualization of data and looking at summary statistic of predictor variable while in pre-processing could have elevated it all aspects.

Reference:

[1] Usama Baig Assignment 3, MSCI 718.