Multiple Linear Regression and Time Series Analysis

Rebecca Santana de Pina
Msc in Data Analytics
National College of Ireland
Dublin, Ireland
x23436786@student.ncirl.ie

I. Introduction

Multiple linear regression is a statistical method used to analyze the relationship between a dependent variable and independent variables. It is important for making predictions. But for the model to be valid, it must satisfy the Gauss-Markov assumptions, which are as follows: linearity, homoscedasticity, independence of errors, absence of multicollinearity, and normality of residuals.

This report analyses a dataset which contains 1,000 observations, using multiple linear regression to predict a continuous dependent variable (y) based on three independent variables (x1,x2,x3).

II. Explanatory Data Analysis

The dataset used contains 1,000 rows and 4 columns:

- *y*: continuous dependent variable
- *x1,x2*: continuous independent variable
- *x3*: categorical independent variable with three levels (A,B,C)

TABLE I
Descriptive Statistics for Variables

Variable	y	x1	x2	$x3_B$	$x3_C$
Count	1000	1000	1000	1000	1000
Mean	15,461.12	49.99	199.92	0.403	0.291
Std Dev	1,795.13	4.91	4.35	0.490	0.454
Min	7,972.73	33.15	186.23	0	0
25	14,434.33	47.27	197.44	0	0
Median	15,521.01	49.96	199.91	0	0
75	16,538.97	53.00	202.46	1	0
Max	21,565.78	69.42	214.46	1	1

Key observations:

- The categorical variable $x3$ was one-hot encoded into binary columns $x3_B$ and $x3_C$, with $x3_A$ as the reference category.

The correlation matrix in Fig.1 reveals strong linear relationships between y, x1 and a small relationship between y and x2. No multicollinearity concerns are observed, as correlations between predictors are low.

A histogram of y in Fig.2 indicates a normal distribution, though a few extreme values can be observed on the left and right ends of the distribution, indicating potential outliers.

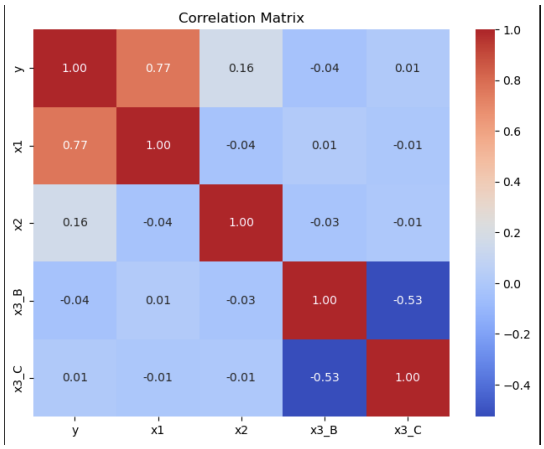

Fig. 1. Correlation Heatmap

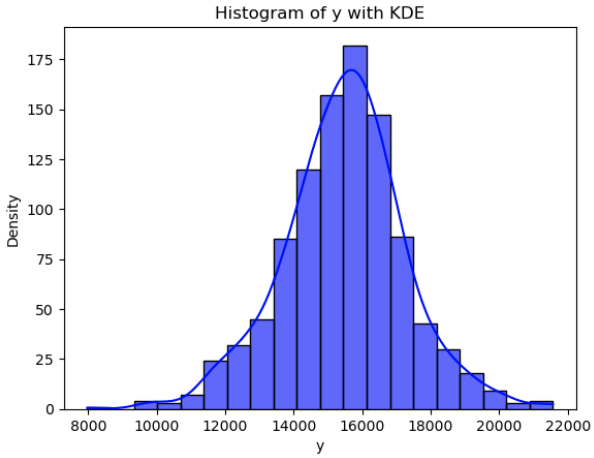

Fig. 2. Histogram of y with KDE Overlay

III. Data Preparation

A. Missing values

Upon inspection, no missing values were detected in the dataset, as shown by the isnull() function. Therefore, no values were removed or replaced.

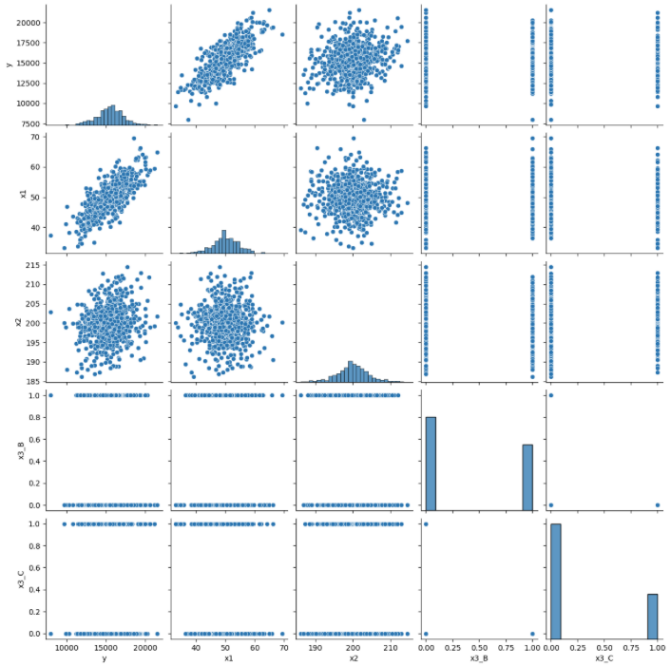

Fig. 3. Pairplot of the train data

B. Encoding Categorical Variables

The independent variable $x3$ was categorical with three levels (A,B,C). To incorporate it into the regression model, one-hot encoding was applied.

Using the getdummies() method, the categorical variables was transformed into binary indicators and later for integers. This approach avoids multicollinearity issues and allows comparisons of the effects of $x3_B$ and $x3_C$ relative to $x3_A$

C. Outlier Detection

Two methods were used to detect outliers in the continuous variables (y,x1,x2):

1) Z-score Method: Outliers were identified as data points with absolute z-score greater than 3. This method detected 16 outliers.
2) Interquartile Range (IQR) Method: Outliers were identified as data points outside the range [Q1 - 1.5 x IQR, Q3 + 1.5 x IQR]. This method detected 68 outliers.

After evaluating the impact of removing outliers on model performance, the Z-Score Method was chosen as it retained more data and improved model stability.

D. Data Splitting

The dataset was split into training (80) and (20) subsets using the traintestsplit() function, with the student number as the random seed for reproducibility.

This split ensures that the model can be trained and evaluated on separate data to assess its generalization capability.

IV. Modelling

This section details the development of the multiple linear regression model using the prepared dataset.

An initial Ordinary Least Squares (OLS) regression model was built using all the variables, including outliers. The results showed:

- Significant p-values for $x1, x2$ and $x3_B$ ($p < 0.05$)
- Some predictors, such as $x3_C$, were not statistically significant ($p > 0.05$)

A. Building the Final Model

Built the OLS Model using the cleaned training dataset (Z-Score method).

- R^2: the final model achieved an R^2 of 0.628, indicating it explained 62.8 percent of the variation in y.
- Significant Predictors:
  - $x1$ and $x2$ were highly significant ($p < 0.001$)
  - $x3_B$ was significant ($p = 0.004$)
  - $x3_C$ was not significant ($p = 0.289$)

B. Intermediate Models and Their Rejection

- Baseline Model (No Cleaning):
  - R^2= 0.650, indicating that 65 percent of the variance in y was explained by the independent variables ($x1, x2, x3_B, x3_C$).
  - Residual plots indicated potential outliers.
  - Retained for comparison but rejected as the final model due to the presence of outliers
- IQR-Cleaned Model:
  - Achieved a lower R^2 (0.576).
  - Removed a large number of observations, reducing the model's variability.
  - Rejected in favor of the Z-Score-cleaned model.

V. Interpretation

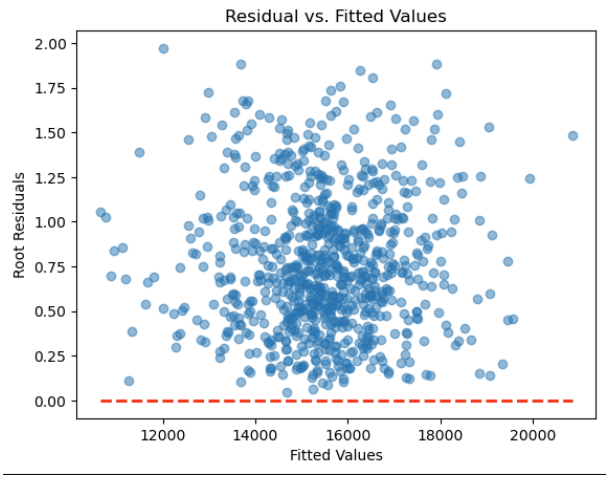

Fig. 4. Residual vs. Fitted

In the Fig. 4 the residuals appear randomly scattered around 0. There doesn't appear to be a clear systematics pattern. The

variance of the residuals appears constant across the range of the fitted values, supporting the assumption of *homoscedasticity*. A few residuals deviate significantly from the rest, suggesting potential outliers.

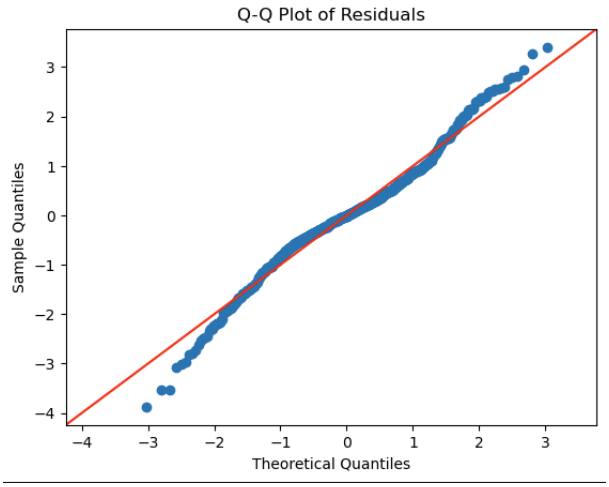

Fig. 5. Q-Q Plot of Residuals

In the Fig. 5 the residuals generally follow the red line, indicating that they are approximately normally distributed. But, there are some deviations at both ends of the plot, suggesting the presence of a few outliers.

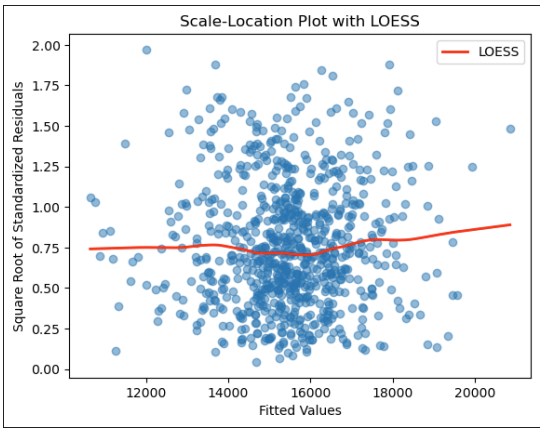

Fig. 6. Scale-Location Plot with LOESS

In the Fig. 6 the LOESS curve in the second plot is relatively flat, though it shows a slight upward trend at higher fitted values. This could suggest a increase in variance but is not strong. The plots suggest that the residuals generally satisfy the assumption of homoscedasticity.

In the Fig.7 there are several large spikes in Cook's Distance plot, indicating the presence of influential observations. More attention and futher investigation required.

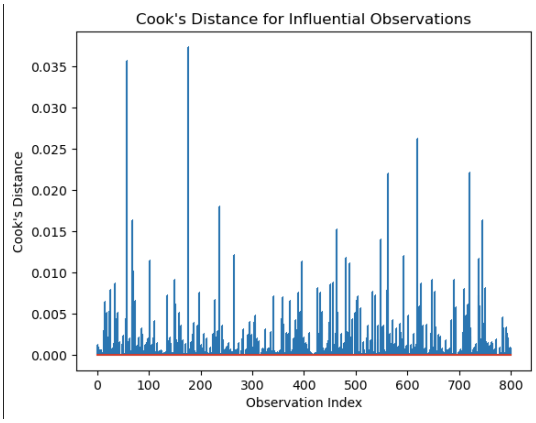

Fig. 7. Cook's Distance

VI. Diagnostics

A. Homoscedasticity - The Breusch-Pagan test

Breusch-Pagan Test Results:

- Lagrange Multiplier Statistic: 0.21187216164513245
- p-value: 0.9947697531205559
- F-statistic: 0.05265093424278067
- F-test p-value: 0.9948187960355617

Since the p-value is 0.9948, which is much greater than the 0.05, there is no evidence of heteroscedasticity on the model. The residuals appear to have a constant variance, satisfying the homoscedasticity assumption.

B. Multicollinearity

TABLE II
VIF Table

Variable	VIF
const	2281.571599
x1	1.002543
x2	1.002694
$x3_B$	1.378945
$x3_C$	1.380044

Even that the const VIF its high, its normal because the correlation with the others variables, but doesn't affect the model's validity. All predictors have VIF values below 5, indicating no multicollinearity concerns.

C. Condition Number

Condition Number: 1.7934583745508974. The condition number can indicate the presence of multicollinearity, in this case, its below 10, so suggest that multicollinearity its not a significant problem and the model is stable.

D. Shapiro-Wilk

The test statistic is 0.978. The p-value is 0.000. The W-statistic (0.978) is close to 1, which suggests that the residuals are approximately normally distributed. A p-value of 0.000 indicates that you reject the null hypothesis. This means that the residuals of the model do not follow a normal distribution.

VII. Evaluation

A. Performance Metrics

- Mean Absolute Error (MAE):
  - measures the average magnitude of the errors between actual and predicted values
  - lower MAE indicates better predictive performance.

  The MAE for the model is 911.5279174892966.

References

[1] Breusch, T., and Pagan, A. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. Econometrica, 47(5), 1287-1294.

[2] Durbin, J., and Watson, G. (1951). Testing for Serial Correlation in Least Squares Regression. II. Biometrika, 38(1/2), 159-177.

[3] Shapiro, S. S. and Wilk, M.B (1965). An analysis of variance test for normality (complete samples), Biometrika, Vol. 52, pp. 591-611.