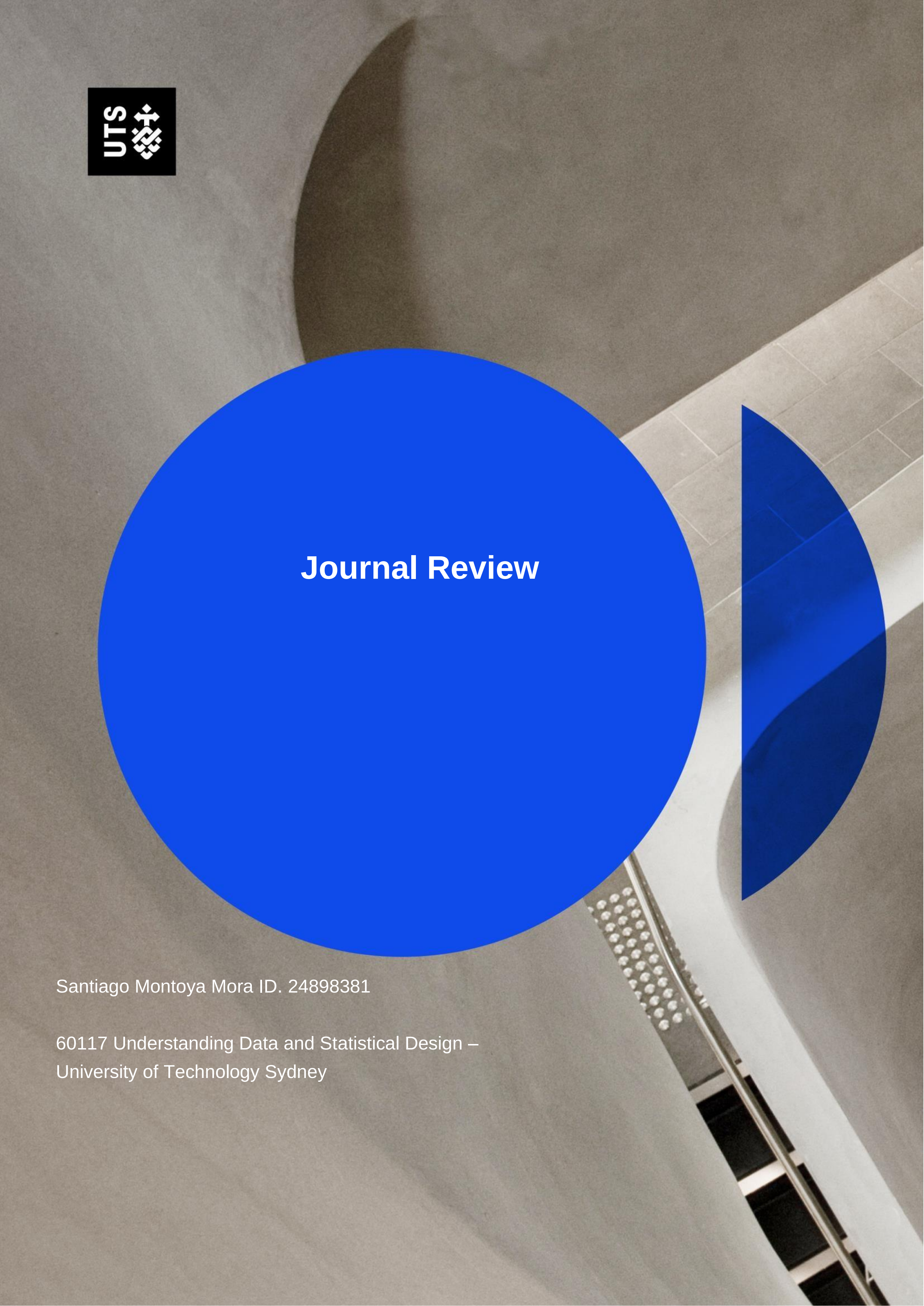


Journal Review

Santiago Montoya Mora ID. 24898381

60117 Understanding Data and Statistical Design –
University of Technology Sydney

Introduction:

This report takes an in-depth look at the research conducted by Teresa Angela Trunfio et al., published in 2022 in the journal BMC Medical Informatics and Decision Making under the title 'Multiple Regression Model to Analyse the Total LOS for Patients Undergoing Laparoscopic Appendectomy'. The study focuses on implementing a multiple linear regression model to predict the Length of Stay (LOS) of patients undergoing laparoscopic appendectomy. Using a dataset including demographic and clinical details of 357 patients from the University Hospital 'San Giovanni di Dio e Ruggi d'Aragona' in Salerno, Italy, the researchers developed a predictive model, first making sure to verify six hypotheses essential for statistical validity:

Each step of this process was crucial to ensure the robustness of the statistical model and its ability to predict the LOS effectively.

1. **Linear Relationship:** Verification of a linear relationship between the independent and dependent variables through scatter plots.
2. **Absence of Multicollinearity:** Assessment of multicollinearity by calculating the Tolerance and the Variance Inflation Factor (VIF).
3. **Independence of Residuals**: Analysis of the independence of the residuals using the Durbin-Watson statistical test.
4. **Homoscedasticity of Residuals:** Testing the homogeneity of variance of the residuals by plotting the standardised residuals against the standardised predicted values.
5. **Normality of Residuals:** Assessment of the normal distribution of the residuals using quantile-quantile (Q-Q) plots.
6. **Absence of Outliers:** Identification of outliers using Cook's distance, with a threshold set at 1.

Each step in this process was crucial to ensuring the robustness of the statistical model and its ability to predict effectively.

'LOS - measured in days - is defined as the difference between the date of admission and the date of discharge of the patient' Trunfio et al. (2022); according to the authors, this measure is used as an indicator of efficiency as it helps in the planning of patient admissions, directly impacting organisation and costs. For this reason, the prediction of this variable for different prognoses and procedures becomes of great importance for the hospital sector, in this study specifically for the procedure of appendectomy by laparoscopy, a surgical procedure widely used in cases of appendicitis.

Experiment design:

As previously mentioned, the authors used data from 357 patients extracted from the QuaniSDO system, a comprehensive hospital information management system of the University Hospital 'San Giovanni di Dio e Ruggi d'Aragona' in Salerno, Italy. The extracted information includes patient demographic and clinical data, such as gender, age, and Diagnostic Related Group (DRG), as well as procedure-specific details, such as the presence or absence of complications and the dates of admission and the laparoscopy (LC) procedure. In addition, the preoperative LC time was also calculated using the formula: date of LC procedure - date of admission, and the LOS was calculated. Additional data related to comorbidities were extracted from the DRG, including:

- Presence of comorbidities (yes/no).
- Heart disease (yes/no)
- Diabetes (yes / no)
- Hypertension (yes / no)
- Obesity (yes/no)
- Peritonitis (yes / no)
- Cancer (yes/no)

The dictionary of variables used by the authors and their data type can be found in **Table 1.**

Variable Name	Data Type	Description
Gender	Categorical	Patient's gender: Male or Female
Age	Continuous	Patient's age in years
Complicated diagnosis	Binary	Indicates if the patient’s diagnosis was complicated: Yes or No
Complications	Binary	Indicates if there were any complications during surgery: Yes or No
Pre-operative LOS	Continuous	Calculated as 'Date of LC Procedure' minus 'Date of Admission', measured in days
LOS	Continuous	Length of Stay: Total number of days from admission to discharge
Presence of Comorbidities	Binary	Indicates presence of any comorbid conditions: Yes or No
Heart Disease	Binary	Indicates if the patient has heart disease: Yes or No
Diabetes	Binary	Indicates if the patient has diabetes: Yes or No
Hypertension	Binary	Indicates if the patient has hypertension: Yes or No
Obesity	Binary	Indicates if the patient is clinically obese: Yes or No
Peritonitis	Binary	Indicates if the patient had peritonitis: Yes or No
Cancer	Binary	Indicates if the patient has cancer: Yes or No

Table1. *Data dictionary*

According to the authors, these chosen variables represent a significant advantage in predicting LOS. However, in addition to this, they do not pose any modelling difficulties, as only simple transformations to the categorical variables are needed without creating new columns. The careful selection of these variables strengthens the robustness of the statistical model and its ability to predict LOS effectively, thus providing valuable insights for hospital management and resource planning.

The authors chose the multiple linear regression (MLR) models for predicting LOS, which they built using IBM SPSS (Statistical Package for Social Science) software ver.27.

The model used for this study is represented in **Formula 1.**

$$y = \beta 0 + \beta 1 \ast x1 + \beta 2 \ast x2 + \beta 3 \ast x3 + \beta 4 \ast x4 + \beta 5 \ast x5 + \beta 6 \ast x6 + \beta 7 \ast x7 + \beta 8 \ast x8 + \beta 9 \ast x9 + \beta 10 \ast x10 + \beta 11 \ast x11 + \beta 12 \ast x12 + \varepsilon$$

***Formula1.** Model equation.*

Where “y” represents the variable to be predicted, LOS, β0 the intercept of the model, xi and βi represent the 12 independent variables with their coefficients (preoperative LOS, presence of complications, complicated diagnosis, gender, age, presence of comorbidities, heart disease, diabetes, hypertension, obesity, peritonitis and cancer) and ε represents the model error. This model includes numerical and categorical variables, considering their direct impact on LOS prediction without including interactions between them.

The approach of this model has certain limitations, the most important being:

- As the authors themselves mention, they have minimal values of the specific comorbidities; other comorbidities may explain more of the model's variance, which is not being captured.
- The use of the variable Preoperative LOS: This variable is directly linked to LOS; this may induce collinearity in the model, affecting the stability of the model, in addition to the fact that it is a variable that is only known at the time of admission to the procedure, so it does not contribute any logistic value to the hospital.
- The model does not consider inter-relationships between predictors, which may cause the model to omit meaningful relationships between predictors. For example, the combination of specific comorbidities and their joint impact on LOS could be significant but not detected in the current analysis.

Alternatives to improve these limitations are:

- Expand the data extraction, specifically the section on comorbidities; it can be based on the literature and extract those believed to be more influential in LOS.
- Change the calculation of LOS to date of discharge - date of LC procedure, so that the Preoperative LOS variable will not be directly linked to the final LOS value and the model will be able to capture the true importance of this variable.
- Create interaction terms between variables that are suspected to have significant joint effects. For example, interactions between different comorbidities (such as diabetes and hypertension) and demographic factors and comorbidities (such as age and presence of heart disease) should be included.

In terms of the selected model, MLR is appropriate for this type of analysis, given its power to handle multiple independent variables and its ability to quantify the effect of each predictor on the dependent variable (LOS). In addition, MLR is relatively simple to interpret and apply using statistical software such as IBM SPSS. However, further modelling could be considered for situations where the outcome is not linearly related to the predictors.

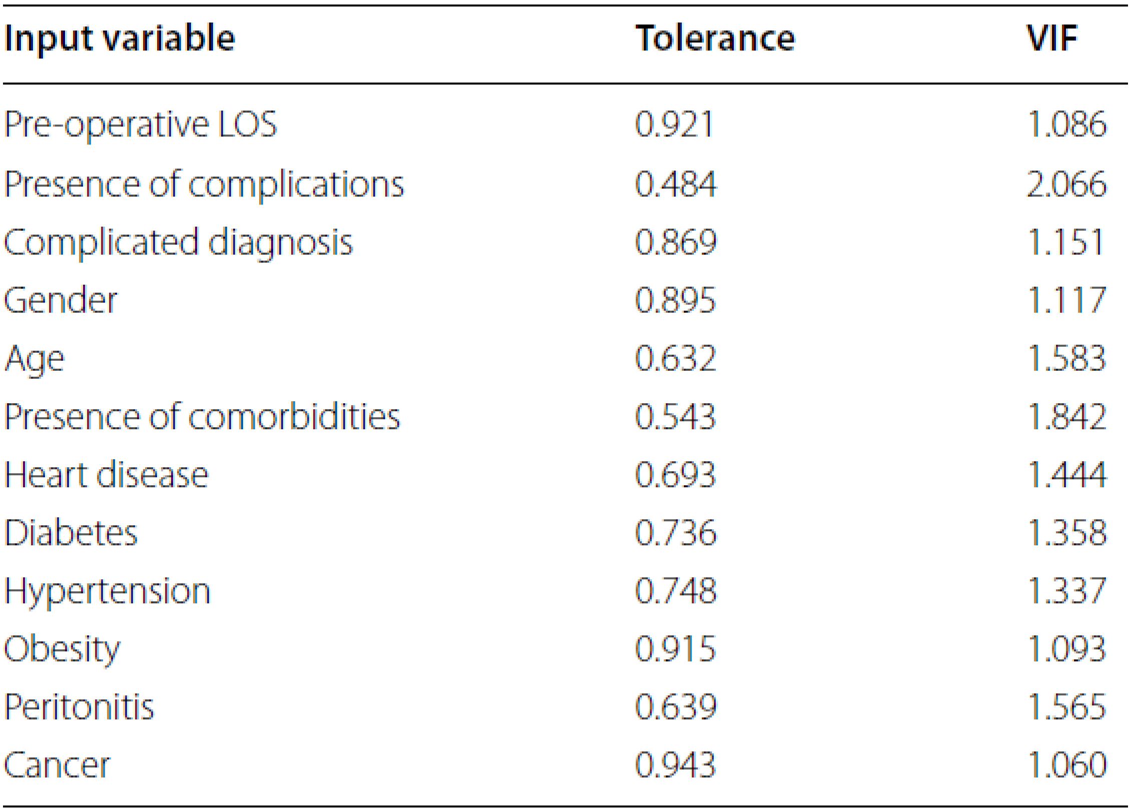
Statistical Analysis:

As previously mentioned, the authors sought to verify six hypotheses before creating the model:

1. **Linearity:** The authors propose verifying this hypothesis using scatter plots. This methodology is appropriate when analysing this hypothesis. However, no evidence of scatter plots is found in the paper. These graphs are extremely useful, as they allow us to identify each variable's relationship with LOS and indicate whether or not it is

necessary to carry out transformations; not having the graphs prevents us from extracting this type of information.

2. **Multicollinearity:** To ensure that there were no multicollinearity problems in the data, the authors used two methods, using the **formula2**, to calculate the variance inflation ratio (VIF) and also calculate the tolerance (1-R2). The results obtained by the authors are given in **table 2**:



Input variable	Tolerance	VIF
Pre-operative LOS	0.921	1.086
Presence of complications	0.484	2.066
Complicated diagnosis	0.869	1.151
Gender	0.895	1.117
Age	0.632	1.583
Presence of comorbidities	0.543	1.842
Heart disease	0.693	1.444
Diabetes	0.736	1.358
Hypertension	0.748	1.337
Obesity	0.915	1.093
Peritonitis	0.639	1.565
Cancer	0.943	1.060

Table2. *Collinearity statistics*

$$\boldsymbol{VIF = \frac{1}{1 - R^2}}$$

Formula2. *VIF*

- **VIF:** Quantifies how much the variance of a regression coefficient is inflated due to multicollinearity. The authors suggested ten as their threshold value for high multicollinearity. In this study, all VIF values were well below this threshold, with the highest VIF being 1.56 for preoperative LOS.
- **Tolerance:** Tolerance is the reciprocal of VIF (Tolerance = 1/VIF) and indicates the proportion of variability in an independent variable that is not explained by other independent variables. Low tolerance values (less than 0.2, as defined by the authors) suggest high multicollinearity. In this study, all tolerance values were well above 0.1, confirming that multicollinearity is not a significant problem.

The implication of these calculations is correct to support this hypothesis. However, using VIF and tolerance may be redundant, and that tolerance is the reciprocal of VIF.

3. **Independence of residuals:** For this hypothesis, the authors performed the Durbin-Watson test to detect autocorrelation in the residuals. The results obtained were:

   - **Ho:** There is no autocorrelation in the residuals.
   - **Ha:** There is autocorrelation in the residuals.
   - **Statistic:** 1.505
   - **Acceptable Range by the authors:** [1.5-2.5].
   - **Decision:** Since the value of 1.505 falls within the acceptable, we do not reject the null hypothesis
   - **Conclusion:** There is no significant autocorrelation in the residuals, indicating that the residuals are independent.

   Using this hypothesis test is fit for purpose; if you want to analyse this visually, a 'Fitted values vs residuals' graph would be appropriate, but it is not necessary once the test is used.

4. **Homoscedasticity of the residuals**: To analyse this hypothesis, the authors made the graph 'standardised residuals vs the standardised predicted value', **Figure 1**. This graph allows us to observe the behaviour of the residuals where the variance changes across different levels of predicted values.

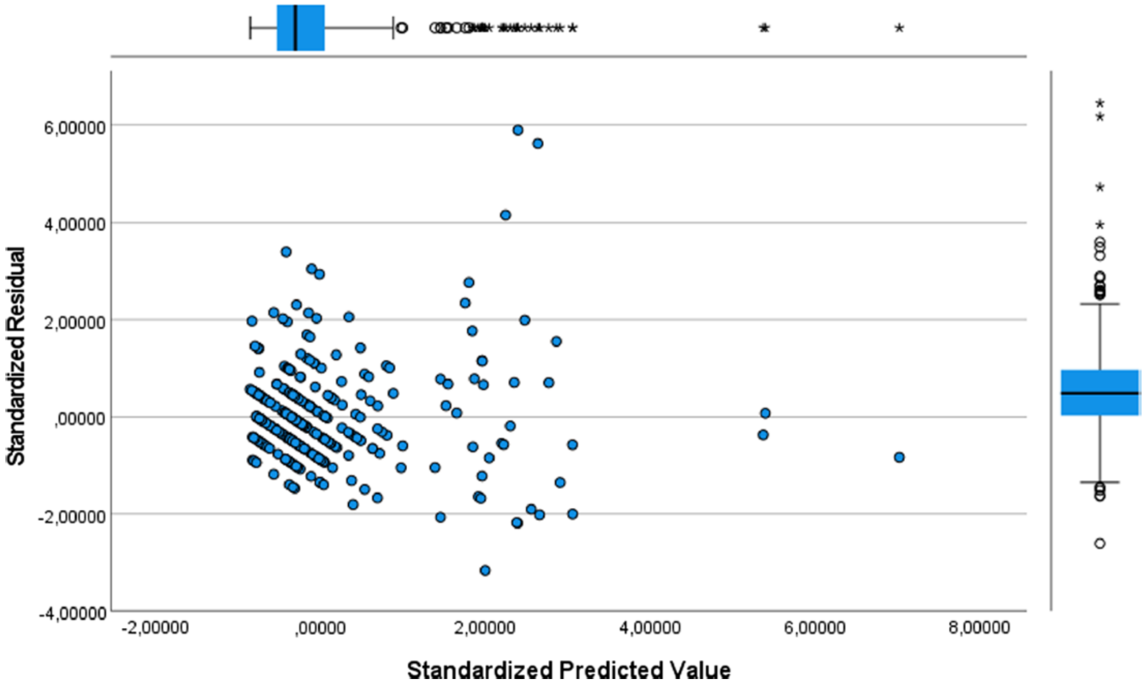


***Figure1**. Plot of "standardized residuals" against the "standardized predicted value"*

The graph shows that this hypothesis is not fulfilled, as the variance of the residuals tends to increase as the predicted value increases. However, the authors took this as a slight violation, and it was accepted that the model should be continued. No further tests other than the visual test were performed.

Although the graph is telling because it shows a possible violation of the hypothesis, performing a hypothesis test, such as Breusch-Pagan, would be a better option to analyse this hypothesis.

5. **Normality of the residuals:** For this hypothesis, the authors chose to use another visual method, the Q-Q plot, which compares the quantiles of the residuals with the expected quantiles of a normal distribution. If the points lie on the reference line, their distribution is normal. **Figure2** indicates that the residuals have an approximately normal distribution, as most of the points lie close to the diagonal reference line, a deviation can be seen in the tails, which can be a problem, but because most of the points lie very close to the line, the authors defined it to be a normal distribution. No additional tests other than visual testing were done.

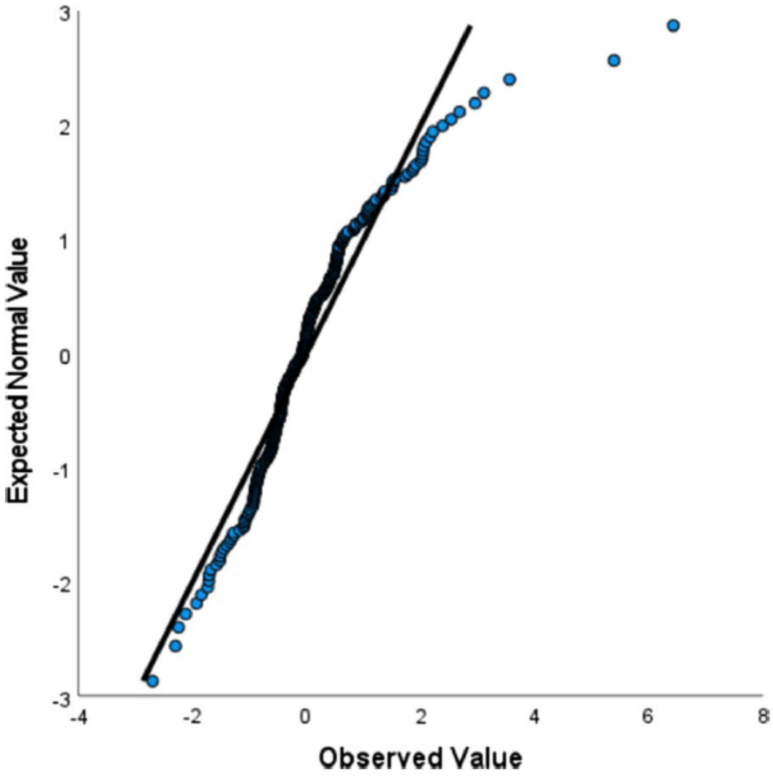
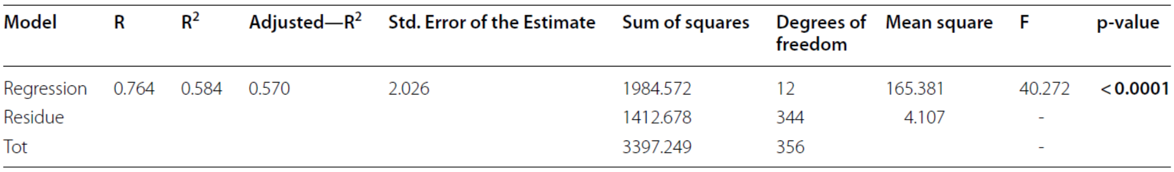


Figure 2. *Normal Q-Q Plot of Standardized Residual*

In the same way as the previous hypothesis, for greater certainty that this hypothesis is fulfilled, the Shapiro-Wilk test for the normality of the data should be carried out.

6. **Presence of outliers:** The authors used Cook's distance method to assess the presence of outliers. Cook's distance is a diagnostic measure to identify data points that substantially influence the estimated regression coefficients. No specific results of this test are presented in the article, but the authors mention that the maximum value found was 0.8, which is lower than the three-hold of 1, concluding that no outliers are present.

Once the six hypotheses were confirmed, the model was developed, and its performance was tested using an ANOVA and a Fisher's test. These results can be seen in **Table 3.**



Model	R	R^2	Adjusted—R^2	Std. Error of the Estimate	Sum of squares	Degrees of freedom	Mean square	F	p-value
Regression	0.764	0.584	0.570	2.026	1984.572	12	165.381	40.272	**< 0.0001**
Residue					1412.678	344	4.107	-	
Tot					3397.249	356		-	

Table 3. *Model summary and Fisher's exact test*

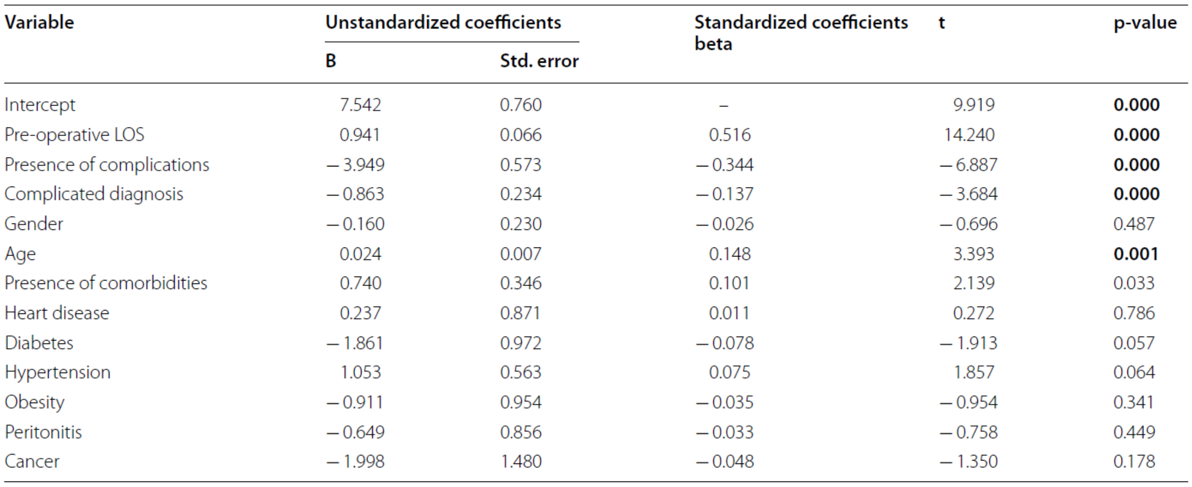
The F-test is used to determine whether there is a significant relationship between the dependent variable and the independent variables. The statistic is calculated by comparing the MSR of the model vs. the MSE of the residuals.

Ho: All regression coefficients are equal to zero.
Ha: At least one regression coefficient is not equal to zero.
Statistic: 40.272
p_value: <0.0001
Decision: The null hypothesis is rejected since the p_value < ($\alpha = 0.05$).
Conclusion: At least one of the model coefficients are different to zero.

This test makes it clear that the model performed by the authors is significant, following this, the ANOVA results indicate that:

- R of 0.764 indicates a strong correlation between observed and predicted values.
- A coefficient of determination R2 of 0.584 means that the model explains 58.4% of the variance.
- Standard Error of the Estimate: 2.026, indicating the average distance between observed values and the regression line.
- The sum of Squares: 1984.572 for regression and 1412.678 for residuals, respectively, indicating the explained and unexplained variance.
- MSR: 165.381 represents the average variance explained by each predictor in the regression model. A higher MSR indicates that the model explains a substantial portion of the variance in the dependent variable.
- MSE: 4.107 represents the unexplained variance in the model. A lower MSE indicates less unexplained variance, suggesting the model fits the data well.

Finally, having already built and tested a model to ensure that it is adequate, or at least significant, the article provides **Table 5**, which presents the regression coefficients, standard errors, t-values and p-values of the multiple linear regression model. With this, it is possible to understand the model and its predictors to define which are significant in predicting LOS and the magnitude and direction affecting these predictors.



Variable	Unstandardized coefficients		Standardized coefficients beta	t	p-value
	B	Std. error			
Intercept	7.542	0.760	–	9.919	**0.000**
Pre-operative LOS	0.941	0.066	0.516	14.240	**0.000**
Presence of complications	−3.949	0.573	−0.344	−6.887	**0.000**
Complicated diagnosis	−0.863	0.234	−0.137	−3.684	**0.000**
Gender	−0.160	0.230	−0.026	−0.696	0.487
Age	0.024	0.007	0.148	3.393	**0.001**
Presence of comorbidities	0.740	0.346	0.101	2.139	0.033
Heart disease	0.237	0.871	0.011	0.272	0.786
Diabetes	−1.861	0.972	−0.078	−1.913	0.057
Hypertension	1.053	0.563	0.075	1.857	0.064
Obesity	−0.911	0.954	−0.035	−0.954	0.341
Peritonitis	−0.649	0.856	−0.033	−0.758	0.449
Cancer	−1.998	1.480	−0.048	−1.350	0.178

Table 5. *Standardised and Unstandardized coefficients with p-values of the MLR analysis*

Table 5 presents the unstandardised coefficients, explaining the magnitude and direction of the predictor's relationship with LOS, and the standardised coefficients (based on the standard deviation), explaining how strong the relationship is.

The intercept tells us that LOS equals 7.542 when all other predictors are 0, being a considerable base value. By analysing its p_value, we can understand that this is significant for the model.

As for the predictors, the variables gender, Heart disease, Diabetes, hypertension, Obesity, Peritonitis and Cancer present p_values > 0.05, so they are not significant for the model.

Among the variables that do significantly affect the model are those that affect LOS positively:

- **Pre-operative LOS** (B = 0.941, p = 0.000): Each additional day in pre-operative LOS increases the total LOS by 0.941 days, making it a strong and significant predictor. According to the authors, this result was expected, given that this variable is directly related to LOS.
- **Age** (B = 0.024, p = 0.001): Each additional year increases LOS by 0.024 days, showing a significant positive effect.
- **Presence of Comorbidities** (B = 0.740, p = 0.033): Comorbidities increase LOS by 0.740 days, a significant predictor.

And variables with negative interaction with LOS:

- **Presence of Complications** (B = -3.949, p = 0.000): Complications decrease LOS by 3.949 days, indicating a significant negative impact.
- **Complicated Diagnosis** (B = -0.863, p = 0.000): A complicated diagnosis reduces LOS by 0.863 days, which is also a significant negative predictor.

Finally, the article describes the construction of a model with an R2 greater than 0.5, thus showing support for the authors of its usefulness and straightforward interpretation by medical staff.

However, this result also shows that the model cannot explain more than 40% of the variance, possibly due to the flaws in the model design mentioned above.

Conclusion:

The study by Teresa Angela Trunfio et al. provides valuable insight into the factors influencing length of hospital stay (LOS) in patients undergoing laparoscopic appendectomy. Using a multiple linear regression (MLR) model, significant predictors such as pre-operative LOS time, age and presence of comorbidities were identified, explaining 58.4% of the variance in LOS.

However, the study has significant limitations. Including the variable pre-operativepre-operative LOS, which is directly related to LOS, could induce collinearity and affect the stability of the model. In addition, not considering interactions between predictors may have omitted essential relationships. Also, the limited focus on specific comorbidities may restrict the model's ability to capture all variance explained by other medical conditions.

Several improvements are suggested: expanding the set of comorbidities, replacing the pre-operative LOS variable with measures not intrinsically related to LOS, and including interaction terms in the model. In addition, further checking the assumptions of the model, such as homoscedasticity and normality of the residuals, by additional tests, such as Breusch-Pagan and Shapiro-Wilk, would strengthen the validity of the results.

In summary, while the study is valuable, addressing these limitations and improvements may provide a more accurate and helpful application of the results in hospital management and clinical decision-making, thereby optimising hospital resources and improving patient care by improving the model to have an R2 that better explains the variance of LOS.

Reference

Trunfio, T. A., Bertolino, G., Liguori, A., & Colosimo, C. (2022). Multiple regression model to analyze the total LOS for patients undergoing laparoscopic appendectomy. BMC Medical Informatics and Decision Making, 22(1), 1-10. https://doi.org/10.1186/s12911-022-01764-5