**Introduction**

Este informe profundiza en la investigación realizada por Teresa Angela Trunfio et al., detallada en su publicación de 2022 en “BMC Medical Informatics and Decision Making,” titulada «Multiple Regression Model to Analyze the Total LOS for Patients Undergoing Laparoscopic Appendectomy». El objetivo del articulo es la implementación de una regresión lineal multiple para predecir “the Length Of Stay (LOS)” de los pacientes sometidos a apendicectomía laparoscópica, para este estudio, los autores usaron un conjunto de datos que contiene detalles demográficos y clínicos de 357 pacientes del Hospital Universitario «San Giovanni di Dio e Ruggi d'Aragona» en Salerno, Italia. Previo al desarrollo del modelo los autores buscaron verificar las siguientes 6 hipotesis:

1. Relación lineal entre las variables, mediante grafico de dispersión
2. Ausencia de milticolinearidad, mediante el calculo de Tolerancia y VIF
3. Independencia de los esiduales, mediante el test de Durbin-Watson
4. Residuales constantes, mediante un grafico "standardized residuals" against the "standardized predicted value"
5. Normalidad de los residuales, usando un grafico QQ
6. Presencia de outliers usando Cook’s distance, con un treshold de 1.

“La LOS -medida en días- se define como la diferencia entre la fecha de ingreso y la fecha de alta del paciente” Trunfio et al. (2022), de acuerdo a los autores esta medida es usada como indicador de eficiencia ya que ayuda a la planeación de ingresos de pacientes, impactando directamente la organización y costos. Es por esta razón, que la predicción de esta variable para diferentes pronósticos y procedimientos se vuelve de gran importancia para el sector hospitalario, en este estudio en especifico, para el procedimiento de apendicectomia por lamparoscopia; un procedimiento quirúrgico ampliamente utilizado en casos de apendicitis.

La importancia de este estudio, según Trunfio et al. (2022), radica en su potencial para mejorar la eficiencia operativa y la atención al paciente en los hospitales mediante la modelización predictiva de la duración de la estancia (LOS). Al predecir con exactitud la LOS, los centros sanitarios pueden optimizar su dotación de personal y la asignación de camas, reduciendo así los tiempos de espera innecesarios y disminuyendo potencialmente los costes sanitarios. Trunfio et al. subrayan que estas capacidades predictivas son esenciales para mejorar la asignación de recursos y la planificación operativa, sobre todo en el contexto de cirugías de urgencia impredecibles como las apendicectomías. Además, la identificación de variables clave que influyen significativamente en la LOS -como la edad, la duración de las intervenciones y las complicaciones- proporciona información esencial que puede ayudar a adaptar las prácticas de atención al paciente a las necesidades individuales y, en última instancia, mejorar los resultados y la satisfacción del paciente. Este modelo no sólo ayuda en la gestión operativa, sino que también respalda la toma de decisiones clínicas mediante la predicción de posibles complicaciones y tiempos de recuperación (Trunfio et al., 2022).

**Experiment design:**

Como se mencionó previamente, para el estudio los autores utilizaron datos de 357 pacientes extraídos del sistema QuaniSDO, un sistema integral de gestión de información hospitalaria del Hospital Universitario 'San Giovanni di Dio e Ruggi d'Aragona' en Salerno, Italia. La información extraída incluye datos demográficos y clínicos de los pacientes, tales como: género, edad, y grupo relacionado con el diagnóstico (Diagnostic Related Group, DRG), así como detalles específicos del procedimiento, como la presencia o ausencia de complicaciones, y las fechas de admisión y del procedimiento de laparoscopia (LC). Además, el tiempo preoperatorio de LC fue calculado mediante la fórmula: fecha del procedimiento LC – fecha de admisión; y el LOS también fue calculado. Del DRG se extrajeron datos adicionales relacionados con comorbilidades, incluyendo:

* **Presencia de comorbilidades** (sí / no)
* **Enfermedad cardíaca** (sí / no)
* **Diabetes** (sí / no)
* **Hipertensión** (sí / no)
* **Obesidad** (sí / no)
* **Peritonitis** (sí / no)
* **Cáncer** (sí / no) En la **Tabla 1** se pueden encontrar el diccionario de las variables usadas por los autores junto a su tipo de dato.

Tabla1

Estas variables elegidas representan una gran ventaja según los autores para predecir el LOS, pero además de esto, no suponen ninguna dificultad a la hora de modelar, ya que solo es necesario realizar transformaciones sencillas a las variables categóricas, sin necesidad de crear nuevas columnas. La cuidadosa selección de estas variables fortalece la robustez del modelo estadístico y su capacidad para predecir eficazmente el LOS, proporcionando así insights valiosos para la gestión hospitalaria y la planificación de recursos

El modelo que los autores escogieron fue el de regresión lineal multiple (MLR) para predecir LOS, este modelo lo construyeron usando el software IBM SPSS (Statistical Package for Social Science) ver.27.

El modelo usado para este estudio esta representado en la formula1.

En donde y representa la variable a predecir, LOS, β 0 el intercepto del modelo, xi y βi representan las 12 variables independientes con sus coeficientes (pre-operative LOS, presence of complications, complicated diagnosis, gender, age, presence of comorbidities, heart disease, diabetes, hypertension, obesity, peritonitis and cancer) y ε representa el error del modelo. Este modelo incluye tanto variables numéricas como categóricas, considerando su impacto directo en la predicción del LOS sin incluir interacciones entre estas variables.

El planteamiento de este modelo presenta ciertas limitaciones, la mas importantes son:

* Como los mismos autores lo mencionan, tienen valores muy limitados de las comorbolidades específicas, puede que otras comorbilidades expliquen mas la varianza del modelo y esto no se esta captando.
* El uso de la variable Pre-operative LOS: esta vriable esta ligada directamente a LOS, esto puede inducir colinearidad al modelo afectando la estabilidad del modelo, además de que esta en una variable que solo se conoce al momento de ingresar a procedimiento, por lo cual no aporta ningún valor logístico al hospital.
* El modelo no tiene en cuenta interacion entre los predictores, esto puede hacer que el modelo omita relaciones importantes entre los predictores. Por ejemplo, la combinación de ciertas comorbilidades y su impacto conjunto en el LOS podría ser significativo pero no detectado en el análisis actual.

Alternativas para poder mejorar estas limitaciones son:

* Ampliar la extracción de datos, en especifico del apartado de comorbilidades, se puede basar en la literatura y extraer las que se crea que son mas influyentes en LOS
* Cambiar el calculo de LOS por, date of discharge – date of LC procedure, de esta forma la variable de Pre-operative LOS no estará ligada al valor final de LOS de forma directa y el modelo podrá captar la verdadera importancia de esta variable.
* Crear términos de interacción entre variables que se sospecha puedan tener efectos conjuntos significativos. Por ejemplo, incluir interacciones entre diferentes comorbilidades (como diabetes y hipertensión) y entre factores demográficos y comorbilidades (como edad y presencia de enfermedades cardíacas).

En términos del modelo seleccionado, MLR, es apropiada para este tipo de análisis dado su poder para manejar múltiples variables independientes y su capacidad para cuantificar el efecto de cada predictor sobre la variable dependiente (LOS). Además, la MLR es relativamente sencilla de interpretar y aplicar utilizando software estadístico como IBM SPSS. Sin embargo, podría considerarse el uso de modelos más para situaciones donde el resultado no sea linealmente relacionado con los predictores.

**Statistical análisis:**

Como se menciono previamente, los autores buscaron verificar 6 hipotesis antes de crear el modelo:

1. **Linearidad:** Los autores plantean verificar esta hipótesis mediante el uso de graficas de dispersión. Esta metodologiaes adecuada a la horade analizar esta hipótesis, sin embargo, en el paper no se encuentra ninguna evidencia de estas graficas. Estas graficas son de extrema utilidad, ya que nos permiten identificar la relación que cada variable tiene con LOS, además de que nos indica si es o no necesario realizar transformaciones, no tener las graficas evita extraer este tipo e información.
2. **Multicolinearidad:** para asegurarse que no hubieran problemas de multicolinearidad en los datos, los autores usaron 2 metodos, mediante la formula2, calcular el variance inflation ratio (VIF) y también calcularon la tolerancia (1-R2). En la **tabla2** se proporcionan los resultados obtenidos por los autores:

**A table with numbers and text

Description automatically generated with medium confidence**

**Formula VIF**

* VIF: Cuantifica cuánto se infla la varianza de un coeficiente de regresión debido a la multicolinealidad. Los autores suaron 10 como su valor limite a una alta multicolinealidad. En este estudio, todos los valores de VIF resultaron estar muy por debajo de este umbral, siendo el VIF más alto de 1,56 para la LOS preoperatoria.
* Tolerancia: La tolerancia es el recíproco del VIF (Tolerancia = 1/VIF) e indica la proporción de variabilidad en una variable independiente que no es explicada por otras variables independientes. Los valores bajos de tolerancia (menos de 0,2 definido por los autores) sugieren una alta multicolinealidad. En este estudio, todos los valores de tolerancia fueron muy superiores a 0,1, lo que confirma que la multicolinealidad no es un problema importante.

La implelmentacion de estos cálculos es correcto para porbar esta hipotesisi, sin embargo, el uso de VIF y tolerance puede ser redundante y que te tolerance es el reciproco de VIF.

1. **Independencia de los residuales:** Para esta hipótesis, los autores realizaron la prueba de Durbin-Watson para detectar autocorrelación en los residuos. Los resultados obtenidos fueron:
   * **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 don’t reject the null hypothesis
   * **Conclusion:** There is no significant autocorrelation in the residuals, indicating that the residuals are independent.

El uso de esta prueba de hipótesis es adecuado con el objetivo, si se quiere analizar de forma visual esto un grafico de “Fitted values vs residuals” seria apropiado, pero no es necesario una vez usado el test.

1. **Homocedastisidad de los residuales:** Con el objetivo de analizar esta hipótesis, los autores realizaron la grafica “standardized residuals vs the standardized predicted value”, Figura1. Esta grafica permite observar el comportamiento de los residuales donde la varianza cambia a través de diferentes niveles de valores predichos.

A graph with blue dots

Description automatically generated

**Fig1**

En la grafica se observa que este hipótesis no se cumple ya que la varianza de los residuales tiene a aumentar a medida que aumenta el valor predicho, pero los autores tomaron estaviolacion como leve y fue aceptada para continuar con el modelo. Nos e realizaron mas pruebas además de la visual.

A pesar de que la grafica sea diciente, debido a que se observa una posible violación de la hipótesis, realizar una prueba de hipótesis, como Breusch-Pagan, seria una mejor opción para analizar realmente esta hipótesis.

1. **Normalidad de los residuales:** para esta hipotesisi los autores obtaron por usar otro método visual, grafico Q-Q, el cual compara los cuantiles de los residuos con los cuantiles esperados de una distribución normal, si los puntos se sitúan en la línea de referencia, su distribución es normal. La Fig2 indica que los residuos tienen una distribución aproximadamente normal, ya que la mayoría de los puntos se sitúan cerca de la línea diagonal de referencia, se pueden ver una desviación en las colas lo cual puede suponer un problema, pero debido a que la mayoría de puntos se sitúa muy cerca de la línea, los autores definieron que es una distribución normal. No se hizo ninguna prueba adicional a la visual.

A graph with blue dots

Description automatically generated  
Fig2

De igual forma que la anterior hipótesis, para una mayor seguridad de que esta hipótesis si se cumpla, se debió realizar el test de shapiro wilk para la normalidad de los datos.

1. **Presencia de outliers:** El método de distancia de Cook fue el empleado por los autores para poder evaluar la presencia de estos valores. La distancia de Cook es una medida de diagnóstico utilizada para identificar los puntos de datos que tienen una influencia sustancial en los coeficientes de regresión estimados. En el articulo no se presenta ningún resultado especifico de esta prueba, pero lo autores mencionan que el valor máximo encontrado fue de 0.8, el cual es inferior al treshold de 1, concluyendo que no existe presencia de valores atípicos.

Una vez confirmadas las 6 hipotesis, se desarrollo el modelo anteriormente planteado y mediante una ANOVA y un test de Fisher se probó su desempeño. Estos resultados se pueden ver en la **tabla 3**

A screenshot of a computer

Description automatically generated  
Table3.

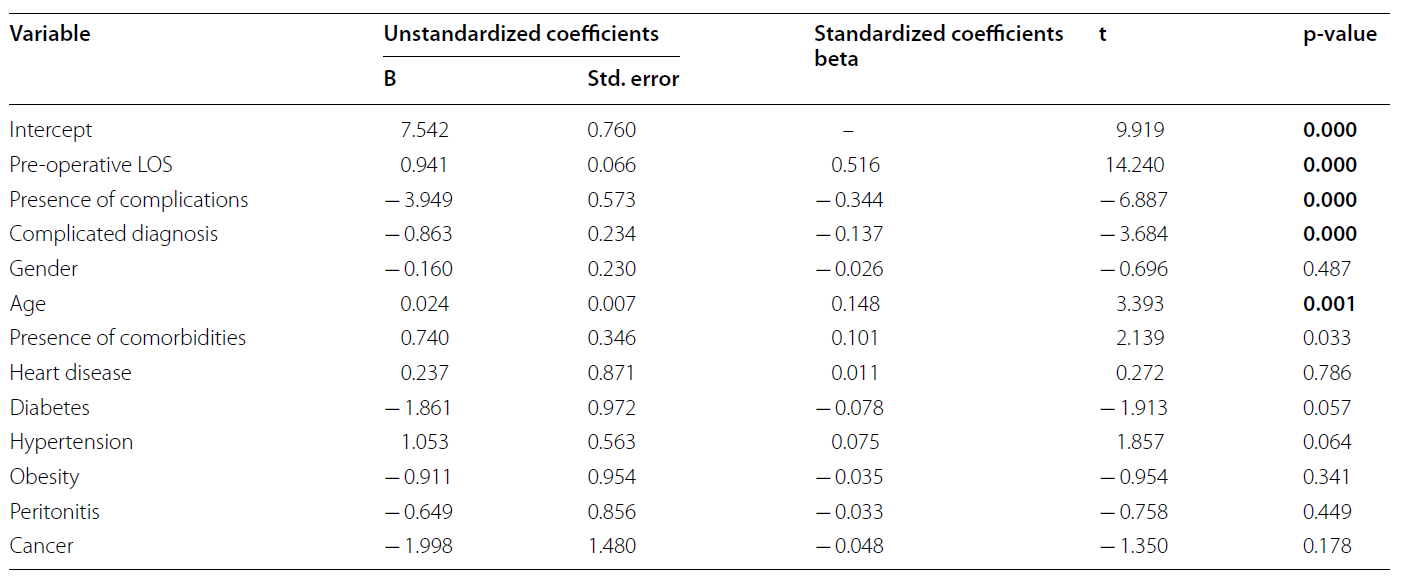
La prueba F se utiliza para determinar si existe una relación significativa entre la variable dependiente y las variables independientes. El estadistico es calculado comparando el MSR del modelo vs MSE de los residuales.

**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 < (α = 0.05).  
**Conclusion:** At least one of the model coefficients are different to zero.

Esta prueba da claridad que el modelo realizado por los autores es significativo, siguiendo esto, los resultados del ANOVA indican que:

* 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.
* 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 that there is less unexplained variance, suggesting that the model fits the data well.

Por ultimo, ya teniendo un modelo construido el cual ya le hicieron las pruebas correspondientes para asegurarse que este si sea adecuado, o como mínimo significativo, el articulo proporciona la tabla 5, la cual presenta los coeficientes de regresión, los errores estándar, los valores t y los valores p del modelo de regresión lineal múltiple. Con esto es posible entender el modelo y sus predictores para definir cuáles son significativos a la hora de predecir LOS, además de comprender la magnitud y dirección, las cuales afectan estos predictores.



**Table 5**

La tabla5 nos presenta los coeficientes sin estandarizar, explicando la magnitud y dirección de la relación del predictor con LOS, y estandarizados (basándose en la desviación estándar), explicando que tan fuerte es la relación.

El intercepto nos indica que LOS es igual a 7.542, cuando el resto de predictores son 0, siendo un valor base grande. Analizando su p\_value damosa entender que este es significativo para el modelo.

En cuanto a los predictores, las variables: gender, Heart diseas, Diabetes, Hyperthension, Obesity, Peritonitis y Cancer, presentan p\_values > 0.05, por lo que no son significativos para el modelo.

Entre las variables que si afectan significativamente el modelo están las que afectan LOS de forma positiva:

* **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 of age increases LOS by 0.024 days, showing a significant positive effect.
* **Presence of Comorbidities** (B = 0.740, p = 0.033): The presence of comorbidities increases LOS by 0.740 days, a significant predictor.

Y las variables con dirección negativa a LOS:

* **Presence of Complications** (B = -3.949, p = 0.000): The presence of complications decreases 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.

Finalmente, el articulo describe la construcción de un modelo con un R2 superior a 0.5, por lo que muestra soporte para los autores de su utilidad, junto a su fácil interpretación por el staff del equipo medico.

**Conclusión**

El estudio de Teresa Angela Trunfio et al. proporciona una comprensión valiosa sobre los factores que influyen en la duración de la estancia hospitalaria (LOS) en pacientes sometidos a apendicectomía laparoscópica. Utilizando un modelo de regresión lineal múltiple (MLR), se identificaron predictores significativos como el tiempo preoperatorio, la edad y la presencia de comorbilidades, explicando el 58.4% de la varianza en el LOS.

No obstante, el estudio presenta limitaciones importantes. La inclusión de la variable pre-operative LOS, directamente relacionada con el LOS, podría inducir colinealidad y afectar la estabilidad del modelo. Además, no considerar interacciones entre predictores puede haber omitido relaciones importantes. También, el enfoque limitado en ciertas comorbilidades puede restringir la capacidad del modelo para capturar toda la varianza explicada por otras condiciones médicas.

Se sugieren varias mejoras: ampliar el conjunto de comorbilidades consideradas, reemplazar la variable pre-operative LOS por medidas no intrínsecamente relacionadas con el LOS, e incluir términos de interacción en el modelo. Además, verificar más exhaustivamente los supuestos del modelo, como la homocedasticidad y la normalidad de los residuos, mediante pruebas adicionales como Breusch-Pagan y Shapiro-Wilk, fortalecería la validez de los resultados.

En resumen, aunque el estudio es valioso, abordar estas limitaciones y mejoras puede proporcionar una aplicación más precisa y útil de los resultados en la gestión hospitalaria y la toma de decisiones clínicas, optimizando así los recursos hospitalarios y mejorando la atención al paciente, esto mejorando el modelo buscando tener un R2 que explique mejor la varianza de LOS.

**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**

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.