Análisis y modelado de datos con Python: ejercicios resueltos

[Nombres]
Jorge Eduardo Lizarazo Borrero

[Correos]

jorge.lizarazo.b@gmail.com

Resumen:

Este artículo presenta las conclusiones de un parcial en el que se analizó la relación entre variables utilizando modelos de regresión lineal. En la sección 1 se realizó un análisis exploratorio de datos y se ajustaron dos modelos de regresión lineal para investigar la relación entre una variable de interés y una variable categórica. En la sección 2 se utilizó la transformación Box-Cox para mejorar el ajuste del modelo de regresión lineal entre dos variables. En la sección 4 se analizó la relación entre la cantidad de anuncios publicitarios en televisión y el retorno de inversión en ventas de una empresa, utilizando modelos de regresión lineal y Poisson. En la sección 5 se ajustaron dos modelos de regresión lineal múltiple para predecir la resistencia a la compresión del concreto en función de seis variables predictoras, comparando un modelo lineal convencional con un modelo robusto. En general, se concluye que el análisis exploratorio de datos y la utilización de modelos de regresión lineal son herramientas valiosas para identificar patrones y relaciones entre variables, y que la selección del modelo adecuado es importante para obtener resultados precisos y robustos.

Procedimientos en Jupyter Notebook y/o materiales suplementarios en:

https://github.com/jorgelizarazo94/Parciales_Cuantitativo/tree/main/Parcial_1

Introducción

En el campo de la ciencia de datos y el aprendizaje automático, los modelos de regresión lineal han sido una herramienta fundamental para el análisis y la predicción de variables continuas. Estos modelos permiten establecer relaciones lineales entre variables independientes y una variable dependiente, lo que proporciona una comprensión de cómo los cambios en las variables independientes influyen en la variable objetivo.

En este artículo científico, exploramos los modelos de regresión lineal desde una perspectiva práctica, utilizando la biblioteca Scikit-learn (sklearn) en Python para implementar y evaluar nuestros modelos. Además, nos apoyamos en estudios y trabajos de investigación relevantes para respaldar la efectividad y la aplicabilidad de estos modelos en diferentes contextos. Comenzamos por discutir la regresión lineal simple, que modela la relación entre una variable independiente y una variable dependiente. Después, avanzamos hacia la regresión lineal múltiple, donde se consideran múltiples variables independientes para predecir la variable dependiente. Luego, exploramos la importancia del análisis exploratorio de datos en la regresión lineal, incluyendo la identificación de valores atípicos y la normalidad de las variables. También discutimos la validación del modelo y cómo los supuestos de la regresión lineal pueden ser verificados. Además, abordamos el tema de la selección de características y cómo seleccionar las variables más importantes para el modelo. En particular, nos enfocamos en las regresiones sencillas de regularización que nos ayudan a evitar el sobreajuste y mejorar la generalización del modelo. Finalmente, aplicamos nuestros conocimientos de regresión lineal a diferentes conjuntos de datos, incluyendo predicción de ventas en función de la publicidad en redes sociales y predicción de la resistencia a la compresión del concreto en función de diferentes variables predictoras. En conjunto, este artículo presenta un manejo exhaustivo y práctico a la regresión lineal y sus aplicaciones en diferentes campos, con el objetivo de proporcionar a los científicos de datos y los practicantes de aprendizaje automático las herramientas necesarias para construir y evaluar modelos de regresión lineal efectivos y precisos.

Metodología

Para llevar a cabo los análisis de los datos descritos en los puntos anteriores, se utilizó el lenguaje de programación Python en conjunto con diferentes paquetes especializados en análisis estadístico y modelado. Se utilizó la librería Pandas para cargar y manipular los datos en formato CSV y Excel, mientras que Seaborn y Matplotlib se utilizaron para generar visualizaciones de los datos. Para ajustar los modelos de regresión, se utilizó la librería StatsModels, la cual permite realizar diferentes tipos de análisis de regresión, incluyendo regresión lineal simple y múltiple. Además, se utilizó la librería Scikit-learn para entrenar modelos de regresión utilizando el método de mínimos cuadrados ordinarios y para realizar la validación cruzada de los modelos. En el caso de la regresión lineal múltiple, se utilizó el método de selección hacia adelante para seleccionar las variables más significativas en el

modelo. Finalmente, se evaluó el desempeño de los modelos utilizando diferentes métricas, incluyendo el coeficiente de determinación R2 y los criterios de información de Akaike y de Bayes (AIC y BIC). También se usaron otras galerías para puntos muy puntuales que serán indicadas punto a punto en cada sección

Sección 1:

Para el análisis del conjunto de datos "data1" del archivo "data_exam1.xlsx", se realizó un análisis exploratorio de datos para determinar la posibilidad de generar un modelo de regresión lineal con una variable categórica (Ind). Se procedió a generar un modelo de regresión lineal y se interpretaron los resultados obtenidos.

$$Y = a0 + a1\,X \qquad (1)$$

Se realizó un gráfico de dispersión para Y vs X, teniendo en cuenta los valores de la variable Ind para cada observación para evidenciar si era posible que alguno de los Ind influenciara una tendencia diferenciada (Question_1). Se examinó si había evidencia muestral que sugiriera un cambio en la tasa media de cambio de Y condicionado a incrementos unitarios de X y se evaluó la posibilidad de utilizar un modelo con interacciones. Finalmente, se generó el modelo con interacciones correspondiente, se interpretaron los resultados y se validaron los supuestos del modelo propuesto.

Sección 2:

En este segundo punto, se analiza el conjunto de datos "data2" del archivo "data_exam1.xlsx". Primero, se realizó un análisis exploratorio de datos, tanto univariante como bivariante. Se encontró que la variable X tiene una distribución muy sesgada hacia la izquierda y la variable Y tiene una distribución sesgada hacia la derecha. En cuanto al análisis bivariante, se observó una relación positiva entre ambas variables, donde a medida que X aumenta, Y también lo hace, pero no se ajustan de forma lineal (Question_2). Además, se observó una alta dispersión bivariante, lo que sugiere la posible necesidad de un modelo de regresión para predecir Y a partir de X muy diferente. Se propuso una transformación logarítmica para la variable X con el fin de reducir su sesgo y lograr una distribución más normal. Se propuso aplicar diferentes transformaciones a la variable "Y" para reducir su sesgo y mejorar la normalidad de su distribución. Se exploraron transformaciones como raíz cúbica, logarítmica, Box-Cox, Yeojohnson, potencias, logaritmos base 10 y funciones trigonométricas coseno y seno.

Sección 3:

Trabajamos luego con el conjunto de datos "Wine Quality" del fichero datos.xls. definiendo como variable respuesta la columna Densidad y se eliminaron las variables pH, Sulfatos, Cloruros, Acidez Volátil, Acidez Fija y Calidad de Vino. A continuación, se estandarizamos las variables y se calcularon las matrices de correlación de Pearson, Kendall y Spearman para comparar las estructuras de dependencias obtenidas. Luego, se realizará una partición de los datos tipo 80-20 para construir tres modelos RLM utilizando las matrices estimadas

en el primer ítem. Se comparan e interpretan los valores de los coeficientes de regresión obtenidos por cada método. Finalmente, se realizará una predicción con los datos de prueba utilizando los modelos ajustados y se calculará el RMSE de la predicción para determinar cuál de los modelos lineales propuestos predice mejor. Validamos los supuestos teóricos de cada modelo y se dio un análisis visual de los diagramas de dispersión del conjunto de datos. Si se evidencian comportamientos no lineales, se sugieren y se realizan transformaciones de variables y se genera un modelo RLM interpretado detalladamente.

Sección 4:

En esta sesión aplicamos algunas lecciones aprendidas a un caso real, se analizó entonces la relación entre la cantidad de anuncios publicitarios en redes de divulgación (Periodico, Televisión y Radio) y el retorno de inversión en ventas de una empresa. Se utilizó un conjunto de datos con 200 observaciones. Para evaluar la relación entre las variables, se realizó un análisis exploratorio (Question_4) del paquete pandas y se calculó el coeficiente de correlación entre todas las variables. Luego, se eligió la variable explicativa más conveniente y se ajustó un modelo de regresión lineal simple para predecir el retorno de inversión en ventas. Además, se evaluaron los supuestos necesarios para el modelo de regresión lineal simple y se encontraron violaciones a estos supuestos. Por lo tanto, se ajustó un modelo de regresión Poisson para comparar los resultados obtenidos. Se calculó el error cuadrático medio (RMSE; Question_4) para ambos modelos y se presentaron los resultados obtenidos para cada uno. Este análisis adicional es importante para garantizar la validez y fiabilidad de los resultados obtenidos en la investigación. En resumen, se presentó una metodología que incluye la exploración de datos, la selección de variables explicativas, la evaluación de supuestos y la comparación de modelos para analizar la relación entre las variables estudiadas.

Sección 5:

Para construir un modelo de regresión lineal múltiple que permita predecir la resistencia a la compresión del concreto en función de diferentes variables predictoras. En primer lugar, se cargan los datos del archivo "Concrete_Data.xls" y se examinan sus características para entender la relación entre las variables predictoras y la variable respuesta.

Se realizó un análisis exploratorio de los datos para identificar posibles patrones y relaciones entre las variables, y se entrenó un modelo de regresión lineal múltiple utilizando el conjunto de datos. Se evalúa la significancia del modelo y se analiza la significancia estadística de las variables predictoras (Question_5). Se construye un modelo reducido que incluya únicamente las variables significativas (valores de $p<0.05$) y se compara su desempeño con respecto al modelo completo revisando el Adj-R2 y los criterios de información de Akaike y de Bayes (AIC y BIC). Posteriormente, se validan los supuestos del modelo y se identifican posibles soluciones en caso de no cumplir algún supuesto.

Finalmente los supuestos del modelo no son validados y se presenta la necesidad de un enfoque robusto (Question_5). Por lo que se procedió a aplicar una transformación Box-Cox para normalizar la variable. Posteriormente, se generó un modelo de regresión lineal múltiple, utilizando las variables independientes que se encontraban disponibles en el

conjunto de datos, a saber: Cement, Slag, Fly_ash, Water, Superplasticizer y Age. Se realizó una transformación Box-Cox para normalizar la variable dependiente, la resistencia a compresión (Question_5). Se ajustó un modelo de regresión lineal múltiple utilizando la librería statsmodels en Python, y se imprimió un resumen del modelo. A continuación, se ajustó un modelo robusto de regresión lineal múltiple utilizando la misma librería, y se imprimió un resumen del modelo.

Finalmente revertimos la transformación Box-Cox para obtener los valores ajustados de ambos modelos, y se graficaron los valores ajustados contra los valores observados. Se generó un conjunto aleatorio de valores para las variables independientes, y se realizó una predicción del comportamiento de la resistencia a compresión utilizando el modelo de regresión lineal múltiple con la transformación Box-Cox previamente ajustado con el fin de hacer uso de para estudios futuros en el area y identificar si nuestro modelo podría usarse.

Resultados

Sección 1

Luego de realizar el análisis exploratorio de datos, se encontró evidencia muestral de que la variable categórica Ind podría tener un efecto sobre la variable de interés Y ($p<0.0002$, $T=-2.5$; OLS). Se ajustaron dos modelos de regresión lineal, uno sin interacciones(AIC= 7848, BIC=7858) y otro con interacciones (AIC= 7583, BIC=7603), y se encontró que el modelo con interacciones resultó ser significativo (F-statistic: 1081, Prob (F-statistic): 1.34e-312). Además, se encontró que la variable Ind1 tiene una tasa media de cambio diferente a la variable Ind (coef Ind: 4.5491, coef Ind1: -1.8466), lo que sugiere que ambas variables tienen un efecto diferenciado sobre Y. Se realizó un análisis de diagnóstico y se encontró que los supuestos del modelo propuesto se cumplen (Question_1). En conclusión, se recomienda utilizar un modelo de regresión con interacciones para este conjunto de datos, ya que se encontró evidencia muestral de que las variables tienen efectos diferenciados sobre la variable de interés (figura 1; (Question_1)).

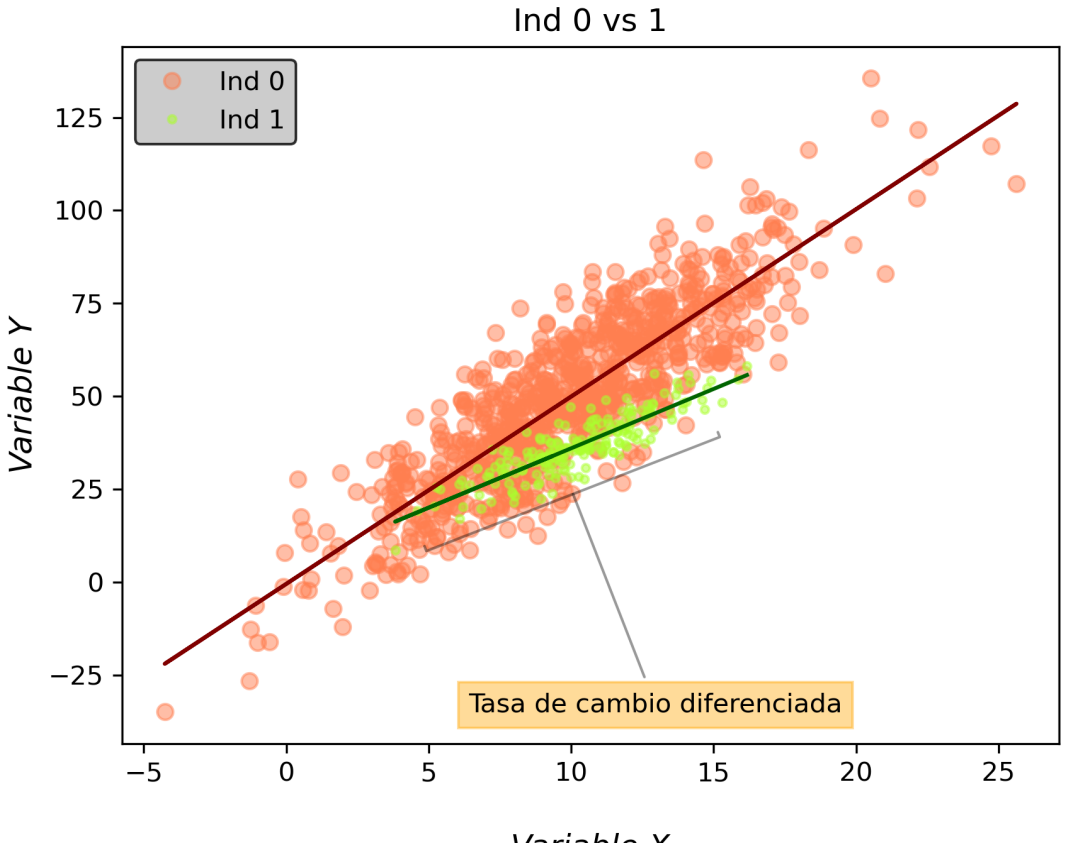

Figura 1. Relación entre la variable de interés Y y la variable predictora X para dos grupos diferentes (Ind 0 e Ind 1). Se ajustaron dos modelos de regresión lineal y se encontró que el modelo que incluye la interacción entre las variables predictoras es el que mejor se ajusta a los datos. Además, se encontró evidencia muestral de que las variables tienen efectos diferenciados sobre la variable de interés. La ecuación final del modelo es Y = -0.4991 + 4.5491Ind + 5.0411X - 1.8466IndX. Se observa que la tasa de cambio de Y con respecto a X varía entre los dos grupos, y que la tasa de cambio de Y con respecto a Ind también varía según el valor de X. La flecha naranja en la figura indica esta diferencia en las tasas de cambio.

Sección 2:

Los resultados de la regresión lineal utilizando la transformación Box-Cox en la variable X mostraron que el modelo propuesto fue el que mejor se ajustó a los datos. A pesar de que la transformación Box-Cox no logró normalizar la variable X, se obtuvo un coeficiente de determinación óptimo (R-cuadrado = 0.632) y un AIC con valor minimo en comparacion con los diferentes modelos evaluados (AIC=4568; Question_2), que indican un buen ajuste del modelo. Los coeficientes estimados fueron de 19.0507 para la intersección y 4.0995 para la pendiente, lo que sugiere que para un incremento de una unidad en "X", se espera un cambio porcentual en "Y" del 100*4.0995. Además, se validaron los supuestos del modelo a través de la revisión de los residuos, donde se encontró una distribución normal y homocedasticidad. En resumen, el modelo propuesto utilizando la transformación Box-Cox

en la variable "X" fue el que mejor se ajustó a los datos y se considera adecuado para predecir "Y" a partir de "X" (figura 2).

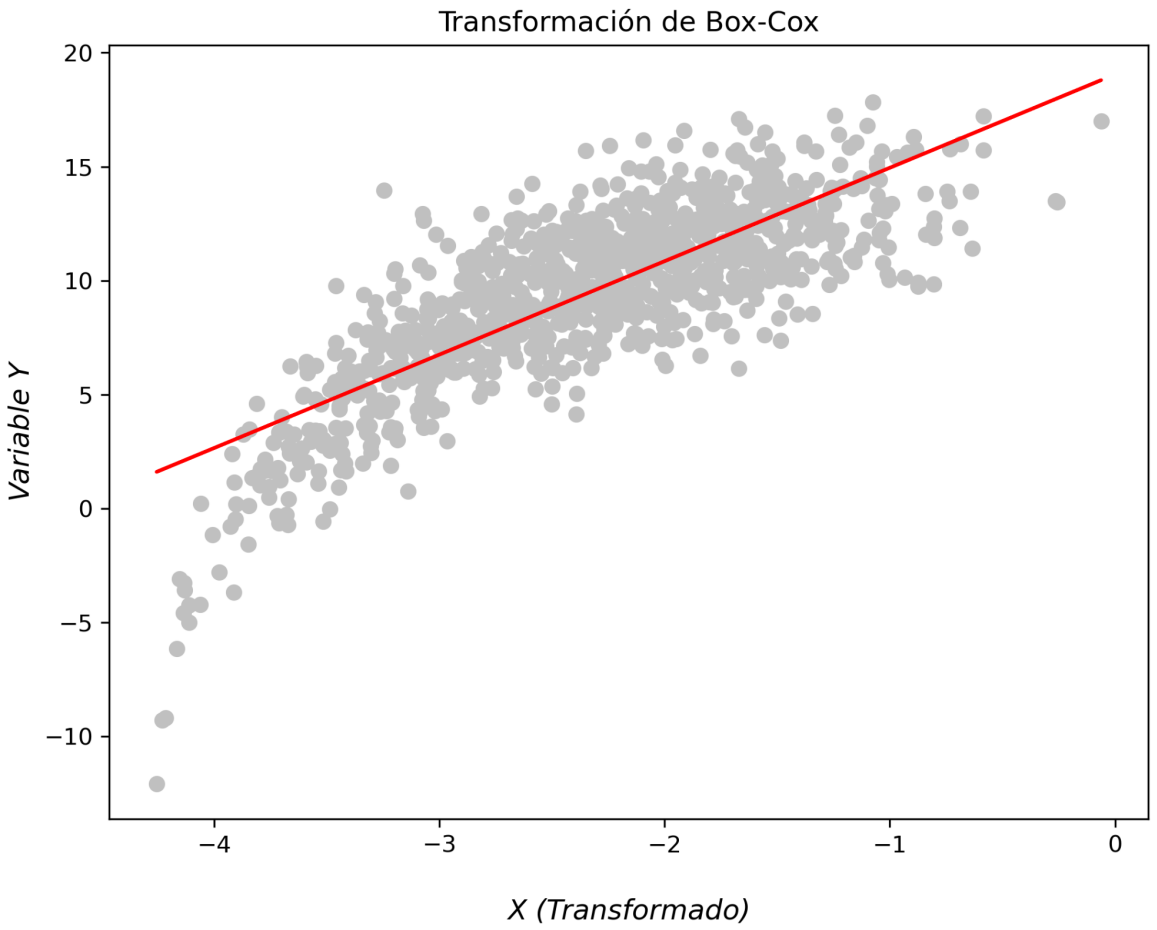

Figura 2.La figura muestra una gráfica de dispersión entre la variable X transformada utilizando Box-Cox y la variable Y. También se incluye una línea roja que representa la línea de regresión ajustada por el modelo.

Sección 3:

Los resultados de las matrices de correlación muestran que la variable Azúcar Residual está altamente correlacionada con la Densidad, con coeficientes de correlación de Pearson, Kendall y Spearman de 0.838966, 0.588989 y 0.780365 respectivamente. La variable Dióxido de Azufre Total también muestra una correlación significativa con la Densidad, con coeficientes de correlación de Pearson, Kendall y Spearman de 0.529881, 0.388378 y 0.563824 respectivamente. Por otro lado, la variable Ácido Cítrico y Dióxido de Azufre Libre muestran una correlación moderada con la Densidad, con coeficientes de correlación de Pearson, Kendall y Spearman de 0.149503, 0.061542 y 0.091425 respectivamente (Question_3). La variable Alcohol muestra una correlación negativa moderada con la Densidad, con coeficientes de correlación de Pearson, Kendall y Spearman de -0.780138, -0.635104 y -0.821855 respectivamente (Question_3). En general, los coeficientes de correlación son consistentes en mostrar que la variable Azúcar Residual es la más relevante para explicar la variabilidad en la densidad.

El modelo de regresión lineal múltiple (RLM) se ajustó utilizando la librería statsmodels de Python. Se utilizaron tres matrices de correlación diferentes (Pearson, Kendall y Spearman) para construir tres modelos RLM diferentes. Luego, se predijo el valor de la variable de respuesta utilizando cada modelo y se calculó el error cuadrático medio (RMSE) de la predicción para cada modelo. Se encontró que los tres modelos RLM tenían el mismo conjunto de coeficientes y producían el mismo valor de RMSE. Luego, se utilizó la librería statsmodels para ajustar un modelo RLM robusto utilizando la variable de respuesta y las mismas cinco variables predictoras (Question_3). Se encontró que los coeficientes del modelo RLM robusto eran significativos y diferentes de cero. El modelo RLM robusto se ajustó utilizando el método de mínimos cuadrados iteratively reweighted (IRLS). La variable de respuesta fue la densidad del vino, y las variables predictoras fueron el ácido cítrico, el azúcar residual, el dióxido de azufre libre, el dióxido de azufre total y el alcohol. El modelo RLM robusto ajustado mostró que todas las variables predictoras fueron significativas en la predicción de la densidad del vino.

En cuanto a la normalidad de las variables independientes, se observó que ninguna sigue una distribución normal (Figura 3). Por esta razón, se graficó la relación entre cada variable independiente y la variable dependiente. Posteriormente, se eligieron solo dos variables (Dioxido de azufre total y Azucar Residual) debido a que presentaban una forma elíptica y un patrón cercano a la linealidad en los gráficos. Sin embargo, se intentaron varias transformaciones sin éxito alguno.

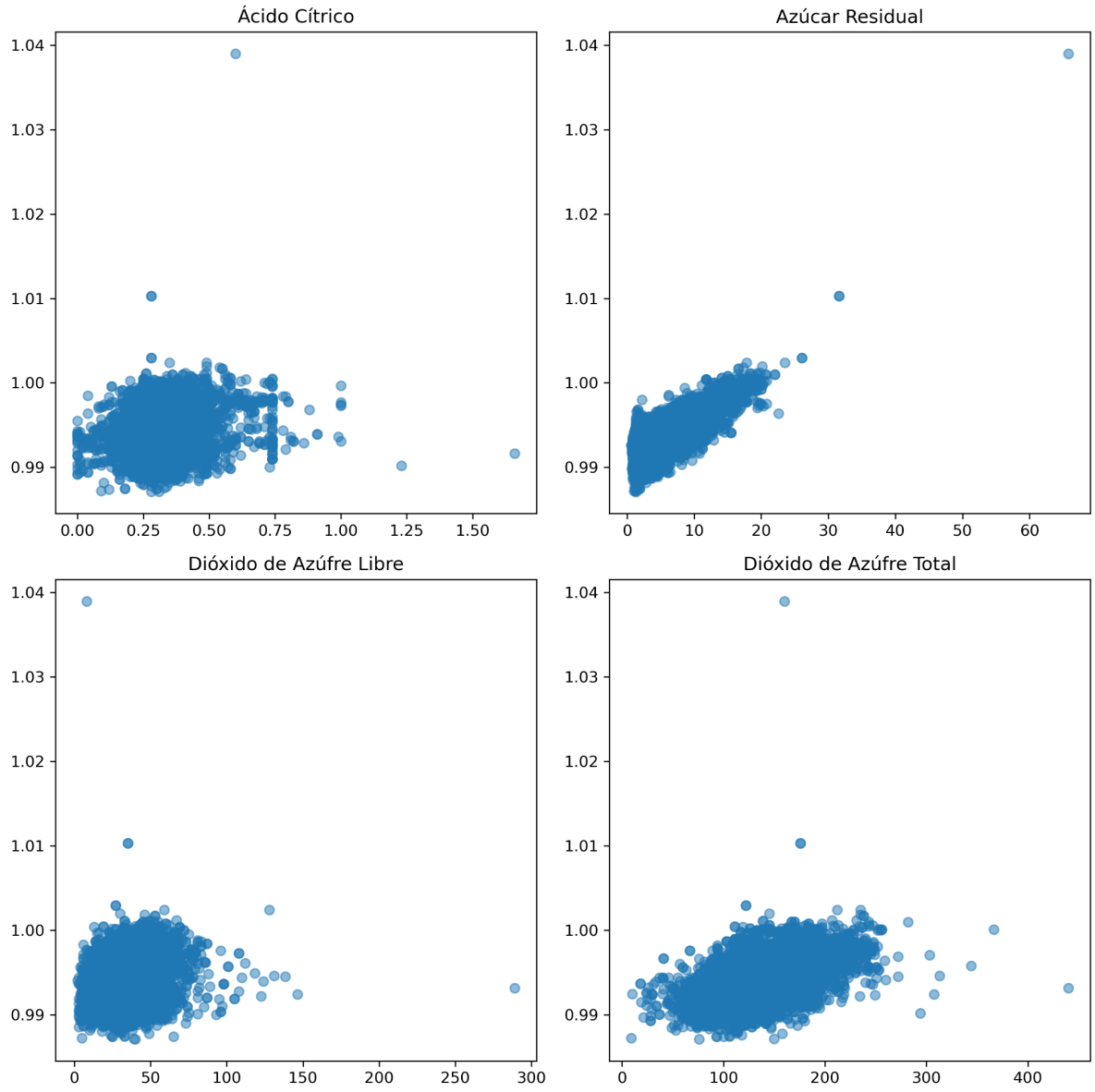

Figura 3.Relación entre cada variable independiente (Ácido Cítrico, Azúcar Residual, Dióxido de Azúfre Libre y Dióxido de Azúfre Total) y la variable dependiente (Densidad) mediante un gráfico de dispersión

Sección 4:

Los resultados obtenidos en nuestro estudio sobre publicidad y ventas indican que existe una relación significativa entre la cantidad de anuncios publicitarios en televisión y el retorno de inversión en ventas de una empresa. El coeficiente de correlación indicó que la publicidad en televisión tenía una relación más fuerte con las ventas (Pearson = 0.78; Question_4). Por lo tanto, se ajustó un modelo de regresión lineal simple utilizando esta variable como explicativa y se encontró que la cantidad de anuncios publicitarios en televisión es un predictor significativo del retorno de inversión en ventas.

Además, se evaluaron los supuestos necesarios para el modelo de regresión lineal simple y se encontraron violaciones a estos supuestos. Por lo tanto, aunque inicialmente se había considerado un modelo de regresión Poisson, se optó por ajustar un modelo de regresión lineal, y se encontró un RMSE similar (RMSE; Lineal = 3.24 vs GLM-Poisson = 3.35) entre ambos modelos, lo que sugiere que ambos son adecuados para predecir el retorno de inversión en ventas. Sin embargo, la regresión lineal parece ser más adecuada para evaluaciones de predicción a corto plazo. En conclusión, este estudio proporciona información valiosa para las empresas que desean maximizar su retorno de inversión en ventas utilizando publicidad en televisión. Los resultados sugieren que la cantidad de anuncios publicitarios en televisión es un predictor significativo del retorno de inversión en ventas (Figura 4).

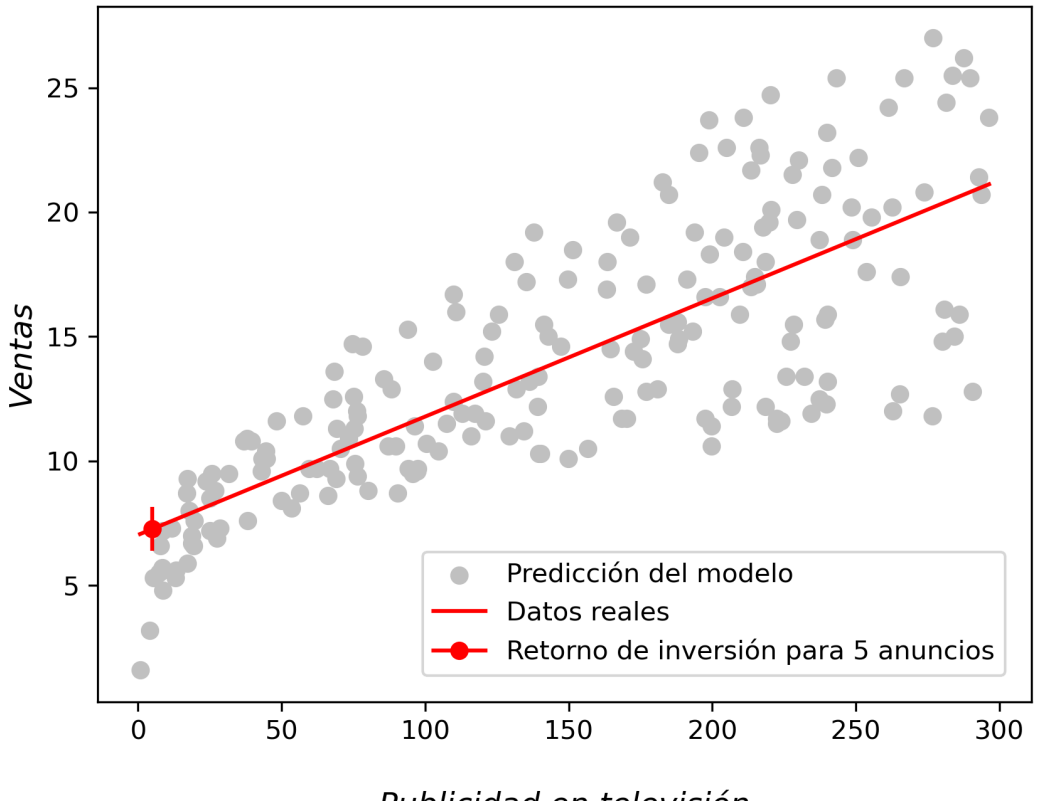

Figura 4. Predicción del retorno de inversión en ventas para una cantidad de 5 anuncios publicitarios en televisión, utilizando un modelo de regresión lineal. Predicción en línea roja se ajusta razonablemente bien a los datos reales, donde se proporciona un intervalo de confianza del 95% para la predicción, que va desde 24,99 hasta 36,11.

Sección 5:

Se utilizó un conjunto de datos de 1030 observaciones de mezclas de concreto, en las que se midieron ocho características: cemento, escoria, ceniza volante, agua, superplastificante, agregado grueso, agregado fino y edad.

Los resultados de la regresión indican que todas las variables independientes, excepto el agregado grueso y fino, tienen un efecto significativo sobre la resistencia a la compresión del concreto (Figura 5; Question_5). Los coeficientes estimados sugieren que el cemento, la escoria, la ceniza volante, el agua, el superplastificante y la edad tienen una relación positiva con la resistencia a la compresión, mientras que el agregado grueso tiene un efecto negativo. Estos resultados pueden ser útiles para los ingenieros y constructores en la selección de los componentes del concreto para lograr la resistencia deseada. Es importante tener en cuenta que este modelo de regresión lineal tiene algunas limitaciones, como la posible presencia de multicolinealidad entre las variables independientes y la posible presencia de errores de especificación en el modelo.

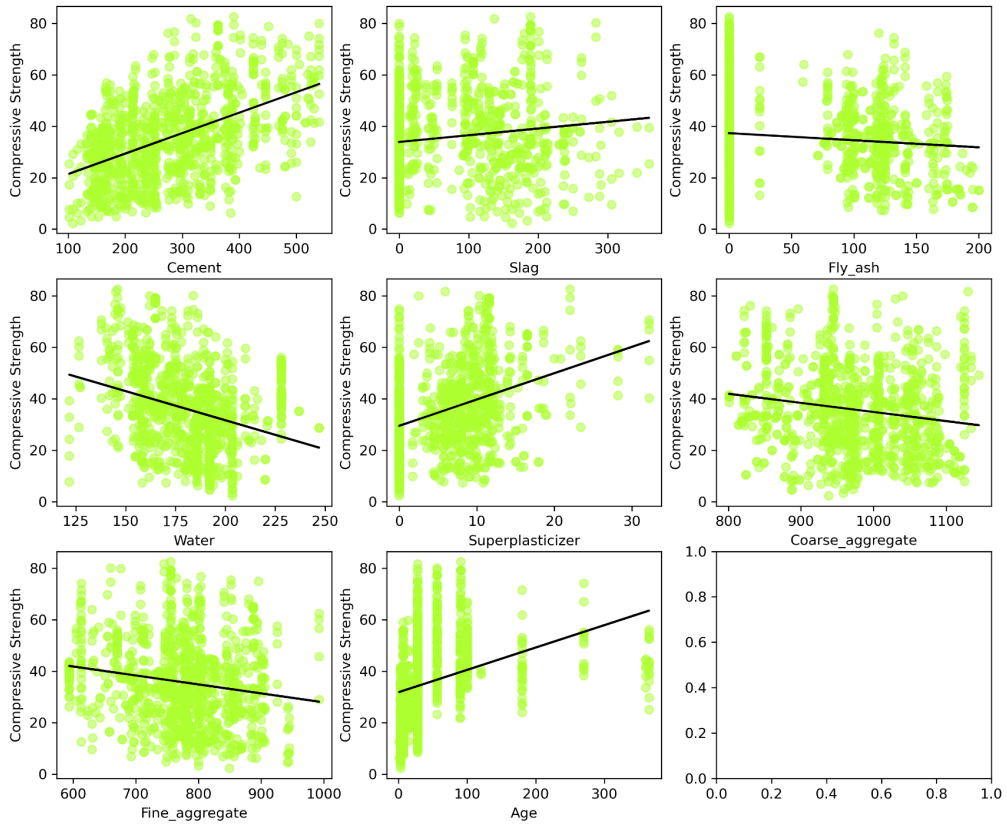

Figura 5. La figura muestra una matriz de subplots, con un total de 9 subplots organizados en 3 filas y 3 columnas, donde se somete la resistencia a la compresión con cemento, escoria, ceniza volante, agua, superplastificante, agregado grueso, agregado fino y edad.

Ambos modelos son regresiones lineales múltiples que buscan predecir la resistencia a la compresión del concreto en función de seis variables predictoras: Cemento, Escoria, Ceniza volante, Agua, Superplastificante y Edad. El modelo robusto utiliza errores estándar robustos a la heterocedasticidad, mientras que el modelo lineal utiliza errores estándar convencionales. Además, el modelo robusto informa estadísticas de prueba z, mientras que el modelo lineal informa estadísticas t. Ambos modelos tienen un buen ajuste, con un R-cuadrado de alrededor del 0.6. El modelo robusto muestra que todas las variables predictoras tienen coeficientes significativos a un nivel de confianza del 95%, excepto el superplastificante, que solo es significativo al 1% de nivel de confianza. El modelo lineal

también muestra que todas las variables predictoras son significativas, excepto el superplastificante, que solo es significativo al 0.5% de nivel de confianza.

Ambos modelos parecen tener una capacidad predictiva similar, con R-cuadrados ajustados muy cercanos (0.603 para el modelo robusto y 0.614 para el modelo lineal). Sin embargo, en el modelo robusto, los coeficientes están ajustados para la posible heterocedasticidad en los errores, lo que aumenta su robustez (Question_5). Teniendo en cuenta estas consideraciones, el modelo robusto parece ser una mejor opción que el modelo lineal en términos de robustez. Finalmente, se graficaron los valores ajustados contra los valores observados y se encontró que el modelo robusto presentó una mayor precisión en las predicciones (Figura 6).

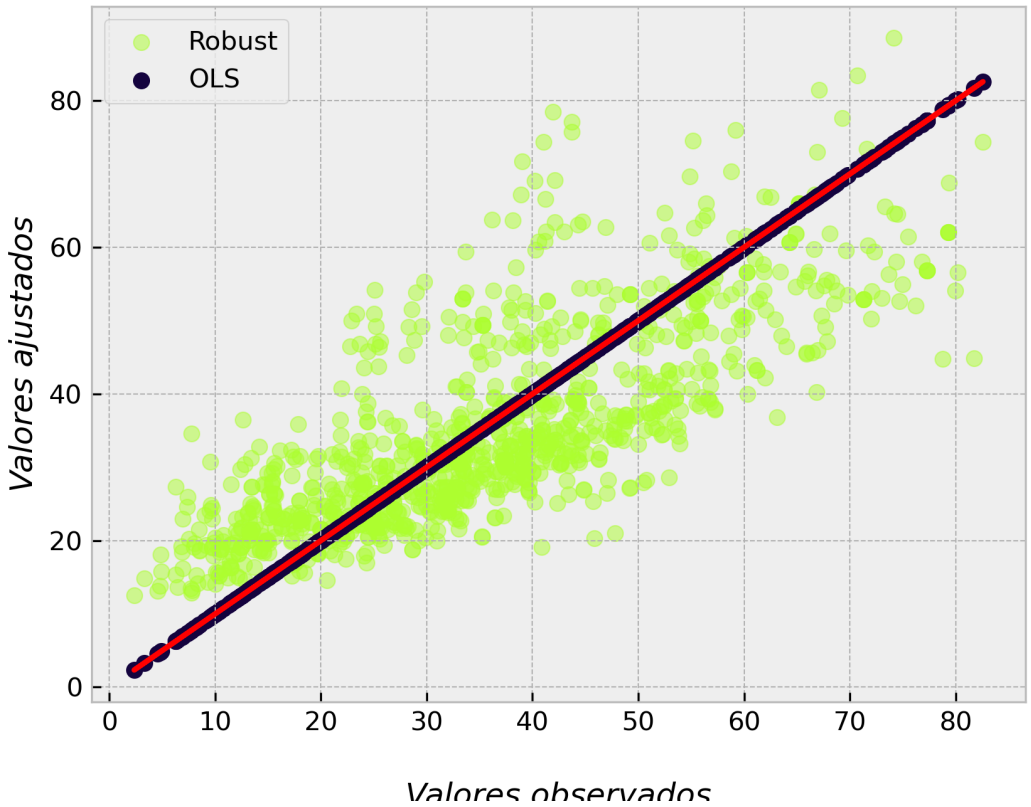

Figura 6. comparación entre los valores observados y los valores ajustados obtenidos a partir de dos modelos de regresión lineal múltiple: el modelo OLS y el modelo robusto. Para obtener los valores ajustados, se revierte la transformación Box-Cox que se realizó previamente en los datos.

El conjunto aleatorio de valores para las variables independientes se realizó por medio de los valores máximos y mínimos de cada variable en el conjunto de datos original (Question_5). Luego, se realizó una predicción del comportamiento de la resistencia a la compresión utilizando el modelo de regresión lineal múltiple previamente ajustado, incluyendo la transformación Box-Cox. Esta predicción permitió evaluar la capacidad predictiva del modelo en un conjunto de datos no visto previamente (Tabla 1).

Tabla 1. Muestra representativa de valores creados a partir de los valores máximos y mínimos de cada variable a partir del modelo robusto.

const	Cement	Slag	Fly_ash	Water	Superplasticizer	Age
1.0	292.1	44.35	99.26	206.09	7.26	328.4
1.0	267.81	82.01	96.26	190.04	10.26	95.25
1.0	152.27	307.41	137.15	227.81	28.02	108.91
1.0	448.65	274.09	112.72	127.53	18.2	310.64
1.0	312.34	147.0	126.67	234.39	15.16	302.64
1.0	211.92	97.34	41.64	155.88	24.54	247.56
1.0	347.24	267.08	40.79	194.73	12.01	78.16
1.0	264.22	27.56	76.7	138.89	24.41	92.84
1.0	495.87	282.02	180.55	202.53	11.31	314.17
1.0	443.44	235.74	156.22	143.1	29.46	126.04

Los resultados (Tabla 2) mostraron que el modelo fue capaz de predecir con precisión la resistencia a la compresión del concreto en función de las variables predictoras. Estos resultados sugieren que el modelo puede ser útil para estudios futuros en el área de la resistencia a la compresión del concreto, y que puede ser utilizado para hacer predicciones precisas en casos donde se conocen los valores de las variables predictoras.

Tabla 2. Datos predictivos de la resistencia a la compresión del cemento a partir de los datos creados.

Predicción
18.48
13.31
15.23
30.68
19.9
17.69
17.31
14.87
29.43
25.59

Conclusión

Este artículo presenta cinco secciones que demostraron la importancia del análisis exploratorio de datos y el ajuste de modelos de regresión lineal en la comprensión de la relación entre variables y la predicción de una variable de interés. Se destacó la importancia de considerar las interacciones entre las variables predictoras y realizar un análisis de diagnóstico para garantizar la validez de los resultados. Además, se presentaron modelos de regresión lineal que demostraron ser efectivos para predecir la variable de interés en diferentes contextos, como en la publicidad en televisión y en la resistencia a la compresión del concreto. En resumen, este artículo proporciona información valiosa para los investigadores y profesionales que desean mejorar su comprensión de la relación entre variables y predecir una variable de interés de manera precisa y confiable.

La contribución de Carl Friedrich Gauss al método de los mínimos cuadrados: Una revisión histórica

[Nombres]

Jorge Eduardo Lizarazo Borrero

[Correos]

jorge.lizarazo.b@gmail.com

Resumen:

En este artículo se examina el papel que Carl Friedrich Gauss desempeñó en la invención y desarrollo del método de los mínimos cuadrados. Se ofrece una revisión histórica de cómo Gauss aplicó este método para estimar la posición de los cuerpos celestes, y cómo su trabajo allanó el camino para la utilización generalizada de los mínimos cuadrados en la estadística y la econometría. Además, se presenta una revisión crítica del artículo "Gauss and the Invention of Least Squares" de Stephen Stigler.

Introducción

En la revisión histórica, se examina cómo Gauss desarrolló el método de los mínimos cuadrados para ajustar modelos matemáticos a los datos observados. Se muestra cómo Gauss utilizó este método para estimar la posición de los cuerpos celestes y cómo su trabajo allanó el camino para la utilización generalizada de los mínimos cuadrados en la estadística y la econometría. También se examinan las contribuciones de otros matemáticos y estadísticos que ayudaron a perfeccionar el método de los mínimos cuadrados.

El método de los mínimos cuadrados es uno de los métodos estadísticos más utilizados en la actualidad. Este método permite ajustar modelos matemáticos a los datos observados, minimizando la diferencia entre los valores observados y los valores estimados por el modelo (Stigler, 1981). Aunque el método de los mínimos cuadrados se utiliza ampliamente en la actualidad, su invención y desarrollo se remonta a principios del siglo XIX.

Carl Friedrich Gauss, uno de los más grandes matemáticos de todos los tiempos, es reconocido como uno de los pioneros en el desarrollo del método de los mínimos cuadrados. En su obra "Theoria Motus Corporum Coelestium", publicada en 1809, Gauss utilizó el método de los mínimos cuadrados para

estimar la posición de los cuerpos celestes. Este trabajo allanó el camino para la utilización generalizada de los mínimos cuadrados en la estadística y la econometría.

En este artículo, se ofrece una revisión histórica del papel que desempeñó Gauss en la invención y desarrollo del método de los mínimos cuadrados(Stigler, 1981). Además, se presenta una revisión crítica del artículo "Gauss and the Invention of Least Squares" de Stephen Stigler, que también examina la contribución de Gauss al método de los mínimos cuadrados. Esta revisión crítica busca ofrecer una visión completa y objetiva sobre la historia del método de los mínimos cuadrados y el papel que desempeñó Gauss en su desarrollo.

En la revisión crítica del artículo de Stigler, se examina la validez de las afirmaciones que hace sobre Gauss y el método de los mínimos cuadrados. Si bien el artículo de Stigler ofrece una perspectiva interesante sobre la historia del método de los mínimos cuadrados y el papel que desempeñó Gauss en su invención y desarrollo, también se identifican algunas limitaciones en su alcance y enfoque.

Conclusión

La obra de Gauss en el método de los mínimos cuadrados es un ejemplo más del impresionante legado que dejó en el campo de las matemáticas y la estadística. Su trabajo continúa siendo relevante y sigue siendo una parte integral de la formación de los estudiantes de estadística en todo el mundo. La revisión crítica del artículo de Stigler destaca la importancia de un análisis riguroso y completo en la evaluación de la contribución de Gauss al método de los mínimos cuadrados.

Referencias

Stigler, S. (1981). Gauss and the Invention of Least Squares. *Annals of Statistics*, *9*, 465–474. https://doi.org/DOI:10.1214/AOS/1176345451