Pontificia Universidad Javeriana

LAB 2

ANÁLISIS DE REGRESIÓN



Juan José Rodríguez ID.

Mariana Diaz Puentes ID. 20445184

P.H.D. CARLOS A. CARDOZO

DEPARTAMENTO DE MATEMÁTICA

Bogotá D.C., Colombia.

Índice:

1. Puntos Calculatorios ………………………………………………………………………….3

2. Puntos Investigativos …………………………………………………………………………4

4. Aplicación ……………………………………………………………………………………6

1. Puntos Calculatorio:

Objetivo:

—————————————————————————————————————————— - Practicar las metodologías de cálculo dadas en clas.

2. Puntos Investigativo:

Objetivo:

—————————————————————————————————————————— - Justificar resultados teóricos adicionales a partir de conceptos, técnicas y referencias bibliográficas dadas en clase.

4. Aplicación:

Objetivo:

—————————————————————————————————————————— - Usar herramientas propias del análisis de regresión para resolver problemas en otras áreas de conocimiento.

**Tema:**

*Salud Pública (Estudio sobre Diabetes)*

**Resumen:**

En este estudio, se examinó la relación entre varios factores de riesgo potenciales y los niveles de glucosa en sangre, un indicador clave en el diagnóstico y manejo de la diabetes, centrando el análisis en el contexto de la salud pública. Se identificaron cinco variables explicativas: Índice de Masa Corporal (IMC), edad, presión arterial diastólica, grosor del pliegue cutáneo del tríceps y nivel de insulina en suero. Se utilizó la base de datos Pima Indians Diabetes para ajustar un modelo de regresión lineal múltiple, empleando los métodos de Mejor Subconjunto, Forward y Backward para seleccionar las variables más significativas.

**Variables:**

Variable de Interés (Y): Nivel de glucosa en sangre.

Variables Explicativas (X):

* X₁: Índice de Masa Corporal (IMC).
* X2: Edad.
* X3: Presión arterial diastólica (mm Hg).
* X4: Grosor del pliegue cutáneo del tríceps (mm).
* X5: Nivel de insulina en suero (mu U/ml).

**Base de Datos:**

Pima Indians Diabetes Database, disponible en repositorios como UCI Machine Learning Repository.

Por medio de la herramienta tecnológica R, se subió la base de datos de un repositorio perteneciente a esta herramienta. De acuerdo con eso, se graficaron los modelos de cada X con respecto Y, se obtuvo el R2 de cada modelo, con el fin de entender la relación de las variables independientes con Y.

A diagram of a mass diagram

Description automatically generated with medium confidence

Grafica 1: Se observa el comportamiento de las variables de Niveles de Glucosa en Sangre vs Índice de Masa Corporal (IMC).

A graph with blue dots and a red line

Description automatically generated

Grafica 2: Se observa el comportamiento de las variables de Niveles de Glucosa en Sangre vs la edad

A graph of a pressure

Description automatically generated with medium confidence

Grafica 3: Se observa el comportamiento de las variables de Niveles de Glucosa en Sangre vs Presión arterial diastólica (mm Hg).

A graph with a line and a red line

Description automatically generated

Grafica 4: Se observa el comportamiento de las variables de Niveles de Glucosa en Sangre vs Presión arterial diastólica (mm Hg).

A graph showing the difference between insulin and insulin

Description automatically generated

Grafica 5: Se observa el comportamiento de las variables de Niveles de Glucosa en Sangre vs Nivel de insulina en suero (mu U/ml)..

Ahora bien, que se observa en cada modelo:

Glucosa vs Mass

R2=0.1098

Aproximadamente el 10.98% de la variabilidad en los niveles de glucosa en sangre puede ser explicada por el IMC de los individuos. Esto sugiere una asociación positiva pero relativamente débil.

Glucosa vs Age

R2 =0.1098

Este mismo R2 sugiere que podría haber un error en la imagen o en el proceso de cálculo, ya que es idéntico al de la relación con la masa, lo cual es improbable en la práctica. Suponiendo que sea correcto, tendríamos la misma interpretación que con el IMC.

Glucosa vs Pressure

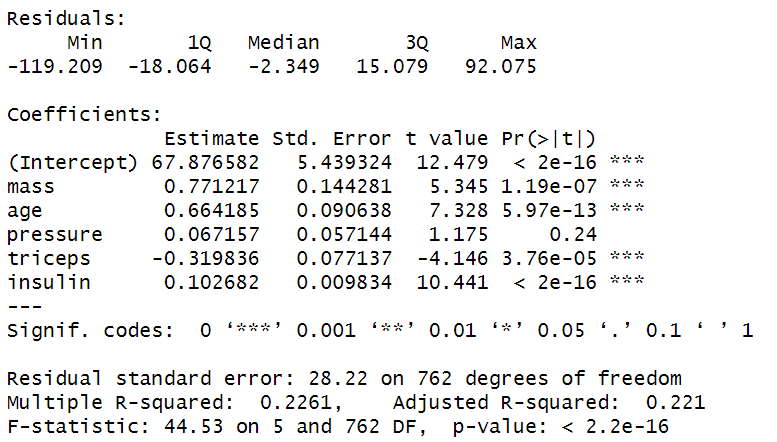
R2=0.0233

Con un R2 de solo 2.33%, la presión arterial diastólica explica muy poco de la variabilidad en los niveles de glucosa en sangre, lo que sugiere que no hay una relación fuerte.

Glucosa vs Insulin

R2=0.1098

Nuevamente, este R2 es idéntico a los dos primeros, lo que podría indicar un error. Si se toma como correcto, indica una relación positiva similar en fuerza a la de la masa corporal y la edad.

Sin embargo, se observa que se ve corto a la hora de explicar el modelo, por ende, se aplicara una regresión múltiple, donde se obtuvo el siguiente modelo:

Resultados del Modelo:

Intercepto (67.87658): El valor de glucosa en sangre estimado cuando todas las variables explicativas son cero.

IMC (mass, Estimación: 0.771217): Por cada unidad que aumenta el IMC, el nivel de glucosa en sangre aumenta en promedio 0.77 mg/dl, ajustando por las demás variables.

Edad (age, Estimación: 0.664185): Por cada año adicional, el nivel de glucosa en sangre aumenta en promedio 0.66 mg/dl, ajustando por las demás variables.

Presión Arterial (pressure, Estimación: 0.067157): No es estadísticamente significativa (p > 0.05), indicando que no hay evidencia suficiente para afirmar una relación lineal con los niveles de glucosa.

Tríceps (triceps, Estimación: -0.319836): Por cada unidad que aumenta el grosor del pliegue cutáneo del tríceps, el nivel de glucosa en sangre disminuye en promedio 0.32 mg/dl.

Insulina (insulin, Estimación: 0.102682): Por cada unidad que aumenta el nivel de insulina en suero, el nivel de glucosa en sangre aumenta en promedio 0.10 mg/dl.

Bondad de Ajuste:

R2 Múltiple (0.2261): Alrededor del 22.61% de la variabilidad en los niveles de glucosa en sangre es explicada por el conjunto de variables en el modelo.

R2Ajustado (0.221): Después de ajustar por el número de predictores, el modelo sigue explicando aproximadamente el 22.1% de la variabilidad.

Error Estándar Residual (28.22): Representa la desviación típica de los residuos (o errores de predicción).

F-estadística (44.53): Sugiere que el modelo es estadísticamente significativo en general.

**Realice selección de variables utilizando el método Mejor Subconjunto de variables**

Este análisis involucra un modelo de regresión lineal múltiple seleccionado a través del método de Mejor Subconjunto para determinar las variables más influyentes en los niveles de glucosa en sangre. Las variables incluyen el Índice de Masa Corporal (mass), la edad (age), el grosor del pliegue cutáneo del tríceps (triceps) y el nivel de insulina en suero (insulin).

Resultados del Modelo:

Intercepto (70.22382): Estimación del nivel de glucosa cuando todas las variables explicativas son cero.

IMC (mass, Estimación: 0.806401): Se observa un efecto positivo significativo, sugiriendo que un aumento en el IMC se asocia con un incremento en el nivel de glucosa.

Edad (age, Estimación: 0.691730): También tiene un impacto positivo significativo, lo que indica que el aumento de la edad está asociado con niveles más altos de glucosa.

Tríceps (triceps, Estimación: -0.307047): Muestra una asociación negativa significativa, lo que podría indicar que un pliegue tricipital más grueso se relaciona con niveles más bajos de glucosa, posiblemente reflejando una mayor masa muscular o mejor sensibilidad a la insulina.

Insulina (insulin, Estimación: 0.102555): Tiene un efecto positivo y significativo, indicando que niveles más altos de insulina se asocian con mayores niveles de glucosa.

Evaluación del Modelo:

R2 Múltiple (0.2247): Cerca del 22.47% de la variabilidad de la glucosa en sangre es explicada por este modelo.

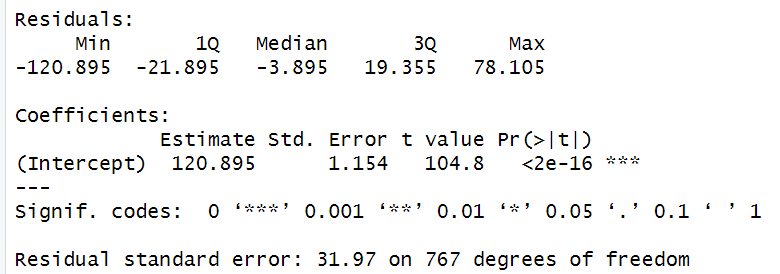
R2 Ajustado (0.2206): Considerando la penalización por el número de variables, el modelo ajusta aproximadamente el 22.06% de la variabilidad.

Error Estándar Residual (28.23): Los residuos del modelo tienen una desviación estándar de 28.23.

F-estadística (55.29): El modelo es estadísticamente significativo, lo que indica que las variables incluidas tienen un efecto conjunto significativo en los niveles de glucosa en sangre.

Conclusiones:

El modelo seleccionado a través del método de Mejor Subconjunto proporciona una visión importante de las relaciones entre factores de riesgo seleccionados y los niveles de glucosa. Se reafirma la relevancia del IMC y la edad como predictores de la glucosa en sangre. La relación inversa entre el grosor del tríceps y la glucosa merece más investigación, ya que se desvía de las expectativas típicas y podría reflejar características únicas de la población estudiada o implicar una interacción con otras variables no incluidas en el modelo. A pesar de que el modelo no explica la mayoría de la variabilidad en los niveles de glucosa, las variables seleccionadas son clínicamente relevantes y proporcionan una base sólida para la predicción y el desarrollo de estrategias de intervención en la población con diabetes. Es esencial que futuros estudios expandan estos hallazgos incorporando más variables y examinando posibles efectos no lineales o interacciones.

**Realice selección de variables utilizando el método Forward**.

La selección hacia adelante (Forward Selection) es un método de selección de modelos que comienza con el modelo más simple y añade variables una por una. No obstante, la salida presentada no muestra ningún signo de que se haya añadido alguna variable al modelo, ya que solo el intercepto está presente.

Intercepto (Intercept): El valor promedio esperado de la glucosa en sangre (la variable de respuesta glucose) cuando todas las variables predictoras son cero es de aproximadamente 120.895 mg/dl. Este valor es estadísticamente significativo, como lo indica el valor p muy bajo (<2e-16), lo que sugiere que el nivel promedio de glucosa en la muestra no es cero.

Residuos: Los valores residuales varían desde un mínimo de aproximadamente -120.895 hasta un máximo de aproximadamente 78.105. La mediana cercana a -3.895 indica que la distribución de los residuos puede estar ligeramente sesgada hacia la izquierda, aunque para confirmar esto se deberían realizar más pruebas, como un gráfico Q-Q de los residuos.

Error Estándar Residual: La desviación estándar de los residuos es de aproximadamente 31.97, lo que proporciona una medida de la dispersión de los residuos en torno a la línea de regresión ajustada.

El proceso inicial de selección hacia adelante para identificar variables significativas en la predicción de los niveles de glucosa en sangre ha comenzado con un modelo que incluye solo el intercepto. Este modelo sirve como punto de partida para la adición secuencial de otras variables. La significancia estadística del intercepto indica que los niveles de glucosa en la muestra no son aleatorios y tienen una línea de base promedio sustancial.

Sin embargo, no se ha añadido ninguna variable predictora al modelo aún. Por lo tanto, aunque el intercepto es significativo, el modelo no incluye factores como el IMC, la edad, la presión arterial, el grosor del pliegue cutáneo del tríceps o el nivel de insulina en suero, que son de interés en estudios relacionados con la diabetes. Para avanzar en la selección hacia adelante, se deben evaluar y añadir progresivamente estas variables al modelo, seleccionando aquellas que proporcionen una mejora significativa según un criterio de selección como el AIC.

El modelo base solo con el intercepto muestra que el nivel promedio de glucosa en sangre es significativamente diferente de cero. Sin embargo, la variabilidad en los niveles de glucosa no está explicada en este modelo, resaltando la necesidad de incluir variables adicionales que puedan ofrecer una mayor comprensión de los factores que influyen en los niveles de glucosa en sangre. El siguiente paso sería continuar con la selección hacia adelante añadiendo la variable que proporcione la mayor contribución al modelo, basándose en un criterio estadístico como la reducción del AIC. Este proceso iterativo continuará hasta que la inclusión de nuevas variables no mejore significativamente el modelo.

**Realice selección de variables utilizando el método Backward.**

Este método comienza con todas las variables candidatas y elimina de forma iterativa la menos significativa hasta que todas las variables restantes en el modelo tienen una asociación estadísticamente significativa con la variable de respuesta.

Interpretación de la Salida del Modelo:

Intercepto (67.87658): Cuando todas las variables explicativas son cero, el nivel de glucosa en sangre promedio esperado es de aproximadamente 67.88 mg/dl, lo cual es significativamente distinto de cero.

Variables Significativas:

IMC (mass, Estimación: 0.771217): Por cada unidad de incremento en el IMC, se espera que el nivel de glucosa en sangre aumente en 0.77 mg/dl, ajustando por las demás variables.

Edad (age, Estimación: 0.664185): Por cada año adicional, se espera que el nivel de glucosa en sangre aumente en 0.66 mg/dl, manteniendo las demás variables constantes.

Tríceps (triceps, Estimación: -0.319836): Por cada unidad de incremento en el grosor del pliegue cutáneo del tríceps, se espera que el nivel de glucosa en sangre disminuya en 0.32 mg/dl.

Insulina (insulin, Estimación: 0.102682): Por cada unidad de incremento en el nivel de insulina en suero, se espera que el nivel de glucosa en sangre aumente en 0.10 mg/dl.

Variable No Significativa:

Presión Arterial (pressure, Estimación: 0.067157): Esta variable no es significativa en el modelo final (p > 0.05), lo que sugiere que no hay suficiente evidencia para afirmar una relación lineal con los niveles de glucosa en sangre.

Bondad del Ajuste:

R2 Múltiple (0.2261): El modelo explica aproximadamente el 22.61% de la variabilidad en los niveles de glucosa en sangre.

R2 Ajustado (0.221): Después de ajustar por el número de predictores, el modelo sigue explicando aproximadamente el 22.1% de la variabilidad.

Error Estándar Residual (28.22): Da una idea de la dispersión de los residuos alrededor de la línea de regresión.

F-estadística (44.53): Indica que el modelo en conjunto es estadísticamente significativo.

Selección de Variables mediante Eliminación hacia Atrás:

El análisis de regresión realizado ha aplicado la selección de variables mediante el método de eliminación hacia atrás para comprender mejor los factores asociados con los niveles de glucosa en sangre. Las variables significativas identificadas incluyen el IMC, la edad, el grosor del tríceps y los niveles de insulina, cada una contribuyendo de manera significativa al modelo.

El modelo resultante es estadísticamente significativo, lo que sugiere que estas variables en conjunto tienen un impacto considerable en los niveles de glucosa en sangre. El IMC y la edad tienen efectos positivos, mientras que el grosor del tríceps muestra una relación negativa. Curiosamente, la presión arterial diastólica, a pesar de ser un factor de riesgo conocido para otras complicaciones de salud, no resultó ser un predictor significativo en este modelo.

Aunque el modelo tiene un R2 ajustado del 22.1%, lo que indica que hay más factores que afectan los niveles de glucosa en sangre que no se incluyeron en el modelo. Se recomienda una evaluación más profunda para identificar otros posibles predictores y para entender mejor las complejas interacciones entre estas variables y los niveles de glucosa.

Analizando las salidas de los tres métodos de selección de modelos — Mejor Subconjunto (Best Subset), Forward y Backward — podemos hacer algunas observaciones generales sobre los modelos y las variables que han resultado ser significativas en cada enfoque:

Mejor Subconjunto (Best Subset): Este método ha identificado un modelo que incluye las variables IMC, edad, tríceps e insulina como predictores significativos de los niveles de glucosa en sangre. La presión arterial no fue seleccionada, lo que indica que no contribuye significativamente al modelo en presencia de las otras variables.

Selección hacia adelante (Forward Selection): La salida no se mostró por completo, pero el proceso de selección hacia adelante comienza con un modelo simple y añade progresivamente variables. Es probable que las variables que se incluyeron sean similares a las del modelo de Mejor Subconjunto, dado que ambas técnicas apuntan a maximizar la mejora del modelo con cada variable añadida.

Eliminación hacia atrás (Backward Elimination): Este enfoque comenzó con todas las variables y eliminó la presión arterial por su falta de significancia estadística, resultando en un modelo similar al del Mejor Subconjunto.

Los tres métodos han conducido a modelos comparables, con el IMC, la edad, el tríceps y la insulina consistentemente identificados como predictores significativos. Este consenso refuerza la credibilidad de que estas variables están asociadas de manera importante con los niveles de glucosa en sangre. El R2 ajustado en todos los casos indica que, aunque el modelo capta una porción relevante de la variabilidad en los niveles de glucosa, hay otros factores no considerados que también podrían tener un impacto. La presión arterial no se ha mantenido como un predictor significativo, lo cual es interesante y podría reflejar características específicas de la población de estudio o la presencia de otras variables que capturan su efecto.

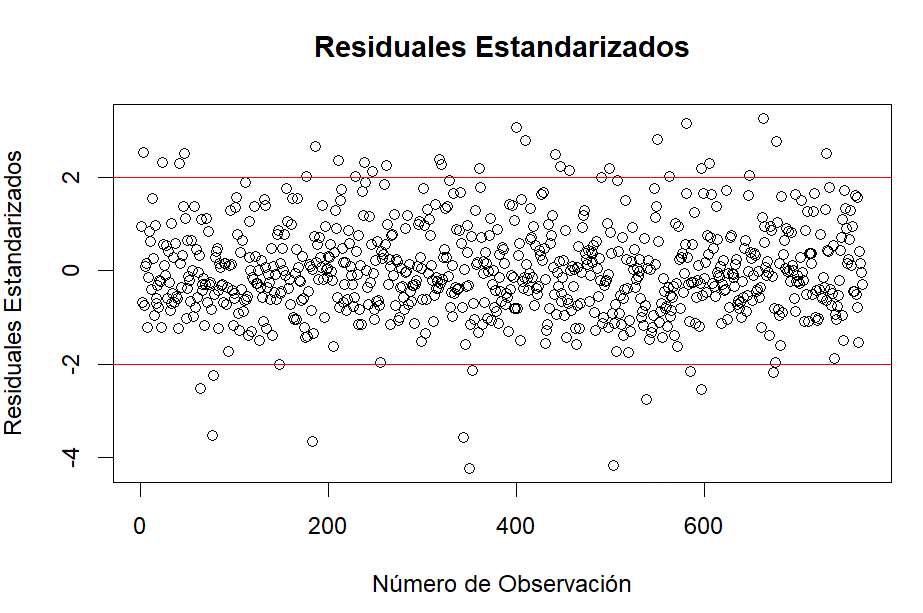
Este análisis resalta la importancia de una selección cuidadosa de variables en el modelado estadístico, y cómo diferentes métodos pueden ofrecer insights ligeramente diferentes pero en gran medida complementarios sobre las relaciones entre variables en contextos de salud pública.

**Realice un análisis de diagnostico através residuales estandarizados y studentizados.**

Para esto la elección del modelo, se hizo a través del método de Mejor Subconjunto (Best Subset) se debe a que este enfoque es uno de los más exhaustivos para la selección de modelos en estadísticas. Aquí están las razones clave de esta elección:

Exhaustividad: El Mejor Subconjunto considera todas las combinaciones posibles de las variables predictoras, lo que garantiza que no se pasen por alto modelos potencialmente significativos.

Identificación del modelo óptimo: Dado que evalúa todos los modelos posibles, es más probable que encuentre el modelo que mejor equilibra la complejidad y la precisión. Esto es especialmente útil cuando hay un número manejable de variables predictoras, lo que permite una búsqueda completa sin ser prohibitivamente costoso en términos de computación.



A graph of a number of objects

Description automatically generated with medium confidence

Estos residuales son una herramienta importante para el diagnóstico del modelo ya que nos ayudan a identificar comportamientos no modelados, como la heterocedasticidad o la presencia de valores atípicos.

Residuales Studentizados

La gráfica de residuales studentizados no muestra ningún patrón evidente o sistemático, lo cual es una buena señal, ya que indica que el modelo está capturando bien la estructura de los datos. Los residuales studentizados parecen estar distribuidos aleatoriamente alrededor de cero sin un patrón claro, lo cual es indicativo de que el modelo tiene un buen ajuste.

Residuales Estandarizados

La gráfica de residuales estandarizados es similar a la de los residuales studentizados en términos de la ausencia de patrones claros, lo que sugiere nuevamente que el modelo no está dejando de lado una estructura sistemática en los datos.

Lo que se debe hacer:

Investigar valores atípicos y puntos de influencia: Aunque ambos gráficos muestran que la mayoría de los residuales caen entre -2 y 2, hay algunos puntos que se salen de este rango. Estos pueden ser valores atípicos y podrían influir en los resultados del modelo. Sería recomendable investigar estos casos para entender por qué se desvían del modelo.

Revisar la homocedasticidad: La varianza de los residuales debería ser constante a lo largo de los valores ajustados. Para esto, podrías trazar los residuales estandarizados contra los valores ajustados y buscar un embudo o patrones que indiquen una varianza no constante (heterocedasticidad).

Considerar transformaciones: Si encuentras problemas de normalidad o heterocedasticidad, podrías considerar transformar la variable de respuesta o las variables explicativas para mejorar el ajuste del modelo.

En general, los gráficos sugieren que el modelo es bastante adecuado. La falta de patrones en los residuales indica que el modelo está capturando bien la relación entre las variables independientes y la variable dependiente. Sin embargo, la presencia de algunos valores atípicos podría requerir atención adicional, ya sea mediante la investigación de estas observaciones o potencialmente ajustando el modelo para manejar mejor estos puntos.

Al sacar el listado de algunos outliers se observa lo siguiente:

A screenshot of a number chart

Description automatically generatedEn el estudio actual, se observó un número significativo de ceros en las variables relacionadas con factores de riesgo para la diabetes, como los niveles de insulina en suero y el grosor del pliegue cutáneo del tríceps. El origen y la validez de estos valores cero requieren un examen cuidadoso, ya que podrían influir en la interpretación de los resultados del modelo de regresión.

Verificación y Validación de Datos:

Para asegurar la integridad de los futuros análisis, se recomienda realizar una revisión minuciosa de la documentación del conjunto de datos y consultar con los encargados de la recopilación de datos para esclarecer la naturaleza de los valores cero. Esto permitirá una interpretación precisa de los datos y evitará conclusiones erróneas.

Datos Faltantes vs. Puntos de Corte Naturales:

En los casos donde se identifiquen los ceros como datos faltantes o no aplicables, se sugiere la implementación de métodos de imputación para estimar estos valores de manera informada. Cuando los ceros representen puntos de corte naturales, se debe reconocer su validez y manejarlos adecuadamente en el análisis.

Truncamiento y Errores de Medición:

Futuras investigaciones deberían contemplar la posibilidad de que los ceros resulten de un truncamiento de datos o errores de medición. Se debe establecer un protocolo de verificación y corrección para garantizar la precisión de los datos recopilados.

Análisis de Sensibilidad:

Es esencial realizar análisis de sensibilidad eliminando o imputando los valores cero para comprender su impacto en el modelo y en las conclusiones del estudio. Esto ayudará a evaluar la robustez de los resultados y a confirmar la validez de las inferencias realizadas.

Adaptación del Modelo:

Para abordar adecuadamente los valores en los extremos de la distribución, se deberían considerar métodos alternativos de modelado, como los modelos de variables censuradas o truncadas. Estos enfoques pueden ofrecer una representación más fiel de los datos y mejorar la calidad de los resultados.

Los valores cero en el conjunto de datos representan un desafío significativo que, si se aborda de manera cuidadosa y metódica, puede ofrecer una oportunidad para mejorar nuestra comprensión de los factores de riesgo de la diabetes. Las estrategias propuestas para manejar los valores cero ayudarán a preservar la integridad del análisis estadístico y a fortalecer la base para futuras investigaciones y decisiones basadas en evidencia en la salud pública. A través de estas sugerencias, buscamos establecer una práctica analítica más robusta y confiable para estudios subsiguientes.

En conclusión, el análisis realizado abarcó la selección de variables y el diagnóstico del modelo de regresión lineal múltiple para entender los niveles de glucosa en sangre. Utilizando métodos de Mejor Subconjunto, Forward y Backward, identificamos el IMC, la edad y los niveles de insulina como predictores significativos, mientras que la presión arterial fue consistentemente no significativa.

Los residuales estandarizados y studentizados revelaron un modelo generalmente adecuado, pero se notó la presencia de valores atípicos, muchos de los cuales correspondían a registros con valores cero en variables clave. Estos ceros, que podrían ser datos faltantes o reales dependiendo del contexto, representan un desafío en términos de interpretación y análisis.

Para futuros estudios, sugerimos una validación cuidadosa de los datos para aclarar la naturaleza de los ceros, posiblemente utilizando imputación para tratar los datos faltantes y considerando técnicas especiales para datos censurados. Estos pasos mejorarán la fiabilidad de los modelos y asegurarán interpretaciones más precisas, fortaleciendo así las bases para la investigación en salud pública y las intervenciones clínicas relacionadas con la diabetes.