Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matemáticas

Modelación del rendimiento académico de los estudiantes de la FCFM durante la pandemia

Tesis presentada al

Colegio de Matemáticas

como requisito parcial para la obtención del grado de

Licenciado en Actuaría

por

Oscar Espinosa Cuahuizo

Asesorado por

Dra. Hortensia Josefina Reyes Cervantes
Dr. Flaviano Godínez Jaimes

Puebla Pue.

5 de agosto de 2024

Título: Modelación del rendimiento académico de los estudiantes de la FCFM durante la pandemia

Estudiante: Oscar Espinosa Cuahuizo

COMITÉ

Mtra. Brenda Zavala López

Presidente

Dr. Rei Israel Ortega Gutiérrez

Secretario

Mtra. Rosalba Mercado Ortiz

Vocal

Dr. Flaviano Godínez Jaimes

Co Asesor

Dra. Hortensia Josefina Reyes Cervantes

Asesor

Agradecimientos

La finalización de esta etapa tan valiosa en mi vida ha sido posible gracias al inmenso apoyo que he recibido incondicionalmente por parte de mi familia, profesores y amigos. Por este motivo, deseo expresar mi profunda gratitud a todas aquellas personas que han sido parte de este logro.

En primer lugar, quiero agradecer a mi madre, Tomasa Cuahuizo Castillo, y a mi padre, Oscar Espinosa Fuentes, quienes son el pilar fundamental de este logro y de cada etapa de mi vida. Gracias por todo el apoyo, el amor incondicional, la confianza y la motivación que me han ofrecido desde el primer día de mi vida. Su fortaleza y su fe en mí han sido esenciales para alcanzar esta meta y muchas más. Ellos son, sin duda, la riqueza más importante que me ha sido otorgada.

A mi asesora de tesis, Dra. Hortensia Josefina Reyes Cervantes, le debo un agradecimiento especial por su gran apoyo, dirección y paciencia durante la realización de este trabajo. Sus enseñanzas, tanto en el aspecto académico como personal, han sido de inmenso valor para mí. Ha sido un pilar fundamental en la definición del camino profesional que deseo seguir.

También quiero agradecer a mi co-asesor de tesis, Dr. Flaviano Godínez Jaimes, por su compromiso, tiempo, observaciones y conocimientos brindados durante la elaboración de esta tesis. Su apoyo fue esencial para la finalización de la misma.

Agradezco profundamente al jurado por su tiempo, dedicación y valiosas aportaciones durante este proceso. Su experiencia y conocimientos han enriquecido mi trabajo y han sido fundamentales para la culminación de esta tesis.

A mi hermana, Mitzi Fernanda Espinosa Cuahuizo, por ser mi compañera de vida, siempre presente en los momentos más importantes, tanto en las buenas como en las malas. Gracias por brindarme tu apoyo, cariño, amistad y mucho más. Tus consejos, risas y, sobre todo, ese amor incondicional, han sido un refugio para mí. Te admiro profundamente por todas tus cualidades y capacidades, las cuales han impulsado mi deseo de ser mejor cada día. Gracias por ello y por mucho más.

A Andrea López Victoria, quien me ha otorgado compañía en momentos importantes de mi vida, tanto en los buenos como en los malos. Me has dado un apoyo constante, fortaleciendo mis debilidades e impulsando mis capacidades. Has sido un refugio de seguridad al que siempre podía acudir. Agradezco y reconozco lo importante que fuiste para la finalización de esta tesis.

Agradezco el apoyo constante, aliento, cariño y confianza, que ha sido un estímulo importante que me ha ayudado a finalizar esta tesis.

A mis amigos que he tenido en varias etapas de mi vida, quienes han sido mi red de apoyo constante. Gracias por cada conversación, por cada ayuda y su increíble amistad, la cual ha sido verdaderamente un tesoro en este proceso.

A mis abuelos, Ezequiel Espinoza y Lucía Fuentes Sánchez. Agradezco de corazón todo el cariño y el apoyo que me brindaron durante mi infancia, siendo una parte fundamental en mi desarrollo personal. Su amor y dedicación han dejado una huella imborrable en mi vida. En especial, a mi abuelita Lucía, quien nos dejó hace tiempo, este logro es también para usted. La llevo siempre en mi corazón.

Finalmente, a todos aquellos quienes contribuyeron en mi desarrollo personal, académico y profesional, ayudándome a finalizar esta fase. Quienes no menciono explícitamente por este medio, sin embargo, generalizo como ciertos familiares cercanos que me han brindado su cariño, compañeros y amigos

de mi primera experiencia profesional que han sido una fuente de conocimientos y consejos, aquellos profesores que marcaron para bien mi vida académica. A todos ellos, les extiendo mi más sincero agradecimiento, dado que han aportado un grano de arena para este proyecto. Su apoyo y aliento han sido cruciales para llegar hasta aquí.

Índice general

Capítulo 1

Introducción

En la presente tesis se tiene como objetivo identificar las principales variables que influyeron en el rendimiento académico de los estudiantes universitarios en la Facultad de Ciencias Físico Matemáticas de la Benemérita Universidad Autónoma de Puebla (BUAP) durante la pandemia entre el periodo de Primavera 2019 y Primavera 2022. Para lograr este propósito, se desarrolló un modelo de regresión lineal múltiple que considera diversas variables, desde la disponibilidad de recursos tecnológicos hasta la situación socioeconómica de los estudiantes, además de las variables académicas individuales. La información necesaria se obtuvo mediante una encuesta aplicada a una muestra representativa de la población objetivo. La elección cada una de las variables seleccionadas se eligieron con base a una lógica fundamentada en la expectativa de que cada una tiene una correlación directa con el rendimiento académico de los estudiantes durante la pandemia. Se utilizaron tanto teorías previas como análisis exploratorios de datos para determinar la relevancia de cada variable.

El promedio de un estudiante es un valor de gran importancia que permite evaluar su rendimiento académico de manera general durante una etapa educativa específica. Este se calcula a partir de las calificaciones obtenidas por el estudiante en las diferentes materias y asignaturas que cursa a lo largo de su trayectoria escolar. El rendimiento académico de un estudiante es influido de manera significativa por diversas variables en su vida cotidiana, lo que a su vez tiene un impacto directo en su desempeño escolar.

Éstas pueden ser tan evidentes como el tiempo que un alumno dedica al estudio. Sin embargo, existen factores que a primera vista podrían parecer no estar relacionados con el rendimiento escolar, como una relación sentimental o la distancia que existe entre el hogar de un estudiante y la institución educativa, etc. Inclusive el nivel de ingresos familiares que percibe el estudiante puede tener una mayor influencia. La existencia de múltiples variables que afectan el rendimiento académico de los estudiantes abre la puerta a la realización de análisis y estudios para identificar y comprender estas variables a profundidad.

Su estudio y relación con el rendimiento académico es fundamental para tomar decisiones informadas y desarrollar estrategias educativas efectivas. A través de un análisis cuidadoso, es posible identificar qué factores tienen un impacto en el rendimiento académico de los estudiantes.

No obstante, eventos inesperados pueden alterar significativamente este panorama. Un ejemplo de esto sucedió con la pandemia de COVID-19, que desde su aparición ha transformado radicalmente la educación a nivel mundial.

La Organización Mundial de la Salud (OMS) define al COVID-19 como: “la enfermedad infecciosa causada por el coronavirus que se ha descubierto más recientemente. Tanto el nuevo virus como la enfermedad eran desconocidos antes de que estallara el brote en Wuhan (China) en diciembre de 2019”. Es un virus que causa enfermedades respiratorias que hasta el 6 de marzo ha matado a 74,565 personas, con un total de 1,345,046 confirmados y 276,515 recuperados según el portal tradingview.com (Britez, 2020).

La pandemia COVID-19 ha venido generando cambios y disrupciones en amplios sectores de la actividad humana. La educación ha sido uno de los más afectados debido a la imposición administrativa

del cierre total de los centros educativos en gran parte de los países del mundo. La modalidad de educación a distancia, fundamentalmente en soporte digital, vino a ofrecer soluciones de emergencia a dicha crisis (García, 2021).

La crisis sanitaria desencadenada por la pandemia de COVID-19 ha acelerado la transición hacia la educación en línea, obligando a estudiantes y docentes a adaptarse a nuevos entornos virtuales de aprendizaje. Ciertas herramientas como Moodle y Google Apps for Education se han convertido en aliados clave para la enseñanza virtual, ofreciendo opciones flexibles y adaptadas a las necesidades de profesores y estudiantes (García, Rivero & Ricis, 2020).

La crisis sanitaria respecto al COVID-19 ha generado una reevaluación importante en el entorno educativo. La adaptación de la educación en línea y el uso de ciertas herramientas digitales han causado cambios en de las clases y la interacción entre profesores y estudiantes. Esta transición también ha resaltado la importancia a la accesibilidad de la tecnología en el proceso educativo, evidenciando disparidades en el acceso a recursos entre distintos grupos de alumnos e incluso entre los propios docentes.

La adaptación a la educación en línea ha sido un desafío tanto para docentes como para estudiantes. Muchas instituciones han optado por la transmisión de clases en vivo a través de internet, así como la utilización de aplicaciones y programas informáticos para satisfacer las necesidades educativas en todos los niveles del sistema (García Aretio, 2020). Sin embargo, la falta de preparación y recursos tecnológicos ha sido un obstáculo significativo. Según un estudio exploratorio en Iberoamérica, el mayor problema identificado por los docentes fue el desconocimiento de los modelos pedagógicos, seguido de la evaluación del alumnado y la falta de plataformas y recursos tecnológicos (Fardoun et al., 2020).

A pesar de los desafíos, este período también ha estimulado la innovación en la enseñanza y el aprendizaje. La situación de pandemia ha impulsado la utilización de elementos de la virtualidad para no frenar los procesos de enseñanza-aprendizaje, lo que ha generado una "tormenta perfecta" en lo que respecta a la evaluación en línea (García-Peñalvo, 2020a).

Los educadores han buscado formas creativas de mantener el compromiso de los estudiantes y fomentar la participación activa, en contra de las barreras físicas. Además, se ha dado un énfasis mayor en el desarrollo de habilidades digitales y la adaptabilidad, competencias esenciales en el mundo actual.

Sin embargo, la implementación de estas tecnologías ha puesto de manifiesto las desigualdades en el acceso a recursos digitales entre diferentes grupos de estudiantes. Según datos del último informe PISA de la OCDE, en España, las familias de bajos ingresos enfrentan dificultades significativas para acceder a dispositivos y conexión a Internet, lo que compromete su participación efectiva en la educación en línea. Estas disparidades socioeconómicas subrayan la necesidad urgente de abordar la brecha digital para garantizar una educación equitativa para todos los estudiantes (García, Rivero & Ricis, 2020).

En definitiva, la pandemia ha acelerado cambios profundos en la educación, destacando la necesidad de flexibilidad en el sistema educativo. A medida que las instituciones y los educadores continúan navegando por estos desafíos, es esencial considerar cómo mantener los aspectos positivos de estas transformaciones y al mismo tiempo abordar las barreras y desafíos que han surgido para garantizar una educación equitativa.

El cambio radical provocado por la pandemia de COVID-19 ha introducido nuevas variables que han tenido un impacto directo en el rendimiento académico de los estudiantes. La alteración del entorno educativo tradicional ha desencadenado una revisión profunda de cómo las variables previamente identificadas interactúan y afectan los resultados académicos.

La introducción repentina de modalidades de educación en línea, la accesibilidad a nuevas herramientas tecnológicas y la adaptación a un aislamiento social son solo algunas de las nuevas variables a considerar en estos estudios que se realizan o nuevos que surgirán. La naturaleza sin precedentes de esta situación ha requerido una evaluación cuidadosa de cómo las variables tradicionales, como el tiempo dedicado al estudio o el acceso a recursos tecnológicos, interactúan ahora en el contexto de la educación en línea.

Es de gran interés identificar las variables que afectaron el rendimiento académico de los estudiantes durante la pandemia y determinar cuáles tuvieron un impacto positivo o negativo en cada estudiante en general. El análisis de los coeficientes estimados en el modelo de regresión que se obtendrá puede

proporcionar información sobre cómo cada variable independiente afecta el rendimiento académico de los estudiantes durante la pandemia, permitiendo identificar aquellas variables que tuvieron un impacto positivo o negativo en general.

Se busca explorar cómo las nuevas condiciones impuestas por la pandemia han interactuado con las variables tradicionales que afectan el rendimiento académico. Se pretende entender cómo la adaptación a la educación en línea, el acceso a la tecnología, el apoyo familiar y otras variables relevantes han contribuido a tener la capacidad de resolver problemas, interactuar y trabajar en equipo por parte de los estudiantes.

A través de un análisis riguroso y un enfoque metodológico sólido, se espera arrojar luz sobre los cambios en el rendimiento académico de los estudiantes debido a la pandemia, lo que a su vez puede contribuir al desarrollo de estrategias más efectivas para abordar situaciones similares en el futuro y garantizar una educación de alta calidad en cualquier circunstancia.

Si bien se reconoce que el modelo de regresión lineal múltiple utilizado tiene sus limitaciones en cuanto a la capacidad de abordar todas las preguntas planteadas en el estudio, su aplicación proporciona una comprensión inicial de las relaciones entre las variables independientes y el rendimiento académico. Este enfoque permite identificar las variables que podrían tener una influencia significativa en el rendimiento académico. Desafortunadamente, este trabajo carece de la infraestructura necesaria para resolver muchas de estas preguntas, ya que solo se cuenta con un cuestionario que refleja una parte de la problemática. Sin embargo, con un mayor desarrollo y apoyo, podría explorarse otros niveles de análisis, más allá del entorno académico. A pesar de estas limitaciones, este estudio representa un primer paso importante hacia una comprensión más profunda de la problemática y la identificación de áreas clave para futuras investigaciones o acciones.

Capítulo 2

Preliminares

En el presente capítulo, se brinda una visión panorámica de los conceptos teóricos que se emplearon como cimientos para construir el modelo de regresión lineal múltiple. Estos conceptos abarcan tanto datos cuantitativos como categóricos, un compendio introductorio sobre los modelos de regresión en sí, asimismo, métodos de selección de variables para los modelos de regresión, y la transformación de Box Cox.

Éstos fundamentos teóricos proporcionan la base esencial para la interpretación, construcción y validación de un modelo de regresión lineal múltiple sólido y confiable. Su comprensión adecuada es crucial para tomar decisiones informadas durante todo el proceso de análisis y modelado estadístico.

2.1. Clasificación de variables

Una variable es una característica medible la cual puede tomar diferentes valores. Existen dos diferentes tipos de variables, la primera clasificación nos describe si ellas pueden ser cuantitativas o cualitativas. Según Luis Rincón (2017), se tienen las siguientes definiciones.

Definición 2.1.1. Una variable es cuantitativa si sus valores son números y representan una cantidad.
Definición 2.1.2. Una variable es cualitativa si sus valores representan una cualidad, un atributo o una categoría. Se les llama también variables categóricas.

Las variables cuantitativas se clasifican en dos categorías de acuerdo con el tipo de valores que toman: discretas o continuas.
Definición 2.1.3. Una variable cuantitativa es continua si puede tomar todos los valores dentro de un intervalo (a, b) de números reales y no toma valores aislados.
Definición 2.1.4. Una variable cuantitativa es discreta si el conjunto de todos sus posibles valores tiene un número finito de elementos, o bien es infinito, pero se pueden numerar uno por uno de acuerdo al conjunto de número naturales.

2.2. Modelos de regresión

El análisis de regresión es una técnica estadística para investigar y modelar la relación entre una variable dependiente y una o varias variables independientes. Son numerosas las aplicaciones de la regresión, en cualquier campo del conocimiento. De hecho, puede decirse que el análisis de regresión es la técnica estadística más usada (Montgomery, D.; Peck, E. & Vining, E., 2006).

En general el análisis de regresión permite estudiar el efecto de una o más variables que se conocen comúnmente independientes (o predictora), sobre otra que llamamos dependiente. Si se incluyen dos o más variables independientes se tiene un modelo de Regresión Lineal Múltiple.

Un modelo de regresión lineal que relaciona una variable dependiente o respuesta aleatoria y en un conjunto de variables independientes $x_1, x_2, x_3, ..., x_k$ tiene la forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon, \tag{2.1}$$

donde $\beta_0, \beta_1, \beta_2, ..., \beta_k$ son parámetros desconocidos, ε es una variable aleatoria, por otro lado $x_1, x_2, x_3, ..., x_k$ pueden ser aleatorias o no. En donde se supone que $E[\varepsilon] = 0$, en consecuencia,

$$E(y) = \beta_0 + \beta_1 + \beta_2 x_2 + ... + \beta_k x_k. \tag{2.2}$$

2.2.1. Modelo de regresión lineal simple

Mediante el modelo de regresión lineal se pretende describir la relación entre dos variables, además de realizar inferencias sobre el comportamiento de la variable dependiente. En lo sucesivo se denota por x la variable independiente (también llamada predictora o regresora) y por y a la dependiente (o variable de respuesta).

El primer paso en un análisis de regresión es elaborar el diagrama de dispersión de datos, y se busca identificar la ecuación matemática que describa la relación entre las variables independientes y dependientes. De acuerdo con el diagrama de dispersión de los datos, ésta puede ayudar a identificar la ecuación en un caso específico, de forma general, considera la relación entre las variables con una ecuación de línea recta, donde se tiene que β_0 es la ordenada al origen y β_1 es la pendiente. Ahora bien, los datos no caen exactamente en la ecuación de la recta, por lo que se debe de modificar la ecuación (2.3), por lo que se denota por ε al error en la ecuación. Se tiene que ε es un error estadístico en el modelo, en otras palabras, es una variable aleatoria que explica por qué el modelo no ajusta exactamente los datos. Por lo tanto, un modelo que expresa lo mencionado es el siguiente.

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{2.3}$$

A la anterior ecuación se le llama **modelo de regresión lineal simple.**

Para entender los supuestos de un modelo de regresión lineal simple, es necesario comprender conceptos sobre las distribuciones de probabilidad. Haciendo referencia a las propiedades de que las observaciones sean Independientes e Idénticamente Distribuidas Normalmente ($IIDN$) .

La propiedad de ser $IIDN$ implica que las observaciones siguen una distribución normal con una media y una varianza constantes. En el caso específico de una distribución normal $N(\mu, \sigma^2)$, la media μ es cero y la varianza σ^2 es constante.

Se supone que los errores son $IIDN(0, 2)$, esto es, cumplen son $E[\varepsilon] = 0$ y $Var(\varepsilon) = \sigma^2$, donde $IIDN$ significa Ïndependientes e Idénticamente Distribuidas Normalmente".

Las suposiciones que deben cumplir los modelos de regresión lineal simple son:

1. $x_i, i = 1, 2, 3, ..., n$ son observaciones de la variable independiente tomada y que son consideradas como no aleatorias.
2. $\beta_i, i = 0, 1$ son parámetros desconocidos que determinan la recta de regresión.
3. $\varepsilon_i, i = 1, 2, 3, ..., n$ llamados errores, son variables aleatorias no observables, independientes idénticamente distribuidas en forma de $N(0, \sigma^2)$
4. Como consecuencia de que ε_i son $IIDN(0, \sigma^2)$, los y_i son variables aleatorias independientes con distribución normal con media $\beta_0 + \beta_1 x_i$ y varianza σ^2, $IIDN(\beta_0 + \beta_1 x_i, \sigma^2)$. Esto es, $E[y|x] = \beta_0 + \beta_1 x$ y $\text{Var}(y|x) = \text{Var}(\beta_0 + \beta_1 x) = \sigma^2$.

2.2.2. Estimación de los parámetros del modelo de regresión lineal simple

Se usa el método de mínimos cuadrados, esto es, se estiman β_0 y β_1 tales que la suma de los cuadrados de las diferencias entre las observaciones y_i y la línea recta $\widehat{y}_i$ sea mínima. La ecuación (2.3), se escribe:

$$y_i = \beta_0 + \beta_1 x + \varepsilon_i, \quad i = 1, ..., n. \tag{2.4}$$

La ecuación (2.3) es un **modelo poblacional de regresión**, mientras que la ecuación (2.4) es un **modelo muestral de regresión**.

Los estimadores por mínimos cuadrados de β_0 y β_1 son:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.5}$$

y

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{2.6}$$

en donde:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.7}$$

y

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2.8}$$

Entonces, tenemos que el modelo ajustado de regresión lineal simple es,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{2.9}$$

La diferencia entre el valor observado y_i y el valor ajustado correspondiente $\hat{y}_i$ se llama residual. Matemáticamente, el $i-$ésimo **residual** es:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i. \tag{2.10}$$

Los residuales tienen un papel importante para investigar la adecuación del modelo de regresión ajustado y para detectar problemas con respecto a los supuestos. Estos supuestos, detallados en la sección 2.2.1 Modelo de regresión lineal simple.

Los estimadores por mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen propiedades estadísticas importantes. Primero, obsérvese que $\hat{\beta}_0$ y $\hat{\beta}_1$ son **combinaciones lineales** de las observaciones y_i. Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ por mínimos cuadrados son **estimadores insesgados** de los parámetros β_0 y β_1 del modelo. Para demostrarlo con $\hat{\beta}_i$ considérese

$$E(\hat{\beta}_i) = E\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i E(y_i) = \sum_{i=1}^{n} c_1 + (\beta_0 + \beta_i x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

ya que, se supone que, $E(e_i) = 0$. Ahora se puede demostrar en forma directa que $\sum_{i=1}^{n} c_i = 0$ y que $\sum_{i=1}^{n} c_i x_i = 1$, y entonces,

$$E(\hat{\beta}_i) = \beta_i, \text{ para } i = 0, 1.$$

Sea $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ y $c_i = \frac{x_i - \bar{x}}{S_{xx}}$ para $i = 1, 2, ..., n$. Entonces

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i y_i. \tag{2.11}$$

La varianza de $\widehat{\beta}_1$ se calcula como sigue:

$$\text{Var}(\widehat{\beta}_1) = \text{Var}(\sum_{i=1}^{n} c_i y_i) = \sum_{i=1}^{n} c_i^2 \text{Var}(y_i) \tag{2.12}$$

ya que las observaciones y_i son no correlacionadas, esto debido a la suposición de independencia entre las unidades de observación en el modelo de regresión, por lo que la varianza de la suma es igual a la suma de las varianzas. La varianza de cada término en la suma es $c_i^2 Var(y_i)$ y se ha supuesto que $\text{Var}(y_i) = \sigma^2$, en consecuencia

$$\text{Var}(\widehat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \overline{x}^2)}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}. \tag{2.13}$$

La varianza de $\widehat{\beta}_0$ es

$$\text{Var}(\widehat{\beta}_0) = \text{Var}(\overline{y} - \widehat{\beta}_1 \overline{x}) = \text{Var}(\overline{y}) + \overline{x}^2 \text{Var}(\widehat{\beta}_1) - 2\overline{x} Cov(\overline{y}, \widehat{\beta}_1). \tag{2.14}$$

Ahora bien, la varianza de $\overline{y}$ no es más que $\text{Var}(\overline{y}) = \frac{\sigma^2}{n}$, y la covarianza entre $\overline{y}$ y $\widehat{\beta}_1$ es cero. Así,

$$\text{Var}(\widehat{\beta}_0) = \text{Var}(\overline{y}) + \overline{x}^2 \text{Var}(\widehat{\beta}_1) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right) \tag{2.15}$$

Estimación de σ^2

Es necesario un estimador de σ^2 para probar hipótesis y formar estimadores de intervalo pertinentes para el modelo de regresión. El estimador de σ^2 se obtiene de la suma de cuadrados de residuales:

$$SS_R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{2.16}$$

Sea $S_{xy} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$.

Sustituyendo la ecuación (2.9) en la ecuación (2.16) y simplificando se tiene:

$$SS_{Res} = SS_R = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 - \widehat{\beta}_1 S_{xy} \tag{2.17}$$

pero

$$\sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2 \equiv SS_T$$

donde SS_T representa la suma total de cuadrados, por lo que

$$SS_{Res} = SS_T - \widehat{\beta}_1 S_{xy}. \tag{2.18}$$

La suma de cuadrados de residuales tiene $n - 2$ grados de libertad, porque dos grados de libertad se asocian con los estimados $\widehat{\beta}_0$ y $\widehat{\beta}_1$ que se usan para obtener $\hat{y}_i$. El valor esperado de SS_{Res} es $E(SS_{Res}) = (n-2)\sigma^2$, por lo que un **estimador insesgado de** σ^2 es

$$\widehat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res}. \tag{2.19}$$

La cantidad MS_{Res} se llama **cuadrado medio residual**. La raíz cuadrada de σ^2 se llama, a veces, el **error estándar de la regresión** y tiene las mismas unidades que la variable de respuesta y.

Prueba de hipótesis de la pendiente y de la ordenada al origen

El juego de hipótesis que se ocupa para esta prueba es el siguiente

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0. \tag{2.20}$$

Esta hipótesis se relaciona con la **significancia de la regresión**. El no rechazar $H_0 : \beta_1 = 0$ implica que no hay relación lineal entre x y y. Si se rechaza $H_1 : \beta_1 \neq 0$, equivale a decir que la variable x explica la variabilidad de la variable de respuesta.

Para construir **intervalos de confianza** y realizar prueba de hipótesis para el parámetro β_1, el estadístico apropiado tiene distribución t con $n-2$ grados de libertad

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}, \tag{2.21}$$

en donde el denominador del estadístico t_0 en la ecuación (2.21) se llama con frecuencia el error estándar estimado. Esto es, $se(\hat{\beta}_1) = \sqrt{MS_{Res}/S_{xx}}$.

La hipótesis de la significancia de la regresión se rechaza si $|t_0| > t_{\alpha/2,n-2}$.

En el contexto de las pruebas de hipótesis y la construcción de intervalos de confianza, α representa el nivel de significancia, que es la probabilidad de cometer un error de Tipo 1 al rechazar incorrectamente la hipótesis nula cuando en realidad es verdadera.

Un intervalo de confianza de $100\,(1-\alpha)$ por ciento para la pendiente β_1 se determina por:

$$\hat{\beta}_1 - t_{\alpha/2,n-2} se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2} se(\hat{\beta}_1) \tag{2.22}$$

donde $t_{\alpha/2,n-2}$ es el punto porcentual de una variable t con $n-2$ grados de libertad que deja una probabilidad de $(\alpha/2)$ en la cola superior.

De forma similar, un intervalo de confianza de $100\,(1-\alpha)$ por ciento para la ordenada al origen β_0 se determina por:

$$\hat{\beta}_0 - t_{\alpha/2,n-2} se(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2} se(\hat{\beta}_0). \tag{2.23}$$

Valor p.

La práctica común al reportar los resultados de pruebas de hipótesis estadísticas es incluir el valor p de la prueba. Si tenemos el valor p de una prueba, podemos determinar el resultado de la prueba comparando el valor p con el nivel de significancia elegido, α, sin necesidad de buscar o calcular los valores críticos. La regla es la siguiente:

Regla del Valor p. Rechazar la hipótesis nula cuando el valor p es menor o igual que el nivel de significancia α. Es decir, si $p \leq \alpha$, entonces se rechaza H_0. Si $p >\leq \alpha$, entonces no se rechaza H_0.

Análisis de varianza

Se puede usar un método de **análisis de varianza** para probar el significado de la regresión. Éste se basa en una partición de la variabilidad total de la variable de respuesta. Para obtener esta partición se comienza con la identidad

$$y_i - \overline{y} = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i). \tag{2.24}$$

Se elevan al cuadrado ambos lados de la ecuación (2.24) y se suma para todas las n observaciones. Así se obtiene:

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+2\sum_{i=1}^{n}(\hat{y}_i-\overline{y})(y_i-\hat{y}_i) \tag{2.25}$$

Nótese que el tercer término del lado derecho de esta ecuación se escribe de la siguiente forma:

$$\begin{aligned}2\sum_{i=1}^{n}(\hat{y}_i-\overline{y})(y_i-\hat{y}_i)&=2\sum_{i=1}^{n}\hat{y}_i(y_i-\hat{y}_i)-2\overline{y}\sum_{i=1}^{n}(y_i-\hat{y}_i)\\&=2\sum_{i=1}^{n}\hat{y}_ie_i-2\hat{y}\sum_{i=1}^{n}e_i\\&=0\end{aligned}$$

ya que la suma de los residuales siempre es igual a cero y la suma de los residuales ponderados por el valor ajustado $\hat{y}_i$ correspondiente también es igual a cero. Por lo anterior,

$$\sum_{i=1}^{n}(y_i-\overline{y}_i)^2=\sum_{i=1}^{n}(\hat{y}_i-\overline{y}_i)^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2. \tag{2.26}$$

La ecuación (2.26) es la **identidad fundamental del análisis de varianza para un modelo de regresión.** Se observa que el lado izquierdo de la ecuación es la suma corregida de cuadrados de las observaciones, SS_T, que mide la variabilidad total en las observaciones.

Se ve que $SS_{Res}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ es la **suma de cuadrados de los residuales o la suma de cuadrados de error** de la Ecuación (2.16).

De igual forma, se tiene que $SS_R=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$ es la **suma de cuadrados de regresión**.

La suma de cuadrados de regresión se calcula como sigue:

$$SS_R=\hat{\beta}_1S_{xy}. \tag{2.27}$$

La cantidad de **grados de libertad** (gl) se determina como sigue:

$$gl_T=gl_R+gl_{Res} \tag{2.28}$$
$$n-1=1+(n-2) \tag{2.29}$$

esto es, la suma total de cuadrados, SS_T tiene $gl_T=n-1$ grados de libertad, la suma de cuadrados de regresión, SS_R, tiene $gl_R=1$ grado de libertad y SS_{Res} tiene $gl_{Res}=n-2$ grados de libertad.

Es posible aplicar la prueba F normal del **análisis de varianza** para probar la hipótesis $H_0:\beta_1=0 \quad \text{vs} \quad H_1:\beta_1\neq 0$.

Se tienen siguientes puntos importantes que mencionar:

1. $SS_{Res}/\sigma^2=(n-2)MS_{Res}/\sigma^2$ sigue una distribución χ^2_{n-2}.
2. Si es cierta la hipótesis nula $H_0:\beta_1=0$, entonces SS_R/σ^2 tiene una distribución χ^2_{n-2}.
3. SS_{Res} y SS_R son independientes por el Teorema de Cochran.

La estadística F se obtiene de la siguiente manera:

$$F_0 = \frac{SS_R/gl_R}{SS_{Res}/gl_{Res}} = \frac{SS_R/1}{SS_{Res}/(n-2)} = \frac{MS_R}{MS_{Res}} \tag{2.30}$$

sigue una distribucuón $F_{1,n-2}$.

Así, se tiene la tabla de **análisis de varianza para probar la significancia de la regresión.**

Tabla 2.1: Análisis de Varianza (ANOVA)(Montgomery, D.; Peck, E. & Vining, E., 2006).

Fuente de Variación	Suma de Cuadrados	gl	Cuadrado Medio	Estadístico F
Regresión	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	$\frac{MS_R}{MS_{Res}}$
Residual	$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}$	n-2	MS_{Res}	
Total	SS_T	n-1		

2.2.3. Modelo de regresión Lineal Múltple

Un modelo de regresión donde interviene más de una variable regresora se llama **modelo de regresión múltiple**. Un modelo de regresión lineal múltiple es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \tag{2.31}$$

donde y representa la variable dependiente, $x_i, i = 1, 2, ..., k$ las variables independientes. El modelo se llama **modelo de regresión lineal múltiple** con k regresores, los parámetros $\beta_j, j = 0, 1, 2, 3, ..., k$ se llaman **coeficientes de regresión** y cada uno, excepto β_0, representa el cambio esperado de acuerdo con el cambio unitario de la variable x_j cuando todas las demás variables regresoras x_j $(j \neq i)$ se mantienen constantes, observar que $j \neq i$. Este modelo describe un hiperplano en el espacio de k dimensiones de las variables regresoras x_i.

Reescribiendo el modelo muestral de regresión que corresponde a la ecuación (2.17)

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, i = 0, 1, 2, ..., n. \tag{2.32}$$

Puede expresarse el modelo por notación matricial como se muestra a continuación

$$y = X\beta + \varepsilon \tag{2.33}$$

donde:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}; \tag{2.34}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \tag{2.35}$$

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}; \tag{2.36}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}. \tag{2.37}$$

En general, y es un vector de n $\times$ 1 de las observaciones, X es una matriz de n $\times$ (k+1) de los niveles de las variables regresoras, β es un vector de k $\times$ 1 de los coeficientes de regresión y ε es un vector de n $\times$ 1 de errores aleatorios.

El estimador de β por mínimos cuadrados es

$$\hat{\beta} = (X'X)^{-1}X'y \tag{2.38}$$

siempre y cuando exista la matriz inversa $(X^tX)^{-1}$. Esta matriz existe si los regresores son linealmente independientes, esto es, si ninguna columna de la matriz X es una combinación lineal de las demás columnas.

El modelo ajustado de regresión que corresponde a los niveles de las variables regresoras $x' = [1, x_1, x_2, ..., x_k]$ es 1 x k y se tiene

$$\hat{y} = x'\hat{\beta} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j. \tag{2.39}$$

El vector de valores ajustados $\hat{y}_i$ que corresponden a los valores observados y_i es

$$\hat{y} = X\hat{\beta} = X\left(X'X\right)^{-1}X'y = Hy, \tag{2.40}$$

donde $H = X\left(X'X\right)^{-1}X'$ es matriz sombrero o matriz proyección.

La diferencia entre el valor observado y_i y el valor ajustado $\hat{y}_i$ correspondiente es el residual $e_i = y_i - \hat{y}_i$. Los n residuales se escriben con notación matricial como:

$$e = y - \hat{y}. \tag{2.41}$$

Hay otras maneras de expresar el vector de residuales e que son útiles, como:

$$e = y - X\hat{\beta} = y - Hy = (I - H)y. \tag{2.42}$$

Estimación de σ^2

Como en la regresión lineal simple, se desarrolla un estimador de σ^2 a partir de la suma de cuadrados de residuales:

$$SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = e'e. \tag{2.43}$$

Sustituyendo $e = y - X\hat{\beta}$ se obtiene

$$SS_{Res} = y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}. \tag{2.44}$$

Como $X'X\hat{\beta} = X'y$, la última ecuación se transforma en

$$SS_{Res} = y'y - \hat{\beta}'X'y. \tag{2.45}$$

El **cuadrado medio residual** es

$$MS_{Res} = \frac{SS_{Res}}{n-p} \tag{2.46}$$

donde p representa el número de parámetros en el modelo.

También se tiene que el **estimador insesgado** de σ^2 es

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res} \tag{2.47}$$

donde el estimador de σ^2 depende del modelo.

Prueba de significancia de la regresión

La prueba de la significancia es la extensión del caso univariado de la regresión y se usa para determinar si hay una relación lineal entre la respuesta y con cualquiera de las variables regresoras $x_1, x_2, ..., x_k$. Las hipótesis son:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0 \quad \text{para al menos una } j. \tag{2.48}$$

El rechazo de la hipótesis nula implica que al menos uno de los regresores $x_1, x_2, ..., x_k$ contribuye al modelo en forma significativa.

El procedimiento de prueba es una generalización del **análisis de varianza**. La **suma total de cuadrados** SS_T se divide en una **suma de cuadrados de residuales**, SS_{Res}. Así,

$$SS_T = SS_R + SS_{Res}. \tag{2.49}$$

Por definición SS_R es $\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$.

Fijando un nivel de significancia de 0.05, para probar la hipótesis H_0 se calcula el estadístico de prueba F_0 como sigue:

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}} \tag{2.50}$$

tiene una distribución $F_{k,n-k-1}$, donde MS_R es el cuadrado medio de la regresión.

Para probar la hiótesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, se calcula el estadístico de prueba F_0 y se rechaza H_0 si:

$$F_0 > F_{\alpha,k,n-k-1}. \tag{2.51}$$

El procedimiento de prueba se resume normalmente en una **tabla de análisis de varianza** que se muestra en el Cuadro 2.2.

R^2 y R^2 ajustada

Tabla 2.2: Tabla de Análisis de Varianza (ANOVA)(Montgomery, D.; Peck, E. & Vining, E., 2006)

Fuente de Variación	Suma de Cuadrados	gl	Cuadrado Medio	F_0
Regresión	SS_R	k	MS_R	$\frac{MS_R}{MS_{Res}}$
Residual	SS_{Res}	$n-k-1$	MS_{Res}	
Total	SS_T	$n-1$		

Una manera de evaluar la adecuación general del modelo son los estadísticos R^2 y R^2 ajustada, o bien, R^2_{Adj}.

La cantidad

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} \tag{2.52}$$

se le llama **coeficiente de determinación**. R^2 se interpreta como la proporción de la variación en la variable respuesta explicada por el modelo de regresión lineal múltiple con los regresores x. Ya que $0 \leq SS_{Res} \leq SS_T$, entonces, $0 \leq R^2 \leq 1$. Los valores de R^2 cercanos a 1 implican que la mayor parte de la variabilidad de y está explicada por el modelo de regresión.

La magnitud de R^2 también depende del intervalo de variabilidad de la variable regresora. En general, R^2 aumenta a medida que aumenta la dispersión de las x y disminuye cuando disminuye la dispersión de las x, siempre y cuando sea correcta la forma supuesta del modelo.

El estadístico R^2 ajustada se define:

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)}. \tag{2.53}$$

En vista de que $SS_{Res}/(n-p)$ es el cuadrado medio de residuales y $SS_T/(n-1)$ es constante, independientemente de cuantas variables hay en el modelo, R^2_{Adj} sólo aumentará al agregar una variable al modelo si esa adición reduce el cuadrado medio de residuales.

La R^2 ajustada penaliza la adición de términos que no son útiles, además que es ventajoso para evaluar y comparar los modelos posibles de regresión.

Pruebas sobre coeficientes individuales de regresión.

Las hipótesis para probar la significancia de cualquier coeficiente individual de regresión son:

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0. \tag{2.54}$$

Si no se rechaza la hipótesis nula $H_0 : \beta_j = 0$, quiere decir que se puede eliminar el regresor x_j del modelo. El **estadístico de prueba** para esta hipótesis es:

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma^2} C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{2.55}$$

donde C_{jj} es el elemento diagonal de $(X'X)^{-1}$ que corresponde a $\hat{\beta}_j$. Se rechaza la hipótesis nula $H_0 : \beta_j = 0$ si $|t_0| > t_{\alpha/2,n-k-1}$. Nótese que ésta es una prueba parcial o marginal.

2.3. Comprobación de la adecuación del modelo

Las principales premisas que se han hecho hasta ahora al estudiar el análisis de regresión son las siguientes:

1. La relación entre la respuesta y y los regresores es lineal, al menos en forma aproximada.

2. El término de error ϵ tiene media cero.
3. El término de error ϵ tiene varianza σ^2 constante.
4. Los errores no están correlacionados.
5. Los errores tienen distribución normal.

En conjunto, las hipótesis 4 y 5 implican que los errores son variables aleatorias independientes.

Es necesario tener en cuenta que la validez de estas premisas es dudosa, y se deben hacer análisis para examinar la adecuación del modelo que se haya desarrollado en forma tentativa. Las grandes violaciones a las premisas producen un modelo inestable, en el sentido que una muestra distinta podría conducir a un modelo totalmente diferente, y obtener conclusiones opuestas. En general, no se pueden detectar desviaciones respecto a las premisas básicas examinando los estadísticos estándar de resumen, como por ejemplo los estadísticos t, F o R^2. Éstas son propiedades "globales"del modelo, y como tal no aseguran la adecuación del mismo.

Existen distintas pruebas para realizar la comprobación de la adecuación del modelo algunas de las cuales son explicadas a continuación.

2.3.1. Prueba de Shapiro-Wilks (Shapiro & Wilk, 1965)

La prueba de Shapiro-Wilk es una prueba estadística utilizada para evaluar si una muestra de datos sigue una distribución normal. La hipótesis nula de la prueba es que los datos se distribuyen normalmente.

Se realiza la formulación de un juego de hipótesis de la siguiente manera:

$$\begin{aligned} H_0 &= \text{Los datos siguen una distribución normal.} \\ &\quad \text{vs} \\ H_1 &= \text{Los datos no siguen una distribución normal.} \end{aligned}$$

Se procede a realizar el cálculo del estadístico de prueba llamado "W" basado en los valores ordenados de la muestra.

Sea $m' = (m_1, m_2, ..., m_n)$ denotada por el valor esperado del valor de la normal estandar, y $V = (v_{ij})$ sea la correspondiente matriz de covarianza $n \times n$. Es decir, si $x_1 \leq x_2 \leq \cdots leqx_n$ denota una muestra aleatoria ordenada de tamaño n de una distribución normal con media 0 y varianza 1.

Sea $y' = (y_1, y_2, ..., y_n)$ denota un vector de observaciones aleatorias ordenadas. El objetivo es derivar una prueba para la hipótesis de que se trata de una muestra de una distribución normal con media desconocida μ y varianza desconocida σ^2.

Claramente, si y_i es una muestra normal, entonces y_i se expresa como:

$$y_i = \mu x_i + \epsilon \text{ donde } (i = 1, 2, ..., n).$$

El estadístico de prueba W para normalidad se define por:

$$W = \frac{R^4 \hat{\sigma}^2}{C^2 S^2} = \frac{b^2}{S^2} = \frac{(a'y)^2}{S^2} = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{2.56}$$

donde

$$
\begin{aligned}
&R^2 = M'V^{-1}m, \\
&C^2 = m'V^{-1}V^{-1}M, \\
&a' = (a_1, a_2, ..., a_n) = \frac{m'V^{-1}}{(M'V^{-1}V^{-1}M)^{1/2}} \\
&b = R^2\hat{\sigma}/C.
\end{aligned}
$$

Posteriormente de haber calculado el estadístico W, la prueba compara el valor W con una distribución de referencia que se deriva bajo la suposición de que los datos siguen una distribución normal.

Si el valor calculado W es significativamente menor que el valor crítico correspondiente (determinado por el nivel de significancia elegido y el tamaño de la muestra), entonces se rechaza la hipótesis nula, lo que significa que los datos no siguen una distribución normal. Por otro lado, si el valor calculado de W no es lo suficientemente pequeño como para rechazar la hipótesis nula, se concluye que los datos se distribuyen normalmente.

2.3.2. Prueba Breusch-Pagan (Hill, Griffiths y Lim, 2017)

Dos pruebas que son relativamente generales en heterocedasticidad son la prueba de Goldfeld-Quandt (1965) y la del multiplicador de Lagrange (LM) de Breusch-Pagan.

Se sabe que la prueba de Goldfeld-Quandt es sumamente útil para identificar correctamente la variable a utilizar en la separación de la muestra. Sin embargo, este requisito limita su generalidad.

Breusch y Pagan desarrollaron la prueba del multiplicador de Lagrange para la hipótesis de que $\sigma_i^2 = \sigma^2 f(\alpha_0 + \alpha' z_i)$ donde z_i es un vector de variables independientes.

La versión de la prueba LM se llama típicamente la prueba de Breusch-Pagan para heterocedasticidad (test BP). Breusch y Pagan propusieron una forma diferente de la prueba en donde se asume que los errores están distribuidos normalmente.

El modelo es homocedástico si $\alpha = 0$. La prueba se lleva a cabo con una regresión simple:

$LM = 1/2$ suma explicada de los cuadrados en la regresión de $e_i^2/(e'e/n)$ en z_i.

Para fines de cálculo, sea Z la matriz $n \times k$ de observaciones en $(1, z_i)$, y sea g el vector de observaciones de $g_i^2/(e'e/n)$. Luego:

$$LM = \frac{1}{2}[g'Z(Z'Z)^{-1}Z'g].$$

Bajo la hipótesis nula de homocedasticidad, LM sigue una distribución chi-cuadrada límite con grados de libertad iguales al número de variables en z_i.

Resumimos los pasos para realizar la prueba de heterocedasticidad utilizando la prueba Breusch-Pagan:

1. Estimar el modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$ mediante MCO (Mínimos Cuadrados Ordinarios) y posteriormente obtener los residuos al cuadrado estimados por MCO, u^2 (uno para cada observación).

2. Realizar la regresión en $u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + \epsilon$ y conservar el coeficiente de determinación R cuadrado de esta regresión.

3. Calcular la estadística F o la estadística LM y obtener el valor p. Si el valor p es lo suficientemente pequeño, es decir, está por debajo del nivel de significancia elegido, entonces se rechaza la hipótesis nula de homocedasticidad.

Si la prueba BP resulta en un valor p lo suficientemente pequeño, se toman medidas correctivas. Una posibilidad es simplemente utilizar los errores estándar robustos a la heterocedasticidad.

Si sospechamos que la heterocedasticidad depende solo de ciertas variables independientes, podemos modificar fácilmente la prueba de Breusch-Pagan: simplemente regresamos u_2 en las variables independientes que elijamos y realizamos la prueba F o LM apropiada. Recordemos que los grados de libertad apropiados dependen del número de variables independientes en la regresión con u_2 como variable dependiente; el número de variables independientes que aparecen en la ecuación $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$ es irrelevante.

Cuando se realiza una regresión lineal de los residuos al cuadrado con una única variable independiente, la prueba de heterocedasticidad se reduce a la estadística t estándar asociada con esa variable. La significancia de este valor t sugiere la posible presencia de heterocedasticidad. Un valor t estadísticamente significativo indica fuertemente que las variaciones en los residuos no son constantes, respaldando así la presencia de heterocedasticidad en el modelo.

2.3.3. Prueba RESET (Hill, Griffiths y Lim, 2017)

La prueba RESET (Regresión Equation Specification Error Test) es una prueba estadística en análisis de regresión que está diseñada para detectar variables omitidas y forma funcional incorrecta del modelo de regresión, así como posibles errores en la especificación del modelo, específicamente en relación con la forma funcional de la ecuación.

El propósito fundamental de la prueba RESET es evaluar si el modelo de regresión especificado es adecuado para representar la relación entre las variables predictoras y la variable de respuesta. Específicamente, la prueba se utiliza para detectar la presencia de posibles términos no lineales omitidos en la ecuación de regresión.

Se procede de la siguiente manera:

Se supone que se tiene especificado y estimado el siguiente modelo de regresión:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

Sean (b_0, b_1, b_2) las estimaciones de mínimos cuadrados, y sea

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

los valores ajustados de y. Se considera los siguientes dos modelos artificiales:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 f(x_1) + \epsilon. \tag{2.57}$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 f(x_1) + \gamma_3 f(x_2) + \epsilon \tag{2.58}$$

donde γ_1 y γ_2 son parámetros establecidos y $f(x_1)$ y $f(x_2)$ representa una función no lineal de las variables predictoras.

En la ecuación (2.57), el juego de hipótesis establecido es el siguiente

$$H_0 : \gamma_1 = 0 \quad \text{vs} \quad H_1 : \gamma_1 \neq 0.$$

En la ecuación (2.58), se tiene el siguiente juego de hipótesis

$$H_0 : \gamma_1 = \gamma_2 = 0 \quad \text{vs} \quad H_1 : \gamma_1 \neq 0 \text{ y/o } \gamma_2 \neq 0.$$

En el primer caso, se utiliza una prueba t o F. Se requiere una prueba F para el segundo caso. Si se rechaza H_0 implica que el modelo original es inadecuado y puede mejorarse. En caso contrario, de no rechazar H_0 indica que la prueba no ha podido detectar ninguna especificación errónea.

2.4. Transformación Box Cox (Box, & Cox, (1964)).

En algunas aplicaciones las transformaciones son importantes para modificar algún problema que esté ocurriendo. Existen diferentes técnicas para su corrección, en particular, se describe un procedimiento analítico que puede aplicarse tanto a la variable de respuesta o a las variables regresoras.

Supóngase que se desea transformar y para corregir que los datos no tengan normalidad y/o que la varianza de los errores sea no constante. Una clase útil de transformaciones es la **transformación de potencia** y^λ, donde λ, es un parámetro que se debe determinar. Box Cox indican cómo se estima en forma simultánea los parámetros del modelo de regresión y y^λ con el método de la máxima posibilidad.

La transformación es en realidad una familia de transformaciones indexadas por un parámetro λ:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, \lambda \neq 0 \\ \ln(y), \lambda = 0 \end{cases} \tag{2.59}$$

Esto también se conoce como transformación de potencia, ya que la variable se eleva a una potencia particular.

La familia "bcnPower", o Box-Cox con negativos, propuesta por Hawkins y Weisberg (2017) permite valores no positivos, al mismo tiempo que permite que los datos transformados se interpreten de manera similar a la interpretación de Box-Cox que transformó los valores.

Esta familia es la transformación Box-Cox de $z = .5 * (y + (y^2 + \gamma^2)^{1/2})$ que depende de un parámetro de ubicación γ. La cantidad de z es positiva para todos los valores de y. Si $\gamma = 0$ y y es estrictamente positiva, entonces las transformaciones de Box - Cox y bcnPower son idénticas. Al ajustar Box - Cox con la familia negativos, lambda se restringe al rango [-3,3] y gamma se restringe al rango desde .1 hasta el valor positivo más grande de la variable.

Un intervalo de confianza aproximado para λ

También se puede determinar un intervalo de confianza aproximado para el parámetro de transformación λ. Este intervalo de confianza puede servir para seleccionar el valor definitivo de λ.

Al aplicar el método de la máxima posibilidad al modelo de regresión, lo que en esencia se está maximizando es

$$L(\lambda) = -\frac{1}{2} n \ln[SS_{Res}(\lambda)] \tag{2.60}$$

o bien, lo que es igual, se está minimizando la función suma de cuadrados de residuales $SS_{Res}(\lambda)$. Un intervalo adecuado de confianza de $100\,(1-\alpha)$ por ciento para λ es el de todos aquellos valores que satisfacen la desigualdad

$$L(\hat{\lambda}) - L(\lambda) \leq \frac{1}{2}\lambda^2_{\alpha,1}/n \tag{2.61}$$

en donde $\lambda^2_{\alpha,1}$ es el punto porcentual superior de la distribución ji cuadrada con un grado de libertad. Para trazar el intervalo de confianza en realidad se trazaría, sobre una gráfica de $L(\lambda)$ en función de λ, una recta horizontal a la altura

$$L(\hat{\lambda}) - \lambda^2_{\alpha,1} \tag{2.62}$$

en la escala vertical. Esa línea cortaría a la curva de $L(\lambda)$ en dos puntos, cuyos lugares en el eje de λ definen los dos extremos del intervalo aproximado de confianza. Si se está minimizando la suma de cuadrados de residuales y graficando $SS_{Res}(\lambda)$ en función de λ, entonces la línea se debe graficar a la altura.

$$SS^* = SS_{Res}(\hat{\lambda}) e \lambda^2_{\alpha,1}/n. \tag{2.63}$$

2.5. Variables Dummy (Draper & Smith, 1998)

Las variables dummy o variables ficticias son variables que se usan para cuantificar los niveles de una variable de predicción cualitativa para su empleo en el análisis de regresión. A cada una de estas variables se les asignas los valores 0 y 1.

Las variables consideradas en un modelo de regresión usualmente toman valores sobre un rango continuo (variables cuantitativas). Ocasionalmente se debe introducir variables categóricas (o cualitativas) con dos o más categorías. Los dos valores significan que la observación pertenece a una de dos categorías. Las variables dummy o indicadoras sirven para identificar categorías o clase a las que pertenecen las observaciones.

En particular, una forma para explicar las variables dummy, si la variable cualitativa x_1 tiene dos valores de interés, A y B, se pueden crear dos variables indicadoras x_2 y x_3. Entonces, dado que sólo se tienen dos categorías, es conveniente definir dos variables indicadoras x_2 y x_3 tales, que

$$x_2 = \begin{cases} 1 & \text{si } x_1 = A \\ 0 & \text{de otro modo,} \end{cases}$$
$$x_3 = \begin{cases} 1 & \text{si } x_1 = B \\ 0 & \text{de otro modo.} \end{cases}$$

Por ende, para obtener una sola ecuación de regresión para ambas opciones, se deberá ajustar el modelo

$$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

Pero si se hace esto, entonces la matriz $X'X$ no tendría inversa. Una manera de evitar este problema es eliminar una de las dos variables indicadores y emplear solamente una, por ejemplo, x_2 en donde al igual que antes,

$$x_2 = \begin{cases} 1 & \text{si la variable toma el valor A} \\ 0 & \text{de otro modo.} \end{cases}$$

En otras palabras, para cualquier dato que tenga el valor A ($x_2 = 1$) o si toma el valor de B ($x_2 = 0$).

En general, si una variable cualitativa tiene m niveles, puede representarse por medio de $m-1$ variables indicadoras, asignando a cada una los valores de 0 y 1.

2.6. Métodos de selección de variables (Draper & Smith, 1998)

Existen distintos métodos estadísticos para seleccionar variables independientes en el modelo de regresión. Se idealiza realizar estos métodos para así obtener únicamente las variables regresoras más importantes y así obtener la mejor ecuación de regresión posible. No existe un procedimiento único para realizar esta acción.

El método Stepwise es una técnica de selección de variables en análisis de regresión que busca construir un modelo de regresión múltiple utilizando las variables más relevantes y significativas. Se comienza eligiendo una ecuación que contiene la mejor variable regresora x y posteriormente se intenta aumentar con adiciones de variables x a la vez, siempre y cuando estas adiciones sean relevantes para el modelo.

El método Stepwise es una estrategia sistemática para construir un modelo de regresión múltiple al seleccionar y refinar las variables predictoras de manera secuencial, buscando la mejor combinación de variables que explique la variabilidad en la variable de respuesta.

El procedimiento básico es el siguiente:

1. **Selección de la variable inicial:** Comienza eligiendo la variable independiente x que tiene la mayor correlación con la variable dependiente y. Luego, se ajusta un modelo de regresión lineal simple con esta variable. Se evalúa si la variable seleccionada es estadísticamente significativa en el modelo. Esto se hace mediante pruebas de hipótesis, como la prueba F o t, para verificar si la variable tiene un impacto significativo en la variable de respuesta.

2. **Añadir o Eliminar Variables:** Si la variable inicial es significativa, se mantiene en el modelo. Luego, se busca la siguiente variable predictora más relevante para agregar al modelo. Esto se hace mediante pruebas de F parciales que comparan el modelo actual con la inclusión de una nueva variable. Se examina todos los valores F parciales de todas las variables independientes que no están en la regresión. Se selecciona la variable independiente x con el valor más alto y se ajusta una segunda ecuación de regresión. Si la nueva variable mejora significativamente el modelo, se agrega. Se ajusta la significancia de la regresión general, se anota la mejora en el valor R^2 y los valores de las F parciales para ambas variables que ahora se encuentran en la ecuación (no solo la variable que recién ingresó) y se examina. Si alguna variable se vuelve no significativa, se considera su eliminación.

3. **Iteración:** El proceso se repite iterativamente, añadiendo y eliminando variables, hasta que no se pueda mejorar más el modelo o hasta que se cumpla algún criterio predefinido para detener el proceso.

4. **Registro de Cambios en R^2:** A medida que se agregan o eliminan variables, se registra cómo estos cambios afectan al coeficiente de determinación ($\mathbf{R^2}$) del modelo. Esto ayuda a evaluar la mejora en la capacidad de explicar la variabilidad en la variable de respuesta.

5. **Criterio de Parada:** El método se detiene cuando no se pueden agregar más variables o cuando se alcanza un punto donde no se pueden eliminar más variables sin empeorar significativamente el modelo.

6. **Selección del Modelo Final:** El modelo resultante es el que se considera más adecuado según los criterios de selección definidos previamente, como el valor de $\mathbf{R^2}$, la significancia de las variables y otros indicadores de ajuste.

Cabe aclarar que un predictor que haya sido el mejor candidato de entrada en una etapa anterior puede, en una etapa posterior, no ser una candidata para el modelo final, esto debido a las relaciones entre él y otras variables que se ingresan a la ecuación.

A medida que se ingresa cada variable en la regresión, generalmente se registra e imprime su efecto sobre R^2.

Capítulo 3

Aplicación del modelo

En el actual capítulo se presenta algunas situaciones que se presentaron durante la pandemia COVID - 19 en el ámbito estudiantil. También se exhibe información obtenida de una encuesta realizada a alumnos de la Facultad de Ciencias Físico Matemáticas de la Benemérita Universidad Autónoma de Puebla durante el periodo de Primavera 2023.

Siguiendo una estructura similar a la presentación del análisis estadístico descriptivo, se precede a explicar el proceso de desarrollo del modelo de regresión. Una vez que se han realizado las modificaciones necesarias a las variables consideradas y se han llevado a cabo las pruebas de validez preliminares, se entra en la etapa de construcción del modelo final.

3.1. Desafíos Educativos en Tiempos de COVID-19

La pandemia COVID-19 generó un impacto significativo en diversos aspectos en la vida cotidiana, laboral y académica. En específico, el ámbito académico presento una transformación como resultado a dicha problemática.

Durante años, el entorno estudiantil ha seguido ciertas temáticas recurrentes. En la presente sección se presenta el impacto que tuvo la pandemia en el entorno académico. La propagación del virus COVID-19 y las medidas de distanciamiento social que fueron requeridas para frenar su avance ocasionaron una transición abrupta hacia la educación a distancia y virtual. Dicha transformación obligó a estudiantes, docentes e instituciones en todo el mundo a adaptar de forma rápida a nuevas modalidades de estudio.

La pandemia COVID-19 ha venido generando cambios y disrupciones en amplios sectores de la actividad humana. La educación ha sido uno de los más afectados debido a la imposición administrativa del cierre total de los centros educativos en gran parte de los países del mundo. La modalidad de educación a distancia, fundamentalmente en soporte digital, vino a ofrecer soluciones de emergencia a dicha crisis (García Aretio, 2020).

Los sistemas educativos de todo el mundo han sufrido la pandemia tanto a nivel escolar como académico y, algo que no carece de importancia, las respuestas educativas y tecnológicas que han dejado ver una marcada disparidad. Los cierres de las instituciones educativas como medida para contener la pandemia de COVID-19 han llevado a un despliegue acelerado de soluciones de educación a distancia para asegurar la continuidad pedagógica en todos los niveles (López-Noriega et al., 2022).

La crisis evidenció las desigualdades en el acceso a recursos educativos que disponían los estudiantes y cuerpo académico. Algunos estudiantes no contaban con acceso a Internet con alta velocidad o dispositivos adecuados para el aprendizaje en línea, lo que pudo impactar negativamente su participación y aprendizaje.

3.1.1. Contextualización de la Pandemia COVID-19

Desde que fue necesario la suspensión de actividades escolares por motivos de la pandemia COVID-19, se tuvo un tiempo de espera considerablemente largo para la reanudación de actividades presenciales. Durante este largo tiempo, tanto el cuerpo académico de la universidad como la comunidad universitaria enfrentaron la necesidad de adaptarse a plataformas digitales y aprender a gestionar su tiempo y estudio de vida de forma diferente a la que se tenía acostumbra antes de la pandemia.

Este periodo de ajuste, que abarcó desde inicios de 2020 hasta la eventual transición a actividades presenciales, presentó un reto de forma importante para todos los individuos involucrados. Los docentes tuvieron la necesidad de adquirir habilidades en tecnología que permitieran continuar con un proceso de enseñanza de forma efectiva.

La crisis está exacerbando las disparidades educativas preexistentes al reducir las oportunidades para muchos de los niños, jóvenes y adultos más vulnerables, como aquellos que viven en áreas pobres o rurales, niñas, refugiados, personas con discapacidades y personas desplazadas forzosamente, para continuar con su aprendizaje (UN, 2020a).

Dicha modificación provocó importantes cambios en el rendimiento académico de los estudiantes, dado que se enfrentaron a diversas problemáticas que arrojo el encierro en los hogares, desde su nivel de compromiso, interacción con compañeros y profesores.

La reanudación gradual de las actividades presenciales se llevó a cabo de forma escalonada, impulsada por diversas circunstancias que variaban según cada institución educativa. En el caso específico de la Benemérita Universidad Autónoma de Puebla (BUAP), se implementó un proceso de regreso a las aulas por facultades, considerando las particularidades de cada una. Como parte de esta estrategia, se limitó el número de estudiantes que podían asistir presencialmente a la universidad en un mismo momento, a fin de mantener la sana distancia requerida.

En consonancia con esta dinámica, la Facultad de Ciencias Físico Matemáticas (FCFM) adoptó medidas específicas para abordar la situación. Se ofrecieron permisos especiales a aquellos estudiantes que enfrentaban dificultades para regresar a clases presenciales. Esta medida permitió a los estudiantes que lo necesitaban continuar con sus estudios de forma virtual durante un semestre adicional, reconociendo las circunstancias individuales y brindando flexibilidad en el proceso de regreso.

Este enfoque estratégico refleja la adaptabilidad y consideración que las instituciones educativas debieron ejercer en medio de la pandemia. El equilibrio entre la necesidad de continuar con la educación presencial de manera segura y la comprensión de las limitaciones individuales que presentaban los estudiantes, esto resultó en una respuesta que priorizaba tanto la salud como el bienestar académico.

El proceso de retorno a las actividades presenciales fue caracterizado por un enfoque gradual y flexible, adecuándose a las circunstancias específicas de cada institución.

3.1.2. Cambio en la Dinámica de Estudio

Jamás en la historia se produjo un cierre universal de instalaciones educativas presenciales como el sucedido con motivo de la pandemia provocada por el COVID-19. Según datos actualizados de la UNESCO, gobiernos de casi 200 países decretaron el cierre total o parcial de centros educativos (García, 2021).

La adaptación a la modificación radical que afectó a los sistemas educativos a nivel global conllevó la necesidad de implementar la enseñanza y el aprendizaje de manera remota en todos los niveles académicos en todo el país.

Se ha visto que muchas instituciones educativas han realizado una adaptación de su docencia pasando la modalidad presencial en remoto, con una interacción cara a cara pero mediada por videoconferencia, en muchos de los casos, sin realizar una transformación real de su docencia a la virtualización de la modalidad (Fardoun et al., 2020).

Aunque las preferencias difieren, y a menudo de forma muy marcada, para muchos estudiantes universitarios la educación en sí no se ve necesariamente perjudicada por el aprendizaje en línea, siempre que, por supuesto, se disponga del equipo adecuado y de acceso a internet. Lo que muchos

estudiantes echan de menos es la vida social en el campus. Las encuestas muestran que los niños de las escuelas de primaria a secundaria también echan de menos estar físicamente con otros niños, pero el cierre de los edificios escolares ha supuesto el fin de la educación para muchos de ellos. En todo el mundo, millones de niños carecen de acceso a computadores y/o a internet. Para muchos profesores, sobre todo encargados de grupos desde el último año del jardín de niños hasta el quinto grado de primaria, el hecho de verse obligados a impartir clases en línea resultaba, en el mejor de los casos, desorientador, e incluso hacía que los conocimientos adquiridos en el aula durante décadas resultaran repentinamente irrelevantes (Pinar, 2022).

Esta adaptación a los cambios generó una serie de dificultades de gran relevancia para todas las personas involucradas en el ámbito educativo. En el contexto universitario, los profesores que enfrentaron la necesidad de realizar esta transición se vieron ante un desafío fundamental que nunca antes había enfrentado. Aquellos que no pudieron adaptarse eficientemente fueron a veces percibidos como menos efectivos en la enseñanza por parte de la comunidad universitaria durante la pandemia. Por otro lado, aquellos docentes que contaban con la capacitación adecuada y lograron adaptarse de manera exitosa experimentaron una mejora en sus métodos de enseñanza que ayudó fuertemente a los estudiantes para sobrellevar esta situación.

La dinámica de estudio impuesta durante la pandemia resultó ser un factor crucial en el desempeño académico. Las dificultades surgidas debido a esta adaptación repercutieron en la forma en que los estudiantes y profesores interactuaron con el proceso educativo, lo que influyó en los resultados académicos. Esta nueva forma de aprendizaje a distancia transformó significativamente la concepción de la educación y resaltó la importancia de la flexibilidad y la preparación en tiempos de cambios inesperados (Pinar, 2022).

En el lado de los estudiantes, la modificación en la modalidad de estudio, que había sido constante a lo largo de su vida académica, trajo consigo dificultades notables. Mantener el mismo nivel de compromiso con los estudios se convirtió en un reto importante y esto sin mencionar los diversos problemas derivados de la transformación en sí. Cada estudiante experimentó desafíos especiales y únicos, como la lucha por encontrar tiempo suficiente para estudiar debido a responsabilidades laborales que ocupaban un espacio significativo en su rutina diaria. Además, la falta de acceso a redes de comunicación para asistir a clases virtuales. Estos y muchos otros problemas similares añadieron obstáculos adicionales que debieron enfrentarse los estudiantes.

La dinámica de estudio que se impuso durante la pandemia resultó ser un factor crucial en el desempeño académico. Las dificultades que surgieron debido a esta adaptación repercutieron en la forma en que los estudiantes y profesores interactuaron con el proceso educativo, lo que a su vez influyó en los resultados académicos. Esta nueva forma de aprendizaje a distancia transformó significativamente la manera en que se concebía la educación y resaltó la importancia de la flexibilidad y la preparación en tiempos de cambios inesperados.

En lo que respecta a las universidades, más allá de que muchas de ellas contaban en sus planes estratégicos con previsiones de futuro para la enseñanza online, la realidad es que muy pocas de ellas estaban realmente preparadas para implementar de urgencia un modelo educativo plenamente digitalizado (u-Multirank, 2020)

3.1.3. Desigualdad en el Acceso a Recursos

La UNESCO destaca que el cierre de las escuelas lleva también “desigualdades educacionales”. “En general, las familias con un nivel económico medio – alto, tienden a tener niveles más altos de educación y más recursos para compensar la falta de clases presenciales. En cambio, las familias de clase baja dedican más tiempo a la televisión y menos a las actividades interactivas y estimulantes a nivel cognitivo”. Por todo esto, en muchos casos, la educación online no funciona (García et al., 2020).

La transición obligada hacia las clases en línea durante la pandemia acentuó de manera sobresaliente la desigualdad existente de los recursos disponibles que contaban los estudiantes, a medida que se requerían para participar en las actividades escolares, así como clases, reuniones entre compañeros, por mencionar algunas. Esta disparidad se hizo aún más evidente debido a que, a pesar de que algunos estudiantes tenían los recursos necesarios para adquirir dispositivos electrónicos que les permitieran enfrentar de manera más efectiva este nuevo sistema de educación, como tablets, laptops y otros

dispositivos similares, un gran porcentaje de estudiantes no contaba con las mismas oportunidades para acceder a tales dispositivos.

En efecto, mientras algunos estudiantes tenían la capacidad de adaptarse a la educación en línea de manera más fluida gracias a la disponibilidad de tecnología debido a una disponibilidad financiera, otro segmento significativo de la población estudiantil se veía desfavorecido por la falta de acceso a estos recursos. Incluso se presentaron casos extremos en los cuales los estudiantes carecían por completo de dispositivos electrónicos necesarios para participar en sus clases virtuales.

Esta brecha tecnológica tuvo implicaciones profundas en el aprendizaje y el rendimiento académico de los estudiantes durante la pandemia. Aquellos con recursos limitados se encontraron en una posición desventajosa, ya que no solo tenían dificultades para asistir a las clases en línea, sino que también experimentaban obstáculos en la entrega de tareas, exámenes, proyectos y la participación en actividades interactivas. Como resultado, la igualdad de oportunidades educativas se vio comprometida, lo que generó que el rendimiento académico de los estudiantes tuviera un efecto de forma negativa, lo que indica de forma alarmante su bajo desarrollo.

La adopción forzada de las clases en línea durante la pandemia subrayó las disparidades existentes en cuanto a recursos tecnológicos entre los estudiantes. Lo que influyó de manera importante en el desarrollo académico de los estudiantes.

3.1.4. Impacto en la Salud Mental

En la región de las Américas, la población ha tenido que enfrentar la pérdida de sus seres queridos por COVID-19, una crisis económica sin precedentes que en muchas ocasiones ha causado la pérdida de trabajos y/o medios de sustento. Adicionalmente, todos hemos experimentado también el cierre de las escuelas, el aislamiento, el teletrabajo, el miedo a contagiarnos y una gran incertidumbre. Estas condiciones han influido y son factores que contribuyen a la generación de una serie de problemas de salud mental, entre ellos, un aumento de la ansiedad, depresión, problemas para dormir, aumento de consumo de alcohol, tabaco, drogas, sustancias y situaciones de violencia intrafamiliar. (Organización Panamericana de la Salud, Boletín Desastres N.131, s. f.).

El cambio hacia las clases en línea generó otro gran problema en términos de la salud mental de los estudiantes universitarios. Esta situación fue resultado de la ausencia de las relaciones sociales que solían tener en el entorno presencial y la limitación de actividades externas a la universidad. Esto es solo un ejemplo de los múltiples desafíos que el aislamiento social debido al COVID-19 trajo consigo.

Aunque había opciones para mantener cierto nivel de conexión entre los estudiantes a través de plataformas virtuales, que permitían a los estudiantes realizar reuniones o llamadas en línea, esta interacción no podía compararse a las actividades sociales que solían formar parte de la rutina diaria de forma presencial antes de la pandemia. La falta de contacto humano directo y la limitación de las experiencias compartidas afectaron negativamente la calidad de vida de los estudiantes.

Estos problemas contribuyeron a que los estudiantes experimentaran fatiga y falta de concentración, entre otros problemas, lo que repercutió en una disminución del rendimiento académico y a su vez, en su aprendizaje obtenido. En muchos casos, estas dificultades tuvieron consecuencias más graves en la salud mental de los estudiantes. La sensación de aislamiento, la falta de motivación y la dificultad para mantener una rutina equilibrada contribuyeron al deterioro del bienestar emocional y psicológico.

El cambio a las clases en línea durante la pandemia exacerbó los problemas relacionados con la salud mental de los estudiantes. La ausencia de relaciones sociales significativas y la falta de actividades extracurriculares afectaron su bienestar de manera profunda. La adaptación a una nueva forma de interacción y el manejo de los desafíos emocionales se convirtieron en aspectos cruciales en la vida de los estudiantes en este período de cambios y dificultades sin precedentes.

3.1.5. Perspectivas a Futuro

Todas las problemáticas mencionadas anteriormente son solo una muestra de los múltiples desafíos que surgieron tras el inicio de la pandemia. Este evento tuvo un impacto significativo en el mundo

contemporáneo, lo que llevó a la implementación forzada de modificaciones para afrontar la situación, no solo en un aspecto educativo, sino, en la vida cotidiana en sí.

En el ámbito educativo, el proceso que se llevó a cabo durante la pandemia permitió la adopción de nuevas herramientas y enfoques en las técnicas de enseñanza. La oportunidad de desarrollar competencias tecnológicas persistió incluso después de la reanudación de las actividades escolares presenciales en diversas instituciones educativas, esto fue ocupado como una herramienta más para el mejoramiento de técnicas de enseñanza por parte de los docentes, y al mismo tiempo, por parte de los estudiantes, que aprendieron a utilizar estas herramientas para un mejor desempeño académico. Estas herramientas se convirtieron en recursos valiosos, facilitando diversas tareas en distintas áreas de trabajo.

Asimismo, se abrió una perspectiva a considerar para futuros eventos que impidan la realización de actividades presenciales, como fue el caso ante la pandemia COVID-19. A través de la experiencia vivida, se estableció una base sólida para la implementación de nuevas modalidades de estudio ante adversidades de gran envergadura. Esta experiencia sirve como fundamento sólido para estar mejor preparados en caso de que se presente un acontecimiento similar en el futuro.

La pandemia catalizó una serie de cambios y desafíos en diversos aspectos de la sociedad, entre ellos el educativo. Aunque los obstáculos fueron notables, también se presentaron oportunidades para el desarrollo de habilidades tecnológicas y la adaptación a nuevas formas de enseñanza. Estas lecciones aprendidas resultarán valiosas para afrontar situaciones similares y mejorar la resiliencia en el sistema educativo en el futuro.

De cara al futuro, las instituciones deben desarrollar planes educativos sostenibles que puedan resistir los desafíos y las incógnitas de este u otros escenarios similares que pudieran acaecer (Johnson, Veletsianos y Seaman, 2020).

3.2. Carreras en la Facultad de Ciencias Físico Matemáticas

En la BUAP se encuentra la Facultad de Ciencias Físico Matemáticas (FCFM) que ofrece diversas carreras. Estas notables disciplinas incluyen:

- Actuaría.
- Matemáticas.
- Matemáticas Aplicadas.
- Física.
- Física Aplicada.

En el contexto del análisis propuesto, se asigna como variable dependiente (y) como la diferencia del promedio que existe al continuar con las actividades presenciales en la facultad posterior de la pandemia, con respecto del promedio previo la pandemia, dado que para obtener dicha variable es necesario que la muestra de estudiantes encuestados estuvieran inscritos en la facultad antes del inicio de la pandemia COVID-19. Al considerar esta restricción, la encuesta tuvo que ser aplicada a estudiantes de las generaciones del 2019 y anteriores.

3.3. Proceso

La encuesta se llevó a cabo durante los meses de enero y febrero de 2023, dirigiéndose a los estudiantes de la FCFM. A pesar de que el objetivo inicial era aplicar un muestreo aleatorio que reflejara la diversidad de carreras y generaciones, esta meta resultó inalcanzable debido a la amplitud de materias que los estudiantes eligen. Como resultado, la encuesta no pudo realizarse de manera aleatoria y, en consecuencia, se tuvo una muestra de conveniencia.

El cuestionario diseñado para la encuesta abarcó un conjunto de variables clave que se consideraron esenciales para el análisis y para el modelo. Estas variables comprenden diversos aspectos relevantes para el estudio y proporcionan una visión completa de la realidad estudiantil que se desea analizar.

Las preguntas del cuestionario se establecieron considerando los objetivos del estudio y la relevancia de cada variable. Las respuestas se eligieron mediante un análisis preliminar de datos, agrupando aquellas con baja participación para garantizar representatividad. Se buscó obtener información significativa para el análisis. Dichas variables que se incluyeron en la encuesta son las siguientes:

1. Carrera: Esta variable describe la carrera que el encuestado se encontraba estudiando en la Universidad en el momento que respondió la encuesta.
2. Ámbito familiar: Esta variable indica si el estudiante se encontraba en un entorno familiar durante el período de la pandemia, estableciendo una conexión entre su situación familiar y su entorno de estudio.
3. Género: Esta variable captura el género del encuestado, para caracterizar mejor a los alumnos en sus licenciaturas.
4. Originario de Puebla: Esta variable identifica si el estudiante es originario del estado de Puebla, lo que podría influir en su situación durante la pandemia debido a posibles diferencias en términos de movilidad y apoyo familiar.
5. Trabajo: Indica si el estudiante tuvo que trabajar durante el período de estudio, lo que permite entender cómo las responsabilidades laborales pueden haber afectado su rendimiento académico.
6. Ingreso: Esta variable cuantifica el ingreso económico promedio en el hogar del estudiante, lo que podría relacionarse con su capacidad para acceder a recursos adicionales durante la pandemia.
7. Internet: Describe si el estudiante tenía acceso a Internet en su hogar, un factor crucial dado el entorno virtual de la educación durante la pandemia.
8. Problemas familiares y/o sociales: Identifica si el estudiante enfrentó problemas familiares y/o sociales durante la pandemia, lo que podría tener un impacto en su bienestar emocional, e inclusive en su salud mental, lo que podría reflejarse en su rendimiento académico.
9. Covid: Indica si el estudiante contrajo la enfermedad COVID-19 durante la pandemia, lo que permite considerar cómo la salud personal podría haber influido en su desempeño.
10. Horas de estudio: Cuantifica las horas diarias que el estudiante dedicaba al estudio, proporcionando información sobre su nivel de dedicación académica.
11. Enfermedad: Esta variable aborda si el estudiante enfrentó otras enfermedades además de COVID-19 durante la pandemia, lo que podría tener un impacto adicional en su rendimiento.
12. Dificultad de profesores: Captura la percepción del estudiante sobre la dificultad de los cursos y la enseñanza de los profesores durante la pandemia.
13. Promedio de materias: Representa el promedio de materias que el estudiante cursaba por cada semestre, brindando información sobre la cantidad de trabajo escolar que realizaba el estudiante.
14. Beca: Indica si el estudiante recibió algún tipo de beca durante la pandemia, lo que podría estar relacionado con su situación financiera y dedicación académica.
15. Uso de Computadora: Describe si el estudiante disponía de su propia computadora, si la compartía o si no tenía acceso a una, lo que puede influir en su capacidad para participar en el aprendizaje en línea.
16. Conocimientos: Captura una apreciación personal del nivel de conocimientos y habilidades que el estudiante adquirió durante las clases en línea, en relación con las materias cursadas.
17. Capacitación de los profesores: Refleja la percepción del estudiante sobre la preparación de sus clases y capacitación tecnológica que tenían los profesores para impartir clases en línea.
18. Aplicaciones: Indica si el estudiante utilizó aplicaciones o herramientas tecnológicas en su proceso de aprendizaje durante la pandemia. Estas aplicaciones fueron divididas en variables dummy, las aplicaciones que fueron consideradas son: WhatsApp, Microsoft Teams, Classroom, Zoom, Virtual Horizon y BlackBoard.

19. Promedio 2022: Esta variable refleja el rendimiento académico del estudiante por medio de sus calificaciones obtenidas durante el período de Otoño de 2022. Proporciona una medida cuantitativa de cómo le fue al estudiante en sus asignaturas y cursos hasta dicho semestre. Este promedio puede influir en la evaluación global del desempeño académico del estudiante en comparación a otros periodos.

20. Promedio 2019: Esta variable se relaciona con el desempeño académico del estudiante en el período de Primavera de 2019. Ofrece una perspectiva retrospectiva sobre el rendimiento del estudiante en un momento anterior a la pandemia. Comparar este promedio con el desempeño posterior a la pandemia podría brindar información valiosa sobre cómo la interrupción educativa causada por la COVID-19 afectó su rendimiento.

Es importante destacar que las variables mencionadas son las variables originales que son consideradas, no obstante, esto no implica que serán incorporadas en el modelo de regresión lineal múltiple, ya que en la modelación estadística pueden surgir nuevas agrupaciones de las clases con el objetivo de presentar un modelo adecuado, el cual es presentado en el presente capítulo. Estos cambios han llevado a una reevaluación de las variables a incluir en el modelo final, lo que ha motivado la exclusión de algunas de las variables previamente consideradas. Esta selección cuidadosa es esencial para garantizar la precisión y relevancia del modelo resultante.

3.4. Análisis Descriptivo

De acuerdo a la información obtenida de las encuestas realizadas a estudiantes, es posible llevar a cabo un análisis descriptivo con respecto a las respuestas proporcionadas por los alumnos de la FCFM.

La realización de un análisis descriptivo de los datos se considera como una etapa de vital trascendencia en la preparación para el modelo de regresión lineal múltiple. A través de este análisis, se obtiene un entendimiento de la naturaleza de los datos antes de la realización de modificaciones para la construcción del modelo. De esta manera, se busca responder a una serie de interrogantes acerca de los encuestados y su realidad. El análisis descriptivo provee información crucial para la toma de decisiones informadas en el proceso de modelado estadístico.

En un primer paso, se extrae la información concerniente a la variable dependiente (y), la cual representa la diferencia entre los promedios de los estudiantes antes y después del surgimiento de la pandemia. Esta relación es expresada mediante la siguiente fórmula, la cual fue utilizada para obtener la variable dependiente.

$$y = \text{Promedio 2022} - \text{Promedio 2019} \tag{3.1}$$

donde Promedio 2022 es el promedio del estudiante al inicio de 2022, y Promedio 2019 es el promedio que llevaba el estudiante hasta el 2019.

En la Figura 3.1, se observa que la mayoría de los estudiantes, constituyendo el 35.1 % del total, lograron mantener o incluso mejorar sus promedios en al menos 0.5 puntos, ubicándose en el rango de 0 a 0.5. Asimismo, un 31.9 % de los estudiantes experimentó una disminución en sus promedios en el intervalo de 0.01 a 0.5. Es relevante destacar que esta información se basa en una muestra, por lo que no se pueden extraer conclusiones definitivas sobre la totalidad de la población.

El tercer rango de frecuencia incluye a aquellos estudiantes cuyos promedios aumentaron en el intervalo de 0.05 a 1.00. En este rango se encuentra un 15.4 % de los estudiantes encuestados.

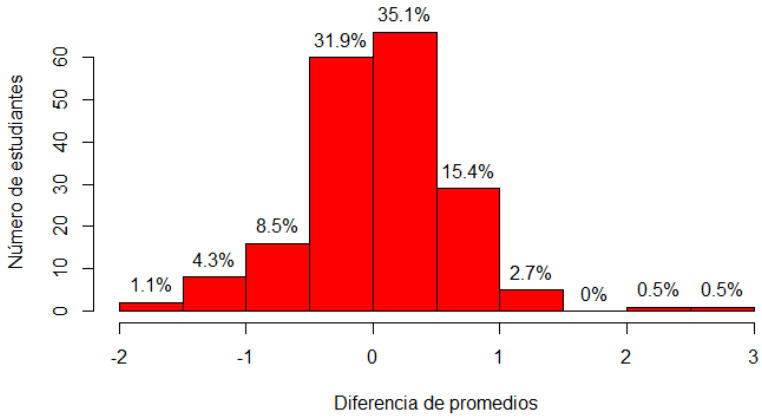


Figura 3.1: Rendimiento de los estudiantes antes y después de la pandemia. Fuente[Propia].

Estos resultados proporcionan una visión de cómo se distribuyen los cambios en los promedios de los estudiantes y muestran una variedad de respuestas a las condiciones cambiantes durante la pandemia.

Cabe mencionar nuevamente que esta información se basa en una muestra específica, por lo que cualquier generalización a la población completa no puede realizarse con este análisis.

En la Figura 3.2 se muestra el análisis de las horas diarias que los estudiantes destinaban al estudio. Se puede observar que la mayoría de los estudiantes dedicaban entre 0 y 2 horas al día, representando aproximadamente el 41.5 % de los encuestados.

Mientras tanto, un 25 % de los estudiantes invertía entre 2 y 4 horas, mientras que un pequeño porcentaje, alrededor del 17.6 %, destinaba entre 4 y 6 horas diarias al estudio. Sin embargo, el restante 15.9 % de la población encuestada, estudiaba más de 6 horas al día. Esto sugiere que la mayoría de los encuestados se concentraba en un rango moderado de horas de estudio diario.

En general, se puede inferir que a medida que los estudiantes invierten más horas en sus estudios, es probable que obtengan un rendimiento académico más sólido y positivo. Sin embargo, es importante destacar que esta relación puede verse afectada por otros factores y variables que también deben ser considerados en el análisis final.

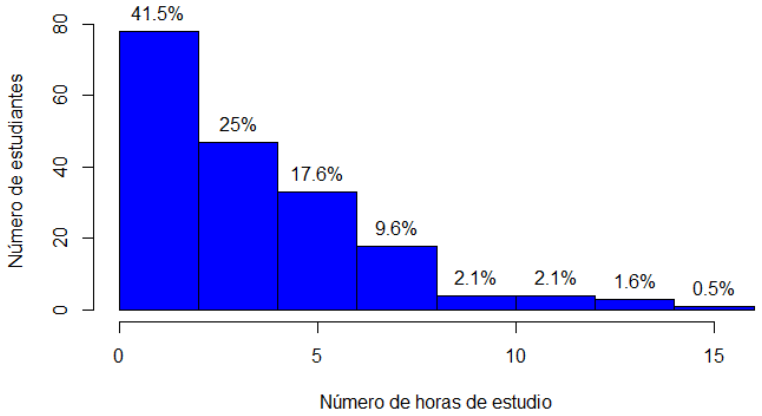


Figura 3.2: Horas de estudio diarias de los estudiantes. Fuente[Propia].

Continuando con la descripción, resulta importante identificar las principales carreras a las que pertenecen los estudiantes encuestados. Esto se ilustra mediante un gráfico de pastel en la Figura 3.3.

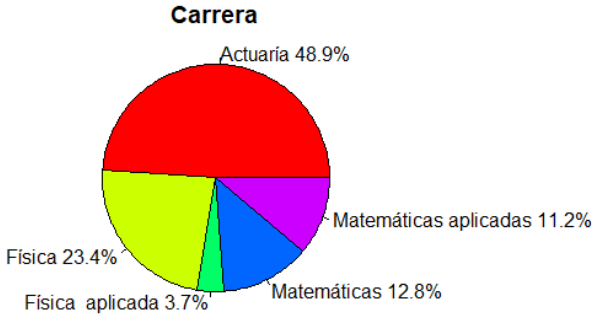


Figura 3.3: Carreras de los estudiantes encuestados. Fuente[Propia].

Al observar detalladamente la Figura 3.3, se desprende que la mayor proporción de estudiantes encuestados está matriculada en la carrera de Actuaría, constituyendo un 48.9 % del total. En un segundo plano se encuentra la carrera de Física, con un 23.4 % de representación. El resto de carreras contribuyen con el 27.7 % restante de la muestra, aunque su distribución no es equitativa.

La precisión y confiabilidad del modelo son elementos fundamentales para obtener conclusiones sólidas y respaldadas por datos. Por esta razón, la selección cuidadosa de las variables que se incorporan al modelo emerge como una estrategia esencial para asegurar la validez y robustez de los resultados. Es importante reconocer que la toma de decisiones en la construcción del modelo debe ser guiada por el compromiso de alcanzar conclusiones significativas y confiables en el análisis de los datos recopilados.

Se prosigue a realizar un análisis sobre el género de los estudiantes encuestados, esto se muestra a continuación:

La Figura 3.4 muestra que el 37,8 % de los encuestados se identificaron como género femenino, mientras que el 61,7 % se identificaron como género masculino. Sin embargo, se observa que una pequeña proporción, solo el 0,5 %, se identificó con un género diferente. En esta categoría, se identificó una única respuesta etiquetada como "Otro". Es importante señalar que esta respuesta proviene de un solo encuestado, por lo que si se mantiene esta categoría, no hay información suficiente para estimar el parámetro correspondiente en el modelo de regresión múltiple.

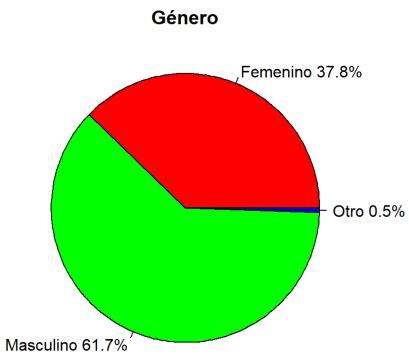


Figura 3.4: Género de los estudiantes encuestados. Fuente[Propia].

Para asegurar la integridad y la confiabilidad del análisis, se ha tomado la decisión de excluir las respuestas asociadas a esta única persona. Esta medida se toma con el propósito de preservar la validez

y la coherencia de los resultados del modelo, sin ninguna intención discriminatoria. La precisión y la integridad del modelo son esenciales para obtener resultados significativos y confiables en el análisis estadístico. Esta acción busca salvaguardar la robustez de los resultados y garantizar que el análisis se base en datos sólidos y representativos.

Prosiguiendo con la descripción, se realiza una visualización gráfica de la variable "Ámbito Familiar", la cual informa si el estudiante se encontraba dentro de su entorno familiar durante la pandemia.

La Figura 3.5 revela que el 89 % de los encuestados estaban en un entorno familiar durante la pandemia, en contraste con el 11 % que no lo estaba. Debido a que una categoría tiene un porcentaje menor a 20 %, esta variable no se considerará en el modelo.

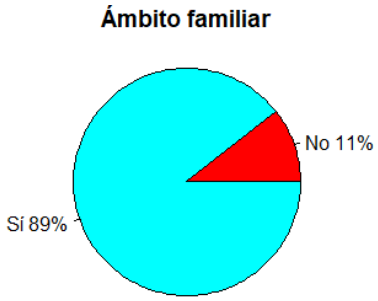


Figura 3.5: Ámbito familiar del estudiante. Fuente[Propia].

Con ello, se continua con una visualización gráfica de la variable "Trabajo", la cual muestra que porcentaje de los estudiantes trabajaron durante la pandemia.

La Figura 3.6 ilustra que la variable "Trabajo" presenta porcentajes similares de respuestas. Aproximadamente el 55 % de los estudiantes indicaron que no trabajaron, mientras que el restante 45 % afirmó estar trabajando durante la pandemia.

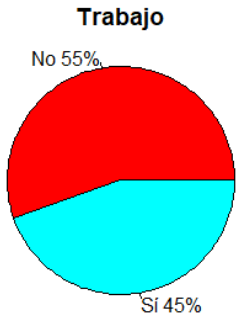


Figura 3.6: Trabajo. Fuente[Propia].

Por otro lado, en cuanto a la variable "Internet", los resultados arrojaron los siguientes porcentajes de respuestas. No sorprendentemente, la Figura 3.7 resalta un alto porcentaje de estudiantes, llegando al 91 %, que tenían acceso a Internet en sus hogares. Por otro lado, un modesto 9 % carecía de esta herramienta.

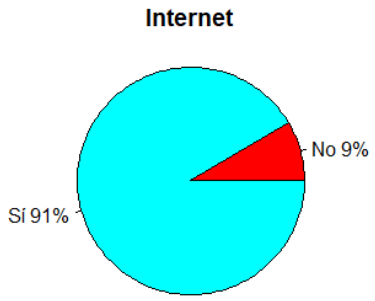


Figura 3.7: Internet. Fuente[Propia].

Se procede a realizar un análisis en la siguiente figura de la variable "Dificultad de Profesores".

Al analizar el contexto de la Figura 3.8, que presenta el nivel de dificultad percibido por los estudiantes de los profesores en las materias cursadas durante la pandemia, se puede inferir que una mayoría con el 77 %, consideró que el nivel de dificultad era "Medio". A su vez, aproximadamente el 20 % expresó que la dificultad fue clasificada como "Alto". Por otro lado, un modesto 3 % indicó que el nivel de dificultad era "Bajo".

Figura 3.8: Nivel de dificultad de los profesores. Fuente[Propia].

Debido a que una categoría de esta variable tiene un porcentaje menor a 20 % se combinaron las categorías "Medio" y "Bajo" en "Medio-Bajo". Este enfoque busca abordar el problema de la escasez de respuestas en la categoría "Bajo" y al mismo tiempo, permitir que la variable sea considerada para su inclusión en el modelo de regresión.

En la Figura 3.9, se presenta el uso de la computadora que el estudiante tenía disponible para su uso durante el período de la pandemia. Esto abarca aspectos como si el estudiante tenía una computadora propia, compartida, entre otros casos.

Figura 3.9: Uso de computadora. Fuente[Propia].

Aproximadamente un notable 62.2 % de los encuestados afirmó contar con su propia computadora, lo que señala una amplia disponibilidad de esta herramienta. Por otro lado, el 21.8 % compartía ocasionalmente una computadora, mientras que el 12.2 % tenía acceso a una computadora de uso compartido. Es interesante notar que solo un 3.7 % de los encuestados no tenía acceso a una computadora.

Dado que el porcentaje correspondiente a la respuesta "No tenía" es menor al 20 % que se marcó como aceptable, la estrategia ideal sería combinar esta respuesta con otra categoría para optimizar la viabilidad del modelo. Sin embargo, debido a las limitaciones de la herramienta, se optó por mantener esta respuesta en el análisis. Este aspecto resalta la importancia de considerar diversos factores al evaluar la pertinencia de las variables en un modelo de regresión, teniendo en cuenta tanto la calidad de los datos como su representatividad.

Figura 3.10: Competencias Tecnológicas de los Profesores. Fuente[Propia].

La Figura 3.10 revela que aproximadamente el 73 % de los estudiantes perciben un nivel medio en las competencias tecnológicas de los profesores durante la pandemia. Contrariamente, el 25 % considera que estas habilidades son "bajas", mientras que un modesto 2 % de los estudiantes opina que son "altas". Con el mismo criterio mencionado en variables anteriores, se considera la unificación de las respuestas "Medio" y "Alto" por "Medio-Alto". Estos resultados ofrecen una visión de la percepción general de los estudiantes con respecto a la adaptación y habilidades tecnológicas de sus profesores durante un período desafiante como la pandemia.

La variabilidad en las respuestas puede proporcionar información valiosa sobre las expectativas y necesidades de los estudiantes en entornos de enseñanza en línea. Esto puede influir en la forma en

que se desarrollan y ofrecen los cursos en línea en el futuro, con el fin de abordar las demandas y expectativas de los estudiantes de manera más efectiva y brindar una experiencia educativa en línea de alta calidad.

Asociación de las variables independientes con la variable respuesta

Una vez concluido la descripción gráfica en relación con datos que serán empleados para la construcción del modelo, surge una necesidad de determinar la asociación de las variables independientes con la variable respuesta.

Estas asociaciones desempeñan un papel en la metodología del modelo, ya que brindan una visión adelantada de cómo el modelo se comporta en comparación con los datos reales.

En la siguiente etapa, se procederá a la representación visual de gráficos clave.

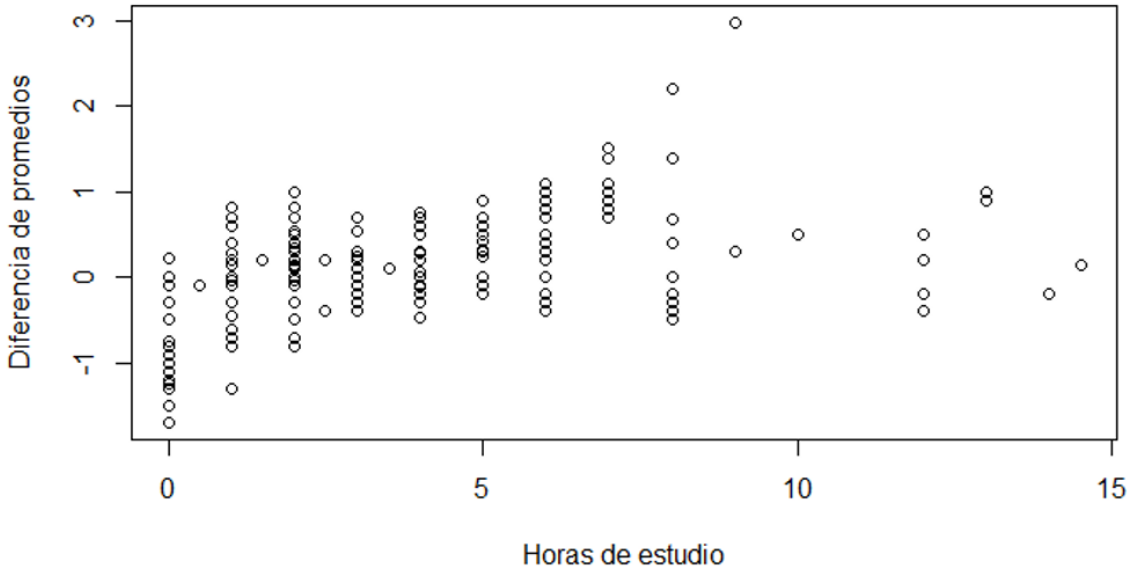


Figura 3.11: Diagrama de dispersión entre *Horas de estudio* y *Diferencia de promedio*. Fuente[Propia].

La Figura 3.11 presenta un gráfico de dispersión que muestra la relación entre la variable independiente *Horas de estudio* y la variable dependiente "y". Aunque se observa un patrón general de dispersión, la relación lineal se puede observar a simple vista para los primeros valores de *Horas de estudio*, sin embargo, no es completamente evidente dicho patrón conforme aumenta los valores en la variable independiente. Los puntos tienden a dispersarse en ciertas direcciones, lo que sugiere cierta variabilidad en la relación entre estas dos variables.

Es importante notar que aunque no sea inmediatamente claro, este tipo de visualización proporciona un primer vistazo sobre la posible relación entre las variables. Sin embargo, para tomar decisiones fundamentadas, será necesario llevar a cabo un análisis más profundo, como la construcción del modelo de regresión lineal y la interpretación de sus resultados.

Continuando con el análisis visual, en la Figura 3.12 se exhibe otro gráfico de dispersión que ilustra la relación entre la variable independiente *Promedio de Materias* y la variable dependiente "y". Aquí también se busca identificar patrones y tendencias visuales que puedan indicar una posible relación lineal entre estas dos variables.

En la Figura 3.12, se aprecia una cierta tendencia lineal en la dispersión de puntos. A simple vista, parece que a medida que aumenta la variable "y" (diferencia en el promedio antes y después de la pandemia), tiende a disminuir el número de materias tomadas por los alumnos.

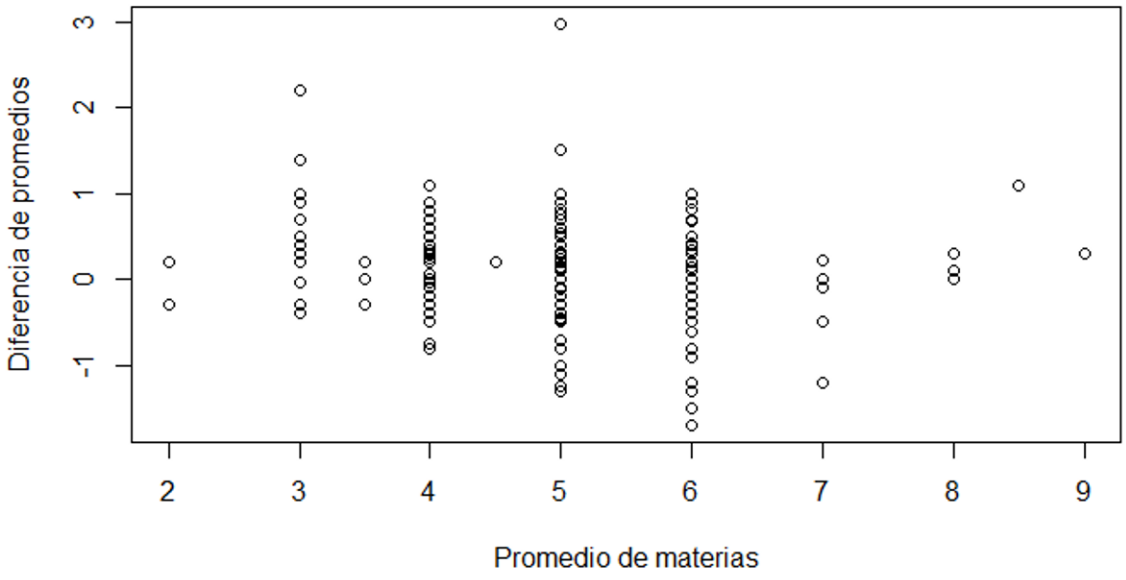


Figura 3.12: Diagrama de dispersión entre *Promedio de Materias* y *Diferencia de Promedio*. Fuente[Propia].

La interpretación visual de la gráfica es solo un primer paso en el proceso. La siguiente fase involucrará la creación y evaluación de un modelo de regresión lineal múltiple para confirmar si esta relación es estadísticamente significativa y si puede ser generalizada a través de más datos. La conclusión final se basará en un enfoque más riguroso y completo de la estadística y el análisis.

En este subcapítulo, se ha usado a la estadística descriptiva que ha permitido resumir de manera concisa las características clave de los datos, y ha brindado una base sólida para tomar decisiones. Las representaciones visuales han añadido una dimensión adicional a la comprensión.

3.5. Desarrollo del modelo

En esta siguiente etapa del proceso, se inicia la construcción y desarrollo del modelo de regresión lineal múltiple. Esta sección marca la transición de la descripción general de los datos hacia la elaboración de un marco analítico sólido y riguroso que permitirá entender y evaluar las relaciones entre las variables involucradas.

Dando inicio con el desarrollo del modelo, solo serán tomadas en consideración únicamente las variables que han superado las etapas de análisis y selección previas.

Para la estructuración de los modelos, se usara la siguiente notación:

- x_1 = Genero Masculino
- x_2 = Originario_de_Puebla Sí
- x_3 = Trabajo Sí
- x_{4-1} = Ingreso Entre \$17,500 y \$22,000
- x_{4-2} = Ingreso Mas de \$22,000
- x_{4-3} = Ingreso Menos de \$11,500
- x_5 = Problemas_fam_soc Sí
- x_6 = COVID Sí

- x_7 = Horas_estudio
- x_8 = Enfermedad Sí
- x_9 = Nivel_Dificultad Medio-Bajo
- x_{10} = Promedio_Materias
- x_{11} = BecaSí
- x_{12-1} = Uso_Computadora No tenía
- x_{12-2} = Uso_Computadora Solo para mi
- x_{12-3} = Uso_Computadora Todos la usabamos
- x_{13} = ConocimientosSí
- x_{14} = Cap_Profesores Medio Alto
- x_{15} = BlackBoard
- x_{16} = Virtual_Horizon
- x_{17} = Zoom
- x_{18} = Classroom
- x_{19} = Microsoft_Teams
- x_{20} = WhatsApp

Es importante destacar que en el modelo de regresión se incluirán variables dummy. Un ejemplo específico es el caso de la variable x_1, la cual representa el género masculino. En la ecuación del modelo de regresión, si esta variable es verdadera, se asignará un valor de 1, que se multiplicará por el correspondiente valor de β. Similarmente, existen varias variables de naturaleza similar que se incorporarán de manera análoga en el modelo. Estas variables dummy son esenciales para capturar la influencia de factores categóricos en la variable dependiente.

Al considerar estas variables como dummy, se establece una categoría de referencia para cada una de las variables independientes. En el contexto específico:

x_1 (*Género*) tiene como categoría de referencia "Femenino".
x_2 (*Originario de Puebla*) tiene como categoría de referencia "No".
x_3 (*Trabajo*) tiene como categoría de referencia "No".
x_4 (*Ingreso*) tiene como categoría de referencia el rango "Entre 11500 y 17500".
x_5 (*Problemas familiares o sociales*) tiene como categoría de referencia "No".
x_6 (*COVID*) tiene como categoría de referencia "No".
x_8 (*Enfermedad*) tiene como categoría de referencia "No".
x_9 (*Nivel de Dificultad*) tiene como categoría de referencia "Alto".
x_{11} (*Beca*) tiene como categoría de referencia "No".
x_{13} (*Conocimientos*) tiene como categoría de referencia "No".
x_{14} (*Capacidad de los Profesores*) tiene como categoría de referencia "Bajo".

Las variables 15 a 20 asignarán el valor 1 si son verdaderas y el valor 0 si no lo son. Estas asignaciones se basan en la comparación de cada variable con su respectiva categoría de referencia. Este enfoque permite incorporar estas variables categóricas de manera efectiva en el modelo de regresión, cuantificando su impacto en la variable dependiente.

La elección de incluir solamente estas variables está fundamentada en su capacidad para ofrecer información relevante y significativa en la predicción de la variable dependiente. Al enfocarnos en

este conjunto específico de variables, nuestro objetivo es simplificar el modelo y mejorar su eficacia al descartar aquellas que no han demostrado una relación sólida con la variable que se desea predecir. Este enfoque contribuye a crear un modelo más conciso y preciso, permitiéndonos concentrarnos en las variables que realmente influyen en nuestro objetivo.

El modelo ajustado inicial se presenta a continuación:

$$\begin{aligned}\widehat{y} = 0.389 &- 0.006x_1 - 0.018x_2 - 0.125x_3 - 0.013x_{4-1} + 0.199x_{4-2} \\ &- 0.192x_{4-3} - 0.091x_5 + 0.069x_6 + 0.061x_7 - 0.081x_8 \\ &+ 0.015x_9 - 0.052x_{10} + 0.071x_{11} + 0.046x_{12-1} - 0.018x_{12-2} \\ &- 0.415x_{12-3} + 0.083x_{13} + 0.016x_{14} + 0.107x_{15} + 0.058x_{16} \\ &- 0.107x_{17} + 0.068x_{18} - 0.143x_{19} - 0.04x_{20}. \quad (3.2)\end{aligned}$$

Estos resultados representan pilares clave en la evaluación y validación del modelo de regresión lineal múltiple, por ello es de vital importancia analizarlos.

Tabla 3.1: Resultados de la regresión. Fuente [Propia].

Estadístico F	p-value
5.32	2.7e-11

El modelo de regresión lineal múltiple exhibe los resultados mostrados en el Cuadro 3.1, el cual se observa el estadístico F, el cual es significativo y el valor p asociado es muy bajo. Este hallazgo permite concluir que el modelo en su conjunto es significativo y agrega información en términos de predecir la variable de respuesta. Esto sugiere que al menos una de las variables predictoras está teniendo un impacto en la variable dependiente, lo cual es una base sólida para proceder con un análisis más detallado de los resultados.

A continuación, se presenta el Cuadro 3.2 con los resultados de la regresión:

Tabla 3.2: Resultados de la regresión. Fuente [Propia].

Variable	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3890	0.5895	0.6599	0.5102
GeneroMasculino	-0.0065	0.0838	-0.0779	0.9380
Originario_de_PueblaSí	-0.0183	0.0796	-0.2305	0.8180
TrabajoSí	-0.1255	0.0931	-1.3481	0.1795
IngresoEntre $17,500 y $22,000	-0.0135	0.1172	-0.1149	0.9087
IngresoMas de $22,000	0.1997	0.1173	1.7027	0.0906
IngresoMenos de $11,500	-0.1921	0.1034	-1.8582	0.0650
Problemas_fam_socSí	-0.0918	0.0869	-1.0561	0.2925
COVIDSí	0.0697	0.0852	0.8188	0.4141
Horas_estudio	0.0612	0.0139	4.4000	0.0000
EnfermedadSí	-0.0818	0.0882	-0.9280	0.3548
Nivel_DificultadMedio Bajo	0.0151	0.1048	0.1442	0.8855
Promedio_Materias	-0.0526	0.0365	-1.4422	0.1512
BecaSí	0.0713	0.1056	0.6753	0.5004
Uso_ComputadoraNo tenía	0.0465	0.2323	0.2004	0.8415
Uso_ComputadoraSolo para mi	-0.0182	0.1124	-0.1618	0.8717
Uso_ComputadoraTodos la usabamos	-0.4152	0.1428	-2.9073	0.0042
ConocimientosSí	0.0832	0.1002	0.8308	0.4073
Cap_PorfesoresMedio Alto	0.0169	0.0932	0.1813	0.8563

Tabla 3.2: Resultados de la regresión (continuación).

Variable	Estimate	Std. Error	t value	Pr(>|t|)
BlackBoard	0.1074	0.0782	1.3730	0.1717
Virtual_Horizon	0.0583	0.0986	0.5917	0.5549
Zoom	-0.1080	0.1773	-0.6090	0.5434
Classroom	0.0683	0.1591	0.4290	0.6685
Microsoft_Teams	-0.1435	0.6183	-0.2321	0.8168
WhatsApp	-0.0402	0.0910	-0.4422	0.6589

Avanzando en el análisis, se llevan a cabo una serie de comprobaciones para evaluar si se cumplen los supuestos asociados a un modelo de regresión lineal múltiple. Estas verificaciones son esenciales para asegurar la validez y la robustez de los resultados obtenidos a través del modelo.

Verificación de los supuestos.

Prueba de Normalidad:

En primer lugar, se realiza una evaluación de la normalidad de los residuos mediante la utilización de un gráfico que facilita la observación de la distribución de las colas de dichos residuos. Esta representación gráfica inicial ofrece una indicación visual de si los residuos siguen una distribución normal o si presentan desviaciones notables.

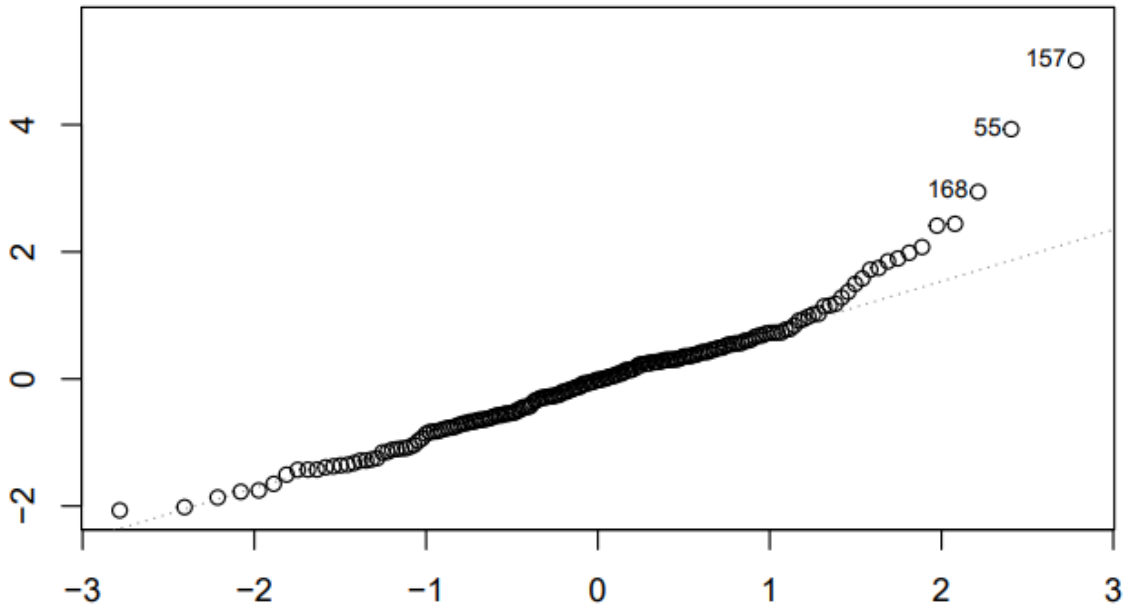


Figura 3.13: Gráfica de probabilidad normal. Fuente [Propia].

Inicialmente, al analizar la Figura 3.13, se observa una asimetría positiva en la distribución de los datos, aunque en ambas colas no siguen la tendencia.

Para una evaluación más precisa de esta cuestión, resulta de gran interés llevar a cabo la prueba de normalidad de Shapiro-Wilk, la cual permite determinar si los residuos siguen una distribución normal. A continuación, en el Cuadro 3.3 se presentan los resultados de dicha prueba:

Tabla 3.3: Resultados de la prueba Shapiro-Wilk. Fuente [Propia].

W	p.value
0.93364	1.506e-07

Los resultados de la prueba de Shapiro-Wilk brindan información crucial acerca de la normalidad de los residuos, y a partir de estos resultados, se pueden destacar los siguientes puntos relevantes:

Estadístico W: Un valor cercano a 1 indica que los datos se ajustan bien a una distribución normal. En este caso, se obtuvo que el valor de W fue 0.93364.

Basado en los resultados del test de Shapiro-Wilk, donde el valor p es muy bajo, con un valor de $1.506e-07$, la hipótesis nula de que los residuos siguen una distribución normal es rechazada. Esto sugiere que los residuos del modelo no cumplen con la suposición de normalidad y esta discrepancia podría tener implicaciones en la interpretación de los resultados del modelo de regresión, ya que la suposición de normalidad es importante para muchas pruebas estadísticas y análisis.

La falta de normalidad en los residuos puede afectar la validez de las pruebas de hipótesis, la precisión de las estimaciones de los coeficientes y la confiabilidad de las inferencias realizadas a partir del modelo. Para abordar esta situación, es necesario aplicar ajustes al modelo con el objetivo de lograr una normalización de la distribución de los residuos. Esto podría implicar la exploración de transformaciones en los datos o la consideración de métodos estadísticos alternativos que no dependan tanto de la suposición de normalidad.

La normalidad de los residuos es fundamental ya que influye en la validez de las pruebas de hipótesis y en la precisión de las estimaciones de los coeficientes. Para corregir esta desviación, pueden implementarse técnicas como transformaciones en los datos o la inclusión de variables adicionales que capturen las no linealidades presentes. Hay que tomar medidas para garantizar que los resultados del modelo sean confiables y sólidamente fundamentados.

Transformación de la variable dependiente $\widehat{y}$.

Después de aplicar la transformación, es esencial volver a realizar las pruebas de normalidad de los residuos y verificar si se ha logrado una mejora significativa en la normalidad. Además, se deben evaluar los resultados del modelo transformado y considerar si los coeficientes y las inferencias resultantes siguen siendo válidos y coherentes con la teoría y el contexto del problema.

La variable respuesta contiene valores negativos, lo que plantea desafíos al aplicar la transformación de Box-Cox. Para abordar esta situación, se utilizó la función powerTransform con la opción family = "bcnPower". Esta elección permite determinar dos valores: lambda y gamma.

De acuerdo a lo anterior, se procede a analizar los resultados obtenidos para la transformación Box-Cox, donde se destaca lo siguiente:

- El valor de λ estimado es -0.1793 para la variable de respuesta $\widehat{y}^{\lambda}$.
- El valor estimado de γ es 0.9714 para la variable de respuesta $\widehat{y}^{\lambda}$.

Estos resultados indican que se aplicó una transformación de potencia de Box-Cox con un valor λ estimado de -0.1793 para ajustar la distribución de la variable de respuesta y. La transformación de la variable dependiente $\widehat{y}$ se expresa como:

$$\widehat{y}^{\lambda} = \frac{\widehat{y}^{-0.1793} - 1}{-0.1793}.$$

Con la transformación de potencia de Box-Cox y el valor estimado de lambda, se ajusta la distribución de la variable de respuesta $\widehat{y}^{\lambda}$ para cumplir con el supuesto de normalidad en un modelo de regresión lineal múltiple.

Una vez realizada la variable dependiente $\widehat{y}$ se transformó y se obtuvo, $\widehat{y}^{\lambda}$, el modelo de regresión ajustado se muestra de la siguiente manera:

$$\begin{aligned}\widehat{y}^{\lambda} = -0.3 - 0.033x_1 - 0.013x_2 - 0.135x_3 - 0.002x_{4-1} - 0.169x_{4-2} \\ - 0.193x_{4-3} - 0.082x_5 + 0.04x_6 + 0.054x_7 - 0.109x_8 \\ + 0.012x_9 - 0.041x_{10} + 0.099x_{11} + 0.039x_{12-1} - 0.045x_{12-2} \\ - 0.465x_{12-3} + 0.114x_{13} + 0.032x_{14} + 0.136x_{15} + 0.098x_{16} \\ - 0.116x_{17} + 0.058x_{18} - 0.234x_{19} - 0.052x_{20}. \quad (3.3)\end{aligned}$$

En el desarrollo de la investigación, se utiliza la notación yt para referirse a la variable dependiente transformada mediante la transformación de Box-Cox. Esta notación es equivalente a $\widehat{y}^{\lambda}$, donde λ es el parámetro de la transformación aplicado a la variable.

Solo se analizarán los resultados generales del modelo y de la prueba de normalidad para verificar si el supuesto de normalidad fue corregido de forma exitosa, que se visualiza en el Cuadro 3.4.

Tabla 3.4: Resultados de la regresión para segundo modelo. Fuente [Propia].

Estadístico F	p-value
6.17	2.6e-13

Los resultados presentados en la tabla muestran los siguientes valores:

Estadístico F: El valor del estadístico F es 6.175183. Este valor se utiliza para realizar una prueba de hipótesis sobre la significancia del modelo.

p-value: El valor de p asociado al estadístico F es 2.609691e-13, que es extremadamente pequeño. Esto indica que al menos una de las variables predictoras tiene un efecto significativo sobre la variable de respuesta después de la transformación de potencia de Box-Cox. Se procede a analizar los resultados de la prueba Shapiro-Wilk que se muestra en el Cuadro 3.5.

Tabla 3.5: Resultados de la prueba de Shapiro-Wilk para el segundo modelo. Fuente [Propia].

W	p.value
0.98	0.003

En la prueba de Shapiro-Wilk, se observa que el valor de p sigue siendo bajo, con un valor de 0.003183, lo que indica que el supuesto de normalidad no se cumple de manera satisfactoria para los residuos del modelo, pero este es mayor que en el modelo de regresión con la variable de respuesta en su escala original.

Dado que el supuesto de normalidad es fundamental para un modelo de regresión lineal múltiple, es necesario realizar ajustes en el modelo con el objetivo de asegurar el cumplimiento de todos los supuestos necesarios. La falta de normalidad en los residuos puede tener implicaciones en la validez de las inferencias estadísticas y en la interpretación de los resultados del modelo.

Método de selección de variables: Stepwise.

Para respuesta a esta necesidad, se opta por la implementación del método de selección de variables "Stepwise". A través de este proceso, se obtiene un nuevo modelo de regresión ajustado, el que se presenta en la Ecuación (3.4), donde nuevamente se presentan únicamente los valores más relevantes para el análisis del modelo, presentados en el Cuadro 3.6, y los resultados de la prueba de Shapiro-Wilk ajustada a este nuevo modelo, en el Cuadro 3.7.

El método selecciona solamente las siguientes variables:

- x_3 = Trabajo Sí
- x_{4-1} = Ingreso Entre \$17,500 y \$22,000

- x_{4-2} = Ingreso Más de $22,000
- x_{4-3} = Ingreso Menos de $11,500
- x_5 = Problemas_fam_soc Sí
- x_7 = Horas_estudio
- x_{10} = Promedio_Materias
- x_{12-1} = Uso_Computadora No tenía
- x_{12-2} = Uso_Computadora Solo para mi
- x_{12-3} = Uso_Computadora Todos la usábamos

El modelo y sus resultados se presentan a continuación:

$$y = -0.348 - 0.150x_3 - 0.0001x_{4-1} + 0.214x_{4-2} - 0.201x_{4-3} - 0.134x_5 + 0.059x_7 - 0.057x_{10} + 0.087x_{12-1} - 0.005x_{12-2} - 0.434x_{12-3}. \quad (3.4)$$

Tabla 3.6: Resultados de la regresión para tercer modelo. Fuente [Propia].

Estadístico F	p-value
5.32	2.74e-11

Tabla 3.7: Resultados de la prueba de Shapiro-Wilk para tercer modelo. Fuente [Propia].

W	p-value
0.95	7.8e-06

Se obtuvo un valor de 2.74e-11 para el valor p de la prueba F, lo que indica que el modelo vuelve a cumplir con las condiciones esenciales de un modelo de regresión lineal. No obstante, los resultados derivados de la prueba de Shapiro-Wilk indican que el supuesto de normalidad en los residuos no se cumple adecuadamente. Como resultado de esta limitación, es necesario volver a ajustar el modelo para abordar este inconveniente.

No obstante, en este escenario, se optará por una estrategia similar a la aplicada previamente. Se procederá a implementar una transformación de Box-Cox sobre el modelo que ha emergido después de la aplicación del método de selección de variables de la Ecuación 3.4. Esta nueva variable se denotará como $yt2$ que se obtiene de transformar y.

Transformación de la variable dependiente y.

Los valores obtenidos para el método Box-Cox fueron los siguientes:

- El valor estimado de λ para la variable de respuesta $yt2$ es -0.1103.
- El valor estimado de γ para la variable de respuesta $yt2$ es 1.0267.

Estos resultados indican que se aplicó una transformación de potencia de Box-Cox con un valor λ estimado de -0.1103 para ajustar la distribución de la variable de respuesta $yt2$. La transformación de la variable dependiente y se expresa como:

$$yt2 = \frac{y^{-0.1103} - 1}{-0.1103}.$$

Con dicha transformación, se puede obtener el modelo ajustado resultante, sin embargo, para esta ecuación se eligió modificar las variables $x_{4-1}, x_{4-2}, x_{4-3}$ para asegurar que la categoría de referencia asociada a los ingresos sea la de menor nivel. Las variables representan los siguientes rangos de ingresos:

- x_{4-1} = Ingreso Entre \$11,500 y \$17,500
- x_{4-2} = Ingreso Entre \$17,500 y \$22,000
- x_{4-3} = Ingreso Más de \$22,000

Esta modificación simplifica la interpretación de los resultados del modelo, ya que las otras categorías se compararán con la de menor ingreso. Proporciona una perspectiva más clara de cómo las diferentes categorías afectan la variable dependiente en relación con la categoría de menor ingreso.

Dicho esto, se obtiene el modelo ajustado resultante:

$$\begin{aligned} yt2 = -0.56 - 0.15x_3 + 0.18x_{4-1} + 0.19x_{4-2} + 0.36x_{4-3} \\ - 0.13x_5 + 0.049x_7 - 0.046x_{10} + 0.09x_{12-1} - 0.01x_{12-2} - 0.433x_{12-3}. \end{aligned} \quad (3.5)$$

Con ella se tienen los siguientes resultados en el Cuadro 3.8:

Tabla 3.8: Resultados de la regresión para cuarto modelo. Fuente [Propia].

R_2	R_2 ajustada	Estadístico F	p-value
0.4352419	0.4031533	13.56378	1.5515e-17

El modelo de regresión lineal muestra un ajuste razonable a los datos, explicando un porcentaje significativo de la variabilidad en la variable de respuesta, con un 43.5 %. La significancia global del modelo se respalda con el estadístico F y el valor p muy bajo, lo que indica que el modelo tiene un efecto colectivo en la variable de respuesta.

Como siguiente paso, se llevarán a cabo una serie de pruebas y análisis para verificar el cumplimiento de los supuestos fundamentales de un modelo de regresión lineal múltiple.

Las pruebas incluirán la evaluación de la linealidad, la homocedasticidad, la normalidad y la independencia de los residuos. Cada uno de estos aspectos será examinado detenidamente para confirmar si el modelo seleccionado es apropiado y proporciona resultados confiables.

Verificación de Supuestos

Prueba de Normalidad

Se procede a realizar la prueba Shapiro-Wilk para probar normalidad en los residuos. Se obtuvieron los siguientes resultados presentados en el Cuadro 3.9 de dicha prueba:

Tabla 3.9: Resultados de la prueba de Shapiro-Wilk para cuarto modelo. Fuente [Propia].

W	p-value
0.9855	0.05104

Los resultados presentados en el Cuadro 3.9 con respecto a la prueba de Shapiro-Wilk resaltan que el estadístico de prueba W muestra una proximidad a 1, lo que sugiere una posible aproximación de los residuos hacia una distribución normal. Además, el valor calculado de p, que es 0.05104, indica que no se rechaza la hipótesis nula de que los residuos siguen una distribución normal.

La cercanía del valor p al nivel de significancia convencional (0.05) indica que es viable considerar que los residuos siguen una distribución normal, lo cual es fundamental para obtener resultados estadísticamente confiables y significativos en el modelo de regresión lineal múltiple.

A continuación, se presenta la gráfica de probabilidad normal que deseas observar:

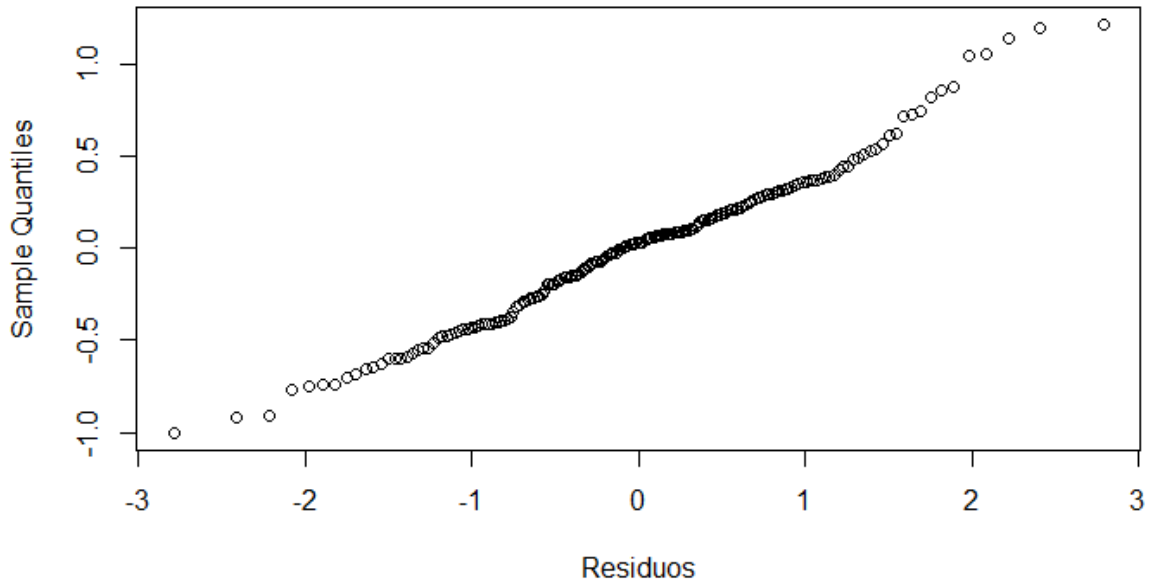


Figura 3.14: Gráfica de probabilidad normal del modelo final. Fuente [Propia].

Con respecto a la Figura 3.14, que muestra la gráfica de probabilidad normal del modelo final, se observa que los valores podrían considerarse ideales y, de acuerdo con la prueba de Shapiro-Wilk realizada, se puede concluir que la verificación del supuesto de normalidad es satisfactoria.

Prueba de homocedasticidad.

Se realiza la prueba de homocedasticidad, la cual se ejecuta con la prueba de Breusch-Pagan. En ella se obtuvieron los siguientes resultados:

Tabla 3.10: Resultados de la prueba Breusch-Pagan para el cuarto modelo. Fuente [Propia].

BP	GL	p-value
10.874	10	0.3675

Con la información presentada en el Cuadro 3.10, se observa que el valor calculado de la estadística de Breusch-Pagan es 10.874. Este valor es el resultado de la prueba que compara la relación entre los residuos y las variables independientes en términos de su varianza. Los grados de libertad asociados con la prueba son 10, que se relaciona con la cantidad de variables independientes en el modelo que podrían estar contribuyendo a la heterocedasticidad.

Dado que el valor p (0.37) es mayor que el nivel de significancia comúnmente utilizado (como 0.05), no hay evidencia suficiente para concluir que existe heterocedasticidad en los residuos del modelo.

Este resultado sugiere que la suposición de homocedasticidad (varianza constante de los residuos) se cumple en este modelo ajustado. Sin embargo, es importante recordar que las pruebas estadísticas pueden tener limitaciones y que la interpretación debe basarse en una combinación de análisis estadísticos y conocimiento del entorno.

Un apoyo visual para la prueba, se tiene la siguiente Figura 3.16, donde se presenta la gráfica de residuales e_i en función de los valores ajustados, esto para detectar algunos tipos frecuentes de inadecuaciones del modelo.

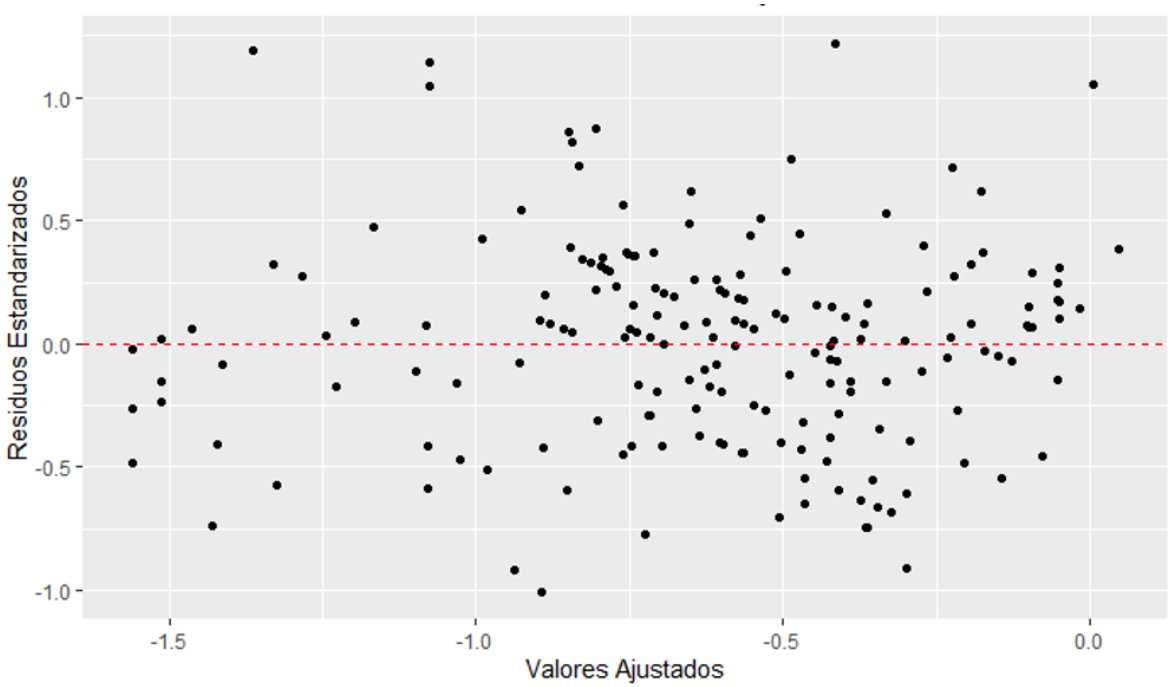


Figura 3.15: Gráfico de Residuos Estandarizados vs Valores Ajustados del modelo final. Fuente [Propia].

En la Figura 3.15, se observa que los residuos se encuentran dentro de una banda horizontal, lo que sugiere que no hay defectos obvios en el modelo.

Prueba de linealidad.

Finalmente, se realiza la prueba de linealidad, esta se obtiene a través de la prueba RESET.

Tabla 3.11: Resultados de la prueba RESET para el cuarto modelo. Fuente [Propia].

RESET	df1	df2	p-value
0.03655	1	175	0.8486

Con respecto a los resultados del Cuadro 1.11, se observa que el valor calculado de la estadística RESET es 0.036, y el p-value obtenido en la prueba fue de 0.85. Dado que el p-value es alto y mucho mayor que el nivel de significancia de 0.05, se puede concluir que el modelo actual parece capturar adecuadamente la relación entre las variables independientes y la variable dependiente.

Este resultado sugiere que no es necesario agregar términos polinómicos de orden superior al modelo para mejorar su ajuste, según la prueba RESET.

3.6. Resultados

Con respecto al modelo de regresión lineal presentado, y una vez aprobados los supuestos que debe cumplir dicho modelo, procedemos a realizar un análisis más detallado de los resultados obtenidos. Para esto, se observa el Cuadro 3.12 que muestra los resultados de las variables del modelo final:

En el proceso de ajuste del modelo de regresión, se optó por transformar las variables para garantizar que la categoría de referencia asociada a los ingresos sea la de menor nivel. Estas variables representan distintos rangos de ingresos, facilitando así la interpretación de los resultados. La categoría de menor ingreso actúa como referencia, permitiendo una comparación clara con las demás categorías y proporcionando una visión más precisa de cómo afectan estas variables a la respuesta del modelo.

Tabla 3.12: Estimaciones de la regresión $yt2$. Fuente [Propia].

Variable	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.57	0.17	-3.24	0.00
TrabajoSí	-0.15	0.07	-2.02	0.04
Ingreso Entre $11,500 y $17,500	0.18	0.08	2.15	0.03
Ingreso Entre $17,500 y $22,000	0.19	0.11	1.77	0.08
Ingreso Más de $22,000	0.36	0.11	3.37	0.00
Problemas_fam_socSí	-0.13	0.07	-1.87	0.06
Horas_estudio	0.05	0.01	4.46	0.00
Promedio_Materias	-0.05	0.03	-1.62	0.10
Uso_Computadora No tenía	0.09	0.18	0.51	0.61
Uso_Computadora Solo para mi	-0.01	0.09	-0.12	0.90
Uso_Computadora Todos la usabamos	-0.43	0.12	-3.69	0.00

En este sentido, destacamos los siguientes puntos de relevancia:

1. Se tuvo que el modelo presentó un $R^2 = 43.52\,\%$, lo que significa que el modelo exhibe un nivel razonable de ajuste a los datos, siendo capaz de explicar aproximadamente dicho $43.5\,\%$ de la variabilidad presente en los datos. En otras palabras, cerca de la mitad de la variabilidad en la variable dependiente puede ser explicada por las variables independientes incluidas en el modelo.

2. El término de intersección (β_0) representa el valor estimado de la variable dependiente cuando todas las variables independientes son cero. En este caso, cuando todas las otras variables son cero, el valor estimado de la variable dependiente $yt2$ es -0.567.

3. La variable x_3, representa si el estudiante trabajó durante la pandemia. Un coeficiente estimado de -0.151 sugiere que, manteniendo todas las demás variables constantes, los estudiantes que tuvieron que trabajar durante la pandemia en promedio tienen 0.151 unidades menos de $yt2$ en comparación con los que no tuvieron que trabajar.

4. La variable x_{4-1}, indica ingresos familiares entre $11,500 y $17,500, observamos un coeficiente estimado de 0.185 con un p-value de 0.03. Esto sugiere que los estudiantes con este rango de ingresos experimentaron, en promedio, un rendimiento 0.185 unidades superior de $yt2$ en comparación con aquellos cuyos ingresos eran inferiores a $11,500.

5. La variable x_{4-2}, que representa ingresos entre $17,500 y $22,000, muestra un coeficiente estimado de 0.191 con un p-value de 0.08. Esto indica que los estudiantes con este rango de ingresos experimentaron, en promedio, un rendimiento de 0.191 unidades más de $yt2$ en comparación con aquellos cuyos ingresos eran inferiores a $11,500.

6. La variable que representa un ingreso mayor de $22,00 ($x_{4-3}$) obtuvo un coeficiente estimado de 0.36 con un p-value de 0.001. Esto sugiere que los estudiantes en este rango de ingresos experimentaron, en promedio, un rendimiento 0.36 unidades más de $yt2$ en comparación con aquellos cuyos ingresos eran inferiores a $11,500.

7. Las interpretaciones de 4-6 indican que los estudiantes con ingresos más altos tienden a tener un rendimiento promedio estimado más alto durante la pandemia en comparación con aquellos con los ingresos más bajos.

8. De acuerdo a la variable x_5, que representa aquellos estudiantes que enfrentaron problemas familiares o sociales, con un coeficiente de -0.13, sugiere que, manteniendo todas las demás variables constantes, los estudiantes que tuvieron problemas familiares o sociales durante la pandemia en promedio tienen un rendimiento de 0.13 unidades más bajo de $yt2$ en comparación de aquellos estudiantes que no tuvieron dichos problemas.

9. En la variable x_7, que representa el número de horas que el estudiante dedicaba al día para estudiar, tiene un coeficiente estimado de 0.05 con un p-value extremadamente bajo, indica que manteniendo todas las demás variables constantes, si el número de horas de estudio incrementa una unidad, en promedio $yt2$ incrementa 0.05 unidades, y este término es significativo.

10. El promedio de materias que curso el estudiante (x_{10}), con un coeficiente estimado de -0.0462 con un p-value de 0.11. Indica que, manteniendo todas las demás variables constantes, si el promedio de las materias incrementa una unidad en el promedio $yt2$ disminuye 0.05 unidades, aunque este término no es significativo.

11. La variable x_{12-1} indica si el estudiante no tenía una computadora personal durante la pandemia, se observa un coeficiente estimado de 0.09 con un p-value de 0.61. Esto sugiere que los estudiantes que se encontraban en este grupo experimentaron, en promedio, un rendimiento de 0.09 unidades superior de $yt2$ en comparación con aquellos estudiantes que compartian su computadora en ciertas ocasiones.

12. La variable x_{12-2} indica si el estudiante contaba con una computadora propia durante la pandemia, presenta un coeficiente estimado de -0.01 con un p-value de 0.9. Esto indica que los estudiantes con una computadora propia experimentaron, en promedio, un rendimiento menor a $yt2$ de 0.01 unidades en comparación de aquellos estudiantes que compartían su computadora en ciertas ocasiones.

13. La variable que representa a los estudiantes que contaba con una computadora compartida por toda su familia (x_{12-3}) obtuvo un coeficiente estimado de -0.43 con un p-value de 0.0002. Esto sugiere que los estudiantes que se encontraban en esta situación experimentaron, en promedio, un rendimiento de 0.43 unidades menos que $yt2$ en comparación con aquellos estudiantes que compartían su computadora en ciertas ocasiones.

Capítulo 4

Conclusiones

Como se menciona en la introducción, resulta de gran importancia analizar los efectos que tuvo la pandemia por la pandemia COVID-19, para así identificar cuales fueron aquellas variables más relevantes que influyeron en el rendimiento de un estudiante.

De acuerdo con el análisis exhaustivo realizado sobre el impacto de la pandemia COVID-19 en el rendimiento académico de los estudiantes, se pueden extraer conclusiones significativas que arrojan luz sobre los efectos de este evento global en la educación. A través de una metodología se trabajaron métodos estadísticos y modelos de regresión lineal múltiple, con lo cual con esta metodología nos hemos acercado a comprender mejor la situación que se vivió durante esta crisis.

El presente estudio resaltó la importancia de una serie de variables que se encontraron tener un efecto relevante en el rendimiento académico. Entre ellas, se destacan las horas de estudio, donde se encontró una relación positiva significativa, lo que sugiere que aquellos estudiantes que invirtieron más tiempo en el estudio lograron un mejor rendimiento durante la pandemia.

Asimismo, se identificó una relación entre el ingreso familiar más bajo y el rendimiento académico. A medida que los ingresos aumentan, se observa un incremento estadísticamente significativo en el rendimiento de los estudiantes, lo que indica que los estudiantes con recursos económicos más limitados enfrentaron mayores desafíos en su desempeño académico.

La adaptación a la educación en línea también emergió como un factor crucial. La disponibilidad de recursos tecnológicos, representada por las variables relacionadas con el uso de computadoras, no mostró una relación estadísticamente significativa con el rendimiento académico. Esto podría indicar que, en un contexto de educación virtual, la igualdad de acceso a la tecnología podría haberse nivelado en cierta medida.

Por otro lado, variables como el trabajo durante la pandemia, los problemas familiares o sociales y el género no demostraron una relación clara y significativa con el rendimiento académico, lo que resalta la complejidad de estos factores y la necesidad de un análisis más profundo.

En general, aunque el modelo de regresión lineal múltiple presentó un ajuste razonable a los datos, es importante considerar que la pandemia tuvo efectos multifacéticos y que algunas relaciones pueden haber sido influenciadas por factores no contemplados en este estudio. No obstante, los resultados obtenidos permiten tener una visión más completa y cuantificada de cómo diferentes variables intervinieron en el rendimiento académico de los estudiantes.

Esta investigación no solo contribuye al entendimiento de cómo la pandemia afectó a los estudiantes universitarios en términos de su rendimiento académico, sino que también resalta la necesidad de enfoques educativos flexibles y adaptativos para enfrentar situaciones de crisis que pueden aparecer en cualquier momento. Estar preparados en construir plataformas, bancos de ejercicios resueltos, etc.

Esta tesis ofrece contribuciones significativas al campo de la actuaría y las ciencias actuariales al proporcionar un análisis detallado del impacto de la pandemia en el rendimiento académico de los

estudiantes universitarios. Al aplicar métodos estadísticos avanzados, como modelos de regresión lineal múltiple, se amplía la comprensión sobre cómo eventos externos pueden influir en resultados específicos.

Además, fortalece el perfil de egreso al demostrar la capacidad para aplicar técnicas estadísticas avanzadas en el análisis de fenómenos complejos, como el impacto de la pandemia en la educación superior. Al identificar y evaluar variables relevantes en situaciones de crisis, se resalta la competencia esencial de los profesionales para adaptarse y tomar decisiones informadas en entornos cambiantes.

En el mercado laboral, esta tesis proporciona una ventaja competitiva al demostrar la capacidad para realizar análisis rigurosos y proporcionar información valiosa sobre el impacto de eventos externos en resultados específicos.

Aunque esta investigación proporciona una comprensión de los factores que afectaron el rendimiento académico durante la pandemia, el panorama sigue siendo dinámico. Si bien no se contempla un trabajo futuro en este momento, se podría requerir una revisión continua y adaptativa de las estrategias educativas. Queda abierta la posibilidad de investigaciones adicionales para abordar aspectos específicos, como el desarrollo de métodos de aprendizaje en entornos virtuales y el impacto de la salud mental en el desempeño académico durante crisis prolongadas.

Anexos

Anexos A

Encuesta

Encuesta para saber el rendimiento de los estudiantes en la FCFM durante la pandemia

1. Matricula:
2. ¿De qué carrera eres?
   - Actuaría
   - Física
   - Física Aplicada
   - Matemáticas
   - Matemáticas Aplicadas
3. ¿Durante la pandemia te encontrabas en un ámbito familiar?
   - Sí
   - No
4. Género:
   - Masculino
   - Femenino
   - Otro
5. Promedio general al finalizar otoño 2019:
6. Promedio general al finalizar primavera 2022:
7. ¿Eres originario de Puebla?
   - Sí
   - No
8. ¿Trabajaste durante la pandemia?
   - Sí
   - No
9. ¿Cuál es el ingreso promedio de tu familia?
   - Menos de $\$11,500$
   - Entre $\$11,500$ y $\$17,500$
   - Entre $\$17,500$ y $\$22,000$

- Más de $\$22,000$

10. ¿Tuviste internet en tu hogar durante la pandemia?
    - Sí
    - No
11. ¿Tuviste problemas familiares o sociales?
    - Sí
    - No
12. ¿Tuviste COVID en alguna ocasión?
    - Sí
    - No
13. ¿Cuántas horas en promedio al día ocupabas para estudiar?
14. ¿Padeciste de alguna enfermedad durante la pandemia?
    - Sí
    - No
15. ¿Cuál es el nivel de dificultad que tuvieron las imparticiones de los cursos de los profesores?
    - Alto
    - Medio
    - Bajo
16. Promedio de materias que tomaste por semestre en pandemia:
17. ¿Tuviste alguna beca durante la pandemia?
    - Sí
    - No
18. ¿Usabas alguna computadora propia o la compartías con otros miembros de la familia?
    - No tenía
    - Sólo para mi
    - Compartía de vez en cuando
    - Todos la usábamos
19. Tomando en cuenta lo que aprendiste en el tiempo de COVID. ¿Obtuviste los conocimientos suficientes para enfrentar las materias que estás llevando ahora?
20. ¿Qué tan capacitados los profesores estaban en sus materias y en el uso de sus plataformas con las cuales impartieron sus clases?
    - Alto
    - Medio
    - Bajo
21. ¿Qué aplicaciones ocuparon tus profesores durante las clases en la pandemia?
    - WhatsApp
    - Microsoft Teams
    - Classroom

- Zoom
- Virtual Horizon
- Blackboard
- Otro:

Anexos B

Anexo II: Código de RStudio

Modelo de regresión

Oscar Espinosa Cuahuizo

A continuación, se procede a desarrollar el modelo de regresión lineal múltiple, el código se realiza en el ambiente de RStudio.

Para llevar a cabo este proceso, se requiere la utilización de diversas bibliotecas que son empleadas en diferentes etapas del análisis. Las siguientes librerías son llamadas a continuación:

```
library(readxl)
library(lmtest)
library(knitr)
library(kableExtra)
library(ggplot2)
library(tidyverse)
library(leaps)
library(MASS)
library(car)
library(lmtest)
library(gridExtra)
library(knitr)
library(broom)
library(pander)
library(dplyr)
```

Modelo de regresión lineal múltiple

En una fase inicial, se requiere la extracción de los datos para llevar a cabo el proceso de construcción del modelo de regresión lineal múltiple. Para lograrlo, es esencial acceder a los datos alojados en un archivo en formato "xlsx". El proceso de extracción de los datos se ejecuta de la siguiente manera:

```
Encuesta <- read_excel("Data.xlsx")
```

Las dimensiones del dataframe son las siguientes:

```
ncol(Encuesta)
```

```
## [1] 26
```

```
nrow(Encuesta)
```

```
## [1] 188
```

A través de este análisis, se puede determinar que el dataframe posee inicialmente 188 filas y 26 columnas.

Además, resulta esencial obtener una observación visual de los datos que se emplearán en la construcción del modelo. Debido a la considerable cantidad de variables (columnas), se realizará una segmentación de los datos en cuatro tablas distintas, con el fin de optimizar la visualización. Estas divisiones se presentan de la siguiente manera:

```
# Tablas
tabla1 <- kable(head(Encuesta[,1:5], 10), format = "latex", booktabs = TRUE, align = "c")
tabla2 <- kable(head(Encuesta[,6:10], 10), format = "latex", booktabs = TRUE, align = "c")
tabla3 <- kable(head(Encuesta[,11:15], 10), format = "latex", booktabs = TRUE, align = "c")
tabla4 <- kable(head(Encuesta[,16:20], 10), format = "latex", booktabs = TRUE, align = "c")
tabla5 <- kable(head(Encuesta[,21:26], 10), format = "latex", booktabs = TRUE, align = "c")

# Imprimir tablas en el orden deseado
pander(tabla1)
```

Matricula	Carrera	Ambito_familiar	Genero	Promedio2019
201866549	Actuaría	Sí	Femenino	8.4
201844127	Actuaría	Sí	Femenino	8.2
201540430	Actuaría	Sí	Femenino	7.8
201934397	Actuaría	Sí	Femenino	9.2
• 201739265	Actuaría	Sí	Masculino	8.3
201859626	Actuaría	Sí	Femenino	8.8
201802808	Actuaría	Sí	Femenino	8.8
201866820	Actuaría	Sí	Masculino	8.2
201931065	Actuaría	Sí	Femenino	9.0
201873928	Actuaría	Sí	Femenino	8.8

```
pander(tabla2)
```

Promedio2022	Originario_de_Puebla	Trabajo	Ingreso	Internet
8.30	Sí	Sí	Entre $11,500 y $17,500	Sí
8.50	Sí	Sí	Entre $11,500 y $17,500	Sí
8.00	No	Sí	Entre $11,500 y $17,500	Sí
8.40	Sí	Sí	Mas de $22,000	Sí
• 8.40	Sí	No	Entre $11,500 y $17,500	Sí
8.40	Sí	Sí	Entre $11,500 y $17,500	Sí
9.05	No	No	Mas de $22,000	Sí
8.00	No	No	Entre $11,500 y $17,500	Sí
8.90	No	No	Mas de $22,000	Sí
8.60	No	No	Mas de $22,000	Sí

```
pander(tabla3)
```

Problemas_fam_soc	COVID	Horas_estudio	Enfermedad	Nivel_Dificultad
Sí	Sí	0.5	Sí	Medio
Sí	Sí	3.0	No	Medio
Sí	Sí	2.0	Sí	Alto
Sí	No	0.0	No	Medio
• No	No	3.0	No	Medio
Sí	Sí	3.0	No	Medio
Sí	No	3.0	Sí	Medio
No	No	4.0	Sí	Medio
Sí	No	0.0	No	Alto
Sí	No	3.0	No	Medio

```
pander(tabla4)
```

Promedio_Materias	Beca	Uso_Computadora	Conocimientos	Cap_Porfesores
5	No	Solo para mi	No	Medio
5	No	Solo para mi	No	Medio
2	No	No tenía	No	Medio
5	No	Solo para mi	No	Medio
• 8	No	Solo para mi	No	Medio
6	No	Solo para mi	No	Medio
5	No	Solo para mi	Sí	Medio
6	No	Solo para mi	Sí	Medio
6	No	Solo para mi	No	Medio
6	No	Solo para mi	Sí	Medio

```
pander(tabla5)
```

WhatsApp	Microsoft_Teams	Classroom	Zoom	Virtual_Horizon	BlackBoard
0	1	1	1	1	1
1	1	1	1	0	1
1	0	0	0	0	0
1	1	0	1	0	0
• 0	1	1	1	1	0
1	1	1	1	0	0
1	1	1	1	0	0
0	1	1	1	1	1
1	1	1	1	1	0
1	1	1	1	0	1

Tras la carga exitosa de los datos y su visualización preliminar, es imperativo adquirir la variable dependiente, que será empleada en la construcción del modelo de regresión. Esta variable será calculada siguiendo la siguiente fórmula:

$$Y = Promedio2022 - Promedio2019$$

Esto se realiza a continuación:

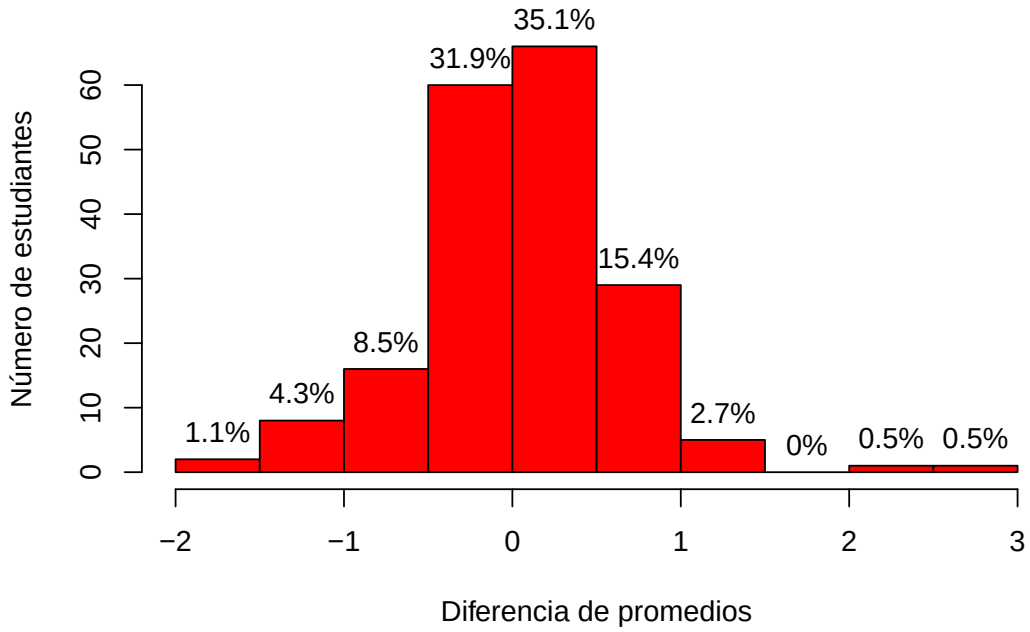
```r
Encuesta$Y <- Encuesta$Promedio2022 - Encuesta$Promedio2019
```

En este punto, se inicia el proceso de fragmentación de los datos en dos categorías: cualitativos y cuantitativos. Esta división es esencial para la generación de gráficos que aportarán significativamente al análisis estadístico. El procedimiento se detalla a continuación:

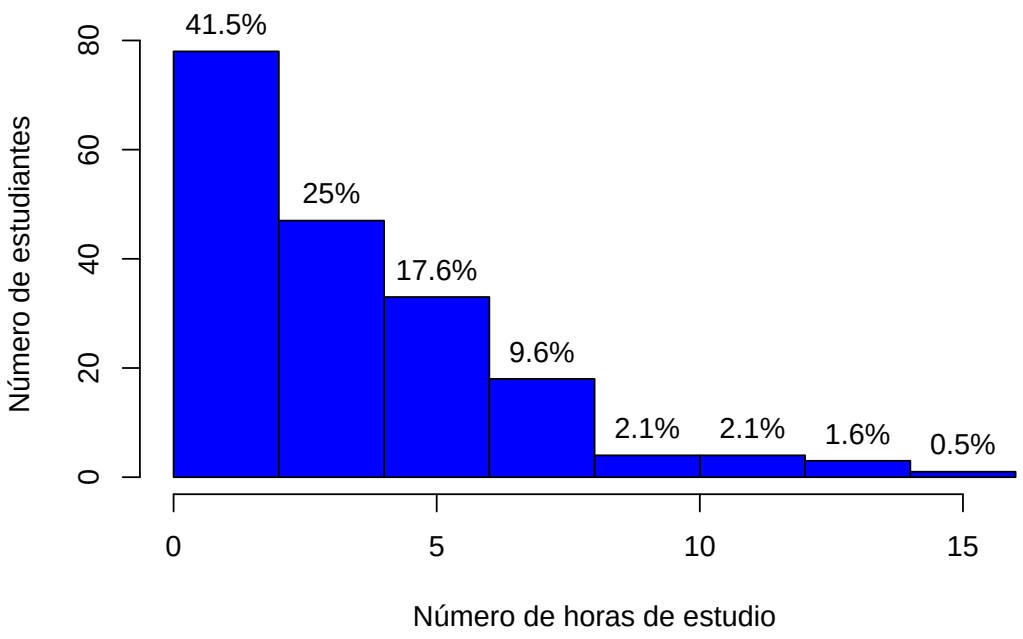
```r
Cuantitativos <- data.frame(Encuesta$Y, Encuesta$Horas_estudio, Encuesta$Promedio_Materias)
Cualitativos <- data.frame(Encuesta$Carrera,
                           Encuesta$Ambito_familiar,
                           Encuesta$Genero,
                           Encuesta$Originario_de_Puebla,
                           Encuesta$Trabajo,
                           Encuesta$Ingreso,
                           Encuesta$Internet,
                           Encuesta$Problemas_fam_soc,
                           Encuesta$COVID,
                           Encuesta$Enfermedad,
                           Encuesta$Nivel_Dificultad,
                           Encuesta$Beca,
                           Encuesta$Uso_Computadora,
                           Encuesta$Conocimientos,
                           Encuesta$Cap_Porfesores)
```

Después de completar la segregación de datos, el siguiente paso involucra la creación de gráficos para llevar a cabo un análisis más efectivo. Comenzaremos obteniendo histogramas para dos variables específicas: la variable dependiente "Y" y la variable independiente "Horas_estudio". Estos gráficos permitirán un examen más profundo de la distribución de los datos en relación con estas dos variables.

```r
#Histograma del rendimiento de los estudiantes

hist_data <- hist(Encuesta$Y, plot = FALSE)
bar_percentages <- prop.table(hist_data$counts) * 100
hist(Encuesta$Y, main = "", xlab = "Diferencia de promedios",
     ylab = "Número de estudiantes", col = "red")
text(hist_data$mids, hist_data$counts, labels =
       paste0(round(bar_percentages, 1), "%"), pos = 3)
mtext(text = paste0(round(bar_percentages[5], 1), "%"),side = 3,
      at = hist_data$mids[5],line = -0.2)
```



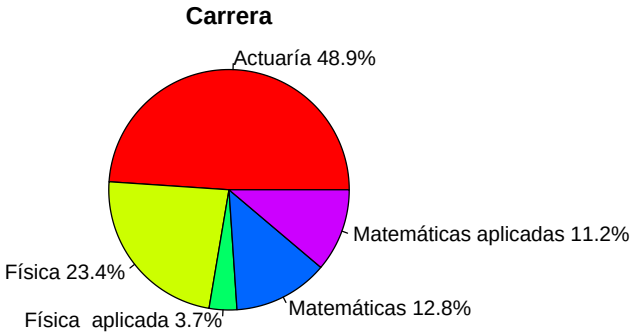

```r
#Histograma de las horas de estudio de los estudiantes

hist_data <- hist(Encuesta$Horas_estudio, plot = FALSE)
bar_percentages <- prop.table(hist_data$counts) * 100
hist(Encuesta$Horas_estudio, main = "", xlab = "Número de horas de estudio",
     ylab = "Número de estudiantes", col = "blue")
text(hist_data$mids, hist_data$counts, labels =
       paste0(round(bar_percentages, 1), "%"), pos = 3)
mtext(text = paste0(round(bar_percentages[1], 1), "%"),side = 3,
      at = hist_data$mids[1],line = -0.2)
```

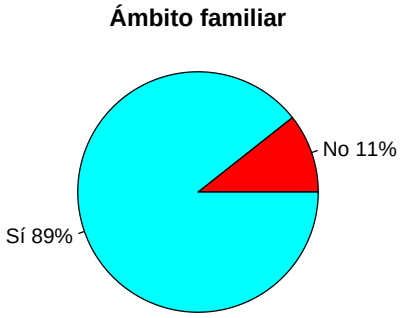


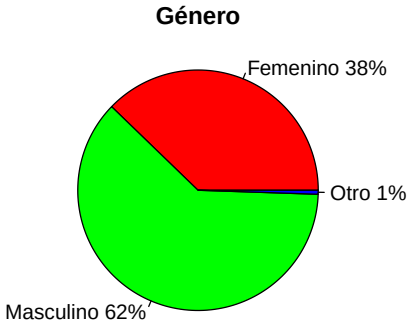
Ahora, se continua con la creación de las gráficas de pastel para el mismo fin:

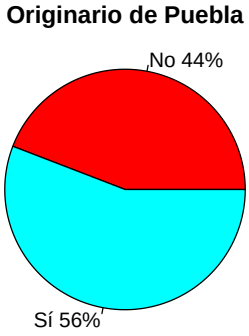
```
# Número de columnas en el dataframe Cualitativos

nombres_personalizados <- c(
  "Carrera", "Ámbito familiar", "Género", "Originario de Puebla",
  "Trabajo", "Ingreso", "Internet", "Problemas familiares y/o sociales",
  "Covid", "Enfermedad", "Dificultad de profesores",
  "Beca", "Uso de Computadora", "Conocimientos", "Capacitación de los profesores"
)

num_columnas <- ncol(Cualitativos)
columnas_por_fila <- 3
filas <- ceiling(num_columnas / columnas_por_fila)
grupos_de_columnas <- split(1:num_columnas, rep(1:filas, each = columnas_por_fila,
                                                length.out = num_columnas))
par(mfrow = c(1,1), mar = c(1, 1, 1, 1))
for (grupo in grupos_de_columnas) {

  for (i in grupo) {
    columna_actual <- colnames(Cualitativos)[i]

    A <- Cualitativos %>% count(!!sym(columna_actual))
    proporciones <- A$n / sum(A$n)

    etiquetas <- paste(A[[columna_actual]], scales::percent(proporciones))

    pie(proporciones, labels = etiquetas, col = rainbow(length(etiquetas)),
```

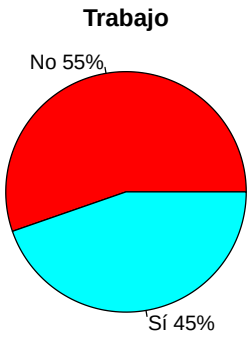
```
        main = nombres_personalizados[i])
  }
}
```

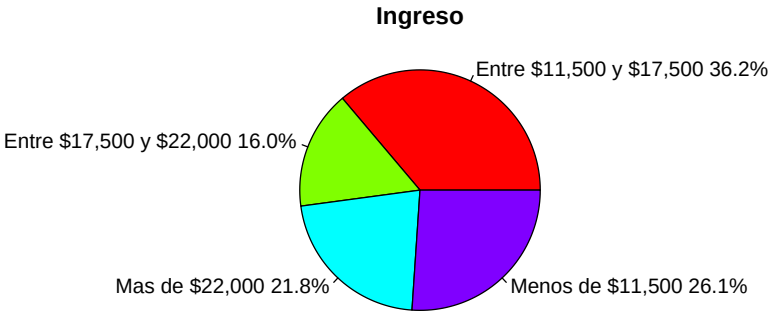


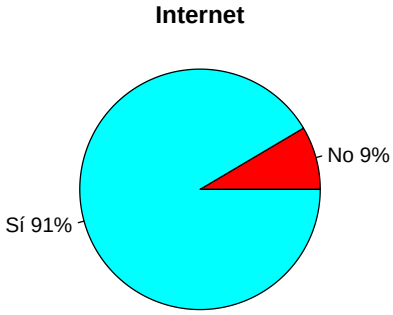


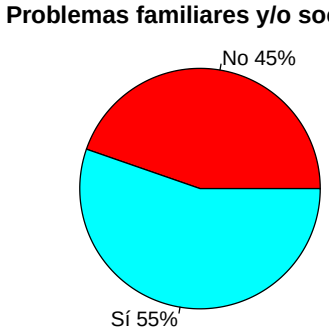




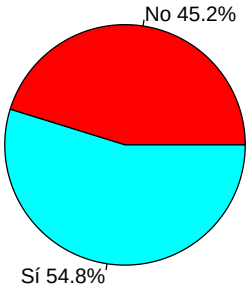

Trabajo
No 55%
Sí 45%


Ingreso
Entre $11,500 y $17,500 36.2%
Entre $17,500 y $22,000 16.0%
Mas de $22,000 21.8%
Menos de $11,500 26.1%

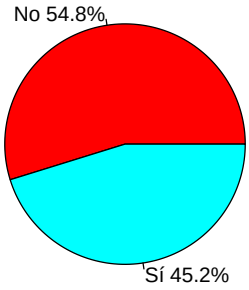

Internet
No 9%
Sí 91%


Problemas familiares y/o sociales
No 45%
Sí 55%

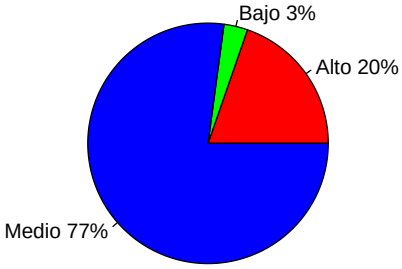
Covid



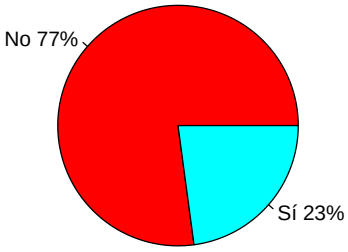
Enfermedad

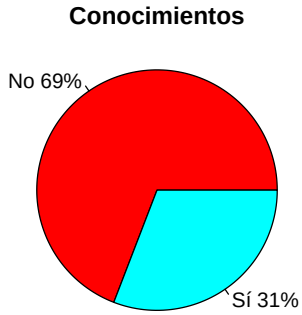


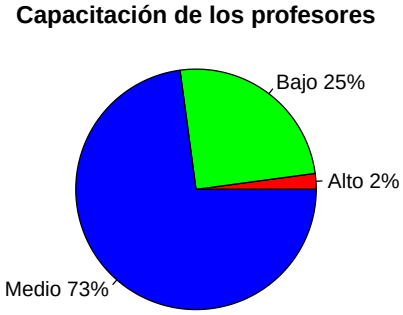
Dificultad de profesores



Beca







Este proceso arroja varios aspectos cruciales que deben ser considerados. Al mismo tiempo, se proporciona el código que implementa las modificaciones necesarias para abordar las problemáticas señaladas.

1. La variable "Carrera" muestra una prevalencia significativa de encuestados en la disciplina de Actuaría. Como resultado, es prudente descartar esta variable, ya que su contribución podría resultar irrelevante en el modelo.

```
Encuesta <- subset(Encuesta, select = -Carrera)
```

2. La variable "Ambito_familiar" proporciona escasa información de relevancia para el análisis, por lo tanto, se opta por no incorporarla en el modelo de regresión.

```
Encuesta <- subset(Encuesta, select = -Ambito_familiar)
```

3. En relación a la variable "Género", se identificó una única respuesta como "Otro". Dado que esta respuesta proviene de un solo encuestado, existe la posibilidad de que su inclusión altere el modelo. Por este motivo, se decide eliminar todas las respuestas correspondientes a esta persona, enfatizando que esta acción no se realiza por razones discriminatorias.

```r
Encuesta <- subset(Encuesta, Encuesta$Genero != "Otro")
```

4. Los datos recolectados para la variable "Internet" presentan una cantidad insuficiente de información para considerarla en el modelo. Por tanto, se concluye que esta variable no aporta el nivel de significancia necesario y no será incorporada en el análisis.

```r
Encuesta <- subset(Encuesta, select = -Internet)
```

5. En relación a la variable "Nivel_Dificultad", se destaca que solo el 3% de los encuestados indicaron un nivel de dificultad bajo. Dado que esta proporción resulta insuficiente para contribuir al modelo, se opta por agrupar esta respuesta con la categoría "Medio". Esta acción se lleva a cabo para mejorar la utilidad y relevancia de la variable en el análisis..

```r
unique_levels <- unique(Encuesta$Nivel_Dificultad)
Encuesta$Nivel_Dificultad[Encuesta$Nivel_Dificultad == "Bajo"] <- "Medio"
#Se cambia el nombre de los datos iguales a medio
Encuesta$Nivel_Dificultad <- ifelse(Encuesta$Nivel_Dificultad == "Medio", "Medio-Bajo",
                                    Encuesta$Nivel_Dificultad)
unique_levels <- unique(Encuesta$Nivel_Dificultad)
print(unique_levels)
```

```
## [1] "Medio-Bajo" "Alto"
```

6. En la variable "Uso_Computadora", se identifican respuestas con un porcentaje reducido. Debido a esta distribución, se toma la decisión de consolidar algunas respuestas para mejorar la representatividad y utilidad de la variable en el análisis.

```r
#unique_levels <- unique(Encuesta$Uso_Computadora)
#Encuesta$Uso_Computadora[Encuesta$Uso_Computadora == "No tenía"] <- "Todos la usabamos"
#Encuesta$Uso_Computadora <- ifelse(Encuesta$Uso_Computadora == "Todos la usabamos",
#                                   "No tenía o uso compartido", Encuesta$Uso_Computadora)
#unique_levels <- unique(Encuesta$Uso_Computadora)
#print(unique_levels)
```

7. Similar al caso mencionado en el punto 3, se observa que en la variable "Cap_Profesores" solo un 2% de los encuestados indicó un nivel "alto". Siguiendo el razonamiento previo, estas respuestas se reasignan a la categoría "Medio" y se modifica esta respuesta para reflejar "Medio-Alto". Esta acción se ejecuta con el propósito de mejorar la coherencia y eficacia de la variable en el análisis.

```r
unique_levels <- unique(Encuesta$Cap_Porfesores)
Encuesta$Cap_Porfesores[Encuesta$Cap_Porfesores == "Alto"] <- "Medio"
Encuesta$Cap_Porfesores <- ifelse(Encuesta$Cap_Porfesores == "Medio", "Medio-Alto",
                                  Encuesta$Cap_Porfesores)
unique_levels <- unique(Encuesta$Cap_Porfesores)
print(unique_levels)
```

```
## [1] "Medio-Alto" "Bajo"
```

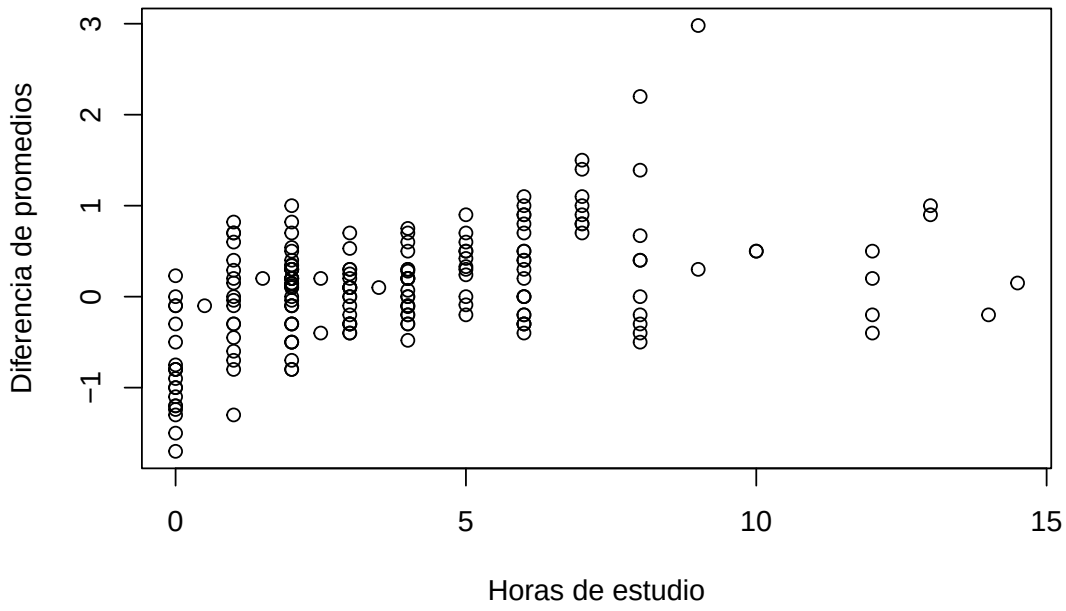
8. En el último paso, se realiza la eliminación de las variables que no serán utilizadas en la construcción del modelo. Estas variables son "Matricula", "Promedio2019" y "Promedio2022".

```r
Encuesta <- subset(Encuesta, select = -Matricula)
Encuesta <- subset(Encuesta, select = -Promedio2019)
Encuesta <- subset(Encuesta, select = -Promedio2022)
```

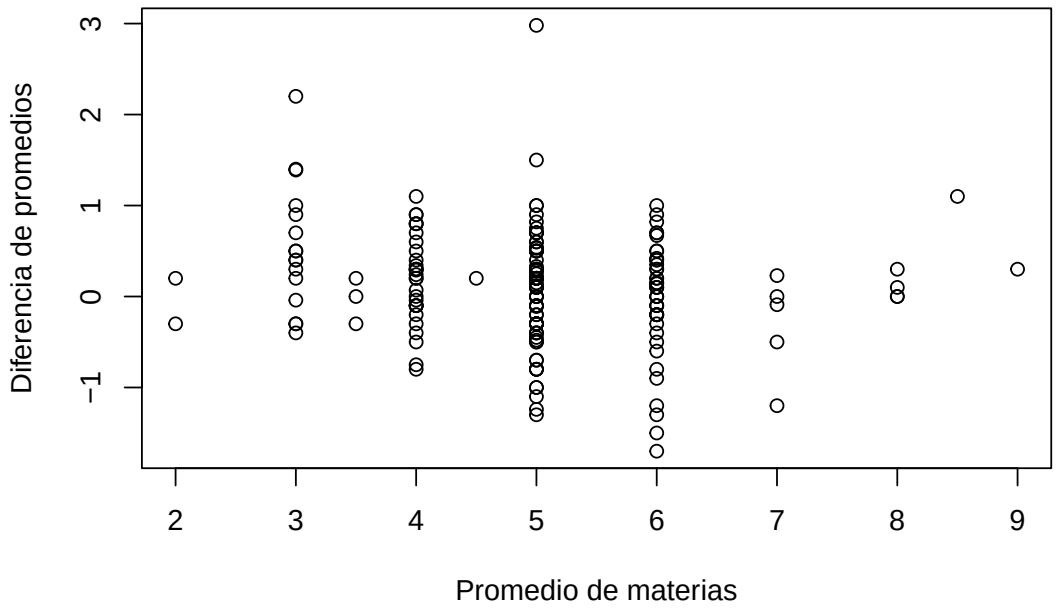
Gráficas

Este bloque de código generará tres gráficos. Los primeros dos son gráficos de dispersión entre "Horas_estudio" y "Y", el segundo modelo es entre "Promedio_Materias" y "Y". Este enfoque visual facilitará identificar patrones y tendencias en los datos, así como las diferencias en la distribución según el género.

```r
plot(Encuesta$Horas_estudio, Encuesta$Y,
     xlab = "Horas de estudio", ylab = "Diferencia de promedios")
```




```r
plot(Encuesta$Promedio_Materias, Encuesta$Y,
     xlab = "Promedio de materias", ylab = "Diferencia de promedios")
```




```
data <- Encuesta
attach(data)
```

Creación del modelo.

En este punto, procedemos a la construcción del modelo de regresión, tomando en consideración únicamente las variables que han superado las etapas de análisis y selección previas. El siguiente código implementa el modelo de regresión y al mismo tiempo exporta sus resultados mediante dos tablas distintas. En la primera tabla se presentan los coeficientes estimados para cada variable independiente, proporcionando una visión clara de sus efectos en las variables dependientes. En la segunda tabla se incluyen valores estadísticos relevantes.

```
modelo <- lm(Y ~ Genero + Originario_de_Puebla +
               Trabajo + Ingreso + Problemas_fam_soc +
               COVID + Horas_estudio + Enfermedad + Nivel_Dificultad +
               Promedio_Materias + Beca + Uso_Computadora + Conocimientos +
               Cap_Porfesores + BlackBoard + Virtual_Horizon + Zoom +
               Classroom + Microsoft_Teams + WhatsApp)
summary_modelo <- summary(modelo)

# Convertir la tabla de coeficientes en un data frame
coef_table <- as.data.frame(summary_modelo$coefficients)

# Crear la tabla con kable
tabla_formateada <- kable(coef_table,
                          caption = "Resumen del Modelo",
                          format = "latex", booktabs = TRUE, align = "c") %>%
```

```
  kable_styling(latex_options = c("striped", "hold_position"))
# , font_size = 10 para ajustar el tamaño
tabla_formateada
```

Table 1: Resumen del Modelo

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3890442	0.5895306	0.6599220	0.5102410
GeneroMasculino	-0.0065235	0.0837545	-0.0778880	0.9380132
Originario_de_PueblaSí	-0.0183414	0.0795603	-0.2305340	0.8179677
TrabajoSí	-0.1254644	0.0930666	-1.3481133	0.1795039
IngresoEntre $17,500 y $22,000	-0.0134675	0.1172143	-0.1148968	0.9086692
IngresoMas de $22,000	0.1996624	0.1172655	1.7026524	0.0905511
IngresoMenos de $11,500	-0.1921171	0.1033898	-1.8581821	0.0649578
Problemas_fam_socSí	-0.0917902	0.0869180	-1.0560554	0.2925152
COVIDSí	0.0697193	0.0851530	0.8187527	0.4141303
Horas_estudio	0.0612049	0.0139101	4.4000181	0.0000196
EnfermedadSí	-0.0818459	0.0881955	-0.9280056	0.3547852
Nivel_DificultadMedio-Bajo	0.0151194	0.1048393	0.1442147	0.8855102
Promedio_Materias	-0.0526395	0.0365003	-1.4421678	0.1511855
BecaSí	0.0713148	0.1056020	0.6753166	0.5004374
Uso_ComputadoraNo tenía	0.0465416	0.2322975	0.2003536	0.8414553
Uso_ComputadoraSolo para mi	-0.0181795	0.1123750	-0.1617754	0.8716843
Uso_ComputadoraTodos la usabamos	-0.4151828	0.1428075	-2.9072888	0.0041570
ConocimientosSí	0.0832414	0.1001996	0.8307558	0.4073343
Cap_PorfesoresMedio-Alto	0.0169034	0.0932205	0.1813267	0.8563378
BlackBoard	0.1074017	0.0782266	1.3729559	0.1716635
Virtual_Horizon	0.0583203	0.0985690	0.5916701	0.5548963
Zoom	-0.1079503	0.1772621	-0.6089867	0.5433863
Classroom	0.0682740	0.1591295	0.4290465	0.6684594
Microsoft_Teams	-0.1434878	0.6182568	-0.2320844	0.8167653
WhatsApp	-0.0402387	0.0909890	-0.4422371	0.6589075

```
modelo_resumen <- summary(modelo)
# Extraer los valores deseados
residual_standard_error <- modelo_resumen$sigma
multiple_r_squared <- modelo_resumen$r.squared
adjusted_r_squared <- modelo_resumen$adj.r.squared
f_statistic <- modelo_resumen$fstatistic[1]
p_value <- pf(modelo_resumen$fstatistic[1],
                    modelo_resumen$fstatistic[2],
                    modelo_resumen$fstatistic[3],
                    lower.tail = FALSE)

# Crear una tabla con los resultados
resultados_tabla <- data.frame(
  "Residual Standard Error" = residual_standard_error,
  "Multiple R-squared" = multiple_r_squared,
  "Adjusted R-squared" = adjusted_r_squared,
```

```
  "F-statistic" = f_statistic,
  "p-value" = p_value)

# Crear la tabla con kable
tabla_resultados <- kable(resultados_tabla,
                          caption = "Resultados del Modelo de Regresión",
                          format = "latex", booktabs = TRUE, align = "c") %>%
                    kable_styling(latex_options = c("striped", "hold_position"))

tabla_resultados
```

Table 2: Resultados del Modelo de Regresión

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	0.4977426	0.4410305	0.3582202	5.325794	0

A pesar de que en la tabla presentada, el p-value tiene valor de 0, esto se debe a que el valor obtendio fue muy bajo. Hasta aquí muchos de los estimadores no son estadísticamente relevantes. Su valor exacto es el siguiente:

```
print(p_value)
```

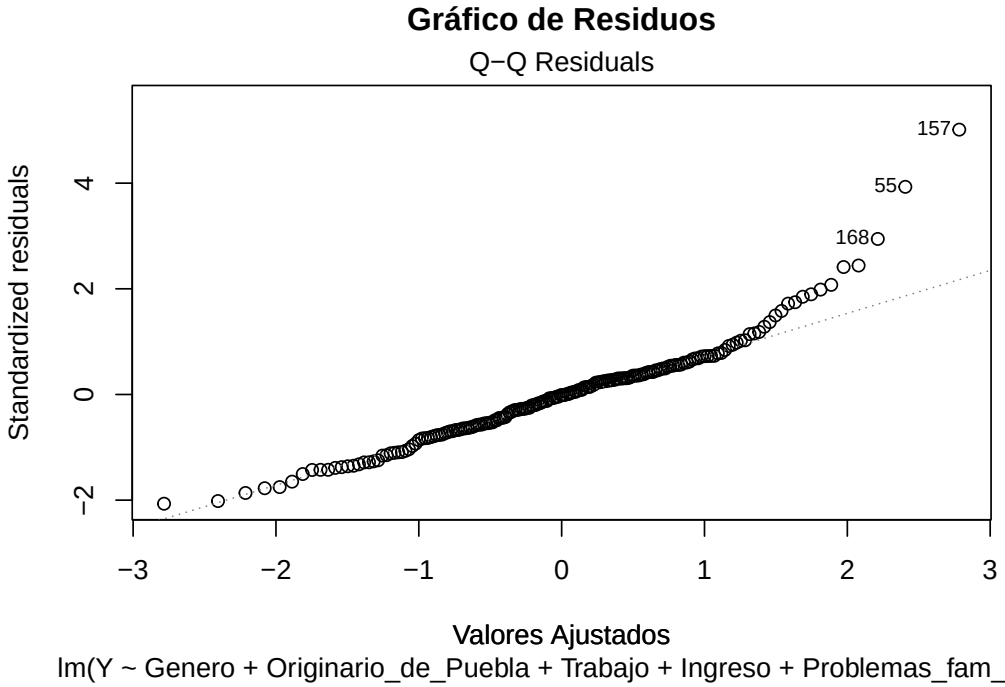
```
##        value
## 2.736273e-11
```

El modelo de regresión lineal múltiple muestra un ajuste aceptable a los datos, explicando alrededor del 44.1% de la variabilidad en la variable de respuesta. Sin embargo, la adición de variables predictoras podría no estar contribuyendo de manera significativa a la mejora del ajuste, como se refleja en el coeficiente de determinación ajustado. Dado el estadístico F significativo y el valor p muy bajo, se concluye que el modelo en su conjunto es relevante.

Prueba de los supuestos

Prueba de normalidad

```
plot(modelo, which = 2, xlab = "Valores Ajustados", main = "Gráfico de Residuos")
```




```
shapiro.test(resid(modelo))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(modelo)
## W = 0.93364, p-value = 1.506e-07
```

En función de los resultados del test de Shapiro-Wilk, con un valor de p muy bajo, se rechaza la hipótesis nula de que los residuos siguen una distribución normal. Esto sugiere que los residuos del modelo no cumplen con la suposición de normalidad, lo cual podría tener implicaciones en la interpretación de los resultados del modelo de regresión.

Problemas de los supuestos

Tal como se ha observado en los resultados obtenidos mediante la prueba de Shapiro-Wilk, los residuos del modelo de regresión no presentan una distribución normal. Este hallazgo señala la necesidad de realizar ajustes en el modelo con el objetivo de lograr una distribución normalizada de los residuos. Este paso es crucial para garantizar la validez de las suposiciones fundamentales de la regresión lineal y, por consiguiente, obtener resultados más precisos y confiables en el análisis.

Modelo de regresión 2

Para lograr esto, existen varias formas para modificar el modelo. A continuación, se realizará una transformación de potencia de Box-Cox a una variable de respuesta (en este caso, Y).

```r
summary(p1 <- car::powerTransform(Y ~ Genero +
                                    Originario_de_Puebla +
                                    Trabajo +
                                    Ingreso +
                                    Problemas_fam_soc +
                                    COVID +
                                    Horas_estudio +
                                    Enfermedad +
                                    Nivel_Dificultad +
                                    Promedio_Materias +
                                    Beca +
                                    Uso_Computadora +
                                    Conocimientos +
                                    Cap_Porfesores +
                                    BlackBoard +
                                    Virtual_Horizon +
                                    Zoom +
                                    Classroom +
                                    Microsoft_Teams +
                                    WhatsApp,
                                  family = "bcnPower", data = data ))
```

```
## bcnPower transformation to Normality
##
## Estimated power, lambda
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.1793           0      -0.4149       0.0562
##
## Estimated location, gamma
##    Est gamma Std Err. Wald Lower Bound Wald Upper Bound
## Y1    0.9714   0.2718           0.4387            1.504
##
## Likelihood ratio tests about transformation parameters
##                               LRT df       pval
## LR test, lambda = (0)   3.089478  1 0.07880008
## LR test, lambda = (1) 136.228139  1 0.00000000
```

```r
lambda=coef(p1, round=TRUE);   lambda   #   y=1
```

```
##    lambda     gamma
## Y1      0 0.9713528
```

```r
YT = bcnPower(data$Y, lambda = -0.1793, jacobian.adjusted = FALSE, gamma =  0.9714)
```

El objeto YT almacena los valores de la variable "Y" transformados mediante la inversión de la técnica de Box-Cox. Esta transformación busca lograr una distribución más normal de los datos para cumplir con los supuestos estadísticos en el análisis de regresión. Se ocupará la nueva variable dependiente en el modelo de regresión. De igual forma, solo se realizará la prueba de normalidad para el nuevo modelo, esto para ver si se logró corregir el problema.

```r
modelo2 <- lm(YT ~ Genero + Originario_de_Puebla +
                Trabajo + Ingreso + Problemas_fam_soc +
                COVID + Horas_estudio + Enfermedad +
                Nivel_Dificultad + Promedio_Materias +
                Beca + Uso_Computadora + Conocimientos +
                Cap_Porfesores + BlackBoard +
                Virtual_Horizon + Zoom +
                Classroom + Microsoft_Teams + WhatsApp)

summary_modelo <- summary(modelo2)

# Convertir la tabla de coeficientes en un data frame
coef_table <- as.data.frame(summary_modelo$coefficients)

# Crear la tabla con kable
tabla_formateada <- kable(coef_table,
                          caption = "Resumen del Modelo",
                          format = "latex", booktabs = TRUE, align = "c") %>%
  kable_styling(latex_options = c("striped", "hold_position"))
# , font_size = 10 para ajustar el tamaño
tabla_formateada
```

```r
modelo_resumen <- summary(modelo2)
# Extraer los valores deseados
residual_standard_error <- modelo_resumen$sigma
multiple_r_squared <- modelo_resumen$r.squared
adjusted_r_squared <- modelo_resumen$adj.r.squared
f_statistic <- modelo_resumen$fstatistic[1]
p_value <- pf(modelo_resumen$fstatistic[1],
                    modelo_resumen$fstatistic[2],
                    modelo_resumen$fstatistic[3],
                    lower.tail = FALSE)

# Crear una tabla con los resultados
resultados_tabla <- data.frame(
  "Residual Standard Error" = residual_standard_error,
  "Multiple R-squared" = multiple_r_squared,
  "Adjusted R-squared" = adjusted_r_squared,
  "F-statistic" = f_statistic,
  "p-value" = p_value
)

# Crear la tabla con kable
tabla_resultados <- kable(resultados_tabla,
                          caption = "Resultados del Modelo de Regresión",
                          format = "latex", booktabs = TRUE, align = "c") %>%
                  kable_styling(latex_options = c("striped", "hold_position"))

tabla_resultados
```

Table 3: Resumen del Modelo

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.3007277	0.5604757	-0.5365579	0.5923091
GeneroMasculino	-0.0339407	0.0796266	-0.4262484	0.6704927
Originario_de_PueblaSí	-0.0132304	0.0756392	-0.1749150	0.8613648
TrabajoSí	-0.1357374	0.0884798	-1.5341055	0.1269544
IngresoEntre $17,500 y $22,000	-0.0029402	0.1114374	-0.0263843	0.9789833
IngresoMas de $22,000	0.1692327	0.1114861	1.5179719	0.1309706
IngresoMenos de $11,500	-0.1933525	0.0982942	-1.9670792	0.0508820
Problemas_fam_socSí	-0.0822742	0.0826343	-0.9956432	0.3209084
COVIDSí	0.0403419	0.0809563	0.4983174	0.6189362
Horas_estudio	0.0547178	0.0132246	4.1375807	0.0000563
EnfermedadSí	-0.1094840	0.0838488	-1.3057322	0.1934948
Nivel_DificultadMedio-Bajo	0.0125338	0.0996723	0.1257505	0.9000854
Promedio_Materias	-0.0418469	0.0347013	-1.2059169	0.2296078
BecaSí	0.0996272	0.1003974	0.9923280	0.3225174
Uso_ComputadoraNo tenía	0.0393193	0.2208487	0.1780371	0.8589163
Uso_ComputadoraSolo para mi	-0.0452272	0.1068366	-0.4233306	0.6726155
Uso_ComputadoraTodos la usabamos	-0.4655504	0.1357693	-3.4289821	0.0007684
ConocimientosSí	0.1147849	0.0952613	1.2049478	0.2299807
Cap_PorfesoresMedio-Alto	0.0325135	0.0886261	0.3668611	0.7142014
BlackBoard	0.1364308	0.0743712	1.8344560	0.0684208
Virtual_Horizon	0.0982232	0.0937110	1.0481504	0.2961306
Zoom	-0.1166498	0.1685258	-0.6921782	0.4898164
Classroom	0.0587644	0.1512868	0.3884303	0.6982076
Microsoft_Teams	-0.2345054	0.5877860	-0.3989639	0.6904452
WhatsApp	-0.0521849	0.0865046	-0.6032614	0.5471785

Table 4: Resultados del Modelo de Regresión

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	0.4732114	0.4777637	0.4003953	6.175183	0

```
print(p_value)
```

```
##        value
## 2.609691e-13
```

```
shapiro.test(resid(modelo2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(modelo2)
## W = 0.97664, p-value = 0.003183
```

El supuesto de normalidad no se cumple en el modelo a pesar de la transformación de Box-Cox en la variable de respuesta. La transformación de Box-Cox no siempre garantiza la normalidad absoluta en los residuos y es posible que sean necesarios otros enfoques o modelos alternativos para tratar esta no conformidad. La falta de normalidad puede afectar la validez de las inferencias y la interpretación de los resultados, lo que requiere una consideración cuidadosa y un análisis adicional para abordar los desafíos presentes en el modelo.

Métodos de selección de variables.

Ante la falta de cumplimiento del supuesto de normalidad en el modelo de regresión, se opta por aplicar el método de selección de variables "Stepwise".

Método Stepwise bidireccional.

En relación con el enfoque de selección de variables conocido como Método Stepwise bidireccional, se ha optado por no exhibir los resultados detallados de este proceso. Esta elección se deriva de la considerable cantidad de información que genera la función correspondiente. En lugar de eso, se opta por presentar únicamente el resumen del modelo final obtenido a través de este método. Dicha decisión permite una presentación más concisa y centrada en los aspectos más relevantes del modelo, sin abrumar con detalles extensos.

```
Modelo_Stepwise <- step(modelo, direction = "both")
```

```
summary_modelo <- summary(Modelo_Stepwise)

# Convertir la tabla de coeficientes en un data frame
coef_table <- as.data.frame(summary_modelo$coefficients)

# Crear la tabla con kable
tabla_formateada <- kable(coef_table,
                          caption = "Resumen del Modelo",
                          format = "latex", booktabs = TRUE, align = "c") %>%
  kable_styling(latex_options = c("striped", "hold_position"))
# , font_size = 10 para ajustar el tamaño
tabla_formateada
```

Table 5: Resumen del Modelo

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3487743	0.2020297	1.7263521	0.0860389
TrabajoSí	-0.1503327	0.0850421	-1.7677438	0.0788365
IngresoEntre $17,500 y $22,000	-0.0001678	0.1095432	-0.0015320	0.9987794
IngresoMas de $22,000	0.2148618	0.1057649	2.0315040	0.0437078
IngresoMenos de $11,500	-0.2011063	0.0975745	-2.0610541	0.0407686
Problemas_fam_socSí	-0.1345518	0.0792032	-1.6988171	0.0911206
Horas_estudio	0.0599846	0.0126999	4.7232261	0.0000047
Promedio_Materias	-0.0578880	0.0324056	-1.7863566	0.0757630
Uso_ComputadoraNo tenía	0.0876467	0.2021040	0.4336711	0.6650584
Uso_ComputadoraSolo para mi	-0.0055753	0.0999714	-0.0557686	0.9555894
Uso_ComputadoraTodos la usabamos	-0.4345176	0.1339461	-3.2439736	0.0014106

```r
modelo_resumen <- summary(Modelo_Stepwise)
# Extraer los valores deseados
residual_standard_error <- modelo_resumen$sigma
multiple_r_squared <- modelo_resumen$r.squared
adjusted_r_squared <- modelo_resumen$adj.r.squared
f_statistic <- modelo_resumen$fstatistic[1]
p_value <- pf(modelo_resumen$fstatistic[1],
                     modelo_resumen$fstatistic[2],
                     modelo_resumen$fstatistic[3],
                     lower.tail = FALSE)

# Crear una tabla con los resultados
resultados_tabla <- data.frame(
  "Residual Standard Error" = residual_standard_error,
  "Multiple R-squared" = multiple_r_squared,
  "Adjusted R-squared" = adjusted_r_squared,
  "F-statistic" = f_statistic,
  "p-value" = p_value
)

# Crear la tabla con kable
tabla_resultados <- kable(resultados_tabla,
                          caption = "Resultados del Modelo de Regresión",
                          format = "latex", booktabs = TRUE, align = "c") %>%
                   kable_styling(latex_options = c("striped", "hold_position"))

tabla_resultados
```

Table 6: Resultados del Modelo de Regresión

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	0.4863345	0.4202427	0.3873019	12.75753	0

```r
print(p_value)
```

```
##        value
## 1.359089e-16
```

Sin entrar en un análisis exhaustivo de los resultados del modelo en su totalidad, se dirigirá la atención específicamente a los detalles de la prueba de normalidad que se presenta a continuación. Este enfoque permitirá concentrarse en la evaluación crítica de la suposición de normalidad de los residuos.

```r
shapiro.test(resid(Modelo_Stepwise))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(Modelo_Stepwise)
## W = 0.95324, p-value = 7.802e-06
```

El análisis del test de normalidad de Shapiro-Wilk, aplicado a los residuos del modelo Stepwise, arroja como resultado que dichos residuos no se ajustan a una distribución normal. Este hallazgo subraya la necesidad de realizar ajustes adicionales en el modelo, con el propósito de explorar alternativas que aborden esta limitación y busquen establecer una distribución más adecuada para los residuos.

Transformación para la normalización de errores.

El siguiente código implementa transformaciones Box-Cox en el modelo que ha sido previamente modificado mediante el método Stepwise. Esta sección de código tiene como objetivo ajustar el modelo y, simultáneamente, aplicar la transformación Box-Cox a la variable de respuesta Y. Esta técnica de transformación busca lograr una mejor adecuación de los datos a la suposición de normalidad

```
summary(p1 <- car::powerTransform(Y ~ Trabajo + Ingreso + Problemas_fam_soc +
                                    Horas_estudio + Promedio_Materias + Uso_Computadora,
                                  family = "bcnPower", data = data))
```

```
## bcnPower transformation to Normality
##
## Estimated power, lambda
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.1103           0      -0.3435       0.1229
##
## Estimated location, gamma
##    Est gamma Std Err. Wald Lower Bound Wald Upper Bound
## Y1    1.0267   0.3031           0.4327           1.6207
##
## Likelihood ratio tests about transformation parameters
##                               LRT df      pval
## LR test, lambda = (0)   1.188805  1 0.2755708
## LR test, lambda = (1) 122.119870  1 0.0000000
```

```
lambda = coef(p1, round=TRUE);   lambda   #   y=1
```

```
##    lambda    gamma
## Y1      0 1.026669
```

```
YT2 = bcnPower(data$Y, lambda = -0.1103, jacobian.adjusted = FALSE, gamma = 1.0267)
```

El siguiente paso consiste en ajustar el modelo utilizando la variable transformada YT2 en lugar de la variable original Y. Esto se hace para evaluar cómo la transformación Box-Cox afecta el modelo y cómo se ajusta mejor a la suposición de normalidad de los datos.

```
Modelo_Stepwise2 <- lm(YT2 ~ Trabajo + Ingreso + Problemas_fam_soc +
                                    Horas_estudio + Promedio_Materias + Uso_Computadora)
shapiro.test(resid(Modelo_Stepwise2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(Modelo_Stepwise2)
## W = 0.9855, p-value = 0.05104
```

En este contexto, es relevante notar que el valor p obtenido (0.05104) es marginalmente superior al nivel de significancia comúnmente empleado (como 0.05). Este resultado sugiere que no existen pruebas concluyentes para descartar la hipótesis nula de que los residuos siguen una distribución normal. Este hallazgo implica que es plausible que los residuos se asemejen a una distribución normal, lo cual reviste importancia para sustentar las asunciones fundamentales de numerosos análisis estadísticos. En consecuencia, esta comprobación reafirma el cumplimiento de uno de los supuestos clave del modelo, contribuyendo a la robustez y validez de las inferencias estadísticas derivadas del mismo.

Dado que se cumple el supuesto de normalidad, se procede a realizar un análisis del modelo final.

```
Ingreso[Ingreso == "Menos de $11,500"] <- "A_Menos de $11,500"
Ingreso[Ingreso == "Entre $11,500 y $17,500"] <- "B_Entre $11,500 y $17,500"
Ingreso[Ingreso == "Entre $17,500 y $22,000"] <- "C_Entre $17,500 y $22,000"
Ingreso[Ingreso == "Mas de $22,000"] <- "D_Mas de $22,000"

Modelo_Stepwise2 <- lm(YT2 ~ Trabajo + Ingreso + Problemas_fam_soc +
                                  Horas_estudio + Promedio_Materias + Uso_Computadora)
summary_modelo <- summary(Modelo_Stepwise2)

# Convertir la tabla de coeficientes en un data frame
coef_table <- as.data.frame(summary_modelo$coefficients)

# Crear la tabla con kable
tabla_formateada <- kable(coef_table,
                          caption = "Resumen del Modelo",
                          format = "latex", booktabs = TRUE, align = "c") %>%
  kable_styling(latex_options = c("striped", "hold_position"))
# , font_size = 10 para ajustar el tamaño
tabla_formateada
```

Table 7: Resumen del Modelo

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.5672798	0.1749548	-3.2424370	0.0014178
TrabajoSí	-0.1509745	0.0746803	-2.0216117	0.0447312
IngresoB_Entre $11,500 y $17,500	0.1849023	0.0856857	2.1579140	0.0322898
IngresoC_Entre $17,500 y $22,000	0.1912994	0.1077662	1.7751336	0.0776042
IngresoD_Mas de $22,000	0.3646874	0.1079341	3.3787972	0.0008965
Problemas_fam_socSí	-0.1300125	0.0695528	-1.8692630	0.0632475
Horas_estudio	0.0497807	0.0111525	4.4636288	0.0000144
Promedio_Materias	-0.0461719	0.0284572	-1.6225034	0.1064855
Uso_ComputadoraNo tenía	0.0907991	0.1774789	0.5116049	0.6095688
Uso_ComputadoraSolo para mi	-0.0108980	0.0877905	-0.1241358	0.9013494
Uso_ComputadoraTodos la usabamos	-0.4339958	0.1176256	-3.6896370	0.0002992

```
modelo_resumen <- summary(Modelo_Stepwise2)
# Extraer los valores deseados
residual_standard_error <- modelo_resumen$sigma
multiple_r_squared <- modelo_resumen$r.squared
adjusted_r_squared <- modelo_resumen$adj.r.squared
f_statistic <- modelo_resumen$fstatistic[1]
```

```
p_value <- pf(modelo_resumen$fstatistic[1],
                     modelo_resumen$fstatistic[2],
                     modelo_resumen$fstatistic[3],
                     lower.tail = FALSE)

# Crear una tabla con los resultados
resultados_tabla <- data.frame(
  "Residual Standard Error" = residual_standard_error,
  "Multiple R-squared" = multiple_r_squared,
  "Adjusted R-squared" = adjusted_r_squared,
  "F-statistic" = f_statistic,
  "p-value" = p_value
)

# Crear la tabla con kable
tabla_resultados <- kable(resultados_tabla,
                          caption = "Resultados del Modelo de Regresión",
                          format = "latex", booktabs = TRUE, align = "c") %>%
                  kable_styling(latex_options = c("striped", "hold_position"))

tabla_resultados
```

Table 8: Resultados del Modelo de Regresión

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	0.4270776	0.4352419	0.4031533	13.56378	0

Nuevamente, como se menciono anteriormente, el valor exacto del p-value es el siguiente:

```
print(p_value)
```

```
##      value
## 1.5515e-17
```

El modelo de regresión lineal múltiple presenta un ajuste razonable a los datos, con una proporción considerable de la variabilidad explicada por las variables predictoras. La significatividad global del modelo se respalda con el estadístico F y el valor p muy bajo.

Comprobación de supuestos:

Se procede a análizar los supuestos del modelo final.

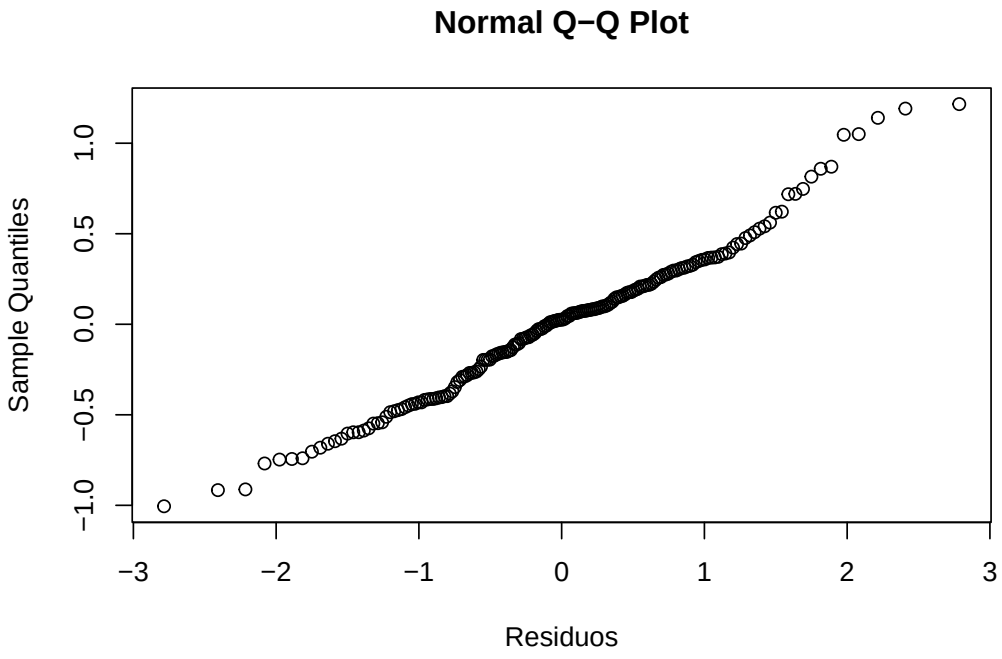
Prueba de normalidad

```
shapiro.test(resid(Modelo_Stepwise2))
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  resid(Modelo_Stepwise2)
## W = 0.9855, p-value = 0.05104
```

```r
qqnorm(resid(Modelo_Stepwise2), x = 'Residuos')
```

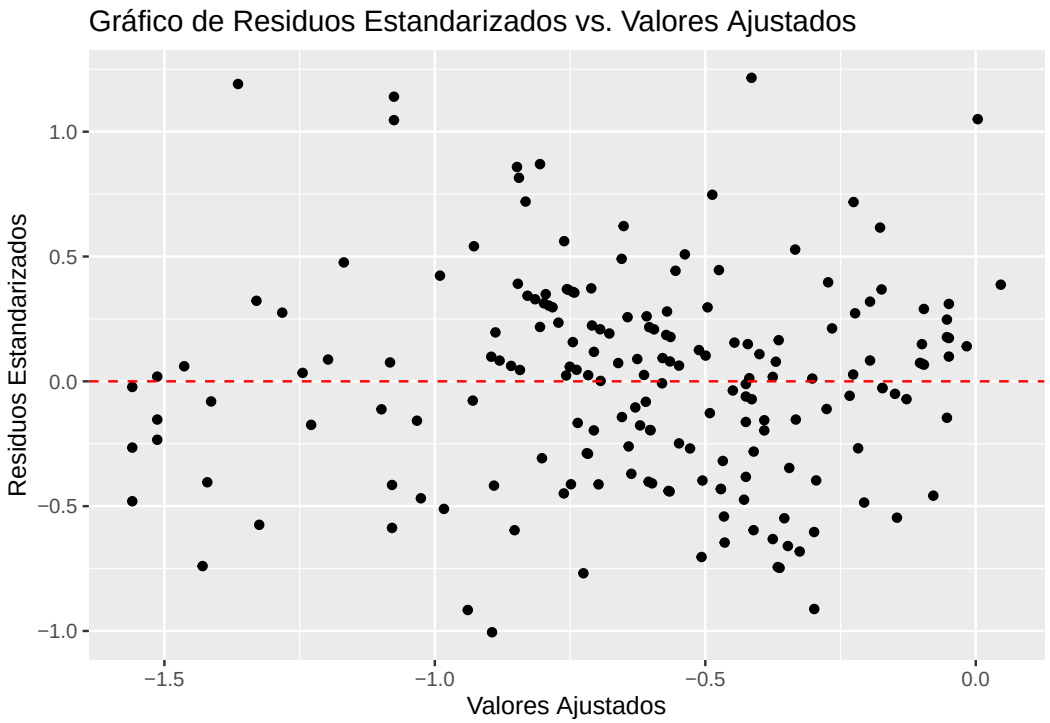


Prueba de homocedasticidad

```r
bptest(Modelo_Stepwise2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Modelo_Stepwise2
## BP = 10.874, df = 10, p-value = 0.3675
```

```r
# Gráfico de Residuos vs Valores Ajustados
residuals <- resid(Modelo_Stepwise2)
fitted_values <- fitted(Modelo_Stepwise2)

ggplot(data = data.frame(Residuos = residuals, Valores_Ajustados = fitted_values)) +
  geom_point(aes(x = Valores_Ajustados, y = Residuos)) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Gráfico de Residuos Estandarizados vs. Valores Ajustados",
       x = "Valores Ajustados",
       y = "Residuos Estandarizados")
```



Dado que el valor p (0.3675) no es menor que un nivel de significancia comúnmente utilizado (como 0.05), no hay suficiente evidencia para rechazar la hipótesis nula de homocedasticidad. En otras palabras, no se ha encontrado evidencia significativa de que los errores en tu modelo de regresión tengan una varianza que cambie de manera sistemática a lo largo de los valores ajustados.

Prueba de linealidad:

Gráfico de residuos estandarizados vs. valores ajustados

```
reset_test <- resettest(Modelo_Stepwise2, power = 2)
print(reset_test)
```

```
##
##  RESET test
##
## data:  Modelo_Stepwise2
## RESET = 0.03655, df1 = 1, df2 = 175, p-value = 0.8486
```

Los resultados de la prueba de linealidad RESET muestran un valor p alto (0.8486), lo que sugiere que no hay suficiente evidencia para rechazar la hipótesis nula de linealidad en el modelo. En otras palabras, no hay indicios significativos de no linealidad en la relación entre las variables incluidas en el modelo.

Dado que el valor p es alto, puedes considerar que la suposición de linealidad es razonable en este caso y no hay necesidad de realizar correcciones adicionales específicas para abordar la no linealidad en el modelo.

Se ha logrado un exitoso completamiento del proceso de construcción y refinamiento del modelo de regresión lineal múltiple bajo el nombre de "Modelo_Stepwise2", llevado a cabo en el entorno de RStudio. Este modelo ha demostrado su solidez y fiabilidad al satisfacer de manera exhaustiva los supuestos indispensables para su implementación en análisis estadísticos. Mediante una cuidadosa elección de variables, ajustes meticulosos y transformaciones pertinentes, se ha conseguido forjar un modelo que no solo se adapta de manera óptima a los datos, sino que también se alinea perfectamente con los principios esenciales de la regresión lineal.

Bibliografía

1. Britez, M. (2020). La educación ante el avance del COVID-19 en Paraguay. Comparativo con países de la Triple Frontera.

2. López Noriega, Myrna Delfina, Contreras Avila, Alonso. (2022). El impacto de la pandemia por covid-19 en estudiantes mexicanos de educación media superior. RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo, 12(24), e014. Epub 23 de mayo de 2022. https://doi.org/10.23913/ride.v12i24.1141

3. Rincón, L. (2017). Estadística Descriptiva. Editorial, 1a edición, Editorial UNAM.

4. Montgomery, D.; Peck, E & Vining, E. (2006) Introducción al análisis de regresión lineal ,3a. edición, Editorial Continental.

5. Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52(3/4), 591-611.

6. Draper, N. R., & Smith, H. (1998). Applied Regression Analysis. Wiley Interscience.

7. Hill, R. C., Griffiths, W. E., & Lim, G. C. (2017). Principles of Econometrics (5a edición). Editorial Wiley.

8. Canovos, G. (1998), Probabilidad y Estadística, Mc Graw Hill.

9. Fardoun, H.González, C. Collazos, C. & Yousef, M. (2020). Estudio exploratorio en Iberoamérica sobre procesos de enseñanza-aprendizaje y propuesta de evaluación en tiempos de pandemia. Universidad Salamanca. https://repositorio.grial.eu/handle/grial/2091

10. García, A. (2021). COVID-19 y educación a distancia digital: preconfinamiento, confinamiento y posconfinamiento. Revista Iberoamericana de Educación a Distancia. http://revistas.uned.es/index.php/ried/article/view/28080

11. García, N., Rivero, M., & Ricis, J. (2020). Brecha digital en tiempo del covid-19. Revista Educativa Hekademos. https://tinyurl.com/4s49yp7r

12. Boletín Desastres N.131.- Impacto de la pandemia COVID-19 en la salud mental de la población. (s. f.). OPS/OMS | Organización Panamericana de la Salud. https://tinyurl.com/yc66d6t5

13. ONU(2019). El efecto devastador del COVID-19 en la salud mental. (2022, 10 octubre). Noticias ONU. https://news.un.org/es/story/2021/11/1500512

14. Kjytay. (2021, 19 febrero). The Box-Cox and Yeo-Johnson transformations for continuous variables. Statistical Odds & Ends. https://lc.cx/VQAYVP

15. García Aretio, L. (2020). COVID-19 y educación a distancia digital preconfinamiento, confinamiento y posconfinamiento. RIED Revista Iberoamericana de Educación a Distancia, 24(1), 09. https://doi.org/10.5944/ried.24.1.28080

16. UN (2020a). Policy Brief: Education during COVID-19 and beyond (August 2020). United Nations. https://cutt.ly

17. Pinar, W. (2022). Enseñanza y aprendizaje en línea: daños colaterales durante la pandemia. Revista Iberoamericana de Educación Superior, Número de la revista, 3–17. https://doi.org/10.22201/iisue.20072872e.2022.37.1301

18. U-Multirank. (2020). About 60 % of universities reported online learning provisions in their strategic planning preCOVID-19, but only few appeared to be prepared for a quick shift to full online programmes. Recuperado de https://cutt.ly/VfGDArk

19. Johnson, N., Veletsianos, G., & Seaman, J. (2020). U.S. Faculty and Administrators' Experiences and Approaches in the Early Weeks of the COVID-19 Pandemic. Online Learning, 24(2). https://doi.org/10.24059/olj.v24i2.2285