
BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA.

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS.


# *"ANÁLISIS DE REGRESIÓN APLICADA AL PESO DE INFANTES EN UN SERVICIO POLIVALENTE."*

T E S I S

PARA OBTENER EL TÍTULO DE:

**LICENCIADO EN MATEMÁTICAS.**

P R E S E N T A


**FROYLAN HERNÁNDEZ DURÁN.**

DIRECTOR DE TESIS:

**DR. FERNANDO VELASCO LUNA.**



**PUEBLA, PUE. MARZO, 2024.**

# *Dedicatoria*

*A mis abuelitas Irene y Francisca*
*quienes lo soportaron todo*
*y se marcharon un dia de verano.*

*A mi familia, gracias por tanto,*
*por estar siempre a mi lado.*

*A mis alumnos, quienes fueron*
*mis grandes maestros y me enseñaron*
*a valorar lo más simple de la vida.*

*A ese pequeño niño que*
*creyó que jamás sería posible*
*y ahora puede volar tan alto como desee.*

# *Agradecimientos*

*Estas líneas no me alcanzaran para agradecer a todos los que me han apoyado durante este tiempo, a Dios quien me ha dado la fortaleza de vivir y encontrar un refugio en él cuando me he sentido perdido.*

*Mi gratitud a la Benemérita Universidad Autonoma de Puebla, así como a la Facultad de Ciencias Físico Matemáticas, por abrirme las puertas en cada uno de sus espacios, por las actividades en las que logré participar, adquirir nuevas experiencias y aprendizajes.*

*La Educación Superior a través de las universidades públicas es una de las formas de garantizar el derecho a la educación a todos los sectores, de esta forma tenemos el compromiso de retribuir a la sociedad un poco de lo que nos ha brindado.*

*Le agradezco infinitamente a mi director de tesis al Dr. Fernando Velasco Luna por todo el apoyo brindado, las orientaciones, la motivación, pero sobretodo la paciencia, gracias por todo el acompañamiento en la realización de esta investigación, quien a pesar del trabajo a distancia me brindó las recomendaciones necesarias y mantuvo toda la disponibilifdad en cada una de mis inquietudes, sin todo esto no sería posible culminarla.*

*Para la revisión de está tesis agradezco el apoyo brindado de mi jurado, M.C Sergio Adán Juárez, Dr. Francisco Solano Tajonar Sanabria y al Dr. Victor Hugo Vázquez Guevara, quienes amablemente mostraron toda la disponibilidad para la revisión y brindar la retroalimentación de este trabajo. ¡Gracias!*

*Mi agradecimiento infinito y mi admiración a mis padres, a mi mamá Flor y a mi papá Tino, jamás olvidaré lo que han hecho por mí, desde aquel primer día: en que se enteraron de que había aprobado el examen para ingresar a la universidad y que no dudaron en hacer todo lo posible para que pudiera aventurarme en una etapa más de mi vida.*

*El camino no ha sido fácil, sin embargo, gracias a ustedes el trayecto ha sido alentador, gracias por todo y por la motivación oportuna en los momentos en donde quería rendirme.*

*Agradezco a mis hermanos: Coni, quien con un mensaje a distancia me enviaba todo su apoyo y quien a veces me brindó palabras de aliento.*

*A mi hermano Paco, quien me apoyo económicamente en estos últimos pasos por la carrera y me ha apoyado emocionalmente, quien me ha hecho confiar en mí mismo y mostrarme la capacidad que tengo para creer en mí, por todo el apoyo brindado y la orientación para correr riesgos, estas palabras no cumplen con todo el agradecimiento que tengo hacia a ti, pero es una pequeña muestra de gratitud de todo lo que has hecho por mí.*

*A mis tías Irene y Mica, Irene quien me ayudó en mis primeros días en la ciudad, ya que era la primera vez que me mudaba a Puebla.*

*A Mica quien me soportó estos últimos años, gracias por las risas y los paseos.*

*A Domi mi prima, quien ha sido una hermana para mí, gracias por apoyarme en la inscripción, aunque ambos no conocíamos bien la ciudad, me alentó a perseguir este sueño.*

*A mis sobrinos, a quienes veo una oportunidad para apoyarlos y alentarnos en su futuro.*

*También quiero agradecer a la comunidad de Tecuahuta, donde presté mi servicio social durante cuatro ciclos escolares a través de los distintos niveles (Preescolar, Primaria y Secundaria), en dos etapas importantes de mi vida la primera como alguien que buscaba la oportunidad de lograr un proyecto y la segunda como egresado, a manera de retribución por todo el apoyo brindado.*

*Donde aprendí mucho de mis primeros alumnos de primaria (hoy jóvenes), quienes nos conocimos un día de verano, juntos comenzamos a vivir la aventura de aprender, jugar y convivir, ustedes aprendieron de mi, pero yo me llevé grandes lecciones de vida.*

*En esta última etapa de crecimiento profesional quiero agradecer a mis pequeños de preescolar y a mis alumnos de secundaria. ¡Gracias por todo!*

*Quiero agradecer a cada uno de los padres de familia de cada uno de los servicios educativos, sin su apoyo no sería posible culminar mi quehacer educativo.*

*También agradezco al Dr. José Jacobo Oliveros Oliveros, por el apoyo brindado durante las práctica profesionales, muchas gracias por la paciencia y sobre todo el ánimo y la comprensión que tuvo hacia a mí en uno de los momentos más difíciles de mi vida.*

*A mis amigos que conocí aquel primer día Belén y Vero, quienes no pudieron continuar. A Ismael ya que fue el primer compañero que habló conmigo el primer día de la carrera.*

*A Jessi quien se sumó a esta amistad años más tarde, pero fue suficiente su apoyo por aquellos días en los que los ánimos no estaban del todo bien, fue un gusto tomar clases contigo*

*A Karim por todo el apoyo académico y emocional a distancia, te agradezco las palabras de aliento y la confianza brindada, gracias por cada uno de los mensajes de apoyo, por los días de plática, clases y demás. ¡Por la experiencia!*

*A Jessi de mate aplicadas, las risas no faltaron en los cursos que coincidimos.*

*A los profesores que confiaron en mí y no dudaron en brindarme su apoyo, a la maestra Miriam Sanchéz, gracias por las clases de italiano.*

*Finalmente, gracias a mí que he tenido la oportunidad de vivir, de creer y de soñar, a cada uno de los desafíos que he enfrentado y me han hecho fuerte, quien ha soportado todo: altas y bajas, obstáculos, fracasos, pero todos esos aprendizajes me han hecho creer en mí mismo y sentirme capaz de todo lo que he logrado.*

*A mi pequeña comunidad: Ixtahuata donde crecí y viví durante mis primeros 19 años.*

*"No importa de donde vengamos y las circunstancias que enfrentemos, sino el coraje y el valor que tengamos para cumplir nuestros sueños".*

# Índice

# Índice de figuras

# Índice de cuadros

# Resumen

La alimentación durante los primeros años de vida es crucial para el desarrollo pleno en los infantes, sin embargo, de acuerdo con los contextos geográficos en los que se logran desarrollar, se presentan situaciones, en donde el sano crecimiento se ve frenado. Por otro lado, el peso, la edad y la estatura, son variables que permiten determinar el estado de salud del menor de acuerdo con estándares establecidos por la OMS.

En este estudio se trató de establecer una posible relación entre el peso como variable respuesta y la edad, la estatura y el número de vasos de leche que consumieron niños menores de $5$ años como variables explicativas, durante el ciclo escolar $2021$-$2022$ en un servicio polivalente. Se llevó a cabo un análisis estadístico entre la variable respuesta y entre cada una de las variables explicativas, para el análisis definitivo se realizó un análisis de regresión múltiple.

Los resultados muestran que la edad presenta una relación lineal con el peso del infante, es decir, a medida que la edad aumenta a través del tiempo, el peso se incrementa, mientras que la estatura y el número de vasos de leche que consumen al día no influyen en el peso de los menores.

# Introducción

## 0.1. Marco Conceptual

La alimentación del infante desde su gestación requiere de una constante aportación de nutrientes necesarios para su desarrollo, no solo en sus primeros meses de vida, sino que este es un procedimiento que se realiza durante todo su crecimiento a través del tiempo, de esta manera podrá mantener la salud en condiciones favorables que lo ayudarán a fortalecer sus condiciones físicas, mentales, psicomotoras, emocionales, sociales, entre otras.

El crecimiento físico se refiere a un aumento en el tamaño del cuerpo (longitud o altura y peso) y en el tamaño de los órganos. Desde el nacimiento hasta la edad de 1 o 2 años. Después de este rápido crecimiento del lactante y del niño, el crecimiento se ralentiza hasta que se llega al crecimiento acelerado de los adolescentes. El crecimiento desde el nacimiento hasta la adolescencia ocurre en 2 fases.

- Fase 1 (desde el nacimiento hasta la edad de 1 o 2 años): esta fase es de crecimiento rápido, aunque la velocidad disminuye a lo largo de ese periodo.
- Fase 2 (desde alrededor de los 2 años hasta el comienzo de la pubertad): en esta fase, el crecimiento se produce con incrementos anuales relativamente constantes [10].

El peso es la medida de la masa corporal que tiene el niño en cantidad de kilogramos, este valor permite determinar las condiciones de nutrición en la que se encuentra el menor. En los primeros meses de vida, para calcular el peso del menor es acostado en un dispositivo como la balanza pediátrica, posteriormente a medida que crece es medido de manera vertical a través de una báscula.

La talla o estatura es la distancia del cuerpo humano desde la cabeza hasta el talón de los pies, durante los primeros meses de vida al igual que en el cálculo del peso el menor es acostado y más adelante es medido de forma vertical.

La edad es el tiempo de vida que ha tenido el ser humano desde su nacimiento, al inicio de su crecimiento se mide en meses y posteriormente en años.

El Índice de Masa Corporal (IMC), es un número el cual es calculado a partir del peso, la estatura y la edad a partir del valor que se calcula se asignan categorías relacionados con el estado de salud como: bajo peso, peso normal, sobrepeso u obesidad.

Por otro lado en las instituciones de salud públicas y privadas se registra desde el nacimiento, el peso y la talla del niño, esto con el fin de contar con antecedentes que les permita detectar a tiempo alteraciones en el crecimiento, de esta manera brinda las recomendaciones necesarias que le ayudarán a mantener la salud y desarrollar su proceso de crecimiento de manera adecuada y de acuerdo con su edad. Cabe señalar que los registros que se obtienen de cada individuo son tomados a partir de ciertos indicadores recomendados por la OMS (Organización Mundial de la Salud).

Durante el crecimiento del bebé es imprescindible que consuma la leche materna, puesto que es uno de los principales alimentos para su desarrollo durante los primeros seis meses de vida o incluso hasta los dos años, en algunos casos existen ciertos sustitutos que pueden emplearse cuando el bebé no logra recibir este alimento al nacer.

Recordemos que los nutrimentos que proporciona la leche materna son esenciales para la salud y de esta manera podrán evitarse enfermedades como la desnutrición, la cual es uno de los principales padecimientos durante los primeros años del menor.

De acuerdo con Guillen López y Vela Amieva [7] "Los expertos en nutrición infantil consideran que la leche entera de vaca no debe ser introducida en la dieta de los niños menores de un año; sin embargo, algunos autores sugieren que puede darse un poco antes, entre el noveno y décimo mes de vida".
Si bien el peso y la estatura del menor son influenciados por la genética, también los hábitos alimenticios que se tengan y qué tan buenos sean, intervienen en el crecimiento del ser humano.

Cabe mencionar que en las zonas rurales e indígenas del país se vive un panorama en donde la alimentación del niño se ve afectada por diversos factores tales como: la pobreza, el bajo aporte alimenticio de productos, la modificación de la alimentación del menor, por mencionar solo algunos.

## 0.2. Antecedentes

Se han llevado a cabo distintas investigaciones dónde algunas variables tales como la edad, la estatura y el peso han sido utilizados para establecer relaciones de interés, con el fin de estudiar un fenómeno

o situación que se presenta en determinados rangos de edad con respecto al crecimiento, la deserción escolar, etc.

Alvarez y Montoya [2] llevaron a cabo un estudio con la finalidad de evaluar la velocidad media de crecimiento de $259$ niños con rango de edad de $2$ a $10$ años, los cuales pertenecen a 150 familias rurales de Marinilla, Colombia. Llevaron a cabo un estudio descriptivo. En una muestra representativa de familias campesinas productoras de hortaliza y que tuvieran niños de 2 a 10 años. El resultado fue que el $26.8\,\%$ de los niños y el $28.9\,\%$ de las niñas tuvieron una velocidad media de ganancia de peso.

Aguilera y Quintana [1], analizaron la asociación entre el peso y las tasas de deserción escolar, la repetición del año y el rezago escolar en niños de $12$ a $14$ años y adolescentes de $15$ a $18$ años en México, dichos datos fueron tomados del año $2002$ y $2006$ de la Encuesta Nacional sobre niveles de Vida de los Hogares. Establecieron un modelo teórico de producción familiar para intentar determinar la relación entre sobrepeso-obesidad y el rendimiento escolar.
De esto, encontraron una pequeña asociación negativa entre peso normal y las tasas de deserción escolar y repetición del año en mujeres de $12$-$14$ años y de $15$-$18$ años que viven en comunidades urbanas, además encontraron una asociación negativa entre tener peso normal y la repetición del año en hombres de $15$ a $18$ años de comunidades rurales.

Yugar y col. [15], llevaron a cabo un estudio para determinar la incidencia de la presencia de talla baja en preescolares de $2$ a $5$ años quienes acudieron a consulta externa del Servicio de Pediatría del Hospital La Paz (Bolivia). Para eso realizaron un estudio descriptivo, el cuál involucró la revisión de $400$ historias clínicas de pacientes preescolares de $2$ a $5$ años. Los resultados fueron que la incidencia global para talla baja fue de $62$ niños con talla baja $(15.5\,\%)$ y $338$ niños con talla normal para la edad $(84.5\,\%)$.

Díaz y Da Costa [6], caracterizaron los hábitos alimentarios y estado nutricional de $125$ binomios padres escolares entre $3$ y $5$ años de un centro de educación inicial en Santa Elena, Ecuador, durante el año $2016$. El estudio fue de tipo descriptivo. Los resultados que obtuvieron fueron que el estado nutricional en los niños fue deficiente, debido al consumo de proteínas de alto valor biológico $(3.97\,\%)$ y $8.32\,\%, 7.94\,\%$y $4.70\,\%$ demostraron alto consumo de baja calidad nutricional, esto de acuerdo con la edad materna, el nivel de instrucción y nivel socioeconómico.

De acuerdo con la información de esos autores, cabe mencionar que la investigación realizada permitirá conocer la manera en que se relacionan las variables de interés estatura, edad y el número de vasos de leche, con respecto al peso, para eso se considera una muestra, de 31 niños con rango de edad de $1$ a $5$ años de una localidad rural e indígena del estado de Puebla de un servicio polivalente.

## 0.3. Planteamiento del Problema

Los estudios entre las variables peso, edad y estatura, se han realizado con mayor frecuencia en unidades de salud públicas y privadas o instituciones de educación, donde se tiene mayor cantidad de datos, sin embargo para servicios educativos pequeños no se tiene el mismo impacto, dado que los datos que se recopilan, son menores, por la característica propia de las localidades que se intregran en cada zona geográfica.

A pesar de que se tienen estudios a nivel mundial y nacional con respecto al comportamiento del peso y la relación existente entre la estatura y la edad de determinada población. Particularmente en el municipio de Cuetzalan, si bien se registran periódicamente estos datos con el fin de valorar la salud de determinado individuo, no se cuenta con un análisis estadístico que permita relacionar además de las variables que se tienen recabadas con el número de vasos de leche que consumen al día los infantes de la localidad de Tecuahuta.

## 0.4. Justificación.

Se espera que los resultados obtenidos en esta investigación sean un antecedente que permita tomar decisiones a futuro con respecto a la alimentación de los infantes. Además de que en estudios posteriores se puedan agregan nuevas variables de interés y que esta información, sea un soporte hacia nuevas investigaciones en localidades rurales o indígenas.

El tener un adecuado desarrollo en menores de 5 años, con respecto a su alimentación permitirá que se cubra un sano crecimiento incidiendo principalmente en los aprendizajes que logre afianzar en su etapa escolar. De esta manera, se podrá prever medidas que ayuden al fortalecimiento en el aumento del peso y la estatura.

## 0.5. Objetivos

### 0.5.1. Objetivo General

Establecer una posible relación entre el peso como variable respuesta y la edad, la estatura y el número de vasos de leche que consumen al día menores de 5 años como variables explicativas.

### 0.5.2. Objetivos Especificos

- Explorar la posible relación entre el peso y la estatura.
- Explorar la posible relación entre el peso y la edad.
- Explorar la posible relación entre el peso y el número de vasos de leche que consumen los infantes.

## 0.6. Breve Descripción del Contenido.

El contenido de la tesis está estructurado como a continuación se describe.

Primero se presenta la introducción donde se establece la fundamentación necesaria para el desarrollo del tema de investigación, así como el objetivo general y los objetivos especificos.

En el Capítulo 1: Se establece la teoría de los modelos de regresión lineal simple y múltiple respectivamente, utilizando el método de mínimos cuadrados para el cálculo de la ecuación de regresión ajustada.

En el Capítulo 2: Se establece la teoría de residuales de los modelos de regresión, para la validación de los supuestos del modelo de regresión lineal.

En el capítulo 3: Se desarrolla la metodología estadística empleada para el análsis de los datos y posteriormente se da a conocer los resultados obtenidos.

En el capítulo 4: Se dan a conocer las conclusiones de la investigación.

# Capítulo 1

# Teoría de Análisis de Regresión Lineal

La Estadística es un área de las Matemáticas que ha mostrado influencia en diversas áreas del conocimiento como: Finanzas, Economía, Biología, en particular en las ciencias de la salud el cual ha permitido analizar y dar respuesta a diversos modelos de interés.

Por otro lado el análisis de regresión lineal es una herramienta estadística que permite analizar la relación entre, distintas variables hacia una variable que depende de ellas, a fin de explicar el comportamiento y la incidencia significativa que tienen las observaciones determinadas en una muestra, de esta manera permite relacionar pronósticos en observaciones que se tengan a futuro.

La primera forma de regresiones lineales documentada fue el método de los mínimos cuadrados, publicado por el matemático francés Adrien-Marie Legendre en 1805, en un apéndice del libro sobre la órbita de los cometas. Más tarde en 1809, el matemático alemán Johann Carl Friederich Gauss publicó sus resultados en *Theoría motus corporum coelestium*, en el cual introdujo el método de los mínimos cuadrados mediante el uso de conceptos estadísticos, como la distribución normal.

Etimológicamente, el término regresión fue acuñado por el estadístico británico Francis Galton en el siglo XIX, para describir un fenómeno biológico, al comparar la estatura de padres e hijos [9].

El análisis de regresión es un método conceptualmente simple para investigar las relaciones funcionales entre variables. La relación se expresa a través de una ecuación o modelo que conecta la variable respuesta y una o más variables explicativas. [5].

## 1.1. Modelo de Regresión Lineal Simple

En el **Modelo de Regresión Lineal Simple** se intenta modelar la relación entre dos variables de interés. Para un modelo de regresión lineal simple, podemos usar un modelo de la forma.

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1.1}$$

Donde:

- $y$ es la variable dependiente o respuesta.
- $x$ es la variable explicativa.
- $\epsilon$ es la variable aleatoria de error, es decir, los errores de medición que se presentan en el modelo [13].
- $\beta_0$ y $\beta_1$ son coeficientes desconocidos, tales que $\beta_0$ es la ordenada al origen y $\beta_1$ es la pendiente de la recta [9].

De (1.1) se tiene que para $n$ observaciones el *modelo de regresión lineal simple* esta dado como sigue:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n. \tag{1.2}$$

Donde los coeficientes de la recta ajustada son:

- La pendiente $\beta_1$ es la media del cambio de la distribución de $y$ producido por un cambio unitario en $x$.
- Si en el intervalo de los datos incluye $x = 0$ entonces la ordenada al origen $\beta_0$ es la media de la distribución de la respuesta $y$. Si no incluye al cero $\beta_0$ no tiene interpretación [12].

## 1.2. Supuestos del Modelo

En este sección, se hará mención de algunos supuestos al modelo de regresión lineal establecido anteriormente, dado que estamos interesados en poder estimar los coeficientes y la varianza del error del modelo (1.2), utilizando el Método de Mínimos Cuadrados que se dará a conocer más adelante.

Suponemos que $y_i, \epsilon_i$ son variables aleatorias y los valores de $x_i$ son constantes conocidas.

Para completar el modelo (1.2), se harán los siguientes supuestos.

1. **Modelo de regresión lineal:** El modelo de regresión es lineal en los coeficientes. Es decir, el modelo como se muestra en la ecuación (1.2)

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

2. **Valores de $x_i$ independientes del término del error:** Los valores que toma la variable explicativa $x$ pueden considerarse fijos en muestras repetidas (el caso de la explicativa fija), o haber sido muestreados junto con la variable respuesta $y$.

   En el segundo caso se supone que las variables $x$ y el termino de error son independientes, esto se expresa como, $cov(x_i, \epsilon_j) = 0$ donde $i$ y $j$ son observaciones diferentes.

   En general la covarianza entre un valor independiente $x_i$ y el termino del error $\epsilon_j$ no siempre es cero, pues en algunas ocasiones puede presentarse multicolinealidad entre los valores independientes.

3. **Esperanza del error $\epsilon_i$ es igual a cero:** Dado el valor de $x_i$, la media o el valor esperado del término del error $\epsilon_i$ es igual a cero. Esto es $E(\epsilon_i) = 0$.

4. **Homocedasticidad o varianza constante de $\epsilon_i$:** La varianza del término de error, es la misma sin importar el valor de $x$. Esto es $var(\epsilon_i) = \sigma^2$.

5. **No hay autocorrelación entre los errores:** Dados dos valores cualesquiera de $x$ y $x_i$ y $x_j$ $(i \neq j)$, la correlación entre dos $\epsilon_i$ y $\epsilon_j$ cualesquiera $(i \neq j)$ es cero. Es decir estas observaciones se muestrean de manera independiente. Esto es $cov(\epsilon_i, \epsilon_j) = 0$.

6. **El número de observaciones $n$ debe ser mayor que el número de coeficientes por estimar:** Sucesivamente, el número de observaciones $n$ debe ser mayor que el número de variables explicativas.

7. **La naturaleza de las variables $x$:** No todos los valores $x$ de una muestra determinada deben ser iguales. Técnicamente, $var(x)$ debe ser un número positivo. Además, no puede haber **valores atípicos** de la variable $x$, es decir, valores muy grandes en relación con el resto de las observaciones [8].

8. **La distribución de los errores:** Los errores $\epsilon_i$, tienen una distribución normal con media $0$ y varianza constante. Es decir,

$$\epsilon_i \sim N(0, \sigma^2).$$

El supuesto 8 es necesario puesto que los estimadores de mínimos cuadrados son los estimadores de regresión lineal que minimizan la suma de los cuadrados de los errores y además permiten la realización de inferencias estadísticas.

## 1.3. Método de Mínimos Cuadrados Ordinarios

El Método de Mínimos Cuadrados Ordinarios es un procedimiento que permite estimar los coeficientes del modelo de regresión lineal simple, se realiza la estimación de $\beta_0$ y $\beta_1$, con la finalidad de minimizar la suma de las diferencias de los cuadrados de las desviaciones, a partir de la recta ajustada bajo los supuestos establecidos en la sección 1.2.

El cuál se expresa como:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x. \tag{1.3}$$

Donde:

- $\widehat{y}$ es el valor pronosticado de la variable respuesta.
- $x$ es la variable independiente o explicativa.
- $\widehat{\beta}_0$ es el valor estimado de la ordenada al origen.
- $\widehat{\beta}_1$ es el valor estimado de la pendiente de la recta.

Si la ecuación (1.3) es el valor pronosticado del $i-$ésimo valor $y$, entonces la desviación (a veces llamado error) del valor observado de $y_i$ a partir de $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ es la diferencia $y_i - \widehat{y}_i$ [14].

Ahora vamos a minimizar a:

$$SCE = \sum_{i=1}^{n} \widehat{\epsilon}_i{}^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))^2. \tag{1.4}$$

Como se desea calcular el mínimo, entonces en cada uno de los coeficientes se debe tener que: $\frac{\partial\, SCE}{\partial\, \widehat{\beta}_0} = 0$ y $\frac{\partial\, SCE}{\partial\, \widehat{\beta}_1} = 0$, calculando las derivadas parciales e igualando a cero se tiene que:

$$\frac{\partial\, SCE}{\partial\, \widehat{\beta}_0} = -2\sum_{i=1}^{n}\left[y_i - (\widehat{\beta}_0 - \widehat{\beta}_1 x_i)\right] = 0$$

$$= \left(\sum_{i=1}^{n} y_i - n\widehat{\beta}_0 - \widehat{\beta}_1 \sum_{i=1}^{n} x_i\right) = 0 \tag{1.5}$$

$$\frac{\partial\, SCE}{\partial\, \widehat{\beta}_1} = -2\sum_{i=1}^{n}\left[y_i - (\widehat{\beta}_0 - \widehat{\beta}_1 x_i)\right] x_i$$

$$= \left(\sum_{i=1}^{n} x_i y_i - \widehat{\beta}_0 \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2\right) = 0 \tag{1.6}$$

Despejando $\sum_{i=1}^{n} y_i$ en (1.5) y $\sum_{i=1}^{n} x_i y_i$ en (1.6) se tiene que:

$$\sum_{i=1}^{n} y_i = n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i \tag{1.7}$$

$$\sum_{i=1}^{n} x_i y_i = \widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 \tag{1.8}$$

Resolviendo (1.7) y (1.8), despejando $\widehat{\beta}_0$ en la primera expresión de (1.7) se tiene:

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - \widehat{\beta}_1 \bar{x} \tag{1.9}$$

Sustituyendo (1.9) y despejando $\widehat{\beta}_1$ en la segunda expresión de (1.7) se tiene:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n})x_i^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.10}$$

Denotaremos como $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ y $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Por lo que $\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$.

Notemos que $S_{xy}$ es la covarianza y $S_{xx}$ es la varianza de los muestrales de $X$ y $Y$.

Finalmente las estimaciones por el Método de Mínimos Cuadrados Ordinarios están dados en (1.9) y (1.10).

## 1.4. Estimación de $\sigma^2$

De acuerdo con el modelo de regresión y sus suposiciones, se puede concluir que $\sigma^2$, la varianza de $\epsilon$, representa también la varianza de los valores de $y$ respecto a la recta de regresión. Recordemos que a las desviaciones de los valores de $y$ de la recta de regresión estimada se les conoce como residuales. Por lo tanto, la suma de $SCE$, (suma de los cuadrados de los residuales), es una medida de la variabilidad de las observaciones reales respecto a la línea de regresión estimada.

El **Cuadrado Medio Residual** (CMR) proporciona una estimación de $\sigma^2$, esta estimación es la Suma del Cuadrado de los Errores $SCE$ dividida entre sus grados de libertad, el cuál se expresó en la ecuación (1.4).

A cada suma de cuadrados le corresponde un número llamado grados de libertad. La $SCE$ tiene $n-2$ grados de libertad porque para calcular $SCE$ es necesario estimar dos coeficientes ($\beta_0$ y $\beta_1$) [3].

Por lo tanto, el cuadrado medio se calcula dividiendo $SCE$ entre $n-2$, el CMR proporciona un estimador insesgado de $\sigma^2$. Como el valor de CMR proporciona un estimado de $\sigma^2$, se emplea la notación $s^2$.

La demostración de este resultado se encuentra en [14].

En el Método de Mínimos Cuadrados estimamos $\beta_0$ y $\beta_1$, estimaremos a la $var(y_i) = \sigma^2$. Para estimar $\sigma^2$ usamos la definición de varianza $\sigma^2 = E[y_i - E(y_i)]^2$. Por el supuesto $4$ en 1.2, $\sigma^2$ es el mismo para cada $y_i$, $i = 1, 2, \ldots, n$. Usando $\widehat{y}$ como un estimador de $E(y_i)$, estimamos $\sigma^2$, por un promedio de la muestra, es decir,

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y})^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i^2)}{n-2} = \frac{SCE}{n-2}. \tag{1.11}$$

Donde $\widehat{\beta}_0$ y $\widehat{\beta}_1$ están dados por (1.9) y (1.10) y $SCE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$.

La desviación $\widehat{\epsilon}_i = y_i - \widehat{yi}$ en ocasiones se denomina **residual** de $y_i$ y $SCE$ se denomina *suma de cuadrados residual o suma de cuadrados de error.*

Con $n-2$ en el denominador, $s^2$ es un estimador insesgado de $\sigma^2$.

$$E(s^2) = \frac{E(SCE)}{n-2} = \frac{(n-2)\sigma^2}{n-2} = \sigma^2. \tag{1.12}$$

## 1.5. Modelo de Regresión Múltiple

En la mayoría de las aplicaciones en distintas áreas del conocimiento, se tiene la necesidad de involucrar distintas variables explicativas que inciden en el modelo de regresión, a dicho modelo se le conoce como Regresión Múltiple, el cual describimos a continuación.

Un modelo de regresión donde interviene más de una variable explicativa se llama modelo de regresión múltiple [12].

Los datos consisten en $n$ observaciones sobre una variable respuesta y $k$ variables explicativas, $x_1, x_2, \ldots, x_k$. La relación entre $y$ y $x_1, x_2, \ldots, x_k$ se establece como un modelo lineal.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \tag{1.13}$$

donde $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ son los coeficientes de regresión y $\epsilon$ es el error, aleatorio. Donde las observaciones están definidas como:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \beta_k x_{1k} + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \ldots + \beta_k x_{2k} + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \ldots + \beta_k x_{nk} + \epsilon_n \end{aligned}$$

Asumimos que para cualquier conjunto de valores fijos de $X_1, X_2, \ldots, X_k$ que caen dentro del rango de los datos, la ecuación lineal (1.13) proporciona una aproximación aceptable de la verdadera relación entre $Y$ y las $X's$ y $\epsilon$ mide la discrepancia en esa aproximación [5].

Para el análisis de regresión lineal múltiple, la notación matricial es una forma más adecuada al momento de analizar los datos y obtener los resultados, de esta manera la notación matricial de este modelo es

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{1.14}$$

donde

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}, \; \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \text{ y } \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{1.15}$$

De (1.14) se tiene que **Y** es la matriz de $n \times 1$ de las observaciones, **X** es la matriz de $n \times (k+1)$ de variables explicativas, $\beta$ es la matriz de $(n \times k)$ de los coeficientes de regresión y $\epsilon$ es la matriz de $n \times 1$ de errores aleatorios.

## 1.6. Supuestos del Modelo de Regresión Múltiple

Los supuestos del modelo que se establecieron en 1.2 son las mismas para este modelo, pues el modelo de Regresión Lineal Simple es un caso especial de este modelo, esto es.

1. **Linealidad:** El modelo de regresión es lineal en los coeficientes. Es decir, el modelo como se muestra en la ecuación (1.14).

2. **Valores de $X$ independientes del término del error:**. En este caso, esto significa que se requiere covarianza cero entre $\epsilon_j$ y cada variable $X$, esto se expresa como $Cov(X, \epsilon_j) = 0$

   Como se mencionó en 1.2 en general la covarianza entre un valor independiente $X$ y el termino del error $\epsilon_j$ no siempre es cero.

3. **Esperanza del error $\epsilon_i$ es igual a cero:** Dado el valor de $x_i$, la media o el valor esperado del término del error $\epsilon_i$ es igual a cero. Esto es $E(\epsilon_i) = 0$ o bien $E(Y) = X\beta$.

4. **Homocedasticidad o varianza constante de $\epsilon_i$:** La varianza del término de error, es la misma sin importar el valor de $x$. Esto es $var(\epsilon_i) = \sigma^2 \quad i = 1, \dots n.$

5. **No hay autocorrelación entre los errores:**: Dados dos valores cualesquiera de $x$ y $x_i$ y $x_j (i \neq j)$, la correlación entre dos $\epsilon_i$ y $\epsilon_j$ cualesquiera $(i \neq j)$ es cero. Es decir estas observaciones se muestrean de manera independiente. Esto es $cov(\epsilon_i, \epsilon_j) = 0.$

6. **El número de observaciones $n$ debe ser mayor que el número de coeficientes por estimar.** [8].

7. **La distribución de los errores:** Los errores $\epsilon_i$, tienen una distribución normal con media $0$ y varianza constante. Es decir.

$$\epsilon_i \sim N(0, \sigma^2)$$

Al igual que el modelo de regresión simple, el supuesto 7 es necesario, puesto que los estimadores de mínimos cuadrados son los estimadores de regresión lineal que minimizan la suma de los cuadrados de los errores y proporcionan mejores estimadores lineales insesgados.

## 1.7. Método de Mínimos Cuadrados Ordinarios

Al igual a como se realizó en el Modelo de Regresión Lineal Simple, estimaremos $\beta$, aquí suponemos que $x_1, x_2, \ldots, x_k$ son fijas.

La función de mínimos cuadrados es:

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \epsilon_i^2 \tag{1.16}$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=i}^{k} \beta_j x_{ij})^2. \tag{1.17}$$

Se debe minimizar la función $S$ respecto a $\beta_0, \beta_1, \ldots, \beta_k$. Los estimadores de $\beta_0, \beta_1, \ldots, \beta_k$ por mínimos cuadrados deben satisfacer

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_k}} = -2 \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \sum_{j=1}^{n} \widehat{\beta_j} x_{ij}) = 0, \tag{1.18}$$

$$\left.\frac{\partial S}{\partial \beta_j}\right|_{\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_k}} = -2 \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \sum_{j=1}^{n} \widehat{\beta_j} x_{ij} = 0), \quad j = 1, 2, \ldots, k. \tag{1.19}$$

Al simplificar la ecuación (1.18) se obtienen las **ecuaciones normales de mínimos cuadrados**.

$$n\widehat{\beta_0} + \widehat{\beta_1} \sum_{i=1}^{n} x_{i1} + \widehat{\beta_2} \sum_{i=1}^{n} x_{i2} + \ldots + \widehat{\beta_k} \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\widehat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \ldots + \widehat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i,$$

$$\vdots$$

$$\widehat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \ldots + \widehat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i. \tag{1.20}$$

Observemos que hay $p = k+1$ ecuaciones normales,una para cada uno de los coeficientes desconocidos de regresión.

La solución de las ecuaciones normales serán los **estimadores por mínimos cuadrados** $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$.

De manera práctica podemos utilizar la notación matricial para expresar los resultados anteriores, esto es.

Se desea determinar el vector $\widehat{\beta}$ que minimice

$$S(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = (y - X\beta)'(y - X\beta) \tag{1.21}$$

Observemos que $S(\beta)$ se puede expresar como sigue:

$$\begin{aligned} S(\beta) &= y'y - \beta'X'y - y'X\beta + \beta'X'X\beta \\ &= y'y - 2\beta'X'y + \beta'X'X\beta. \end{aligned}$$

como $\beta'X'y$ es una matriz de $1 \times 1$, es decir, un escalar, y que su transpuesta $(\beta'X'y)' = y'X\beta$ es el mismo escalar. Los estimadores de mínimos cuadrados deben satisfacer

$$\left.\frac{\partial S}{\partial \beta}\right|_{\widehat{\beta}} = -2X'y + 2X'X\widehat{\beta} = 0$$

el cual se simplifica a

$$X'X\widehat{\beta} = X'y \tag{1.22}$$

Las ecuaciones en (1.22) son las ecuaciones de mínimos cuadrados. Son la forma matricial de la presentación escalar de las ecuaciones dadas en (1.20).

Para resolver las ecuaciones normales se multiplican ambos lados de (1.22) por la inversa de $X'X$.

Así el **estimador de $\beta$ de mínimos cuadrados** es

$$\widehat{\beta} = (X'X)^{-1}X'y \tag{1.23}$$

siempre y cuando la matríz inversa $(X'X)^{-1}$ exista. La matriz, $(X'X)^{-1}$ existe siempre y cuando las variables explicativas sean **linealmente independientes**, esto es, sin ninguna columna de la matriz $X$ es una combinación lineal de las demás columnas [12].

## 1.8. Estimación de $\sigma^2$

. De igual manera como se realizó en el modelo de regresión simple, podemos calcular un estimador de $\sigma^2$ a partir de la suma de cuadrados de residuales

$$SS_{Res} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2 = \epsilon'\epsilon \tag{1.24}$$

Sustituyendo $\epsilon = y - X\widehat{\beta}$ se tiene que

$$\begin{aligned} SS_{Res} &= (y - X\widehat{\beta})'(y - X\widehat{\beta}) \\ &= y'y - \beta'X'y - y'X\beta + \beta'X'X\beta \\ &= y'y - 2\beta'X'y + \beta'X'X\beta. \end{aligned}$$

Como $X'X\widehat{\beta} = X'y$, la ultima igualdad se transforma en

$$SS_{Res} = y'y - \widehat{\beta}X'y. \tag{1.25}$$

Como se desea estimar $k+1$ coeficientes, se tiene que la suma de cuadrados de residuales tiene $n-(k+1)$ grados de libertad las cuáles están asociadas con el modelo establecido. De esta manera el cuadrado medio residual está dado como sigue

$$MS_{Res} = \frac{SS_{Res}}{n-(k+1)}. \tag{1.26}$$

La ecuación (1.26) se le conoce como el error cuadrado medio residual.

La demostración de esté resultado puede revisarse en [12], además se demuestra que el valor esperado de $MS_{Res}$ es $\sigma^2$, de está manera se tiene que un estimador insesgado para $\sigma^2$ es

$$\sigma^2 = MS_{Res} \tag{1.27}$$

# Capítulo 2

# Teoría de Residuales

En el presente capítulo estableceremos las herramientas necesarias que nos permitan determinar las características del análisis residual para la validación de los supuestos de los modelos de regresión lineal simple y múltiple establecidos en el capítulo 1.

El **análisis residual** es la herramienta principal para determinar si el modelo de regresión empleado es apropiado. En el capítulo 1 se estableció que el **residual** de la observación $i$ es la diferencia entre el valor observado de la variable respuesta y el valor estimado (1.4).

Los residuales nos brindan una mejor información acerca de $\epsilon$ pues nos permite validar el ajuste del modelo de regresión, además de poder saber como es el comportamiento de las observaciones, por lo que el análisis de los residuales es muy importante para determinar si las suposiciones hechas sobre $\epsilon$ son las adecuadas. La mayor parte del análisis residual que se realiza se basa en la examinación de gráficas, de esta manera nos permite verificar el cumplimiento de los supuestos.

Con la Teoría de Residuales y teniendo como base los supuestos establecidos para los modelos simple y múltiple en el capítulo anterior, nos permitirá determinar la validación de los mismos y la detección de datos que afectan o influyen en ella.

## 2.1. Análisis Residual

En el capítulo anterior se establecieron los modelos de regresión lineal y múltiple, bajo los siguientes supuestos.

1. **Supuesto de la forma del modelo:**, el modelo de regresión es lineal, la comprobación de la

linealidad en regresión simple, se realiza mediante la examinación de la gráfica de dispersión de $y$ contra $x$ y mediante la gráfica de dispersión se garantiza la linealidad.

2. **Supuesto acerca del error:** Los errores $\epsilon_1.\epsilon_2, \ldots, \epsilon_n$ definidos en la ecuación (1.2) se supone que son variables aleatorias normales, independientes e identicamente distribuidos (iid) cada una con media $0$ y varianza $\sigma^2$. De esto se tienen los siguientes cuatro supuestos.

   - El error $\epsilon_i, i = 1, 2, \ldots, n$ tiene distribución normal, nos referimos a esto como el supuesto de normalidad. La suposición de normalidad no es fácil de validar, especialmente cuando los valores no se replican. Esta validez puede evaluarse examinando las gráficas apropiadas de los residuales, del cuál se explicará más adelante.
   - Los errores $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ tienen media cero.
   - Los errores $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ tienen la misma varianza $\sigma^2$. Esta es la suposición de varianza constante, el cuál también es conocido como el supuesto de homogeneidad o homocedasticidad.
   - Los errores $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ son independientes entre sí, donde la covarianza entre pares es igual a cero. A esto se refiere como el supuesto de errores independientes.

## 2.2. Tipos de residuales

Uno de los métodos que se lleva a cabo para la verificación de los supuestos del modelo de regresión es el empleo de gráficas de residuales, pues es una de los formas que permite examinar y determinar cuál de los supuestos no se está cumpliendo y qué medidas deberán emplearse para el ajuste del modelo. El análisis de residuales puede conducir a sugerencias de estructura o señalar información en los datos que podría perderse o pasarse por alto si el análisis se basa solo en estadísticas resumidas.

La visualización de las gráficas de residuales también permite observar aquellos puntos que se alejan un poco más de la linea de regresión estimada, de esta manera podemos comprender un poco más el comportamiento del modelo que se desea analizar.

En el capítulo anterior hemos establecido el Método de Mínimos Cuadrados, del cuál hemos calculado los valores ajustados de los modelos establecidos

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{i1} + \ldots + \widehat{\beta_k} x_{ik} \quad i = 1, 2, \ldots, n \tag{2.1}$$

y los residuales de los mínimos cuadrados ordinarios correspondientes

$$y_i - \widehat{y_i} \quad i = 1, 2, \ldots, n. \tag{2.2}$$

Cuando se emplea el método de mínimos cuadrados, la media de los residuales es cero.

El **leverage** es una medida que determina la desviación de alguna observación $x_i$ de la media muestral de la variable independiente, es decir que tan lejos se encuentra algún punto de la media muestral. El cuál está definido como:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \tag{2.3}$$

El valor $h_i$ es llamado el **leverage** de la i-ésima observación, los valores del leverage son necesarios en el análisis de regresión, por lo que aparecen con cierta frecuencia.
Existen algunos tipos de residuales los cuales se describen a continuación.

- Residual estandarizado: para calcular el residual estandarizado se divide cada uno de los residuales entre la desviación estándar, esto es:

$$s_{y_i - \widehat{y_i}} = s\sqrt{1 - h_i} \tag{2.4}$$

Donde: $s =$ error estándar de estimación.

El error estándar de estimación para el modelo de regresión lineal simple está expresada en la ecuación (2.5).

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n - 2}} \tag{2.5}$$

En este caso se toman dos valores de libertad, esto se debe a que se restan dos grados de libertad, que corresponden a dos coeficientes estimados en el modelo de regresión lineal.

Los residuales tienen media cero y varianza aproximadamente uno, en consecuencia, un residual estanzarizado grande (por ejemplo $h_i$) indica que se trata de un valor atípico potencial [12].

- Matriz hat. Es el vector de valores ajustados $\widehat{Y}$

$$\widehat{Y} = X\widehat{\beta} = PY$$

donde:

$$\mathbf{P} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T} \tag{2.6}$$

y

$X$ es la matriz de variables explicativas [5].

La varianza del $i-$ésimo residual es

$$Var(e_i) = \sigma^2(1 - h_i) \tag{2.7}$$

donde $h_i$ es el i-ésimo valor del leverage. Para resolver el problema de las varianzas desiguales, se estándariza el i-ésimo residual $\epsilon_i$ dividiéndolo por su desviación estándar, esto es

$$z_i = \frac{\epsilon_i}{\sigma\sqrt{(1 - h_i)}} \tag{2.8}$$

## 2.3. Tipos de gráficas de residuales

Los métodos gráficos juegan un papel importante en el análisis de datos, de manera particular en el ajuste de modelos lineales a los datos.

El empleo de gráficas en el análisis de información, permite obtener conclusiones del conjunto de datos que se estudia, de esta manera se sintetiza lo que se pretende dar a conocer, a partir de un estudio cuidadoso.

La utilización de gráficas, resultan ser muy útiles y tienen diferentes objetivos, alguno de ellos son:

- Detectar errores en los datos (por ejemplo un punto atípico).
- Confirmar o negar suposiciones.
- Evaluar la idoneidad del modelo ajustado.

Después de ajustar un modelo de regresión lineal, las gráficas de residuales ayudan a verificar los supuestos del modelo y evaluar el comportamiento de las observaciones.

Estas gráficas pueden clasificarse en:

1. Gráficas para comprobar los supuestos de linealidad y normalidad.
2. Gráficas para la detección de valores atípicos y observaciones influyentes.
3. Gráficas de diagnóstico para el efecto de variables.

### 2.3.1. Gráfica de residuales contra la variable $x$

La gráfica de residuales contra la variable explivativa $x$ es una gráfica en donde los valores de la variable explicativa se representan en el eje horizontal y los valores de los residuales correspondientes se representan en el eje vertical [3].

Por cada residual que se tenga se gráfica un punto, la pareja está dada por $(x_i, y_i - \widehat{y_i})$, donde la primera coordenada esta dada por el valor $x_i$ y la segunda coordenada esta dada por $y_i - \widehat{y_i}$, que es el valor del residual calculado.

Previamente a la interpretación de los resultados de la gráfica de residuales, es necesario tomar en cuenta algunas formas que pueden llegar a tener al momento de implementarlos en algún software estadístico.

- Si la gráfica de residuales se parece a la gráfica A de la Figura 2.1, entonces se tendrá un conjunto de puntos, que pueden ser encerrados en una banda horizontal, lo cual significa que el modelo está representado de manera adecuada por las variables.

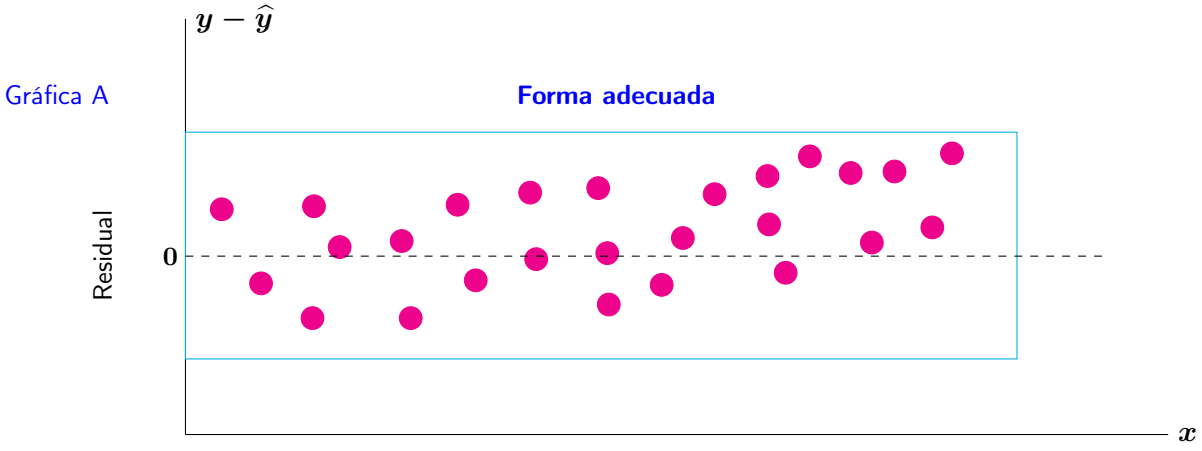


Figura 2.1: Forma general de los residuales.

- Si la gráfica de residuales se parece a la gráfica B de la Figura, 2.2 significa que la varianza del error $\epsilon$ es distinta para todos los valores de $x$ entonces se tiene que la varianza del error no es constante y por lo tanto no cumple el supuesto de varianza constante.
  En este caso se observa mucha variabilidad de los residuales.

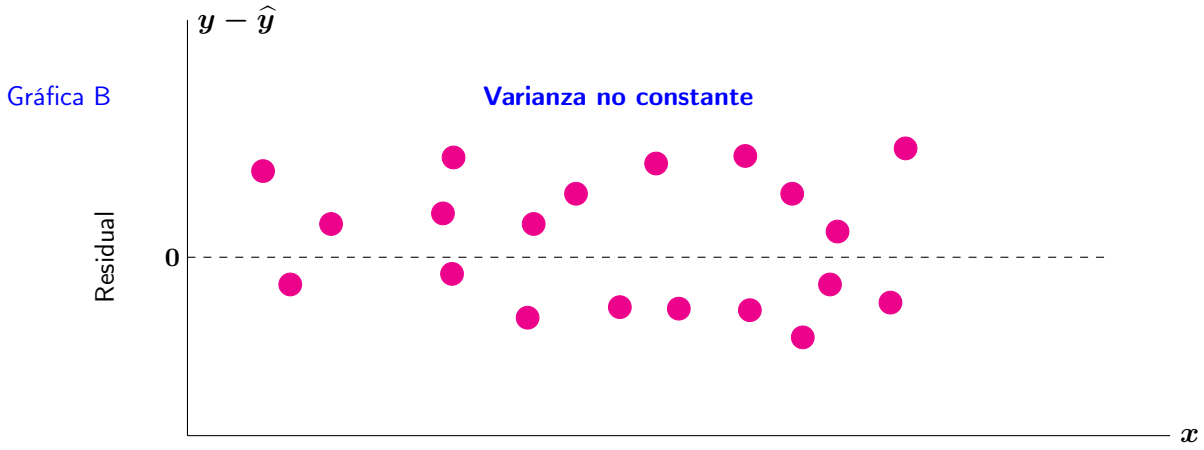


Figura 2.2: Forma general de los residuales.

- Para la gráfica C de la Figura 2.3 se observa la distribución de los residuales en forma de curva, por lo tanto no cumple con el supuesto de linealidad y se puede concluir que el modelo de regresión no representa de manera adecuada la relación entre las variables, por lo que se recomienda emplear otro tipo de modelo de regresión o aplicar una transformación apropiada.

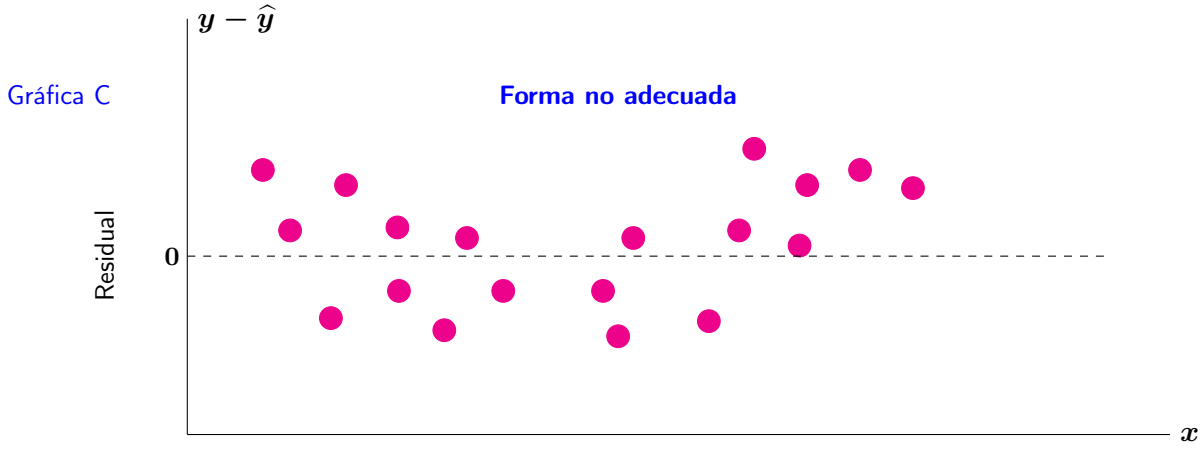


Figura 2.3: Forma general de los residuales.

### 2.3.2. Gráfica de residuales contra $\widehat{y}$

La gráfica de residuales contra los valores ajustados $\widehat{y}$, es una de las gráficas comunmente utilizadas en los modelos de analisis de regresión simple. La curvatura podría indicar que la media ajustada no es la adecuada. Los residuales que parecen aumentar o disminuir en promedio con los valores ajustados puede indicar que la varianza residual no es constante. Si se tiene algunos residuales grandes pueden indicarnos valores atípicos, en algunos casos puede indicarnos que el modelo no es consistente.

En la gráfica de residuales los valores estimados de la variable respuesta $\widehat{y}$ son representados en el eje horizontal y el valor calculado de los residuales se representan en el eje vertical, por lo que cada pareja ordenada representa un punto en la gráfica.

Cada componente de la pareja ordenada, está dada por $(\widehat{y}, y_i - \widehat{y})$. Es importante mencionar que para el modelo de regresión simple, tanto la forma de las gráficas de residuales contra la variable $x$ y contra la variable $\widehat{y}$ muestran la misma información.

Para el análisis de regresión múltiple, la gráfica de residuales contra la variable $\widehat{y}$ es utilizada con mayor frecuencia debido a que se tiene más de una variable explicativa.

### 2.3.3. Gráfica de probabilidad normal

Un método para comprobar el supuesto de normalidad de los residuales, es la **gráfica de probabilidad normal**, para eso se observa el comportamiento de los datos con respecto a una línea recta, si la mayoría de los datos se distribuyen sobre ella, se concluye que los residuales tienen una distribución normal.

Sean $\epsilon_1 < \epsilon_2 < \ldots < \epsilon_n$ los residuales ordenados en orden creciente. Si se grafican $\epsilon_i$ en función de la probabilidad acumulada, es decir

$$P_i = \frac{i - 0.5}{n} \qquad i = 1, \ldots, n \tag{2.9}$$

Donde:

$i =$ es el número de orden de cada observación.

$n =$ es el número total de observaciones.

Los valores que resulten al emplear la ecuación (2.9) deberán estar ubicados sobre una línea recta.

Al momento de obtener la gráfica de probabilidad normal y se observa que estás no caen sobre la línea recta, se necesitaría de mayor experiencia para poder realizar la interpretación adecuada.

Cuando se calcula la desviación estándar de cada uno de los residuales, se calculan los residuales estandarizados dividiendo cada residual entre sus respectivas desviaciones estándar.

Esto es:

$$\frac{y_i - \widehat{y_i}}{s_{y_i - \widehat{y_i}}} \tag{2.10}$$

A la ecuación (2.10) se le conoce **residual estandarizado de la observación $i$.**

La gráfica de los residuales estándarizados también permite visualizar si el termino de error $\epsilon$ tiene distribución normal a través de la validación de los supuestos del modelo de regresión lineal.

## 2.4. Observaciones atípicas y observaciones influyentes

En secciones anteriores de este capítulo, se establecieron herramientas necesarias para realizar el análisis de residuales y de esta manera verificar el cumplimiento de los supuestos del modelo de regresión lineal.

En esta sección se utilizará el análisis de los residuales para la identificación de observaciones atípicas o influyentes que se presentan en la gráfica de residuales.

En ocasiones cuando se obtienen las gráficas de los residuales se observan puntos que se alejan demasiado de la linea de regresión, por lo que su análisis es necesario para determinar el impacto que tiene con respecto al resto de las observaciones y en especial al modelo de regresion de interés.

Una observación es **influyente**, si al omitirla del conjunto de observaciones provoca cambios en el ajuste del modelo estimado.

Por otro lado una observación es **atípica**, si esta no sigue el comportamiento del resto de las observaciones, es decir se aleja de manera drástica del resto de las observaciones.

### 2.4.1. Detección de observaciones atípicas

Como se observa en el ejemplo de la Figura 2.4, la observación atípica se aleja del resto de las observaciones, por lo que resulta necesario realizar un análisis cuidadoso al revisar estás observaciones.

Algunos razones por las cuáles pudiera generarse una observación atípica son:

- Datos erróneos en la entrada de los datos.
- El no cumplimiento de alguno de los supuestos del modelo de regresión.
- Datos inusuales que se presentan en las observaciones.

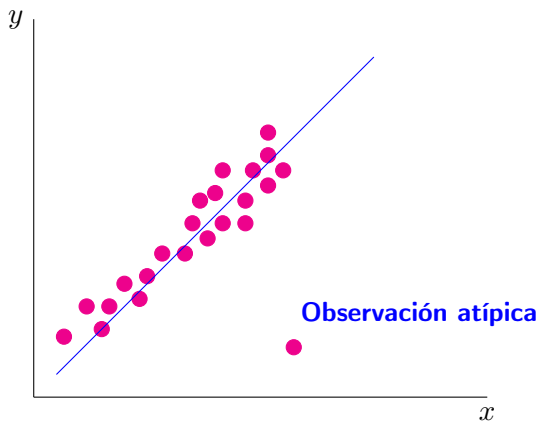


Figura 2.4: Datos con observación atípica.

Si se trata de un valor erróneo al momento de ingresar los datos en algún software estadístico, simplemente se detecta la observación, se corrige y se vuelve ajustar el modelo.
Si fuera el caso donde no se cumpliera alguno de los supuestos del modelo, se necesitaría emplear otras técnicas estadísticas que permitan explicar el modelo de la mejor forma.

En caso de obtener observaciones inusuales, lo más conveniente es conservar el valor para el análisis del comportamiento que tenga con respecto al resto de las observaciones. Los residuales cuyo valor absoluto es bastante mayor que los demás digamos de $3$ a $4$ desviaciones estándar respecto a la media, indica que hay valores atípicos potenciales [12].

La identificación adecuada de las observaciones atípicas permite realizar adecuaciones apropiadas hacia el modelo de regresión, por lo cuál su análisis se vuelve fundamental, previo a la obtención de conclusiones.

### 2.4.2. Detección de observaciones influyentes

En algunas ocasiones se pueden presentar casos en donde más de una observación presenta una gran influencia sobre los resultados, por lo que al omitir estás observaciones en la gráfica de residuales, podría generar cambios drásticos en el comportamiento de la pendiente de la recta de regresión ajustada.

Cuando se tiene una sola variable respuesta, una observación influyente puede ser identificada a través de un diagrama de dispersión, en ocasiones una observación influyente puede ser también una observación atípica, es decir un valor que se desvía del resto de las observaciones o incluso puede ocurrir que se trate de ambos casos.

De manera similar que las observaciones atípicas, las observaciones influyentes deben examinarse de manera cuidadosa, pues como se mencionó anteriormente son puntos que influyen en el comportamiento de la línea de regresión estimada.

Por esta razón es necesario revisar desde la entrada de los datos, si alguna de ellas se ingresó de manera errónea esta debe corregirse y realizar nuevamente el proceso del análisis de regresión para obtener una nueva ecuación de regresión estimada, por otro lado si no se cometió ningún error, esta observación nos ayudará a comprender mejor el modelo de regresión.

Se llaman **puntos de gran influencia** a las observaciones en donde la variable respuesta presenta valores extremos, esta influencia dependerá de la distancia en la que se encuentre el valor de la variable respuesta de su media.

A la influencia también se le conoce como ***leverage*** de la observación $i$ y se calcula utilizando la ecuación (2.3).

De la ecuación (2.3) entre más alejada se encuentre $x_i$ de su media $\bar{x}$ mayor será la influencia (leverage) de la observación $i$.

Las observaciones influyentes debidas a la interacción de una observación de gran influencia y de residuales grandes, suelen ser difíciles de detectar [3].

## 2.5. Análisis residual para el análisis de regresión múltiple

En la sección anterior se establecieron las herramientas que permiten analizar los residuales estandarizados, en este caso la utilización de graficas de residuales y la identificación de observaciones atípicas para su análisis, resultan una manera adecuada de verificar los supuestos del modelo establecido.

En las secciones 2.2 y 2.3.3 se establecieron las ecuaciones (2.4) y (2.10), el cuál son la desviación estándar del residual $i$ y el residual estandarizado de la observación $i$ respectivamente y también se definió a $h_i$ como la influencia de la observación $i$ en la ecuación (2.3).

Para el modelo de regresión lineal múltiple dichas ecuaciones continuán siendo validas, debido a la dificultad de calcular $s_{y_i-\widehat{y_i}}$ y $h_i$ de forma manual, se emplean softwares estadísticos que permitan obtener los resultados de manera precisa.

De igual forma si se desea saber si la distribución de $\epsilon$ es normal, entonces debe utilizarse la gráfica de probabilidad normal, el cuál se estableció en (2.9), la cual resulta ser una gráfica adecuada para el modelo de regresión múltiple

## 2.6. Detección de observaciones atípicas

En la sección 2.4 se mencionó que las observaciones atípicas son aquellas que presentan comportamientos inusuales que el resto de los datos, para el modelo de regresión lineal múltiple de acuerdo con las observaciones obtenidas pueden existir situaciones en donde se presenten más de una observación atípica para la cuál deberá realizarse un análisis que permita revisar la influencia que tienen con el resto de los datos.

Cuando se presentan más de una observación atípica en un conjunto de datos el error estándar de estimación se incrementa y por lo tanto la desviación estandar del residual $i$ también se incrementa. Como el error estándar de estimación es el denominador en la ecuación 2.10 del residual estandarizado, el tamaño del residual estandarizado disminuirá a medida que el error estándar de estimador aumente.

Esto da como resultado que aún cuando un residual sea inusualmente grande, el denominador de la expresión (2.10), que será grande, hará que la regla del residual estandarizado falle para la identificación de una observación como observación atípica [3].

Una forma de resolver esta dificultad es utilizar los **residuales estudentizados**, el cual se describirá a continuación.

## 2.7. Residuales estudentizados

Un **residual estudentizado** es un residuo dividido por su desviación estándar estimada. Para estudentizar un residual se debe determinar el valor dado por la expresión.

$$\frac{\text{residual} - \text{media residual}}{\text{desviación estándar residual}}$$

La desviación estándar residual está dada por la ecuación 2.4.

En [12] los residuales estudentizados están dados por:

$$r_i = \frac{\epsilon_i}{\sqrt{MS_{Res}(1-h_i)}}, \quad i = 1, \ldots, n. \tag{2.11}$$

Donde:

$$MS_{Res} = \frac{\sum_{i=1}^{n} \epsilon_i^2}{n-p}$$

$$n - p = \text{grados de libertad}$$

$$\epsilon_i = y_i - \widehat{y}_i \quad i = 1, \ldots, n.$$

De esta forma tanto los residuales estándarizados como los residuales estudentizados brindan información valiosa, en la detección de observaciones atípicas e influyentes.

Sin embargo, cuando se presenta cualquier punto con un residual muy grande y a su vez con un leverage muy grande sobre el ajuste por mínimos cuadrados ordinarios, se recomienda revisar los residuales estudentizados.

Los residuales estudentizados pueden detectar observaciones atípicas que los residuales estandarizados no, para ello existen diversos software estadísticos que brindan una manera de obtener los residuales estudentizados, tales como R y Python por mencionar sólo algunos.

## 2.8. Observaciones influyentes

En la Sección 2.4 se establecieron ecuaciones que permiten detectar observaciones influyentes y de esta manera también se determinó la importancia de un análisis cuidadoso de la recta de regresión estimada.

Así mismo, para el modelo de regresión múltiple, también se emplea la ecuación del leverage para el análisis de las observaciones influyentes con respecto al resto de los datos.

Debido a la implicación de varias variables en el modelo de regresión múltiple se emplea software estadístico para la examinación del modelo y sus observaciones influyentes.

# Capítulo 3

# Metodología Estadística y Resultados

## 3.1. Aspectos Generales

Para esta Tesis se llevó a cabo la recopilación de datos en el mes de julio de 2022, en una comunidad rural e indígena denominada Tecuahuta, adscrita al municipio de Cuetzalan del Progreso en el estado de Puebla, se aplicó un cuestionario a tutores de menores de 0 a 5 años, que asisten a un servicio polivalente: Primera Infancia y Preescolar, del cual se registraron 40 observaciones.

**Criterios de inclusión:** para la recopilación de datos, se tomaron en cuenta tanto los alumnos que asistían a clases de nivel Preescolar y a menores de edad que recibían atención a la Primera Infancia, donde previamente se les solicitó en una sesión con padres de familia, los datos recientes de peso, estatura y edad, posteriormente al entregar la información requerida se les preguntaba sobre el número de vasos de leche que consumían al día.

**Criterios de exclusión:** de las 40 observaciones que se obtuvieron se eliminaron 9, ya que para el rango de edad 0-11 meses presentaron variaciones en la alimentación del infante, es decir, existieron observaciones donde los menores de edad eran alimentados con leche materna y algunos, con fórmula especial, debido al rechazo de la lactancia materna.

De las 40 observaciones registradas, se incluyeron 31 observaciones con rango de edad de 1-5 años, que de acuerdo con las características de la investigación, permitieron llevar a cabo el estudio.

Para el análisis estadístico se utilizó el Software Estadístico R.

## 3.2. Diseño Estadístico

Para el análisis de las observaciones registradas, se consideraron las siguientes variables: la variable peso se asignó como variable respuesta y las variables edad, estatura y número de vasos de leche fueron asignadas como variables explicativas.

Las variables de estudio se presentan en la siguiente tabla.

Cuadro 3.1: Variables de estudio.

| Variable | Descripción | Valores |
|---|---|---|
| edad | Edad del infante. | $\{1, 2, 3, 4, 5\}$ |
| estatura | Estatura del infante. | $(71, 107)$ |
| peso | Peso del infante. | $(8, 20)$ |
| n_vasos | Número de vasos de leche que consume al día. | $\{1.5, 2, 3.4\}$ |

Los rangos de los valores para la estatura y el peso fueron considerados a partir del menor rango de estatura y peso respectivamente, que se presentaron en la muestra.

## 3.3. Análisis Estadístico

### 3.3.1. Análisis Preliminar

Consistió en la elaboración de gráficas de pastel, histogramas de frecuencias y gráficas de dispersión, con el propósito de visualizar el comportamiento de cada una de las variables de manera univariada y posteriormente de manera conjunta contra la variable peso.

### 3.3.2. Análisis Definitivo

Se realizó un análisis de regresión lineal múltiple, teniendo como variable respuesta al peso, con el objetivo de ver una posible relación lineal entre el peso y las variables explicativas establecidas en el cuadro 3.1. A continuación se presentan los resultados que se obtuvieron del análisis preliminar y del análisis definitivo.

## 3.4. Resultados del análisis preliminar

En la siguiente figura se presenta el peso de los infantes registrados.

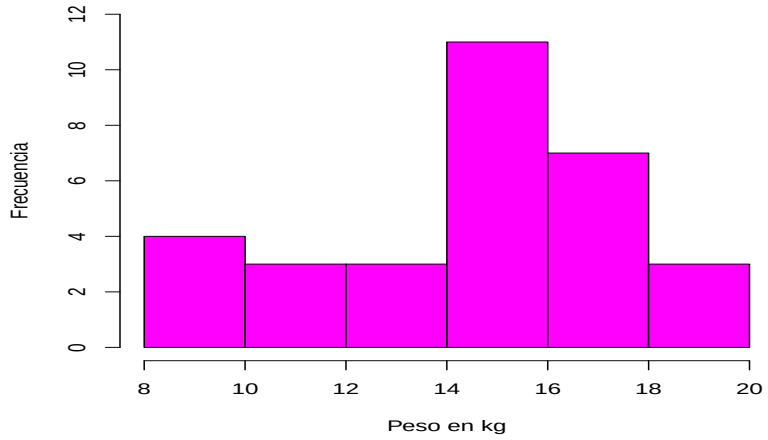


Figura 3.1: Histograma de frecuencias del peso.

De la Figura 3.1, se observa que el peso que se presenta con mayor frecuencia es entre los $14$ y $16$ kilos, hay niños con un peso entre $8$ y $10$ kilos y $3$ niños cuyo peso es entre los $18$ y $20$ kilos. De acuerdo con la distribución de los datos podemos notar una asimetría hacia la derecha.

En la siguiente Figura se presenta la estatura de los infantes participantes.

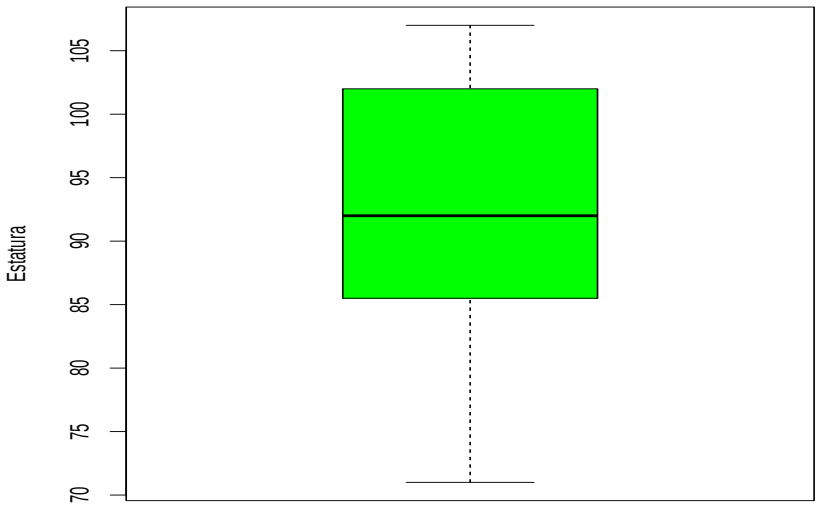


Figura 3.2: Gráfica de cajas y alambres de la estatura.

En la Figura 3.2, se observa que la estatura mínima del infante es de $71cm$, mientras que la estatura máxima que se obtuvo del infante es mayor a $105cm$.

La edad de los infantes, se registró en la siguiente Figura.

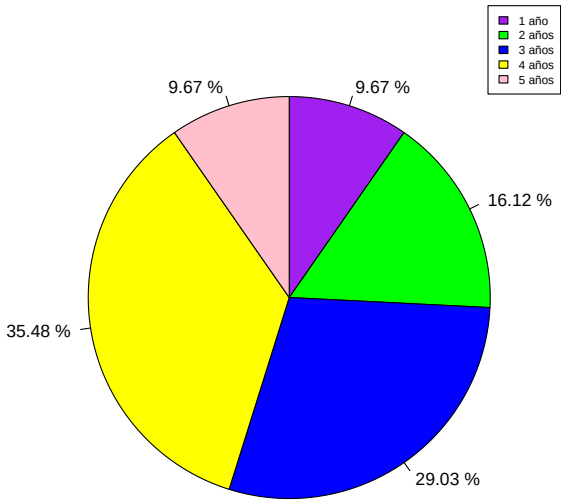


Figura 3.3: Gráfica de edad de los infantes.

En la Figura 3.3 se presenta la distribución por edades de los infantes que recibieron atención en el servicio polivalente, el porcentaje con mayor participación que se presentó en el espacio educativo fue alrededor del $35.48\,\%$ con infantes de $3$ años, posteriormente continua la edad de $4$ años donde se presentó una participación alrededor del $29.03\,\%$.

Por otro lado, la menor participación que se presentó fue aproximadamente del $9.67\,\%$ para las edades de $1$ año y $5$ años respectivamente, mientras que para la edad de 5 años, solo se tenían inscritos a $3$ infantes.

En la siguiente Figura se presenta la distribución de número de vasos de leche que consumen al día los infantes.

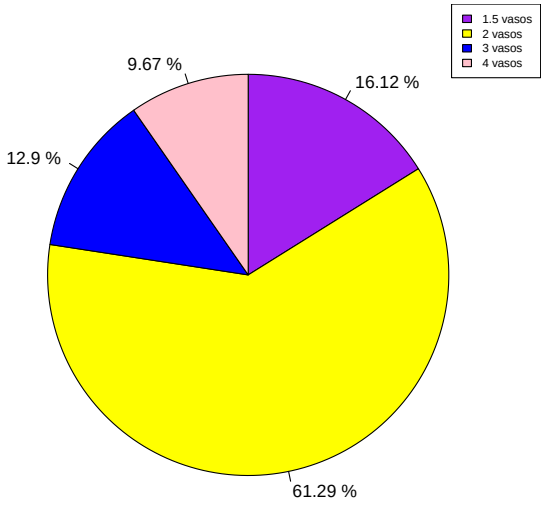


Figura 3.4: Número de vasos de leche que consumen al día.

En la Figura 3.4 se muestra el número de vasos de leche que consumen los infantes, de acuerdo con la información que se registró, puede observarse que el $61.29\,\%$ consumen $2$ vasos de leche al día, mientras que solo el $9.67\,\%$ solo consume $4$ vasos.

Ahora presentaremos las gráficas de dispersión de la variable peso con respecto a cada una de las variables explicativas (estatura, edad y número de vasos de leche), con la finalidad de visualizar la posible relación lineal que existe con cada una de las variables.

En la siguiente Figura se muestra la dispersión entre el peso y la estatura.

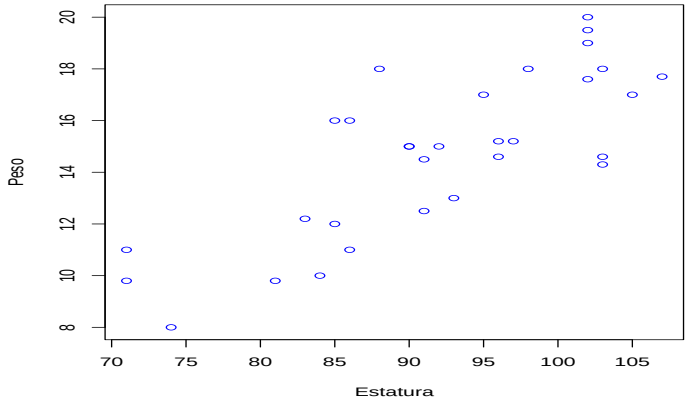


Figura 3.5: Gráfica de dispersión entre el peso y la estatura.

De acuerdo con la Figura se observa visualmente que entre ambas variables podría existir una relación lineal significativa ya que al incrementarse la estatura del infante se observa un incremento en el peso de esté.

Ahora en la siguiente Figura 3.6 se muestra la dispersión entre el peso y la edad.

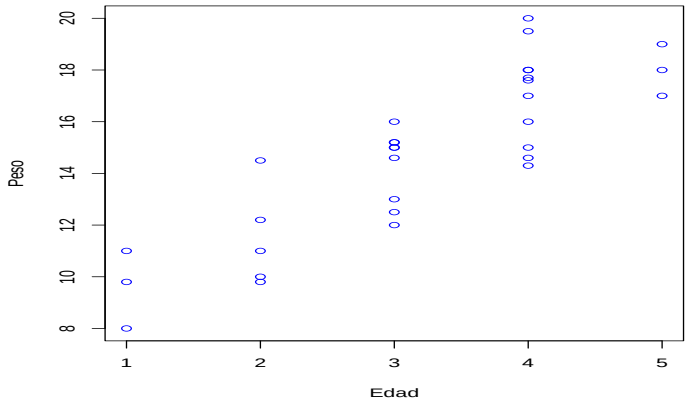


Figura 3.6: Gráfica de dispersión entre el peso y la edad.

De la Figura 3.6 se tiene que entre la variable peso y la variable edad pudiera existir una posible relación, ya que al incrementarse la edad aumenta el peso.

En la siguiente Figura que se muestra a continuación, se presenta la variable peso con respecto al número de vasos de leche que consumen los infantes del servicio polivalente.

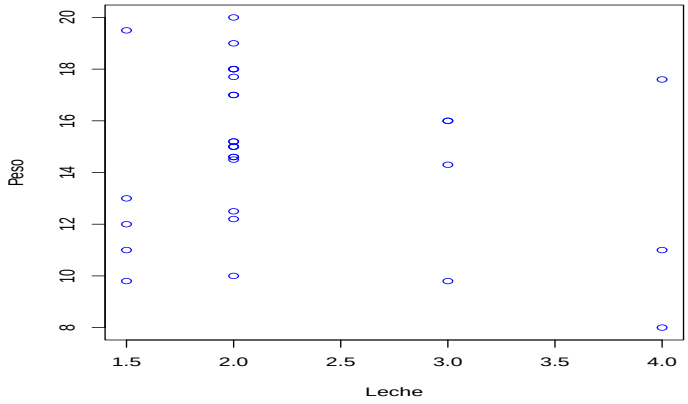


Figura 3.7: Gráfica de dispersión entre el peso y la leche que consumen al día.

Con base a la Figura 3.7 visualmente se observa que entre el peso y el número de vasos de leche que consumen al día, pudiera no existir una relación entre ambas variables.

## 3.5. Resultados del análisis definitivo

A continuación, se presentan los resultados del análisis definitivo, en primer lugar se ajustó el Modelo de Regresión Lineal Múltiple con el *peso* como variable respuesta y como variables explicativas la *edad, estatura y n_vasos*.

En el Cuadro 3.2 se muestran las estimaciones de los coeficientes del modelo de regresión.

Cuadro 3.2: Estimaciones de los coeficientes del modelo de regresión.

| Coeficiente | Estimación | p-valor |
|---|---|---|
| $\beta_0$ | 1.1028 | 0.8134 |
| $\beta_{edad}$ | 1.8745 | $< 0.05$ |
| $\beta_{estatura}$ | 0.0753 | 0.2185 |
| $\beta_{n\_vasos}$ | 0.3142 | 0.4919 |

En el Cuadro 3.2 se tienen los p-valor calculados, se observa que algunos son mayores a $0.05$. A partir de la información del cuadro anterior, se eliminará el coeficiente que tiene un p-valor mayor a $0.05$, esto con la finalidad de determinar la variable significativa con respecto a la variable respuesta peso.

Por lo que ahora se eliminará la variable con mayor p-valor, en este caso la variable *n_vasos*. A continuación se ajusta el modelo con la variable peso como variable respuesta y como variables explicativas las variables edad y estatura, en el Cuadro 3.3 se muestran las estimaciones de los coeficientes del modelo de regresión ajustado.

En el Cuadro 3.3 se presentan el p-valor obtenido, se observa que la variable estatura es mayor a $0.05$. Ahora se eliminará la variable con mayor p-valor, en este caso se eliminará la variable *estatura*.

Cuadro 3.3: Estimaciones de los coeficientes del modelo de regresión ajustado

| Coeficiente | Estimación | p-valor |
|---|---|---|
| $\beta_0$ | 2.6030 | 0.5263 |
| $\beta_{edad}$ | 1.8745 | $< 0.05$ |
| $\beta_{estatura}$ | 0.0663 | 0.2612 |

A continuación se ajusta el modelo con la variable peso como variable respuesta y como variable explicativa la variable edad, en el Cuadro 3.4 se muestran las estimaciones de los coeficientes del modelo de regresión.

Cuadro 3.4: Estimaciones de los coeficientes del modelo de regresión ajustado

| Coeficiente | Estimación | p-valor |
|---|---|---|
| $\beta_0$ | 7.1410 | $< 0.05$ |
| $\beta_{edad}$ | 2.38 | $< 0.05$ |

De los ajustes del modelo, la edad es el coeficiente que tiene una relación con la variable peso.

Por lo cuál el modelo de regresión ajustado esta dado como a continuación se establece:

$$\widehat{P}eso = 7.1410 + 2.38\beta_{edad} \tag{3.1}$$

Donde: $\widehat{P}eso$ es el peso ajustado.

Del ajuste del modelo se tiene que al incrementarse la edad del infante en un año, el peso del infante incrementa aproximadamente en $2.38\ kg$.

## 3.6. Análisis de los residuales

A continuación se presenta el análisis de los residuales, para la verificación de los supuestos del modelo de regresión lineal establecido en la ecuación (3.1).

En la siguiente Figura se muestran los residuales contra los valores ajustados del modelo de regresión lineal.

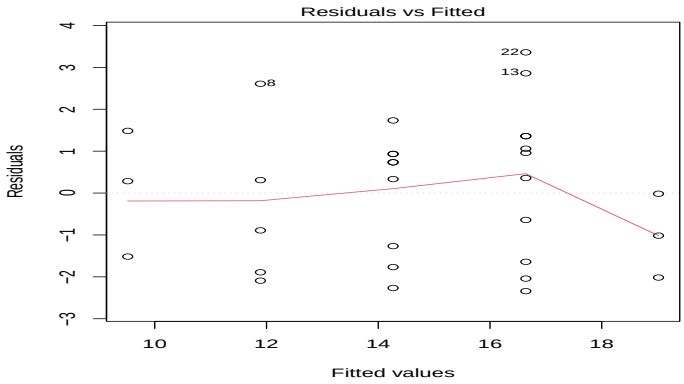


Figura 3.8: Gráfica de residuales vs Valores ajustados.

De la Figura 3.8 se observa que la linealidad suele mantenerse de manera adecuada, ya que la linea roja está cerca de la línea discontinua en gris, la dispersión de los residuales se mantienen de manera casi simétrica entre ambos ejes de los residuales. Los puntos $8, 13$ y $22$ pueden ser valores atipicos.

En la siguiente figura 3.9 se presenta la gráfica de probabilidad normal del modelo de regresión lineal, de esta manera se determinará cómo es la distribución de los residuales del modelo de regresión.

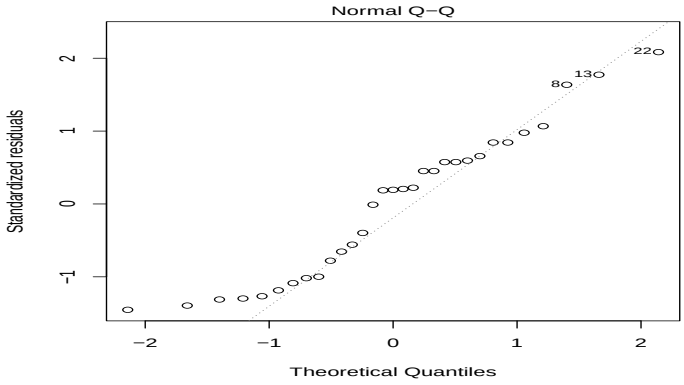


Figura 3.9: Gráfica de probabilidad normal.

## 3.7. Pruebas de hipótesis

A continuación se establecerán algunas pruebas de hipótesis con la finalidad de evaluar los supuestos del modelo de regresión. Evaluaremos el supuesto de homoscedasticidad, con nivel de significancia

$\alpha = 0.05$, para eso se plantea la siguiente hipótesis.

$$H_0 : p > \alpha \quad \text{(La varianza es constante)}$$
$$H_a : p < \alpha \quad \text{(La varianza no es constante)}$$

```
#studentized Breusch-Pagan test
data:  modelo
BP = 0.24898, df = 1, p-value = 0.6178
```

En este caso se tiene que el p-valor $= 0.6178 > 0.05$, entonces se rechaza la hipótesis alternativa en favor de la hipótesis nula, por lo que se concluye que los errores mantienen varianza constante. Se evaluará el supuesto de normalidad, con nivel de significancia $\alpha = 0.05$, se plantea la siguiente hipótesis.

$$H_0 : p > \alpha \quad \text{(Los errores tienen distribución normal)}$$
$$H_a : p < \alpha \quad \text{(Los errores no tienen distribución normal)}$$

```
#Shapiro-Wilk normality test
data:  resid(modelo)
W = 0.94299, p-value = 0.09984
```

De acuerdo con el test de normalidad de Shapiro-Wilks se tiene que el p-valor $= 0.09984 > 0.05$, por lo que se rechaza la hipótesis alternativa en favor de la hipótesis nula y se concluye que los errores tienen distribución normal. Ahora evaluaremos el supuesto de independencias de los errores, por lo que se plantea la siguiente hipótesis nula.

$$H_0 : p = 0 \quad \text{(No existe autocorrelación)}$$
$$H_a : p > 0 \quad \text{(Existe autocorrelación)}$$

```
#Durbin-Watson test
data:  modelo
DW = 1.751, p-value = 0.21
alternative hypothesis: true autocorrelation is greater than 0
```

Como $p = 0.21 > 0.05$ se rechaza la hipótesis alternativa en favor de la hipótesis nula y se concluye que no existe autocorrelación.

Con base a los gráficos de los residuales y las pruebas de hipótesis, se concluye que el modelo (3.1) satisface los supuestos del modelo. El análisis preliminar brindó información que apoyó sustancialmente a la realización del análisis definitivo, pues se revisó el comportamiento de las observaciones de manera univariada, posteriormente a través de las gráficas de dispersión permitió visualizar la posible relación entre la variable respuesta y entre cada una de las variables explicativas, en está parte se logró visualizar que la variable explicativa de número de vasos de leche no mantenía una relación lineal.

Al tener un análisis previo de la posible relación de la variable respuesta y la variable explicativa, con el análisis definitivo se estimaron los coeficientes del modelo y confirmar cuál de las variables explicativas eran significativas y de esta manera se estableció el modelo de regresión ajustado y la verificación de los supuestos.

Al estimar los coeficientes y determinar el modelo de regresión ajustado nos permitirá pronosticar el comportamiento del peso con base a la edad, el peso al ser pronósticado permitirá brindar recomendaciones para mantener condiciones de salud favorables en los infantes.

# Capítulo 4

# Conclusiones

La alimentación del infante es un proceso continuo, el cual comienza desde la gestación hasta la tercera edad, el primer alimento que consume al nacer es la leche materna o alguna leche de fórmula y a medida que va desarrollándose se complementa con alimentos que el sistema digestivo puede procesar sin generarle algún tipo de malestar. Sin embargo, conforme el menor va creciendo se enfrenta a diversos factores que impiden un sano desarrollo, donde el peso y la estatura se ven frenados por estos factores.

A partir de esta situación las instituciones de salud realizan un seguimiento periódico con el propósito de brindar recomendaciones que ayuden a mejorar el crecimiento del menor, de esta manera logran prevenir enfermedades o padecimientos que puedan desarrollarse a corto o largo plazo.

Es por eso que en el servicio polivalente se brinda atención desde la primera infancia con la finalidad de reflexionar acerca de las prácticas de crianza entre las madres, padres o cuidadores y la transición del infante al nivel preescolar.

Está investigación tuvo como objetivo general realizar un análisis estadístico a partir del establececi-miento de una posible relación lineal entre el peso como variable respuesta y la edad, estatura y número de vasos de leche que consumían al día, como variables explicativas, a través de un modelo de regresión lineal.

De esta manera los resultados que se obtuvieron fue que la edad influye en el peso del menor, es decir, conforme el infante va cumpliendo cierta edad, el peso también aumenta aproximadamente $2.38\ kg$, mientras que la estatura y el número de vasos de leche no influyen en el peso.

Sin embargo, existe la posibilidad de que en determinado tiempo el peso se vea frenado por diversos factores tales como: los cambios de hábitos alimenticios, enfermedades o padecimientos que puede presentar el infante.

Por ello es necesario que se le brinden recomendaciones desde la primera infancia a fin de hacer reflexionar a padres, madres y cuidadores en la importancia de llevar a cabo prácticas de alimentación saludable.

Un estudio que podría realizarse a futuro es analizar la influencia del ingreso económico en la alimentación en zonas rurales e indígenas, donde la conformación de los integrantes de la familia son en promedio de $7$ a $12$ integrantes aproximadamente y la ocupación laboral de los jefes del hogar son mayormente dedicados al campo y a la construcción.

# Apéndice

A continuación se presentan algunos tests o pruebas para la verificación de los supuestos del modelo de regresión, para más detalles puede consultar [8] y [4].

## Test de Breusch-Pagan.

El test de Breuch-Pagan ajusta un modelo de regresión lineal a los residuos del modelo ajustado, con las variables explicativas adicionales sospechosas de inducir varianza no constante, y rechaza si una buena parte de la varianza es explicada por dichas variables. El estadístico de contraste de Breusch-Pagan (bajo homocedasticidad) sigue una distribución chicuadrado con tantos grados de libertad como variables explicativas introducidas para justificar la falta de varianza constante [4].

## Test de Durbin-Watson.

La prueba más conocida para detectar correlación serial es la de los estadísticos Durbin y Watson. Se le conoce como estadístico $d$ de Durbin-Watson, que se define como

$$d = \frac{\sum_{t=2}^{t=n} (\widehat{\epsilon}_t - \widehat{\epsilon}_{t-1})^2}{\sum_{t=1}^{t=n} \widehat{\epsilon}_t^2}$$

que es simplemente la razón de la suma de las diferencias al cuadrado de residuos sucesivos sobre la SCR. Observe que, en el numerador del estadístico d, el número de observaciones es $n$ - $1$ porque se pierde una observación al obtener las diferencias consecutivas. Una gran ventaja del estadístico d es que se basa en los residuos estimados, que se calculan de manera rutinaria en los análisis de regresión [8].

## Test de Shapiro-Wilks.

El test de Shapiro-Wilks para normalidad se basa, más o menos, en una correlación entre los cuantiles empíricos de los residuos y los teóricos según una distribución normal. Cuanta mayor correlación, más indicios de normalidad para los residuos [4].

# Bibliografía


[1] Aguilera, N., & Quintana, M. (2011). *El peso de los niños y adolescentes y el rendimiento escolar en México*. El trimestre económico, 78(309), 115-141.

[2] Álvarez Uribe, M. C., & Montoya Puerta, E. C. (2004). *Velocidad media de ganancia de peso y estatura en niños de 2 a 10 años pertenecientes a familias del área rural del municipio de Marinilla-Antioquia, Colombia*. Revista española de salud pública, 78(2), 257-266.

[3] Anderson, D. R. & Sweeney, D. J. & Williams, T. A. M. D. C. H., & Álvarez, T. L. (2008). *Estadística para administración y economía.* Decima Edición. CENGAGE Learning.

[4] Aparicio J., Martinez, M., & Morales J. (2004). *Modelos lineales aplicados en R*. Dto. Estadística, Matemáticas e Informática.

[5] Chatterjee, Samprit & Hadi Ali S. (2012). *Regression Analysis by Example* (Wiley series in probability and statistic). John Wiley & Sons, Inc.

[6] Díaz Amador, Y., & Da Costa Leites Da Silva, L. (2019). *Caracterización de hábitos alimentarios y estado nutricional de preescolares*. Revista Cubana de Enfermería, 35(2).

[7] Guillén-López, L. S., & Vela-Amieva, M. (2010). *Desventajas de la introducción de la leche de vaca en el primer año de vida*. Acta Pediátrica de México, 31(3), 123-8.

[8] Gujarati, D. N & Porter, C. Dawn. (2008) Fifth Edition. *Basic Econometrics.* Mc GrawHill.

[9] Gutiérrez González E. & Vladimirovna Panteleeva O. (2016). *Estadística Inferencial 1 para Ingeniería y Ciencias*. Editorial Patria, 272-273

[10] Manual MSD.Crecimiento Físico de Lactantes y Niños (Consultado el día 08 de abril de 2023) *https://www.msdmanuals.com/es-mx/professional/pediatr%C3%ADa/crecimiento-y-desarrollo/crecimiento-f%C3%ADsico-de-lactantes-y-ni%C3%B1os*

[11] Mayorga, J. H., & Soto, O. F. (1988). *El análisis de regresión: Perspectiva histórica*. Revista Colombiana de Estadística, 9(17-18).

[12] Montgomery, Douglas C., Peck Elizabeth A. & Vining G. Geoffrey *Introduction to Linear Analysis* (Wiley series in probability and statistic). John Wiley & Sons, Inc.

[13] Rencher C. Alvin & Schaalje Bruce (2008). *Linear Models In Statistics*. John Wiley & Sons, Inc.

[14] Wackerly Dennis D., Mendenhall III William & Scheaffer Richard L. *Mathematical Statistics With Applications* (2008) Seventh Edition. CENGAGE Edition.

[15] Yugar, F., Flores, E., Vargas, N., & Vásquez, P. K. (2009). *Estudio de talla baja en preescolares de 2 a 5 años atendidos en consulta externa de pediatría en el hospital la paz*. Revista Médica La Paz, 15(2), 15-20.