ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

Modelos lineales generalizados para el análisis de defunciones en menores de un año en la región Sierra del Ecuador

VERÓNICA JANETH ARGÜELLO PAZMIÑO

Trabajo de Titulación modalidad: Proyectos de Investigación y Desarrollo, presentado ante el Instituto de Posgrado y Educación Continua de la ESPOCH, como requisito parcial para la obtención del grado de:

MAGÍSTER EN MATEMÁTICA MENCIÓN MODELACIÓN Y DOCENCIA

Riobamba-Ecuador

Julio - 2022

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

CERTIFICACIÓN:

EL TRIBUNAL DEL TRABAJO DE TITULACIÓN CERTIFICA QUE:

El **Trabajo de Titulación modalidad Proyectos de Investigación y Desarrollo**, titulado: Modelos lineales generalizados para el análisis de defunciones en menores de un año en la región Sierra del Ecuador, de responsabilidad de la señora Verónica Janeth Argüello Pazmiño, ha sido prolijamente revisado y se autoriza su presentación.

Tribunal:

Firmado electrónicamente por:
LUIS EDUARDO HIDALGO ALMEIDA

Ing. Luis Eduardo Hidalgo Almeida Ph. D.

PRESIDENTE

AMALIA ISABEL ESCUDERO VILLA

Firmado digitalmente por AMALIA ISABEL ESCUDERO VILLA
DN: CN=AMALIA ISABEL ESCUDERO VILLA, SERIALNUMBER=181021092802, OU=ENTIDAD DE CERTIFICACION DE INFORMACION, O=SECURITY DATA S.A. 2, C=EC
Razón: He revisado este documento
Foxit Reader Versión: 10.1.1

Ing. Amalia Isabel Escudero Villa Mag.

DIRECTORA

Firmado electrónicamente por:
JENNY PATRICIA PAREDES FIERRO

Ing. Jenny Patricia Paredes Fierro Mag.

MIEMBRO

Firmado electrónicamente por:
SALOMON RODRIGO CARGUA SUAREZ

Fis. Salomón Rodrigo Cargua Suarez Mag.

MIEMBRO

Riobamba, julio de 2022

DERECHOS INTELECTUALES

Yo, VERÓNICA JANETH ARGÜELLO PAZMIÑO, declaro que soy responsable de las ideas, doctrinas y resultados expuestos en el presente Trabajo de Titulación modalidad Proyectos de Investigación y Desarrollo, y que el patrimonio intelectual generado por la misma pertenece exclusivamente a la Escuela Superior Politécnica de Chimborazo.

Firmado electrónicamente por:
VERONICA JANETH ARGUELLO PAZMIÑO

VERÓNICA JANETH ARGÜELLO PAZMIÑO
No. Cédula: 0201976545

DECLARACIÓN DE AUTENTICIDAD

Yo, VERÓNICA JANETH ARGÜELLO PAZMIÑO, declaro que el presente Trabajo de Titulación Proyectos de Investigación y Desarrollo, es de mi autoría y que los resultados del mismo son auténticos y originales. Los textos constantes en el documento que provienen de otra fuente están debidamente citados y referenciados.

Como autor, asumo la responsabilidad legal y académica de los contenidos de este proyecto de investigación de maestría.

Firmado electrónicamente por:
VERONICA JANETH
ARGUELLO PAZMIÑO

VERRÓNICA JANETH ARGÜELLO PAZMIÑO
No. Cédula: 0201976545

DEDICATORIA

La concepción de este proyecto de tesis lo dedico a Dios, a mis padres, a mis hermanos, a mi esposo e hijos. A Dios porque siempre ha estado presente en toda esta etapa académica cuidándome y brindándome sus bendiciones, a mis padres María y Pedro quienes han sido el pilar de apoyo, protección y confianza desde mi niñez hasta el día de hoy, a mis hermanos Alexandra y Andrés que con su apoyo y consejos han hecho posible que siga adelante, a mi esposo Vinicio por apoyarme en cada etapa de preparación académica, a mis hijos Leandro y Samantha que son el motor fundamental que inundan mi corazón con ganas de seguir adelante venciendo obstáculos que a lo largo del camino se presentan en mi vida apoyándome con todo su amor.

Verónica

AGRADECIMIENTO

Agradezco en primera instancia a Dios, porque con su misericordia e inmenso amor me ha permitido llegar a este punto importante venciendo y superando diferentes obstáculos que se han presentado, a mis familiares y amigos por brindarme su apoyo incondicional, sus consejos y ayudas. También quiero extender mi agradecimiento a todos mis maestros y en especial a la Mg. Patricia Paredes, Mg. Isabel Escudero y Fis. Salomón Cargua por sus conocimientos, sugerencias, apoyo y experiencias que han hecho posible la realización del proyecto y así poder llegar a culminar una etapa de mi vida profesional.

Verónica

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ABREVIATURAS

AIC: Criterio de Información Akaike (Akaike Information Criterion),

GLM: Modelos Lineales Generalizados (GLM Generalized Linear Models)

MV: Máxima Verosimilitud

MCG: Mínimos Cuadrados generalizados

MCP: Mínimos Cuadrados Ponderados

OMS: Organización Mundial de la Salud

TOL: Tolerancia

VIF: Factor de Inflación de la Varianza

RESUMEN

El objetivo fue caracterizar y modelar las muertes en menores de un año, mediante modelos lineales generalizados. Con la estadística descriptiva y exploratoria se identificó como relevantes 2 variables cuantitativas (Muertes y Nacidos) y 3 cualitativas nominales (Género, Causas de Muerte, Provincias), también se obtuvo que, a nivel provincial, el comportamiento de la tasa de mortalidad no es homogéneo, observándose fuertes diferencias entre las entidades en cuanto a la mortalidad durante el primer año de vida. La provincia de Bolívar es la que lidera; sin embargo, hay un número considerable de muertes en Pichincha, en general en hombres; por último, se concluyó que la principal causa de muerte son las afecciones por periodo prenatal. Mediante el empleo de técnicas de regresión simple, se identificó los factores que han contribuido en la mortalidad infantil, así como las variables que explican las diferencias interprovinciales de dicho indicador a excepción de género. De igual manera se hizo un análisis de regresión lineal múltiple, obteniendo un Criterio de Información Akaike (AIC) de 427.66 y un R^2 de 0.9666, sin embargo, el modelo no cumplió con los supuestos, por lo tanto, no es adecuado para predecir, por ello se aplicó modelos lineales generalizados, en especial la regresión de Poisson, obteniendo un destacado ajuste, un R^2 de 0.858 y un AIC de 650.72. Finalmente se realizó simulaciones comparativas con el propósito de evaluar la posibilidad de aplicar la misma metodología a datos de mortalidad de todo el país, y se concluyó que sí es posible.

Palabras claves: <MATEMÁTICAS>, <MORTALIDAD INFANTIL >, <MODELOS LINEALES>, <MODELOS LINEALES GENERALIZADOS>, <REGRESION MULTINOMIAL>, <REGRESION DE POISSON>.

ABSTRACT

The goal of this study was to characterize and model deaths in children under one year of age, using generalized linear models. With the descriptive and exploratory statistics, two (2) quantitative variables (Deaths and Births) and three (3) nominal qualitative variables (Gender, Causes of Death, Provinces) were identified as relevant. It was also obtained that, at the provincial level, the behavior of the mortality rate is not homogeneous. It was observed strong differences between the entities in terms of mortality during the first year of life. The province of Bolívar is the one that leads. However, there is a considerable number of deaths in men in Pichincha province. It was concluded that the main cause of death is prenatal period conditions. Using simple regression techniques, the factors that have contributed to infant mortality were identified, as well as the variables that explain the interprovincial differences of this indicator except for the gender. In the same way, a multiple linear regression analysis was made getting an Akaike Information Criterion (AIC) of 427.66 and a R^2 of 0.9666, However, the model did not meet the assumptions. For this reason, it was not suitable to predict. Therefore, generalized linear models were applied, especially Poisson regression, getting an outstanding adjustment, a R^2 of 858 and an AIC of 650.72. Finally, comparative simulations were carried out with the purpose of evaluating the possibility of applying the same methodology to mortality data from all over the country, and it was concluded that it is possible.

Keywords: *<Mathematics>, <Childish Mortality>, <Linear Models>, <Generalized Linear Models>, <Mulnomial Regression> < Poisson Regression>.*

CAPÍTULO I

1. INTRODUCCIÓN

1.1. Situación Problemática

La mortalidad infantil en menores de un año continúa siendo una problemática desde hace muchas décadas en la salud, a nivel general. Según la Organización Mundial de la Salud (2019), las defunciones de 1 a 11 meses de edad representaron 1,5 millones. El parto y el posparto son los momentos en que las mujeres y los recién nacidos son más vulnerables. Se estima que cada año mueren 2,8 millones de embarazadas y recién nacidos, es decir 1 cada 11 segundos.

Las políticas que se han empleado sobre la salud en Ecuador han permitido la disminución de la tasa de mortalidad en menores de un año (Arguello, 2020), gracias a intervenciones específicas tales como: Programas de vacunación, prevención, tratamiento oportuno de enfermedades e infecciones, vigilancia y seguimiento continuo a mujeres embarazadas, entre otras. Dicho lo anterior en el año 1990 la tasa de mortalidad infantil en el Ecuador fue de 21,8 por cada 1000 nacidos vivos, en donde aplicando las debidas atenciones en los diferentes centros de salud en el año 2000, la tasa de mortalidad disminuyó en 6,3 puntos porcentuales.

En el año 2019 se registra una tasa de mortalidad infantil en menores de un año del 10,1 por cada 1000 nacidos vivos, disminuyendo en 0,1 puntos porcentuales con respecto al 2018 según el Instituto Nacional de Estadísticas y Censos.
En la región sierra del Ecuador la provincia de Pichincha tiene la tasa de mortalidad infantil más alta con 12,5 muertes, seguida por Bolívar con 11,2 por cada 1000 nacidos vivos correspondiente al año 2019. (INEC, 2019).

Con estos antecedentes este modelo lineal general surge de la necesidad de cuantificar las relaciones entre un conjunto de variables con otra, en la que una de ellas se denomina respuesta o dependiente, y las restantes son las explicativas o independientes. Se asume que la variable dependiente sigue una distribución normal y es homocedástica (McCulloch C., 2001), sin embargo, en muchas ocasiones estos supuestos no se cumplen y debe hacerse una transformación a la variable respuesta. Ciertas

investigaciones se enfrentan a problemas de la no normalidad debido a que las variables dependientes se obtienen por conteo, surgiendo la necesidad de estudiar otros modelos.

Según (Li, 1977) se puede usar diversas transformaciones a las variables para así poder ajustar a una normal, sin embargo, estas transformaciones no siempre ayudan a corregir algún supuesto como la falta de normalidad, heterocedasticidad o la no linealidad y su interpretación resulta muchas veces difícil. Un ejemplo muy común es en ecología, que a medida que aumenta la media de la muestra, aumenta también su varianza. Estos problemas se pueden llegar a solucionar mediante la transformación de la variable respuesta (por ejemplo, tomando logaritmos, raíz cuadrada) para ajustar a una normal (Cayuela, 2010). Una alternativa para que la variable respuesta no se transforme es aplicar modelos lineales generalizados que permiten utilizar distribuciones no normales de los errores y varianzas no constantes. Además, que son útiles para trabajar con cualquier distribución de la familia exponencial.

Los (GLM) es la extensión natural del Modelo Lineal clásico, inicialmente propuesto por Nelder y Wedderburn (1972), que ha llegado a suponer "una auténtica revolución estadística" (Ato, 2007), convirtiéndose en una solución especialmente adecuada para modelos de dependencia con datos no métricos. (López, 2011)

En el presente trabajo se pretende aplicar modelos lineales generalizados a las variables: provincias de la región sierra, causa de muerte en menores de un año y número de nacidos vivos puesto que son complejas y difíciles de definir, este indicador permite que las políticas públicas adoptadas por gobiernos de un país sean eficientes, debido a ello, resulta de vital importancia contar con reportes estadísticos a partir de GLM que faciliten la interpretación sobre las variables que influyen en las defunciones en menores de un año (Arguello, 2020) y que aporten a la toma de decisiones de un país.

1.2. Formulación del Problema

¿Los modelos lineales generalizados son adecuados para el análisis de defunciones en menores de un año en la región sierra del Ecuador?

1.3. Preguntas Directrices

¿Cómo el análisis estadístico descriptivo permite identificar las variables que tienen mayor influencia en el estudio de las defunciones en menores de un año?

¿Cómo los modelos lineales permiten identificar las variables más significativas en el estudio?

¿Cómo podemos modelar situaciones experimentales donde la variable respuesta es de tipo Binomial y como modelar la variable respuesta cuando el modelo lineal no es adecuado?

¿Cómo se puede validar los modelos mediante el criterio de AIC y realizando simulaciones comparativas?

1.4. Justificación de la Investigación

En el Ecuador entre los años 2000-2018 la tasa de mortalidad infantil ha disminuido considerablemente pero no en su totalidad esto gracias a la gratuidad de los servicios de salud, cobertura hospitalaria, infraestructura adecuada en los establecimientos de salud, accesibilidad de los servicios de saneamiento; sin embargo, no se ha logrado reducir la tasa de mortalidad infantil en dos terceras partes por lo cual ha provocado estancamiento en los últimos ocho años, por eso surge la necesidad de realizar este estudio, además se podrá analizar las causas que influyen en la mortalidad, e identificar en que provincias se presentan más casos para poder formular políticas y adoptar decisiones sobre la accesibilidad como la calidad de los servicios de asistencia.

Teórica: En cualquier tipo o área de investigación surge la necesidad de cuantificar relaciones entre las variables y poder explicar el comportamiento de una en función de otras. Este tipo de análisis generalmente se lo realiza utilizando los modelos lineales y en particular los de regresión y análisis de varianza, pero suponiendo que la variable respuesta es siempre continua y en general normalmente distribuida con media y varianza constante. Sin embargo, no cumple los supuestos necesarios, ya no es factible y no funciona. Ante esta limitación se analizan los datos mediante los GLM como una alternativa de procesamiento estadísticos.

Esta investigación pretende aportar al estudio la tasa de mortalidad para así medir el grado de desarrollo del país y en qué zona se debe poner mayor énfasis ya que es un indicador útil de las condiciones socioeconómicas en las que vive el país.

Práctica: La aplicación del modelo óptimo permitirá conocer en qué provincia de la región sierra hay mayor porcentaje de defunciones, establecer las causas que influyen en la mortalidad y realizar predicciones a corto plazo, esto ayudará a los diferentes centros de salud formular políticas de decisiones sobre la accesibilidad y la calidad de los servicios de asistencia en los lugares con alta tasa de mortalidad.

1.5. Objetivos de la Investigación

1.5.1. Objetivo General

Aplicar modelos lineales generalizados para el análisis de defunciones en menores de un año en la región sierra del Ecuador.

1.5.2. Objetivos Específicos

a) Diagnosticar las características de las variables mediante un análisis estadístico descriptivo y exploratorio.
b) Diseñar un esquema para identificar las variables significativas.
c) Proponer al menos dos modelos lineales generalizados adecuados.
d) Validar el modelo lineal generalizado óptimo mediante simulaciones comparativas.

1.6. Hipótesis

Los modelos lineales generalizados basado en simulaciones son adecuados para analizar las defunciones en menores de un año en diferentes provincias de la región sierra del Ecuador.

CAPÍTULO II

2. MARCO TEÓRICO

2.1. Antecedentes del Problema

Para la presente investigación se realizó la búsqueda de estudios que tengan relación con el problema de indagación como antecedentes bibliográficos, dentro de los cuales se detalla:

Bolancé C.,Vernic R. (2019), en su trabajo de investigación aplicaron modelos lineales generalizados de recuento multivariante basado en el enfoque de distribución Sarmanov que hace hincapié en distribuciones triviales, en este caso dichas distribuciones fueron representadas en tres tipos de reclamos que percibía una aseguradora, representado en un algoritmo de estimación basado en el método de máxima verosimilitud que permitieron mejorar su enfoque clásico.

Giuseppe De Luca, Jan R., Magnus, Peracchi F. (2018), en su investigación se enfocaron a extender el tratamiento de WALS para enfrentar la incertidumbre sobre especificación del predictor lineal en la clase más amplia de GLM. Esta clase incluye una variedad de modelos no lineales para resultados discretos y categóricos, como: logit, probit y regresión de Poisson. El tratamiento se basó desde un enfoque integral de WALS para GLM considerando que los enfoques WALS a los modelos lineales Gaussianos continúan manteniéndose en la clase amplia de la linealización de los estimadores de máxima verosimilitud restringida.

Minchón C., Vizconde T., Minchón D., & Minchón M. (2015), en su trabajo de investigación "Modelos lineales generalizados para pronóstico de la anemia infantil mediante factores asociados" En el análisis de distribución logística binaria la prevalencia de anemia fue asociada con los factores: área de residencia, región natural, edad, sexo, orden de nacimiento, periodo intergenésico y nivel de educación de la madre. Para cada factor se determinó las categorías que incrementan o disminuyen la prevalencia de anemia, y en todos los casos el modelo estimado fue adecuado. El análisis de regresión ordinal para la severidad de la anemia mostró también que los factores estaban asociados, y con excepción de la edad de la niña o niño como el nivel de educación de la madre, los modelos estimados fueron adecuados. El estudio revela que las técnicas incluidas en los GLM pueden ser empleados para pronosticar adecuadamente tanto la prevalencia como la severidad de la anemia infantil.

2.2. Bases Teóricas

2.2.1. Estadística Descriptiva

La estadística descriptiva es la rama de las Matemáticas que recolecta, presenta y caracteriza un conjunto de datos con el fin de describir apropiadamente las diversas características de ese conjunto.

Al conjunto de los distintos valores numéricos que adopta un carácter cuantitativo se llama variable estadística.
Las variables pueden ser de dos tipos:

- Cualitativas o categóricas: no se pueden medir numéricamente (por ejemplo: nacionalidad, color de la piel, sexo).
- Cuantitativas: tienen valor numérico (edad, precio de un producto, ingresos anuales).

La estadística para las variables cuantitativas es:

- Media
- Mediana
- Moda
- Varianza
- Desviación estándar

La estadística para las variables cualitativas es:

- Tablas de frecuencias
- Tablas de contingencia

2.2.2. Modelos de regresión lineal

2.2.2.1. Regresión lineal simple

La regresión lineal tiene como propósito modelar el comportamiento de una variable respuesta en función de una o más variables predictoras, este modelo puede usarse para realizar predicciones. (Gutiérrez, 2008)

Definición: Conforme Zurita Herrera (2010) la distribución lineal se basa en generar un modelo de regresión que ayude a explicar la relación lineal que existe entre dos variables. A la variable respuesta se la identifica como ***Y*** y a la predictora como ***X***. El modelo lineal simple esta descrita por la ecuación:

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X} + \varepsilon$$
$$\beta_0 = Es\ la\ ordenada\ en\ el\ origen$$
$$\beta_1 = Es\ la\ pendiente$$
$$\varepsilon = Error\ aleatorio\ con\ media\ 0\ y\ varianza\ \sigma^2$$
$$\mathrm{X} = Variable\ predictora$$

Dado el modelo condicional se trabaja en los siguientes supuestos:

$$E(Y_i|X = X_i) = B_0 + B_1 X_i$$
$$E(\varepsilon_i) = 0 \quad Normalidad$$
$$Var(\varepsilon_j) = \sigma^2 \quad Varianza\ Constante$$
$$cov(\varepsilon_i, \varepsilon_j) = 0;\ i \neq j \quad Independencia$$

Dónde:

$$E(Y_i|X = X_i) = Parte\ sistemática\ o\ determinística\ del\ modelo$$

Los valores $B_0, B_1\ y\ \sigma^2$, son constantes desconocidas, pero estadísticamente estimables. El hecho de que la varianza σ^2 del error sea constante durante todo el proceso es un supuesto fuerte y hace que el modelo utilizado sea considerado como homocedástico, es decir que la variabilidad es constante (Zurita Herrera, 2010). El error aleatorio es la diferencia entre los valores estimados y los valores reales.

En la mayoría de los estudios, los valores β_0 y β_1 poblacionales son desconocidos, debido a ello, a partir de una muestra, se obtienen sus estimaciones como:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$$

Estas estimaciones se las conoce como coeficientes de regresión, puesto que toman valores que minimizan la suma de cuadrados de los residuos (Amat Rodrigo, 2016).

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Dónde:

$$S_{xx} = \frac{\sum(x_i - \bar{x})^2}{n-1} \; ; \quad S_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Dando como resultado:

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

Mientras que $\hat{\beta}_0$ es:

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

Dónde $\boldsymbol{S_{xx}}$ y $\boldsymbol{S_{xy}}$ son las desviaciones de la variable X y Y respectivamnete. $\boldsymbol{\hat{\beta}_0}$ es el valor esperado de la variable Y cuando la variable $X = 0$. Una recta de regresión puede aplicarse en diferentes propósitos y para ello es necesario cumplir distintas condiciones. Siendo $\bar{x}$ y $\bar{y}$ las medias aritméticas de los valores observados de X y de Y; $\boldsymbol{\hat{y}_i}$ es el valor que el modelo estima para Y_i, dado $X = x_i$ con lo efectuado anteriormente se puede obtener el coeficiente de Correlación $\boldsymbol{r_{xy}}$ siendo:

$$S_{yy} = \frac{\sum_{i=1}^{n}\ (y_i - \bar{y})^2}{\text{n}}$$

Dando como resultado:

$$r_{xy} = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}}$$

Por otro lado, el Coeficiente de determinación R^2 se define como el cociente de la suma Cuadrática de Regresión para la suma Total:

Dónde la suma cuadrática de Regresión (SCR) se define de la siguiente manera:

$$SCR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

La suma cuadrática del Error (SCE) se define de la siguiente manera:

$$SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

La suma cuadrática Total (SCT) se define de la siguiente manera:

$$SCT = SCR + SCE$$

$$SCT = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Dando como resultado:

$$R^2 = \frac{SCR}{SCT} = \frac{SCT - SCE}{SCT} = 1 - \frac{SCE}{SCT}$$

Por otro lado, se puede comprobar en la regresión lineal simple la relación entre el coeficiente de Correlación de X con Y y el coeficiente de determinación que viene dado por:

$$r_{xy} = \pm\sqrt{R^2}$$

Tabla de Análisis de Varianza (Anova)

De acuerdo Zurita Herrera (2010) menciona el teorema de Cochran que establece que si cada una de las n observaciones y_i son tomadas de una misma Población Normal con parámetros μ_i y σ^2, la SCT se descompone en K sumas Cuadráticas que se denota SC_q, $q = 1; 2; \ldots; k$. Cada una de ellas con gl_q grados de libertad; por lo que el cociente SC_q/σ^2 tiene distribución Ji-Cuadrado con gl_q grados de libertad, teniendo la SCT, (n-1) grados de libertad esto significa:

$$\sum_{q=1}^{k} gl_q = n - 1$$

Se ha efectuado una partición de la SCT, donde K = 2, la una parte de la SCT es la SCR (q=1) y la otra es la SCE, con $gl_1 = (p - 1)$ y la SCE $(q = 2 = k)$ con $gl_2 = (n - p)$.

En la tabla 1-2 se visualiza la tabla de análisis de varianza o ANOVA que consiste en un arreglo rectangular cuyos componentes son: las fuentes de variación, grados de libertad, las sumas y medias cuadráticas y un valor adicional que es el estadístico de Prueba F.

Tabla 1-2: Análisis de Varianza o ANOVA

Fuentes de Variación	**Grados de libertad**	**Sumas Cuadráticas**	**Medias Cuadráticas**	**Estadístico de Prueba F**
Regresión	$\boldsymbol{p-1}$	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	$\frac{SCR}{p-1}$	$F_0=\frac{MCR}{MCE}$
Error (residuos)	$\boldsymbol{n-p}$	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$\frac{SCE}{n-p}$	
Total	$\boldsymbol{n-1}$	$\sum_{i=1}^{n}(y_i-\bar{y})^2$		

Fuente: Adaptado de Zurita Herrera, (2010)

Distribución F de Fisher

Es el cociente de:

$$F_0=\frac{MCR}{MCE}=\frac{SCR/1}{SCE/n(n-2)}$$

Es una variable aleatoria F con $(p-1)=1\ gl$ en el numerador y $(n-p)\ gl$ en el denominador. Si la Hipótesis Nula es verdadera $(H_0: B_1=0)$ entonces, $E(MCR)=E(MCE)=\sigma^2$, por lo que un valor del estadístico de prueba F, cercano a "uno" sería evidencia de que de que la Hipótesis Nula es verdadera, en tanto que un valor alejado de uno, evidencia la necesidad de rechazar H_0.

Por lo tanto, establecido el contraste:

$$H_0: \beta_1=0\ Vs.\ H_1: \beta_1\neq 0$$

Con $(1-\alpha)$ 100% de confianza la Hipótesis Nula dede ser rechazada si el Estadístico de Prueba:

$$F_0=\frac{MCR}{MCE}>F_{(\alpha;p-1,n-p)}$$

Una recta de regresión puede emplearse para diferentes propósitos y dependiendo de ellos es necesario satisfacer distintas condiciones. No obstante, si se quiere predecir el valor de una variable en función de la otra, no solo se requiere obtener los coeficientes del modelo, también es necesario verificar el ajuste del modelo a los datos (Amat Rodrigo, 2016).

2.2.2.2. Inferencia de Regresión lineal para β0 y β1

El modelo generado es una estimación de la relación de la población a partir de la relación de la muestra. Para los parámetros del modelo de regresión lineal simple (β0 y β1) se calcula su significancia (p valor) y su intervalo de confianza. El estadístico utilizado es el t-test. (Arguello, 2020)

Las hipótesis para la pendiente (β_1) son:

H_0: No hay relación lineal entre las variables. $\beta_1 = 0$

H_1: Sí hay relación lineal entre las variables. $\beta_1 \neq 0$

Los intervalos de confianza para cada $\boldsymbol{\beta_1}$

Con (1-α) 100% de confianza para β_i, el intervalo para cada β_i es:

$$\widehat{\beta_i} \pm SE(\hat{\beta}_i) t_{(\frac{\alpha}{2},n-2)}$$

El intervalo quedaría como:

$$\widehat{\beta_i} - SE(\hat{\beta}_i) t_{(\frac{\alpha}{2},n-2)} \leq \beta_i \leq \widehat{\beta_i} + SE(\hat{\beta}_i) t_{(\frac{\alpha}{2},n-2)}$$

dónde $t_{(\frac{\alpha}{2},n-2)}$ es el percentil de la distribución t de Student con n-2 grados de libertad que deja a su derecha un área de $\alpha/2$. Cuando n tiene menos observaciones, menor es su capacidad de calcular el error estándar del modelo y debido a ello la exactitud de los coeficientes de regresión estimados se reduce. (Alexandra, 2020)

2.2.2.3. Regresión lineal múltiple

En varias situaciones, existe la posibilidad de tener varias variables predictoras (X_1,X_2,X_3 ….) que pueden estar influyendo con la variable respuesta (Y). La regresión múltiple es una extensión de la regresión simple. (Gutierrez, 2018). Los modelos de regresión múltiple se aplican para predecir la variable respuesta o para evaluar la influencia que tienen las variables predictoras sobre ella (Amat Rodrigo, 2016).

Los modelos de regresión múltiples esta descrita por la siguiente ecuación:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_n X_n + \varepsilon_i}$$, donde

$$\varepsilon_i \sim N(0, \sigma^2 I_n)$$

β_0: es la ordenada en el origen, es el valor que toma la variable Y cuando todas las variables predictoras son cero.

β_i: es el efecto promedio que tiene al incrementar una unidad en X_i sobre la variable Y, manteniéndose constantes el resto de las variables.

ε_i : es el error aleatorio.

Esta vez con los parámetros $\boldsymbol{\beta_0, \beta_1, \beta_2}$, dos variables de explicación $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ y el error $\boldsymbol{\varepsilon_i}$. Para determinar la influencia que tienen en el modelo cada una de las variables predictoras, se utilizan los coeficientes estandarizados.(Amat Rodrigo, 2016).

Supuestos del Modelo

Joaquín Amat (2016) señala que los modelos de regresión múltiple solicitan que se cumplan los siguientes supuestos:

Multicolinealidad:

Las variables predictoras deben ser independientes, es decir, que un predictor no debe estar linealmente relacionado con uno o varios de los otros predictores y al no cumplirse esto no se puede identificar de forma precisa la influencia que tiene cada una de las variables predictoras sobre la variable respuesta. Cuando el coeficiente de correlación es 1 puede que exista multicolinealidad, esto no ocurre con frecuencia.

No existe un método estadístico que permita determinar la existencia de multicolinealidad entre los predictores, sin embargo, hay algunos pasos que se deben seguir para detectarla.

a) Si R^2 es alto, pero ninguna de las variables predictoras resulta significativa, hay sospechas de multicolinealidad.
b) Calcular una matriz de correlación y estudiar la relación entre las variables predictoras.
c) Realizar una regresión lineal simple entre cada uno de las variables predictoras frente al resto. Si en alguno el coeficiente de determinación R^2 es alto, estaría señalando a una posible multicolinealidad.
d) Tolerancia (TOL) y Factor de Inflación de la Varianza (VIF). Se los calcula con las siguientes formulas:

$$VIF_{\hat{\beta}j} = \frac{1}{1 - R^2}$$

$$Tolerancia\ _{\hat{\beta}j} = \frac{1}{VIF_{\hat{\beta}j}}$$

Los límites de referencia del VIF son:

a) VIF = 1: Ausencia total de multicolinealidad
b) 1 < VIF < 5: La regresión se ve afectada por la multicolinealidad.
c) 5 < VIF < 10: Cuando el valor esta entre 5 y 10 debe tener preocupación

d) El termino tolerancia es 1/VIF y los límites recomendados están entre 1 y 0.1.

En caso de que se presente multicolinealidad entre las variables predictoras, se puede dar dos posibles soluciones. La primera es excluir la variable predictora que presenta problemas e identificar si está influyendo realmente en la variable respuesta, al aplicar esta medida no influye en la capacidad predictiva de modelo. La segunda opción es combinar las variables multicolineales en un único predictor. (Amat Rodrigo, 2016)

Independencia

Para verificar si los errores son independientes, se aplica el test Durbin- Watson, y puede tomar valores entre 0 y 4, donde un valor cercano a 2 significa que los residuos no están correlacionados, por otro lado, un valor superior a 2 indica una correlación positiva entre errores subyacentes, mientras que un valor menor a 2 representa una correlación negativa entre errores.

Distribución normal

Los residuos están distribuidos de forma normal con media cero y varianza σ^2. Para comprobarlo se realizan gráficas de cuantiles normales y test de hipótesis de normalidad de Shapiro Wills o Kolmogorov Smirnov. (Amat Rodrigo, 2016)

Homocedasticidad

Los residuos deben tener una varianza constante a lo largo del tiempo. Para comprobarlo se gráfica l los valores ajustados vs los residuos, esta debe mantener una misma dispersión y sin ningún patrón específico. Una forma de cubo es un claro ejemplo de falta de homocedasticidad. También se puede recurrir a contrastes de homocedasticidad como el test de Breusch-Pagan. (Amat Rodrigo, 2016)

2.2.3. Modelo Lineal Generalizado (GLM)

Los modelos lineales generalizados son una extensión de los modelos lineales clásicos, son una alternativa a transformaciones de la respuesta, justificadas por la falta de linealidad y homogeneidad de la varianza, y tiene como objetivo describir el efecto de una o más variables explicativas sobre una o más variables respuestas. (Martínez, 2001)

2.2.3.1. Componentes del Modelo Lineal Generalizado

Existen tres componentes en el modelo lineal generalizado. (Bueno, 2013)

a) **Componente aleatorio:** Las variables respuestas $Y_i, i = 1, \dots, n$ comparten la misma distribución en la familia exponencial.

b) **Componente sistemático:** Está dado por el predictor lineal, define la relación entre η, que es una función del valor esperado de Y, y las variables independientes en el modelo.

$$\boldsymbol{\eta_{=}\beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_p X_p}$$

Por lo tanto, los coeficientes de regresión se pueden interpretar de manera idéntica a los de la regresión lineal: un cambio de 1 unidad en X1 da como resultado un cambio de b1 unidades en η, manteniendo todas las demás variables constantes. (Little, T, 2013)

c) **Función de enlace**: La función de enlace $g()$, relaciona la media condicional de Y, también conocido como el valor esperado de Y, $E(Y/X) = \mu$, a la combinación lineal de predictores de η.

$$g(\mu_i) = \eta_i = X_i'\beta$$

2.2.3.2. Propiedades del modelo lineal Generalizado

Hay dos propiedades del modelo lineal generalizado que se consideran importantes:

a) Estructura de los errores: Cuando los datos tienen una estructura no normal, las únicas herramientas para poder tratar la ausencia de normalidad era la transformación de la variable respuesta o la aplicación de métodos no paramétricos, pero hoy se puede aplicar los GLM, los cuales permiten especificar distintos tipos de distribución de errores. En la tabla 2-2 se visualiza la distribución de errores que más se utilizan en los modelos lineales generalizados. (Cayuela, 2010).

Tabla 2-1: Distribución de errores del GLM

Distribución	Descripción de la función
Poisson	Son muy útiles para conteos (p. e. número de muertos por accidentes de tráfico, número de días con heladas en el mes de enero, etc.)
Binomial	Son de gran utilidad para proporciones y datos de presencia/ausencia (p. e. tasas de mortalidad, tasas de infección, presencia o ausencia de una determinada especie, etc.)
Gamma	Son muy útiles con datos que muestran un coeficiente de variación constante, esto es, en donde la varianza aumenta según aumenta la media de la muestra de manera constante (p. e. número de presas comidas por un predador en función del número de presas, etc.)
Exponenciales	Son muy útiles para los análisis de supervivencia

Fuente: Adaptado de Luis Cayuela, (2010)

b) Función de vínculo: Otra razón por la que un modelo lineal puede no ser adecuado para describir un fenómeno determinado es que la relación entre las variables repuesta e independiente(s) no es siempre lineal.

La función de vínculo, por lo tanto, se encarga de linealizar la relación entre la variable respuesta e independiente(s) mediante la transformación de la variable respuesta. En la tabla 3-2 se muestra las funciones de vínculos más comunes utilizadas por los modelos lineales generalizados. (Cayuela, 2010).

Tabla 3-2: Funciones de vínculo más comunes utilizados por los GLM

Función de vínculo	Fórmula	Uso
Identidad	μ	Datos continuos con errores normales (regresión y ANOVA).
Logarítmica	$Log(\lambda)$	Conteos con errores de tipo Poisson
Logit	$Log\left(\frac{\mu}{n-\mu}\right)$	Proporciones (datos entre 0 y 1) con errores binomiales
Recíproca	$\frac{1}{\mu}$	Datos continuos con errores gamma
Raíz cuadrada	$\sqrt{\mu}$	Conteos
Exponencial	μ^n	Funciones de potencia

Fuente: Luis Cayuela, (2010)

Las funciones de vínculo canónicas se aplican por defecto a cada una de las distribuciones de errores. Pero esto no significa que siempre se deba usar una única función de vínculo para una determinada distribución. De hecho, puede ser recomendable comparar diferentes funciones de vínculo para un mismo modelo y ver con cual se obtiene un mejor ajuste. En la tabla 4-2 se puede observar las funciones de vínculo canónicas para cada una de las distribuciones de errores, así como otras posibles funciones de vínculo que pueden ser usadas.

Tabla 4-2: Funciones de vínculo canónico para distribuciones de errores en GLM

Distribución de errores	Función de vínculo canónica	Otras funciones de vínculo posibles
Normal	Identidad	Logarítmica
Poisson	Logarítmica	Identidad, Raíz cuadrada
Binomial	Logit	Logarítmica
Gamma	Recíproca	Identidad, Logarítmica

Fuente: Luis Cayuela, (2010)

2.2.3.3. Construcción y evaluación de un GLM

Al construir un GLM es importante tener en cuenta que no existe un único modelo que sea válido, y este es un error común al aplicar regresión, debido a que se usa el mismo modelo varias veces sin una perspectiva crítica. En muchos casos, habrá varios modelos aceptables que puedan ajustarse al conjunto de datos. Parte de la labor de construcción y evaluación del modelo es establecer cuál es el más adecuada (Cayuela, 2010).

Los pasos a seguir para la construcción y evaluación de un GLM se describen a continuación:

1. **Exploración de los datos:** Es recomendable conocer nuestros datos. Se puede realizar algunos gráficos que muestren la relación entre la variable respuesta y cada una de las predictoras.
   Lo importante es:
   - Buscar posibles relaciones de la variable respuesta con las predictoras.
   - Aplicar transformaciones de las variables de ser el caso.
   - Eliminar las variables predictoras que estén correlacionadas.

2. **Elección de la estructura de errores en función de vínculo**: Elegir las propiedades del modelo, analizar residuos y ver su idoneidad en la distribución de errores elegida.

3. **Ajuste del modelo a los datos:**

   Conocer la significancia de los estimadores del modelo mediante test, y se verifica mediante la cantidad de varianza explicada por el modelo. En los GLM se determina la conocida desviación D^2. La devianza de la idea de variabilidad explicada por el modelo, se compara la devianza del modelo nulo (Null deviance) con la desviación residual (Residual deviance):

$$D^2 = \frac{Devianza\ modelo\ nulo - Devianza\ residual}{Devianza\ modelo\ nulo} * 100$$

4. **Análisis de los residuos:** Muchas veces utilizan los residuos estandarizados que deben seguir una distribución normal. (Cayuela, 2010), debido a ello es importante analizar los siguientes gráficos:
   a) Histograma de los residuos
   b) Gráfico de los residuos frente a los valores estimados
   c) El gráfico probabilístico de normalidad

2.2.3.4. Estimación de los Modelos Lineales Generalizados

Los parámetros desconocidos del GLM se pueden estimar mediante dos métodos clásicos. (Bueno, 2013)

a) Método de Máxima Verosimilitud: Permite estimar el vector de parámetros desconocidos β y una característica importante de los modelos lineales generalizados es que todos pueden ajustarse a los datos utilizando el mismo algoritmo, una forma de mínimos cuadrados ponderados iterativamente. A continuación, describimos el algoritmo.

Dada una estimación de prueba de los parámetros $\hat{\beta}$, calculamos el predictor lineal estimado $\hat{\eta}_i = X_i'\hat{\beta}$ y usa eso para obtener los valores ajustados $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$. Usando estas cantidades, calculamos la variable de trabajo.

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{d\eta_i}{d\mu_i} \tag{1.1}$$

dónde el término más a la derecha es la derivada de la función de enlace evaluada en la estimación de prueba.

A continuación, calculamos los pesos iterativos

$$w_i = p_i/[b''(\theta_i)(\frac{d\eta_i}{d\mu_i})^2] \tag{1.2}$$

Donde $b''(\theta_i)$ es la segunda derivada de $b(\theta_i)$ evaluada en la estimación de prueba y hemos supuesto que $a_i(\phi)$ tiene la forma usual ϕ/p_i. Este peso es inversamente proporcional a la varianza de la variable dependiente de trabajo z_i dadas las estimaciones actuales de los parámetros, con factor de proporcionalidad ϕ.

Finalmente, obtenemos una estimación mejorada de β haciendo una regresión de la variable dependiente de trabajo z_i sobre los predictores x_i usando los pesos w_i, es decir, calcular la estimación de mínimos cuadrados ponderados.

$$\hat{\beta} = (X'WX)^{-1}X'\mathrm{W}z \tag{1.3}$$

dónde X es la matriz del modelo, Wi es una matriz diagonal de pesos con entradas wi dadas por (1.2) y $\boldsymbol{z}$ es un vector de respuesta con entradas zi dadas por (1.1). (Little, 2013)

b) Método de Mínimos Cuadrados Generalizados: Se obtiene minimizando la suma de cuadrados de los errores

$$\min_{\boldsymbol{\beta}\in\Re^p} \boldsymbol{S}(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}\in\Re^p} (\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{Y}-\boldsymbol{X\beta}) \tag{1.4}$$

Derivando e igualando a cero $S(\beta)$ se obtiene:

$$(X'X)\beta = X'Y \tag{1.5}$$

Resolviendo la ecuación (1.5) se tiene:

$$\beta = (X'X)^{-1}X'Y \tag{1.6}$$

2.2.4. Criterios de evaluación del modelo

a) **Criterio de Información de Akaike (AIC):** El criterio más usado para evaluar este tipo de modelos es el criterio de Información Akaike (AIC es Akaike Information Criterion), el cual es un índice que evalúa el ajuste de los datos, como su complejidad. Y mientras más pequeño es este valor mejor es el ajuste. Es apropiado para comparar modelos similares con diferentes grados de complejidad o iguales (Cayuela ,2010).

Este modelo está definido como:

$$AIC = 2k - 2\ln(L) \tag{1.7}$$

Dónde:

k: Es el número de parámetros en el modelo

L: Es el valor máximo de la probabilidad de función para el modelo estimado

b) Criterio de información Bayesiana (BIC): Es similar al AIC, pero incorpora el tamaño de la muestra, penaliza la adición de parámetros en mayor medida que el AIC. Es recomendable utilizar cuando la muestra es grande y el número de parámetros es pequeño. Está definido de la siguiente manera.

$$BIC = 2\ln(n) - 2\ln(L) \quad (1.8)$$

2.2.5. Tipos de Modelos Lineales Generalizados

Se puede visualizar los principales GLM en la tabla 5-2.

Tabla 5-2: Modelos Lineales Generalizados

Naturaleza de la variable de Respuesta	Componente		Función de enlace	Modelo Lineal	Siglas
	Sistemático	Aleatorio			
Numérica cuantitativa	• Numérico • Categórico • Mixto	• Normal • Normal • Normal	• Identidad • Identidad • Identidad	• Regresión Lineal • ANOVA o de diseño experimental • ANCOVA o de diseño experimental con variables concomitantes	ML
Categórica binaria • No agrupada • Agrupada (frecuencias)	• Mixto • Categórico	• Binomial (1) • Bernoulli • Binomial (n)	• Logit • Logit • Probit	• Regresión Logística • Análisis Logit • Regresión probit	GLM
Categórica politómica • No agrupada • Agrupada (frecuencias)	• Mixto • Categórico	• Multinomial	• Logit Generalizado	• Regresión Logística multinomial • Análisis logit multinomial	
Recuento	• Mixto	• Poisson	• Logarítmica	• Regresión de Poisson	
Frecuencia	• Categórico	• Poisson	• Logarítmica	• Análisis Loglineal	

Fuente: Adaptado de López González y Ruíz Soler, (2011)

2.2.6. Regresión logística binaria

La regresión logística binaria es un análisis apropiado y de uso común cuando la variable de resultado es binaria, lo que significa que el resultado toma uno de dos valores mutuamente excluyentes. Los ejemplos de variables binarias en el área de salud más comunes son: vivos o muertos, infectados o saludables. La regresión logística binomial es un GLM con estructura de distribución binomial y función de enlace logit. La función de masa de probabilidad para la distribución binomial.

$$P(Y = y|n, \pi) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y} \quad (1.9)$$

Da la probabilidad de observar un valor dado, y, de la variable Y que se distribuye con una distribución binomial con parámetros n y π. Considere una variable binaria que tiene dos valores mutuamente excluyentes; uno de estos valores es el valor de resultado de interés, a menudo llamado "éxito" o "caso", y ocurre con probabilidad π. La distribución binomial da la probabilidad de un número específico de éxitos, y, en un conjunto de n ensayos independientes, donde cada éxito ocurre con probabilidad π y cada fracaso con probabilidad $(1 - \pi)$.

La función de enlace canónico para la distribución binomial es el logit. Esta función permite que el modelo de regresión logística tenga una forma lineal.

El logit se define como el logaritmo natural de las probabilidades, de que ocurra un evento dividido por la probabilidad de que no ocurra, y la función de logit es:

$$logit = ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$$

dónde $\hat{\pi}$ es la probabilidad predicha de que ocurra un evento. Una ventaja de GLM es que permite una relación no lineal entre los valores predichos y los predictores.

La forma lineal del modelo de regresión logística binaria es de la siguiente forma:

$$\begin{aligned} logit &= ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \eta \\ &= \beta_o + \beta_1 X_1 + \beta_2 X_2 + \dots\dots\dots + \beta_p X_p \end{aligned} \tag{2.0}$$

La ecuación 2.0 indica por qué la regresión logística binaria a menudo se denomina "lineal en el logit".

Al elevar ambos lados de la ecuación a potencia tenemos:

$$e\left(ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)\right) = e^{\beta_o+\beta_1 X_1+\beta_2 X_2+\dots\dots\dots+\beta_p X_p}$$

$$odds = \frac{\hat{\pi}}{1-\hat{\pi}} = e^{\beta_o+\beta_1 X_1+\beta_2 X_2+\dots\dots\dots+\beta_p X_p}$$

$$odds = \frac{\hat{\pi}}{1-\hat{\pi}} = e^{\beta_o} e^{\beta_1 X_1} e^{\beta_2 X_2} \dots\dots e^{\beta_p X_p} \tag{2.1}$$

Se puede examinar el efecto del cambio de 1 unidad en X_1 en las probabilidades.

$$e^{b_1(X_1+1)} = e^{b_1 X_1+b_1} = e^{b_1 X_1} e^{b_1} \tag{2.2}$$

El término e^{b_1} se conoce como razón de probabilidades y es útil cuando se interpreta el efecto de los predictores.

Al manipular la ecuación 2.1 se podrá interpretar los coeficientes de regresión en términos de la probabilidad.

$$odds = \frac{\hat{\pi}}{1-\hat{\pi}}$$

$$(1-\hat{\pi}) * \ odds = \hat{\pi}$$

$$\hat{\pi} = \frac{odds}{1-odds} \tag{2.3}$$

Al sustituir odds por los coeficientes tenemos:

$$\hat{\pi} = \frac{e^{\beta_o} e^{\beta_1 X_1} e^{\beta_2 X_2} \ldots\ldots e^{\beta_p X_p}}{1-(e^{\beta_o} e^{\beta_1 X_1} e^{\beta_2 X_2} \ldots\ldots e^{\beta_p X_p})} \tag{2.4}$$

La relación entre las probabilidades y logit se representa en una gráfica con una curva en S. (Little, T, 2013)

2.2.6.1. Regresión Logística Multinomial

La regresión logística multinomial es la generalización de la regresión logística, se emplea en modelos con variable dependiente de tipo nominal que tiene tres o más categorías (politómica), y las variables independientes suele ser variables continuas, discretas y categóricas. (Gonzalez, 2015).

Las variables dependientes politómicas clásicamente fueron modeladas mediante análisis discriminante, pero, gracias al avance en el desarrollo de técnicas de cálculo, es más habitual utilizar modelos de regresión logística multinomial (Gómez & Palacios, 2013), implementados en paquetes estadísticos en R, debido a la fácil interpretación de los resultados que proporcionan. (Roque Cruz , 2018).

2.2.6.2. Formulación del Modelo

Sea una variable aleatoria Y_i y toma un número finito de valores, 1, 2,…, J. Sea $p_{ij} = P(Y_i = j)$ entonces $\sum_{j=1}^{J} p_{ij} = 1$.

El modelo multinomial está definido de la siguiente manera:

$$\left.\begin{array}{llll} p_1(X_1, X_2) = & p_1 = & E(Y_1) = & \frac{\exp{(Z_1)}}{1+\exp(Z_1)+\exp{(Z_2)}} \\ p_2(X_1, X_2) = & p_2 = & E(Y_2) = & \frac{\exp{(Z_2)}}{1+\exp(Z_1)+\exp{(Z_2)}} \end{array}\right\} \tag{2.5}$$

Dónde $Z_1 = \beta_{01} + \beta_{11} * X_1 + \beta_{21} * X_2 \ \ y \ Z_2 = \beta_{02} + \beta_{12} * X_1 + \beta_{22} * X_2$

Siendo $\beta_{01}, \beta_{11}, \beta_{21} \beta_{02}, \beta_{12}, \beta_{22}$ parámetros que se desea estimar.

Los modelos logit-multinomial vendrán definidos en términos de los log-odds (logaritmo del cociente). Al predecir cociente de probabilidad se suele elegir uno de los niveles de las variables a explicar cómo nivel de referencia, generalmente el primero o el último.

Al cociente p_1/p_2 se le llama "odds" y se representa por $O_1(X_1, X_2) = O_1(idem.\,para\ O_2)$. De esta manera se puede observar que la razón de cambio en O_1 cuando X_1 se incrementa en unidad manteniéndose constante

X_2 viene dada por $\frac{O_1(X_1+1,X_2)}{O_1(X_1,X_2)} = \exp(\beta_{11})$, que recibe el nombre de "odds-ratio" de la categoría 1 respecto de la X_1 y se representa por $OR_1(X_1), para\ OR_1(X_2), OR_2(X_1) y\ OR_2(X_2)$.

Es interesante observar que estas "odds-ratio" dependen de las unidades de medida en que vengan las variables regresoras (si se multiplica X_1 por 10, $OR_1(X_1)$pasaría a ser $\sqrt[10]{\exp(\beta_{11})}$. Por tanto, la influencia de cada regresora debería medirse por el valor de la odds-ratio suponiendo que esten estandarizadas. Cuando el valor sea más grande más relevante es la variable en el modelo. También es importante definir las proporciones de cambio en las "odds" con respecto a cada variable regresora

$$\frac{O_1(X_1+1,X_2)-O_1(X_1,X_2)}{O_1(X_1,X_2)} = OR_1(X_1) - 1 = \exp(\beta_{11}) - 1 \tag{2.6}$$

Otra alternativa y más conocida, se obtiene tomando logaritmos en ambas ecuaciones del modelo:

$$\left.\begin{aligned} ln\left(\frac{p_1}{p_3}\right) = \quad Z_1 = \quad \beta_{01} + \beta_{11} * X_1 + \beta_{21} * X_2 \\ ln\left(\frac{p_2}{p_3}\right) = \quad Z_2 = \quad \beta_{02} + \beta_{12} * X_2 + \beta_{22} * X_2 \end{aligned}\right\}$$

Dónde las ecuaciones de la parte izquierda se denominan "logits" (como en la regresión logística binaria) y los parámetros representan las tasas de cambio en los "logits" , es decir, cuando una de las variables explicativas se incrementa en una unidad manteniéndose constante la otra. (Roque Cruz , 2018).

2.2.6.3. Estimación de parámetros

Dado unos datos $(Y_{1i}, Y_{21}, X_{1i}, X_{2i})$ con $i = 1,2 \ldots, n$ se puede definir, en función de los parámetros del modelo, las funciones $Z_{1i}, Z_{2i}, P_{1i}, P_{2i}$ y tocar el problema de la estimación de los mismos mediante el método de máxima verosimilitud.

Al tener el modelo planteado, la función de verosimilitud viene dada por la siguiente ecuación.

$$L = \prod_{i=1}^{n} \left(p_{1i}^{Y_{1i}} * p_{2i}^{Y_{2i}} * p_{3i}^{1-Y_{1i}-Y_{2i}}\right) = \prod_{i=1}^{n} \left(\left(\frac{p_{1i}}{p_{3i}}\right)^{Y_{1i}} * \left(\frac{p_{2i}}{p_{3i}}\right)^{Y_{21}} * p_{3i}\right)$$

Y en vez de trabajar de manera directa con esta expresión se utiliza la función auxiliar:

$$\Lambda = -2 * \ln(L) = -2 * \sum_{i=1}^{n} \left(Y_{1i} * ln\left(\frac{p_{1i}}{p_{3i}}\right) + Y_{2i} * ln\left(\frac{p_{2i}}{p_{3i}}\right) + ln(p_{3i}) \right)$$

$$= 2 * \sum_{i=1}^{n} \left(ln\left(1 + \exp\left(Z_{1i}\right) + exp(Z_{2i})\right) - Y_{1i} * Z_{1i} - Y_{2i} * Z_{2i} \right)$$

Ahora la maximizar la verosimilitud equivale minimizar la función auxiliar Λ_Λ y se puede resolver por métodos numéricos de forma iterativa partiendo de la estimación inicial $\beta_{11} = \beta_{21} = \beta_{12} = \beta_{22} = 0$, $\beta_{01} = \ln(n_1) - \ln(n - n_1 - n_2)$ y $\beta_{02} = ln(n_2) - ln(n - n_1 - n_2)$ siendo $n_1\, y\, n_2$ el número de observaciones en las categorías 1 y 2 respectivamente. Los estimadores iniciales se obtienen asumiendo que no hay una influencia de variables regresoras en el modelo planteado y debido a ello el valor inicial de la función auxiliar se minimiza:

$$\Lambda_0 = -2 * \left(n_1 * ln\left(\frac{n_1}{n}\right) + n_2 * ln\left(\frac{n_2}{n}\right) + (n - n_1 - n_2) * ln\left(\frac{n - n_1 - n_2}{n}\right) \right) \quad (2.7)$$

Una vez alcanzada la convergencia del método iterativo, se designa por Λ_0 al mínimo coseguido y por $\hat{\beta}_{01}, \hat{\beta}_{11}, \hat{\beta}_{21}, \hat{\beta}_{02}, \hat{\beta}_{12}, \hat{\beta}_{22}$ a los valores estimados de los parámetros del modelo. (Roque Cruz , 2018)

2.2.6.4. Significatividad global del modelo

Se puede comparar la hipótesis de no existencia de un efecto significativo global en las variables regresoras asumiendo que la diferencia entre el valor inicial y el valor final de la función auxiliar Λ tiene una distribución x^2 con 4 grados de libertad. El p-valor de la prueba para la hipótesis nula de que no existe efecto de las variables regresoras ($(\beta_{11} = \beta_{21} = \beta_{12} = \beta_{22} = 0)$ vendrá dado por $p(X_4^2 > \Lambda_0 - \Lambda_f)$. (Roque Cruz , 2018)

2.2.6.5. Significatividad del efecto de cada variable regresora

Se denomina Λ_{-1} al mínimo de la función auxiliar que se obtendrá excluyendo del modelo la $X_1(\beta_{11} = \beta_{12} = 0)$, se confirma que la diferencia entre los mínimos de la función auxiliar en el modelo reducido y en el completo tiene una distribución x^2 con 2 grados de libertad. Por tanto, el valor p de la prueba para la hipótesis nula de que no existe efecto de la $X_1(\beta_{11} = \beta_{12} = 0)$ vendrá dado por $p(x_2^2 > \Lambda_{-1} - \Lambda_0)$. De modo similar se calcula:

Λ_o(mínimo de la función auxiliar eliminando $\beta_{01} y\, \beta_{02}\ del\ modelo$) y $\Lambda_{-2}(mínimo\ de\ la\ función\ auxiliar\ eliminando\ del\ modelo\ la\ variable\ x_2)$ y construir tests de hipótesis para $\beta_{01} = \beta_{02} = 0$ y $\beta_{21} = \beta_{22} = 0$ respectivamente. (Roque Cruz , 2018)

2.2.6.6. Significatividad de cada parámetro

Se tiene en cuenta que el cuadrado de cada estimador dividido por su error estándar tiene una distribución x^2 con 1 grado de libertad se construye la prueba de hipótesis para la igualdad de cada parámetro a cero y se determinar los estimadores de los parámetros del modelo son significativamente distintos de cero. La prueba de hipótesis $\beta_{11} = 0$ el p-valor sería $p\left(x_1^2 > \left(\frac{\beta_{11}}{s.e(\beta_{11})}\right)^2\right)$, siendo $s.e.\left(\hat{\beta}_{11}\right)$ el valor correspondiente al error estándar del estimador del parámetro β_{11}. (Roque Cruz , 2018).

2.2.7. *Regresión de Poisson*

La regresión de Poisson es un análisis adecuado cuando la variable respuesta es un recuento del número de eventos en un período fijo de tiempo y se pude modelar en términos de tasas de incidencia que depende de algunas variables explicativas.

La función de masa de probabilidad de la distribución de Poisson es:

$$P(Y = y|\lambda) = \frac{\lambda^y}{y!} e^{-\lambda} \qquad (2.8)$$

Da la probabilidad de observar un valor dado, y, de la variable de resultado Y que se distribuye como una distribución de Poisson con parámetro λ.

Los elementos básicos para plantear una regresión de Poisson son: una variable dependiente $\boldsymbol{Y}$ basadas en conteos, para la que se asume una distribución de Poisson y se va a estudiar la relación con las variables explicativas que determinan las condiciones específicas para la observación.
Se intenta construir un modelo $\lambda(x) = \mathrm{E}((Y|X = x)$, para la media de $\boldsymbol{Y}$ condicionada a cada valor de la variable explicativa. Como $\boldsymbol{Y}$ no toma valores negativos, no se puede utilizar un modelo lineal directo, por ello, se necesita una función de enlace que es el logaritmo natural (ln).
Se expresaría de la siguiente forma la regresión de Poisson de forma lineal:

$$\mathbf{ln}\left(\hat{\lambda}\right) = \boldsymbol{X'\beta}$$

$$\ln(\hat{\lambda}) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_p X_p \quad (2.9)$$

dónde $\hat{\lambda}$ es el recuento predicho de la variable respuesta, dados los valores específicos de los predictores.

Como los valores solo pueden tomar valores enteros no negativos, los residuos toman solo un conjunto limitado de valores.

Al hacer la interpretación como la regresión lineal tiene la desventaja de interpretar el cambio en las unidades de una transformación del resultado, por ello elevando a potencia a ambos lados de la ecuación (2.9) tenemos:

$$e^{\ln(\hat{\lambda})} = e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_p X_p}$$

$$\hat{\lambda} = e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_p X_p}$$

$$\hat{\lambda} = e^{\beta_o} e^{\beta_1 X_1} e^{\beta_2 X_2} \ldots\ldots e^{\beta_p X_p} \quad (3.0)$$

Ahora se ve que el cambio en el valor de un predictor da como resultado un cambio multiplicativo en el conteo predicho, para examinar efecto de un cambio de 1 unidad en X_1 en el resultado:

$$e^{b_1(X_1+1)} = e^{b_1 X_1 + b_1} = e^{b_1 X_1} e^{b_1} \quad (3.1)$$

El término e^{b_1} es el efecto de un cambio de 1 unidad en X_1 en el resultado. Para un aumento de 1 unidad en X_1, el recuento predicho ($\hat{\lambda}$) se multiplica por e^{b_1}, manteniendo constantes todas las demás variables. (Little, 2013)

Para estimar los parámetros del modelo se utiliza la función de máxima verosimilitud, como se muestra a continuación:

$L(\beta) = \prod_{i=1}^{n} \left[e^{-\lambda(x_i,\beta)} \frac{\lambda(x_i,\beta) y_i}{y_i!} \right]$ siendo su logaritmo (en términos dependientes de β)

$$L(\beta) = \sum_{i=1}^{n} \left(y_i x_i' \beta - e^{x_i' \beta} \right)$$

Derivando la función y se iguala a cero, se obtienen las ecuaciones de verosimilitud:

$$\frac{\partial log L(\beta)}{\partial \beta} = \sum\nolimits_{i=1}^{n} x_i' \left[y_i - \lambda(x_i, \beta) \right] = 0,$$

La matriz hessiana es:

$$\frac{\partial^2 log L(\beta)}{\partial \beta \partial \beta^t} = -\sum\nolimits_{i=1}^{n} x_i x_i' \, \lambda(x_i, \beta)$$

Las ecuaciones de verosimilitud son no lineales en los parámetros, debido a ello, se recurre a procedimientos iterativos para el cálculo de sus estimaciones. Newton-Raphson y el IRLS (Iterative

Re-weighted Least Squares) son los métodos iterativos que se emplean en este modelo. Además, se sabe que los estimadores de máxima verosimilitud son asintóticamente normales, centrados y su matriz de varianzas-covarianzas es la inversa de la matriz de información (la matriz hessiana cambiada de signo), esto permite hacer inferencias sobre los parámetros.

La desviación para un modelo de Poisson se da:

$$D = -2\sum_{i=1}^{m_i}\left[y_i\, ln\left(\frac{m_i}{y_i}\right) + (y_i - m_i)\right]$$

Las estadísticas D se pueden usar como medidas de bondad de ajuste, ya que se pueden calcular a partir de los datos y del modelo ajustado, y para medir la significancia del modelo se utiliza el análisis de la desvianza.

Tabla 6-2: Análisis de la Devianza

	Grados de libertad	Devianza	Devianza media
Modelo nulo	$\boldsymbol{k-1}$	$\boldsymbol{D(\lambda(\hat{\beta}); \hat{\lambda}_{nulo})}$	$\boldsymbol{\frac{D(\lambda(\hat{\beta}); \hat{\lambda}_{nulo})}{k-1}}$
Modelo residual	$\boldsymbol{N-k}$	$\boldsymbol{D(y; \lambda(\hat{\beta}))}$	$\boldsymbol{\frac{D(y; \lambda(\hat{\beta}))}{N-k}}$
Total corregido	$\boldsymbol{N-1}$	$\boldsymbol{D(y; \hat{\lambda}_{nulo})}$	$\boldsymbol{\frac{D(y; \hat{\lambda}_{nulo})}{N-1}}$

Fuente: Adaptado de Elmis Garcia, Jeanette Gonzáles y Eloy López, (2014)

La medida de ajuste es el Pseudo R^2

$$R^2 = \frac{(Devianza\ nulo - Devianza\ residual)}{Devianza\ nulo}$$

2.2.7.1. Sobredispersión

La sobredispersión es un problema en los GLM con media y varianza dependientes, la regresión de Poisson asume la equidispersión es decir asume que la media y la varianza son iguales, sin embargo, los datos reales pueden estar dispersos, la varianza de los residuos es mayor que la media e Ignorar la sobredispersión da como resultado una subestimación de los errores estándar

La sobredispersión ocurre cuando la regresión de Poisson supone que cada evento que ocurre para un individuo es un evento independiente, que no puede ser el caso o cuando se omite un predictor importante para el modelo.

El ajuste más simple para la sobredispersión es el modelo de Poisson sobredisperso. (Little, 2013).

2.2.7.2. Regresión de Poisson sobredispersa

El modelo de regresión de Poisson sobredispersión incluye un parámetro adicional que se utiliza en la estimación de la varianza conocido como parámetro de escala de sobredispersión ϕ.

El modelo estimado con esta corrección ahora asume esencialmente una distribución de error de Poisson con media λ y varianza $\phi\lambda$.

El parámetro de escala ϕ será mayor a 1 si hay sobredispersión en los datos, igual a 1 si hay equidispersión y menor a 1 si los datos están subdispersión. La cantidad de dispersión en el modelo es determinada por el estadístico de bondad de ajuste chi-cuadrado de Pearson. El parámetro de escala se calcula como: (Little, 2013)

$$\phi = \frac{X^2_{Pearson}}{df}$$

CAPÍTULO III

3. METODOLOGÍA DE INVESTIGACIÓN

3.1. Tipo y diseño de investigación

Este estudio tendrá una investigación descriptiva puesto que se aplicará un análisis exploratorio (AED) que permitirá conocer la tasa de mortalidad y que causas de muerte son las más influyentes, también contará con una investigación correlacional, es decir, se realizará un análisis de correlación para identificar la relación entre variables y descartar multicolinealidad que puede ser un problema a la hora de aplicar modelos de regresión, se aplicará regresión lineal y múltiple para identificar las variables individuales y conjuntas que contribuyen significativamente, además, se utilizará regresión logística multinomial para conocer los factores asociados a las causas de muerte, y una regresión de Poisson con número de eventos en menores de un año que va a permitir modelar estos datos de conteo y se verificará la medida de calidad del modelo mediante AIC, para finalizar se tomará en cuenta la investigación explicativa debido a que una vez encontrando el modelo óptimo se simulará eventos similares, para verificar que tan adecuado es con estas bases simuladas y dar una conclusión sobre si aplicar o no este modelo a los datos de mortalidad infantil a nivel Ecuador.

3.2. Métodos de investigación

La presente investigación iniciará con él método cuantitativo de los datos de las defunciones en menores de un año proporcionados por el INEC, iniciando con un análisis descriptivo, luego un inferencial y por último multivariante.

3.3. Enfoque de la investigación

Tendrá un enfoque cuantitativo puesto que es una investigación empírica-analista, ya que utiliza datos para probar las hipótesis utilizando análisis estadístico descriptivo y analítico que permita establecer patrones de comportamiento para predecir la media de defunciones en menores de un año objeto de estudio.

3.4. Técnicas de investigación

El INEC para recolectar la información lo realiza mediante Informes y registros clínicos obtenidos facilitados por las instituciones de salud en el año 2019, sin embargo, también se realizará la consulta

de documentos públicos (reportes oficiales), la revisión de artículos científicos, bibliografía web y libros que ayudarán al sustento para el desarrollo del estudio.

3.5. Selección de la muestra

Dado que es una base de datos que se encuentra en el INEC, se tomará los registros del INEC del año 2019, considerando las 11 provincias de la región sierra que son: Azuay, Bolívar, Cañar, Carchi, Cotopaxi, Chimborazo Imbabura, Loja, Pichincha, Tungurahua y Santo Domingo de los Tsáchilas con un total de 119840 nacidos vivos y 1517 defunciones en menores de un año, no se aplicará ninguna técnica de muestreo.

3.6. Datos

La base datos que se manejará para el estudio de modelos lineales y generalizados son las defunciones de menores de un año del 2019 recolectadas por el Instituto Nacional de Estadísticas y Censos. Se cuenta con la siguiente información: provincias, causa de muerte, número de muertes, número de nacidos vivos; todo esto correspondiente a la región sierra del Ecuador.

3.7. Modelos lineales

Suponga que un modelo lineal explica la relación entre una variable de respuesta continua y un conjunto de variables regresoras $X_1, X_2, \ldots, X_n$.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \ldots + \beta_n X_{ni} + \varepsilon_i$$

Gran parte del modelado de regresión requiere la suposición de que las ε_i son variables aleatorias independientes e idénticamente distribuidas (i.i.d.) $N(O, \sigma^2)$.

Supuestos del modelo lineal

- Normalidad
- Independencia
- Homocedaticidad
- Multicolinealidad

En Rstudio se tiene la función $lm()$ que genera las estimaciones de los parámetros β y el estadístico t obtenido para cada uno de los parámetros junto con los valores p correspondientes permiten verificar la significancia de estos.

3.8. Modelos Lineales Generalizados

El modelo de regresión lineal utiliza la linealidad para describir la relación entre la media de la variable de respuesta y un conjunto de variables explicativas, asumiendo que la distribución de la respuesta es normal. Los (GLM) amplían los modelos de regresión lineal para abarcar distribuciones de respuesta no normales.

En los estudios de salud es frecuente trabajar atributos, que se miden de forma no métrica (discreta, nominal u ordinal), no ajustándose al Modelo Lineal clásico e incumpliendo los supuestos de linealidad y normalidad, estas situaciones requieren de modelos que trabajen con datos dicotómicos, es decir, de modelos de probabilidad de un evento (modelos logit, probit, modelos de regresión de Poisson y modelos de regresión ordinal). Estos modelos son parte integrante de los Modelos Lineales Generalizados y, junto con la regresión lineal, el análisis de varianza, la regresión logística, los modelos de respuesta multinomial, e incluso ciertos análisis de supervivencia y de series temporales, son extensiones del Modelo Lineal clásico.

Ahora bien, para hablar del Modelo Lineal Generalizado debemos explicar cómo se desarrolla, el *modelado estadístico*, que pretende explicar la variación de una respuesta a partir de la relación conjunta de dos fuentes de variabilidad, una de carácter determinista y otra aleatoria, lo que responde a la expresión: (López, 2011)

$$\textit{Respuesta} = \textit{componente sistemático} + \textit{componente aleatorio.}$$

3.9. Modelo Binomial

La regresión logística se caracteriza por preguntas de investigación binaria (sí/no o éxito/fracaso) o binomial (número de síes o éxitos en n intentos). Y se asume que la variable respuesta Y_i tiene una distribución binomial $Y_i \sim B(n_i, \pi_i)$, con n_i eventos y probabilidad π_i. Esto define la estructura del modelo.

Supone que el logit de la probabilidad es una función lineal de los predictores

$$logit(\pi_i) = X'\beta$$

Los coeficientes de regresión se pueden interpretar de la misma manera que en los modelos lineales, teniendo en cuenta que el lado izquierdo es un logit. (Rodriguez, 2016)

Supuestos de regresión logística

La regresión logística para hacer inferencias requiere suposiciones del modelo.

- **Respuesta binaria.** La variable respuesta es dicotómica
- **Independencia.** Las obervaciones deben ser independientes entre sí.
- **Estructura de Varianza**
- **Linealidad.** El $ln\left(\frac{\pi}{1-\pi}\right)$ debe ser lineal en función de X. (Roback, 2021)

3.10. Modelo de Poisson

La regresión de Poisson es el modelo más básico para variables respuesta de recuento. Morales (2018) considera que este modelo es usual en datos discretos que proceden de conteos de sucesos que se producen al azar con cierta periodicidad y son modelizables como tasas de incidencia que dependen de ciertas variables explicativas. Para modelizar este tipo de fenómenos se utiliza la distribución de Poisson, $X \sim Po(\lambda)$, donde λ representa el número medio de ocurrencias y la regresión de Poisson de forma lineal se expresa de la siguiente forma:

$$\mathbf{ln}(\hat{\boldsymbol{\lambda}}) = \boldsymbol{X'\beta}$$

Supuestos del modelo de Poisson

- Equidispersión

Rstudio cuenta con la función glm() y permite aplicar diferentes modelos lineales generalizados, especificando en el argumento family, que distribución aplicar. (Alexandra, 2020)

a) binomial (link= "logit")
b) gaussian (link= "identity")
c) gamma (link= "inverse")
d) inverse.gaussian (link= "1/mu ^2")
e) poisson (link= "log")
f) quasi (link= "identity",variance="constant")
g) quasibinomial (link= "logit")
h) quasipoisson (link= "log")

3.11. Modelos Multinomiales

La regresión logística multinomial se utiliza para predecir la probabilidad de pertenencia a una categoría en una variable dependiente en función de múltiples variables independientes. Las variables independientes pueden ser dicotómicas o continuas. La regresión logística multinomial es una extensión simple de la regresión logística binaria que permite más de dos categorías de la variable dependiente. Al igual que la regresión logística binaria, la regresión logística multinomial utiliza la estimación de máxima verosimilitud para evaluar la probabilidad de pertenencia categórica (Starkweather, 2011).

Los modelos multinomiales vendrán definidos en términos de (logaritmo del cociente) las probabilidades condicionadas con dos niveles de respuesta.

$$\ln\left(\frac{\pi_{i|jk}}{\pi_{I|jk}}\right)$$

Supuestos del modelo multinomial

- Independencia entre variables dependientes.
- Separación no perfecta

Los modelos multinomiales se determinaron en Rstudio gracias a la función multinom. Para ello se utilizó la librería library(nnet) y así plantear el modelo.

CAPÍTULO IV

4. RESULTADOS Y DISCUSIÓN

En el presente capítulo, se presentan los resultados obtenidos en el proceso de diagnóstico sobre las características de las variables e identificar las más significativas, la propuesta de dos modelos lineales generalizados y su validación mediante simulaciones comparativas.

4.1. Características de las variables mediante un análisis descriptivo y exploratorio.

Para determinar las características de las variables se realizó un mapa que permite observar las provincias con las que se va a trabajar.

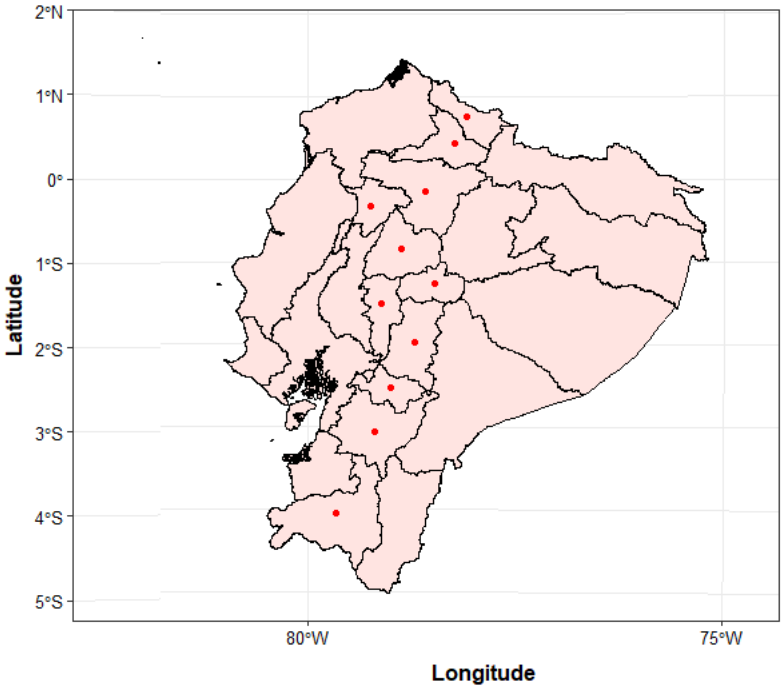


Figura 1-4: Mapa de las provincias de la región sierra ecuatoriana
Realizado por: Arguello, Verónica, 2021

Mediante un resumen descriptivo se identificó que: Muertes y Nacidos son variables cuantitativas, y Provincias, Género, Causas de muerte son cualitativas nominales, y sus estadísticos se presentan en las siguientes tablas de contingencia.

Tabla 1-4: Tabla de contingencia entre las variables Provincia y Género respecto al número de muertes en menores de un año.

Provincias	Género		Total
	Hombre	Mujer	
Azuay	78	69	**147**
Bolívar	25	18	**43**
Cañar	29	22	**51**
Carchi	22	9	**31**
Chimborazo	51	33	**84**
Cotopaxi	62	43	**105**
Imbabura	38	34	**72**
Loja	51	31	**82**
Pichincha	342	273	**615**
Tungurahua	53	34	**87**
Santo Domingo de los Tsáchilas	46	40	**86**
Total	**797**	**606**	**1403**

Realizado por: Arguello, Verónica, 2021

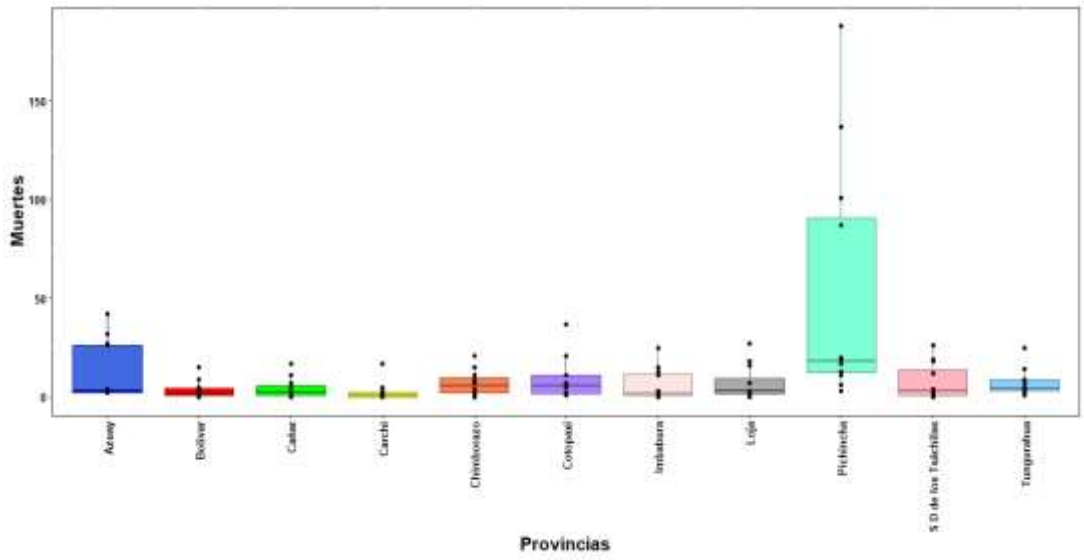


Figura 2-4: Gráfica de cajas de las provincias (Azuay, Bolívar, Cañar, Carchi, Cotopaxi, Chimborazo Imbabura, Loja, Pichincha, Tungurahua y Santo Domingo de los Tsáchilas), respecto a las muertes en menores de un año.
Realizado por: Arguello, Verónica, 2021

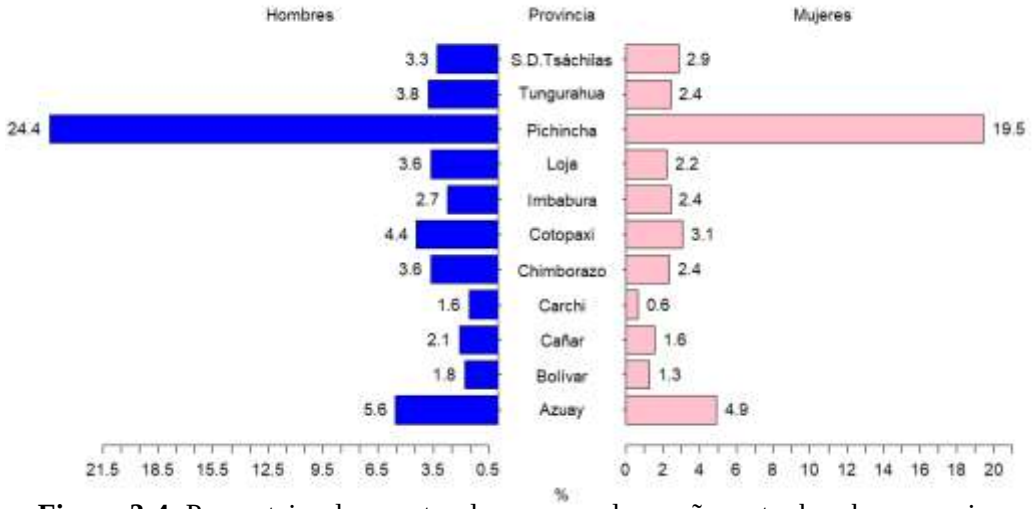


Figura 3-4: Porcentajes de muertes de menores de un año, entre hombres y mujeres por provincia

Realizado por: Arguello, Verónica, 2021

De acuerdo a la tabla 1-4, la figura 2-4 y 3-4 la provincia que tiene alto porcentaje de muertes es Pichincha que acumula un 43.83% de la población de estudio, mientras que el Carchi presenta un 2.21%, también se visualizó un incremento de muertes en hombres que en mujeres en toda la sierra.

Tabla 2-4: Tabla de contingencia entre las variables Provincia y Causas de muerte respecto al número de muertes en menores de un año.

Provincias	Causas de Muerte						Total
	Afecciones periodo prenatal	Malformaciones	Obstrucción respiratoria	Influenza y neumonía	Resto de causas	Causas mal definidas	
Azuay	69	58	4	8	4	4	**147**
Bolívar	24	9	1	4	2	3	**43**
Cañar	28	12	6	1	2	2	**51**
Carchi	22	5	1	1	2	0	**31**
Chimborazo	36	16	5	13	2	12	**84**
Cotopaxi	58	13	22	7	3	2	**105**
Imbabura	40	24	4	2	2	0	**72**
Loja	45	23	2	5	3	4	**82**
Pichincha	325	188	37	32	24	9	**615**
Tungurahua	39	18	10	9	7	3	**86**
S D de los Tsáchilas	45	30	2	4	2	4	**87**
Total	**731**	**396**	**94**	**86**	**53**	**43**	**1403**

Realizado por: Arguello, Verónica, 2021

En la tabla 2-4 apunta que la principal causa de muerte más común en niños menores de un año son las afecciones originadas en el periodo prenatal en todas las provincias de la sierra, y sigue siendo Quito la que cuenta con mayor número de muertes.

Tabla 3-4: Tabla de contingencia entre las variables Género y Causas de muerte respecto al número de muertes en menores de un año.

Género	Causas de Muerte						Total
	Afecciones periodo prenatal	Malformaciones	Obstrucción respiratoria	Influenza y neumonía	Resto de causas	Causas mal definidas	
Hombre	440	210	48	53	22	24	**797**
Mujer	291	186	46	33	31	19	**606**
Total	**731**	**396**	**94**	**86**	**53**	**43**	**1403**

Realizado por: Arguello, Verónica, 2021

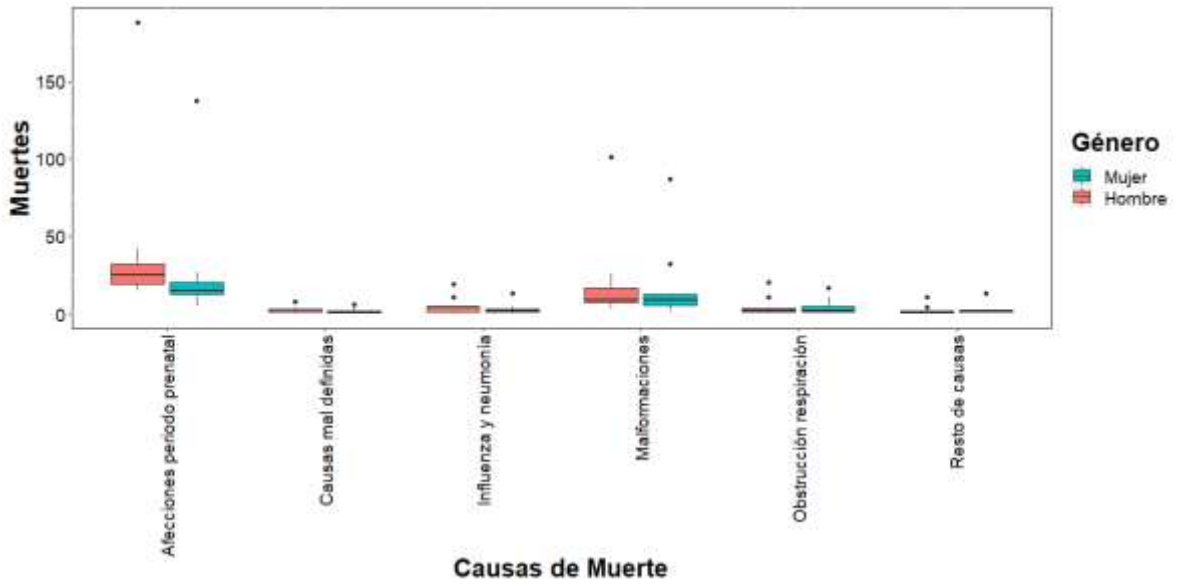


Figura 4-4: Gráfico de cajas de las causas de muerte (Afecciones por periodo prenatal, Malformaciones, Influenza y neumonía, obstrucción respiratoria, causas mal definidas, resto de causas) y la variable genero (hombre, mujer) con respecto a las muertes de menores de un año
Realizado por: Arguello, Verónica, 2021

La tabla 3-4 y la figura 4-4, se identificó que en hombres se presenta con mayor porcentaje la principal causa de muerte que es afecciones en periodo prenatal, seguida por malformaciones y se concluye que los hombres son más propensos.

Tabla 4-4: Tabla de contingencia entre las variables Provincia y Género respecto a la variable Nacidos vivos.

Provincias	Género		Total
	Hombre	Mujer	
Azuay	6685	6545	**13230**
Bolívar	1508	1441	**2949**
Cañar	2373	2270	**4643**
Carchi	1331	1276	**2607**
Chimborazo	3804	3758	**7562**
Cotopaxi	3840	3821	**7661**
Imbabura	3716	3741	**7457**
Loja	4128	3896	**8024**
Pichincha	23825	23193	**47018**
Tungurahua	4533	4497	**9030**
Santo Domingo de los Tsáchilas	4908	4751	**9659**
Total	**60651**	**59189**	**119840**

Realizado por: Arguello, Verónica, 2021

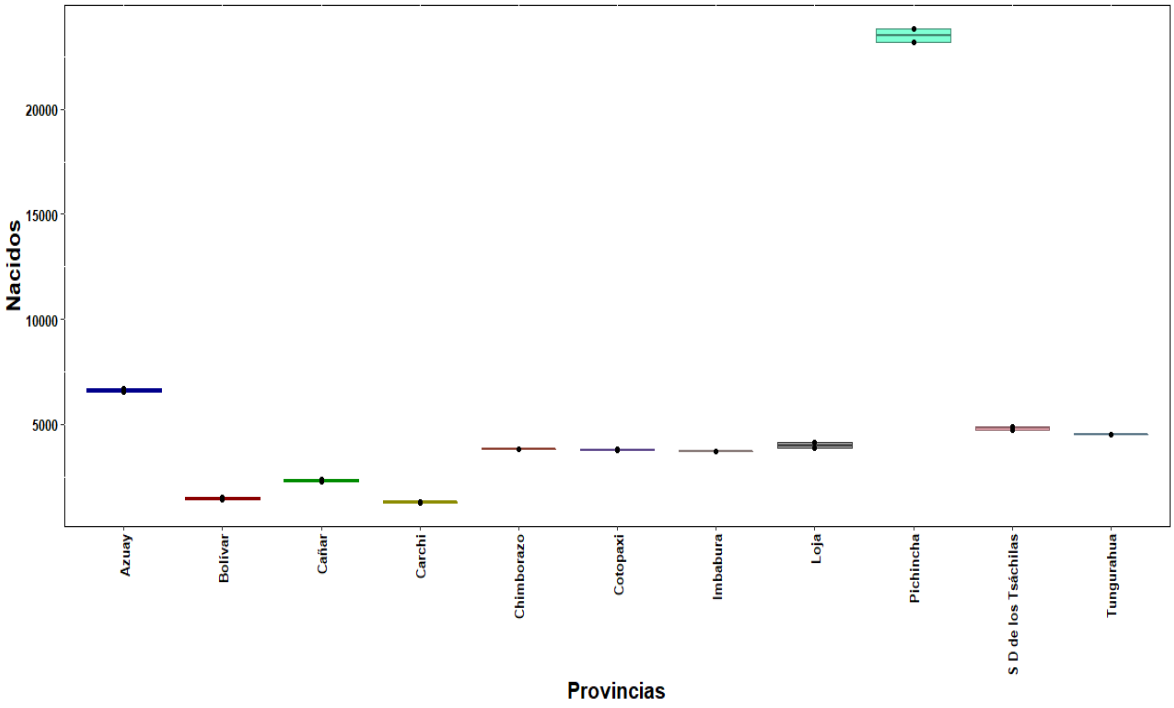


Figura 5-4: Gráfica de cajas de las provincias (Azuay, Bolívar, Cañar, Carchi, Cotopaxi, Chimborazo Imbabura, Loja, Pichincha, Tungurahua y Santo Domingo de los Tsáchilas), respecto a nacidos vivos.

Realizado por: Arguello, Verónica, 2021

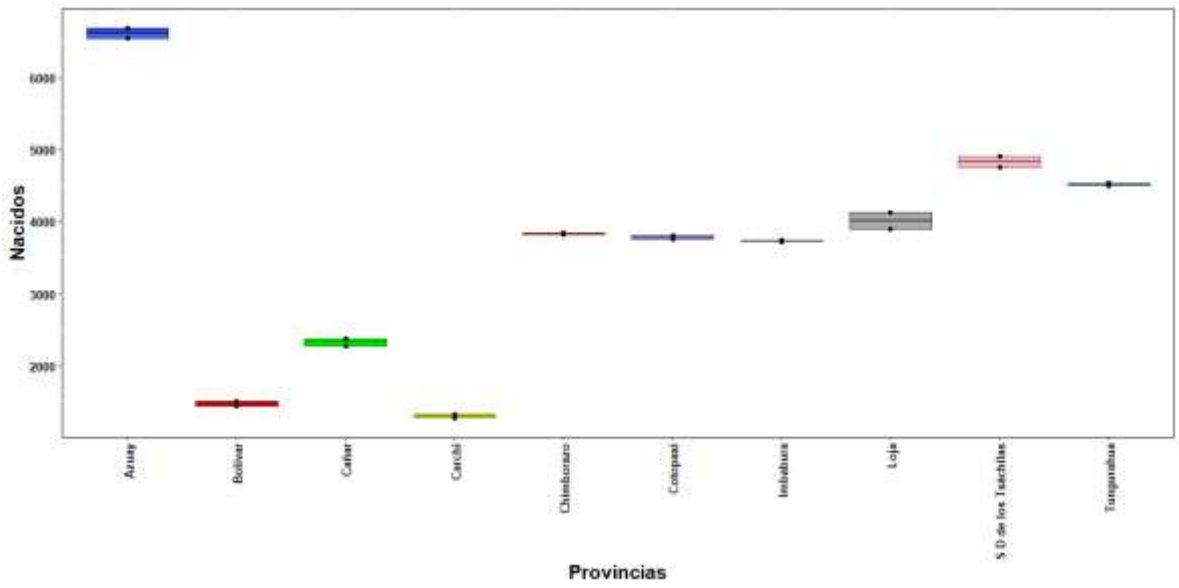


Figura 6-4: Gráfica de cajas de las provincias (Azuay, Bolívar, Cañar, Carchi, Cotopaxi, Chimborazo Imbabura, Loja, Tungurahua y Santo Domingo de los Tsáchilas), respecto a nacidos vivos.
Realizado por: Arguello, Verónica, 2021

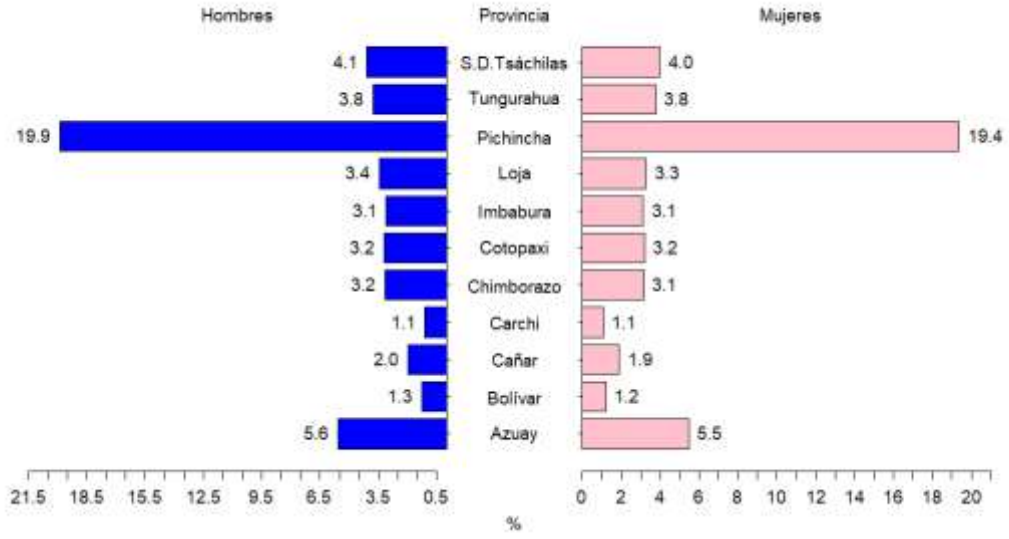


Figura 6-4: Porcentajes de nacidos vivos, entre hombres y mujeres por provincia
Realizado por: Arguello, Verónica, 2021

De acuerdo con la tabla 4-4 y la figura 5-4 y 6-4, el mayor porcentaje de nacidos en la provincia de Pichincha con un 39.33% y un menor en Carchi con 2.18%, y acorde a la tabla y gráfica no hay una diferencia significativa entre nacimientos de hombre con mujeres.

Tabla 5-4: Tasa de mortalidad por provincias de la región sierra del Ecuador

Provincias	Nacidos	Muertes	Tasa de Mortalidad
Azuay	13230	159	12.018
Bolívar	2949	48	16.277
Cañar	4643	52	11.200
Carchi	2607	35	13.425
Cotopaxi	7562	115	15.208
Chimborazo	7661	93	12.139
Imbabura	7457	79	10.594
Loja	8024	88	10.967
Pichincha	47018	657	13.973
Tungurahua	9030	99	10.963
Santo Domingo de los Tsáchilas	9659	92	9.525
Total	119840	1517	12.659

Realizado por: Arguello, Verónica, 2021

De acuerdo a la tabla 5-4 se tiene que Bolívar es la provincia que reporta la mayor tasa de mortalidad con 16.28% lo que quiere decir que por 1000 nacidos hay 16 muertos y Santo Domingo de los Tsáchilas tiene menor tasa de mortalidad con 9.52%.

4.2. Identificar las variables significativas.

Conforme con los resultados obtenidos se realizó un análisis de regresión lineal de muertes con cada una de las variables explicativas para identificar su significancia.

- **Análisis del número de muertes de menores de un año respecto a las causas de muerte.**

$$Y_i = 33.227 - 31.273 * Causas\ mal\ definidas - 29.318 * Influenza\ y\ Neumonía - 15.227 * Malformaciones - 28.955 * Obstrucción\ y\ Respiración - 30.818 * Resto\ de\ Causas.$$

Tabla 6-4: Coeficientes de la regresión lineal simple de la variable Causas de Muerte y su valor de significancia

Variables Independientes	Coeficientes	Error Estándar	Valor T	Valor p
Constante	33.227	4.451	7.465	0.000
Causas mal definidas	-31.273	6.295	-4.968	0.000
Influenza y Neumonía	-29.318	6.295	-4.657	0.000
Malformaciones	-15.227	6.295	-2.419	0.017
Obstrucción y Respiración	-28.955	6.295	-4.6	0.000
Resto de Causas	-30.818	6.295	-4.896	0.000
R-SQ.(ADJ.)	**0.211**	**SE**	**20.88**	

Realizado por: Arguello, Verónica, 2021

Existe una relación lineal significativa entre las Causas de Muerte y Muertes, excepto una de las Causas de muerte que es Malformaciones, puesto que, el valor p es menor que el nivel de significancia α=0.05.

Tabla 7-4: Análisis de varianza de la regresión lineal simple de la variable Causas de Muerte

Fuentes de Variación	Suma de cuadrados	GL	Cuadrados medios	Estadístico F	Valor p
Modelo	17454	5	3490.800	8.009	0.000
Error	54922	126	435.890		
Total	72376	131			

Realizado por: Arguello, Verónica, 2021

Al comprobar si el modelo es significativo, se obtiene un valor p menor que el nivel de significancia de 0.05, por lo tanto, se concluyó que si es significativo.

- **Análisis del número de muertes de menores de un año respecto a las provincias.**

$$Y_i = 12.250 - 8.667\ Bolívar - 8.000\ Cañar - 9.667\ Carchi - 5.250\ Chimborazo - 3.500\ Cotopaxi - 6.250\ Imbabura - 5.417\ Loja + 39.000\ Pichincha - 5.000\ Santo\ Domingo\ de\ los\ Tsáchilas - 5.083\ Tungurahua$$

Tabla 8-4: Coeficientes de la regresión lineal simple de la variable Provincia y su valor de significancia

Variables Independientes	Coeficientes	Error Estándar	Valor T	Valor p
Constante	12.250	5.855	2.092	0.039
Bolívar	-8.667	8.280	-1.047	0.297
Cañar	-8.000	8.280	-0.966	0.336
Carchi	-9.667	8.280	-1.167	0.245
Chimborazo	-5.250	8.280	-0.634	0.527
Cotopaxi	-3.500	8.280	-0.423	0.673
Imbabura	-6.250	8.280	-0.755	0.452
Loja	-5.417	8.280	-0.654	0.514
Pichincha	39.000	8.280	4.710	0.000
Santo Domingo de los Tsáchilas	-5.000	8.280	-0.604	0.547
Tungurahua	-5.083	8.280	-0.614	0.540
R-SQ.(ADJ.)	**0.255**	**SE**	**20.28**	

Realizado por: Arguello, Verónica, 2021

El valor p indica que las provincias de Bolívar y Pichincha son significativas, mientras que las demás no lo son.

Tabla 9-4: Análisis de varianza de la regresión lineal simple de la variable Provincia

Fuentes de Variación	Suma de cuadrados	GL	Cuadrados medios	Estadístico F	Valor p
Provincia	22604	10	2260.400	5.495	0.000
Error	49772	121	411.340		
Total	72376	131			

Realizado por: Arguello, Verónica, 2021

Al verificar si el modelo es significativo, se obtuvo un valor p menor que el nivel de significancia de 0.05, por lo tanto, se concluyó que si es significativo.

- **Análisis del número de muertes de menores de un año respecto a los nacidos vivos.**

$$Y_i = -1.445860 + 0.0022164\ Nacidos$$

Tabla 10-4: Coeficientes de la regresión lineal simple de los Nacidos vivos y su valor de significancia

Variables Independientes	Coeficientes	Error Estándar	Valor T	Valor p
Constante	-1.445	2.321	-0.622	0.535
Nacidos	0.002	0.000	7.664	0.000
R-SQ.(ADJ.)	**0.3059**	**SE**	**19.58**	

Realizado por: Arguello, Verónica, 2021

El valor p es menor que el nivel de significancia 0.05, e indica que la variable Nacidos es significativa.

Tabla 11-4: Análisis de varianza de la regresión lineal simple de la variable Nacidos vivos

Fuentes de Variación	Suma de cuadrados	GL	Cuadrados medios	Estadístico F	Valor p
Modelo	22525	1	22525.000	58.739	0.000
Error	49852	130	383.480		
Total	72377	131			

Realizado por: Arguello, Verónica, 2021

Al verificar si el modelo es significativo, se obtuvo un valor p menor que el nivel de significancia de 0.05, por lo tanto, se concluyó que si es significativo.

- **Análisis del número de muertes de menores de un año respecto al género.**

$$Y_i = 12.076 - 2.894\ Mujer$$

Tabla 12-4: Coeficientes de la regresión lineal simple de la variable Género

Variables Independientes	Coeficientes	Error Estándar	Valor T	Valor p
Constante	12.076	2.899	4.166	0.000
Género Mujer	-2.894	4.100	-0.706	0.482
R-SQ.(ADJ.)	**-0.0038**	**SE**	**23.55**	

Realizado por: Arguello, Verónica, 2021

El valor p es mayor que el nivel de significancia 0.05, eso indica que la variable Género no es significativa.

Tabla 13-4: Análisis de varianza de la regresión lineal simple de la variable Género

Fuentes de Variación	Suma de cuadrados	GL	Cuadrados medios	Estadístico F	Valor p
Modelo	276	1	276.00	0.498	0.482
Error	72100	130	554.62		
Total	72376	131			

Realizado por: Arguello, Verónica, 2021

El valor p es mayor que el nivel de significancia 0.05, eso indica que el modelo no es significativo.

Al realizar el análisis de regresión lineal se identificó que las variables significativas son: Provincia, Causas de Muerte, Muerte y Nacidos, por lo que se propuso un modelo más amplio que el de regresión múltiple.

$$
\begin{aligned}
Y_i = {} & -62.05 + 36.53\ ProvinciaBolívar + 36.58\ ProvinciaCañar + 9.858\ ProvinciaCarchi \\
& - 593.2\ ProvinciaChimborazo - 192.9\ ProvinciaCotopaxí \\
& + 164.1\ ProvinciaImbabura + 7.815\ ProvinciaLoja \\
& - 317.9\ ProvinciaPichincha \\
& - 31.54\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas \\
& - 59.62\ ProvinciaTungurahua + 6.003\ CMuerteCausas\ mal\ definidas \\
& + 5.124\ CMuerteInfluenza\ y\ Neumonía - 0.0701\ CMuerteMalformaciones \\
& + 4.991\ CMuerteOstrucción\ y\ Respiración \\
& + 4.491\ CMuerteResto\ de\ Causas + 0.015\ Nacidos \\
& - 0.007\ CMuerteCausas\ mal\ definidas : Nacidos \\
& - 0.006\ CMuerteInfluenza\ y\ neumonía : Nacidos \\
& - 0.003\ CMuerteMalformaciones : Nacidos \\
& - 0.006\ CMuerteOsbtrucción\ respiración : Nacidos \\
& - 0.006\ CMuerteRestos\ de\ causas : Nacidos \\
& + 0.007\ ProvinciaBolívar : Nacidos + 0.001\ ProvinciaCañar : Nacidos \\
& + 0.029\ ProvinciaCarchi : Nacidos + 0.147\ ProvinciaChimborazo : Nacidos \\
& + 0.058\ ProvinciaCotopaxí : Nacidos - 0.037\ ProvinciaImbabura : Nacidos \\
& + 0.004\ ProvinciaLoja : Nacidos + 0.007\ ProvinciaPichincha : Nacidos \\
& + 0.009\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas : Nacidos \\
& + 0.017\ ProvinciaTungurahua : Nacidos
\end{aligned}
$$

Tabla 14-4: Coeficientes de la regresión lineal múltiple con las variables significativas (Provincia, nacidos, causa de muerte) y sus interacciones.

Variables Independientes	Coeficientes	Error Estándar	Valor T	Valor p
Intercepto	-62.050	132.600	-0.468	0.640
ProvinciaBolívar	36.530	146.300	0.250	0.803
ProvinciaCañar	36.580	146.900	0.249	0.804
ProvinciaCarchi	9.858	148.300	0.066	0.947
ProvinciaChimborazo	-539.200	581	-0.928	0.356
ProvinciaCotopaxi	-192.900	266	-0.725	0.470
ProvinciaImbabura	164.100	439	0.374	0.709
ProvinciaLoja	7.815	141.200	0.055	0.956
ProvinciaPichincha	-317.900	168.7	-1.884	0.063
ProvinciaSanto Domingo de los Tsáchilas	-31.540	158.2	-0.199	0.842
ProvinciaTungurahua	-59.620	376.1	-0.159	0.874
CMuerteCausas mal definidas	6.003	1.995	3.009	0.003
CMuerteInfluenza y Neumonía	5.124	1.995	2.568	0.012
CMuerteMalformaciones	-0.070	1.995	-0.035	0.972
CMuerteObstrucción y Respiración	4.991	1.995	2.502	0.014
CMuerteResto de Causas	4.491	1.995	2.251	0.027
Nacidos	0.015	0.020	0.773	0.441
CMuerteCausas mal definidas: Nacidos	-0.007	0.000	-27.526	0.000
CMuerteInfluenza y neumonía: Nacidos	-0.006	0.000	-25.434	0.000
CMuerteMalformaciones: Nacidos	-0.003	0.000	-11.191	0.000
CMuerteObstrucción respiración: Nacidos	-0.006	0.000	-25.068	0.000
CMuerteRestos de causas: Nacidos	-0.006	0.000	-26.074	0.000
ProvinciaBolívar: Nacidos	0.007	0.046	0.144	0.886
ProvinciaCañar: Nacidos	0.001	0.034	0.018	0.986
ProvinciaCarchi: Nacidos	0.029	0.055	0.523	0.602
ProvinciaChimborazo: Nacidos	0.147	0.149	0.988	0.326
ProvinciaCotopaxi: Nacidos	0.058	0.064	0.905	0.367
ProvinciaImbabura: Nacidos	-0.037	0.114	-0.328	0.744
ProvinciaLoja: Nacidos	0.004	0.023	0.156	0.876

ProvinciaPichincha: Nacidos	0.007	0.021	0.364	0.716
ProvinciaSanto Domingo de los Tsáchilas: Nacidos	0.009	0.027	0.352	0.725
ProvinciaTungurahua: Nacidos	0.017	0.080	0.212	0.833
R-SQ.(ADJ.)	**0.957**	**SE**	**4.86**	

Realizado por: Arguello, Verónica, 2021

La variable significativa es: Causas de Muerte debido a que el valor p es menor que el nivel de significancia 0.05.

Tabla 15-4: Análisis de varianza de la regresión lineal múltiple con el modelo anterior

Fuentes de Variación	Suma de cuadrados	GL	Cuadrados medios	Estadístico F	Valor p
Modelo	70015	31	2258.550	95.628	0.000
Error	2361.800	100	23.620		
Total	72376.800	131			

Realizado por: Arguello, Verónica, 2021

Se comprobó que el modelo es adecuado debido a que el R^2 ajustado es de 0.957 y se verificó que es significativo.

Se utilizó el método paso a paso mediante la dirección backward para eliminar variables y mejorar el modelo, mediante un AIC de 427.66 que resultó ser:

$$\begin{aligned} Y_i = -109.300 &+ 83.150\ ProvinciaBolívar + 68.690\ ProvinciaCañar \\ &+ 85.200\ ProvinciaCarchi + 44.480\ ProvinciaChimborazo \\ &+ 47.120\ ProvinciaCotopaxí + 45.310\ ProvinciaImbabura \\ &+ 41.080\ ProvinciaLoja - 262.700\ ProvinciaPichincha \\ &+ 26.890\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas \\ &+ 32.420\ ProvinciaTungurahua + 6.003\ CMuerteCausas\ mal\ definidas \\ &+ 5.124\ CMuerteInfluenza\ y\ Neumonía - 0.070\ CMuerteMalformaciones \\ &+ 4.991\ CMuerteOstrucción\ y\ Respiración \\ &+ 4.491\ CMuerteResto\ de\ Causas + 0.023\ Nacidos \\ &- 0.007\ CMuerteCausas\ mal\ definidas: Nacidos \\ &- 0.006\ CMuerteInfluenza\ y\ neumonía: Nacidos \\ &- 0.003\ CMuerteMalformaciones: Nacidos \\ &- 0.006\ CMuerteOsbtrucción\ respiración: Nacidos \\ &- 0.006\ CMuerteRestos\ de\ causas: Nacidos \end{aligned}$$

Tabla 16-4: Coeficientes de la regresión lineal múltiple del modelo mejorado con las variables (Provincias, causas de muerte, nacidos) y sus interacciones excepto (provincia: nacidos)

Variables Independientes	Coeficientes	Error Estándar	Valor T	Valor p
Intercepto	-109.300	24.880	-4.393	0.000
ProvinciaBolívar	83.150	19.380	4.291	0.000
ProvinciaCañar	68.690	16.220	4.235	0.000
ProvinciaCarchi	85.200	20.020	4.256	0.000
ProvinciaChimborazo	44.480	10.620	4.188	0.000
ProvinciaCotopaxi	47.120	10.800	4.363	0.000
ProvinciaImbabura	45.310	11	4.119	0.000
ProvinciaLoja	41.080	9.950	4.129	0.000
ProvinciaPichincha	-262.700	63.400	-4.144	0.000
ProvinciaSanto Domingo de los Tsáchilas	26.890	6.966	3.860	0.000
ProvinciaTungurahua	32.420	8.106	4.000	0.000
CMuerteCausas mal definidas	6.003	1.924	3.120	0.002
CMuerteInfluenza y Neumonía	5.124	1.924	2.663	0.009
CMuerteMalformaciones	-0.070	1.924	-0.036	0.971
CMuerteObstrucción y Respiración	4.991	1.924	2.594	0.011
CMuerteResto de Causas	4.491	1.924	2.334	0.021
Nacidos	0.023	0.004	6.031	0.000
CMuerteCausas mal definidas: Nacidos	-0.007	0.000	-28.548	0.000
CMuerteInfluenza y neumonía: Nacidos	-0.006	0.000	-26.379	0.000
CMuertesMalformaciones: Nacidos	-0.003	0.000	-11.606	0.000
CMuertesObstrucción respiración: Nacidos	-0.006	0.000	-25.999	0.000
CMuertesRestos de causas: Nacidos	-0.006	0.000	-27.042	0.000
R-SQ.(ADJ.)	**0.966**	**SE**	**4.685**	

Realizado por: Arguello, Verónica, 2021

Al comparar el valor p de cada una de las variables con respecto al nivel de significancia de 0.05, se identificó que todas son significativas.

Tabla 17-4: Análisis de varianza de la regresión lineal múltiple del modelo mejorado

Fuentes de Variación	Suma de cuadrados	GL	Cuadrados medios	Estadístico F	Valor p
Modelo	69962.100	21	3331.530	151.759	0.000
Error	2414.800	110	21.950		
Total	72376.900	131			

Realizado por: Arguello, Verónica, 2021

Conforme a los resultados del valor p, se comprueba que el valor del modelo es significativo, además de acuerdo a la tabla 16-4 el R^2 ajustado es de 0.966, lo que indica que se ajusta bien a los datos. Sin embargo, se procede comprobar si del primer ejemplar se puede eliminar el predictor Provincia: Nacidos.

$$H_0: B_{Provincia:Nacidos} = 0$$
$$H_1: B_{Provincia:Nacidos} \neq 0$$

Modelo 1: Muertes ~ Provincia + Enfermedad + Nacidos + Nacidos * Enfermedad+ Nacidos * Provincia.

Modelo 2: Muertes ~ Provincia + Enfermedad + Nacidos + Nacidos * Enfermedad

Tabla 18-4: Análisis de varianza del modelo de regresión lineal múltiple con todas las variables significativas y el modelo mejorado

Fuentes de Variación	Res.Df	RSS	Suma de cuadrados	Estadístico F	Valor p
1	100	2361.8			
2	110	2414.8	-10	0.224	0.993
Total	210	4776.6			

Realizado por: Arguello, Verónica, 2021

El valor p de 0.993 indica que no se puede rechazar la hipótesis nula, concluyendo que el ejemplar 2 es la mejor propuesta, ahora se debe verificar que los supuestos se cumplan.

- Normalidad

$$H_o: Los\ \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

$$H_o: Los\ \varepsilon_i \neq N(0, \sigma_\varepsilon^2)$$

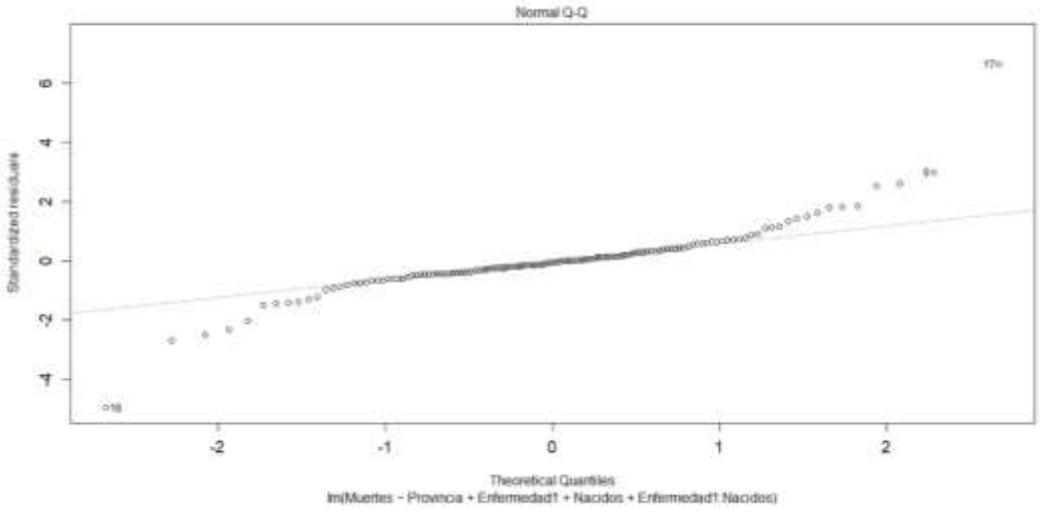


Figura 7-4: Gráfica de los residuales para probar la normalidad
Realizado por: Arguello, Verónica, 2021

Visualmente pareciera que se cumple el supuesto de normalidad, sin embargo, hay algunos datos atípicos que pueden estar influyendo, para ello se aplicó el test de Shapiro Wilks y Kolmogorov-Smirnov.

Tabla 19-4: Test de normalidad Shapiro-Wilk

Shapiro-Wilk normality test
W = 0.89821**,** **p-value** = 5.176e-08

Realizado por: Arguello, Verónica, 2021

Tabla 20-4: Test de normalidad Kolmogorov-Smirnov

One-sample Kolmogorov-Smirnov test
D = 0.12093**,** **p-value** = 0.0421
Alternative hypothesis: two-sided

Realizado por: Arguello, Verónica, 2021

Ambos test comprueban que no hay normalidad en los residuos, debido a que el valor p es menor que el nivel de significancia de 0.05, esto puede deberse a la presencia de datos atípicos.

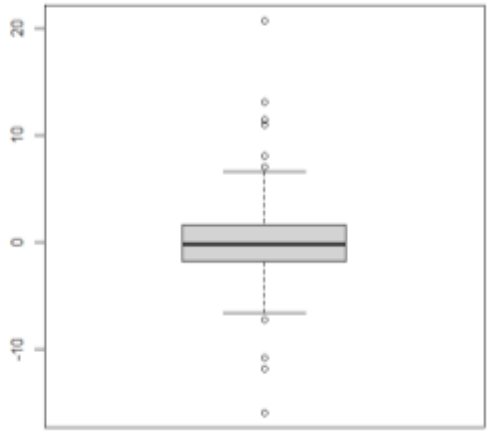


Figura 4-1: Diagrama de caja de los Residuos para identificar los datos atípicos
Realizado por: Arguello, Verónica, 2021

El gráfico de normalidad parecía indicar atípicos, el diagrama de cajas confirma la presencia de estos, de acuerdo a este resultado se aplicó el test de Jarque Vera para verificar si estos datos atípicos están afectando la normalidad y localizar los puntos.

$$H_o: La\ normalidad\ no\ se\ ve\ afectada\ por\ los\ datos\ atípicos$$
$$H_1: La\ normalidad\ si\ se\ ve\ afectada\ por\ los\ datos\ atípicos$$

Tabla 21-4: Test de Jarque Bera

Jarque Bera Test		
X-squared = 175.38,	**df** = 2**,**	**p-value** < 2.2e-16

Realizado por: Arguello, Verónica, 2021

Al obtener un valor p menor que el nivel de significancia de 0.05 no se puede aceptar la hipótesis nula y se concluye que la normalidad si se ve afectada por los atípicos.
Los datos atípicos identificados son:

Tabla 22-4: Tets para determinar datos atípicos

	Rstudent	unadjusted p-value	Bonferroni p
17	8.490	1.1517e-13	1.5202e-11
18	-5.612	1.5303e-07	2.0201e-05

Realizado por: Arguello, Verónica, 2021

Estos datos pertenecen a la provincia de Pichincha y al tener este tipo de estudio no se consideró eliminarlos.

- Varianza constante

Para verificar la homocedasticidad, se realizó la siguiente gráfica

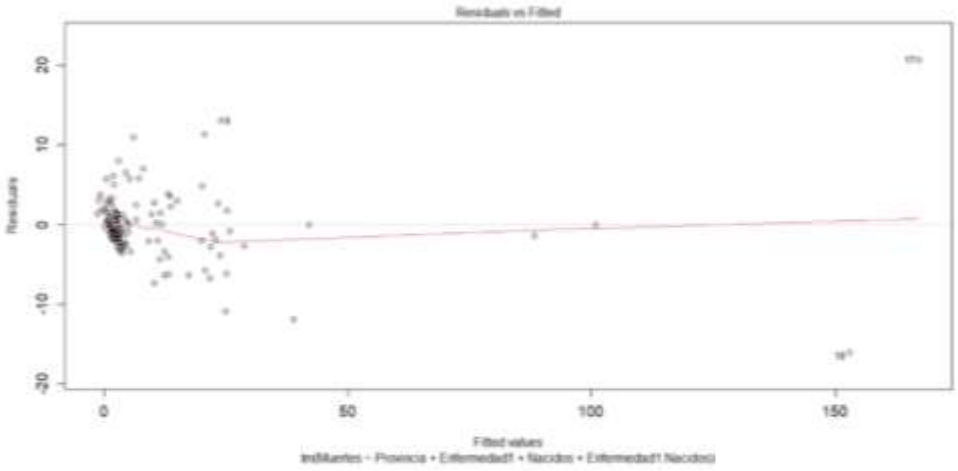


Figura 9-4: Gráfica de los residuos vs los predichos.
Realizado por: Arguello, Verónica, 2021

Conforme a la gráfica 9-4, los datos no tienen una varianza constante a lo largo del tiempo, ahora se aplicó el test Breusch-Pagan para corroborar lo dicho en la imagen.

$$\boldsymbol{H_o}: \boldsymbol{V(\varepsilon_i) = \sigma_\varepsilon^2}$$
$$\boldsymbol{H_1}: \boldsymbol{V(\varepsilon_i) \neq \sigma_\varepsilon^2}$$

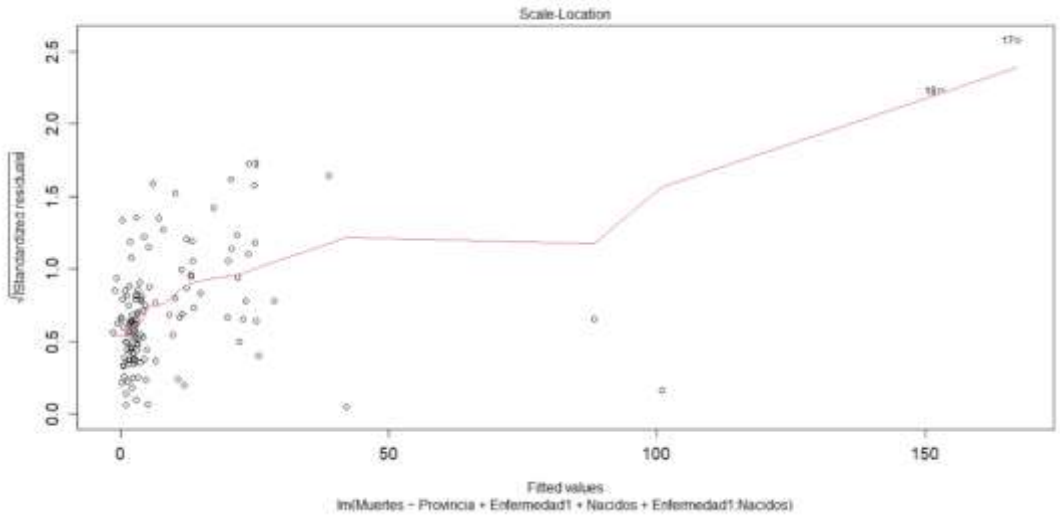
Tabla 23-4: Test de Breusch-Pagan

Studentized Breusch-Pagan test		
BP = 95.802**,**	**df** = 21**,**	**p-value** = 1.582e-11

Realizado por: Arguello, Verónica, 2021

Como el valor p es menor que el nivel de significancia de 0.05, se rechaza la hipótesis nula y se concluye que los errores no tienen una varianza constante.

- Independencia



$$H_o: Los\ \varepsilon_i, \varepsilon_j\ son\ independientes\ \ para\ i \neq j.$$

$$H_1: Los\ \varepsilon_i, \varepsilon_j\ son\ dependientes\ \ para\ i \neq j.$$

Figura 9-4: Gráfica de los predichos vs la normalidad de los residuos
Realizado por: Arguello, Verónica, 2021

Tabla 24-4: Test de Durbin Watson

Durbin-Watson test
DW = 2.3198**,** **p-value =** 0.747
alternative hypothesis: true autocorrelation is greater than 0

Realizado por: Arguello, Verónica, 2021

El valor p es mayor que el nivel de significancia, por lo tanto, no hay evidencia para rechazar Ho y se concluye que no hay independencia en los errores.

Aunque el modelo es capaz de explicar el 96.6% de la variabilidad observada en las muertes de menores de un año (R^2: 0.966, R^2-Adjusted: 0.9603). El test F muestra que es significativo (p-value: 000). No satisfacen los supuestos, por lo tanto, no es apto para hacer predicciones, Debido a que no cumple con los supuestos, se aplicó GLM.

4.3. Modelos lineales generalizados

4.3.1. Modelo Lineal Generalizado (Distribución Poisson)

Para realizar esta modelación se utilizó la matriz conformada por 132 datos y el modelo es el siguiente.

$$Y_i = e^{0} * e^{9.467\ ProvinciaBolívar} * e^{5.154\ ProvinciaCañar} * e^{8.248\ ProvinciaCarchi} * e^{2.884\ ProvinciaChimborazo} * e^{3.716\ ProvinciaCotopaxí} * e^{2.640\ ProvinciaImbabura} * e^{2.499\ ProvinciaLoja} * e^{0.002\ ProvinciaPichincha} * e^{1.611\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas} * e^{1.925\ ProvinciaTungurahua} * e^{0.059\ CM\ Causas\ mal\ definidas} * e^{0.118\ CM\ Influenza\ y\ Neumonía} * e^{0.543\ CM\ Malformaciones} * e^{0.129\ CM\ Obstrucción\ respiración} * e^{0.073\ CM\ Resros\ de\ causas}$$

Tabla 25-4: Coeficientes de la regresión de Poisson y su valor de significancia

Variables Independientes	Estimaciones	Error Estándar	Valor Z	Valor p
Intercepto	-7.702	0.838	-9.186	0.000
ProvinciaBolívar	2.248	0.671	3.351	0.001
ProvinciaCañar	1.640	0.565	2.902	0.004
ProvinciaCarchi	2.110	0.698	3.022	0.003
ProvinciaChimborazo	1.059	0.377	2.810	0.005
ProvinciaCotopaxi	1.313	0.380	3.458	0.001
ProvinciaImbabura	0.971	0.391	2.479	0.013
ProvinciaLoja	0.916	0.355	2.579	0.010
ProvinciaPichincha	-6.343	2.136	-2.970	0.003
ProvinciaSanto Domingo de los Tsáchilas	0.477	0.262	1.818	0.069
ProvinciaTungurahua	0.655	0.298	2.200	0.028
CMuerteCausas mal definidas	-2.827	0.157	-18.017	0.000
CmuerteInfluenza y Neumonía	-2.135	0.114	-18.724	0.000
CmuerteMalformaciones	-0.610	0.062	-9.777	0.000
CmuerteObstrucción y Respiración	-2.046	0.110	-18.669	0.000
CmuerteResto de Causas	-2.618	0.142	-18.406	0.000
Nacidos	0.000	0.000	3.052	0.002

Realizado por: Arguello, Verónica, 2021

En la tabla 25-4 se visualiza las estimaciones de los coeficientes del modelo de Poisson y se verificó que todas las variables son significativas debido a que tiene el valor p menor que el nivel de significancia de 0.05.

Tabla 26-4: Distribución de Poisson a partir de la estimación exponencial

Variables Independientes	Estimaciones	Exp(Estimación)	Interpretación
Intercepto	-7.702	0.000	
ProvinciaBolívar	2.248	9.467	La media de muertes en menores de un año en Bolívar es 90.5% menos que la media de Azuay.
ProvinciaCañar	1.640	5.154	La media de muertes en menores de un año en Cañar es 94.8% menos que la media de Azuay.
ProvinciaCarchi	2.110	8.248	La media de muertes en menores de un año en Carchi es 91.8% menos que la media de Azuay.
ProvinciaChimborazo	1.059	2.884	La media de muertes en menores de un año en Chimborazo es 97.1% menos que la media de Azuay.
ProvinciaCotopaxi	1.313	3.716	La media de muertes en menores de un año en Cotopaxi es 96.3% menos que la media de Azuay.
ProvinciaImbabura	0.971	2.640	La media de muertes en menores de un año en Imbabura es 97.4% menos que la media de Azuay.
ProvinciaLoja	0.916	2.499	La media de muertes en menores de un año Loja es 97.5% menos que la media de Azuay.
ProvinciaPichincha	-6.343	0.002	La media de muertes en menores de un año en Pichincha es 99.998% más que la media de muertes de Azuay.
ProvinciaSanto Domingo de los Tsáchilas	0.477	1.611	La media de muertes en menores de un año en Santo Domingo de los Tsáchilas es 98.4% menos que la media de Azuay.
ProvinciaTungurahua	0.655	1.925	La media de muertes en menores de un año en Tungurahua es 98.1% menos que la media de Azuay.
CmuerteCausas mal definidas	-2.827	0.059	La media de muertes en menores de un año por muertes mal definidas aumenta 6% que la media de muertes por afecciones en el periodo prenatal.
CmuerteInfluenza y Neumonía	-2.135	0.118	La media de muertes en menores de un año por influenza y neumonía aumenta 11.8% que la media de muertes por afecciones en el periodo prenatal.
CmuerteMalformaciones	-0.610	0.543	La media de muertes en menores de un año por malformaciones aumenta 54.3% que la media de muertes por afecciones en el periodo prenatal.
CmuerteObstrucción y Respiración	-2.046	0.129	La media de muertes en menores de un año por obstrucción y respiración aumenta 12.9% que la media de muertes por afecciones en el periodo prenatal.
CmuerteResto de Causas	-2.618	0.073	La media de muertes en menores de un año por otras causas de muerte aumenta 7.3% que la media de muertes por afecciones en el periodo prenatal.
Nacidos	0.000	1.000	

Realizado por: Arguello, Verónica, 2021

En la tabla 26-4 representa el modelo a partir de la estimación exponencial en donde se realiza una comparación de medias entre las muertes en menores de un año de las regiones objeto de estudio.

Tabla 27-4: Análisis de Devianza para la variable muertes

	Grados de libertad	Devianza	Devianza media
Modelo Nulo	131	1676.990	12.801
Modelo Residual	115	208.930	1.817
Total Corregido	210	4776.6	

Realizado por: Arguello, Verónica, 2021

El análisis de devianza, reportan que el modelo es estadísticamente significativo y concluyó lo siguiente: El modelo de regresión de Poisson tiene un pseudo-R^2 de 0,858 en comparación con el de regresión múltiple que tenía un R^2 de 0.966. la hipótesis referente al pseudo-R^2, es que H_0: $R^2 = 0 \ \ vs \ \ H_1$: $R^2 \neq 0$, y H_0 es rechazada lo que involucra que los datos observados se ajustan al modelo de regresión de Poisson, además se cumple con la equidispersión ya que se obtuvo un valor de $\phi = 1.002$.

4.3.2. Modelo lineal Generalizado (Distribución Multinomial).

Para la elaboración del modelo multinomial se trabajó con seis grupos que representan las causas de las muertes en menores de un año. Esta variable es cualitativa de tipo nominal con seis niveles.

Tabla 28-4: Diferentes categorías de la variable causas de muerte

N°	Enfermedad o causas de Muerte
1	Afecciones del periodo prenatal
2	Malformaciones
3	Obstrucción respiratoria
4	Influenza y neumonía
5	Resto de causas
6	Causas mal definidas

Realizado por: Arguello, Verónica, 2021

La regresión logística multinomial consideró a la variable Causas de Muerte como dependiente puesto que es de tipo categórica nominal con seis categorías (politómica).

$$\left.\begin{aligned} ln\left(\frac{p_2}{p_1}\right) = Z_1 = \beta_{01} + \beta_{11} * X_1 + \beta_{21} * X_2 + \beta_{31} * X_3 + \beta_{41} * X_4 + \beta_{51} * X_5 \\ ln\left(\frac{p_3}{p_1}\right) = Z_2 = \beta_{02} + \beta_{12} * X_1 + \beta_{22} * X_2 + \beta_{32} * X_3 + \beta_{42} * X_4 + \beta_{52} * X_5 \\ ln\left(\frac{p_4}{p_1}\right) = Z_3 = \beta_{03} + \beta_{13} * X_1 + \beta_{23} * X_2 + \beta_{33} * X_3 + \beta_{43} * X_4 + \beta_{53} * X_5 \\ ln\left(\frac{p_5}{p_1}\right) = Z_4 = \beta_{04} + \beta_{14} * X_1 + \beta_{24} * X_2 + \beta_{34} * X_3 + \beta_{44} * X_4 + \beta_{54} * X_5 \\ ln\left(\frac{p_6}{p_1}\right) = Z_5 = \beta_{05} + \beta_{15} * X_1 + \beta_{25} * X_2 + \beta_{35} * X_3 + \beta_{45} * X_4 + \beta_{55} * X_5 \end{aligned}\right\}$$

Tabla 29-4: Coeficientes de la regresión multinomial

Coeficientes	Causas de Muertes					Error Estándar				
	2	3	4	5	6	2	3	4	5	6
Intercepto	6.416	273.414	268.267	276.106	276.435	1.195	5.146	5.224	5.418	6.141
ProvinciaBolívar	137.739	-22.544	-34.244	-3.353	-40.407	0.072	0.716	2.808	0.749	3.037
ProvinciaCañar	136.072	-34.015	6.770	16.080	0.404	0.001	0.001	0.865	0.759	0.949
ProvinciaCarchi	106.179	-147.043	-144.566	-139.035	-334.308	9.999	13.963	8.918	1.042	0.000
ProvinciaChimborazo	339.470	78.103	73.575	100.047	71.594	3.831	2.857	3.062	0.328	3.738
ProvinciaCotopaxi	157.587	-205.378	-56.986	-59.665	-56.999	21.711	0.001	8.269	8.314	8.745
ProvinciaImbabura	315.020	116.215	110.569	119.608	-293.780	1.140	4.625	7.648	6.828	0.000
ProvinciaLoja	257.586	43.737	43.556	31.939	38.658	0.474	3.370	3.731	5.664	4.366
ProvinciaPichincha	271.404	66.151	68.969	70.303	71.864	0.000	2.175	2.700	2.573	3.859
ProvinciaS. D. de los Tsáchilas	349.674	136.397	102.299	129.930	106.316	16.623	4.105	15.567	4.136	2.873
ProvinciaTungurahua	112.662	50.201	59.704	54.558	105.594	0.851	2.923	3.015	3.219	6.222
Muertes	-0.202	-12.349	-10.578	-13.463	-13.857	0.037	0.956	0.498	1.444	2.380
ProvinciaBolívar: Muertes	-24.325	-54.855	-40.015	-74.906	-37.883	0.223	0.716	1.562	0.749	1.204
ProvinciaCañar: Muertes	-16.718	-27.212	-49.157	-62.856	-47.015	0.005	0.005	0.865	0.759	0.949
ProvinciaCarchi: Muertes	-26.984	-25.073	-24.189	-34.083	-5.433	7.924	13.791	8.030	1.042	0.000
ProvinciaChimborazo: Muertes	-26.792	-16.029	-15.992	-27.977	-13.471	0.762	0.965	0.755	0.657	1.749
ProvinciaCotopaxi: Muertes	-15.683	7.689	-14.054	-13.772	-15.790	4.889	0.016	2.399	2.655	3.227
ProvinciaImbabura: Muertes	-23.516	-24.619	-22.508	-28.934	9.700	14.823	2.961	3.642	5.554	0.000
ProvinciaLoja: Muertes	-15.524	-16.568	-15.684	-11.713	-13.760	7.586	1.442	1.335	2.097	1.976
ProvinciaPichincha: Muertes	-2.207	8.832	7.189	9.432	9.535	0.000	0.803	0.834	1.116	1.763

ProvinciaS. D. de los Tsáchilas : Muertes	-19.003	-22.494	-10.334	-18.951	-10.102	1.274	7.222	2.249	7.181	2.635
ProvinciaTungurahua: Muertes	-9.710	-21.586	-24.240	-22.097	-43.223	7.664	3.632	3.614	3.730	12.530

Realizado por: Arguello, Verónica, 2021

De acuerdo a la tabla 29-4 se identificó los coeficientes del modelo de regresión multinomial.

$$\text{Ln}\left(\frac{P(Enfermedad = Malformación)}{P(Enfermedad = Afecciones\ \ periodo\ \ prenatal)}\right)$$
$$= 6.416 + 137.739 * Provincia\ Bolívar\ + 136.072\ ProvinciaCañar$$
$$+ 106.179 ProvinciaCarchi + 339.470\ ProvinciaChimborazo$$
$$+ 157.587\ ProvinciaCotopaxí + 315.020\ ProvinciaImbabura$$
$$+ 257.586\ ProvinciaLoja + 271.404\ ProvinciaPichincha$$
$$+ 349.674\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas$$
$$+ 112.662\ ProvinciaTungurahua - 0.020 Muertes$$
$$- 24.325\ ProvinciaBolívar: Muertes - 16.718\ ProvinciaCañar: Muertes$$
$$- 26.984\ ProvinciaCarchi: Muertes$$
$$- 26.792\ ProvinciaChimborazo: Muertes$$
$$- 15.683\ ProvinciaCotopaxí: Muertes$$
$$- 23.516\ ProvinciaImbabura: Muertes$$
$$- 15.524\ ProvinciaLoja: Muertes - 2.207\ ProvinciaPichincha: Muertes$$
$$- 19.003\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas: Muertes$$
$$- 9.710\ ProvinciaTungurahua: M\text{uertes}$$

$$\ln\left(\frac{P(Enfermedad = Obstucción\ respiratoria)}{P(Enfermedad = Afecciones\ \ periodo\ \ prenatal)}\right)$$
$$= 273.414 - 22.544 * Provincia\ Bolívar - 34.015\ ProvinciaCañar$$
$$- 147.043\ ProvinciaCarchi + 78.103\ ProvinciaChimborazo$$
$$- 205.378\ ProvinciaCotopaxí + 116.215\ ProvinciaImbabura$$
$$+ 43.737\ ProvinciaLoja + 66.151\ ProvinciaPichincha$$
$$+ 136.397\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas$$
$$+ 50.201\ ProvinciaTungurahua - 12.349\ Muertes$$
$$- 54.8555\ ProvinciaBolívar: Muertes$$
$$- 27.212\ ProvinciaCañar: Muertes - 25.073\ ProvinciaCarchi: Muertes$$
$$- 16.029\ ProvinciaChimborazo: Muertes$$
$$+ 7.689\ ProvinciaCotopaxí: Muertes$$
$$- 24.619\ ProvinciaImbabura: Muertes$$
$$- 16.568\ ProvinciaLoja: Muertes + 8.832\ ProvinciaPichincha: Muertes$$
$$- 22.494\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas: Muertes$$
$$- 21.586\ ProvinciaTungurahua: M\text{uertes}$$

$$\ln\left(\frac{P(Enfermedad = Influenza\ y\ neumonía)}{P(Enfermedad = Afecciones\ \ periodo\ \ prenatal)}\right)$$
$$= 268.267 - 34.244 * Provincia\ Bolívar\ + 6.770\ ProvinciaCañar$$
$$- 144.566\ ProvinciaCarchi + 73.575\ ProvinciaChimborazo$$
$$- 56.986\ ProvinciaCotopaxí + 110.569\ ProvinciaImbabura$$
$$+ 43.556\ ProvinciaLoja + 68.969\ ProvinciaPichincha$$
$$+ 102.299\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas$$
$$+ 59.704\ ProvinciaTungurahua - 10.578Muertes$$
$$- 40.015\ ProvinciaBolívar: Muertes - 49.157\ ProvinciaCañar: Muertes$$
$$- 24.189\ ProvinciaCarchi: Muertes$$
$$- 15.992\ ProvinciaChimborazo: Muertes$$
$$- 14.054\ ProvinciaCotopaxí: Muertes$$
$$- 22.508\ ProvinciaImbabura: Muertes$$
$$- 15.684\ ProvinciaLoja: Muertes + 7.189\ ProvinciaPichincha: Muertes$$
$$- 10.334\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas: Muertes$$
$$- 24.240\ ProvinciaTungurahua: M\text{uertes}$$

$$\ln\left(\frac{P(Enfermedad = Otras\ causas)}{P(Enfermedad = Afecciones\ \ periodo\ \ prenatal)}\right)$$
$$= 276.106 - 3.353 * Provincia\ Bolívar\ + 16.080\ ProvinciaCañar$$
$$- 139.035\ ProvinciaCarchi + 100.047\ ProvinciaChimborazo$$
$$- 59.665\ ProvinciaCotopaxí + 119.608\ ProvinciaImbabura$$
$$+ 31.939\ ProvinciaLoja + 70.303\ ProvinciaPichincha$$
$$+ 129.930\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas$$
$$+ 54.558\ ProvinciaTungurahua - 13.463Muertes$$
$$- 74.906\ ProvinciaBolívar: Muertes - 62.856\ ProvinciaCañar: Muertes$$
$$- 34.083\ ProvinciaCarchi: Muertes$$
$$- 27.977\ ProvinciaChimborazo: Muertes$$
$$- 13.772\ ProvinciaCotopaxí: Muertes$$
$$- 28.934\ ProvinciaImbabura: Muertes$$
$$- 11.713\ ProvinciaLoja: Muertes + 9.432\ ProvinciaPichincha: Muertes$$
$$- 18.951\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas: Muertes$$
$$- 22.097\ ProvinciaTungurahua: M\text{uertes}$$

$$\ln\left(\frac{P(Enfermedad = Causas\ mal\ definidas)}{P(Enfermedad = Afecciones\ \ periodo\ \ prenatal)}\right)$$
$$= 276.435 - 40.407\ Provincia\ Bolívar\ + 0.404\ ProvinciaCañar$$
$$- 334.308\ ProvinciaCarchi + 71.594\ ProvinciaChimborazo$$
$$- 56.999\ ProvinciaCotopaxí - 293.780\ ProvinciaImbabura$$
$$+ 38.658\ ProvinciaLoja + 71.864\ ProvinciaPichincha$$
$$+ 106.316\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas$$
$$+ 105.594\ ProvinciaTungurahua - 13.857Muertes$$
$$- 37.883\ ProvinciaBolívar: Muertes - 47.015\ ProvinciaCañar: Muertes$$
$$- 5.433\ ProvinciaCarchi: Muertes$$
$$- 13.471\ ProvinciaChimborazo: Muertes$$
$$- 15.790\ ProvinciaCotopaxí: Muertes$$
$$+ 9.700\ ProvinciaImbabura: Muertes - 13.760\ ProvinciaLoja: Muertes$$
$$+ 9.535\ ProvinciaPichincha: Muertes$$
$$- 10.102\ ProvinciaSanto\ Domingo\ de\ los\ Tsáchilas: Muertes$$
$$- 43.223\ ProvinciaTungurahua: Muertes$$

Un aumento en la probabilidad de que el bebé muera por malformaciones frente a afecciones en periodo prenatal se ve influenciado por la variable provincia Santo Domingo de los Tsáchilas.

Hay un aumento en la probabilidad de que el nacido muera por obstrucción respiratoria frente a afecciones en periodo prenatal debido a la influencia de la variable provincia Santo Domingo de los Tsáchilas.

Se evidencia un aumento en la probabilidad de que el bebé muera por influenza y neumonía frente a afecciones en periodo prenatal debido a la influencia de la variable provincia Imbabura.

Se visualiza un aumento en la probabilidad de que el recién nacido muera por otras causas frente a afecciones en periodo prenatal debido a la influencia de la variable provincia Santo Domingo de los Tsáchilas.

Por último, presenta un aumento en la probabilidad de que el bebé muera por causas mal definidas frente a afecciones en periodo prenatal debido a la influencia de la variable provincia Santo Domingo de los Tsáchilas.

4.4. Simulaciones comparativas

Tabla 30-4: Simulaciones que permiten verificar cuál de los modelos se ajusta mejor a los datos

Simulaciones	Modelo Lineal R^2	Modelo de Poisson R^2
Primera	0.120	0.207
Segunda	0.251	0.383
Tercera	0.230	0.360
Cuarta	0.206	0.440
Quinta	0.128	0.315
Sexta	0.285	0.240
Séptima	0.132	0.330
Octava	0.139	0.253
Novena	0.233	0.411
Décima	0.199	0.299

Realizado por: Arguello, Verónica, 2021

De acuerdo con la tabla 30-4 se obtuvo los valores de R^2 de los modelos de regresión lineal múltiple y de regresión de Poisson con 10 simulación con una base de datos creada para ambos modelos, obteniendo que los valores de R^2 de los modelos de regresión lineal múltiple son muy bajos y que los valores para el modelo de Poisson tienen valores más altos en R^2.

CONCLUSIONES

a) Mediante el análisis exploratorio se identificó las características de las variables, obteniendo como resultado 2 cuantitativas (Muertes y Nacidos) y 3 cualitativas nominales (Género, Causas de Muerte, Provincias), también se obtuvo que a nivel provincial, el comportamiento de la tasa de mortalidad no es homogéneo, observándose fuertes diferencias entre las entidades en cuanto a la mortalidad durante el primer año de vida y Bolívar es la que lidera, sin embargo hay un ascenso de número de muertes en Pichincha, y en general en hombres, además se concluyó que la principal causa de muerte son las afecciones por periodo prenatal.

b) La regresión lineal simple permitió identificar los factores que han contribuido en la mortalidad infantil, así como las variables que más explican las diferencias interprovinciales de dicho indicador y se determinó que la única no significativa es Género. De igual manera se realizó un análisis de regresión lineal múltiple para ajustar el mejor modelo, obteniendo un AIC de 427.66 y un R^2 de 0.966 sin embrago no cumplió con los supuestos por lo tanto no es adecuado para realizar predicciones.

c) Al no cumplir los supuestos la regresión lineal múltiple queda descartada, y se propone los modelos lineales generalizados, la regresión de Poisson es la que mejor se ajusta a los datos, además tiene un R^2 de 0.858 y un AIC de 650.72.

d) Para terminar, se realiza simulaciones comparativas para poder verificar si el estudio realizado se puede utilizar a nivel de todo el país y se concluye que el modelo puede utilizar datos de mortalidad infantil de menores de un año de todo el Ecuador con regresión de Poisson.

RECOMENDACIONES

a) Respecto a la base de datos, poder contar con la información necesaria para la investigación y antes de diseñarla, determinar su finalidad, organizar y categorizarla según las claves, definir las relaciones entre tablas y elementos, lo cual ayudó a configurar una metodología de trabajo eficiente.

b) Al tener variables cualitativas se hace difícil encontrar la relación entre ellas, y no poder aplicar un coeficiente de correlación e identificar las que tienen mayor relación que otras, lo que se recomienda es hacer regresión lineal múltiple e ir descartando las que no aportan nada al modelo.

c) Los GLM, son una extensión de los modelos lineales, para modelar variables que pertenezcan a la familia de distribución de probabilidad exponencial esta familia contiene las distribuciones: normal, binomial, Poisson y binomial negativa, entre otras sin embargo, se tiene que poner mayor atención al tipo de datos que se va a manejar para ver que método va mejor y referente al análisis de regresión de Poisson antes de interpretar y concluir acerca de la hipótesis de interés , es necesario verificar si el modelo propuesto es el indicado para la modelación de los datos; uno de los supuestos necesarios es la equidispersión.

d) Seguir utilizando más simulaciones comparativas debido a que son rápidas de realizar, permiten tomar decisiones adecuadas, razonadas, y en el tiempo que se requieren.

BIBLIOGRAFÍA

Alvarado, C., & Villamar, C. (2010). *Aplicación del modelo anfis a la sintetizacion de notas musicales y señales de voz*. 7. http://www.dspace.espol.edu.ec/handle/123456789/16119

Amran, M. A. N., Bakar, A. A., Jalil, M. H. A., Wahyu, M. U., & Gani, A. F. H. A. (2020). Simulation and modeling of two-level DC/DC boost converter using ARX, ARMAX, and OE model structures. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(3), 1172–1179. https://doi.org/10.11591/ijeecs.v18.i3.pp1172-1179

Arguello, A. (2020). Modelos lineales generalizados en el análisis de las defunciones fetales de litoral ecuatoriano.

Ding, F., Pan, J., Alsaedi, A., & Hayat, T. (2019). Gradient-based iterative parameter estimation algorithms for dynamical systems from observation data. *Mathematics*, *7*(5). https://doi.org/10.3390/math7050428

Eykhoff, P. (1978). System identification; Advances and case studies. In *IEEE Transactions on Automatic Control* (Vol. 23, Issue 4). https://doi.org/10.1109/tac.1978.1101781

Gutiérrez, H., & Vara, R. (2008). Análisis y diseño de experimentos (Vol. 2).

Hsia, T. C. (1977). *System Identification Least-Squares Methods* (pp. 1–168). Lexington Books.

Hsu, H. P. (2015). *Señales y Sistemas* (Segunda Ed). McGrawHill.

Internacional Centre for Mechanical Sciencie. (1988). Application of System Identification in Engineering. In H. Natke (Ed.), *Application of System Identification in Engineering*. Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-7091-2628-8

Isermann, R., & Munchhof, M. (2019). Identification of Dynamic Systems an Introduction with Applications. In *Journal of Chemical Information and Modelling* (Vol. 53, Issue 9). Springer Heidelberg Dordrecht London New York.

Kamen, E. W., & S., H. B. (2008). *Fundamentos de Señales y Sistemas usando la Web y MATLAB* (Tercera Ed). Prentice Hall.

Kendall, K. E., & Kendall, J. E. (1993). *Análisis y Diseño de Sistemas* (Tercera Ed).

Kunusch, C. (2003). Identificacion De Sistemas Dinamicos. In *Universidad Nacional de la Plata*. Universidad Nacional de la Plata.

Lathi, B. P. (2005). Linear Systems and Signals. In *Linear Systems and Signals*.

Leondes, C. T. (1987). Control and Dynamic System: Advances in Theory and Applications. In *System Identification and Adaptive Control* (Vol. 25, Issue P1). Academic Press.

Little, T (2013). The Oxford Handbook of Quantitative Methods (Vol. 2, Statistical Analysis).

Ljung, L. (1987). *System identification: Theory for the User* (Primera). Prentice Hall.

Ljung, L. (1997). System Identification. *Prencite Hall Lntemational*. http://www.diva-portal.org/smash/get/diva2:316967/FULLTEXT01.pdf%0Ahttp://scholar.google.com/scholar?q=system+identification+1989+soderstrom%2C+t&btnG=&hl=en&as_sdt=0%2C5#1

Lopez, E., & Ruiz, M. (2011). Análisis de datos con el modelo lineal generalizado. Una aplicación con R.

Martínez, A. (2001). Modelos lineales generalizados. Universidad Carlos III de Madrid

Nelles, O. (2001). Nonlinear System Identification. In *Nonlinear System Identification*. https://doi.org/10.1007/978-3-662-04323-3

Oppenheim, A. V., & Willsky, A. S. (1997). *Señales y Sistemas* (P. Roig (ed.); Segunda Ed). Prentice Hall.

Pintelon, R., & Schoukens, J. (2001). *System Identification - A Frequency Domain Approach*. IEEE Press.

Proakis, J. G., & Manolakis, D. G. (2007). *Tratamiento digital de señales* (Cuarta Edi).

Prentice Hall. https://books.google.com.mx/books?id=8rhdNQAACAAJ&dq=tratamiento+digital+de+señales+4+ed+proakis&hl=en&sa=X&ved=0ahUKEwj5m9Wi5-DeAhURCawKHZIuCg8Q6AEIKTAA

Rachad, S., Nsiri, B., & Bensassi, B. (2015). System Identification of Inventory System Using ARX and ARMAX Models. *International Journal of Control and Automation*, *8*(12), 283–294. https://doi.org/10.14257/ijca.2015.8.12.26

Roback, P., & Legler, J. (2021). Beyond Multiple Linear Regression_ Applied Generalized Linear Models And Multilevel Models in R.

Rodriguez, G. (2016). Generalized Lineal Models.

Salah-Eddine, M., Sadki, S., & Bensassi, B. (2020). Microcontroller based data acquisition and system identification of a DC servo motor using ARX, ARMAX, OE, and BJ models. *Advances in Science, Technology and Engineering Systems*, *5*(6), 507–513. https://doi.org/10.25046/aj050660

Sanandaji, B. M., Vincent, T. L., Wakin, M. B., Tóth, R., & Poolla, K. (2011). Compressive System Identification of LTI and LTV ARX Models. *IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, *50*, 791–798. https://doi.org/978-1-61284-801-3/11

Walpole, R., Myers, R., & Myers, S. (2007). *Probabilidad y estadística para ingeniería y ciencias* (Octava Edi). PEARSON EDUCACIÓN, S.A.

Wang, L., & Garnier, H. (2014). *System Identification, Environmental Modelling, and Control System Design* (Liuping Wang & G. Hugues (eds.)). Springer London Dordrecht Heidelberg New York. https://doi.org/10.1007/978-0-85729-974-1

Weyer, E. (2000). Finite sample properties of system identification of ARX models under mixing conditions. *Automatica*, *36*(9), 1291–1299. https://doi.org/10.1016/S0005-1098(00)00039-X

Yankov, K. (2013). Data Structure of Models in System Identification. *Proceedings of the International Conference on Information Technologies*, *September*, 312–328.

ANEXOS

- **ANEXO A. Análisis exploratorio de las variables**

  Para realizar el mapa se necesita los archivos de las provincias en shp y los datos de longitud y latitud de las provincias de estudio en shp.

```
> shp2=st_read(dsn="D:/R/SHP/nxprovincias.shp")
> shp3 = st_read(dsn="D:/R/Volcanes/Prov.shp")


> g <- guide_legend("Muerte")
> ggShp5 = ggplot(data = shp2) +
+   geom_sf(color="black",fill="mistyrose")+
+   labs(x="Longitude",y="Latitude")+
+   theme_bw()+
+   theme(legend.title = element_text(colour="black", size=10, face="bold"), # adjust legend
title
+         legend.position = c(1.01, 0.5), # relative position of legend
+         plot.margin = unit(c(t=0, r=1, b=0, l=1), unit="cm"))+
+   geom_sf(data=shp3,pch=19,col="red",aes(shape=Muerte))+
+   guides(fill=g,size=g, alpha=g)+
+   theme(legend.position = c(0.95, 0.2))+
+   scale_size(trans = "reverse",
+              range=c(0,2),
+              guide = guide_legend(override.aes=list(fill="black")))+
+   scale_fill_gradientn(colours=c("yellow","green","blue","red","black","brown"))+
+   scale_alpha_continuous(guide = guide_legend(override.aes = list(fill = "black")))+
+   theme(axis.text.x=element_text(face="bold",color="black",size=10),
+         axis.text.y=element_text(face="bold",color="black",size=10),
+         axis.title.x=element_text(face="bold",color="black",size=12),
+         axis.title.y=element_text(face="bold",color="black",size=12))
```

- **ANEXO B. Tabla de contingencia entre las variables Provincia y Género respecto al número de muertes en menores de un año.**

```
> xtabs(data$Muertes~data$Provincia+data$Genero)
```

- **ANEXO C. Gráfica de cajas de las variables provincia y muertes en menores de un año.**

```
> p1<-ggplot(data, aes(Provincia,Muertes),fill=Provincia)+
+geom_boxplot(notch=F,color=c("blue4","red4","green4","yellow4","coral4","mediumpurp
le4","mistyrose4","gray24","aquamarine4","lightpink4","lightskyblue4"),fill=c("royalblue","
red1","green","yellow","coral","mediumpurple1","mistyrose","gray67","aquamarine","light
pink","lightskyblue"))+
+   geom_point()+
+   xlab("Provincias")+
+   theme(panel.background=element_rect(fill="white",colour="black"),
+         axis.text.x=element_text(face="bold",color="black",size=10,angle = 90, vjust = 0.5,
hjust=1),
+         axis.text.y=element_text(face="bold",color="black",size=10),
+         axis.title.x=element_text(face="bold",color="black",size=16),
+         axis.title.y=element_text(face="bold",color="black",size=16))
> p1
```

- **ANEXO D. Porcentajes de muertes de menores de un año, entre hombres y mujeres por provincia**

```
> s<-sum(Hombre) + sum(Mujer)
> M2021<-Mujer*100/s; H2021<-Hombre*100/s
> pyramid.plot(H2021,M2021,labels=Provincia,
+              main="",lxcol="blue",rxcol="pink",labelcex=1.2, unit="%",
+              gap=3.5,show.values=T, top.labels=c("Hombres","Provincia","Mujeres"),ndig=1,
xlim=c(25,25))
```

- **ANEXO E. Tabla de contingencia entre las variables Provincia y Causas de muerte respecto al número de muertes en menores de un año.**

  > xtabs(data$Muertes~data$Provincia+data$Enfermedad)

- **ANEXO F. Tabla de contingencia entre las variables Género y Causas de muerte respecto al número de muertes en menores de un año.**

```
> xtabs(data$Muertes~data$Genero+data$Enfermedad)
```

- **ANEXO G. Gráfica de cajas de las variables Causas de muerte y género en menores de un año.**

```
> p2<-ggplot(data, aes(Enfermedad1,Muertes,fill=Genero))+
+   geom_boxplot()+
+   xlab("Causas de Muerte")+
+   theme(panel.background=element_rect(fill="white",colour="black"),
+         axis.text.x=element_text(color="black",size=14,angle = 90, vjust = 0.5, hjust=1),
+         axis.text.y=element_text(color="black",size=14),
+         axis.title.x=element_text(face="bold",color="black",size=20),
+         axis.title.y=element_text(face="bold",color="black",size=20))+
+   scale_fill_discrete("Género",guide = guide_legend(reverse=TRUE))+
+   theme(legend.title = element_text(colour="black", size=20, face="bold"))+
+   theme(legend.text = element_text(colour="black", size = 14))
> p2
```

- **ANEXO H. Tabla de contingencia entre las variables Provincia y Género respecto a la variable Nacidos vivos.**

```
> xtabs(data$Nacidos~data$Provincia+data$Genero)
```

- **ANEXO I. Gráfica de cajas de las variables provincia y nacidos**

```
> p3<-ggplot(data, aes(Provincia,Nacidos),fill=Provincia)+
+   geom_boxplot(notch=F,color=c("blue4","red4","green4","yellow4","coral4","mediumpur
ple4","mistyrose4","gray24","aquamarine4","lightpink4","lightskyblue4"),fill=c("royalblue"
,"red1","green","yellow","coral","mediumpurple1","mistyrose","gray67","aquamarine","ligh
tpink","lightskyblue"))+
+   geom_point()+
+   xlab("Provincias")+
+   theme(panel.background=element_rect(fill="white",colour="black"),
+         axis.text.x=element_text(face="bold",color="black",size=10,angle = 90, vjust = 0.5,
hjust=1),
+         axis.text.y=element_text(face="bold",color="black",size=10),
+         axis.title.x=element_text(face="bold",color="black",size=16),
+         axis.title.y=element_text(face="bold",color="black",size=16))
> p3
```

- **ANEXO J. Porcentajes de nacidos, entre hombres y mujeres por provincia**

```
> s<-sum(Hombre) + sum(Mujer)
> M2021<-Mujer*100/s; H2021<-Hombre*100/s
> pyramid.plot(H2021,M2021,labels=Provincia,
+              main="",lxcol="blue",rxcol="pink",labelcex=1.2, unit="%",
+              gap=3.5,show.values=T, top.labels=c("Hombres","Provincia","Mujeres"),ndig=1,
xlim=c(25,25))
```

- **ANEXO K. Tablas de mortalidad**

```
> tabla_de_mortalidad<-(Muertes/Nacidos)*1000
```

- **ANEXO L. Regresión Lineal Simple**

Coeficientes de la regresión lineal simple y análisis de varianza de la variable Causas de Muerte

```
> mod1<-lm(Muertes~Enfermedad1,data)
> summary(mod1)
> anova(mod1)
```

Coeficientes de la regresión lineal simple y análisis de varianza de la variable Provincia

```
> mod2<-lm(Muertes~Provincia,data)
> summary(mod2)
> anova(mod2)
```

Coeficientes de la regresión lineal simple y análisis de varianza de la variable Nacidos

```
> mod3<-lm(Muertes~Nacidos,data)
> summary(mod3)
> anova(mod3)
```

Coeficientes de la regresión lineal simple y análisis de varianza de la variable Género

```
> mod4<-lm(Muertes~Genero,data)
> summary(mod4)
> anova(mod4)
```

- **ANEXO M. Regresión lineal múltiple**

Coeficientes de la regresión lineal múltiple y análisis de varianza del modelo completo.

```
> mo1<-lm(Muertes~Provincia+Enfermedad1+Nacidos+Nacidos*Enfermedad1+Nacidos*P
rovincia,data)
> summary(mo1)
> anova(mo1)
```

Coeficientes de la regresión lineal múltiple y análisis de varianza del modelo propuesto.

```
> step(mo1, direction = "backward")
> mo2<-lm(Muertes ~ Provincia + Enfermedad1 + Nacidos + Enfermedad1:Nacidos,data)
> summary(mo2)
> anova(mo2)
```

Análisis de varianza del modelo de regresión lineal múltiple completo vs el propuesto

```
> anova(mo1,mo2)
```

- **ANEXO N. Supuestos del modelo**

Normalidad

```
> plot(mo2)
> shapiro.test(residuals(mo2))
> r<-c(residuals(mo2))
> ks.test(r,pnorm,mean(r),sd(r))
> boxplot(residuals(mo2))
> jarque.bera.test(residuals(mo2))
> outlierTest(mo2)
> which.max(mo2$residuals)
```

Varianza constante

```
> plot(mo2)
> bptest(mo2)
```

Independencia

```
> plot(mo2)
> dwtest(mo2)
```

- **ANEXO O. Modelos lineales generalizados**

Regresión de Poisson

```
> total=Muertes+Nacidos
> modelo1<-glm(Muertes~Provincia+Enfermedad1+Nacidos,offset(log(total)),family=pois
son)
> summary(modelo1)
> exp(coef(modelo1))
> (dp <- sum(residuals(modelo1,type="pearson")^2)/modelo1$df.res)
```

Regresión Multinomial

```
> data$Enfermedad2 <- relevel(data$Enfermedad, ref = "1")
> test4 <- multinom(Enfermedad2 ~ Provincia +Muertes +Provincia:Muertes,weights=Muer
tes)
> summary(test4)
```

Simulaciones

```
> n=1000
> rep3=replicate(n=n,{
+   partida=sample(data$Provincia,132,replace=T)})
> datoss<-function(w,x,y,a){
+   n=132
+   e<-sample(w,n,replace=T)
+   m<-sample(x,n,replace=T)
+   na<-sample(y,n,replace=T)
+   p<-sample(a,n,replace=T)
+
+   dat<-data.frame(e,m,na,p)
+   m2<-lm(m~e+na+p+e*na)
+   su<-summary(m2)
+   total=m+n
+
+   g2<-glm(m~p+e+na+offset(log(total)),family=poisson)
+   su1<-summary(g2)
+
+   R2m2<-su$r.squared
+   R2g2<-1-(g2$deviance/g2$null.deviance)
+
+   print(R2m2)
+   print(R2g2)
+ }
> datoss(data$Enfermedad1,data$Muertes,data$Nacidos,data$Provincia)
```

espoch | Dirección de Bibliotecas y Recursos del Aprendizaje

UNIDAD DE PROCESOS TÉCNICOS Y ANÁLISIS BIBLIOGRÁFICO Y DOCUMENTAL

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

Fecha de entrega: 05 / 08 / 2022

INFORMACIÓN DEL AUTOR/A (S)
Nombres – Apellidos: *Verónica Janeth Argüello Pazmiño*
INFORMACIÓN INSTITUCIONAL
Instituto de Posgrado y Educación Continua
Título a optar: *Magíster en Matemática mención Modelación y Docencia*
f. Analista de Biblioteca responsable: Lic. Luis Caminos Vargas Mgs.

Firmado electrónicamente por:
LUIS ALBERTO CAMINOS VARGAS

UNIDAD DE PROCESOS TÉCNICOS DBRA ESPOCH

0090-DBRA-UPT-IPEC-2022

Traducción Resumen

6

Andrea Sofía Ribadeneria Vacacela

1

Para: Centro de Idiomas; Veronica Janeth Arguello Pazmino

Lun 1/8/2022 13:57

ABSTRACT Veronica Janeth A...
14 KB

Estimada señorita estudiante,

Con un saludo cordial y felicitándole por la etapa de estudios en la que se encuentra, por favor procedo a entregar el abstract entregado traducido para que por favor continúe con el proceso.

Warm regards,

M.T.E.F.L. Andrea Sofía Ribadeneira V.
ENGLISH TEACHER
CENTRO DE IDIOMAS MODALIDAD EXTENSIÓN
ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO
Celular # +593 987727600
Coreo Electrónico: andrea.ribadeneira@espoch.edu.ec