Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Posgrado
Maestría en Estadística Aplicada

ELABORACIÓN DE UN MODELO DE REGRESIÓN BINOMIAL NEGATIVA APLICADO A LA INCIDENCIA DE COVID19, EN FUNCIÓN DE LOS CAMBIOS EN LA MOVILIDAD POBLACIONAL, EN GUATEMALA

Dr. José Luis Alvarado Sosa
Asesorado por la MSc. Dra. Ingrid Fabiola Castillo Angel

Guatemala, octubre de 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

ELABORACIÓN DE UN MODELO DE REGRESIÓN BINOMIAL NEGATIVA APLICADO A LA INCIDENCIA DE COVID19, EN FUNCIÓN DE LOS CAMBIOS EN LA MOVILIDAD POBLACIONAL, EN GUATEMALA

TRABAJO DE GRADUACIÓN
PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

Dr. JOSÉ LUIS ALVARADO SOSA
ASESORADO POR LA MSC. DRA. INGRID FABIOLA CASTILLO ANGEL

AL CONFERÍRSELE EL TÍTULO DE

MAESTRO EN ESTADÍSTICA APLICADA

GUATEMALA, OCTUBRE DE 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Armando Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

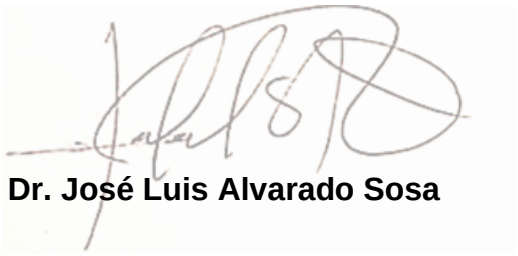
DECANA	Inga. Aurelia Anabela Cordova Estrada
EXAMINADOR	Mtro. Ing. Edgar Darío Álvarez Cotí
EXAMINADOR	Mtro. Ing. Edwin Adalberto Bracamonte Orozco
EXAMINADOR	Mtro. Ing. Luis Carlos Bolaños
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

ELABORACIÓN DE UN MODELO DE REGRESIÓN BINOMIAL NEGATIVA APLICADO A LA INCIDENCIA DE COVID19, EN FUNCIÓN DE LOS CAMBIOS EN LA MOVILIDAD POBLACIONAL, EN GUATEMALA

Tema que me fuera asignado por la Dirección de la Escuela de Estudios de Posgrado, con fecha 31 de enero de 2022.



Dr. José Luis Alvarado Sosa

Decanato
Facultad de Ingeniería
24189101- 24189102
secretariadecanato@ingenieria.usac.edu.gt

LNG.DECANATO.OI.667.2022

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Estudios de Posgrado, al Trabajo de Graduación titulado: **ELABORACIÓN DE UN MODELO DE REGRESIÓN BINOMIAL NEGATIVA APLICADO A LA INCIDENCIA DE COVID19, EN FUNCIÓN DE LOS CAMBIOS EN LA MOVILIDAD POBLACIONAL, EN GUATEMALA,** presentado por: **José Luis Alvarado Sosa,** que pertenece al programa de Maestría en artes en Estadística aplicada después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANA
FACULTAD DE INGENIERÍA

Inga. Aurelia Anabela Cordova Estrada
Decana

Guatemala, octubre de 2022

AACE/gaoc

Guatemala, octubre de 2022

LNG.EEP.OI.667.2022

En mi calidad de Director de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor, verificar la aprobación del Coordinador de Maestría y la aprobación del Área de Lingüística al trabajo de graduación titulado:

"ELABORACIÓN DE UN MODELO DE REGRESIÓN BINOMIAL NEGATIVA APLICADO A LA INCIDENCIA DE COVID19, EN FUNCIÓN DE LOS CAMBIOS EN LA MOVILIDAD POBLACIONAL, EN GUATEMALA"

presentado por **José Luis Alvarado Sosa** correspondiente al programa de **Maestría en artes en Estadística aplicada** ; apruebo y autorizo el mismo.

Atentamente,

"Id y Enseñad a Todos"

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA · FACULTAD DE INGENIERÍA · DIRECCIÓN · ESCUELA DE POSTGRADO ·

Mtro. Ing. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería

Guatemala 30 de mayo 2022.

M.A. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Presente

M.A. Ingeniero Álvarez Cotí:

Por este medio informo que he revisado y aprobado el **INFORME FINAL y ARTICULO CIENTÍFICO** titulado: **ELABORACIÓN DE UN MODELO DE REGRESIÓN BINOMIAL NEGATIVA APLICADO A LA INCIDENCIA DE COVID19, EN FUNCIÓN DE LOS CAMBIOS EN LA MOVILIDAD POBLACIONAL, EN GUATEMALA** del estudiante **José Luis Alvarado Sosa** quien se identifica con número de carné **100021434** del programa de Maestría en Estadística Aplicada.

Con base en la evaluación realizada hago constar que he evaluado la calidad, validez, pertinencia y coherencia de los resultados obtenidos en el trabajo presentado y según lo establecido en el *Normativo de Tesis y Trabajos de Graduación aprobado por Junta Directiva de la Facultad de Ingeniería Punto Sexto inciso 6.10 del Acta 04-2014 de sesión celebrada el 04 de febrero de 2014*. Por lo cual el trabajo evaluado cuenta con mi aprobación.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

Atentamente,

UNIVERSIDAD DE SAN CARLOS
ESCUELA DE POSTGRADO
FACULTAD DE INGENIERIA
DE GUATEMALA

MSc. Ing. Edwin Adalberto Bracamonte Orozco
Coordinador
Maestría Estadística Aplicada
Escuela de Estudios de Postgrado

Guatemala, 16 de mayo de 2022.

M.A. Ing. Edgar Darío Álvarez Cotí

Director

Escuela de Estudios de Postgrado

Presente

Estimado M.A. Ing. Álvarez Cotí

Por este medio informo a usted, que he revisado y aprobado el Trabajo de Graduación y el Artículo Científico: "**ELABORACIÓN DE UN MODELO DE REGRESIÓN BINOMIAL NEGATIVA APLICADO A LA INCIDENCIA DE COVID19, EN FUNCIÓN DE LOS CAMBIOS EN LA MOVILIDAD POBLACIONAL, EN GUATEMALA**" del estudiante **JOSÉ LUIS ALVARADO SOSA** del programa de Maestría en **Estadística Aplicada**, identificado con número de carné: **100021434.**

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

Dra. Ingrid Fabiola Castillo Angel
MEDICO Y CIRUJANO
COL. 15865

MSc. Dra. Ingrid Fabiola Castillo Angel

Colegiado No. 15,865

Asesora de Tesis

ACTO QUE DEDICO A:

Dios	Por quien todo fue hecho.
Mis padres	José Alvarado y Lucrecia Sosa.
Mis hermanos	Rodolfo, Marcos y Manuel Alvarado.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por las oportunidades.
Facultad de Ingeniería	Por las enseñanzas.
Dr. Ingrid F. Castillo	Por el apoyo y asesoría.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
$\boldsymbol{r}$	Coeficiente de correlación de Pearson
$\boldsymbol{R^2}$	Coeficiente de determinación
$\boldsymbol{R^2_{aj}}$	Coeficiente de determinación ajustado
$\boldsymbol{b_i}$	Coeficientes de regresión en un modelo de regresión
$\boldsymbol{BIC}$	Criterio de información Bayesiano
$\boldsymbol{C_p}$	Criterio de información Cp de Mallows
$\boldsymbol{AIC}$	Criterio de información de Akaike
K	Curtosis de un conjunto de datos
$\boldsymbol{s}$	Desviación estándar
DP	Distribución de Poisson
$\boldsymbol{\varepsilon}$	Error aleatorio
$\boldsymbol{\tau}$	Estimador Tau de Kendall
$\boldsymbol{H_a}$	Hipótesis alternativa
$\boldsymbol{H_0}$	Hipótesis nula
$\boldsymbol{\beta_o}$	Intercepto en un modelo de regresión
Ln	Logaritmo natural
μ	Media aritmética
$\boldsymbol{\overline{x}}$	Media aritmética de la variable x
$\boldsymbol{\overline{y}}$	Media aritmética de la variable y
$\boldsymbol{\hat{\mu}}$	Media muestral
$\boldsymbol{Mo}$	Mediana

$\boldsymbol{k}$	Número de parámetros incluidos en el modelo de regresión
$\boldsymbol{e}$	Número e
$\boldsymbol{\beta_i}$	Parámetro (pendiente) de variable regresora en modelo de regresión
$\boldsymbol{\alpha}$	Parámetro de sobredispersión en un modelo de regresión de Poisson
$\boldsymbol{t}$	Período de tiempo determinado
$\boldsymbol{Pr}$	Probabilidad
$\boldsymbol{\lambda}$	Promedio de eventos en un tiempo determinado de una variable de conteo
$\mathbf{RBN}$	Regresión binomial negativa
$\mathbf{RP}$	Regresión de Poisson
$\boldsymbol{e_i}$	Residuos en un modelo de regresión
$\boldsymbol{skp}$	Sesgo o asimetría de un conjunto de datos
$\boldsymbol{n}$	Tamaño de muestra
$\boldsymbol{\hat{y}_i}$	Valor de y estimado en un modelo de regresión
$\boldsymbol{y_i}$	Valor de y medido
$\boldsymbol{x_i}$	Variable aleatoria
$\boldsymbol{y}$	Variable dependiente o respuesta
$\boldsymbol{s^2}$	Varianza muestral

GLOSARIO

Android	Sistema operativo de algunos dispositivos electrónicos.
APdA	ARN polimerasa dependiente de ARN.
ARN	Ácido ribonucleico.
Astra/Zéneca	Empresa farmacéutica de origen británico.
CDC	Centros para el Control de Enfermedades.
CoV	Coronavirus.
COVID-19	Enfermedad causada por el SARS-CoV2.
Delta	Variante patogénica del SARS-CoV2.
DP	Distribución de Poisson.
ECA2	Enzima convertidora de angiotensina 2.
FDA	Agencia de drogas y alimentos.

Google	Compañía subsidiaria de *Alphabet*, especializada en productos y servicios de internet, *softwares* y otros electrónicos.
Incidencia	Casos nuevos de una enfermedad en una unidad de tiempo determinada, per cápita.
IRR	Tasa de razones de incidencia.
MASS	Paquete de *software* para análisis estadístico, en el lenguaje R.
MERS	Síndrome respiratorio del medio este.
MMC	Método de mínimos cuadrados.
Moderna	Empresa farmacéutica de origen estadounidense.
OR	*Odds ratio* (proporción de riesgos).
Ómicron	Variante patogénica del SARS-CoV2.
Pscl	Paquete de *software* para análisis estadístico, en el lenguaje R.
Pfizer	Empresa farmacéutica de origen estadounidense.
R	Lenguaje de programación de *software* estadístico.

RBN	Regresión binomial negativa.
RECOVERY	Ensayo clínico de Fase III que comparó la eficacia de múltiples estrategias terapéuticas en la COVID-19.
RLM	Regresión lineal múltiple.
RLS	Regresión lineal simple.
Ro	Número efectivo de reproducción.
RP	Regresión de Poisson.
RStudio	Entorno de desarrollo integrado para el *software* de programación R.
SARS	Síndrome agudo respiratorio severo.
SARS-CoV2	Virus causante del síndrome agudo respiratorio severo por coronavirus 2.

RESUMEN

El propósito de este estudio fue analizar la incidencia de COVID-19 en Guatemala, de acuerdo con la variabilidad de la movilidad poblacional.

El objetivo general fue construir un modelo de regresión binomial negativa para explicar la incidencia de COVID-19 en Guatemala, en función de la variabilidad en la movilidad poblacional a categorías de sitios específicos.

La metodología empleada consistió en la obtención de los datos de fuentes *web* de acceso abierto, lo que permitió analizar 780 días consecutivos de datos. El modelo fue estimado con todas las variables pre-especificadas y se depuró por AIC y R^2aj. Posteriormente, se exponenciaron los coeficientes regresores, para estimar las tasas de razones de incidencia, para cada una de las variables incluidas en el modelo final.

Los resultados muestran que la variante Delta es el predictor más fuerte (IRR 2.7296, IC 95 % 2.2931-3.2482, $p=2^{-16}$) y la afluencia a áreas residenciales (0.9775, IC 95 % 0.9666-0.9886, $p=2^{-16}$) es el otro predictor significativo. El modelo estimado presenta un R^2=0.357, AIC=11815 y BIC=11833.

El mejor modelo estimado incluyó el estatus de la variante Delta y la afluencia a áreas residenciales.

Esto permite considerar que limitar la movilidad comunitaria es una medida efectiva para reducir la incidencia de COVID-19 en Guatemala.

PLANTEAMIENTO DEL PROBLEMA

- Contexto general

La pandemia de la COVID-19 ha conllevado la necesidad de regular las actividades poblacionales que antes se consideraban cotidianas. Al coartar libertades individuales, el poder medir el impacto de estas medidas, en términos de eficacia, es una necesidad fundamental, pues es la única forma de justificar el seguir implementando las o el modificarlas.

Una de las actividades poblacionales que se ha visto restringida es la de la movilidad libre. En Guatemala, la mayor parte del 2020 se caracterizó por la implementación de toques de queda, cordones sanitarios localizados, restricciones de horario y afluencia de personas en sitios comerciales y no comerciales, implementación de trabajo en casa. Todas estas medidas han obligado a la población a adaptar sus vidas, y las de sus familias, a lo que se ha llamado la nueva normalidad. En 2021 las medidas, aunque menos estrictas, no fueron aliviadas del todo.

Esto ha tenido un costo social importante, con disrupción de la economía, pérdida de empleos, cierre de negocios y alteración en la salud mental de la población. Por lo anterior, se considera necesario estimar el impacto de dicha medida, y valorar si realmente contribuye a la disminución de la incidencia de casos de COVID-19 que están dirigidas a lograr, a modo de poder justificar su continuidad, o suspensión, en los meses siguientes.

- Descripción del problema

Las medidas de restricción de la movilidad poblacional han sido la norma en 2020 y 2021. La empresa Google, gracias al registro de datos de localización geográfica, ha hecho pública una extensa base de datos que permite estimar la variabilidad en movimiento o afluencia de personas hacia los diferentes lugares tanto dentro de sus países como de sus provincias. Esta base de datos recibe el nombre de *Reporte de Movilidad Comunitaria*, y ha categorizado los lugares en: tiendas de ocio, supermercados y farmacias, parques, estaciones de transporte, lugares de trabajo y residenciales.

A la vez, tanto el Ministerio de Salud, como entidades internacionales, han hecho públicas bases de datos que incluyen la incidencia diaria de casos de los diferentes países del mundo.

Con lo anterior, se pretende desarrollar un modelo de regresión con la incidencia bruta de casos como la variable respuesta, y las variaciones de movilidad comunitaria hacia las diferentes categorías de lugares como las diferentes variables regresoras.

Para el respectivo análisis, se considerará un período de incubación de 10 días, y se podrá ajustar el modelo por la presencia de las variantes Delta y Ómicron, así como la tasa *per cápita* de vacunación pues, de otra forma, podrían ser fuentes de error.

Esto también permitirá evaluar cuáles de las categorías de lugares cuya afluencia poblacional se medirán guardan más relación con la variable

dependiente, de acuerdo con la información que brinden los coeficientes del modelo de regresión que se estime.

- Formulación del problema

    - Pregunta central

¿Qué modelo de regresión mejor explica la incidencia de COVID-19 en función de los cambios de la movilidad poblacional en Guatemala?

    - Preguntas auxiliares

        - ¿Cuál ha sido el comportamiento de la incidencia de casos nuevos de COVID-19 en la población de Guatemala?

        - ¿Cuál ha sido el comportamiento y la variabilidad de la movilidad comunitaria, durante la pandemia de COVID-19, en la población de Guatemala?

        - ¿Cuál es el mejor modelo de regresión a elaborar mediante la evaluación de los criterios de información pertinentes, para explicar la incidencia de COVID-19 en función de los cambios de movilidad poblacional en Guatemala?

- Delimitación del problema

Se utilizaron los datos del período comprendido entre febrero de 2020 (período previo al momento en que la pandemia fue detectada en Guatemala, por

primera vez) hasta abril de 2022. Para abordar el problema, se utilizaron dos bases de datos ya existentes, de acceso abierto:

- Reporte de movilidad comunitaria, perteneciente a la empresa Google, disponible en la web, https://www.google.com/COVID-19/mobility/?hl=es.

- Reporte diario de casos, disponible a la organización Our World in Data, disponible en la web, https://ourworldindata.org/.

OBJETIVOS

General

Explicar la incidencia de COVID-19 en función de los cambios en la movilidad poblacional en Guatemala, por medio de un modelo de regresión.

Específicos

- Describir cuál ha sido el comportamiento de la incidencia bruta de casos nuevos de COVID-19 en la población de Guatemala, por medio de un análisis de estadística descriptiva.

- Caracterizar cuál ha sido el comportamiento de la movilidad comunitaria, durante la pandemia de COVID-19, en la población de Guatemala, mediante un análisis de estadística descriptiva.

- Determinar cuál es el mejor modelo de regresión a elaborar de acuerdo con la evaluación de los criterios de información apropiados para la naturaleza de dicho modelo, para explicar la incidencia de COVID-19 en función de los cambios de movilidad poblacional en Guatemala.

RESUMEN DEL MARCO METODOLÓGICO

A continuación, se resumen las principales características metodológicas del proceso de investigación llevado a cabo.

- Enfoque

Se trata de una investigación de enfoque cuantitativo, en el que se midió la correlación entre variables continuas y categóricas, mediante la estimación de un modelo de regresión binomial negativa.

- Diseño

Diseño no experimental, con una proyección retrospectiva, descriptiva, de corte transversal. Es de enfoque correlacional.

- Tipo

Se trata de un estudio correlacional de enfoque cuantitativo, en el que se estudió la relación entre variables cuantitativas y cualitativas predictoras para explicar la variable respuesta, discreta, de conteo, mediante la creación de un modelo de regresión binomial negativa, de diseño no experimental.

- Variables e indicadores

Tabla I. **Variables del estudio**

Variable	Definición teórica	Definición operativa	Escala de medición
Incidencia de COVID-19.	Cantidad de casos nuevos de COVID-19 por día, en la República de Guatemala.	Casos de pacientes nuevos diagnosticados y registrados de COVID-19 en la República de Guatemala.	Cuantitativa, discreta, de conteo.
Variabilidad de movilidad para estaciones de transporte.	Variabilidad en visita y permanencia para centros de transporte.	Porcentaje de variabilidad en la movilidad comunitaria de individuos guatemaltecos con teléfono Android y permiso para registrar su actividad de movilización, según asistencia y permanencia para centros de transporte.	Cuantitativa, proporción porcentual.
Variabilidad de movilidad para lugares de trabajo.	Variabilidad en visita y permanencia para sitios de trabajo.	Porcentaje de variabilidad en la movilidad comunitaria de individuos guatemaltecos con teléfono Android y permiso para registrar su actividad de movilización, según asistencia y permanencia para lugares de trabajo.	Cuantitativa, proporción porcentual.
Variabilidad de movilidad para zonas residenciales	Variabilidad en visita y permanencia para sitios de residencia	Porcentaje de variabilidad en la movilidad comunitaria de individuos guatemaltecos con teléfono Android y permiso para registrar su actividad de movilización, según asistencia y permanencia para zonas residenciales.	Cuantitativa, proporción porcentual.
Variabilidad de movilidad para Parques	Variabilidad en visita y permanencia a parques	Porcentaje de variabilidad en la movilidad comunitaria de individuos guatemaltecos con teléfono Android y permiso para registrar su actividad de movilización, según asistencia y permanencia para parques.	Cuantitativa, proporción porcentual.

Continuación tabla I.

Variabilidad de movilidad para Tiendas y lugares de ocio	Variabilidad en visita y permanencia a tiendas y lugares de ocio	Porcentaje de variabilidad en la movilidad comunitaria de individuos guatemaltecos con teléfono Android y permiso para registrar su actividad de movilización, según asistencia y permanencia para tiendas y lugares de ocio.	Cuantitativa, proporción porcentual.
Variabilidad de movilidad para Supermercados y farmacias	Variabilidad en visita y permanencia a supermercados y farmacias	Porcentaje de variabilidad en la movilidad comunitaria de individuos guatemaltecos con teléfono Android y permiso para registrar su actividad de movilización, según asistencia y permanencia para supermercados y farmacias.	Cuantitativa, proporción porcentual.
Tiempo de incubación de COVID-19, desde la exposición hasta el diagnóstico.	Tiempo estimado que transcurre entre la exposición del sujeto y su diagnóstico formal de COVID-19, entre 2-14 días, según el "*Centers for Disease Control*" (CDC) de Estados Unidos.	10 días de estimación de incubación usual, que Funcionarán para correlacionar la variabilidad en la movilidad con la incidencia de casos de COVID-19 de 10 días después	Cuantitativa, proporción porcentual.
Presencia de variante Delta.	Valoración dicotómica para establecer si la variante delta del SARS-CoV2 se ha detectado o no en Guatemala.	Dicotomización de la variable, según si se ha detectado en Guatemala o no, 0=no y 1=sí	Cualitativa, dicotómica.
Presencia de variante Ómicron	Valoración dicotómica para establecer si la variante delta del SARS-CoV2 se ha detectado o no en Guatemala.	Dicotomización de la variable, según si se ha detectado en Guatemala o no, 0=no y 1=sí	Cualitativa, dicotómica.
Tasa de vacunación	Cantidad de vacunas administradas en una cantidad de tiempo y población específicas	Dosis de vacuna anti COVID-19 por cada 100,000 habitantes	Cuantitativa, tasa por 100,000 habitantes

Fuente: elaboración propia, empleando Microsoft Excel.

- Fases del estudio

    - Fase 1: revisión de literatura

Esta fase tuvo como objetivo primordial la búsqueda y lectura de las fuentes relevantes referentes a la pandemia de la COVID-19, así como a la metodología estadística planteada, con el propósito de dar soporte documental a la selección del método, así como brindar un marco referencial ante el cual contextualizar y comparar los resultados que se obtengan.

    - Fase 2: obtención de la información

Esta fase consistió en la descarga de las bases de datos del registro de casos de COVID-19, la movilidad comunitaria y el registro de dosis de vacunas administradas en Guatemala. Se extendió hasta abril de 2022 de forma intencional, buscando contar con la mayor cantidad de datos posible. Se incluyeron 780 días consecutivos de datos ininterrumpidos.

    - Fase 3: análisis e interpretación de la información

La primera parte de esta fase consistió en analizar el comportamiento de la incidencia de casos de COVID-19 en Guatemala y la variabilidad de la movilidad comunitaria, con la categorización referente a la presencia o ausencia de las variantes de interés estudiadas (Delta y Ómicron). En este momento se modificó la variable de casos nuevos de COVID-19 para incluir el retraso de 10 días correspondiente al tiempo seleccionado para representar el período de incubación del SARS-CoV2. Esto se realizó en los *softwares:* Excel 365, Minitab y RStudio.

Posterior a esto, se procedió a construir el modelo de regresión binomial negativa (que es el producto principal de esta investigación), con la consecuente depuración guiada por criterios de información seleccionados, en este caso, el criterio de Akaike y el R2 ajustado. Esto fue realizado en el *software* RStudio (ver apéndices para repasar la sintaxis).

Finalmente, se procedió a interpretar los resultados, fase en la que se buscó la significancia epidemiológica de los resultados obtenidos mediante el análisis estadístico realizado.

- Fase 4: redacción de informe final

Se redactó el informe final, de acuerdo con especificaciones previamente definidas de estilos pertinentes, incluyendo tablas e ilustraciones explicativas, estadísticos calculados e interpretación de resultados con conclusiones y recomendaciones pertinentes.

INTRODUCCIÓN

El presente trabajo de investigación consiste en una sistematización, interpretación crítica de los datos disponibles, que busca la estimación de un modelo de regresión para explicar la incidencia de COVID-19 en Guatemala (variable dependiente) en función de la variabilidad de la movilidad comunitaria categorizada en seis tipos de lugares diferentes: tiendas y ocio, supermercados y farmacias, lugares de trabajo, áreas residenciales, parques y estaciones de transporte. Esto, a la vez que se ajustó para la presencia o ausencia de las principales variantes patogénicas detectadas en el país (Delta y Ómicron), así como la incidencia de vacunación en el tiempo. En este sentido, está enmarcado en la línea investigativa de análisis multivariado y regresión.

El aparecimiento de la COVID-19 ha representado un problema global. En mayor o menor medida, todos los países han sido afectados por la pandemia. Esta ha tenido efectos directos y evidentes, como el colapso de sistemas de salud completos o elevada carga de morbilidad y mortalidad, así como efectos indirectos, como ralentización de la economía, alteración de las cadenas de suministros y necesidad de recurrir a medidas estrictas y severas para contener la expansión del virus.

En este sentido, las medidas de restricción de movilidad comunitaria coartan una libertad humana y, de esta cuenta, muchos han sido sus detractores. De esta cuenta, el poder estimar el impacto que tiene la movilidad de los habitantes en la expansión del SARS-CoV2 puede ser considerado como una necesidad que debe de ser resuelta de forma urgente pues, tras un análisis crítico, la decisión de prolongar y fortalecer estas medidas o, por otro lado, apresurar su terminación

o reducir su drasticidad, debe de tener como objetivo el restringir la movilidad únicamente si esto implica salvar vidas y/o preservar la funcionalidad del sistema de salud.

La restricción en la movilidad comunitaria es una de las medidas aplicadas y, sin embargo, es también causa de controversia, pues restringe libertades individuales. En este contexto, comprender la relevancia de la movilidad sobre la incidencia de COVID-19 representa una necesidad fundamental para soportar estas medidas.

Para lo anterior, y tras la respectiva investigación documental, se obtuvieron bases de datos extensas que incluían 780 días consecutivos de registro de las variables de interés, y, mediante el uso de *softwares* especializados, se construyó el modelo de regresión binomial negativa, según lo sugirió la naturaleza de la variable dependiente. Tras esto, dicho modelo fue depurado, guiado por los criterios de información seleccionados (AIC y R^2aj) y redactado en el informe que es el producto final de este trabajo.

El modelo final, depurado según se comentó en el párrafo anterior, permitió determinar los principales determinantes de la variabilidad de casos diarios de COVID-19. Las variantes Delta y Ómicron representan la principal razón de variabilidad en la incidencia de COVID-19. Referente a la movilidad comunitaria, el incremento en la afluencia a los sitios residenciales (IRR 0.89) y a las estaciones de transporte (IRR0.94) representan reducción en la probabilidad de incremento de la incidencia de COVID-19, mientras que el incremento en la afluencia a sitios de trabajo (IRR 1.009) y a sitios de tiendas y ocio (IRR 1.01) se asociaron con mayor probabilidad de incremento en la incidencia de COVID-19.

Estos resultados pueden servir como referencia a las autoridades encargadas de diseñar las políticas públicas que se implementan para mitigar la pandemia reduciendo la transmisibilidad de casos de COVID-19 mediante medidas de salud pública, con el fin de diseñar medidas nuevas de acuerdo con la eficacia de las previamente implementadas.

Las cuatro fases en que realizó este estudio son: 1° la investigación documental para generar un marco de referencia contra el cual se pudieran comparar y entender los resultados a obtener, 2° la obtención de los datos necesarios, mediante la descarga de estos de fuentes confiables en la *web*, 3° el análisis e interpretación de la información, actividades que se realizaron de forma simultánea y 4° la redacción del informe final.

Desde las diferentes perspectivas que se analizan, el estudio fue factible. Económicamente, la asesoría se realizó *ad honorem*, los *softwares* usados para el análisis estadístico fueron RStudio (de acceso abierto) y Minitab y Microsoft Excel, bajo licencia anual. El equipo utilizado fue modesto (una única computadora personal) y no requirió movilidad ni personal adicional al investigador.

Este informe consta de cuatro capítulos. El primero se titula marco referencial, y presenta antecedentes nacionales e internacionales que abordan la problemática mediante diferentes métodos, y se comentan sus resultados principales.

El segundo capítulo marco teórico, incluye los principios teóricos necesarios para sustentar el análisis estadístico realizado, así como la teoría fundamental sobre el SARS-CoV2 y la COVID-19.

El tercer capítulo presenta los resultados obtenidos, tanto la descripción de los datos, incluyendo las gráficas respectivas, como numérica y estadísticamente, así como las estimaciones que dan respuesta a los objetivos planteados.

En el cuarto capítulo se presenta la discusión de los resultados, presentando el análisis interno y externo del trabajo realizado y sirve de sustento para las conclusiones y recomendaciones planteadas posteriormente.

1. MARCO REFERENCIAL

En esta sección, se discutirán la naturaleza y evolución de la pandemia, y algunas medidas adoptadas en el mundo y en Guatemala para mitigarla. La información que se presentará en los siguientes párrafos permite: contextualizar los resultados, validar las bases de datos y la metodología estadística empleadas, así como comprender la necesidad de la realización del estudio.

1.1. Generalidades

La pandemia de la COVID-19, producto de la diseminación descontrolada del SARS-CoV2, ha derivado en una situación caótica, con resultados nefastos tanto en términos de pérdida de vidas (Huang *et. al.*, 2020).

Como de disrupción de la economía global esto ha impulsado la adopción de importantes medidas de salud pública, muy heterogéneas entre los países, muchas de las cuales han estado sujetas a críticas constantes de un buen número de detractores (Chu *et. al.*, 2020).

Obviamente, al tratarse de una situación de emergencia global, en la que la experiencia reciente se limitaba a brotes epidémicos limitados a regiones geográficas determinadas, no existió consenso en cómo debía de manejarse el balance entre medidas restrictivas preventivas y el mantenimiento del estatus quo de la calidad de vida de la población, con lo que la contextualización de la situación de Guatemala es un paso necesario para entender los resultados, más allá de lo que las estadísticas y números puedan aportar (Huremović, 2019).

Una de estas medidas, y, también, una de las más polémicas, ha sido la que implica restricciones o limitaciones en la movilidad de la población. La medición de la eficacia de estas medidas, o de indicadores que permitan evaluarlas (aún si sólo de forma indirecta), se hace, entonces, algo necesario.

1.1.1. Análisis de resultados de investigaciones previas

A continuación, se resumen algo de la evidencia generada, tanto en el mundo como en Guatemala.

1.1.1.1. Análisis a nivel internacional

En el caso del impacto de la movilidad comunitaria, Sulyok y Walker (2020) realizaron un interesante estudio. Al analizar la movilidad comunitaria en 135 países, estimaron la correlación de la variabilidad de esta con la incidencia de casos por país, mediante el estimador τ de Kendall, (data no paramétrica). Estimaron un modelo aditivo generalizado de interceptos de efectos mixtos aleatorios, para lo que emplearon una distribución tipo Tweedie, con los países considerados como los interceptos aleatorios.

En dicho estudio, Sulyok y Walker (2020) encontraron correlaciones fuertemente negativas en América del Norte, Europa Occidental, Rusia y Australia, respecto al movimiento en Tiendas y ocio, Abarrotes y farmacias, Laboral y Transporte, con tiempos variables entre los cambios en la movilidad y los cambios en la incidencia de casos. Cabría concluir que la movilidad hacia sitios no indispensables es una posible causa de aumentos en la incidencia de casos. Este estudio permitió entender que los patrones de movilidad comunitaria importan. En este sentido, los lugares a los que se mueven los ciudadanos son importantes, pero en diferentes medidas, teniendo diferente valor predictivo al

momento de intentar comprender la variabilidad de la incidencia de COVID-19. De esta forma, la estimación de un modelo de regresión, que permitirá obtener los coeficientes respectivos para cada uno de los diferentes lugares, junto a sus parámetros de significancia estadística, así como ajustar el modelo para reducir el error que podrían causar la presencia de otras variables de interés (variantes del virus, vacunación) se justifica.

Un estudio diferente (Kramer *et. al.*, 2020), pero con motivaciones similares, limitado a China, en el que analizaron movilidad humana en tiempo real en conjunto con datos epidemiológicos de diferentes provincias, logró encontrar una fuerte asociación entre la movilidad de los sujetos y la diseminación de la enfermedad, aunque cabe mencionar que esto fue al inicio del brote, tiempo en el que los casos comunitarios aún no eran el principal factor que determinase la expansión de la enfermedad.

En este, los resultados sugieren que las medidas de restricción de movilidad tuvieron gran impacto en reducir la cadena de contagios en dicho país. En el modelo estadístico utilizado por los investigadores, la data referente a movilidad y capacidad de testeo fueron factores fundamentales que fortalecieron dicho modelo, en comparación con un modelo inicial que consideraba sólo transmisión autóctona de los casos y un período de incubación de 2-8 días. Ellos concluyeron que el cordón sanitario fue fundamental para controlar la epidemia (Kramer *et. al.*, 2020).

Este estudio permite entender, a diferencia del anterior, que este tipo de análisis, aplicado a regiones geográficas pequeñas, mantiene validez estadística y metodológica, con los ajustes pertinentes, según se considere.

Por otro lado, al analizar las causas en los cambios en los patrones de movilidad comunitaria, un estudio que busca correlacionar los cambios de movilidad con la preferencia individual a tomar o no un riesgo, logró encontrar correlación entre la actitud de los individuos y los cambios en movilidad, a partir de la generación de un modelo lineal de efectos aleatorios (Chan *et. al.*, 2020).

Esto nos sugiere que las preferencias individuales tienen impacto en la eficacia de las medidas adoptadas por las autoridades. Aunque este estudio no contribuye a construir el modelo estadístico, sí brinda información que puede ser relevante al momento de entender los patrones de movilidad comunitaria en Guatemala. Esto ayuda a responder el objetivo que busca caracterizar este comportamiento.

Por otro lado, Oztig y Askin (2020) realizaron un estudio en el que analizaron la incidencia de la COVID-19 de acuerdo con indicadores de movilidad humana internacional, en el que encontraron que esta es también una fuerte predictora de la evolución de la incidencia. Para esto, se basaron sobre un análisis de regresión binomial negativa. Este artículo apoya fuertemente la elección de la distribución de Poisson y la construcción del modelo de regresión binomial negativa para llevar a cabo el análisis estadístico del estudio.

Como se discute en un artículo publicado durante los primeros meses de la pandemia, las medidas de distanciamiento social son efectivas, y las restricciones a la movilidad poblacional son una forma bastante fuerte de forzar este distanciamiento (Chu *et. al.*, 2020).

A pesar de ser un evento muy reciente (a la fecha en que esto fue escrito, la enfermedad tiene 21 meses de existir), se han logrado avances muy importantes en la atención de la enfermedad, así como en la modulación de los

factores patológicos de las personas y conductuales de las poblaciones que determinan la diseminación y secuelas de esta. El estudio de las medidas de restricción de la movilidad es sólo un ejemplo de lo anterior.

1.1.1.2. Análisis a nivel nacional

Guatemala es un país relativamente pequeño, ubicado en la región norte de Centroamérica, con una población estimada en poco más de 16,860,000 habitantes, un área territorial de 108.9 mil Km^2, con indicadores propios de naciones subdesarrolladas, esto de acuerdo con datos del Banco Mundial (CountryProfile, 2021).

La COVID-19 llegó a Guatemala en marzo de 2020. La carga ha sido muy difícil de llevar, por las deficiencias del sistema, conocidas desde antes.

De acuerdo con la Organización Mundial de la Salud (OMS), Guatemala muestra algunos indicadores insatisfactorios: bajo índice de desarrollo humano (puesto 128 en el ranquin mundial, para 2014), baja inversión pública en salud (62.4 % de los gastos se sufragan de forma privada), con escaso personal (0.897 médicos por cada 1000 habitantes) y tasas de mortalidad superiores para grupos vulnerables mayores a la media regional (GHO, 2021).

Estos datos permiten comprender la vulnerabilidad del sistema de salud en diferentes perspectivas: por un lado, la capacidad diagnóstica y de registro y comunicación de datos y de información pueden verse limitadas, lo que ha llevado al autor de este estudio a elegir bases de datos internacionales como fuente primaria de la incidencia de COVID-19. Así mismo, la precariedad de la capacidad de atención permite entender la necesidad de implementar medidas de contención, lo que ayuda a comprender los resultados respectivos, necesarios

para alcanzar el objetivo que busca describir el comportamiento de la pandemia en Guatemala.

A la fecha en que esto fue escrito (19 de septiembre de 2021), se contabilizan en Guatemala casi 530,000 casos totales, 13,040 fallecidos (2.4 % de letalidad), con una proporción muy similar entre hombres y mujeres. El Departamento de Guatemala concentraba muchos de estos casos. El comportamiento de la pandemia ha sido en olas de contagios, o brotes epidémicos. Al analizar estos datos, ha podido entenderse que no se trata de una serie temporal, permitiendo descartar este tipo de metodología estadística.

Medidas de restricción de movilidad, que propician distanciamiento social, uso de equipo de barrera y paliación económica han sido aplicados en Guatemala de forma que ha sido considerada arbitraria y poco efectiva, por parte de algunos autores (Hanvoravongchai, *et. al.*, 2020).

Esta información ha demostrado la necesidad de buscar bases de datos sobre la movilidad comunitaria libres de sesgo (en la medida de lo posible). Considerando que los teléfonos con sistema operativo Android predominan en Guatemala, con más del 85 % del *market share* en nuestro país, se consideró que esta información era la ideal para proceder con el estudio (Statcounter Global Stats, 2022).

Por otra parte, en un estudio de 314 ciudades en América Latina, en el que Guatemala tomó parte, Kephart et al. (2021) concluyen que hay un importante incremento en la incidencia de casos de COVID-19, con una ratio de tasa de incidencia de 2.35 (IC 95 % 2.12-2.6) por cada incremento de una unidad logarítmica en el cambio de los patrones de movilidad (comparado con el valor basal) de la semana previa. Este artículo permite entender la naturaleza de la

data a estudiar. Era fundamental elegir un tiempo de incubación y este estudio permite entender el impacto que esta decisión tiene en el análisis estadístico.

Finalmente, Sulyok y Walker (2020), en su estudio que incluye a Guatemala, encuentran que la data nacional concuerda con lo observado en el resto de América, esto es, la movilidad comunitaria impacta la incidencia de la COVID-19. En general, en Guatemala se han aplicado medidas sociales para mitigar la pandemia, que han evidenciado tener efecto positivo en reducir la transmisibilidad del virus en nuestro país.

2. MARCO TEÓRICO

Este capítulo se divide en dos partes. Por un lado, se incluye la fundamentación estadística que permite comprender la metodología que se aplicará para alcanzar los objetivos, y la segunda lo referente al problema que se pretende abordar, la COVID-19.

2.1. Fundamentación estadística

El trabajo estadístico, en el presente estudio, se dividirá en partes sucesivas. Inicialmente, se describirá los datos. Posteriormente, en la búsqueda de relación entre variables. Se mencionarán únicamente los métodos de análisis que se han pensado utilizar en este estudio.

2.1.1. Estadística descriptiva

De acuerdo con Vittinghoff *et. al.* (2012) y Walpole *et. al.* (2012), el primer paso de cualquier análisis estadístico comprende una revisión preliminar de la data, con tres objetivos: detectar errores o anomalías, analizar su distribución e iniciar a entender cómo se relacionan los predictores, o si no lo hacen.

Para lo anterior, se emplean diferentes técnicas estadísticas, sencillas, pero de suma importancia, que incluyen análisis por estimadores y análisis gráficos. Describiremos estas técnicas.

2.1.1.1. Revisión de los datos

Walpole *et. al.* (2012) sugiere que esta es una parte fundamental del proceso, si bien no incluye ningún análisis estadístico. Implica una rigurosa revisión de la data a analizar, con el propósito de evitar errores que pueden resultar en la invalidación, inclusive completa, del posterior análisis. Puede resultar en la detección de valores extremos y sus causas (si las hay), la reducción de estos o su exclusión, la estandarización de datos o hasta la necesidad de revisar o de volver a medir los valores incluidos (Barchard y Verenikina, 2013).

El empleo de *softwares* modernos puede facilitar el proceso.

2.1.1.2. Descripción de variables

Vittinghoff *et. al.* (2012) es muy claro al indicarnos que el primer paso para describir las variables es comprender su naturaleza. La selección de los estimadores y los gráficos adecuados para analizarlas nace de la adecuada comprensión de su naturaleza (numérica o categórica, y las subclases de estas). La descripción de variables representa también una oportunidad, para entender y corregir errores en la data introducida, valores extremos o su distribución.

Para el análisis de las variables numéricas, comentamos brevemente los principales estimadores estadísticos, separados en indicadores de tendencia central e indicadores de dispersión.

2.1.1.2.1. Medidas de tendencia central

Las medidas de tendencia central de una muestra son la media o promedio aritmético, la mediana y la moda (Walpole *et. al.*, 2012). Se describen a continuación:

La media aritmética es la más conocida de estas. Se por la fórmula siguiente (Vittinghoff *et. al.*, 2012):

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{(Ec. 1)}$$

Donde n es el tamaño de la muestra y x representa cada uno de los valores del conjunto de datos (Vittinghoff *et. al.*, 2012).

Es única, su estimación e interpretación son sencillos y los valores extremos influyen sobre ella, lo que puede generar distorsión que, en ocasiones, hará que no sea deseable como estimador (Daniel y León, 2014).

La mediana, Mo, es el dato que divide al set de datos en 2 partes iguales. Si el total de valores en análisis es un número par, se toman los dos valores centrales y se estima una media aritmética de estos, siendo esta la mediana del conjunto de datos. Es única y sencilla y los valores extremos no tienen influencia sobre ella (Daniel y León, 2014).

La moda, M_o, es el estimador más sencillo. En una serie finita de datos, es el que se encuentra con más frecuencia. La data puede ser multimodal o carecer de ella, es sencilla de estimar e interpretar y es útil al describir data categórica (Daniel y León, 2014).

2.1.1.2.2. Medidas de dispersión

La dispersión es la variabilidad de los resultados de mediciones de un suceso. Si todos fuesen iguales, no habría dispersión (Daniel y León, 2014).

A continuación, resumimos las principales medidas de dispersión.

La varianza, s^2, se describe como la variación de los valores que adquiere la data, x_i respecto de la media, μ, elevado al cuadrado y dividido por la muestra n menos 1. Se obtiene mediante la ecuación (Vittinghoff *et. al.*, 2012):

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1} \qquad \text{(Ec. 2)}$$

La desviación estándar se define como la raíz cuadrada de la varianza. Esta, a diferencia de la varianza, expresa la magnitud de la dispersión en las dimensionales originales de la serie de datos (Daniel y León, 2014).

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}} \qquad \text{(Ec. 3)}$$

El sesgo, por otra parte, según explican Groeneveld y Meeden (1984), es la asimetría de la distribución de la data. En data normalmente distribuida, $\hat{x}$, Mo y M_o son similares entre sí. Una media considerablemente mayor que la moda, se considera sesgada a la derecha (sesgo positivo), mientras que, si la media es considerablemente menor que la moda, se considera que el set de datos tiene un sesgo a la izquierda (negativo). Se estima mediante la siguiente ecuación 4, donde μ representa la media de la data, x_i representa a cada uno de los valores, n es el tamaño muestral y s es la desviación estándar:

$$skp = \frac{\sum_{i=1}^{n}(x_i-\mu)^3}{(n-1)s^3} \quad \text{(Ec. 4)}$$

Con un sesgo entre -0.5 a +0.5, la data se considera simétrica, un sesgo entre -1 a -0.5 se considera sesgado a la izquierda y un sesgo entre 0.5 a 1 se considera sesgado a la derecha. Sets de datos con sesgos menores a -1 o mayores a +1 se consideran muy sesgados (Groeneveld y Meeden, 1984).

Finalmente, la curtosis, como explica DeCarlo (1997), es el grado de la altitud que alcanza una curva de frecuencia, en su pico central, en comparación a una curva normal. Se estima mediante la siguiente ecuación:

$$k = \frac{\sum_{i=1}^{n}(x_i-\mu)^4}{(n-1)s^4} \quad \text{(Ec. 5)}$$

En la ecuación 5 cada uno de los datos está representado por x_i, μ representa la media aritmética y s la desviación estándar. La curva se puede considerar mesocúrtica si $k = 0$, leptocúrtica, si $k > 0$ o platicúrtica, si $k < 0$ (DeCarlo, 1997).

La curtosis guarda relación con la presencia de valores extremos. Una curva platicúrtica implica menos valores extremos que una distribución normal, lo contrario que una curva leptocúrtica (DeCarlo, 1997).

Diversos autores, sin embargo, cuestionan su validez, al aducir que es un estimador ambiguo y que pretender que la relación entre esta y lo picuda de una distribución sea una estimación útil es, en verdad, una afirmación engañosa (Westfall, 2014).

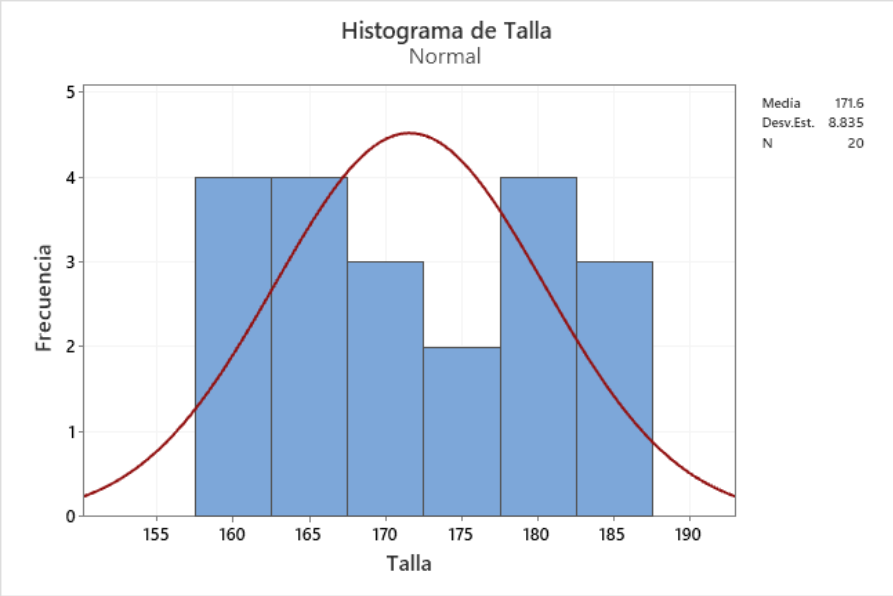
2.1.1.3. Técnicas gráficas

Según Vittinghoff *et. al.* (2012), las gráficas son un método rápido y efectivo para hacerse una idea inicial de la data. Procedemos a detallar las principales, y más utilizadas, técnicas gráficas.

La más conocida es el histograma, que corresponde a la ilustración gráfica, utilizando barras, de la distribución relativa de frecuencias de un set de datos, utilizando los puntos centrales de los intervalos y las frecuencias respectivas (Walpole *et. al.*, 2012).

La figura siguiente presenta un histograma hipotético sobre la talla de una población bimodal (incluye individuos de ambos sexos):

Figura 1. **Ejemplo de un histograma**



Fuente: elaboración propia, empleando RStudio.

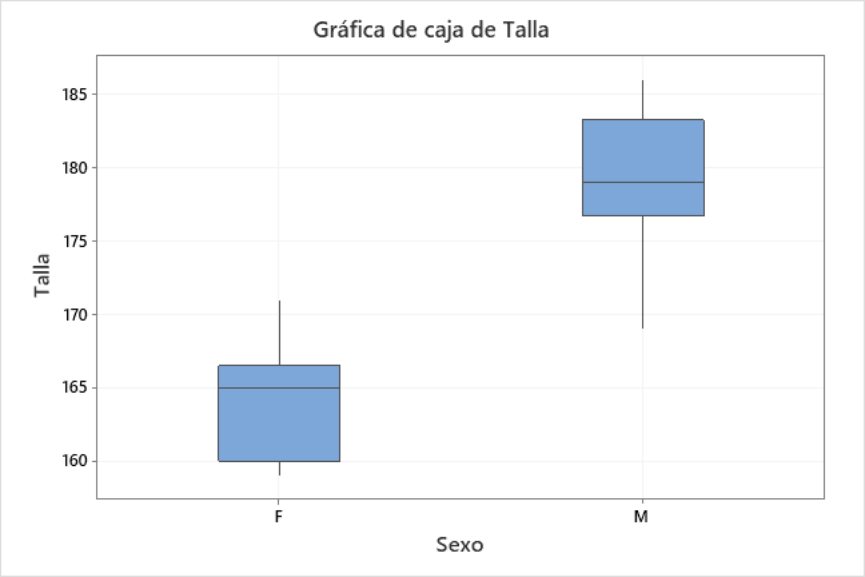
La figura 1 muestra presenta un histograma hipotético que muestra la distribución de datos de talla, en centímetros, de una población imaginaria. Se pueden apreciar dos modas: una entre 160-165 cm y la otra en 180 cm. Esto es consecuencia de que hay dos grupos diferentes: hombres y mujeres.

Otro de los gráficos más comúnmente utilizados es el de cajas y bigotes, o *boxplot*. Esta incluye el rango intercuartil de la data, representada por una caja, la cual tiene representada en su interior la mediana (Walpole *et. al.*, 2012).

Los extremos de la caja son primer y tercer cuartil, mientras que, de estos, se extienden los bigotes, que representan los valores más alejados del centro de los datos (Walpole *et. al.*, 2012).

Se suele considerar, en términos sencillos, que cualquier observación que se aleje, desde la caja, 1.5 veces el valor del rango intercuartil, se considera como un valor extremo, aunque los *softwares* modernos cuentan con diferentes formas de detectar estos valores (Walpole *et. al.*, 2012). La figura 2 muestra la versatilidad de estos gráficos.

Figura 2. **Ejemplo de un *boxplot***



Fuente: elaboración propia, empleando RStudio.

En la figura 2 se muestra el *boxplot* correspondiente a los datos que dan origen al histograma de la figura 1.

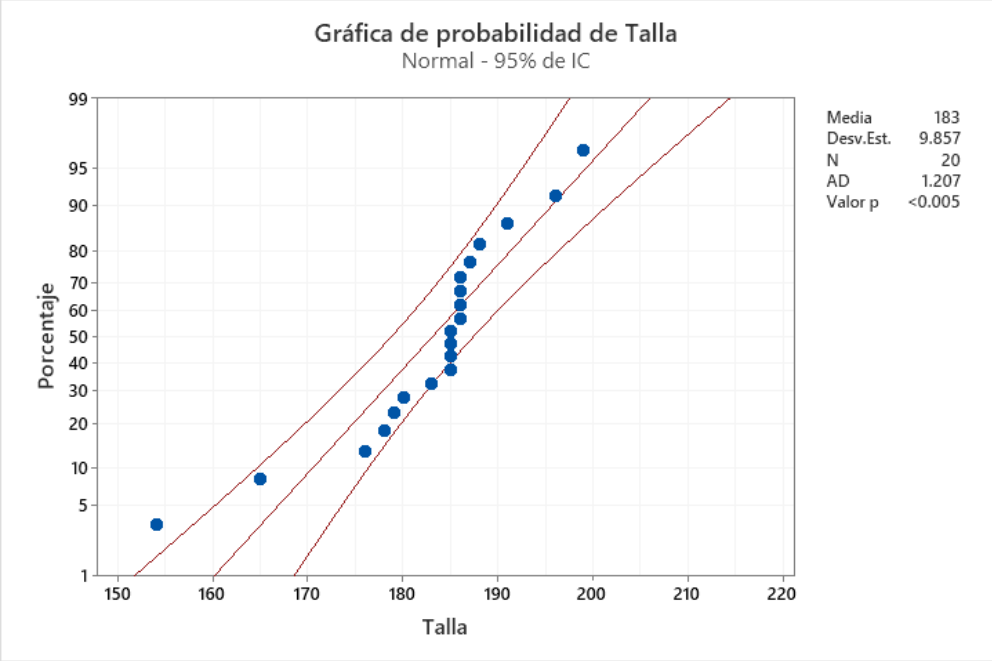
Como se puede observar, esta figura nos da información adicional: muestra la distribución de la talla (numérica, continua) según la variable sexo (categórica). Se observa la mediana (dentro de las cajas), los cuartiles 25 y 75 (extremos de las cajas) y los bigotes (valores más alejados de la mediana).

Para valorar gráficamente si un set de datos se comporta de forma esperada para una distribución de datos normal, se emplea el gráfico cuantil-cuantil (qqplot) (Vittinghoff *et. al.* 2012).

Este se construye al comparar los cuantiles de una distribución normal versus los cuantiles empíricos de la data medida y, si traza una línea recta, los

datos tendrán, muy probablemente, una distribución normal (Walpole *et. al.*, 2012). La figura 3 ejemplifica estos gráficos.

Figura 3. **Ejemplo de un qqplot**



Fuente: elaboración propia, empleando RStudio.

En la figura 6 se observa la distribución de la data de la variable talla, comparada contra los cuantiles de una distribución normal, se aprecia que no coinciden, lo que sugiere una distribución diferente, situación que se confirma con la prueba numérica incluida en el *software* ($p<0.005$).

2.1.2. Análisis de regresión

En su libro, Vittinghoff *et. al.* (2012) explica que el estudio de cómo interactúan las variables, (llamado de regresión), inicia con la representación

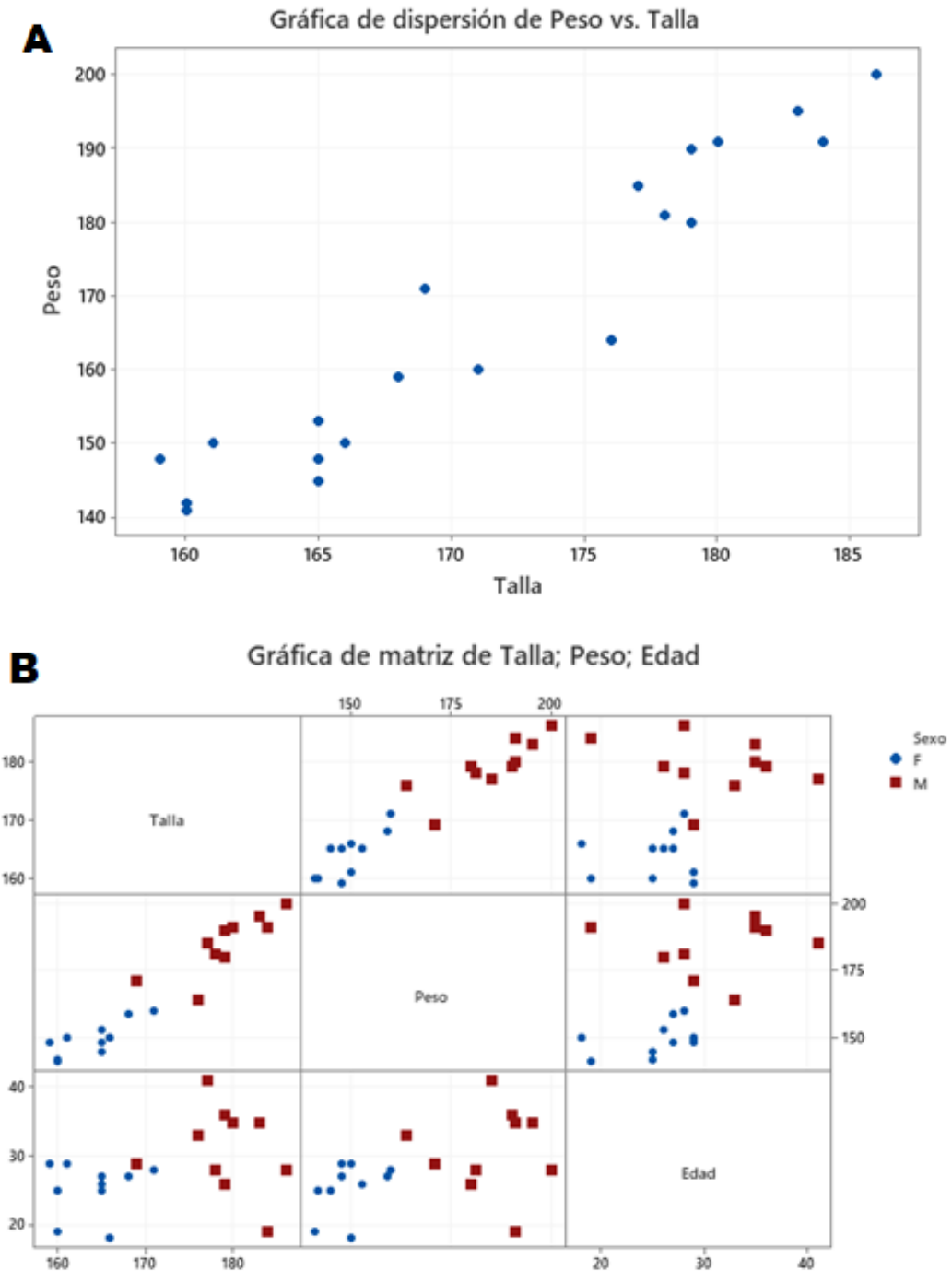
gráfica de variables: la predictora (eje de las x) versus la de respuesta (eje de las y).

Cuando se analizan relaciones entre múltiples variables predictoras y una de respuesta, puede presentarse una matriz de gráficos de dispersión (Vittinghoff *et. al.*, 2012).

Si se desea, también es posible categorizar las variables predictoras, al segmentar los datos, y establecer relaciones entre variables dentro de las categorías creadas, a modo de realizar análisis que más significativos, si se considera pertinente (Vittinghoff *et. al*., 2012).

La figura 4, que se muestra a continuación, ilustra algunos de estos conceptos:

Figura 4. **Ejemplos de gráficos de dispersión**



Fuente: elaboración propia, empleando RStudio.

En la figura 4, imagen A, se ilustra la relación entre dos variables, peso y talla, de una población hipotética, en la que se sugiere una correlación positiva. En la imagen B, se ilustra una matriz de gráficos de dispersión, del mismo set de datos hipotéticos, y se añaden dos variables: sexo y edad, esta última, variable categórica. Como se puede valorar, se aprecian diferentes grados de correlación entre las variables, que también cambian según el sexo, segmentado por colores.

Con frecuencia, es de interés estimar modelos numéricos que permitan, dentro de un grado de error aceptable, predecir el resultado de la variable respuesta de acuerdo con las variables predictoras. Este es el objeto principal del análisis de regresión: (Vittinghoff *et. al.* 2012):

$$y = \beta_0 + \beta_1 x \qquad \text{(Ec. 6)}$$

Donde β_0 es el intercepto y β_1 es la pendiente, x es el valor predictor y y representa el valor predicho.

Cuando se consideren múltiples predictores, esto deberá de añadirse al modelo matemático. La ecuación 7 añade un predictor adicional, x_2, con su pendiente respectiva, β_o. En este caso, obviamente, las pendientes son los coeficientes de regresión para cadea x (Vittinghoff *et. al.*, 2012):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad \text{(Ec. 7)}$$

Al considerar un único predictor, se realiza un análisis de regresión simple, RLS, caso contrario, uno de regresión múltiple, RLM (Vittinghoff *et. al.*, 2012). Se discutirán estos modelos.

2.1.2.1. Análisis de regresión lineal simple

Según Vittinghoff *et. al.* (2012), en la RLS únicamente se considera una variable predictora. También, debe de existir un componente aleatorio que representa factores no medibles o no conocidos por el investigador, el error. La ecuación puede considerarse:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \qquad \text{(Ec. 8)}$$

Donde ε representa el componente del error. Ahora bien, la variable y es aleatoria pues depende de ε, que, en efecto, también lo es. Por otra parte, x no es aleatoria y se puede medir con un error despreciable (Walpole *et. al.*, 2012).

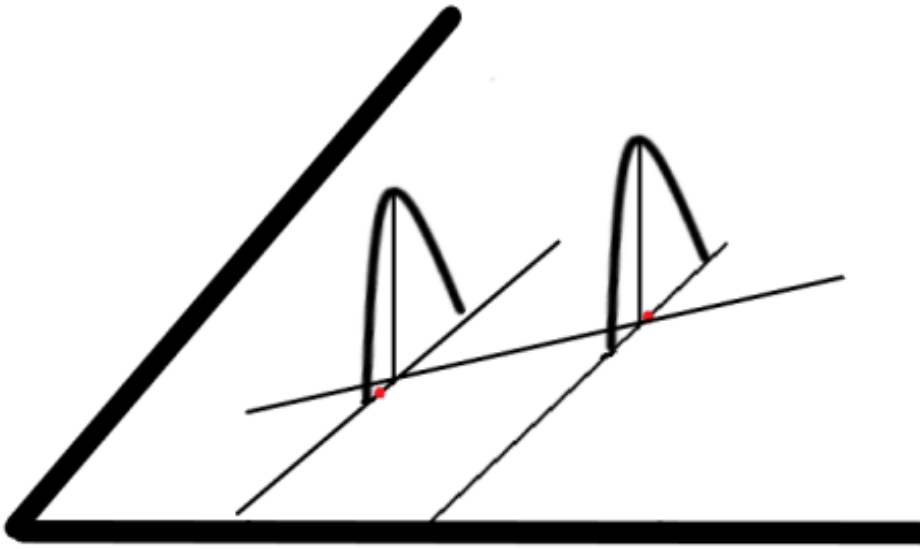
Así mismo, tanto Walpole *et. al.* (2012) como Vittinghoff *et. al.* (2012), aclaran que, en la práctica, la pendiente (β_1) y el intercepto (β_0), en la ecuación de la pendiente del modelo de RLS, no son conocidos, y se estiman mediante los datos: coeficientes de regresión, b_o y b_1.

En resumen, la línea de regresión, o línea de medias, se encontrarán los valores medios del valor de y, resultado de la interacción entre x y ε (Vittinghoff *et. al.*, 2012).

Esto es relevante, pues, al tener estimado el modelo de RLS, los n pares x_i, y_i, colocados en un gráfico de dispersión, figuran puntos dispersos alrededor de la recta resultante y cada conjunto de estos puntos se distribuye normalmente, y es el centro de dicha distribución donde pasa dicha recta (Walpole *et. al.*, 2012).

La figura 5 ejemplifica una verdadera línea de regresión.

Figura 5. **Ilustración de una línea de regresión**



Fuente: elaboración propia, empleando AutoCAD.

En la figura 5 se presenta una verdadera línea de regresión, con los datos reales, medidos, y representados alrededor de esta como puntos de color rojo. Se aprecia que la recta de regresión se dibuja sobre el parámetro central de las respuestas, μ, separado de esta por ε, que es aleatorio (Walpole *et. al.*, 2012).

De acuerdo con lo anterior, se entiende que el modelo de RLS estimado, podrá diferir del real, y, que tan alejada esté la predicción del valor real, medido, representa el error del modelo (Walpole *et. al.*, 2012).

Vittinghoff *et. al.* (2012) nos explica que el modelo de RLS asume, con respecto a la distribución del error, ε, lo siguiente: posee distribución normal, posee una varianza constante (homocedasticidad) y los valores son independientes entre sí.

Al mismo tiempo, tanto Walpole *et. al.* (2012) como Vittinghoff *et. al.* (2012) nos aclaran que no hay presunciones de este tipo con respecto a las variables predictoras, sin embargo, sí nos menciona que la presencia de valores extremos puede afectar la precisión del modelo, a la vez que una mayor variabilidad en dicho predictor puede significar una mayor representatividad de la población estudiada.

Uno de los métodos más empleados para estimar el modelo de regresión es el de mínimos cuadrados, MMC, (Vittinghoff *et. al.*, 2012).

De acuerdo con Walpole *et. al.* (2012), los estimadores b_o y b_1, para β_o y β_1 permiten aproximar los valores de y pronosticados de acuerdo con la recta ajustada $\hat{y} = b_0 + b_1 x$. En este contexto, el error al que hacíamos mención en párrafos anteriores, consistente entre la diferencia entre lo medido y lo estimado, se le conoce como residuo, y está dado por la ecuación 9, en la que e_i representa a cada uno de los residuos, y_i es el valor medido y $\hat{y_i}$ es el valor predicho de y:

$$e_i = y_i - \hat{y}_i,\ i = 1,2,\dots,n \qquad \text{(Ec. 9)}$$

Mientras más pequeños sean los residuos, mejor ajuste tiene el modelo. El MMC implica que se estimarán b_o y b_1 según las fórmulas siguientes (Walpole *et. al.*, 2012):

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{(Ec. 10)}$$

$$b_0 = \bar{y} - b_1\bar{x} \qquad \text{(Ec. 11)}$$

Vittinghoff *et. al.* (2012) es claro al explicar que, cuando el resultado se distribuye normalmente, los coeficientes de regresión tendrán, también, una distribución normal. Esto conlleva que la proporción de b_1 contra s tendrá una distribución t con $n-2$ GdL, para establecer $H_0: \beta_1 = 0$, esto es, que no hay pendiente. En este contexto, las estimaciones del p valor y de los intervalos de confianza se hacen en la forma usual, lo que permite confirmar o descartar H_0 que dice que no hay relación sistemática entre la variable regresora y el resultado.

Complementario al estudio de regresión, es importante estimar los coeficientes respectivos. El más usado, el de Pearson, según Vittinghoff *et. al.* (2012), es una medición adimensional que estima la relación lineal entre dos eventos de interés. Se estima la covarianza de x, y y esta se divide por el producto de las desviaciones estándar de x y y:

$$r(x,y) = \frac{Cov_{(x,y)}}{s_X s_y} = \frac{\sum_{i=1}^{n} {(x_i-\bar{x})(y_i-\bar{y})}/{(n-1)}}{\sqrt{\sum_{i=1}^{n} {(x_i-\bar{x})^2}/{(n-1)}}\sqrt{\sum_{i=1}^{n} {(y_i-\bar{y})^2}/{(n-1)}}} \qquad \text{(Ec. 12)}$$

Tanto Walpole *et. al.* (2012) como Vittinghoff *et. al.* (2012) sugieren que $r = 0$ refleja ausencia de relación, $r = 1$ refleja relación lineal positiva absoluta y $r = -1$ una relación negativa absoluta. Una relación nula, o casi nula, no descarta una relación no lineal (cuadrática o de mayor magnitud, o de otros tipos).

Existen alternativas, que pueden dar una mejor estimación, en caso de distribución no normal de los datos, o presencia de valores extremos (Vittinghoff *et. al.* 2012).

El coeficiente de Spearman es equivalente al coeficiente de Pearson, aplicado a datos ranqueados de x e y, y también toma valores de -1 a +1 (Vittinghoff *et. al.*, 2012).

El coeficiente τ de Kendall, se basa en la concordancia de los pares de x e y (Vittinghoff *et. al.* 2012).

Si $x_i > x_j$ y $y_i > y_j$, hay concordancia, caso contrario, si $x_i > x_j$ y $y_i < y_j$, hay discordancia. Obviamente, no se necesitan conocer los valores de x e y, sólo sus posiciones en el ranqueo y si hay o no concordancia. Si la proporción de los valores concordantes es muy parecida a la de los valores discordantes, $\tau \approx 0$ y esto quiere decir que el ranqueo de los datos da poca información, mientras que, al acercarse a +1 o -1, la información aumenta y el indicador es más útil (Vittinghoff *et. al.*, 2012).

Finalmente, el coeficiente de determinación, R^2, representa qué tanto la variabilidad de y, que es explicada por la variabilidad de los predictores, que puede entenderse como la correlación entre Y y $\hat{y}$ (Vittinghoff *et. al.*, 2012).

Según Walpole *et. al.* (2012), el R^2 ajustado, $R^2_{adj} = 1 - \frac{RSS}{SST}$, es preferente al R^2 usualmente medido (Vittinghoff *et. al.*, 2012).

2.1.2.2. Análisis de regresión lineal múltiple

De acuerdo con Vittinghoff *et. al.* (2012), el modelo RLM es una generalización de RLS, en tanto que puede entenderse como la ecuación 13:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \qquad \text{(Ec. 13)}$$

En la que hay p variables regresoras.

Vittinghoff *et. al.* (2012) explica que los coeficientes de regresión pueden interpretarse así: β_0 es el punto en que todos los predictores son iguales a cero y $\beta_1,\ \beta_2,\ \beta_3$, etc. hacen referencia al cambio en y que es resultado del incremento de una unidad del predictor x respectivo, y ε representa el error promedio asociado a todos los predictores. Se asume, al igual que en MRS, que ε se distribuye normalmente, con una media de $\hat{x} = 0$ y varianzas iguales para los términos de $\hat{x}$ (Walpole *et. al.*, 2012).

El hecho de incluir múltiples predictores sí afecta la varianza de $\hat{b}_1$, que ahora también depende de un factor adicional, r_j, que representa la correlación múltiple de las diferentes x_j entre sí (Vittinghoff *et. al.* 2012):

$$var(\hat{b}_j) = \frac{\sigma^2_{y|x}}{(n-1)\sigma^2_{x_j}(1-r_j^2)} \qquad \text{(Ec. 14)}$$

Donde $\sigma^2_{y|x}$ es la varianza residual del resultado y $\sigma^2_{x_j}$ representa la varianza de x_j, r_j equivale a $r = \sqrt{R^2}$ de un modelo de RLM en el que x_j es regresado ante todos los predictores relacionados (Vittinghoff *et. al.* 2012).

El término $1/(1 - r_j^2)$ se conoce como factor de inflación de la varianza, FIV, ya que $var(\hat{b}_j)$ se incrementa en el grado en que x_j se correlaciona con los predictores del modelo (Vittinghoff *et. al.*, 2012).

2.1.2.3. Criterios de información

Los criterios de información son funciones que analizan la capacidad sintética de un modelo que se moverá en dirección contraria a la complejidad de dicho modelo (Lindsey y Sheather, 2010). Se mencionarán 4 criterios.

2.1.2.3.1. R^2 ajustado

Ya descrito en la sección previa, se considera que, a medida que este incrementa, el modelo se hace más deseable (Lindsey y Sheather, 2010).

Algunos autores sugieren penalizar por mayores números de predictores (Lindsey y Sheather, 2010), por ejemplo:

$$R^2_{adj} = 1 - \frac{n-1}{n-k-1}\frac{RSS}{SST} \qquad \text{(Ec. 15)}$$

Donde RSS representa la suma de los cuadrados de los residuos y SST la suma de los cuadrados totales.

2.1.2.3.2. Criterio de Akaike (AIC)

Al contrario que el R_Adj^2, mientras menor sea el AIC, mejor ajuste tiene el modelo. El poder del modelo, según este criterio, se explica por la maximización de la probabilidad logarítmica de los coeficientes de los predictores y la varianza del error (Lindsey y Sheather, 2010).

$$\begin{aligned} AIC &= 2\left\{-logL\left(\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_p}, \hat{\sigma}^2 | Y\right) + k + 2\right\} \\ &= n \log \frac{RSS}{n} + 2k + n + n \log(2\pi) \end{aligned} \qquad \text{(Ec. 16)}$$

Donde k representa el número de los parámetros del modelo, RSS la suma de los cuadrados de los errores, n el número de observaciones.

2.1.2.3.3. Cp de Mallows

Si $p = k + 1$, y definimos RSS_{FULL} como el RSS del modelo completo (versus el simplificado), con m posibles predictores (Lindsey y Sheather, 2010):

$$C_p = (n - m - 1)\frac{RSS}{RSS_{FULL}} - (n - 2p) \qquad \text{(Ec. 17)}$$

Se considera que los mejores modelos tienen $C_p \approx p$, criterio que siempre será satisfecho por el modelo completo. Se preferirán los modelos con el valor más pequeño de C_p (Lindsey y Sheather, 2010).

2.1.2.3.4. Criterio de información Bayesiano (BIC)

Similar al AIC excepto que ajusta para el término de la penalidad por la complejidad, basado en n, pero utilizando los mismos términos (Lindsey y Sheather, 2010).

$$\begin{aligned} BIC &= -2logL\left(\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_p}, \hat{\sigma}^2 | Y\right) + (k+2)\log n, \\ &= n\log\frac{RSS}{n} + k\log n + n + n\log(2\pi) \end{aligned} \qquad \text{(Ec. 18)}$$

En general, es complicado elegir el mejor criterio de información. Utilizar el R^2_{adj} como criterio único lleva a elegir modelos más complejos de lo necesario, mientras que, para el resto, el modelo que más disminuye el RSS suele ser el

mejor, lo que puede simplificar, en cierto grado, la selección del mejor modelo (Lindsey y Sheather, 2010).

2.1.2.4. Regresión de variables de conteo: modelo de Poisson

De acuerdo con Rodríguez (2007), en los casos de variables respuesta que representan conteo de eventos, no una variable continua, el análisis estadístico puede, y suele, hacerse bajo los supuestos de una distribución discreta de Poisson (DP), que también adquiere relevancia en estudio de tablas de contingencia o de supervivencia.

En general, el producto de los procesos analizados bajo este contexto son valores enteros de conteo, de naturaleza aleatoria, es decir, la frecuencia de la obtención de un resultado específico en un contexto específico (Vittinghoff *et. al.*, 2012).

Según Walpole *et. al.* (2012) , este proceso posee propiedades específicas:

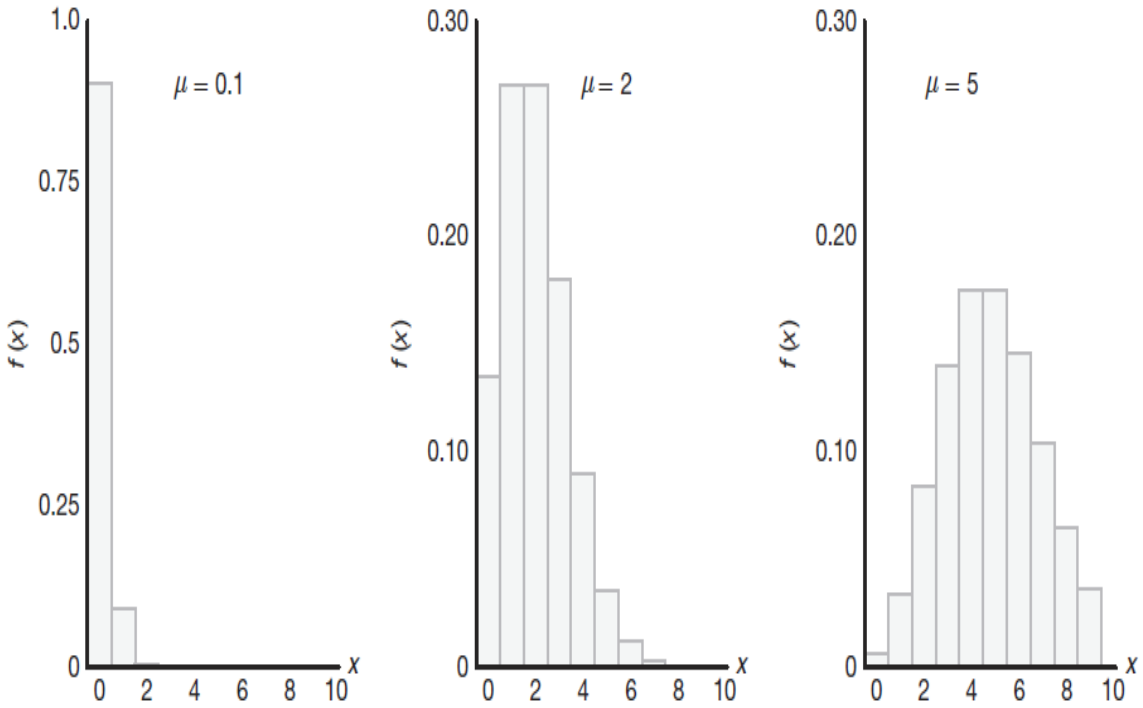
- La frecuencia de eventos observados en un contexto específico no depende de la frecuencia de eventos en otro contexto.

- La probabilidad de un único evento está relacionada con la magnitud del período o región contextuales.

- La probabilidad de más de un evento en un contexto definido y pequeño es despreciable.

Tanto Vittinghoff *et. al.* (2012) como Walpole *et. al.* (2012) aclaran que la frecuencia de eventos en un intervalo o región tiene naturaleza aleatoria, y su distribución de probabilidad es la DP. A continuación, se presenta la DP.

Figura 6. **Funciones de densidad de Poisson para medias diferentes**



Fuente: Walpole, Myers y Myers (2012). *Probabilidad y estadística para ingeniería y ciencias.*

La DP está dada por:

$$Pr\{x; \lambda t\} = \frac{e^{-\lambda t}(\lambda t)^x}{x!} \qquad \text{(Ec. 19)}$$

Donde λ es la frecuencia promedio de eventos en el tiempo t y e es la constante respectiva. De acuerdo con lo descrito hasta ahora, la DP guarda relación con la distribución binomial (Walpole *et. al.*, 2012).

Así, los modelos típicos de regresión lineal, que se estiman sobre supuestos específicos de la distribución de residuos (distribución normal, homocedasticidad e independencia), cuando buscan predecir o explicar una variable de conteo, violan estos supuestos. La principal transgresión es la presencia de heterocedasticidad, pues, como suele observarse, la varianza de la data suele incrementarse conforme aumenta su media y esto lleva a la generación de sesgo del error estándar y de las pruebas de significancia (Coxe, West y Aiken, 2009).

La distribución condicionada también transgrede el primer supuesto, pues suele sesgarse positivamente y mostrarse platicúrtica. En general, esto produce un aumento en el error tipo I, lo que afecta el poder estadístico del modelo de detectar verdaderos positivos (Coxe, West y Aiken, 2009).

Tal y como explican Coxe, West y Aiken (2009), la regresión de Poisson (RP) pertenece a la familia de los modelos lineales generalizados. Estos generalizan la regresión ordinaria de mínimos cuadrados para su uso con diversos tipos de estructura de error y variables dependientes, mediante dos modificaciones importantes:

- Permite la transformación de la variable dependiente, para así linealizar relaciones no lineales, lo que implica que los valores predichos pueden estar dimensionados en unidades diferentes a las de los valores medidos. En la RP, la función de transformación es el logaritmo natural.

- Estos modelos son flexibles al analizar la estructura del error. Mientras que los modelos tradicionales presuponen normalidad de los residuos, la RP se presume sobre la DP.

De acuerdo con lo expuesto en párrafos previos y la ecuación 15, la DP se considera ideal para modelar conteos de sucesos, al ser discreta y que únicamente refleja probabilidad para enteros no negativos. Analizar datos de esta naturaleza mediante un análisis de RLS o RLM implicaría que se puedan obtener resultados negativos cuando estos sucesos sólo pueden presentarse en valores positivos enteros. Mientras que la distribución normal se describe por medio de μ y s^2, la DP se describe sólo por su μ, en tanto que μ y s^2 se suponen iguales (Coxe, West y Aiken, 2009).

Conforme μ aumenta, la DP se asemeja a la normal. Se considera que con $\mu = 10$ la forma y simetría se aproximan a la normal, aunque la distribución de Poisson tiene la ventaja, en los casos que así lo requieran, que permite modelar data discreta (Coxe, West y Aiken, 2009).

El modelo de RP es el siguiente:

$$\ln(\hat{\mu}) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \qquad \text{(Ec. 20)}$$

Empero las semejanzas con el MRL (ecuaciones 6 y 7), y lo dicho hasta el momento, las predicciones en la RP no son valores de conteo, sino el Ln de dicho valor (Coxe, West y Aiken. 2009).

Esto puede conllevar cierta dificultad en la comprensión de los coeficientes, para aquellos que deseen aplicar el modelo. Para solucionar esto, se puede recurrir a una sencilla manipulación algebraica de la ecuación de regresión:

$$e^{\ln(\hat{\mu})} = e^{(b_0+b_1x_1+b_2x_2+\cdots+b_px_p)} \quad \text{(Ec. 21)}$$

Que puede interpretarse así:

$$\hat{\mu} = e^{(b_0+b_1x_1+b_2x_2+\cdots+b_px_p)} \quad \text{(Ec. 22)}$$

Que, a su vez, puede ser simplificada a la ecuación:

$$\hat{\mu} = e^{b_0}e^{b_1x_1}e^{b_2x_2} \ldots e^{b_px_p} \quad \text{(Ec. 23)}$$

La que nos muestra, de forma más clara, el impacto que cada predictor tiene en el resultado predicho, así como expresa el predicho en las dimensionales de interés, lo que facilita su interpretación (Coxe, West y Aiken, 2009).

La RP, sin embargo, presenta limitaciones relevantes que restringen su uso en la práctica. La más importante, es la sobredispersión, esto es, cuando la varianza condicional de los residuos es mayor que la media de los predichos (Vittinghoff *et. al.*, 2012).

En general, las principales razones para que haya sobredispersión son las siguientes:

- Diferencias individuales no tomadas en cuenta por el modelo de regresión.
- Que los conteos de eventos para cada intervalo no sean verdaderamente independientes.

Como solución a lo anterior, existen alternativas a la RP estándar, especialmente los modelos de regresión de Poisson sobredispersa y de regresión binomial negativa (RBN). Se discute esta última en los siguientes párrafos.

2.1.2.4.1. Regresión binomial negativa

La RBN toma en consideración la sobredispersión que el modelo estándar de Poisson ignora, al asumir que habrá heterogeneidad (variabilidad) inexplicada o inexplicable, entre los individuos o intervalos que tengan el mismo valor predicho. Esta heterogeneidad tendrá efecto en la varianza, más no en la media, lo que genera la sobredispersión. Conceptualmente, esta variabilidad adicional es análoga a la inclusión de ε en el MRL (Coxe, West y Aiken, 2009).

En general, el modelo binomial negativo permite a dos intervalos o individuos con los mismos predictores ser modelados por DP con μ diferentes, diferente a la RP estándar. En estos casos, se asume que la distribución de los parámetros μ de los diferentes intervalos o individuos sigue una distribución Gamma (Coxe, West y Aiken, 2009).

De esta forma, la media condicional de los resultados es idéntica en ambos casos (RP y RBN), pero la varianza condicional será mayor en el modelo de RBN (Coxe, West y Aiken, 2009).

En la RBN, la varianza está dada por $\mu + \alpha\mu_2$, donde α representa la sobredispersión. Con $\alpha > 0$, existe sobredispersión y, su magnitud es mayor cuanto mayor es la sobredispersión. Obviamente, con un $\alpha = 0$ el modelo se reduce a una RP. Los coeficientes del modelo se analizan igual en ambos modelos (Coxe, West y Aiken, 2009).

Los modelos de RP y RBN, por su parte, no pueden ser comparados al añadir o remover predictores; al modificarlos, se modifica también α, lo que deriva en que no se pueden considerar como modelos verdaderamente anidados (Coxe, West y Aiken, 2009).

Sin embargo, los criterios de información AIC y BIC sí pueden ser estimados, pues estos modelos se componen de dos partes aditivas: la primera es la función logarítmica de la reducción de probabilidad y la segunda es la referente al número de parámetros (coeficientes), que en el BIC añade la información referente al tamaño de la muestra (Coxe, West y Aiken, 2009).

2.2. Infección por el SARS-CoV2

A fines de 2019, un clúster de casos de neumonía secundarios a un nuevo betacoronavirus, en Wuhan, China, capaz de causar enfermedad severa, similar al SARS, marcó el inicio de una pandemia de importantes repercusiones sanitarias y económicas a nivel global. Más adelante, se determinaría que se trataba de un nuevo coronavirus, el SARS-CoV2 (Huang *et. al.*, 2020).

2.2.1. Coronavirus

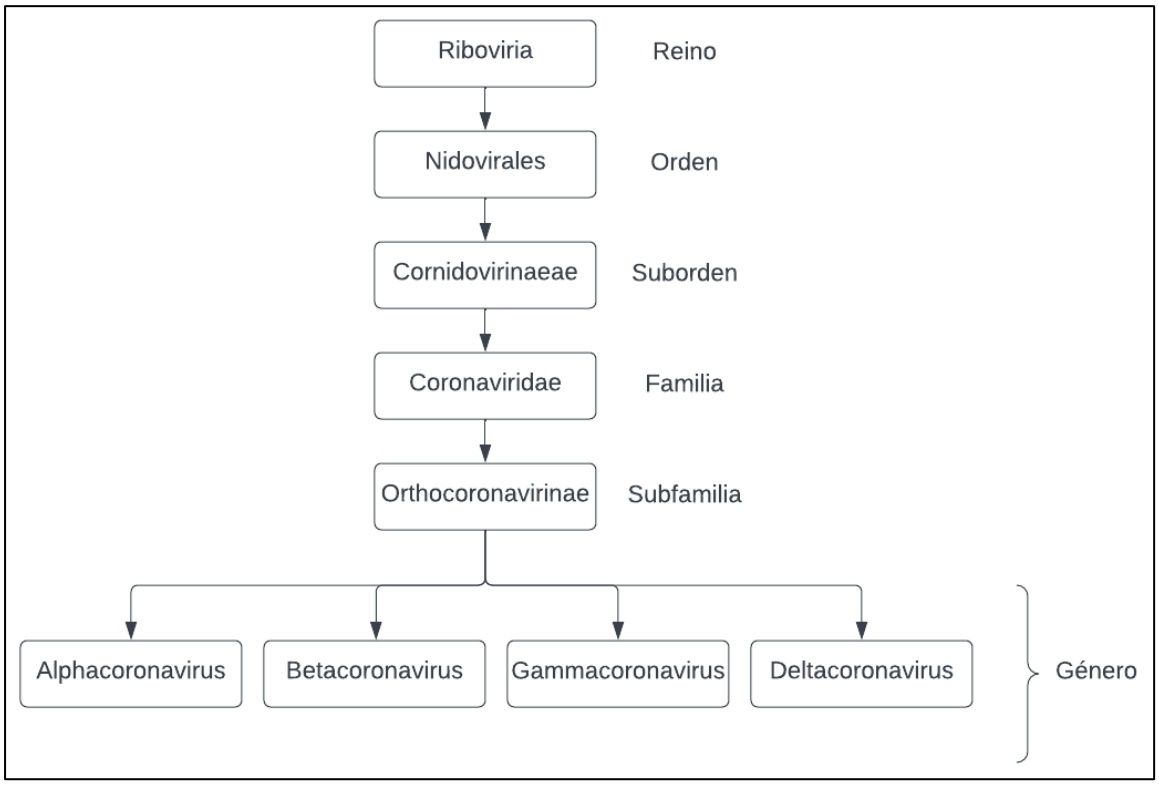
Desde el descubrimiento del primer coronavirus (CoV) en 1,937 (virus de bronquitis infecciosa), el principal punto de inflexión en la cantidad de investigación dirigida a esta familia es el aparecimiento del SARS, en el año 2,008. De esta cuenta, antes del año 2,003, se tenían plenamente identificados y secuenciados los genomas de 10 coronavirus (2 CoV humanos, 7 CoV de otros mamíferos y 1 CoV aviar), mientras que, después del 2,008, se añadieron 16 CoV más con genoma completo secuenciado: 2 CoV humanos, 10 CoV de otros mamíferos y 4 CoV aviares (Woo, Lau, Huang y Yuen, 2009).

Considerando que los CoV son virus ARN, dependientes de ARN polimerasa, con una elevada tasa de errores en su replicación (1,000 – 10,000 por nucleótido replicado) con alta tasa de recombinación homóloga de su material genético, los CoV han alcanzado una importante diversidad de especies y

genotipos, lo que les permite adaptarse a nuevos hospederos y nichos ecológicos y los hace capaces de generar brotes zoonóticos (Woo, Lau, Huang y Yuen, 2009).

Con respecto al virus SARS-CoV2, este pertenece, al igual que los virus causales del SARS y del MERS, a los betacoronavirus (Aydogdu et al., 2021). La taxonomía de los CoV se muestra en la figura 7.

Figura 7. **Taxonomía de los Coronavirus**



Fuente: elaboración propia, con datos de Aydogdu *et. al.*, (2021). *Surface interactions and viability of coronaviruses.*

2.2.2. Estructura del SARS-CoV2, mecanismos de patogenicidad

Hace más de 50 años que se sabe que la estructura básica de los virus consiste en proteína y material genético. Ya en 1956, Watson y Crick, en la revista *Nature*, hacían este tipo de aseveraciones referente al virus del tabaco. El SARS-CoV2 posee 4 proteínas estructurales: S, E, M y N y 16 proteínas no estructurales (ns 1-16) (Wang *et. al.*, 2020).

El SARS-Cov2 es un virus ARN monocatenario, con un genoma de 29.9 kb. El material genético está encapsulado por las proteínas de la nucleocápside (proteína N), que son altamente antigénicas y se asocian con diversas interacciones con el sistema inmune. En general, el complejo de ribonucleoproteína es a lo que nos referimos como nucleocápside (Chen *et. al.*, 2007).

La glicoproteína S es la mediadora de la entrada a la célula. Está compuesta de dos subunidades: S1, que es la que contiene la información de reconocimiento de dominio y la S2, que es la que genera la fusión con la membrana de la célula que van a invadir para poder, en efecto, invadirla. Al igual que el SARS-CoV, mediante su subunidad S1, reconoce al receptor de ECA2 como región de entrada a la célula (Wang *et. al.*, 2020).

Esto ha sido propuesto como potencial mecanismo explicativo para la excesiva severidad en hombres, ancianos e hipertensos, que expresan mayores cantidades de la proteína ECA2 en tejidos críticos (Li, Zhou, Yang y You, 2020).

La proteína de membrana (proteína E) es una proteína corta (entre 76-109 aminoácidos), con actividad de viroporina de canal iónico, con acción relevante

en el ensamblaje del genoma viral de las nuevas partículas que se forman dentro de la célula invadida (Aydogdu *et. al.*, 2021).

Es una proteína transmembrana, con importante rol en la virulencia y la patogenicidad del virus (Schoeman y Fielding, 2019).

La proteína M, transmembrana, guarda alta similitud con la proteína M de todos los miembros de la familia Coronaviridae. Esta conservación a través del árbol taxonómico sugiere que es altamente relevante para la estructura del virus. Su región exterior exhibe la parte N-terminal, mientras que la interior, la parte C-terminal. Esta parte de carboxilo terminal parece ser la más antigénica de la proteína. La región N terminal externa es también responsable de ayudar al proceso de fusión llevado a cabo, en mayor parte, por la proteína S, para la entrada del virus a la célula (Hu *et. al.*, 2003).

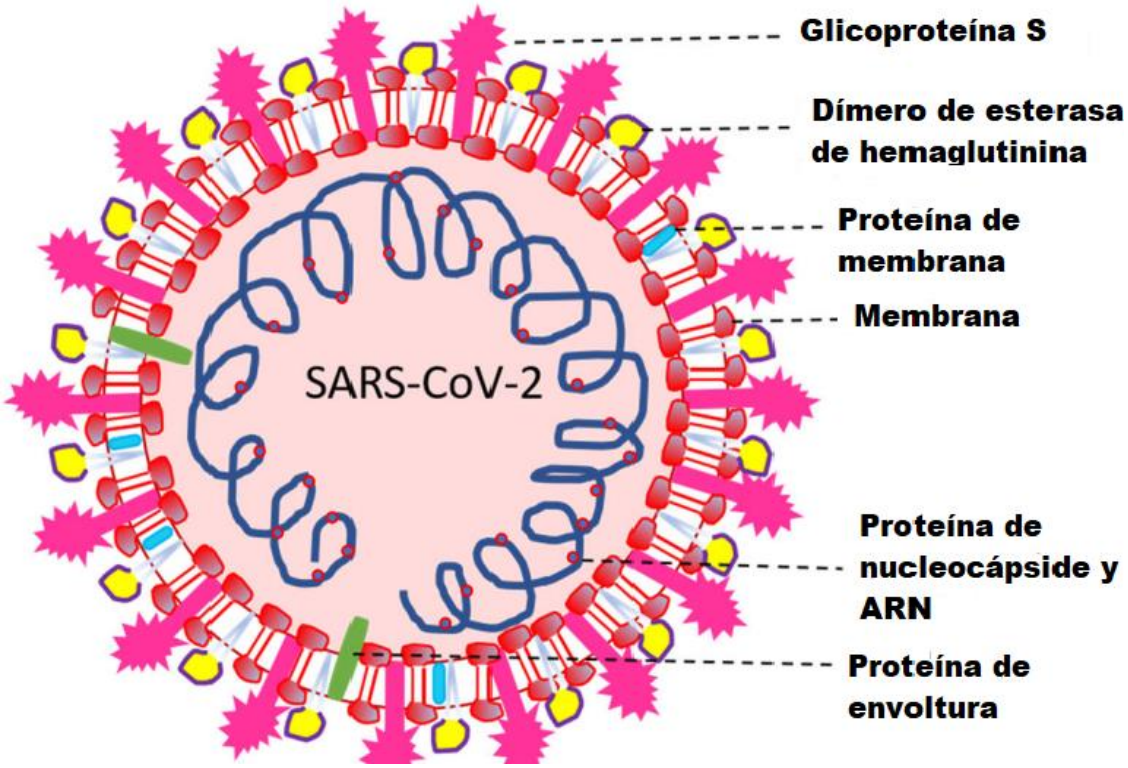
Poseen también una hemaglutinina esterasa (HE), que es común en los betacoronavirus. Esta, que se cree fue adquirida por recombinación no homóloga de ARN, comparte hasta un 30 % de la secuencia de la HE del virus de la influenza tipo C. En general, poseen también una estructura que semeja una espícula, aunque más pequeña que la proteína S Su función es la de reconocer el receptor O-acetyl SA, en el proceso de la interacción virus-hospedero (Kim, 2020).

Así mismo, y en similitud a la mayoría de los virus ARN (los retrovirus son la excepción), el SARS-CoV2 requiere de una ARN polimerasa dependiente de ARN (APdA), proteína de núcleo catalítico. Esta se asocia con algunas de las proteínas no estructurales, formando el complejo de replicación-transcripción (Hillen, 2021).

Esta proteína, crítica para el ciclo del virus, se consideró como posible diana terapéutica desde el descubrimiento de este y, tras varios meses de investigación, la droga Molnupiravir®, comercializada por la farmacéutica MSD, fue aprobada por la FDA para su uso em COVID-19.

La siguiente figura muestra un esquema de la estructura del virus:

Figura 8. **Estructura del SARS-CoV2**



Fuente: Boopathi, Poma y Kolandaivel, (2020). *Novel* 2019 *coronavirus structure, mechanism of action, antiviral drug promises and rule out against its treatment.*

2.2.3. Evolución de la pandemia

Según Lu *et. al.* (2020), al representar una emergencia mundial, importantes esfuerzos permitieron alcanzar hitos importantes en períodos de tiempo relativamente cortos. Por ejemplo, apenas 2 meses después de detectado el inicio de la, en ese entonces, epidemia, ya se había determinado el genoma del nuevo virus, así como ya se hipotetizaba sobre el papel que jugaba la enzima ECA2 en la entrada a la célula de dicho virus. Esto se confirmaría poco tiempo después.

Los investigadores también, rápidamente, notaron que el virus se transmitía de forma exponencial, y, antes que se detectaran casos fuera de China, ya se conocía el potencial del virus de expandirse rápidamente, y el peligro que suponía para los demás países, con lo que sugirieron, inmediatamente, que se necesitarían medidas extremas de salud pública para contener, o evitar, la diseminación del virus (Wu *et. al.*, 2020).

Así mismo, antes que se cumplieran los 2 meses de inicio de la epidemia, análisis epidemiológicos del clúster inicial, soportados por análisis microbiológicos y clínicos, y seguimiento de los casos nuevos, rápidamente demostraron la transmisión persona a persona (Chan *et. al*., 2020).

Prontamente, se conocería que es altamente contagioso, y que cada persona infectada puede contagiar hasta 5 personas más (Ro >1, hasta 5), con un período de incubación de 1-14 días, media de 5-6 (Del-Río *et. al.*, 2020).

Ya en marzo de ese año, poco más de tres meses después de la detección del brote original, se reportaba el primer caso conocido de transmisión persona a

persona en Estados Unidos, así como ya se contaban por decenas de miles los casos y los muertos en China y regiones de Europa (Ghinai *et. al.*, 2020).

La pandemia llegaría a Guatemala a mediados de marzo del mismo año. A finales de enero del 2021, poco más de un año después, las pruebas diagnósticas están disponibles en todo el mundo y el síndrome clínico ha sido caracterizado. Se sabe que la edad avanzada, el sexo masculino, la presencia de enfermedades crónicas, enfermedades cardiacas y/o pulmonares asociadas y mayor concentración de marcadores inflamatorios y de coagulopatía se asocian a enfermedad más severa y mortalidad (Cummings *et. al.*, 2020).

Así mismo, ya se cuentan con algunos tratamientos que ayudan a disminuir, modestamente, la mortalidad y severidad del cuadro. Básicamente, se cuenta con el antiviral remdesivir, capaz de reducir el tiempo a recuperación, aunque no reduzca la mortalidad asociada en forma estadísticamente significativa, y el esteroide dexametasona, que sí logró reducir, en 36 %, la mortalidad a 28 días, comparado con placebo. Luego, los antivirales molnupiravir y paxlovid, también obtendrían aprobación de la FDA (Beigel *et. al.*, 2020).

Otras terapias, como la hidroxicloroquina, azitromicina e ivermectina, aunque populares en redes sociales, no cuentan con autorización de la FDA pues no existen estudios que sustenten su utilización en COVID-19 (Annie *et. al.*, 2020).

2.2.4. Medidas para mitigar la pandemia

A pesar de numerosos estudios de intervención farmacológica, las medidas más efectivas para contener la pandemia han sido, previo a las vacunas, las medidas de salud pública, aplicadas tanto a nivel poblacional, con

distanciamiento social y uso de mascarillas, como a nivel de instituciones de cuidados de la salud, con uso de mascarillas y cobertura de tejidos mucosos expuestos, lo que lleva a estimar reducciones drásticas en la transmisibilidad, de hasta más de 85 %, con la simple aplicación de estos métodos (Chu *et. al.*, 2020).

En el contexto del distanciamiento social, se ha hipotetizado que las medidas de restricción de la movilidad pueden potenciar este factor, y llegar a reducir la transmisibilidad de casos. Se han hecho análisis de los registros de movilidad de dispositivo celulares y su correlación con incidencia de casos, lo que sugiere que estas medidas son efectivas (Sulyok y Walker, 2020).

Finalmente, el método que se espera sea definitivo para resolver la emergencia global, es el de la vacunación. En tiempo relativamente corto, y como un hito en la historia de la medicina moderna, en menos de 1 año se han aprobado tres vacunas, de dos tecnologías diferentes, con adecuados grados de efectividad y seguridad.

Las vacunas, conocidas, de acuerdo con el nombre de la casa farmacéutica que las han desarrollado, como la de Astra/Zéneca, la de Pfizer y la de Moderna, lideran el mundo, que aún espera los resultados de los ensayos de Fase III de otras tantas vacunas que han avanzado en las etapas de investigación clínica y podrían estar disponibles en escasas semanas o meses.

3. PRESENTACIÓN DE RESULTADOS

A continuación, se presentan los resultados que dan respuesta a los objetivos planteados.

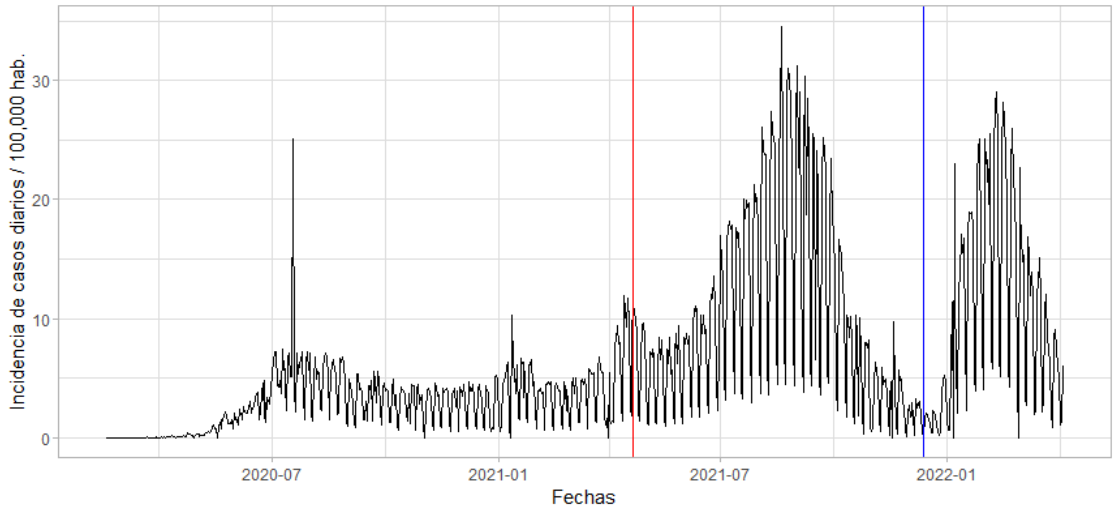
3.1. Objetivo 1. Describir cuál ha sido el comportamiento de la incidencia bruta de casos nuevos de COVID-19 en la población de Guatemala

El primer caso de COVID-19 en Guatemala se registró el día 13 de marzo de 2020. El reporte de casos se ha actualizado cada día desde ese entonces.

Se sabe, de acuerdo con diversos estudios, que la tasa de crecimiento de la incidencia de casos se comporta de forma exponencial, si no hay medidas que mitiguen la transmisión del virus. En Guatemala se implementaron varias medidas que intentaron controlar la tasa de crecimiento de la pandemia.

A continuación, se presenta el Gráfico que muestra el comportamiento de la incidencia de casos diarios, por cada 100,000 habitantes, en el país. Para esto se consideró una población total de 16,860,000 habitantes:

Figura 9. **Evolución temporal de la incidencia de COVID-19 en Guatemala, marzo de 2020 a abril de 2022**

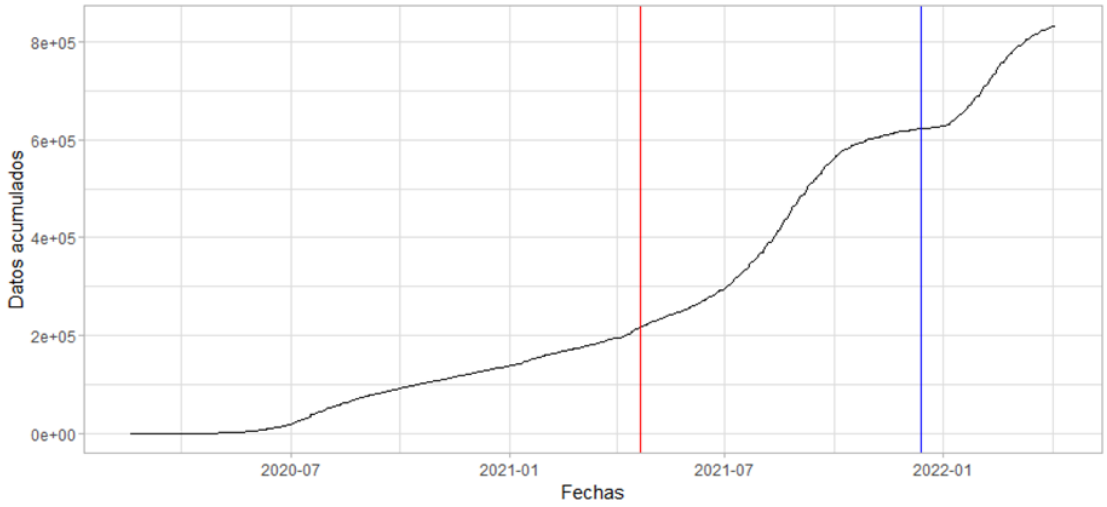


Fuente: elaboración propia, empleando RStudio.

La línea vertical roja hace referencia a la primera detección de la variante Delta en Guatemala (06 de abril de 2021) y la línea azul a la primera detección de la variante Ómicron en Guatemala (12 de diciembre de 2021). Se aprecia claramente que el aparecimiento de ambas variantes se correlaciona cronológicamente con las mayores olas de contagios.

Entre el período correspondiente a la detección del primer caso en Guatemala y la fecha en que esto fue escrito (06 de abril de 2022), transcurrieron 754 días y se registraron 832,956 casos. La gráfica siguiente muestra los casos acumulados durante este tiempo:

Figura 10. **Casos de COVID-19 acumulados en el tiempo, en Guatemala, marzo de 2020 a abril de 2022**

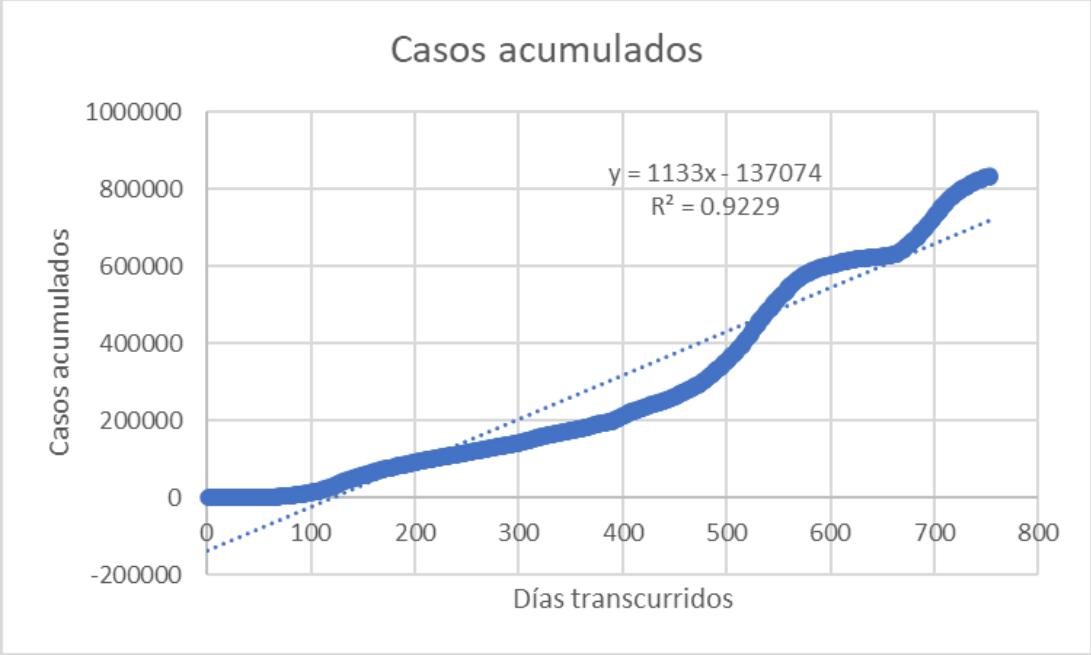


Fuente: elaboración propia, empleando RStudio

La figura también incluye la información referente a las primeras detecciones de casos de las variantes Delta y Ómicron y las acodaduras respectivas en la curva representan aumento en la incidencia secundario al aparecimiento de las variantes.

La tendencia del crecimiento de los casos siguió una línea de tendencia de crecimiento lineal, con $y = 1133x - 137074, R^2 = 0.9229$:

Figura 11. **Crecimiento de los casos acumulados de COVID-19 en Guatemala, marzo 2020 a abril 2022**



Fuente: elaboración propia, empleando Microsoft Excel.

En la siguiente tabla se muestran los datos que resumen la incidencia diaria de casos:

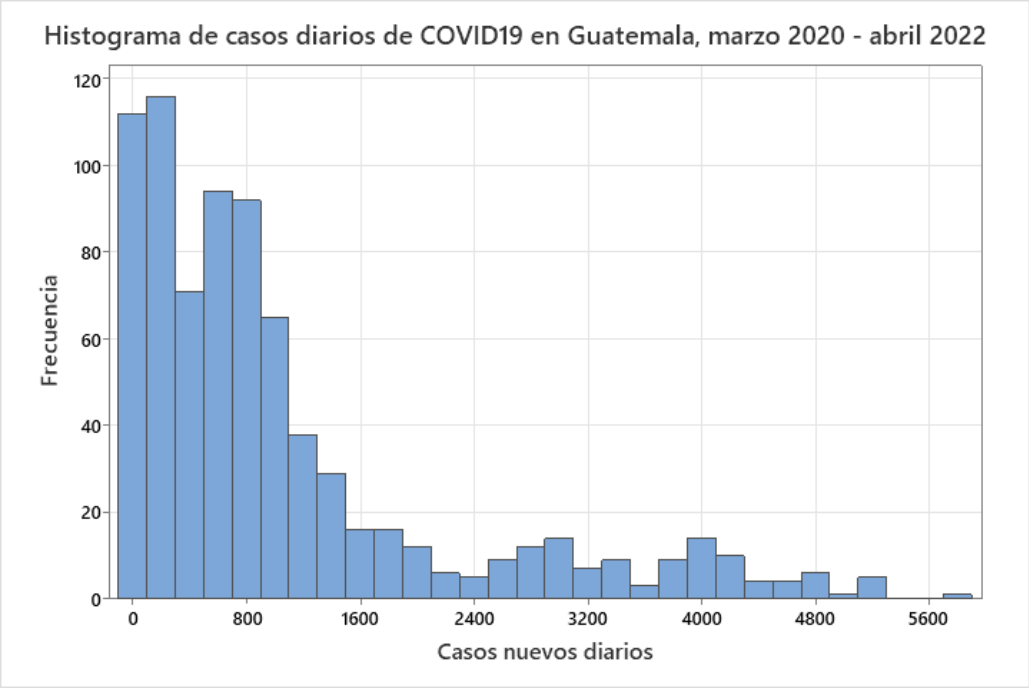
Tabla II. **Resumen de la variable Incidencia diaria de casos de COVID-19, Guatemala, marzo de 2020 a abril de 2022**

Parámetro	Resultado	Parámetro	Resultado
Valor mínimo	0.0	Varianza	1431936.2
1Q	235.2	Desviación estándar	1196.6
Mediana	697.5, IC 95 % (660.4 – 746.9)		
Media	1067.9, IC 95 % (938.8-1152)	Asimetría	1.68454
3Q	1272.5	Curtosis	2.16115
Valor máximo	5826	N	780
Prueba de normalidad de Anderson – Darling		A^2 58.36, p<0.005	

Fuente: elaboración propia, empleando Microsoft Excel.

La prueba de normalidad rechaza la hipótesis nula, permitiendo concluir que se trata de una distribución no normal. El histograma correspondiente lo ilustra claramente.

Figura 12. **Histograma de la distribución de frecuencias de la variable incidencia diaria de casos de COVID-19, Guatemala, marzo de 2020 a abril de 2022**



Fuente: elaboración propia, empleando Minitab.

Ya que la naturaleza de la variable respuesta es la de una de recuento de datos, inicialmente se consideró que podría ajustarse a una distribución de Poisson o a alguna de las que derivan de esta.

Los estadísticos media y varianza fueron útiles para este propósito, pues, en la distribución de Poisson, se considera que estos son estadísticamente iguales y, tal y como se aprecia en la tabla II, estos difieren de forma muy evidente. De acuerdo con esto, se puede descartar que se trate de una distribución de Poisson.

Para corroborar lo anterior, se realizó una prueba de bondad de ajuste, mediante una prueba de chi cuadrado. El valor p resultó menor de 0.05, lo que rechazó la hipótesis nula para una distribución de Poisson, confirmando lo visto anteriormente, lo que llevó a considerar la RBN como el modelo apropiado.

Para poder llevar a cabo la RBN, primero se debe establecer si los datos presentan sobredispersión. Para esto se estimó un modelo de regresión de Poisson y, con una prueba de hipótesis, donde Ho sostiene que no hay sobredispersión y Ha sostiene que no se puede rechazar la sobredispersión, se obtuvieron los siguientes resultados:

Tabla III. **Modelo de regresión de Poisson para la incidencia de casos nuevos de COVID-19 en función de las variables regresoras de interés**

Código en Rstudio	#estimar modelo de Poisson
	pois <- glm(formula = new_cas_lag ~ RRpct + GPpct + PRKSpct + TRSTpct + WRKpct + RESpct + Omicron + Delta + vac_pob, family = poisson, data = full_tasas)
	#valorar si hay sobredispersión
	dispersiontest(pois)

Desviación de Residuos:

Min	1Q	Mediana	3Q	Máx
-64.433	-23.332	-5.052	14.789	106.299

Coeficientes:

	Estimados	Error Std.	valor z	Pr(>|z|)	
Intercept	$5.64x10^{0}$	$5.87x10^{-3}$	960.05	$<2x10^{-16}$	***
RR	$3.21x10^{-2}$	$4.99x10^{-4}$	64.24	$<2x10^{-16}$	***
GP	$1.57x10^{-3}$	$2.59x10^{-4}$	6.04	$1.52x10^{-9}$	***
PRK	$-3.39x10^{-2}$	$3.20x10^{-4}$	-105.71	$<2x10^{-16}$	***
TRST	$5.39x10^{-2}$	$3.19x10^{-4}$	-169.02	$<2x10^{-16}$	***
WRK	$3.94x10^{-2}$	$1.67x10^{-4}$	235.84	$<2x10^{-16}$	***
RES	$-4.77x10^{-2}$	$4.70x10^{-4}$	-101.45	$<2x10^{-16}$	***
Omicron1	$1.44x10^{-2}$	$5.04x10^{-4}$	28.43	$<2x10^{-16}$	***
Delta1	1.59	$4.23x10^{-4}$	375.19	$<2x10^{-16}$	***
vac_pob	$1.57x10^{-5}$	$4.67x10-^{6}$	16.22	$<2x10^{-16}$	***

Códigos de significancia: 0*** 0.001** 0.01* 0.05

(Parámetro de dispersión para familia de Poisson considerado como 1)

Desviación nula: 903839 on 779 degrees of freedom
Desviación con residuos: 483690 on 770 degrees of freedom
AIC: 489672

Número de iteraciones de Fisher: 5

Prueba de sobredispersión:
data: pois
z= 12.134, pvalue < $2.2x10^{-16}$
Ha: dispersión real es mayor que 1
Estimación:
Dispersión:
626.2

Fuente: elaboración propia, empleando Microsoft Excel.

De acuerdo con lo anterior, con un valor $p < 2.2^{-16}$ se rechazó H_o y se estableció que no se puede rechazar la sobredispersión, lo que llevó a realizar un modelo de RBN (ver más adelante).

A continuación, se presenta la tabla que compara la incidencia diaria de casos dependiendo del estatus de las variantes Delta y Ómicron.

Tabla IV. **Resumen de la variable incidencia diaria de casos, de acuerdo con el estatus de detección de las variantes Delta y Ómicron, Guatemala, marzo 2020 a abril 2022**

Variable	Variante Ómicron	Delta	Media	Desviación estándar	Mediana
Incidencia diaria de casos	0	0	546.8	476	533
	0	1	1876	1514	1373
	1	1	1862	1463	1417

Fuente: elaboración propia, empleando Microsoft Excel.

Se aprecia que la media y mediana de casos diarios fue mucho menor antes de que se detectaran las variantes Delta y Ómicron. Se presentan la información más detallada:

Tabla V. **Resumen detallado de la variable incidencia diaria de casos, de acuerdo con el estatus de detección de las variantes Delta y Ómicron, Guatemala, marzo 2020 a abril 2022**

Incidencia diaria de casos	**Variante Delta**	**Mediana**	**IC 95 % de la mediana**	**Prueba de la mediana de Mood (valor p)**	**IC 95 % (mediana(0) – mediana(1))**
	0	533	387.71; 623.29	<0.001	-1276.7; -676.941
	1	1397	1162.29; 1727.32		
	Variante Ómicron				
	0	665	622.414; 699	<0.001	-1409; -330
	1	1417	1008.89; 2080.81		

Fuente: elaboración propia, empleando Microsoft Excel.

La tabla anterior muestra que hay diferencia estadísticamente significativa en la incidencia de casos antes y después de la detección de cada una de las variantes de interés estudiadas.

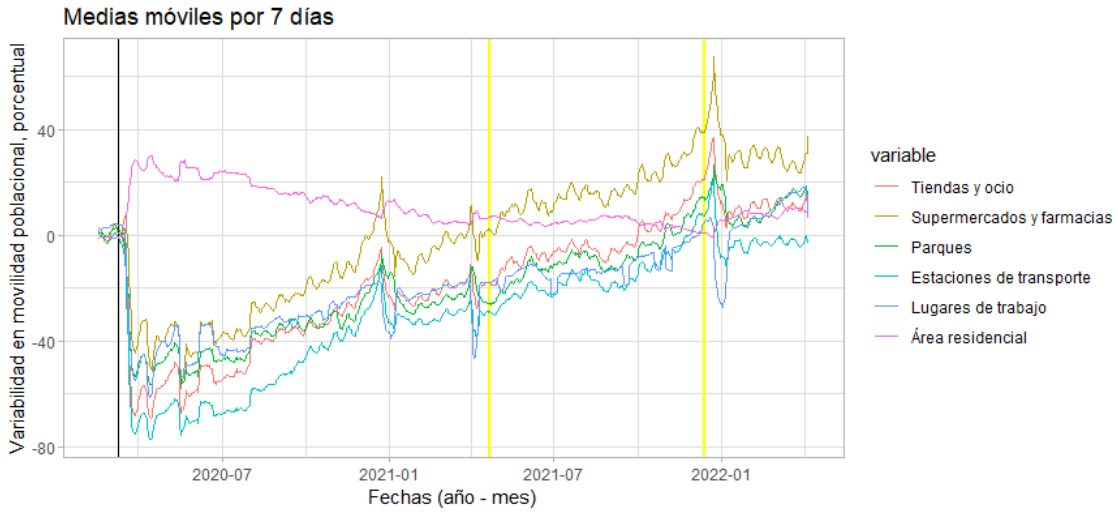
3.2. Objetivo 2. Caracterizar cuál ha sido el comportamiento de la movilidad comunitaria, durante la pandemia de COVID-19, en la población de Guatemala

La movilidad comunitaria es una variable difícil de medir. Para este trabajo, la información se obtuvo de los registros electrónicos hechos públicos por Google que, mediante el análisis de los datos de geolocalización de los teléfonos celulares de las personas que han dado su permiso expreso para el registro de estos datos, estimó la variabilidad porcentual de la visita y permanencia de los individuos en diferentes categorías de lugares.

Se trabajó con la información completa de Guatemala, iniciando 3 semanas antes de la detección del primer caso de COVID-19 en el país, hasta el 06 de abril de 2022. Estas 3 semanas previas al inicio del brote epidémico en Guatemala sirvieron como referencia ante la que se comparó todo el período epidémico referido.

A continuación, se presenta el gráfico que muestra la evolución temporal en la movilidad comunitaria:

Figura 13. **Comportamiento de la movilidad comunitaria de la población guatemalteca, febrero 2020 a abril 2022**



Fuente: elaboración propia, empleando RStudio.

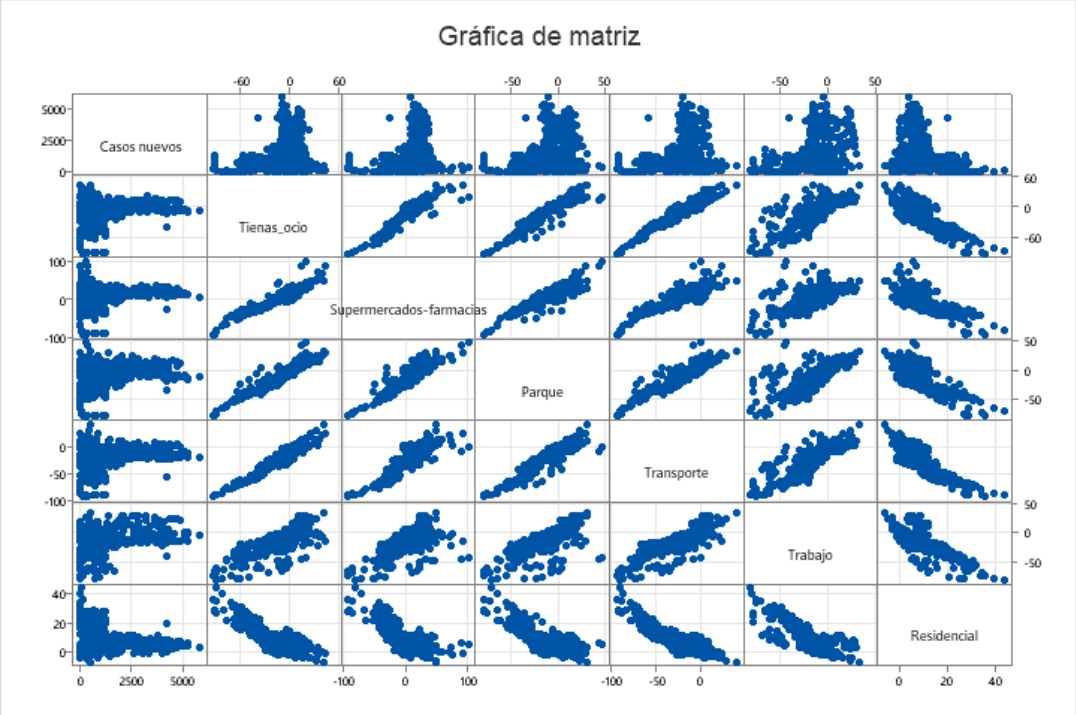
La figura 13 muestra claramente que, tras cierta estabilidad, las medidas de restricción en la movilidad comunitaria fueron aplicadas en Guatemala y fueron acatadas por la población. La línea negra vertical, en la parte izquierda del gráfico, marca el inicio del brote epidémico en el país, representando el día del primer caso (13 de marzo de 2020). Las líneas verticales de color amarillo muestran las primeras detecciones de las variantes Delta y Ómicron, en su respectivo orden cronológico.

En dicha figura se hace evidente que la movilidad comunitaria se vio incrementada a las áreas residenciales alrededor de 30 % de forma pronta tras detectarse el primer caso, mientras que todas las demás áreas vieron reducida la afluencia de personas entre 40 % a 75 %. Sin embargo, la tendencia a lo largo de todo el brote epidémico ha sido el retorno gradual a niveles similares a los prepandemia. Al final del período evaluado, supermercados y farmacias son la

categoría más visitada (casi 40 % por sobre el valor de referencia), los parques representan el área menos visitada (muy próximo al valor de referencia) y las demás áreas se encuentran en un perfil intermedio, entre 10 % - 20 % por sobre el valor de referencia.

Obviamente, la movilidad incrementada a algunas de las áreas implica movilidad reducida en las otras. Esto se puede apreciar en la figura siguiente:

Figura 14. **Gráfico de matriz para mostrar la relación entre movilidad a distintas áreas y, también, incidencia de casos**



Fuente: elaboración propia, empleando Minitab.

Se presentan los parámetros estadísticos de la movilidad comunitaria de forma global y luego se compararon diferentes momentos de estas, para valorar si hubo impacto de la detección de variantes.

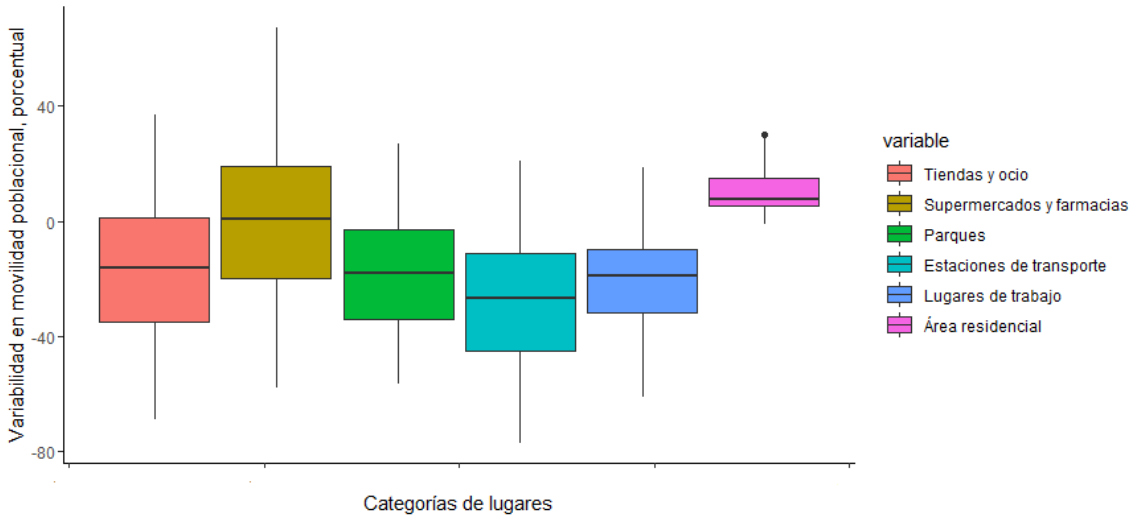
Tabla VI. **Resumen general de la movilidad comunitaria en Guatemala, durante la pandemia de COVID-19, de acuerdo con las categorías de lugares preespecificadas**

Variable	N	Media	Error estándar de la media	Desviación estándar	Mínimo	1Q	Mediana	3Q	Máximo
Tiendas y ocio	780	-17.33	0.89	24.89	-92	-35	-15	1	44
Supermercados y farmacias	780	0.18	0.98	27.34	-92	-21	2	21	101
Parques	780	-18.55	0.75	21.02	-81	-34	-18.5	-2	47
Estaciones de transporte	780	-29.34	0.87	24.34	-93	-48	-26	-10	40
Lugares de trabajo	780	-19.15	0.72	19.99	-81	-33	-19	-5	33
Áreas residenciales	780	10.45	0.29	8.14	-6	-5	8	15	44

Fuente: elaboración propia, empleando Microsoft Excel.

Los datos de la tabla anterior pueden diferir un poco de lo mostrado en la gráfica respectiva, pero debe de recordarse que las estimaciones hechas para la tabla se hicieron a partir de los datos brutos, mientras que la gráfica representa las medias móviles de 7 días. Sin embargo, la información es compatible. A continuación, se muestran los gráficos de cajas respectivos.

Figura 15. **Variabilidad de la movilidad comunitaria en Guatemala, durante la pandemia de COVID-19, acorde a categorías de lugares visitados**



Fuente: elaboración propia, empleando RStudio.

Es claro que sólo hubo incremento significativo en la mediana respectiva a Área Residencial, aunque, debido a la evolución y tendencia, se sabe que también hubo un incremento discreto en las visitas a supermercados y farmacias. Recordando la figura 13, se aprecia que esta última categoría ha incrementado sustancialmente la cantidad y tiempo de estancia de visitas.

Se presenta ahora la comparación entre los períodos pre y post pandemia para cada una de las categorías. Se inicia con la tabla respectiva. Se verificará mediante la prueba de la mediana de Mood, con Ho que implica que las medianas son iguales y Ha que implica que las medianas no son iguales. Se utilizó este estadístico por ser data no paramétrica:

Tabla VII. **Comparación en la movilidad comunitaria hacia las diferentes categorías de lugares, pre *versus* post COVID-19, Guatemala, febrero 2020 a abril 2022**

Variable	Estatus COVID-19	Media	Desviación estándar	Mediana	IC 95 % de la mediana	Prueba de la mediana de Mood (valor p)	IC 95 %
Tiendas y ocio	0	0.23	3.48	0	-1.35; 1	<0.001	14.31; 18
	1	-17.93	25.09	-16	-18; -13		
Supermercados y farmacias	0	0.77	4.49	-0.5	-2; 2.35	0.054	-5; 0.28
	1	0.16	27.80	2.5	0; 6		
Parques	0	1.12	2.41	1.5	0; 3	<0.001	20; 24
	1	-18.82	21.06	-20	-23; -17		
Estaciones de transporte	0	-1.54	3.11	-2	-3; -0.65	<0.001	24; 26.28
	1	-30.29	24.19	-27	-29; -25		
Lugares de trabajo	0	3.77	1.07	4	3; 4	<0.001	22; 24
	1	-19.94	19.86	-19.5	-21; -18		
Áreas residenciales	0	-0.42	0.76	0	-0.35; 0	<0.001	-8.28; -8
	1	10.82	8.02	8	8; 9		

Fuente: elaboración propia, empleando Microsoft Excel.

Considerando que la distribución de los datos de movilidad comunitaria es no paramétrica, se compararon las medianas antes y después del inicio del brote epidémico en Guatemala. Entre las diferentes categorías, es claro que todas las categorías, excepto la de Supermercados y Farmacias, se modificaron de forma estadísticamente significativa, así como que la categoría de Áreas Residenciales fue la única que mostró aumento en la mediana de visitas y permanencia a la misma.

Debe recordarse que, para el período que se abarcó en el estudio, la política para el control de aforos siempre estuvo vigente.

Ahora se presentan las comparaciones de la movilidad según el estatus de la variante Delta.

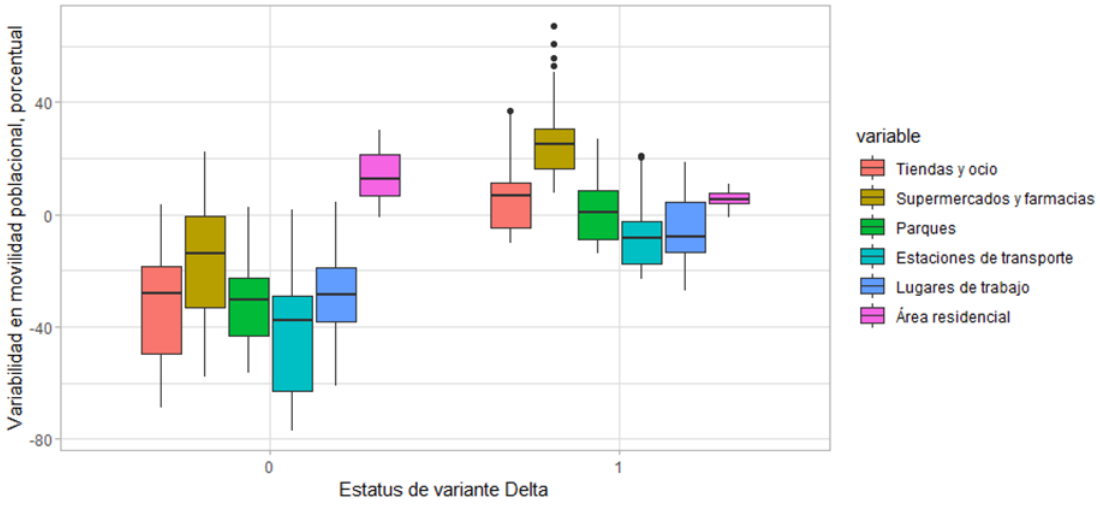
Tabla VIII. **Comparación en la movilidad comunitaria hacia las diferentes categorías de lugares, pre versus post detección de variante Delta en Guatemala, marzo 2020 a abril 2022**

Variable	Estatus Variante Delta	Media	Desviación estándar	Mediana	IC 95 % de la mediana	Prueba de la mediana de Mood (valor p)	IC 95 %
Tiendas y ocio	0	-31.81	19.99	-30	-33; -26	<0.001	-40; -30
	1	-4.99	11.64	6	2.85; 8		
Supermercados y farmacias	0	-15.87	21.23	-15	-18; -12	<0.001	-42; -35.64
	1	24.91	13.99	24	22; 26		
Parques	0	-30.69	15.47	-30	-32; -29	<0.001	-33.36; -27.60
	1	1.16	11.77	0	-3; 3		
Estaciones de transporte	0	-42.69	20.90	-39	-43; -36	<0.001	35; -27
	1	-8.76	11.58	-8	-10; -7		
Lugares de trabajo	0	-28.44	16.76	-27	-29; -25.7	<0.001	-24.36; -19
	1	-4.83	15.36	-5	-7.15; -4		
Áreas residenciales	0	13.58	8.75	12	11; 13.29	<0.001	5.64; 7
	1	5.63	3.52	6	5; 6		

Fuente: elaboración propia, empleando Microsoft Excel.

Hay diferencia significativa en el estatus pre y post Delta en todas las categorías. Importante remarcar que todas las áreas incrementaron las medias de visita o permanencia en el período post Delta, excepto el área residencial, que disminuyó. El gráfico siguiente permite visualizar estos datos:

Figura 16. **Comparación en la movilidad comunitaria hacia las diferentes categorías de lugares, pre *versus* post detección de variante Delta en Guatemala, marzo 2020 a abril 2022**



Fuente: elaboración propia, empleando RStudio.

A continuación, se presentan las comparaciones respectivas al estatus de la variante Ómicron.

Tabla IX. **Comparación en la movilidad comunitaria hacia las diferentes categorías de lugares, pre versus post detección de variante Ómicron en Guatemala, marzo 2020 a abril 2022**

Variable	Estatus Variante Ómicron	Media	Desviación estándar	Mediana	IC 95 % de la mediana	Prueba de la mediana de Mood (valor p)	IC 95 %
Tiendas y ocio	0	-22.41	22.93	-19	-21; -18	<0.001	-32; -29
	1	12.66	10.91	12	11; 12		
Supermercados y farmacias	0	-5.29	24.88	-3	-5; 0	<0.001	-36; -30
	1	32.5	16.88	30	28; 32		
Parques	0	-23.17	18.17	-24	-26; -22.7	<0.001	-38; -32
	1	11.42	8.78	11	8; 13		

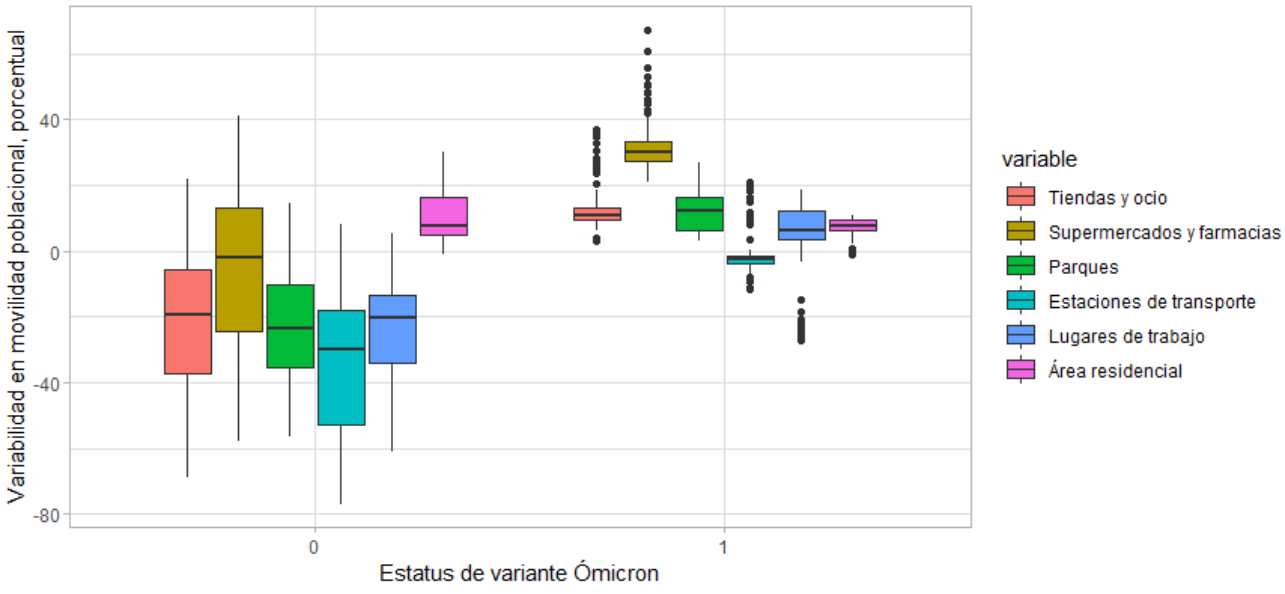
Continuación tabla IX.

Estaciones de transporte	**0**	-33.98	22.92	-30	-32; -28	<0.001	-28; -24
	1	-1.94	10.54	-4	-5; -2		
Lugares de trabajo	0	-23.20	17.43	-21	-23; -20	<0.001	-32; -23
	1	4.74	17.26	5	2; 10		
Áreas residenciales	0	10.97	8.58	8	8; 9	0.38	-1; 1
	1	7.37	3.53	8	7; 9		

Fuente: elaboración propia, empleando Microsoft Excel.

Es claro que, tras la detección de la variante Ómicron, la tendencia que inició previamente se mantuvo, aumentando la movilidad hacia todas las áreas públicas y disminuyendo hacia las áreas residenciales. A continuación, se presenta el gráfico respectivo:

Figura 17. **Comparación en la movilidad comunitaria hacia las diferentes categorías de lugares, pre versus post detección de variante Ómicron en Guatemala, marzo 2020 a abril 2022**



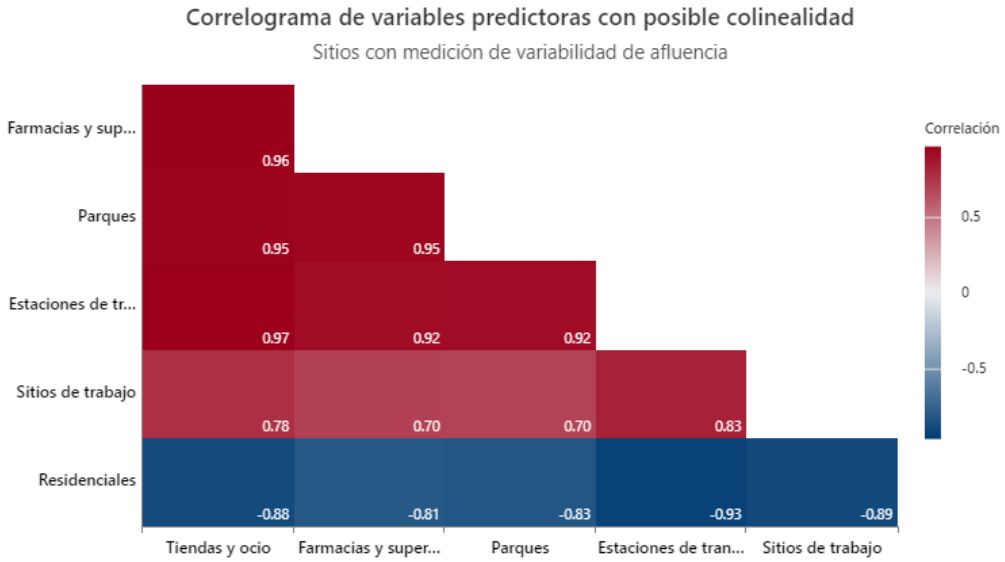
Fuente: elaboración propia, empleando RStudio.

3.2.1. Análisis multivariado de la movilidad comunitaria

Se decidió complementar el análisis de la movilidad comunitaria con un análisis de componentes principales, con el propósito de comprender mejor la naturaleza de la movilidad poblacional, de cara a los análisis que se hicieron posteriormente. Se incluyó para este análisis únicamente la movilidad de la población en el período antes de la detección de las variantes de interés ya que, durante las olas Delta y Ómicron, el crecimiento explosivo de los contagios hacen que el análisis sea difícil de realizar en este período.

Se inicia por evaluar la relación entre variables predictoras, tanto gráfica como numéricamente. A continuación, se muestra el correlograma con los índices respectivos:

Figura 18. **Correlograma de variables predictoras que representan la movilidad poblacional**



Fuente: elaboración propia, empleando Minitab.

Esto se complementó con un análisis de componentes principales, según las tablas siguientes (estimado en Minitab):

Tabla X. **Vectores propios, variables y componentes principales, que describen la movilidad poblacional durante la pandemia (período pre-Delta)**

Vectores propios						
Variable	PC1	PC2	PC3	PC4	PC5	PC6
Tiendas y ocio	0.48	0.13	-0.19	-0.03	0.84	-0.08
Farmacias y supermercados	0.53	0.44	0.59	0.33	-0.24	-0.11
Parques	0.34	0.24	-0.09	-0.86	-0.27	0.11
Estaciones de transporte	0.48	-0.14	-0.59	0.37	-0.33	0.39
Sitios de trabajo	0.33	-0.81	0.44	-0.14	0.04	0.15
Residenciales	-0.20	0.24	0.26	0.01	0.21	0.89

Fuente: elaboración propia, empleando Microsoft Excel.

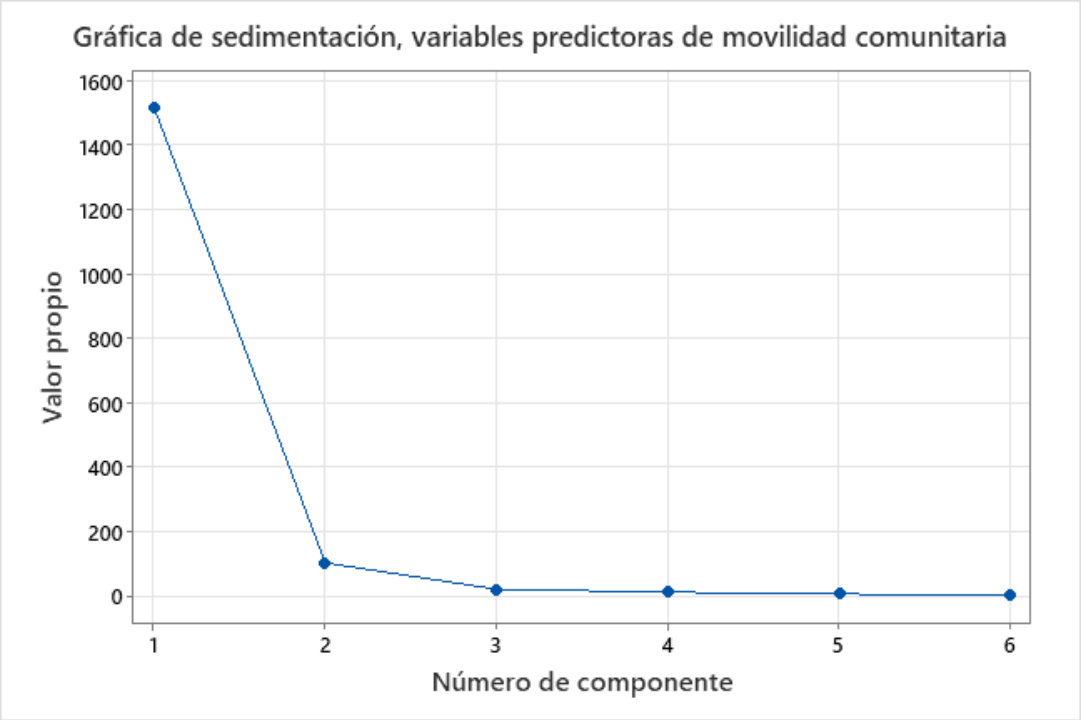
Tabla XI. **Valores propios, relevancia proporcional y acumulada de la matriz de covarianza**

Análisis de los valores y vectores propios de la matriz de covarianza						
Valor propio	1515.5	104.1	22.4	13.2	7.5	4.8
Proporción	0.91	0.06	0.01	0.01	0.05	0.05
Acumulada	0.91	0.97	0.98	0.99	0.99	1.000

Fuente: elaboración propia, empleando Microsoft Excel.

Que genera la siguiente gráfica de sedimentación:

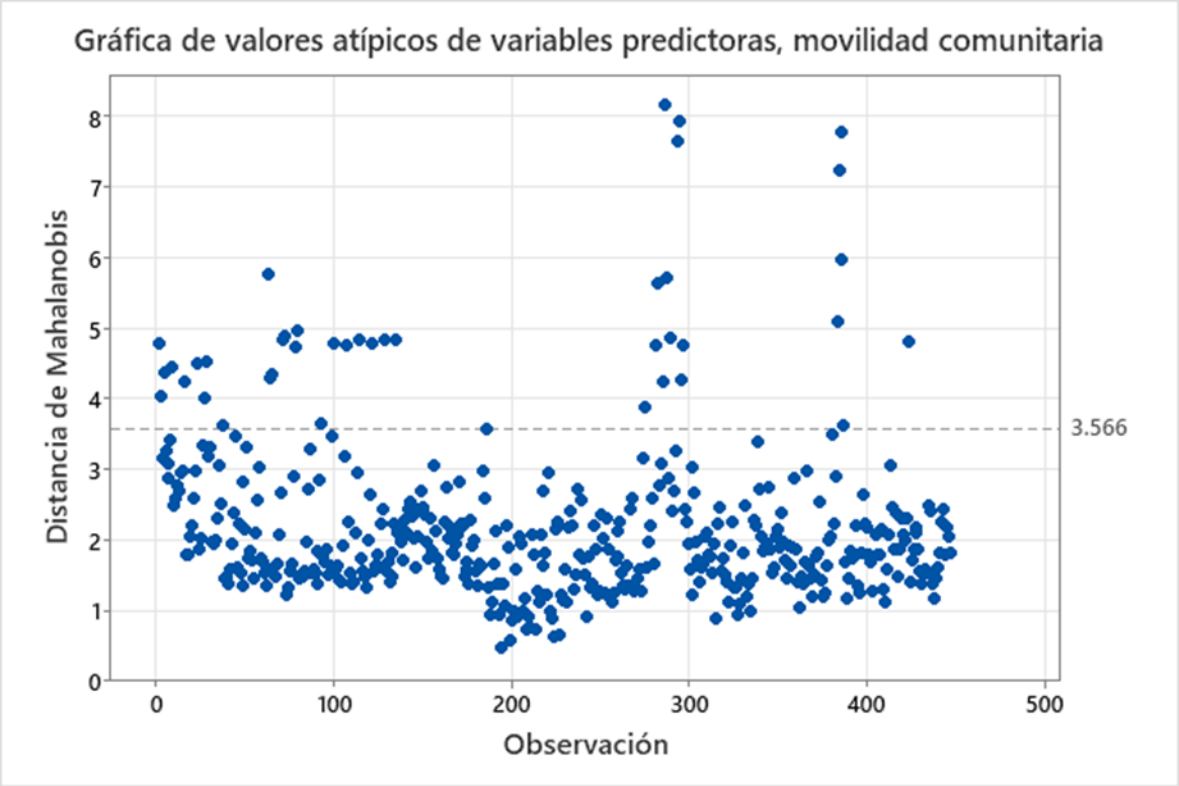
Figura 19. **Gráfico de sedimentación, análisis de componentes principales**



Fuente: elaboración propia, empleando en Minitab.

De acuerdo con las tablas y gráfico, los componentes principales 1 y 2 explican el 97 % de la variabilidad de la movilidad poblacional, con los coeficientes de las variables originales descritos en las columnas respectivas en la tabla, bajo la información de los vectores propios (valores remarcados en negrita), que permite encontrar abundantes valores atípicos, de acuerdo con la distancia de cada observación versus el centroide, según la distancia de *Mahalanobis*, como se aprecia en la gráfica siguiente:

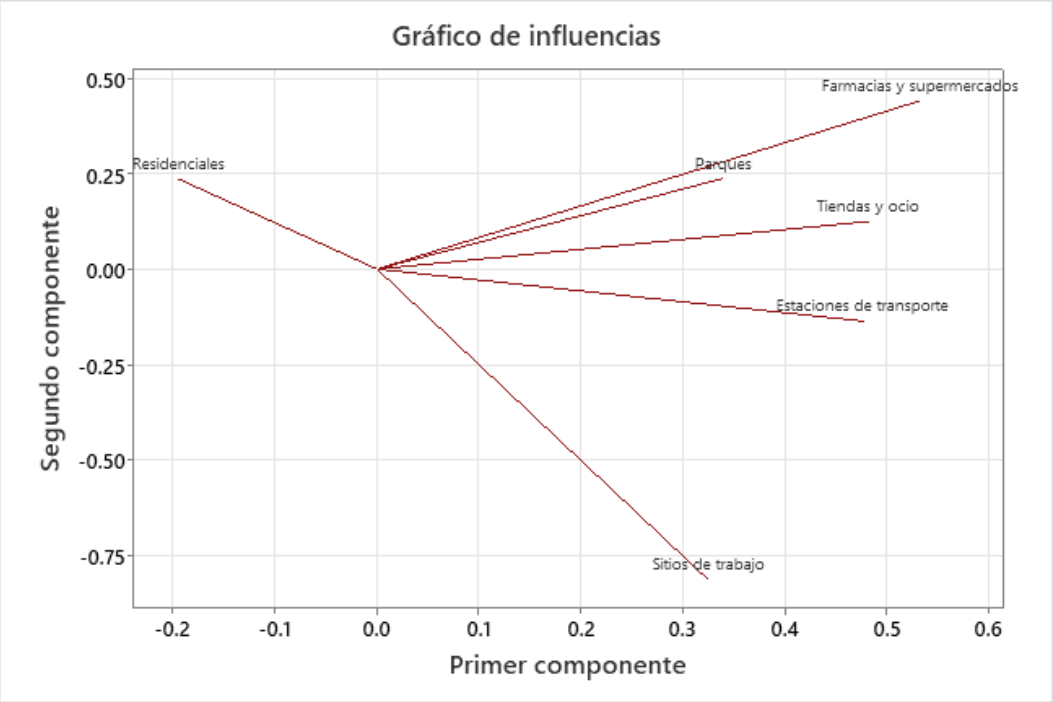
Figura 20. **Valores atípicos, distancia de *Mahalanobis***



Fuente: elaboración propia, empleando Minitab.

Finalmente, se incluye la gráfica de influencias para los componentes principales 1 y 2:

Figura 21. **Gráfico de influencias para componentes principales 1 y 2**



Fuente: elaboración propia, empleando Minitab.

3.3. Objetivo 3. Determinar cuál es el mejor modelo de regresión a elaborar de acuerdo con la evaluación de los criterios de información apropiados para la naturaleza de dicho modelo, para explicar la incidencia de COVID-19 en función de los cambios de movilidad poblacional en Guatemala

Para estimar un modelo de regresión que intente explicar la incidencia de casos de COVID-19 en Guatemala, en función de la variabilidad en la movilidad poblacional a las diferentes categorías de sitios a visitados, se modificaron los datos para establecer un período de incubación de 10 días, así como se omitieron los datos que, para describir la movilidad poblacional, se usaron como referencia. Por lo anterior, la cantidad de días analizados se redujo a 754.

Esto para evitar elevar artificialmente los días con 0 eventos registrados (que generaría un modelo inflado por ceros debido al mal manejo de los datos), a la vez que, sin el retraso establecido como período de incubación, se perdería la correlación cronológica biológica propia del proceso salud-enfermedad.

Lo anterior conlleva desventajas obvias que son inevitables, sobre todo, que se toma el período de incubación como fijo, lo que, desde la perspectiva biológica, es imposible. Sin embargo, la naturaleza de los datos hace que un análisis más específico sea imposible.

Ya que la variable respuesta es de naturaleza de conteo, se abordó el problema desde la perspectiva de las distribuciones discretas. Se estimaron diferentes modelos de regresión, para poder compararlas entre sí y elegir la que se considerase mejor entre estas.

Estos se estimaron en el *software* RStudio. Las variables predictoras incluidas fueron:

Movilidad poblacional hacia área residencial (R), en términos de variabilidad porcentual, comparados contra las tres semanas previas a la fecha en que fue confirmado el primer caso de COVID-19 en Guatemala.

Estatus de las variantes Delta, como variable categórica dicotómica, en la que 1 representa la presencia confirmada de la variante en el país y 0 representa su ausencia.

La variable respuesta fue el conteo de casos reportados diarios, con un retraso de 10 días. Este retraso representa el período de incubación estimado, de acuerdo con el CDC.

Para cada variable, se contó con 754 observaciones válidas, de acuerdo con la frecuencia en que se presentó la variable respuesta.

3.3.1. Análisis por regresión binomial negativa:

El modelo estimado brindó los siguientes resultados.

Tabla XII. **Resumen del modelo de RBN para explicar la incidencia de COVID-19 con todas las variables regresoras**

Desviación de los residuos	**Min: -4.03**	**1Q: -0.89**	**2Q: -0.20**	**3Q: 0.35**	**Máx: 3.53**
Coeficientes:					
	Estimados	**EE**	**Valor z**	**P**	
Interceptos	6.03	1.77×10^{-1}	34.12	$<2 \times 10^{-16}$	***
Tiendas y ocio	2.55×10^{-2}	1.18×10^{-2}	2.17	0.03	*
Farmacias y supermercados	-6.86×10^{-3}	6.25×10^{-3}	-1.10	0.27	
Parques	-8.47×10^{-3}	8.10×10^{-3}	-1.05	0.30	
Estaciones de transporte	-5.88×10^{-2}	1.01×10^{-2}	-5.82	6.07×10^{-9}	***
Sitios de trabajo	6.67×10^{-3}	4.65×10^{-3}	1.44	0.15	
Áreas residenciales	-1.16×10^{-1}	1.36×10^{-2}	-8.58	2×10^{-16}	***
Ómicron_1	9.1×10^{-1}	1.75×10^{-1}	5.20	2×10^{-7}	***
Delta_1	1.27	1.36×10^{-1}	9.38	2×10^{-16}	***
Vacunas	1.76×10^{-6}	1.29×10^{-6}	1.37	0.17	
AIC: 11768					
P = 0.00000					

Fuente: elaboración propia, empleando Microsoft Excel.

Se estimó pseudoR^2, $pR^2 = 0.27$.

Tras esto, se buscó simplificar el modelo. Para ello, en el *software* estadístico RStudio, a través del paquete MASS, se aplicó la función *step* que

permite seleccionar el mejor modelo de acuerdo con el criterio de información AIC. Esto lo hace añadiendo o quitando variables de forma secuencial, partiendo, primero, de un modelo que no incluye ninguna variable explicativa hacia un modelo que las incluye todas, y luego hace el recorrido inverso, seleccionando el modelo más simple que no pierda valor predictivo, seleccionado mediante el referido criterio de información.

De acuerdo con lo observado en las figuras 14 y 18, las variables predictoras presentan colinealidad. El mejor modelo estimado, omitiendo las variables con factor de inflación de la varianza mayor a 10 (sugiere colinealidad), se resume de la siguiente forma:

Tabla XIII. **Modelo depurado, de acuerdo con AIC**

Desviación de los residuos	**Min: -3.78**	**1Q: -0.93**	**2Q: -0.11**	**3Q: 0.42**	**Máx: 3.36**
Coeficientes:					
	Estimados	**EE**	**Valor z**	**P**	
Interceptos	6.68	0.09	73.72	$<2x10^{-16}$	***
Áreas residenciales	-0.02	0.01	-4.18	$2.97x10^{-5}$	***
Delta_1	1.00	0.09	11.11	$2x10^{-16}$	***
AIC: 11815 **BIC: 11833** **R^2 (Nagelkerke): 0.36** **P = 0.00000**					

Fuente: elaboración propia, empleando Microsoft Excel.

Tras esto, se estimaron los IC 95 %:

Tabla XIV. **Intervalos de confianza para los coeficientes del modelo depurado**

	Estimados	**P**	**2.5 %**	**97.5 %**
Interceptos	6.68	$2x10^{-16}$	6.50	6.86
Delta_1	1.00	$2x10^{-16}$	0.83	1.18
Área residencial	-0.02	$2x10^{-16}$	-0.03	-0.01

Fuente: elaboración propia, empleando Microsoft Excel.

Finalmente, ya que la regresión binomial negativa, al igual que otros modelos de regresión de variables de conteo, utiliza la transformación logarítmica para ajustar a la distribución de referencia, se exponencian los coeficientes para obtener la razón de tasa de incidencia (IRR):

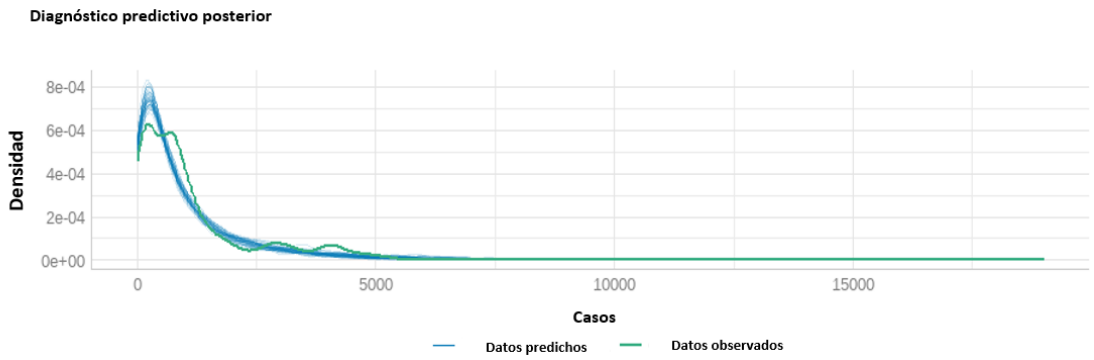
Tabla XV. **Estimación de IRR para el modelo depurado**

	Estimados	**P**	**2.5 %**	**97.5 %**
Interceptos	796.60	$2x10^{-16}$	666.84	957.38
Delta_1	2.73	$2x10^{-16}$	2.29	3.25
Área residencial	0.98	$2x10^{-16}$	0.97	0.99

Fuente: elaboración propia, empleando Microsoft Excel.

A continuación, se muestra el análisis gráfico referente al diagnóstico del modelo estimado. Esto se realizó con el comando *check.model()* del paquete *performance*, del *software* R. Inicialmente, se presenta la comparación entre la función de densidad del modelo y de los datos observados.

Figura 22. **Gráfico de densidad, datos predichos versus observados**



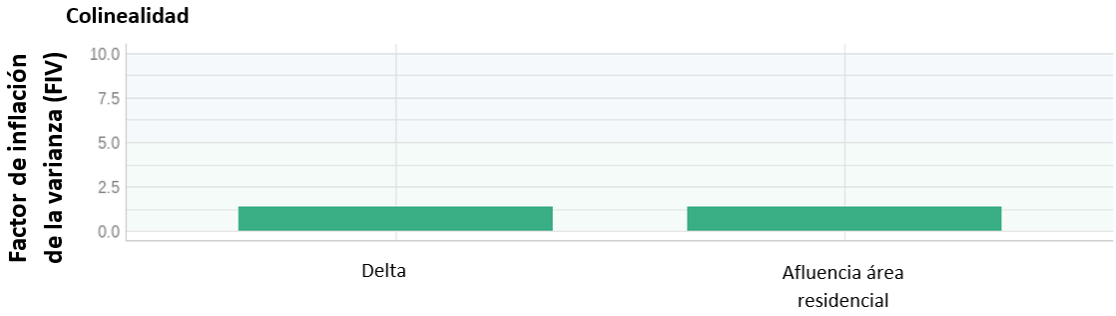
Fuente: elaboración propia, empleando RStudio.

Con respecto a la homogeneidad de la varianza y colinealidad, las figuras 23 y 24 ilustran los resultados.

Figura 23. **Homogeneidad de la varianza del modelo estimado**

Fuente: elaboración propia, empleando RStudio.

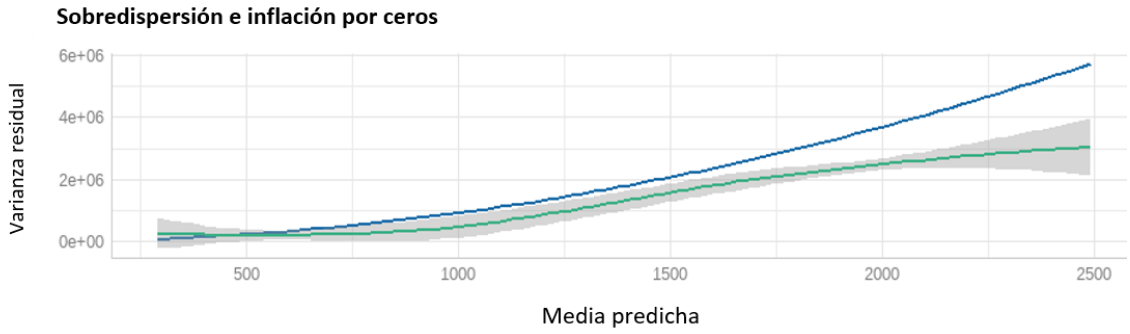
Figura 24. **Análisis de colinealidad, en función de VIF**



Fuente: elaboración propia, empleando RStudio.

Adicionalmente, y de acuerdo con lo expuesto en secciones previas, el gráfico siguiente muestra la sobredispersión del modelo, y muestra que no hay inflación por ceros, lo que fue evidente desde el análisis de los datos (ver sección 3.1 y tabla III).

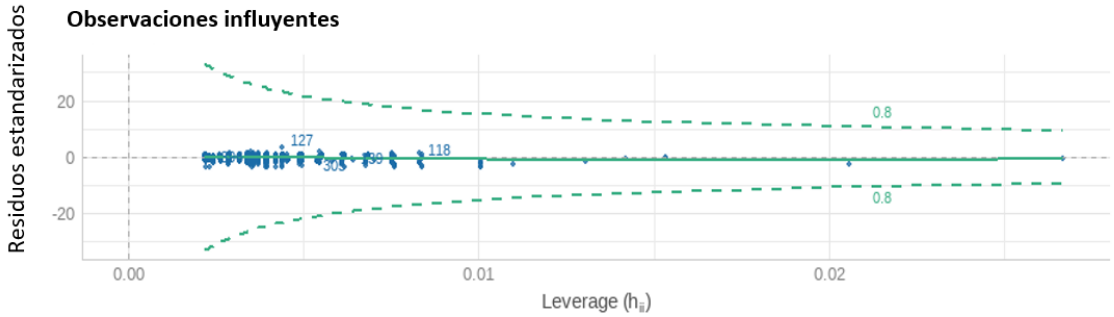
Figura 25. **Análisis gráfico de la sobredispersión**



Fuente: elaboración propia, empleando RStudio.

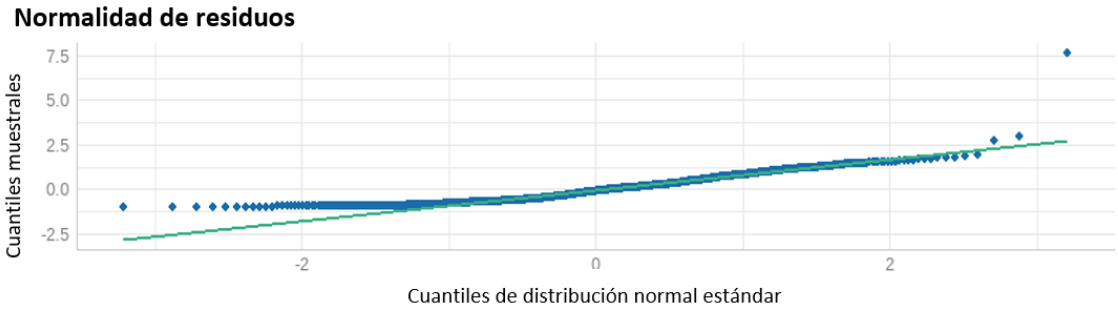
A continuación, se muestra la figura que presenta si hay valores influyentes y, luego, el qqplot de residuos:

Figura 26. **Valores influyentes**



Fuente: elaboración propia, empleando RStudio.

Figura 27. **qqplot de distribución de residuos**



Fuente: elaboración propia, empleando RStudio.

3.4. Objetivo general. Elaborar un modelo de regresión para explicar la incidencia de COVID-19 en función de los cambios en la movilidad poblacional en Guatemala

Se trabajó con los datos de 754 días consecutivos, sin datos faltantes, que incluyó la información referente a movilidad comunitaria a: áreas residenciales, lugares de trabajo, estaciones de transporte, parques, tiendas y ocios y supermercados y farmacias, en términos de variabilidad porcentual, comparado

con la media geométrica de afluencia poblacional de 5 semanas consecutivas en el período de 3 de enero a 6 de febrero de 2020, comparando los días específicos de la semana. Así mismo, se trabajó también la información referente a la presencia de las variantes Delta y Ómicron en el país y la tasa de vacunación, en número de dosis aplicadas *per cápita*, según su evolución en el tiempo. Todo lo anterior como variables regresoras, para generar el modelo de RBN. El número de casos nuevos detectados es la variable respuesta. Se estableció un período de incubación de 10 días.

El modelo fue trabajado en el *software* RStudio, el paquete utilizado para este propósito fue MASS. Se depuró mediante AIC y R^2.

El modelo es el siguiente:

$$\ln(incidencia\_covid19) = \beta_0 + \beta_1 * Delta_1 + \beta_2 * (var_{Residenciales}) \quad \text{(Ec. 24)}$$

Donde β_i corresponde a los coeficientes que multiplican las diferentes variables regresoras. La razón de tasas de incidencia (IRR) se obtienen al exponenciar los coeficientes, tal y como se aprecia en la tabla XV.

3.4.1. Construcción de un modelo alternativo que -también- resuelve el problema de multicolinealidad

La alta correlación entre las variables predictoras es evidente, tanto gráfica como numéricamente. Esto implica que la misma información, en general, se obtiene al analizar cada una de estas.

Se infiere que puede construirse un modelo alternativo que incluya únicamente a la variable afluencia a sitios residenciales, ya que ella correlaciona

de forma negativa y directamente proporcional con las demás variables predictoras. Esto puede analizarse en conjunto con la figura 14.

Al completar el análisis de esta forma, se obtiene que la única variable predictora que mantiene significancia estadística es la presencia / ausencia de la variante Delta. De esto, se puede interpretar que, debido a lo explosivo de las olas de contagio asociadas a las variantes, así como a que delta presentó más casos detectados (no puede descartarse que la población buscó el diagnóstico formal con menos fuerza durante la ola Ómicron), así como al traslape temporal entre las cepas mencionadas, las variantes representan la principal causa de variabilidad en la incidencia de COVID-19.

Como método alternativo, se estimó un modelo adicional mediante un análisis de regresión de mínimos cuadrados parciales. Este método, como se explicó previamente, reduce los predictores a un grupo más pequeño de predictores no correlacionados, a la vez que no presupone precisión en la medición original, lo que le permite mayor robustez ante el error. Esta vez, se incluyó para análisis únicamente el período entre la detección del primer caso (13 de marzo de 2020) y el inicio de la ola de la variante Delta, a modo de validar, en este contexto de COVID-19 antes de las variantes de interés, si la movilidad comunitaria era un predictor más fuerte. Se obtuvieron los resultados siguientes:

Tabla XVI. **ANOVA regresión de mínimos cuadrados parciales**

Fuente	**GL**	**SC**	**MC**	**F**	**P**
Regresión	7	25594313	3656330	20.65	0.000
Error residual	439	77717617	177033		
Total	449	103311930			

Fuente: elaboración propia, empleando Microsoft Excel.

Ya que el análisis de mínimos cuadrados parciales parte de un análisis de componentes principales, la validación pertinente es la siguiente.

Tabla XVII. **Selección y validación del modelo de regresión de cuadrados mínimos parciales**

Componentes	Varianza de X	Error	R^2	R^2 (predicho)
1	0.78	93750894	0.09	0.08
2	0.88	88793132	0.14	0.12
3	0.95	85438896	0.17	0.14
4	0.98	80143443	0.22	0.20
5	0.99	78409577	0.24	0.21
6	0.99	77717617	0.25	0.22
7	1	77717617	0.25	0.22

Fuente: elaboración propia, empleando Microsoft Excel.

Los coeficientes regresores son:

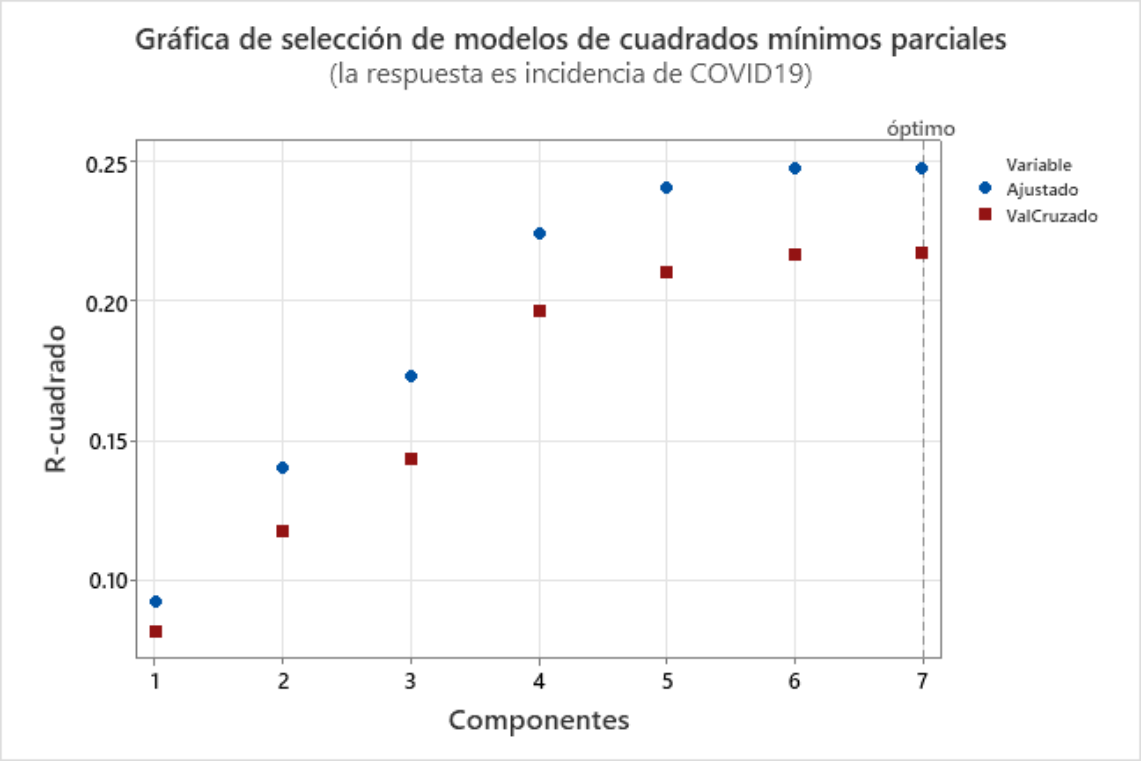
Tabla XVIII. **Coeficientes del modelo de regresión de mínimos cuadrados parciales**

Coeficientes	COVID-19	COVID-19 estandarizados
Constante	1083.95	0.00
Tiendas y ocio	27.65	1.09
Farmacias y supermercados	-21.91	-0.98
Parques	17.30	0.50
Estaciones de transporte	-26.49	-1.04
Sitios de trabajo	1.27	0.04
Residenciales	-36.82	-0.63
Vacunación (incidencia)	3.00	0.16

Fuente: elaboración propia, empleando Microsoft Excel.

Gráficamente, la selección del modelo se aprecia de la siguiente forma:

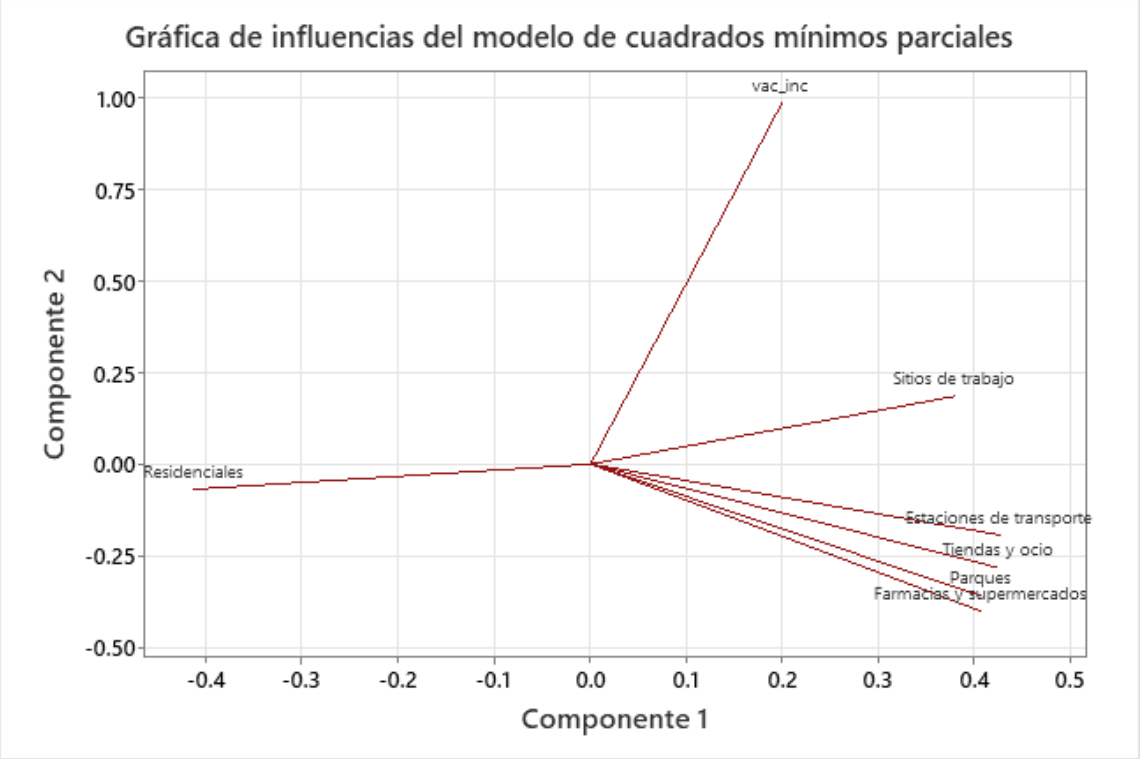
Figura 28. **Selección del modelo**



Fuente: elaboración propia, empleando Minitab.

Y el gráfico de influencias es el siguiente:

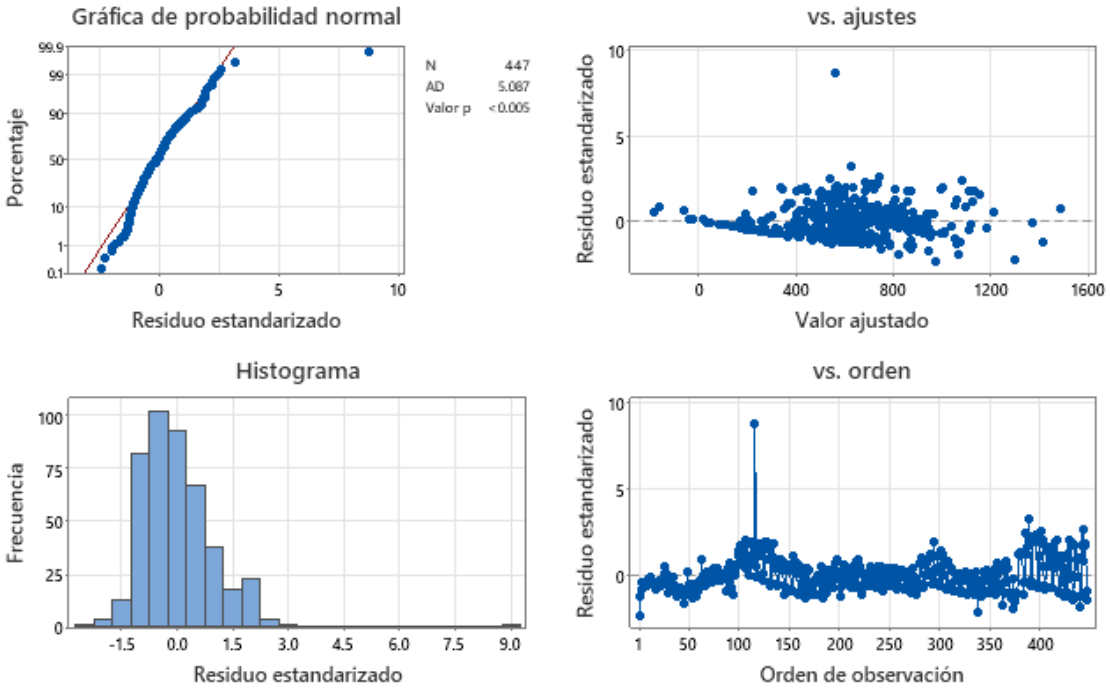
Figura 29. **Gráfico de influencias para el modelo de regresión de cuadrados mínimos parciales**



Fuente: elaboración propia, empleando Minitab.

Y el análisis de residuos muestra los siguientes gráficos:

Figura 30. **Análisis de residuos del modelo de regresión de cuadrados mínimos parciales**



Fuente: elaboración propia, empleando Minitab.

4. DISCUSIÓN DE RESULTADOS

4.1. Análisis interno

A continuación, se presenta la discusión de resultados del análisis interno.

4.1.1. Comportamiento de la incidencia bruta de casos de COVID-19 en Guatemala

La figura 9 es clara al mostrar que, en el período incluido en el estudio, se presentaron 5 olas de contagio. Las primeras tres representan eventos puntuales:

La primera ola, alrededor de julio del 2020, representa la primera ola, producto del ingreso del SARS-CoV2 a Guatemala, y muestra un pico importante que sugiere ser el ingreso acumulado de casos registrados tardíamente al sistema de registro nacional.

La segunda ola, alrededor de enero de 2021, se corresponde con el período después de las fiestas de final de año. De las cinco olas es, también, la de menor relevancia.

La tercera ola se presentó entre marzo y abril de 2021. Coincidió con el período posterior a las fechas de Semana Santa de dicho año.

La cuarta ola, la de mayor relevancia, inició aún antes de que la tercera ola alcanzara el nivel de referencia que correspondería con el final de la esta, se

prolongó desde inicios de mayo a noviembre de 2021, y corresponde al ingreso de la variante Delta.

La quinta ola inició en enero de 2022 y se prolongó hasta, al menos, inicios de abril del mismo año. Se corresponde con el ingreso de la variante Ómicron a Guatemala.

La figura 10 permite ver las acodaduras respectivas en la representación de la incidencia acumulada a lo largo de los 780 días registrados. La figura 11 incluye la línea de casos acumulados y sugiere una tendencia lineal, con un R^2 cercano a 1. Esto difiere de lo visto al inicio de la pandemia, en otros países, cuando todavía no se comprendía de forma apropiada la naturaleza de la enfermedad y se evidenció que el virus podía tener una transmisibilidad tal que la curva de crecimiento siguiera una curva exponencial.

Medidas que restringían la proximidad entre personas, como la restricción en la movilidad, el límite de aforos y el distanciamiento social, así como el uso de barreras respiratorias, como las mascarillas, son las causas de que, con mayor probabilidad, la tendencia fuese lineal y no exponencial en nuestro país.

Al estudiar la naturaleza de la variable respuesta, se encontró que presentó una distribución no normal. Al ser una variable de conteo, discreta, y al analizar el histograma correspondiente (figura 12), lo anterior se hace obvio.

Se analizó como una distribución de Poisson, pero presentó sobredispersión, como mostrado por el estadístico respectivo, con dispersión mayor a 1, $p < 2^{-16}$, (ver figura 20), así como al realizar la prueba de bondad de ajuste respectiva ($x^2 < 0.01$), para DP.

Se analizó también si la incidencia de casos aumentó después del aparecimiento de las variantes Delta y Ómicron. Confirmando lo mostrado en la figura 9, las incidencias post-Delta y post-Ómicron son significativamente mayores que las de los períodos previos. Esto se analizó mediante la prueba de la mediana de Mood, en ambos casos se estimó una $p < 0.001$, con intervalos de confianza significativos, aún si amplios.

4.1.2. Comportamiento de la movilidad poblacional durante la pandemia

La figura 13 es explicativa. La primera porción representa el período antes de la detección del primer caso en el país (13 de marzo de 2020). Se observa claramente que existe poca variabilidad, si se le compara con el período posterior.

La figura 14, por otro lado, muestra que la afluencia a sitios residenciales se correlaciona de forma negativa con la afluencia a todas las demás categorías de lugares.

Se observa un incremento inicial, y bastante relevante, en la visita y permanencia a lugares residenciales. Este alcanzó su punto máximo poco después del inicio del brote epidémico Guatemala, y, a partir de ese momento, ha presentado una disminución lenta, pero constante, hasta alcanzar, a inicios de abril de 2022, valores muy cercanos a los prepandemia.

Al contrario que lo anterior, las visitas y permanencia a las demás categorías de lugares, cayó sustancialmente, en el mismo momento. La caída más significativa la presentaron los lugares correspondientes a estaciones de transporte, que cayó cerca de un 80 % en su promedio semanal (a pesar de esto,

alcanzó días de caídas de más del 90 %). Las caídas fueron menos intensas para tiendas y ocio, supermercados y farmacias, parques y lugares de trabajo.

A pesar de que la disminución en visitas y tiempo de permanencia en estos lugares fue mucho más marcada que el aumento en visitas a residenciales, la recuperación a la línea basal también lo fue. Al final del tiempo registrado, sólo las estaciones de transporte registraban menos visitas que los lugares residenciales. Adicionalmente, tal y como se muestra en dicho gráfico, las visitas a supermercados y farmacias incrementaron también de forma muy relevante.

Muy relevante también notar que la detección de las variantes no fue una razón para revertir las tendencias. Al analizar las medianas respectivas, áreas residenciales muestran incremento en los datos referentes a visita ($Mo = 8$, $IC\ 95\ \%\ 15 - 44$), pero con tendencia a disminuir desde el valor inmediatamente posterior al inicio de la pandemia ($máx\ = 44$), lo que evidencia la tendencia, contrario a todos los demás sitios (ver tabla V y figura 14).

Al hacer la comparación estadística formal (tabla VI), esto fue evidente, al punto en que la variabilidad entre los períodos pre y post inicio de pandemia para, por ejemplo, las visitas a supermercados y farmacias no son estadísticamente diferentes, mientras que tiendas y ocio sí alcanza diferencia, pero por el marcado descenso inicial pues, en la actualidad, su recuperación y aceleración ayudan a evidenciar una recuperación completa de los niveles de actividad pre-pandemia o superiores.

Cuando se hizo el análisis para valorar si la detección de las variantes Delta y Ómicron generaron cambios significativos, los resultados fueron bastante claros: sí hubo diferencia estadísticamente significativa entre las medias de las visitas a todos los lugares entre los períodos pre y post detección de Delta, con

incremento en todas las categorías excepto en áreas residenciales, que disminuyó, y, en el caso de Ómicron, las áreas residenciales no se modificaron de forma estadísticamente significativa, al contrario que el resto de categorías, que mostraron incremento en la etapa posterior a la detección de esta variante (ver tablas VII y VIII y figuras 15 y 16). Así también, al ver los gráficos de cajas que resumen la movilidad pre y post Delta, y pre y post Ómicron, lo estrecho de las cajas después de la detección de las variantes muestra que se ha estabilizado considerablemente la movilidad comunitaria.

Por otro lado, el análisis multivariado de la movilidad comunitaria durante la pandemia mostró resultados interesantes. Como se explicó en la sección correspondiente, únicamente se analizó bajo esta metodología el período previo a la detección de las variantes Delta y Ómicron. Se determinó que un componente principal explica el 91 % de la covarianza y 2 componentes principales el 97.1 %, muy superior al 80 % que suele considerarse suficiente al realizar estudios observacionales.

El primer componente principal presentó asociaciones positivas grandes para farmacias y supermercados, tiendas y ocio y estaciones de transporte (moderadas para parques y sitios laborales), y negativa pequeña hacia área residencial, lo que expresa movilidad hacia afuera del área residencial; el segundo componente principal es un tanto más equilibrado y presentó asociación positiva fuerte con farmacias y supermercados, pero asociación negativa muy fuerte con sitios de trabajo y positiva moderada con residenciales, por lo que expresa movilidad hacia el área residencial, en detrimento de la movilidad a sitios laborales.

Así, se puede entender que la movilidad se dio a tres categorías de lugares: 1° áreas residenciales; 2° lugares de trabajo; 3° todas las demás. Esta tercera

categoría, todas las demás, parece ser la fuente principal de variabilidad en la movilidad de la población (ver figuras 21 y 29), y, por ende, la información pertinente se conservó para la sección subsiguiente del análisis.

4.1.3. Estimación de modelo de RBN

Se estimó un modelo de RBN, incluyendo las variables predictoras relevantes. El modelo fue el siguiente:

$$Casos\_nuevos \sim RR \ + \ Delta$$

Se estimó un modelo con un $R^2aj = 0.27, AIC = 11{,}768$. En este, las variables vacunas (tasa diaria de vacunación), visita a sitios de trabajo, visita a parques y visita a supermercados y farmacias no alcanzaron significancia estadística. Se depuró el modelo, utilizando el criterio de información de Akaike como referencia:

$$Casos\ nuevos \sim Delta \ + \ Residenciales$$

Este modelo depurado estima un $R^2aj = 0.357,\ AIC = 11{,}815,\ BIC = 11{,}833,\ p < 0.001$.

En el diagnóstico del modelo, encontramos que los supuestos tradicionalmente analizados no se cumplen: la varianza incrementa conforme los valores de y se hacen mayores, la distribución de residuos no se ajusta a la normal (se ajusta en un 69 % a la normal, un 19 % a la distribución de Tweedie y un 9 % a la de Weibull) y la comparación de los gráficos de densidad de los predichos versus los observados muestra que el modelo infla los ceros.

Como puntos fuertes, no se detectaron valores influyentes, esto es, observaciones cuya deleción cambiaría notablemente el modelo, y, las principales desviaciones ocurren en valores extremos, lo que justifica que la presencia de variantes de interés sea la principal fuente de variabilidad de y, no hay multicolinealidad (FIV de 1.35 para ambas variables) y la variable dependiente cumple los criterios para un modelo de RBN.

Al momento de analizar el problema como una regresión de cuadrados mínimos parciales, obtenemos que, en comparación contra el modelo de RBN: 1° no se logra mantener ni optimizar el principio de parsimonia (se trabajan con los 7 componentes principales estimados); 2° el R^2 ajustado no se optimiza (22 %); 3° la distribución de residuos no es óptima y 4° igualmente, se alivia el problema de la multicolinealidad.

Adicionalmente, la información obtenida mediante la RBN, esto es, la estimación de los IRR, que se considera relevante, no es mejorada por el modelo de cuadrados mínimos parciales.

4.2. Análisis externo

Los datos al inicio de la pandemia mostraron que el SARS-CoV2 es un virus altamente contagioso, capaz de expandirse en una población no protegida de forma exponencial, incluso más con las variantes de interés clínico que surgieron después de la ola de contagios inicial (Elliot *et. al.*, 2021). En este trabajo, se constató que la tasa de crecimiento del brote epidémico en Guatemala ha seguido una tendencia lineal. Esto puede ser producto de la implementación de medidas de salud pública diseñadas para mitigar la pandemia en nuestro país (figura 11).

Las medidas mitigantes incluyen todas aquellas que reducen el contacto entre individuos, así como la obligatoriedad del uso de mascarillas en todo momento (Chu *et. al.*, 2020).

En lo concerniente a este trabajo, la restricción en la movilidad comunitaria, mandato acatado por la población desde el inicio pero que fue desvaneciendo conforme el paso de los meses (figura 13), es una de las medidas que se consideran pudo haber tenido un impacto significativo en la variabilidad de la incidencia de la COVID-19 (Sulyok y Walker, 2020).

Tal y como lo sugieren Kephart *et. al.* (2020), la movilidad comunitaria, medida por el uso de apartaos de telefonía celular, puede ser considerada un factor relevante con efecto significativo en la incidencia de COVID-19 en las poblaciones de Latinoamérica, incluyendo Guatemala. En este caso, los autores utilizaron la medición del *Odds Ratio (OR)* como medida de la magnitud del riesgo.

De forma análoga, Oztig y Askin (2020) utilizaron un modelo de regresión binomial negativa para estimar los coeficientes de las variables predictoras como medida de la magnitud del impacto que estas tienen sobre la incidencia de casos de COVID-19. En este caso, las variables estudiadas representaron la movilidad entre países, de acuerdo con el volumen de tráfico aéreo. La estimación de coeficientes les permitió estimar los IRR.

El trabajo realizado midió, de esa forma, los IRR para los diferentes sitios en los cuales se registró la movilidad comunitaria. Los coeficientes del modelo de RBN fueron exponenciados, debido a la naturaleza de la transformación logarítmica que sufre el modelo al ser la función transformadora común en los modelos de regresión de data de conteo. Se explican en la tabla siguiente:

Tabla XIX. **Interpretación de los IRR (Estimados)**

	Estimados	**P**	**2.5** %	**97.5** %	**Interpretación**
Interceptos	796.60	$<2x10^{-16}$	666.84	957.38	
Delta_1	2.73	$<2x10^{-16}$	2.29	3.25	Incremento de 1 unidad en y por cada incremento de 2.73 unidades en Delta
Residencial	0.98	$<2x10^{-16}$	0.97	0.99	Incremento de 1 unidad en y por cada decremento de 0.98 unidades porcentuales en afluencia a áreas residenciales

Fuente: elaboración propia, empleando Microsoft Excel.

De acuerdo con lo anteriormente expuesto, la variante Delta representa la principal razón de variabilidad en la incidencia de COVID-19 en Guatemala. Sin embargo, y de acuerdo con cómo se categorizaron las variantes Delta y Ómicron y con cómo se registraron, la información puede resumirse en presencia de variantes de interés, lo que indicaría que, las variantes son, en verdad, la principal fuente de variabilidad.

Referente a la movilidad comunitaria, el incremento en la afluencia a los sitios residenciales (IRR 0.98) representó disminución en el riesgo de aumento de casos de COVID-19.

CONCLUSIONES

1. Al describir la incidencia de casos diarios de COVID-19 en Guatemala, entre marzo de 2020 a abril de 2022, esta muestra variabilidad importante, con 5 olas de contagios, las más relevantes asociadas al ingreso al país de las variantes Delta y Ómicron. El acúmulo de casos presenta un crecimiento lineal con $R^2 = 0.93$.

2. Se caracterizó la movilidad de la población usuaria de dispositivos Android. En el inicio del brote, la afluencia a lugares residenciales incrementó 40 % y disminuyó cerca del 90 % a las demás. La tendencia a lo largo de la pandemia ha sido revertir estos cambios, sin verse afectada por la detección de las variantes Delta y Ómicron.

3. Se estimó y depuró un modelo de RBN (R^2aj = 0.357) para explicar la incidencia de casos de COVID-19 en función de la variabilidad en la movilidad poblacional, que se ve afectado por la variante Delta de forma significativa. La afluencia a los lugares residenciales (IRR=0.98), representa el sitio que tienen mayor correlación con la incidencia de casos de COVID-19 y, por alta correlación con el resto de las variables relacionadas, resume la información de estos de forma apropiada.

4. El mejor modelo estimado para explicar la variabilidad en la incidencia de casos de COVID-19 en Guatemala, marzo 2020 a abril 2022, depurada de acuerdo con AIC y R^2 ajustado, incluyó las variables predictoras referentes al estatus de las variantes Delta y la afluencia a las áreas residenciales.

RECOMENDACIONES

1. Establecer mecanismos de registro de datos que permitan dar seguimiento en tiempo real a la incidencia de COVID-19 en Guatemala, así como potenciar las capacidades diagnósticas para evitar retraso en la captación de pacientes y permitir comprender el comportamiento de la pandemia a mayor profundidad.

2. Diseñar sistemas que permitan registrar, de forma anónima, la afluencia a sitios de interés de toda la población de Guatemala, independientemente del dispositivo electrónico que utilicen, para facilitar este tipo de estudios, de aplicaciones diversas.

3. Adaptar las políticas de restricción de la movilidad comunitaria en Guatemala, como medida para mitigar el brote epidémico de COVID-19, de acuerdo con la información producida en la elaboración del modelo propuesto.

4. Prolongar la captación y análisis de información para optimizar el modelo de regresión elaborado, para medir la eficacia de las medidas implementadas, así como obtener información valiosa de aplicabilidad en la pandemia y otras situaciones sanitarias relevantes.

REFERENCIAS

1. Annie, F., Sirbu, C., Frazier, K., Broce, M., y Lucas, B. (septiembre, 2020). Hydroxychloroquine in Hospitalized Patients with COVID-19: Real-World Experience Assessing Mortality. *Pharmacotherapy: The Journal Of Human Pharmacology And Drug Therapy*, *40*(11), 1072-1081.

2. Aydogdu, M., Altun, E., Chung, E., Ren, G., Homer-Vanniasinkam, S., Chen, B., y Edirisinghe, M. (enero, 2021). Surface interactions and viability of coronaviruses. *Journal Of The Royal Society Interface*, *18*(174), 20200798.

3. Barchard, K., y Verenikina, Y. (febrero, 2013). Improving data accuracy: Selecting the best data checking technique. *Computers In Human Behavior*, *29*(5), 1917-1922.

4. Beigel, J., Tomashek, K., Dodd, L., Mehta, A., Zingman, B., & Kalil, A. et al. (noviembre, 2020). Remdesivir for the Treatment of Covid-19 — Final Report. *New England Journal Of Medicine*, *383*(19), 1813-1826. Recuperado de: https:// https://www.nejm.org/doi/10.1056/NEJMoa2007764.

5. Boopathi, S., Poma, A. B., y Kolandaivel, P. (abril, 2020). Novel 2019 coronavirus structure, mechanism of action, antiviral drug promises and rule out against its treatment. *Journal of Biomolecular Structure and Dynamics*, *39*(9), 3409-3418. Recuperado de:

https://www.tandfonline.com/doi/full/10.1080/07391102.2020.1758788.

6. Buheji, M., da Costa Cunha, K., Beka, G., Mavrić, B., Leandro, Y., y Souza, S. (octubre, 2020). The Extent of COVID-19 Pandemic Socio-Economic Impact on Global Poverty. A Global Integrative Multidisciplinary Review. *American Journal Of Economics*, *10*(4), 213-224. Recuperado de: http://article.sapub.org/10.5923.j.economics.20201004.02.html.

7. Chan, H., Skali, A., Savage, D., Stadelmann, D., y Torgler, B. (noviembre, 2020). Risk attitudes and human mobility during the COVID-19 pandemic. *Scientific Reports*, *10*(1). Recuperado de: https://www.nature.com/articles/s41598-020-76763-2.

8. Chan, J., Yuan, S., Kok, K., To, K., Chu, H., y Yang, J. (enero, 2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*, *395*(10223), 514-523. Recuperado de: https://doi.org/10.1016/s0140-6736(20)30154-9.

9. Chen, C. Y., Chang, C. K., Chang, Y. W., Sue, S. C., Bai, H. I., Riang, L., Hsiao, C. D., y Huang, T. H. (mayo, 2007). Structure of the SARS coronavirus nucleocapsid protein RNA-binding dimerization domain suggests a mechanism for helical packaging of viral RNA. *Journal of molecular biology*, *368*(4), 1075–1086. Recuperado de: https://www.sciencedirect.com/science/article/pii/S0022283607002616?via%3Dihub.

10. Chu, D., Duda, S., Solo, K., Yaacoub, S., y Schunemann, H. (octubre, 2020). Physical Distancing, Face Masks, and Eye Protection to Prevent Person-to-Person Transmission of SARS-CoV-2 and COVID-19: A Systematic Review and Meta-Analysis. *Journal Of Vascular Surgery*, *72*(4), 1500. Recuperado de: https://www.jvascsurg.org/article/S0741-5214(20)31563-9/fulltext.

11. Coxe, S., West, S. G., y Aiken, L. S. (febrero, 2009). The analysis of count data: A gentle introduction to poisson regression and its alternatives. *Journal of Personality Assessment*, *91*(2), 121–136. Recuperado de: https://www.tandfonline.com/doi/abs/10.1080/00223890802634175.

12. Crick, F., y Watson, J. (febrero, 1956). Structure of Small Viruses. *Nature*, *177*(4506), 473-475.

13. Cummings, M., Baldwin, M., Abrams, D., Jacobson, S., Meyer, B., y Balough, E. (mayo, 2020). Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: a prospective cohort study. *The Lancet*, *395*(10239), 1763-1770. Recuperado de: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31189-2/fulltext.

14. Daniel, W., y León Hernández, F. (2014). *Bioestadística*. México: Limusa Wiley.

15. DeCarlo, L. (junio, 1997). On the meaning and use of kurtosis. *Psychological Methods*, *2*(3), 292-307. Recuperado de: https://psycnet.apa.org/doiLanding?doi=10.1037%2F1082-989X.2.3.292.

16. Del-Río, C., Alcocer-Gamba, M., Escudero-Salamanca, M., Galindo-Fraga, A., Guarner, J., y Escudero, X. (octubre, 2020). La pandemia de coronavirus SARSCoV-2 (COVID-19): situación actual e implicaciones para México. *Cardiovascular And Metabolic Science*, *31*(S3), 170-177. Recuperado de: https://www.medigraphic.com/cgi-bin/new/resumen.cgi?IDARTICULO=93943.

17. Famoye, F. (febrero, 2010). On the bivariate negative binomial regression model. *Journal of Applied Statistics*, 37(6), 969–981. Recuperado de: https://www.tandfonline.com/doi/full/10.1080/02664760902984618

18. Ghinai, I., McPherson, T., Hunter, J., Kirking, H., Christiansen, D., y Joshi, K. (marzo, 2020). First known person-to-person transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the USA. *The Lancet*, *395*(10230), 1137-1144. Recuperado de: https://linkinghub.elsevier.com/retrieve/pii/S0140673620306073.

19. Groeneveld, R., y Meeden, G. (diciembre, 1984). Measuring Skewness and Kurtosis. *The Statistician*, *33*(4), 391. Recuperado de: https://www.jstor.org/stable/2987742?origin=crossref.

20. Hanvoravongchai, P., Obando, C., Petrosyan, V., Rao, K. D., Ruano, A. L., Shi, L., Souza, L. E., Spitzer-Shohat, S., Sturgiss, E., Suphanchaimat, R., Uribe, M. V., y Willems, S. (octubre, 2020). Health equity and COVID-19: global perspectives*. International Journal for Equity in Health*, *19*(104), 1–16.

21. Hillen, H. (junio, 2021). Structure and function of SARS-CoV-2 polymerase. Recuperado de *Current Opinion In Virology*, 48, 82-90. https://pubmed.ncbi.nlm.nih.gov/33945951/.

22. Hu, Y., Wen, J., Tang, L., Zhang, H., Zhang, X., Li, Y., Wang, J., Han, Y., Li, G., Shi, J., Tian, X., Jiang, F., Zhao, X., Wang, J., Liu, S., Zeng, C., Wang, J., y Yang, H. (mayo, 2003). The M protein of SARS-CoV: basic structural and immunological properties. *Genomics, proteomics & bioinformatics*, *1*(2), 118–130. Recuperado de https://www.sciencedirect.com/science/article/pii/S1672022903010167?via%3Dihub.

23. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., y Hu, Y. (enero, 2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), 497-506. Recuperado de: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30183-5/fulltext.

24. Huremović D. (2019). *Brief History of Pandemics (Pandemics Throughout History).* Estados Unidos: Springer International Publishing. Recuperado de: https://link.springer.com/chapter/10.1007/978-3-030-15346-5_2.

25. Kephart, J. L., Delclòs-Alió, X., Rodríguez, D. A., Sarmiento, O. L., Barrientos-Gutiérrez, T., Ramirez-Zea, M., Quistberg, D. A., Bilal, U., y Roux, A. V. D. (agosto, 2021). The effect of population mobility on COVID-19 incidence in 314 Latin American cities: a longitudinal ecological study with mobile phone location data. *The Lancet Digital Health*, *3*(11). Recuperado de: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00174-6/fulltext.

26. Kim, C. H. (junio, 2020). SARS-CoV-2 Evolutionary Adaptation toward Host Entry and Recognition of Receptor O-Acetyl Sialylation in Virus-Host Interaction. *International Journal of Molecular Sci*ences, *21*(12), 4549. Recuperado de: https://www.mdpi.com/1422-0067/21/12/4549.

27. Li, Y., Zhou, W., Yang, L., y You, R. (julio, 2020). Physiological and pathological regulation of ACE2, the SARS-CoV-2 receptor. *Pharmacological Research*, 157, 104833. Recuperado de: https://pubmed.ncbi.nlm.nih.gov/32302706/.

28. Lindsey, C., y Sheather, S. (enero, 2010). Variable selection in linear regression. *The Stata Journal*, *10*(4), 650-669. Recuperado de: https://www.researchgate.net/profile/Jos_Feys/post/How_to_use_the_factor_scores_for_regression_analysis/attachment/5e0cc353cfe4a777d4fff099/AS %3A842565758771200 %401577894739782/download/STATA_regression.pdf.

29. Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., y Wu, H. (enero, 2020). Genomic characterization and epidemiology of 2019 novel coronavirus:

implications for virus origins and receptor binding. *The Lancet*, *395*(10224), 565-574. Recuperado de: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30251-8/fulltext.

30. Ministerio de Salud Pública y Asistencia Social (2022). *COVID-19 en Guatemala*. Guatemala: Autor.

31. Oztig, L. I., y Askin, O. E. (agosto, 2020). Human mobility and coronavirus disease 2019 (COVID-19): a negative binomial regression analysis. *Public Health*, *185*, 364–367. Recuperado de: https://www.sciencedirect.com/science/article/pii/S0033350620302973?via%3Dihub.

32. Rodríguez, G. (2007). *Lecture Notes* on g*eneralized linear models*. Estados Unidos: Autor.

33. Schoeman, D., y Fielding, B. (mayo, 2019). Coronavirus envelope protein: current knowledge. *Virology Journal*, *16*(1), 1-20. Recuperado de: https://virologyj.biomedcentral.com/articles/10.1186/s12985-019-1182-0.

34. Statcounter Global Stats. (2022). *Mobile Operating System Market Share Guatemala*. Guatemala: Autor.

35. Sulyok, M., y Walker, M. (noviembre, 2020). Community movement and COVID-19: a global study using Google's Community Mobility Reports. *Epidemiology And Infection*, *148*(e284), 1-9. Recuperado

de: https://www.cambridge.org/core/journals/epidemiology-and-infection/article/community-movement-and-covid19-a-global-study-using-googles-community-mobility-reports/C1EF917377F6013D5E1ECC44E03A3E6E.

36. Vittinghoff. E., Glidden, D., Shiboski, S., y McCulloch, C. (2012). *Regression methods in biostatistics. Linear, Logistic, Survival and Repeated Measures Models*. Estados Unidos: Springer.

37. Walpole, R., Myers, R., y Myers, S. (2012). *Probabilidad y estadística para ingeniería y ciencias*. Mexico: Pearson Educación.

38. Wang, M., Zhao, R., Gao, L., Gao, X., Wang, D., y Cao, J. (noviembre, 2020). SARS-CoV-2: Structure, Biology, and Structure-Based Therapeutics Development. *Frontiers In Cellular And Infection Microbiology*, *10*, 1-20.

39. Westfall, P. H. (enero, 2014). Kurtosis as Peakedness, 1905 – 2014. *R.I.P*. *The American statistician*, *68*(3), 191–195. Recuperado de: https://www.tandfonline.com/doi/abs/10.1080/00031305.2014.917055.

40. Woo, P. C. Y., Lau, S. K. P., Huang, Y., y Yuen, K. Y. (octubre, 2009). Coronavirus diversity, phylogeny and interspecies jumping. *Experimental Biology and Medicine*, *234*(10), 1117–1127. Recuperado de https://journals.sagepub.com/doi/10.3181/0903-MR-94.

41. Wu, J., Leung, K., y Leung, G. (enero, 2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet*, *395*(10225), 689-697. Recuperado de: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30260-9/fulltext.

APÉNDICES

Apéndice 1. **Título: Elaboración de un modelo de regresión binomial negativa splocsdo a la incidencia de COVID-19, en función de los cambios de movilidad poblacional, en Guatemala**

PROBLEMA	PREGUNTAS DE INVESTIGACIÓN	OBJETIVOS	METODOLOGÍA	RESULTADOS	CONCLUSIONES	RECOMENDACIONES
Principal No se ha elaborado un modelo de regresión para explicar la incidencia de covid19 en función de los cambios en la movilidad poblacional en Guatemala.	Principal ¿Qué modelo de regresión permite explicar la incidencia de covid19 en función de los cambios de la movilidad poblacional en Guatemala?	Principal Elaborar un modelo de regresión para explicar la incidencia de covid19 en función de los cambios en la movilidad poblacional en Guatemala.	Tipo de investigación: descriptiva, correlacional, en la línea de pronósticos y análisis multivariado Diseño: No experimental, retrospectiva, transversal, descriptivo, correlacional Variables - Incidencia de COVID19 - Variabilidad semanal en la movilidad comunitaria a los diferentes sitios posibles (según categorías preespecificadas) - Tiempo de incubación o latencia - Presencia de variante Delta - Presencia de variante Ómicron - Vacunación *per cápita*. **Universo:** población de la República de Guatemala **Población:** individuos con teléfono celular con SO Android, con registro de movilidad activado y permitido. **Unidad de análisis:** data de telefonía móvil anonimizada, disponible en la web, provista por Google, resumida en porcentajes, incidencia diaria de casos de COVID19. **Muestreo**: no probabilístico, por conveniencia	Modelo estimado: $\ln(incidencia\_covid19) = \beta_0 + \beta_1 * Delta_1 + \beta_2 * Ômicron_1 + \beta_3 * (var_R) + \beta_4 * (var_{PRK}) + \beta_5 * (var_{TRST}) + \beta_6 (var_{WRK})$, **IRR estimados** Interceptos 386.73 Delta_1 3.80 RES 0.89 TRST 0.94 Omicron_1 2.28 WRK 1.009 RR 1.01	Conclusión general El mejor modelo estimado para explicar la variabilidad en la incidencia de casos de COVID19 en Guatemala, marzo 2020 a abril 2022, depurada de acuerdo con AIC y R2 ajustado, incluyó las variables predictoras referentes al estatus de las variantes Delta y Ómicron, así como la afluencia a los sitios referentes a Áreas residenciales, Sitios de trabajo, Estaciones de transporte y Tiendas y Ocio.	Recomendación general Prolongar la captación y análisis de información para optimizar el modelo de regresión elaborado, para medir la eficacia de las medidas implementadas, así como obtener información valiosa de aplicabilidad en la pandemia y otras situaciones sanitarias relevantes.
Secundarios No se ha caracterizado cuál ha sido el comportamiento de la incidencia de casos nuevos de covid19 en la población de Guatemala.	Preguntas específicas ¿Cuál ha sido el comportamiento de la incidencia de casos nuevos de COVID19 en la población de Guatemala?	Objetivos específicos Describir cuál ha sido el comportamiento de la incidencia bruta de casos nuevos de COVID19 en la población de Guatemala.		Se registraron 832,956 casos en 754 días consecutivos. Se detectaron 5 olas de casos, 3 explicadas por fenómenos sociales y 2, las mayores, como consecuencia de variantes más contagiosas de SARS-CoV2. El crecimiento de la pandemia tuvo un modelo lineal, R=0.93.	Conclusiones específicas Al describir la incidencia bruta de casos diarios de COVID19 en Guatemala, entre marzo de 2020 a abril de 2022, esta muestra variabilidad importante, presentó 5 olas de contagios, de las cuales las más relevantes están asociadas al ingreso al país de las variantes Delta y Ómicron. El acúmulo de casos presenta un crecimiento lineal con $R^2 = 0.93$.	Recomendaciones específicas Establecer mecanismos de registro de datos que permitan dar seguimiento en tiempo real a la incidencia de COVID19 en Guatemala, así como potenciar las capacidades diagnósticas para evitar retraso en la captación de pacientes y permitir comprender el comportamiento de la pandemia a mayor profundidad.
No se ha caracterizado cuál ha sido el comportamiento de la variabilidad de la movilidad comunitaria, durante la pandemia de COVID19, en la población de Guatemala.	¿Cuál ha sido el comportamiento y la variabilidad de la movilidad comunitaria, durante la pandemia de COVID19, en la población de Guatemala?	Caracterizar cuál ha sido el comportamiento de la movilidad comunitaria, durante la pandemia de COVID19, en la población de Guatemala.		Contrastado contra la mediana del día correspondiente, del período 3/1 a 6/2 de 2020, la variabilidad porcentual en la movilidad comunitaria, se detectó incremento de 40% en afluencia a residencias y reducción de 70-90% en afluencia a los demás sitios en semanas posterior a inicio del brote. Tras estas semanas, tendencia a revertir esto, con incremento en todos los sitios, excepto residencia, sin que las olas de contagios ni detección de variantes de interés tuviese influencia en esto.	Se caracterizó la movilidad de la población usuaria de dispositivos Android®. En el inicio del brote, la afluencia a lugares residenciales incrementó en 40%, mientras que disminuyó cerca del 90% a las demás. La tendencia a lo largo de la pandemia ha sido revertir estos cambios, sin verse afectada por la detección de las variantes Delta y Ómicron.	Diseñar sistemas que permitan registrar, de forma anónima, la afluencia a sitios de interés de toda la población de Guatemala, independientemente del dispositivo electrónico que utilicen, para facilitar este tipo de estudios, de aplicaciones diversas.

Continuación apéndice 1.

No se ha determinado cuál es la robustez del modelo de regresión a elaborar mediante la evaluación de los criterios de apropiados, para explicar la incidencia de COVID19 en función de los cambios de movilidad poblacional en Guatemala.	¿Cuál es la robustez del modelo de regresión a elaborar mediante la evaluación de los criterios de información pertinentes, para explicar la incidencia de COVID19 en función de los cambios de movilidad poblacional en Guatemala?	Determinar cuál es la robustez del modelo de regresión a elaborar mediante la evaluación de los criterios de información apropiados para la naturaleza de dicho modelo, para explicar la incidencia de COVID19 en función de los cambios de movilidad poblacional en Guatemala.	**Técnica:** elaboración de modelo de regresión binomial negativa. **Instrumento:** bases de datos en línea, todas de acceso abierto: Ourworldindata.org	Modelo de RBN, con R2aj = 28%, $\ln(incidencia\_covid19) = \beta_0 + \beta_1 * Delta_1 + B_2 * Ômicron_1 + \beta_3 * (var_R) + \beta_4 * (var_{PRKS}) + \beta_5 * (var_{PRVT}) + \beta_6 (var_{WRK})$. Con transformación logit (Ln) de la variable respuesta.	Se estimó y depuró un modelo de RBN (R2aj = 28%), depurado por AIC, para explicar la incidencia de casos de COVID19 en función de la variabilidad en la movilidad poblacional, que se ve afectado por las variantes Delta y Ómicron de forma significativa. La afluencia a los lugares residenciales, de transporte, de trabajo y tiendas y ocio fueron incluidos y representan los sitios que tienen mayor correlación con la incidencia de casos de COVID19.	Adaptar las políticas de restricción de la movilidad comunitaria en Guatemala, como medida para mitigar el brote epidémico de COVID19, de acuerdo con la información producida en la elaboración del modelo propuesto.

Fuente: elaboración propia, empleando Microsoft Excel.

Apéndice 2. **Código en RStudio utilizado para estimación de modelo de regresión**

```
#### Elaboración de modelo de RBN ###

library(performance)
library(boot)
library("reshape2")
library(DHARMa)
library(foreign)
library(AICcmodavg)
library(vcd)
library(tidyverse)
library(igraph)
library(dplyr)
library(dlookr)
library(MASS)
library(pscl)
library(AER)
library(ggplot2)
library(tibble)
library(lubridate)
library(pkr)
library(summarytools)
library(jtools)

#estimar modelo de Poisson
pois <- glm(new_cas_lag ~ RRpct+GPpct+PRKSpct+TRSTpct+WRKpct+RESpct+Omicron+Delta+vac_pob,
data = full_tasas, family = poisson)

summary(pois)

#valorar si hay sobredispersión
dispersiontest(pois)

#Hay sobredispersión, corresponde estimar un modelo de regresión binomial negativa
#Se estima el modelo de reg bin negativa

bneg <- glm.nb(new_cas_lag ~ RRpct+GPpct+PRKSpct+TRSTpct+WRKpct+RESpct+Omicron+Delta,
data = full_df)
bneg
summary(bneg)

##descripción de movilidad comunitaria
head(full_df)

#Incluir tasa (creando nuevo df)
pop_gt <- 16860000
full_tasas <- full_df  %> % mutate(tasa = (new_cases / pop_gt)*100000)
str(covid_dicotomico)
covid_dicotomico <- as.factor(covid_dicotomico)
full_tasas <- full_tasas  %> % mutate(covid=as.factor(covid_dicotomico$COVID-19))
covid_dicotomico

#Incluir fechas respectivas
fechas <- as_date(mdy("02-17-20"):mdy("04-06-22"))
fechas
full_tasas <- full_tasas  %> % mutate(fechas = fechas)

#Incluir columna con suma acumulativa
full_tasas <- full_tasas  %> % mutate (acum = full_tasas$new_cases <- cumsum(full_tasas$new_cases))
```

Continuación apéndice 2.

```
#Visualizar resumen de nuevo df
summary(full_tasas)
full_tasas$fechas
covid_dicotomico
#Gráfico de incidencia diaria, con líneas que representan la detección de variantes
incidencia <- ggplot(full_tasas, aes(fechas,tasa)) + geom_line() + theme_light() +
labs(x = "Fechas", y = "Incidencia de casos diarios / 100,000 hab.") +
geom_vline(xintercept = fechas[430], col="red") +
geom_vline(xintercept = fechas[666], col="blue")
incidencia

#Histograma
histograma <- ggplot(full_tasas, aes(new_cases)) + geom_histogram() + theme_light() +
labs(x="Casos nuevos", y="Conteo") +
geom_density()
histograma
#Gráfico de incidencia acumulada, con líneas que representan la detección de variantes
acum <- ggplot(full_tasas, aes(fechas,acum)) + geom_line() + theme_light() +
labs(x = "Fechas", y = "Datos acumulados") +
geom_vline(xintercept = fechas[430], col="red") +
geom_vline(xintercept = fechas[666], col="blue")
acum


#resumen de casos diarios
summary(full_tasas$new_cases)

#Gráfico de incidencia diaria, con líneas que representan la detección de variantes
incidencia_vac <- ggplot(full_tasas, aes(fechas,vac_pob)) + geom_line() + theme_classic() +
labs(x = "Fechas", y = "Dosis de vacunas administradas por 100,000 habitantes") +
geom_vline(xintercept = fechas[430], col="red") +
geom_vline(xintercept = fechas[666], col="blue")
incidencia_vac

###
summary(semanales)
full_tasas_sem <- semanales  %> % mutate(tasa_sem = (cas_sem_lag / pop_gt)*100000)

incidencia_sem <- ggplot(full_tasas_sem, aes(Semana,tasa_sem)) + geom_line() + theme_classic() +
labs(x = "Semana", y = "Incidencia de casos semanales / 100,000 hab.")
incidencia_sem

dosis_vac
str(dosis_vac)
str(full_tasas)

###Regresión
summary(full_df_lag)



bneg <- glm.nb(lagging ~ RRpct+GPpct+PRKSpct+TRSTpct+WRKpct+RESpct+
as.factor(Omicron)+as.factor(Delta)+vacunas,
data = full_df_lag)

bneg2 <- glm.nb(lagging ~ RRpct+GPpct+PRKSpct+TRSTpct+WRKpct+RESpct+
as.factor(Omicron)+as.factor(Delta)+vacunas,
link=sqrt,
data = full_df_lag)
bneg2
bneg
```

Continuación apéndice 2.

```
pr2 <- (full_df_lag$TRSTpct+full_df_lag$RESpct)/2
summary(pr2)

pr3 <- (full_df_lag$RRpct+full_df_lag$WRKpct)/2

pR3 = 1 - bneg3$deviance / bneg3$null.deviance
pR3


bneg3 <- glm.nb(lagging ~ pr2 + pr3 +
as.factor(Omicron)+as.factor(Delta)+vacunas,
data = full_df_lag)
summary(bneg3)


durbinWatsonTest(bneg)
diag <- glm.diag.plots(bneg)
check_model(modelo_paso_)

summary(bneg)
pR3 = 1 - bneg$deviance /bneg$null.deviance
pR3

modelo_nulo<-glm.nb(lagging~1,data=full_df_lag)
step(modelo_nulo, scope = list(lower = modelo_nulo,
upper = bneg),
direction = "both", ) -> modelo_paso_

(est <- cbind(Estimate = coef(modelo_paso_), confint(modelo_paso_)))
exp(est)

summary(est)


summary(modelo_paso_)

round((est <- cbind(Estimate = coef(modelo_paso_), confint(modelo_paso_))), 4)

round(exp(est),4)

par(mfrow=c(2,2))
plot(modelo_paso_)
```

Fuente: elaboración propia, empleando Microsoft Excel.