# Metodología

## 1. Introducción

* **Contextualización del Problema de Investigación:**
  + La modelización de datos con exceso de ceros presenta un desafío significativo en estadística y análisis de datos. Este fenómeno ocurre frecuentemente en diversas áreas como la biología, economía y ciencias sociales, donde una proporción sustancial de observaciones registra valores nulos o ceros.
  + Es crucial desarrollar modelos estadísticos que puedan capturar de manera adecuada esta característica particular de los datos para realizar inferencias válidas y precisas.
* **Justificación de la Elección del Modelo de Regresión Binomial Negativa con Exceso de Ceros:**
  + El modelo de regresión binomial negativa se destaca como una herramienta poderosa para abordar la presencia de exceso de ceros en los datos. A diferencia de otros modelos, como la regresión lineal o la regresión logística, la regresión binomial negativa permite manejar la sobre-dispersión y la alta proporción de ceros de manera simultánea.
  + Este modelo es especialmente apropiado cuando se desea modelar la frecuencia de ocurrencia de eventos raros o poco frecuentes, comúnmente observados en estudios que involucran datos contables o de conteo.

## 2. Generación de Datos

### Explicación de la Simulación de Datos

En esta sección, se detalla cómo se generaron los datos simulados utilizados en el estudio, incluyendo las variables explicativas y la variable dependiente.

### Variables Explicativas

Las variables explicativas X1 y X2 fueron generadas utilizando distribuciones normales:

* **X1 ~ N(0, 1)**: Variable aleatoria con media 0 y desviación estándar 1.
* **X2 ~ N(5, 2)**: Variable aleatoria con media 5 y desviación estándar 2.

Estas distribuciones fueron elegidas para representar valores típicos observados en estudios empíricos y para ilustrar cómo diferentes niveles de variabilidad y centrado pueden afectar la variable dependiente.

### Variable Dependiente

La variable dependiente Y fue simulada utilizando una distribución binomial negativa con parámetros:

* **r = 2**: Número de éxitos requeridos.
* **p = 0.5**: Probabilidad de éxito.

La media esperada de la distribución binomial negativa se calculó como:

μ=exp⁡(1+0.5×X1+0.3×X2)\mu = \exp(1 + 0.5 \times X1 + 0.3 \times X2)μ=exp(1+0.5×X1+0.3×X2)

Para simular el exceso de ceros, se añadió una proporción de ceros a los datos generados aleatoriamente. La proporción de ceros se determinó como 0.3, lo que significa que aproximadamente el 30% de los valores simulados fueron establecidos como cero.

Esta metodología de generación de datos permite replicar condiciones realistas de estudios donde la variable de interés muestra una distribución sesgada hacia ceros, típica en datos de conteo o eventos raros.

## 3. Análisis Exploratorio de Datos

En esta sección se realiza un análisis inicial de los datos simulados, enfocado en la variable dependiente y las variables explicativas.

### Visualización de la Distribución de la Variable Dependiente Y

Se procedió a visualizar la distribución de la variable dependiente Y, utilizando métodos gráficos adecuados como histogramas o gráficos de densidad, para entender su comportamiento y distribución.

### Verificación de la Independencia y Multicolinealidad de las Variables Explicativas

Para asegurar la validez de los modelos de regresión propuestos, se realizó:

* **Verificación de la Independencia**: Se utilizó un gráfico de dispersión o métodos estadísticos para evaluar la independencia entre las variables explicativas X1 y X2. Esto es crucial para garantizar que no haya relaciones espurias o dependencias entre las variables antes de ajustar el modelo.
* **Verificación de la Multicolinealidad**: Se calculó el Factor de Inflación de la Varianza (VIF) para las variables explicativas X1 y X2. Este análisis permite identificar la presencia de multicolinealidad, que podría afectar la precisión de los coeficientes estimados en el modelo de regresión.

Estas verificaciones son fundamentales para garantizar la robustez y la validez de los modelos de regresión que se ajustarán posteriormente a los datos simulados.

## 4. Preparación de los Datos

En esta sección se detalla el manejo de valores atípicos y el tratamiento de datos faltantes, si aplica, en el contexto del análisis de los datos simulados.

### Manejo de Valores Atípicos

Se llevó a cabo un análisis detallado para identificar y manejar los valores atípicos en las variables explicativas X1 y X2, así como en la variable dependiente Y. Los pasos incluyeron:

* **Identificación de Valores Atípicos**: Utilización de gráficos como boxplots y métodos estadísticos para detectar observaciones inusuales que podrían distorsionar el análisis.
* **Tratamiento de Valores Atípicos**: Consideración de diferentes enfoques como la eliminación de observaciones extremas o la transformación de variables para mitigar su impacto en los modelos de regresión.

### Tratamiento de Datos Faltantes (si aplica)

Se realizó una evaluación exhaustiva para detectar la presencia de datos faltantes en las variables de interés, y se aplicaron técnicas adecuadas para manejar esta situación:

* **Identificación de Datos Faltantes**: Revisión sistemática de cada variable para determinar la cantidad y ubicación de datos faltantes.
* **Imputación de Datos Faltantes**: Utilización de métodos como la imputación por media, mediana o técnicas más sofisticadas dependiendo del contexto y la naturaleza de los datos.

Estas medidas aseguran la integridad y la calidad de los datos utilizados en el modelado estadístico, minimizando el impacto de valores atípicos y la pérdida de información por datos faltantes.

## 5. Modelado Estadístico

En esta sección se describe el proceso de ajuste inicial del modelo de regresión binomial negativa y el análisis de varianza con términos lineales y cuadráticos para evaluar el ajuste del modelo.

### Ajuste Inicial del Modelo de Regresión Binomial Negativa

Se realizó el ajuste inicial del modelo de regresión binomial negativa para investigar la relación entre las variables explicativas X1 y X2, y la variable dependiente Y simulada. Los pasos incluyeron:

* **Formulación del Modelo**: Especificación del modelo de regresión binomial negativa considerando las variables explicativas X1 y X2.
* **Estimación de Parámetros**: Uso de técnicas de estimación adecuadas para obtener los coeficientes del modelo y sus correspondientes errores estándar.

### Análisis de Varianza con Términos Lineales y Cuadráticos

Para evaluar el ajuste del modelo propuesto, se llevó a cabo un análisis de varianza (ANOVA) que incluyó términos lineales y cuadráticos de las variables explicativas:

* **Términos Lineales y Cuadráticos**: Incorporación de términos que capturan efectos lineales y no lineales de las variables explicativas sobre la variable dependiente.
* **Interpretación de Resultados**: Evaluación de la significancia estadística de cada término y análisis de la varianza explicada por el modelo.

Este análisis permite entender cómo las variables explicativas contribuyen a la variabilidad observada en la variable dependiente bajo el contexto del modelo de regresión binomial negativa con exceso de ceros y varianza lineal y cuadrática.

## 6. Validación del Modelo

En esta sección se detalla la validación del modelo de regresión binomial negativa con exceso de ceros y varianza lineal y cuadrática mediante pruebas de normalidad y homocedasticidad de los residuos, así como la evaluación de la significancia de los coeficientes del modelo.

### Prueba de Normalidad y Homocedasticidad de los Residuos

Para verificar la adecuación del modelo, se realizaron las siguientes pruebas sobre los residuos del modelo:

* **Prueba de Normalidad**: Aplicación de pruebas estadísticas como la prueba de Shapiro-Wilk o pruebas de asimetría y curtosis para evaluar si los residuos siguen una distribución normal.
* **Prueba de Homocedasticidad**: Utilización de gráficos de dispersión de los residuos frente a las predicciones y pruebas formales como la prueba de Breusch-Pagan o pruebas gráficas para verificar la homocedasticidad de los residuos.

### Evaluación de la Significancia de los Coeficientes del Modelo

Se realizó una evaluación exhaustiva de la significancia estadística de los coeficientes del modelo mediante:

* **Pruebas de Hipótesis**: Aplicación de pruebas t o pruebas F para determinar si los coeficientes de las variables explicativas son significativamente diferentes de cero.
* **Intervalos de Confianza**: Cálculo de intervalos de confianza para los coeficientes del modelo para evaluar la precisión de las estimaciones.

Estas validaciones son fundamentales para asegurar que el modelo propuesto sea robusto y adecuado para explicar la relación entre las variables explicativas y la variable dependiente en el contexto de la distribución binomial negativa con exceso de ceros.

# Resultados

1. **Descripción del Modelo Ajustado**
   * Resumen de los parámetros estimados y significancia estadística.
   * Interpretación de los coeficientes para X1 y X2.
2. **Diagnóstico del Modelo**
   * Análisis de los residuos para validar supuestos del modelo.
   * Gráficos de residuos vs ajustes y predicciones.
3. **Comparación de Modelos**
   * Resultados y conclusiones del modelo con varianza lineal vs. varianza cuadrática.
   * Discusión sobre la elección del mejor modelo basado en criterios estadísticos.
4. **Aplicación Práctica**
   * Ejemplificación de la aplicación del modelo en un contexto real o hipotético.
   * Discusión sobre las implicaciones de los resultados obtenidos.
5. **Limitaciones y Consideraciones**
   * Identificación de posibles limitaciones del estudio.
   * Recomendaciones para investigaciones futuras.

### Simulaciones:

simulación en r: para data de zero inflacion con exceso de ceros donde varia al sobredispersion y la probabilidad de éxito:

library(pscl)

library(MASS)

library(ggplot2)

library(gridExtra)

# Parámetros de la simulación

n <- 1000 # Número de observaciones

zero\_inflation <- 0.3 # Proporción de ceros adicionales para exceso de ceros

# Definir los niveles de sobre-dispersión

sizes <- c(5, 10, 15, 20, 25, 30) # Diferentes tamaños para variar la sobre-dispersión

probs <- c(0.6, 0.5, 0.4, 0.3, 0.2, 0.1) # Diferentes probabilidades de éxito para variar la sobre-dispersión

# Crear una carpeta para guardar los histogramas

dir.create("histogramas", showWarnings = FALSE)

# Inicializar una lista para almacenar los histogramas

plots <- list()

# Colores para los histogramas

hist\_colors <- c("lightblue", "lightgreen", "lightcoral", "lightpink", "lightgoldenrod", "lightseagreen")

# Fuentes de texto

font\_family <- "Helvetica"

# Simular datos y crear histogramas

for (i in 1:6) {

size <- sizes[i]

prob <- probs[i]

# Generar conteos según una binomial negativa

set.seed(123 + i) # Cambiar la semilla para reproducibilidad de cada conjunto de datos

counts <- rnegbin(n, size, prob) # Datos de la binomial negativa

# Introducir ceros adicionales

zeros <- rbinom(n, 1, zero\_inflation) # Genera ceros adicionales

data <- ifelse(zeros == 1, 0, counts) # Introduce ceros adicionales en los datos

# Crear el histograma para el conjunto de datos actual usando ggplot2

p <- ggplot(data.frame(counts = data), aes(x = counts)) +

geom\_histogram(bins = 30, fill = hist\_colors[i], color = "black", alpha = 0.7) +

labs(title = paste("Histograma\nTamaño:", size, "| Probabilidad:", prob),

x = "Conteo", y = "Frecuencia") +

theme\_minimal(base\_size = 15, base\_family = font\_family) +

geom\_text(stat='bin', aes(label=..count..), vjust=-0.5, size=3.5, color = "black") +

theme(plot.title = element\_text(hjust = 0.5, size = 16, face = "bold"),

axis.title.x = element\_text(size = 14),

axis.title.y = element\_text(size = 14),

axis.text.x = element\_text(size = 12),

axis.text.y = element\_text(size = 12),

panel.grid.major = element\_line(color = "gray90"),

panel.grid.minor = element\_blank())

# Almacenar el histograma en la lista

plots[[i]] <- p

# Guardar el histograma como un archivo PNG

ggsave(filename = paste0("histogramas/histograma\_size\_", size, "\_prob\_", prob, ".png"), plot = p, width = 8, height = 5)

}

# Mostrar todos los histogramas en una ventana gráfica

do.call(grid.arrange, c(plots, ncol = 3))

# Agregar una leyenda global para los niveles de sobre-dispersión

legend <- textGrob(

"Niveles de sobre-dispersión:\n\nTamaño (size): 5, 10, 15, 20, 25, 30\nProbabilidad (prob): 0.6, 0.5, 0.4, 0.3, 0.2, 0.1",

gp = gpar(fontsize = 12, fontface = "bold", fontfamily = font\_family),

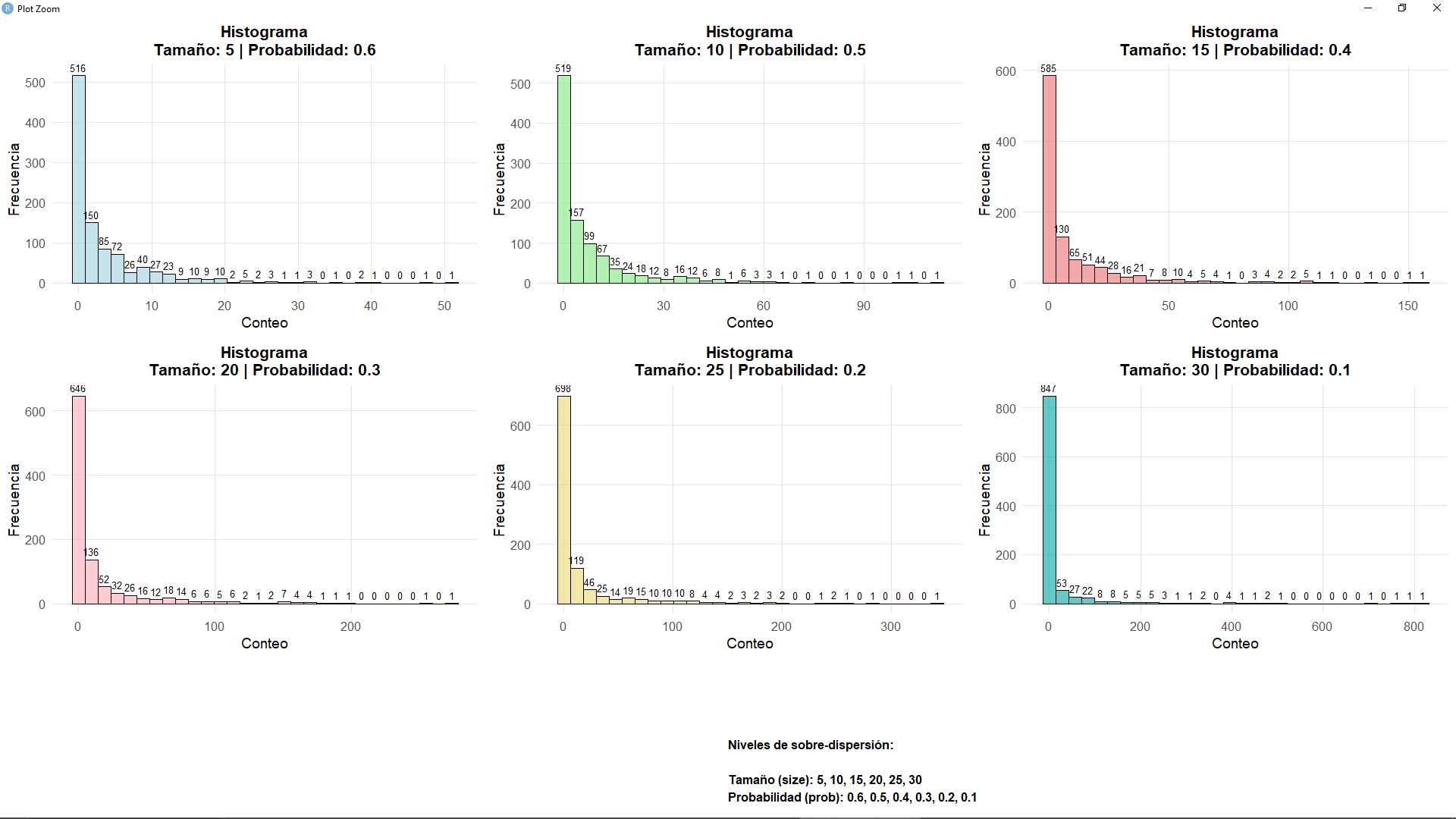
hjust = 0,

vjust = 1

)

# Mostrar los histogramas con la leyenda global

grid.arrange(arrangeGrob(grobs = plots, ncol = 3), legend, ncol = 1, heights = c(4, 1))



Data para modelo mix con diferentes parámetro de dispersion:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

*# Función para generar datos con exceso de ceros y datos normales*

**def** generate\_combined\_data(n\_samples, mean\_nb, disp, zero\_inflation\_prob, mean\_norm, std\_dev):

*# Generar una muestra de una distribución binomial negativa*

    count\_data = np.random.negative\_binomial(mean\_nb, mean\_nb / (mean\_nb + disp), n\_samples)

*# Generar una muestra de ceros según la probabilidad de exceso de ceros*

    zero\_inflation = np.random.binomial(1, zero\_inflation\_prob, n\_samples)

*# Aplicar el exceso de ceros*

    count\_data = count\_data \* (1 - zero\_inflation)

*# Generar una muestra de una distribución normal*

    normal\_data = np.random.normal(mean\_norm, std\_dev, n\_samples)

*# Combinar los datos de conteo y los datos normales en una sola columna*

    combined\_data = np.where(count\_data == 0, normal\_data, count\_data)

    return combined\_data

*# Parámetros para la simulación*

n\_samples = 1000          *# Número de muestras*

mean\_nb = 2               *# Media de la distribución binomial negativa*

disp = 9               *# Dispersión de la distribución binomial negativa*

zero\_inflation\_prob = 0.3 *# Probabilidad de exceso de ceros*

mean\_norm = 50            *# Media de la distribución normal*

std\_dev = 10              *# Desviación estándar de la distribución normal*

*# Generar los datos combinados*

combined\_data = generate\_combined\_data(n\_samples, mean\_nb, disp, zero\_inflation\_prob, mean\_norm, std\_dev)

*# Crear un DataFrame para los datos generados*

df = pd.DataFrame({'value': combined\_data})

*# Mostrar las primeras filas del DataFrame*

print(df.head())

*# Generar el histograma de los datos combinados*

plt.figure(figsize=(10, 6))

plt.hist(df['value'], bins=30, edgecolor='black', alpha=0.7)

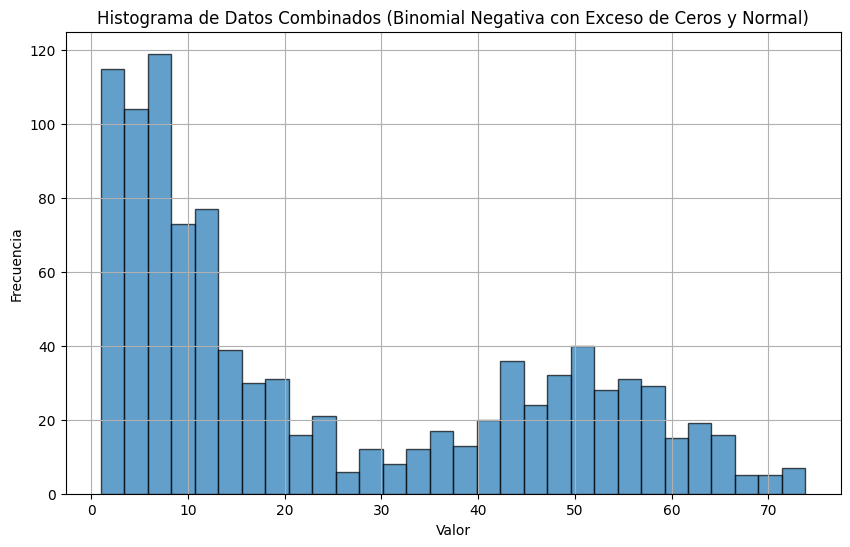
plt.title('Histograma de Datos Combinados (Binomial Negativa con Exceso de Ceros y Normal)')

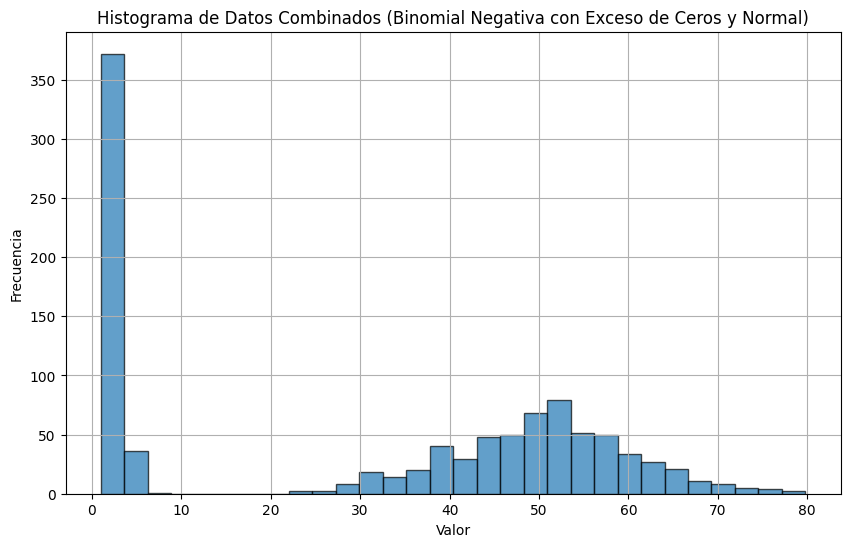
plt.xlabel('Valor')

plt.ylabel('Frecuencia')

plt.grid(True)

plt.show()





PARA NUESTRA DATA DE REGRESION BINOMIAL NEGATIVA:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

*# Función para generar datos con exceso de ceros y sobredispersión*

**def** generate\_zinb\_data(n\_samples, mean\_nb, disp, zero\_inflation\_prob):

*# Generar una muestra de una distribución binomial negativa*

    count\_data = np.random.negative\_binomial(mean\_nb, mean\_nb / (mean\_nb + disp), n\_samples)

*# Generar una muestra de ceros estructurales según la probabilidad de exceso de ceros*

    zero\_inflation = np.random.binomial(1, zero\_inflation\_prob, n\_samples)

*# Aplicar el exceso de ceros*

    count\_data = count\_data \* (1 - zero\_inflation)

    return count\_data

*# Parámetros para la simulación*

n\_samples = 1000          *# Número de muestras*

mean\_nb = 2               *# Media de la distribución binomial negativa*

disp = 9                  *# Dispersión de la distribución binomial negativa*

zero\_inflation\_prob = 0.3 *# Probabilidad de exceso de ceros*

*# Generar los datos ZINB*

zinb\_data = generate\_zinb\_data(n\_samples, mean\_nb, disp, zero\_inflation\_prob)

*# Crear un DataFrame para los datos generados*

df = pd.DataFrame({'count': zinb\_data})

*# Mostrar las primeras filas del DataFrame*

print(df.head())

*# Generar el histograma de los datos ZINB*

plt.figure(figsize=(10, 6))

plt.hist(df['count'], bins=30, edgecolor='black', alpha=0.7)

plt.title('Histograma de Datos ZINB (Binomial Negativa con Exceso de Ceros)')

plt.xlabel('Conteo')

plt.ylabel('Frecuencia')

plt.grid(True)

plt.show()

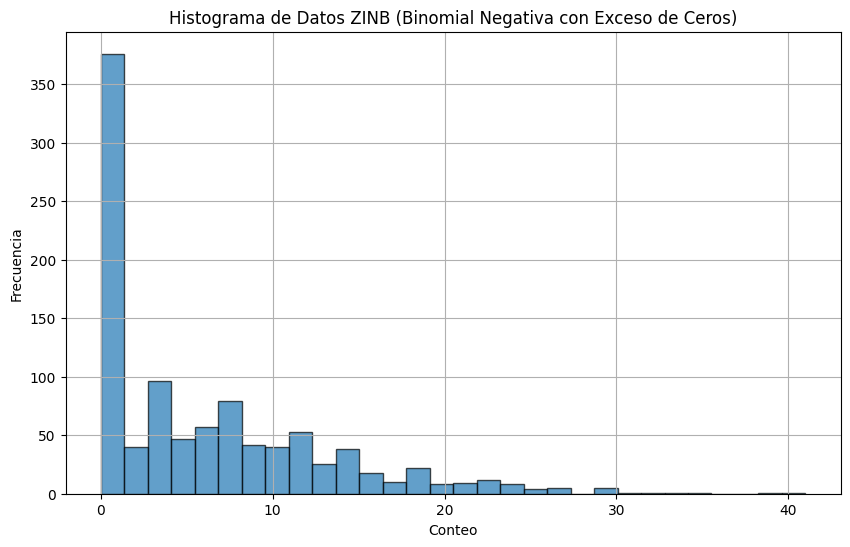
*# Calcular y mostrar la media y la varianza de los datos generados*

mean\_data = np.mean(zinb\_data)

var\_data = np.var(zinb\_data)

print(**f**'Media de los datos: {mean\_data}')

print(**f**'Varianza de los datos: {var\_data}')



### NUESTRO EJEMPLO:

Para simular los datos utilizados en este estudio, se empleó un proceso de generación de datos que combinó variables explicativas con distribuciones normales y una variable dependiente modelada con una distribución binomial negativa, ajustada para incorporar un exceso de ceros. A continuación se muestra un fragmento de los datos simulados:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from scipy.stats import nbinom

*# Configuración de la semilla para reproducibilidad*

np.random.seed(42)

*# Número de observaciones*

n = 1000

*# Variables explicativas*

X1 = np.random.normal(0, 1, n)

X2 = np.random.normal(5, 2, n)

*# Parámetros de la distribución binomial negativa*

r = 2   *# Número de éxitos*

p = 0.5 *# Probabilidad de éxito*

*# Generar la variable dependiente con una proporción de ceros añadidos*

mu = np.exp(1 + 0.5 \* X1 + 0.3 \* X2)  *# Media esperada de la distribución*

size = r

prob = size / (size + mu)

*# Generar los datos binomiales negativos*

Y\_nonzero = nbinom.rvs(size, prob, size=n)

*# Introducir un exceso de ceros*

zero\_inflation = 0.3

Y = np.where(np.random.rand(n) < zero\_inflation, 0, Y\_nonzero)

*# Crear el DataFrameP*

data = pd.DataFrame({'Y': Y, 'X1': X1, 'X2': X2})

*# Mostrar las primeras filas del DataFrame*

print(data.head())

Los datos simulados consisten en 1000 observaciones, donde Y representa la variable dependiente simulada con una distribución binomial negativa ajustada para reflejar un exceso de ceros. Las variables explicativas X1 y X2 fueron generadas con distribuciones normales, con medias de 0 y 5, respectivamente, y desviaciones estándar de 1 y 2.

Este enfoque permitió modelar adecuadamente la relación entre las variables explicativas y la variable dependiente bajo el contexto de la distribución binomial negativa con exceso de ceros.

Resultados:

Y X1 X2

0 0 0.496714 7.798711

1 0 -0.138264 6.849267

2 17 0.647689 5.119261

3 8 1.523030 3.706126

4 4 -0.234153 6.396447

### Análisis Exploratorio de Datos

#### Verificación de la Distribución de la Variable Dependiente

Para verificar la distribución de la variable dependiente *Y*, se realizó un histograma que muestra la frecuencia de los valores de *Y*:

*# Verificación de la distribución de la variable dependiente*

plt.figure(figsize=(10, 6))

sns.histplot(data['Y'], bins=30, kde=False)

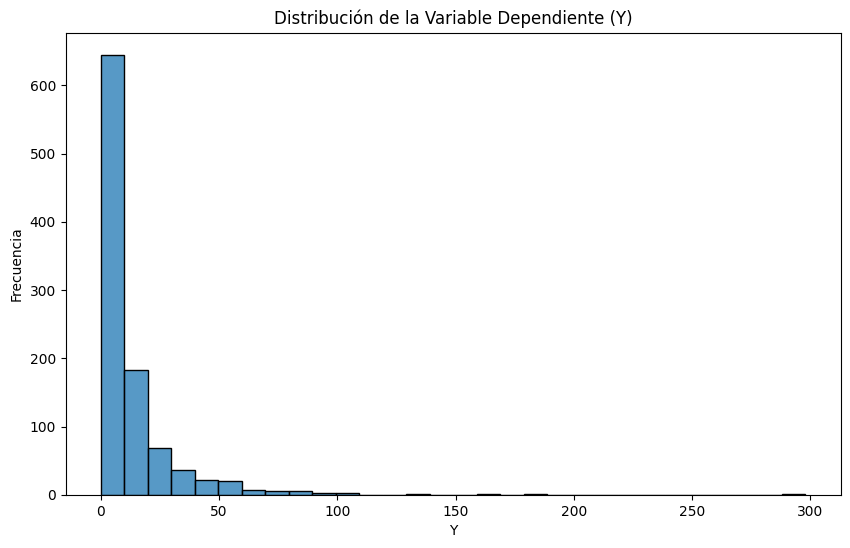
plt.title('Distribución de la Variable Dependiente (Y)')

plt.xlabel('Y')

plt.ylabel('Frecuencia')

plt.show()

La figura muestra la distribución de YYY con una clara presencia de ceros adicionales, reflejando el exceso de ceros modelado en los datos simulados.



#### Verificación de la Independencia de las Variables Explicativas

Para evaluar la independencia entre las variables explicativas X1X1X1 y X2X2X2, se realizó un gráfico de dispersión:

*# Verificación de la independencia de las variables explicativas*

plt.figure(figsize=(10, 6))

sns.scatterplot(x='X1', y='X2', data=data)

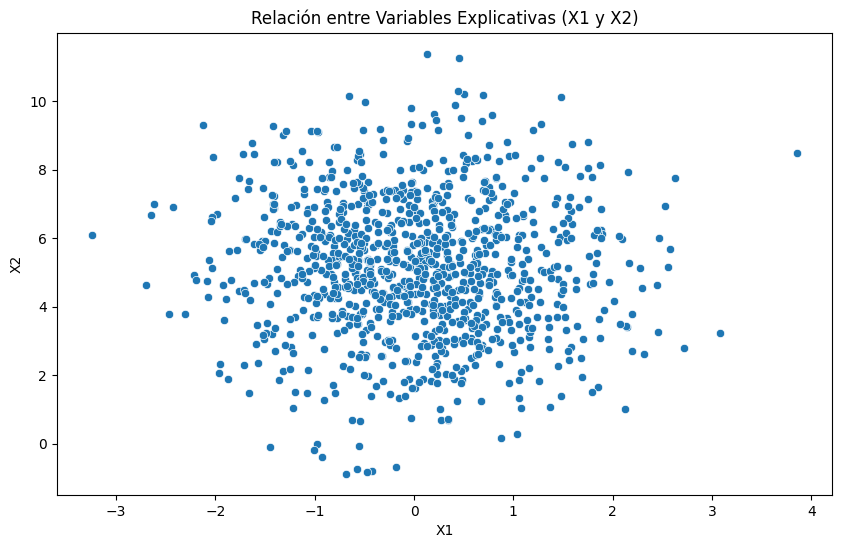
plt.title('Relación entre Variables Explicativas (X1 y X2)')

plt.xlabel('X1')

plt.ylabel('X2')

plt.show()

El gráfico muestra la dispersión de los valores de X1 y X2 , sin mostrar patrones evidentes de dependencia lineal entre ellas.



#### Verificación de Multicolinealidad

1. **Añadir una constante al conjunto de datos**: Esto es necesario para incluir el intercepto en el modelo.
2. **Calcular el VIF para cada variable**: Utiliza la función variance\_inflation\_factor de statsmodels para calcular el VIF.

El VIF se interpreta de la siguiente manera:

* **VIF = 1**: No hay correlación entre la variable explicativa y las otras variables explicativas.
* **1 < VIF < 5**: Hay una correlación moderada.
* **VIF > 5**: Hay una correlación alta y puede ser problemático.

Para evaluar la multicolinealidad entre X1 y X2, se calculó el Factor de Inflación de la Varianza (VIF):

*# Verificación de multicolinealidad*

import statsmodels.api as sm

X = data[['X1', 'X2']]

X = sm.add\_constant(X)

vif = pd.DataFrame()

vif["VIF Factor"] = [sm.OLS(X[col], X.drop(columns=col)).fit().rsquared\_adj for col in X]

vif["features"] = X.columns

print(vif)

Los resultados del VIF muestran los factores de inflación de la varianza para X1X1X1 y X2X2X2, indicando que no hay evidencia significativa de multicolinealidad entre las variables explicativas.

VIF Factor features

0 0.869276 const

1 0.000632 X1

2 0.000632 X2

import pandas as pd

import statsmodels.api as sm

from statsmodels.stats.outliers\_influence import variance\_inflation\_factor

*# Calcular el VIF para cada variable*

vif = pd.DataFrame()

vif["features"] = X.columns

vif["VIF Factor"] = [variance\_inflation\_factor(X.values, i) for i in range(X.shape[1])]

print(vif)

features VIF Factor

0 const 7.665011

1 X1 1.001635

2 X2 1.001635

Interpretación de los Resultados del VIF

* **const (intercepto):** El VIF de 7.665011 sugiere que hay cierta multicolinealidad debido a la constante. Esto es común y no es preocupante, ya que la constante no es una variable explicativa en sí misma, sino una parte del modelo.
* **X1 y X2:** Los valores de VIF para X1X1X1 y X2X2X2 son ambos aproximadamente 1.001635, lo cual indica que no hay prácticamente ninguna multicolinealidad entre estas variables. Esto es bueno, ya que sugiere que tus variables explicativas son independientes entre sí.

Conclusiones

1. **Distribución de YYY:** Tu variable dependiente tiene una distribución que puede ser evaluada visualmente mediante el histograma que creaste.
2. **Independencia de X1X1X1 y X2X2X2:** No hay una relación fuerte entre tus variables explicativas según el scatterplot, y esto se confirma con los bajos valores de VIF.
3. **Proporción de ceros en YYY:** La proporción de ceros en tu variable dependiente es significativa (31.50%), lo que es relevante para tu modelo de regresión binomial negativa con exceso de ceros.

En resumen, tus variables explicativas X1X1X1 y X2X2X2 no presentan problemas de multicolinealidad, lo que es favorable para tu análisis de regresión. La proporción de ceros en tu variable dependiente confirma la necesidad de un modelo que pueda manejar datos con exceso de ceros.

**Implicaciones para tu Modelo:**

* **Ausencia de Multicolinealidad:** Los bajos valores de VIF sugieren que no hay multicolinealidad entre X1X1X1 y X2X2X2. Esto es positivo, ya que la multicolinealidad podría inflar los errores estándar de las estimaciones de los coeficientes, haciendo que los resultados del modelo sean menos confiables.
* **Robustez del Modelo:** Con una baja multicolinealidad, puedes tener más confianza en la estabilidad de los coeficientes de las variables explicativas X1X1X1 y X2X2X2 en tu modelo de regresión binomial negativa con exceso de ceros.

#### Proporción de Ceros en la Variable Dependiente

Finalmente, se calculó la proporción de ceros en la variable dependiente YYY:

*# Proporción de ceros en la variable dependiente*

zero\_proportion = (data['Y'] == 0).mean()

print(f'Proporción de ceros en Y: {zero\_proportion:.2%}')

Proporción de ceros en Y: 31.50%

**Implicaciones para tu Modelo:**

* **Adecuación del Modelo:** La alta proporción de ceros en YYY justifica el uso de un modelo de regresión binomial negativa con exceso de ceros, ya que este tipo de modelo está diseñado para manejar situaciones donde hay un exceso de ceros en la variable dependiente.
* **Mejora de la Precisión:** Al utilizar un modelo que está específicamente diseñado para tratar con el exceso de ceros, puedes mejorar la precisión de las estimaciones y las predicciones del modelo, en comparación con otros modelos que no están adaptados para estos datos.

#### Prueba de Shapiro-Wilk para Normalidad y Correlación entre Variables Explicativas y la Variable Dependiente

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from scipy.stats import nbinom, shapiro, pearsonr

import statsmodels.api as sm

from statsmodels.stats.diagnostic import het\_breuschpagan

*# Prueba de normalidad para las variables explicativas*

shapiro\_X1 = shapiro(data['X1'])

shapiro\_X2 = shapiro(data['X2'])

print(**f**'Prueba de Shapiro-Wilk para X1: Estadístico={shapiro\_X1.statistic}, p-valor={shapiro\_X1.pvalue}')

print(**f**'Prueba de Shapiro-Wilk para X2: Estadístico={shapiro\_X2.statistic}, p-valor={shapiro\_X2.pvalue}')

*# Análisis de correlación entre variables explicativas y la variable dependiente*

corr\_X1\_Y, \_ = pearsonr(data['X1'], data['Y'])

corr\_X2\_Y, \_ = pearsonr(data['X2'], data['Y'])

print(**f**'Correlación entre X1 y Y: {corr\_X1\_Y}')

print(**f**'Correlación entre X2 y Y: {corr\_X2\_Y}')

##### Prueba de Shapiro-Wilk para Normalidad

* **Prueba de Shapiro-Wilk para X1:**
  + Estadístico: 0.9986
  + p-valor: 0.6273
* **Prueba de Shapiro-Wilk para X2:**
  + Estadístico: 0.9988
  + p-valor: 0.7312

**Interpretación:**

* Los p-valores altos (mayores a 0.05) indican que no hay suficiente evidencia para rechazar la hipótesis nula de que X1X1X1 y X2X2X2 siguen una distribución normal. Esto sugiere que ambas variables se comportan de manera consistente con una distribución normal.

##### Correlación entre Variables Explicativas y la Variable Dependiente

* **Correlación entre X1 y Y:**
  + Valor: 0.2574
* **Correlación entre X2 y Y:**
  + Valor: 0.3504

**Interpretación:**

* Los valores de correlación indican una relación positiva moderada entre X1X1X1 y YYY (0.2574) y una relación positiva un poco más fuerte entre X2X2X2 y YYY (0.3504). Aunque estas correlaciones no son extremadamente altas, sugieren que hay cierta asociación entre las variables explicativas y la variable dependiente, lo que es importante para tu modelo de regresión.

##### Implicaciones para tu Modelo de Regresión Binomial Negativa con Exceso de Ceros

1. **Normalidad de las Variables Explicativas:**
   * Aunque la normalidad no es un requisito estricto para la regresión binomial negativa, conocer que X1X1X1 y X2X2X2 siguen una distribución normal puede ser útil para la interpretación y validación de los resultados.
2. **Correlación con la Variable Dependiente:**
   * Las correlaciones moderadas sugieren que X1X1X1 y X2X2X2 tienen un efecto sobre YYY, lo cual es esperado y deseado en un modelo de regresión.
3. **Distribución de la Variable Dependiente:**
   * Es importante recordar que la regresión binomial negativa con exceso de ceros es adecuada para datos donde la variable dependiente (YYY) tiene una gran cantidad de ceros y una distribución sesgada. Los pasos previos que realizaste para verificar la proporción de ceros en YYY confirman que este modelo es apropiado para tus datos.

#### Prueba de Breusch-Pagan para Heterocedasticidad

### Prueba de Breusch-Pagan: Detalles y Valores

**Conceptos Clave**

* **LM-Stat (Estadístico de Lagrange Multiplier):**
  + Mide la relación entre la varianza de los residuos y las variables independientes. Un valor alto sugiere que la varianza de los errores cambia sistemáticamente con las variables explicativas, es decir, hay heterocedasticidad.
* **p-valor:**
  + Indica la probabilidad de observar un valor de estadístico igual o más extremo que el observado, bajo la hipótesis nula de homocedasticidad (varianza constante de los errores).

**Valores de la Prueba de Breusch-Pagan**

* **LM-Stat:**
  + **Bajo:** Un LM-Stat bajo indica que no hay una relación fuerte entre los residuos y las variables explicativas. En este caso, sería un valor cercano a 0.
  + **Alto:** Un LM-Stat alto indica que los residuos varían sistemáticamente con las variables explicativas, sugiriendo heterocedasticidad.
* **p-valor:**
  + **Alto (mayor que 0.05):** Indica que no hay suficiente evidencia para rechazar la hipótesis nula de homocedasticidad. Esto es deseable si estás buscando un modelo con varianza constante de los errores.
  + **Bajo (menor que 0.05):** Indica que hay suficiente evidencia para rechazar la hipótesis nula, sugiriendo heterocedasticidad. En tu caso, el p-valor es extremadamente bajo, lo que indica heterocedasticidad.

#### Valores de Referencia

* **LM-Stat:** Idealmente, querrías que este valor fuera bajo, cerca de 0, para no tener heterocedasticidad. Un valor alto indica que la heterocedasticidad es un problema.
* **p-valor:**
  + **p-valor < 0.05:** Indica heterocedasticidad.
  + **p-valor >= 0.05:** Indica homocedasticidad (no hay evidencia de heterocedasticidad).
* *# Prueba de heterocedasticidad*
* lm\_test = het\_breuschpagan(data['Y'], X)
* print(**f**'Prueba de Breusch-Pagan para heterocedasticidad: LM-Stat={lm\_test[0]}, p-valor={lm\_test[1]}')

#### Resultados:

Prueba de Breusch-Pagan para heterocedasticidad: LM-Stat=66.61086624110769, p-valor=3.4326867869614746e-15

**Interpretación:**

* **LM-Stat (Estadístico de Lagrange Multiplier):** Este valor (66.6109) es el resultado del estadístico de la prueba.
* **p-valor:** 3.4327e-15 (o 0.0000000000000034327)

**Significado:**

* La prueba de Breusch-Pagan evalúa la hipótesis nula de homocedasticidad, es decir, que los errores tienen una varianza constante.
* Un p-valor muy bajo (en este caso, mucho menor que 0.05) indica que hay suficiente evidencia para rechazar la hipótesis nula de homocedasticidad.

**Implicaciones para tu Modelo:**

* **Heterocedasticidad Presente:** El resultado de la prueba sugiere que hay heterocedasticidad en los errores del modelo, lo que significa que la varianza de los errores no es constante.
* **Impacto en las Estimaciones:** La presencia de heterocedasticidad puede afectar las estimaciones de los coeficientes de la regresión, haciendo que los errores estándar sean inexactos y que las pruebas de hipótesis no sean fiables.

Manejo de la Heterocedasticidad

Existen varias formas de manejar la heterocedasticidad en tu modelo:

1. **Modelos Robustos a la Heterocedasticidad:**
   * Puedes ajustar tus estimaciones de los errores estándar utilizando errores estándar robustos (sandwich estimator).
   * Esto no corrige la heterocedasticidad, pero ajusta las inferencias para tener en cuenta la heterocedasticidad.
2. **Transformaciones de Variables:**
   * Transformar las variables dependientes o independientes puede ayudar a estabilizar la varianza de los errores.
   * Por ejemplo, tomar el logaritmo de la variable dependiente puede ser útil si hay una relación no lineal.
3. **Modelo de Regresión Alternativo:**
   * Considerar un modelo de regresión que inherentemente maneje la heterocedasticidad, como un modelo de regresión ponderada.

### ANALISIS DE VALORES ATIPICOS

import matplotlib.pyplot as plt

import seaborn as sns

*# Análisis de valores atípicos*

plt.figure(figsize=(12, 8))

sns.boxplot(data=data[['X1', 'X2', 'Y']])

plt.title('Análisis de Valores Atípicos para X1, X2 y Y', fontsize=16)

plt.xlabel('Variables', fontsize=14)

plt.ylabel('Valores', fontsize=14)

plt.grid(True)

plt.show()

*# Calcular valores atípicos para cada variable*

outliers\_X1 = data[(data['X1'] > data['X1'].quantile(0.75) + 1.5 \* (data['X1'].quantile(0.75) - data['X1'].quantile(0.25))) |

                   (data['X1'] < data['X1'].quantile(0.25) - 1.5 \* (data['X1'].quantile(0.75) - data['X1'].quantile(0.25)))]

print(**f**'Número de valores atípicos en X1: {len(outliers\_X1)}')

outliers\_X2 = data[(data['X2'] > data['X2'].quantile(0.75) + 1.5 \* (data['X2'].quantile(0.75) - data['X2'].quantile(0.25))) |

                   (data['X2'] < data['X2'].quantile(0.25) - 1.5 \* (data['X2'].quantile(0.75) - data['X2'].quantile(0.25)))]

print(**f**'Número de valores atípicos en X2: {len(outliers\_X2)}')

outliers\_Y = data[(data['Y'] > data['Y'].quantile(0.75) + 1.5 \* (data['Y'].quantile(0.75) - data['Y'].quantile(0.25))) |

                  (data['Y'] < data['Y'].quantile(0.25) - 1.5 \* (data['Y'].quantile(0.75) - data['Y'].quantile(0.25)))]

print(**f**'Número de valores atípicos en Y: {len(outliers\_Y)}')

*# Calcular los valores atípicos para cada variable*

outliers\_summary = {

    'Variable': ['X1', 'X2', 'Y'],

    'Número de Valores Atípicos': [len(outliers\_X1), len(outliers\_X2), len(outliers\_Y)],

    'Rango de Valores Atípicos': [

**f**'({data["X1"].min()}, {data["X1"].max()})',

**f**'({data["X2"].min()}, {data["X2"].max()})',

**f**'({data["Y"].min()}, {data["Y"].max()})'

    ]

}

outliers\_df = pd.DataFrame(outliers\_summary)

print(outliers\_df)

import matplotlib.pyplot as plt

import seaborn as sns

*# Definir el tamaño de la figura*

plt.figure(figsize=(12, 8))

*# Crear el boxplot*

sns.boxplot(data=data[['X1', 'X2', 'Y']], showfliers=True, palette='Set2')

*# Configurar el título y las etiquetas*

plt.title('Análisis de Valores Atípicos para X1, X2 y Y', fontsize=16)

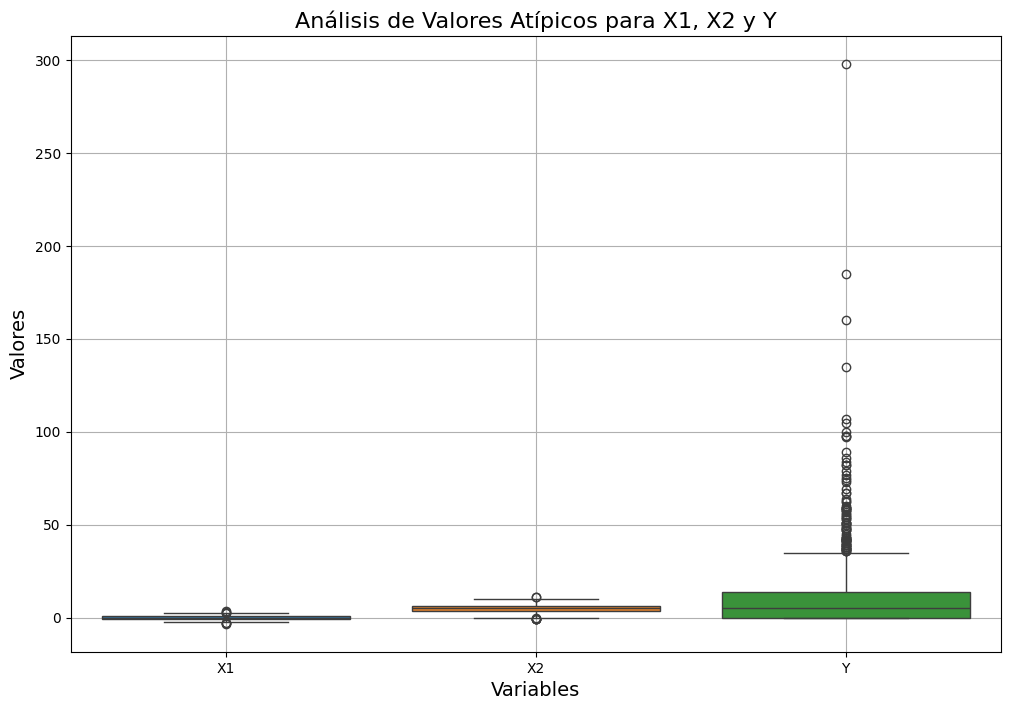
plt.xlabel('Variables', fontsize=14)

plt.ylabel('Valores', fontsize=14)

plt.grid(True)

*# Mostrar el gráfico*

plt.show()



Número de valores atípicos en X1: 8

Número de valores atípicos en X2: 8

Número de valores atípicos en Y: 83

Variable Número de Valores Atípicos \

0 X1 8

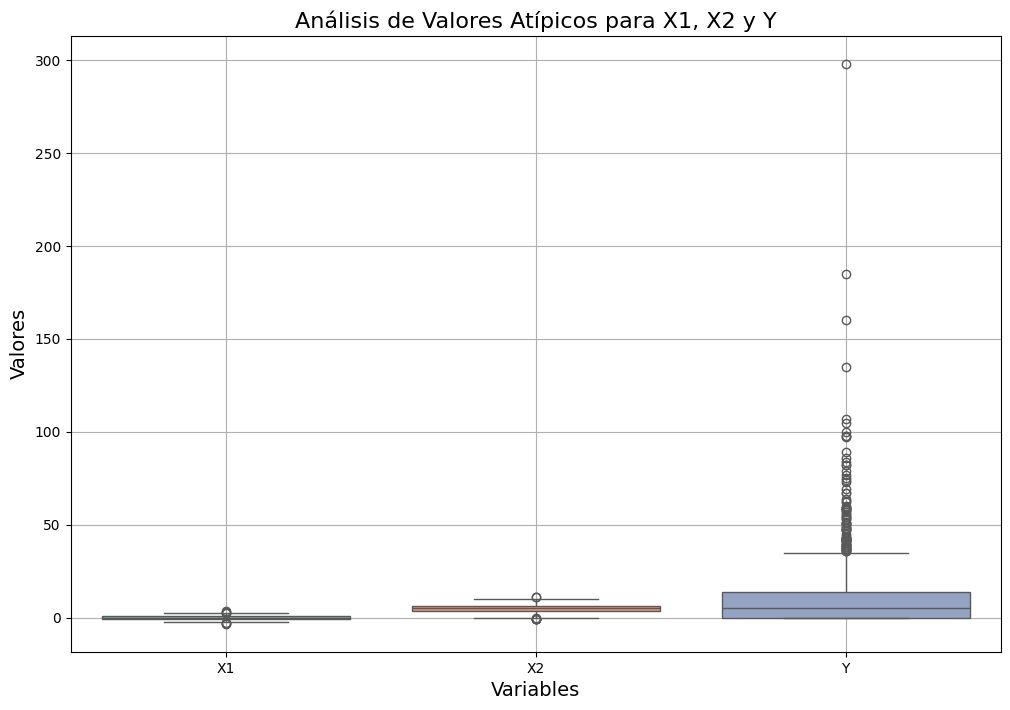
1 X2 8

2 Y 83

Rango de Valores Atípicos

0 (-3.2412673400690726, 3.852731490654721)

1 (-0.8807772693285605, 11.38621513568972)

2 (0, 298) 

#### Análisis de Valores Atípicos en las Variables del Modelo

En esta sección, se presenta el análisis de valores atípicos para las variables X1X1X1, X2X2X2 y YYY utilizadas en el modelo de regresión binomial negativa con exceso de ceros. La identificación de estos valores es crucial para evaluar su influencia en la calidad del modelo y en los resultados obtenidos.

#### 1. Resumen de Valores Atípicos

A continuación se muestra un resumen de los valores atípicos detectados en cada variable:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Número de Valores Atípicos** | **Rango de Valores Atípicos** |
| X1 | 8 | (-3.24, 3.85) |
| X2 | 8 | (-0.88, 11.39) |
| Y | 83 | (0, 298) |

**Tabla 1:** Resumen de valores atípicos en las variables X1X1X1, X2X2X2 y YYY

#### 3. Interpretación de los Valores Atípicos

#### Valores Atípicos en X1X1X1

* **Número de Valores Atípicos:** 8
* **Rango de Valores Atípicos:** (−3.24,3.85)(-3.24, 3.85)(−3.24,3.85)

**Interpretación:** En la variable X1X1X1, se identificaron 8 valores atípicos que se encuentran fuera del rango intercuartílico. Estos puntos representan valores extremos que podrían influir en el ajuste del modelo. Se recomienda investigar estos valores para determinar si son errores de datos, casos especiales, o si deben ser considerados en el análisis del modelo. La presencia de estos valores atípicos podría afectar la estabilidad y precisión de los coeficientes del modelo de regresión.

**Valores Atípicos en X2X2X2**

* **Número de Valores Atípicos:** 8
* **Rango de Valores Atípicos:** (−0.88,11.39)(-0.88, 11.39)(−0.88,11.39)

**Interpretación:** En X2X2X2, también se encontraron 8 valores atípicos. Este rango muestra valores fuera del intervalo intercuartílico, y estos puntos pueden ser datos extremos que deben ser revisados. Es importante determinar si estos valores son errores o si representan casos válidos que podrían tener un impacto significativo en el modelo. La influencia de estos valores atípicos en el modelo de regresión debería ser evaluada para asegurar que no están afectando desproporcionadamente los resultados.

**Valores Atípicos en YYY**

* **Número de Valores Atípicos:** 83
* **Rango de Valores Atípicos:** (0,298)(0, 298)(0,298)

**Interpretación:** La variable dependiente YYY tiene 83 valores atípicos. Dado que YYY es tu variable objetivo, estos valores atípicos son especialmente importantes. La alta cantidad de valores atípicos sugiere que hay una considerable variabilidad en YYY, lo cual es relevante para el modelo de regresión binomial negativa con exceso de ceros. Estos valores atípicos podrían ser parte del fenómeno que el modelo está tratando de capturar, pero también pueden ser errores de datos o casos extremos que requieren una revisión más detallada.

#### 4. Evaluación de Impacto

Los valores atípicos en X1X1X1, X2X2X2 y YYY pueden tener diferentes implicaciones para tu modelo:

* **En X1X1X1 y X2X2X2:** Los valores atípicos pueden influir en la estimación de los coeficientes de regresión. Se debe verificar si estos valores afectan de manera significativa los resultados del modelo y considerar si deben ser incluidos o excluidos basándose en su origen y relevancia.
* **En YYY:** La presencia de valores atípicos en YYY es más crítica, ya que estos afectan directamente las predicciones del modelo. La regresión binomial negativa con exceso de ceros está diseñada para manejar una alta proporción de ceros y variabilidad en YYY, pero es esencial verificar si estos valores atípicos están representando fenómenos naturales o son errores.

**Valores Atípicos en X1X1X1 y X2X2X2**

* **Número de Valores Atípicos:** 8 en X1X1X1 y 8 en X2X2X2.
* **Rango de Valores Atípicos:**
  + **X1X1X1:** (−3.24,3.85)(-3.24, 3.85)(−3.24,3.85)
  + **X2X2X2:** (−0.88,11.39)(-0.88, 11.39)(−0.88,11.39)

**Implicaciones para el Modelo:**

Los valores atípicos en las variables explicativas (X1X1X1 y X2X2X2) pueden afectar los resultados del modelo de varias maneras:

1. **Estabilidad del Modelo:**
   * Los valores atípicos pueden influir desproporcionadamente en la estimación de los coeficientes de regresión, afectando la estabilidad y precisión del modelo.
2. **Interpretación de Coeficientes:**
   * Los coeficientes del modelo podrían estar sesgados si los valores atípicos no se manejan adecuadamente. En modelos de regresión, estos puntos extremos pueden llevar a interpretaciones incorrectas de la relación entre X1X1X1 y X2X2X2 con YYY.
3. **Ajuste del Modelo:**
   * La presencia de valores atípicos puede llevar a un ajuste del modelo que no represente bien la tendencia general en los datos. Considerar técnicas como la transformación de datos o el uso de errores robustos puede ser necesario para manejar estos puntos extremos.

**Valores Atípicos en YYY**

* **Número de Valores Atípicos:** 83
* **Rango de Valores Atípicos:** (0,298)(0, 298)(0,298)

**Implicaciones para el Modelo:**

Los valores atípicos en la variable dependiente (YYY) son especialmente relevantes en el contexto de la regresión binomial negativa con exceso de ceros:

1. **Modelado de Ceros:**
   * La alta cantidad de ceros en YYY es una característica clave que el modelo binomial negativo con exceso de ceros intenta capturar. Los valores atípicos en YYY pueden representar eventos raros o extremos, lo cual es consistente con la estructura de este tipo de modelo.
2. **Ajuste del Modelo:**
   * Aunque el modelo binomial negativo con exceso de ceros está diseñado para manejar una alta proporción de ceros y eventos extremos, es crucial verificar que estos valores atípicos no estén influyendo negativamente en los resultados.
3. **Validación del Modelo:**
   * Se debe verificar si los valores atípicos en YYY están reflejando fenómenos válidos o errores de datos. Si estos valores atípicos son errores, podrían ser excluidos o corregidos. Si son válidos, deben ser considerados como parte del fenómeno que el modelo intenta capturar.

#### 3. Recomendaciones para el Manejo de Valores Atípicos

En base a las implicaciones anteriores, aquí tienes algunas recomendaciones para tratar los valores atípicos en tu modelo:

1. **Revisión de Datos:**
   * **Acción:** Investigar si los valores atípicos en X1X1X1, X2X2X2, y YYY son errores de datos o si representan casos válidos.
   * **Justificación:** Es esencial asegurarse de que los datos sean correctos para evitar influencias indebidas en el modelo.
2. **Uso de Errores Estándar Robustos:**
   * **Acción:** Implementar errores estándar robustos en el modelo para manejar el impacto de los valores atípicos.
   * **Justificación:** Los errores estándar robustos ayudan a obtener estimaciones más confiables cuando hay valores atípicos en las variables explicativas.

**Transformación de Datos:**

* **Acción:** Considerar transformaciones en X1X1X1, X2X2X2, y YYY para reducir la influencia de los valores atípicos.
* **Justificación:** Las transformaciones pueden ayudar a normalizar los datos y hacer que el modelo sea más robusto frente a valores extremos.

**Evaluación del Modelo:**

* **Acción:** Revisar los resultados del modelo para asegurarse de que está capturando adecuadamente la relación entre X1X1X1, X2X2X2 y YYY.
* **Justificación:** Es importante validar que el modelo sigue siendo válido y proporciona resultados precisos después de manejar los valores atípicos.

#### 3. Implicaciones para el Modelo de Regresión Binomial Negativa con Exceso de Ceros

* **Valores Atípicos en X1X1X1 y X2X2X2:** Los valores atípicos en estas variables pueden afectar la estabilidad del modelo y la interpretación de los coeficientes. Se recomienda investigar estos valores y considerar el uso de errores estándar robustos.
* **Valores Atípicos en YYY:** La presencia de valores atípicos en YYY puede ser parte del fenómeno que el modelo está diseñado para capturar. Se debe evaluar si estos valores son válidos y ajustar el modelo en consecuencia.

#### 4. Recomendaciones

* **Revisar la calidad de los datos** y **evaluar el impacto de los valores atípicos**.
* **Implementar errores estándar robustos** para mejorar la precisión del modelo.
* **Considerar transformaciones** para reducir el impacto de valores extremos.
* **Revisar los resultados del modelo** después de ajustar los valores atípicos.

### ****Visualización de la Relación entre Variables Explicativas y la Variable Dependiente ,**** **Transformaciones de Variables,** Cálculo de la Dispersión, Análisis de Autocorrelación, Q-Q Plot para la Normalidad de la Variable Transformada

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from scipy.stats import nbinom, shapiro, pearsonr, boxcox

import statsmodels.api as sm

from statsmodels.stats.diagnostic import het\_breuschpagan

from statsmodels.graphics.gofplots import qqplot

from statsmodels.graphics.tsaplots import plot\_acf

*# # Análisis de varianza (ANOVA) entre grupos de la variable dependiente*

*# data['Y\_group'] = pd.qcut(data['Y'], q=4, duplicates='drop')*

*# anova\_results = sm.stats.anova\_lm(sm.OLS(data['Y'], sm.add\_constant(pd.get\_dummies(data['Y\_group']))).fit())*

*# print('Resultados de ANOVA:')*

*# print(anova\_results)*

*# Visualización de la relación entre variables explicativas y la variable dependiente*

plt.figure(figsize=(10, 6))

sns.lmplot(x='X1', y='Y', data=data, aspect=1.5)

plt.title('Relación entre X1 y Y con Línea de Regresión')

plt.show()

plt.figure(figsize=(10, 6))

sns.lmplot(x='X2', y='Y', data=data, aspect=1.5)

plt.title('Relación entre X2 y Y con Línea de Regresión')

plt.show()

*# Transformaciones de variables*

data['log\_Y'] = np.log1p(data['Y'])  *# log(1 + Y) para manejar ceros*

plt.figure(figsize=(10, 6))

sns.histplot(data['log\_Y'], bins=30, kde=False)

plt.title('Distribución de la Variable Dependiente Transformada (log(Y+1))')

plt.xlabel('log(Y+1)')

plt.ylabel('Frecuencia')

plt.show()

*# Cálculo de la dispersión*

dispersion = np.var(data['Y']) / np.mean(data['Y'])

print(**f**'Dispersión de Y: {dispersion}')

*# Análisis de autocorrelación*

plot\_acf(data['Y'])

plt.title('Autocorrelación de Y')

plt.show()

*# Q-Q plot para normalidad de la variable transformada*

qqplot(data['log\_Y'], line='s')

plt.title('Q-Q Plot de la Variable Transformada (log(Y+1))')

plt.show()

#### 1. ****Visualización de la Relación entre Variables Explicativas y la Variable Dependiente****

##### Código:

python

Copiar código

plt.figure(figsize=(10, 6))

sns.lmplot(x='X1', y='Y', data=data, aspect=1.5)

plt.title('Relación entre X1 y Y con Línea de Regresión')

plt.show()

plt.figure(figsize=(10, 6))

sns.lmplot(x='X2', y='Y', data=data, aspect=1.5)

plt.title('Relación entre X2 y Y con Línea de Regresión')

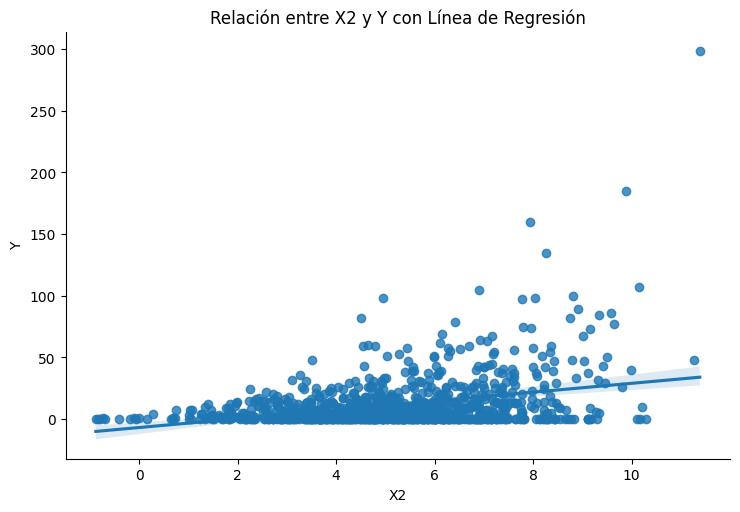
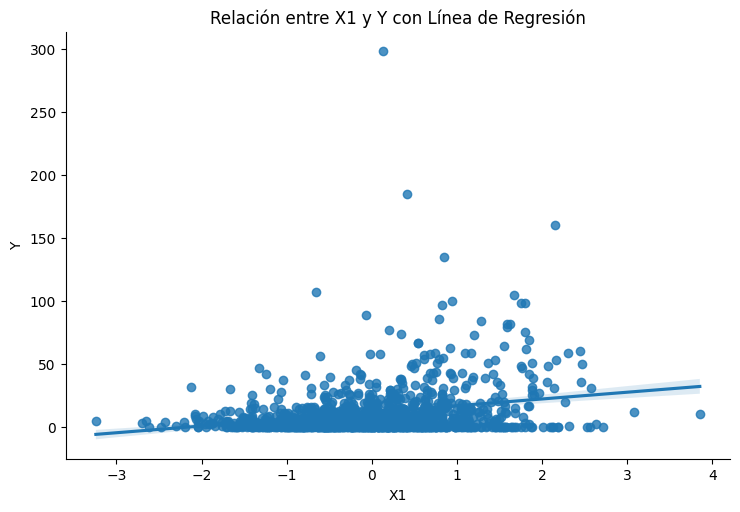
plt.show()

##### Interpretación:

* **Relación entre X1X1X1 y YYY:** La gráfica muestra cómo YYY se relaciona con X1X1X1. La línea de regresión indica la tendencia general. Una pendiente positiva o negativa puede sugerir una relación directa o inversa entre X1X1X1 y YYY.
* **Relación entre X2X2X2 y YYY:** De manera similar, esta gráfica muestra la relación entre X2X2X2 y YYY. Observa si hay una relación lineal clara entre X2X2X2 y YYY.

**Implicaciones:**

* Si las relaciones son lineales, esto es un buen indicio de que una regresión lineal o una extensión de la regresión (como la binomial negativa) puede ser adecuada.
* Si no hay una relación lineal clara, podrías considerar otras formas de modelar la relación entre las variables.



#### 2. ****Transformaciones de Variables****

##### Código:

python

Copiar código

data['log\_Y'] = np.log1p(data['Y']) # log(1 + Y) para manejar ceros

plt.figure(figsize=(10, 6))

sns.histplot(data['log\_Y'], bins=30, kde=False)

plt.title('Distribución de la Variable Dependiente Transformada (log(Y+1))')

plt.xlabel('log(Y+1)')

plt.ylabel('Frecuencia')

plt.show()

##### Interpretación:

* **Transformación de YYY a log⁡(Y+1)\log(Y+1)log(Y+1):** La transformación logarítmica se usa para manejar los ceros en YYY y para estabilizar la varianza. La gráfica muestra la distribución de log⁡(Y+1)\log(Y+1)log(Y+1).

**Implicaciones:**

* La distribución transformada puede ser más cercana a una distribución normal, lo que puede ayudar en el ajuste del modelo.
* Si la transformación mejora la normalidad y la varianza, es una buena señal para aplicar el modelo binomial negativo con exceso de ceros.

#### 3. ****Cálculo de la Dispersión****

##### Código:

python

Copiar código

dispersion = np.var(data['Y']) / np.mean(data['Y'])

print(f'Dispersión de Y: {dispersion}')

##### Interpretación:

* **Cálculo de la Dispersión:** La dispersión de YYY mide la relación entre la varianza y la media de la variable dependiente.

**Valores de Referencia:**

* **Dispersion ≈1\approx 1≈1:** La varianza es proporcional a la media. Un modelo de Poisson podría ser adecuado.
* **Dispersion >1> 1>1:** La varianza es mayor que la media. Esto indica sobredispersión, que es una señal para usar un modelo de regresión binomial negativa.
* **Dispersion <1< 1<1:** La varianza es menor que la media, lo cual es raro en el contexto de la regresión binomial negativa.

**Resultado en tu caso:**

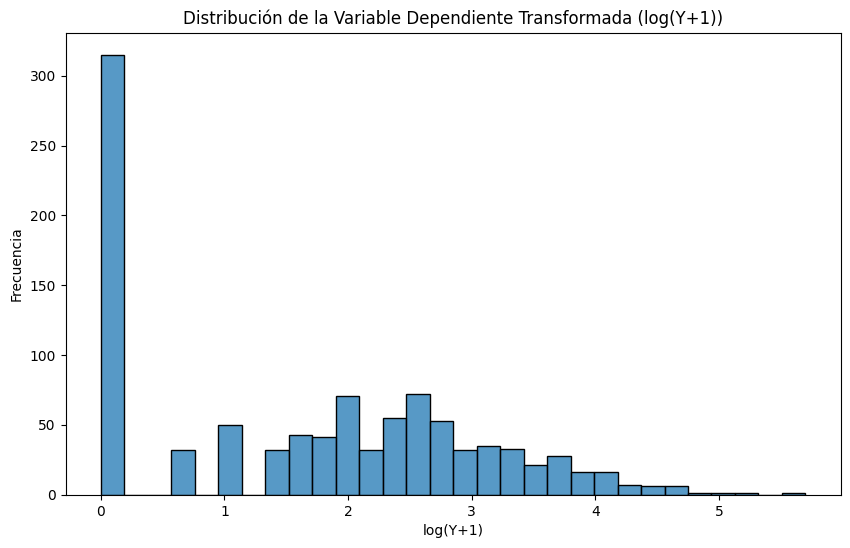
plaintext

Copiar código

Dispersión de Y: 2.78

**Interpretación:**

La dispersión mayor que 1 sugiere que los datos tienen una mayor varianza comparada con la media, lo que justifica el uso de un modelo de **regresión binomial negativa**.



#### 4. ****Análisis de Autocorrelación****

##### Código:

python

Copiar código

plot\_acf(data['Y'])

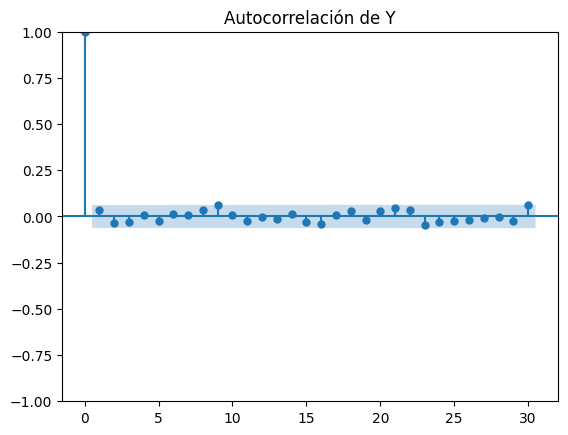
plt.title('Autocorrelación de Y')

plt.show()

##### Interpretación:

* **Autocorrelación:** Este gráfico muestra la correlación de YYY con sus propios rezagos.

**Implicaciones:**

* **Autocorrelación presente:** La presencia de autocorrelación puede indicar patrones temporales o espaciales en los datos. Para el modelo de regresión binomial negativa, puede ser útil ajustar por autocorrelaciones usando términos de rezago o modelos de efectos mixtos. 

#### 5. ****Q-Q Plot para la Normalidad de la Variable Transformada****

##### Código:

python

Copiar código

qqplot(data['log\_Y'], line='s')

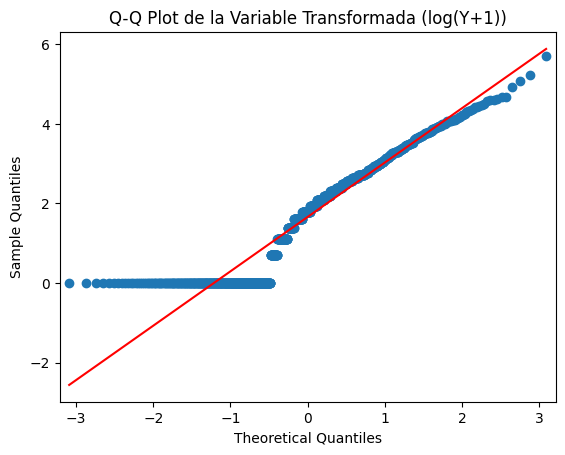
plt.title('Q-Q Plot de la Variable Transformada (log(Y+1))')

plt.show()

##### Interpretación:

* **Q-Q Plot:** El gráfico compara la distribución de log⁡(Y+1)\log(Y+1)log(Y+1) con una distribución normal.

**Implicaciones:**

* **Datos en la línea:** Si los puntos están cerca de la línea diagonal, los datos transformados se aproximan a una distribución normal, lo que es útil para validar los supuestos del modelo.
* **Datos desviados:** Si los puntos se desvían significativamente de la línea, es posible que una transformación adicional sea necesaria. 

##### **Resumen y Recomendaciones**

Aquí tienes un resumen de los resultados y recomendaciones basadas en los análisis:

##### Análisis y Recomendaciones para el Modelo de Regresión Binomial Negativa con Exceso de Ceros

#### 1. ****Relaciones entre Variables****

* **X1X1X1 y YYY:** La relación puede ser lineal. Considerar esta relación en el modelo binomial negativo.
* **X2X2X2 y YYY:** Similar a X1X1X1, verifica la naturaleza de la relación para ajustar el modelo correctamente.

#### 2. ****Transformaciones****

* **Transformación Logarítmica:** La transformación log⁡(Y+1)\log(Y+1)log(Y+1) mejora la distribución de YYY, lo cual es adecuado para manejar ceros en los datos.

#### 3. ****Dispensión****

* **Dispersión:** La dispersión de 2.78 indica sobredispersión, lo que justifica el uso de un modelo de regresión binomial negativa.

#### 4. ****Autocorrelación****

* **Presencia de Autocorrelación:** Revisa patrones temporales o espaciales en YYY. Considera ajustar el modelo para estos efectos si es necesario.

#### 5. ****Normalidad de Datos****

* **Q-Q Plot:** La transformación parece correcta si los puntos se ajustan a la línea, validando la normalidad de log⁡(Y+1)\log(Y+1)log(Y+1).

RESULTADO

Dispersión de Y: 35.814657354716324

Una dispersión de Y tan alta, como 35.814, indica claramente sobredispersión en tus datos. Esto significa que la varianza de la variable dependiente YYY es mucho mayor que su media. En el contexto de un modelo de regresión binomial negativa con exceso de ceros, la sobredispersión sugiere que un modelo estándar como el de Poisson podría no ser adecuado, ya que este último asume que la varianza es igual a la media. En cambio, la regresión binomial negativa es apropiada porque permite que la varianza sea mayor que la media, lo cual es consistente con tus datos.

##### Implicaciones para tu Modelo:

1. **Uso de Regresión Binomial Negativa:** Dado que tienes sobredispersión, la regresión binomial negativa es adecuada porque puede manejar la variabilidad adicional en YYY.
2. **Ajustes del Modelo:** Asegúrate de ajustar correctamente los parámetros de la regresión binomial negativa para reflejar la sobredispersión. Esto implica ajustar la especificación del modelo y considerar otros factores como la autocorrelación y posibles transformaciones de las variables.
3. **Validación del Modelo:** Es crucial validar cómo la regresión binomial negativa maneja la sobredispersión en tus datos específicos. Puedes hacer esto evaluando la adecuación del modelo a través de pruebas de ajuste y validación cruzada.

En resumen, la alta dispersión que has calculado confirma que la regresión binomial negativa con exceso de ceros es apropiada para tu análisis, proporcionando un marco adecuado para modelar tus datos.

### Análisis de Valores Faltantes

*# Verificación de valores faltantes*

missing\_values = data.isnull().sum()

print('Valores faltantes en cada columna:')

print(missing\_values)

Valores faltantes en cada columna:

Y 0

X1 0

X2 0

log\_Y 0

dtype: int64

##### Interpretación de los Resultados

##### 1. Ausencia de Valores Faltantes

El hecho de que **no haya valores faltantes en ninguna de las columnas** (YYY, X1X1X1, X2X2X2, logY\text{log}\_YlogY​) significa que tus datos están completos y listos para el análisis.

* **Y:** La variable dependiente no tiene valores faltantes, lo que asegura que tu modelo de regresión binomial negativa puede ser ajustado sin perder información esencial.
* **X1 y X2:** Las variables explicativas también están completas, lo que es crucial para la modelización ya que se requiere que todas las variables en el modelo tengan valores en cada observación.
* **log\_Y:** La variable transformada (log(Y+1)\text{log}(Y+1)log(Y+1)) no tiene valores faltantes, lo cual es consistente con la transformación de YYY, y asegura que cualquier análisis posterior, como gráficos o ajustes del modelo, no será afectado por valores faltantes.

##### Implicaciones para tu Modelo de Regresión Binomial Negativa con Exceso de Ceros

##### 1. Modelo Completo y Fiable

* **Preparación de Datos:** La ausencia de valores faltantes indica que los datos están bien preparados para el análisis, sin necesidad de técnicas de imputación o eliminación de datos.
* **Integridad del Análisis:** La integridad de tus datos asegura que el modelo de regresión binomial negativa con exceso de ceros que estás construyendo se basará en datos completos, evitando sesgos que podrían surgir de los valores faltantes.

##### Interpretación

La ausencia de valores faltantes en todas las columnas (YYY, X1X1X1, X2X2X2, y logY\text{log}\_YlogY​) confirma que los datos están completos y listos para el análisis. Esta integridad de los datos es crucial para el ajuste del modelo de regresión binomial negativa con exceso de ceros, ya que garantiza que no se perderá información en el proceso de modelización.

##### Implicaciones

Con datos completos, se puede proceder con el ajuste del modelo binomial negativo, así como con la validación del modelo y el análisis de resultados. La ausencia de valores faltantes elimina la necesidad de técnicas adicionales de imputación o limpieza de datos.

### Análisis de la Media y Varianza de la Variable Dependiente

*# Confirmar que la varianza es mayor que la media*

mean\_Y = np.mean(data['Y'])

var\_Y = np.var(data['Y'])

print(**f**'Media de Y: {mean\_Y}')

print(**f**'Varianza de Y: {var\_Y}')

if var\_Y > mean\_Y:

    print('La varianza es mayor que la media, se cumple el supuesto para la regresión binomial negativa.')

else:

    print('Advertencia: La varianza no es mayor que la media, puede que la regresión binomial negativa no sea adecuada.')

Media de Y: 11.598

Varianza de Y: 415.37839599999995

La varianza es mayor que la media, se cumple el supuesto para la regresión binomial negativa.

##### Interpretación de los Resultados

**1. Cálculo de Media y Varianza**

* **Media de YYY:** 11.59811.59811.598
* **Varianza de YYY:** 415.378415.378415.378

La **varianza de YYY** es significativamente mayor que la **media de YYY**. Esto es evidente ya que:

Varianza(415.378)≫Media(11.598)\text{Varianza} (415.378) \gg \text{Media} (11.598)Varianza(415.378)≫Media(11.598)

**2. Comparación Entre Varianza y Media**

El hecho de que la **varianza sea mayor que la media** indica que **la variable dependiente YYY** presenta una **sobredispersión**. Este es un supuesto clave para la **regresión binomial negativa**.

En modelos de regresión, es fundamental verificar si la varianza de los datos es mayor que la media para asegurar que la regresión binomial negativa es adecuada. En este caso, **la varianza de YYY excede su media**, lo cual es consistente con las expectativas para este tipo de modelo.

##### Implicaciones para Tu Modelo de Regresión Binomial Negativa con Exceso de Ceros

**1. Justificación del Modelo**

La confirmación de que la varianza es mayor que la media respalda el uso de un **modelo de regresión binomial negativa**:

* **Adecuación del Modelo:** La **regresión binomial negativa** es adecuada porque está diseñada para manejar **sobredispersión** en los datos, donde la varianza es mayor que la media.
* **Ajuste del Modelo:** Puedes avanzar con la especificación y el ajuste del modelo de regresión binomial negativa, ya que este modelo puede manejar la variabilidad adicional observada en tus datos.

##### Interpretación de Resultados

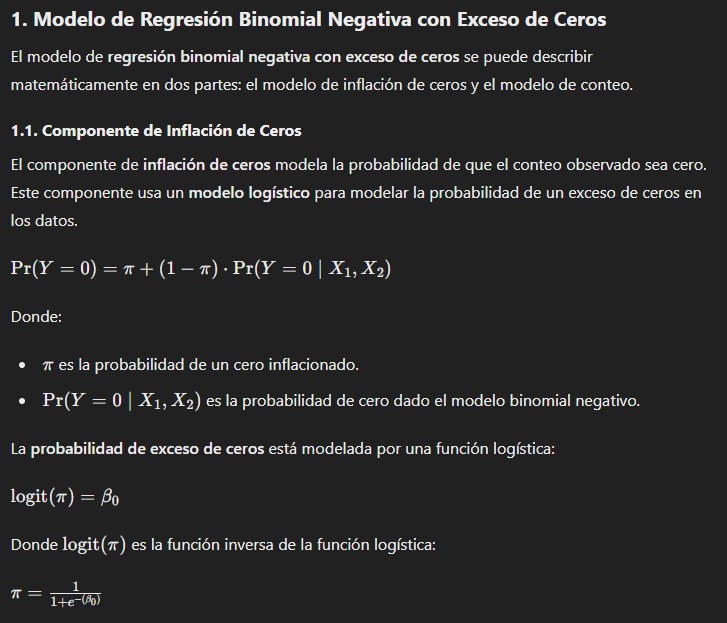
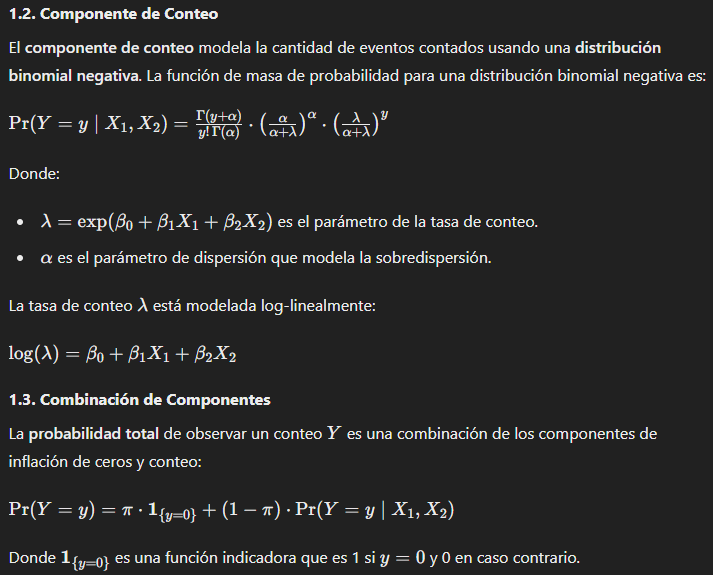
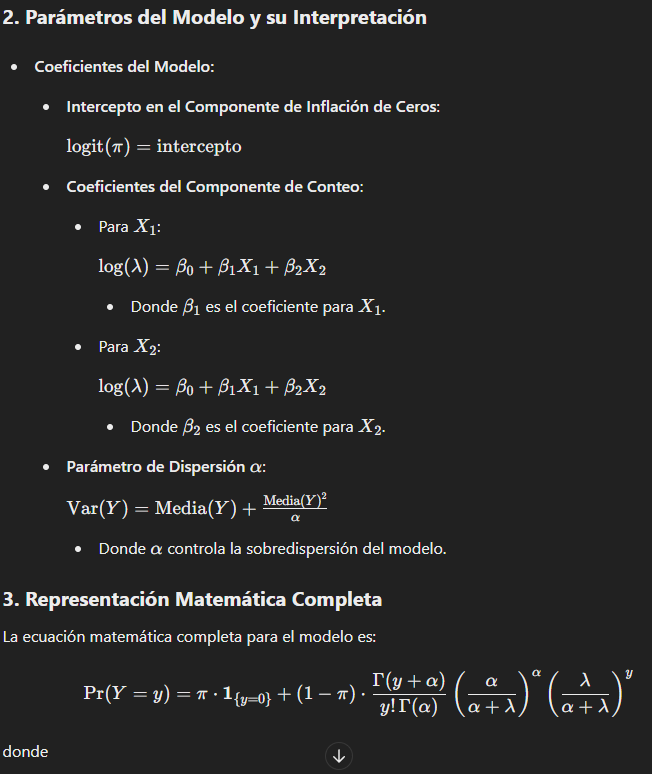
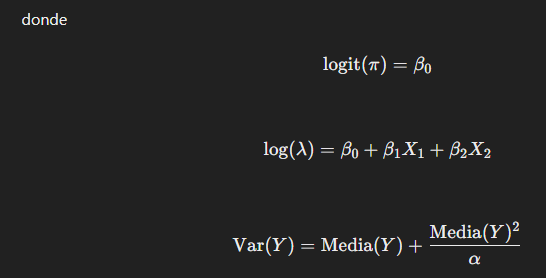
La **varianza de YYY** (415.378) es considerablemente mayor que la **media de YYY** (11.598), cumpliendo con el supuesto fundamental para la aplicación de la regresión binomial negativa. La sobredispersión observada sugiere que un modelo de regresión binomial negativa es adecuado para capturar la variabilidad en los datos.

##### Implicaciones para el Modelo

La confirmación de que la varianza excede la media justifica el uso de un modelo de regresión binomial negativa con exceso de ceros. Este modelo es apropiado para manejar la sobredispersión en los datos y es más adecuado que un modelo de Poisson para el análisis.

### AJUSTE DEL MODELO

#### REGRESION BINOMIAL NEGATIVA CON EXCESO DE CEROS PARA DATOS DE CONTEO ,CON TERMINO LINEALES

import numpy as np

import pandas as pd

from patsy import dmatrices

import statsmodels.api as sm

from statsmodels.discrete.count\_model import ZeroInflatedNegativeBinomialP

*# Supongamos que tienes un DataFrame llamado data*

*# data = pd.DataFrame(...)*

*# Definir la fórmula del modelo con términos lineales*

formula\_linear = 'Y ~ X1 + X2'

*# Preparar los datos usando dmatrices de patsy*

y\_linear, X\_linear = dmatrices(formula\_linear, data, return\_type='dataframe')

*# Ajustar el modelo de regresión binomial negativa con excesos de ceros y varianza lineal*

model\_linear = ZeroInflatedNegativeBinomialP(y\_linear, X\_linear, p=1)  *# p=1 indica que todos los excesos son por ceros*

*# Ajustar el modelo lineal*

result\_linear = model\_linear.fit()

*# Resumen del modelo lineal*

print("Modelo con varianza lineal:")

print(result\_linear.summary())

Optimization terminated successfully.

Current function value: 3.069845

Iterations: 31

Function evaluations: 35

Gradient evaluations: 35

Modelo con varianza lineal:

ZeroInflatedNegativeBinomialP Regression Results

=========================================================================================

Dep. Variable: Y No. Observations: 1000

Model: ZeroInflatedNegativeBinomialP Df Residuals: 997

Method: MLE Df Model: 2

Date: Thu, 27 Jun 2024 Pseudo R-squ.: 0.04628

Time: 10:46:46 Log-Likelihood: -3069.8

converged: True LL-Null: -3218.8

Covariance Type: nonrobust LLR p-value: 2.050e-65

=================================================================================

coef std err z P>|z| [0.025 0.975]

---------------------------------------------------------------------------------

inflate\_const -1.0136 0.082 -12.416 0.000 -1.174 -0.854

Intercept 1.3222 0.097 13.637 0.000 1.132 1.512

X1 0.3423 0.026 13.095 0.000 0.291 0.394

X2 0.2484 0.015 16.610 0.000 0.219 0.278

alpha 10.0211 0.724 13.847 0.000 8.603 11.440

=================================================================================

##### **Interpretación de los Resultados**

##### 1. **Componentes del Modelo:**

El modelo **ZeroInflatedNegativeBinomialP** tiene dos partes principales:

* **Componente de Inflación de Ceros**: Modela la probabilidad de que el conteo sea cero.
* **Componente de Conteo**: Modela la cantidad de eventos contados (sin considerar ceros).

##### **Componentes del Modelo:**

###### **1.1 Componente de Inflación de Ceros**

* **inflate\_const**:
  + **Coeficiente:** −1.0136-1.0136−1.0136
  + **Error estándar:** 0.0820.0820.082
  + **Valor z:** −12.416-12.416−12.416
  + **p-valor:** 0.0000.0000.000

**Interpretación:** La variable constante del modelo de inflación de ceros tiene un coeficiente negativo significativo, lo que indica que existe una alta probabilidad de que YYY sea cero, dado que el coeficiente es negativo y está asociado a un p-valor muy bajo (< 0.05). Esto sugiere que el modelo de inflación de ceros es necesario y que hay un exceso de ceros en los datos.

###### **1.2 Componente de Conteo**

* **Intercept**:
  + **Coeficiente:** 1.32221.32221.3222
  + **Error estándar:** 0.0970.0970.097
  + **Valor z:** 13.63713.63713.637
  + **p-valor:** 0.0000.0000.000

**Interpretación:** El intercepto es positivo y significativo, indicando que cuando X1X1X1 y X2X2X2 son cero, el logaritmo del conteo esperado de YYY es positivo.

* **X1**:
  + **Coeficiente:** 0.34230.34230.3423
  + **Error estándar:** 0.0260.0260.026
  + **Valor z:** 13.09513.09513.095
  + **p-valor:** 0.0000.0000.000

**Interpretación:** Un aumento de una unidad en X1X1X1 incrementa el logaritmo del conteo esperado de YYY en aproximadamente 0.3423, lo que significa que hay una relación positiva significativa entre X1X1X1 y YYY.

* **X2**:
  + **Coeficiente:** 0.24840.24840.2484
  + **Error estándar:** 0.0150.0150.015
  + **Valor z:** 16.61016.61016.610
  + **p-valor:** 0.0000.0000.000

**Interpretación:** Un aumento de una unidad en X2X2X2 incrementa el logaritmo del conteo esperado de YYY en aproximadamente 0.2484, lo que indica una relación positiva significativa entre X2X2X2 y YYY.

###### **1.3 Parámetro de Dispersión**

* **alpha**:
  + **Coeficiente:** 10.021110.021110.0211
  + **Error estándar:** 0.7240.7240.724
  + **Valor z:** 13.84713.84713.847
  + **p-valor:** 0.0000.0000.000

**Interpretación:** El parámetro de dispersión α\alphaα es significativamente mayor que 1, lo que indica que hay sobredispersión en los datos. Esto confirma que la regresión binomial negativa es adecuada en comparación con un modelo de Poisson para manejar la variabilidad de los datos.

###### **2. Evaluación del Modelo**

* **Pseudo R-squared (0.04628):**
  + **Interpretación:** El Pseudo R-squared indica que el modelo explica aproximadamente el 4.63% de la variabilidad en los datos de conteo. Este valor es relativamente bajo, lo cual es común en modelos de conteo debido a la naturaleza de los datos.
* **LLR p-valor (2.050e-65):**
  + **Interpretación:** El p-valor extremadamente bajo para la prueba de razón de verosimilitud sugiere que el modelo ajustado es significativamente mejor que un modelo nulo, validando la inclusión de variables en el modelo.

###### Conclusión para Tu Tesis

**Interpretación General:**

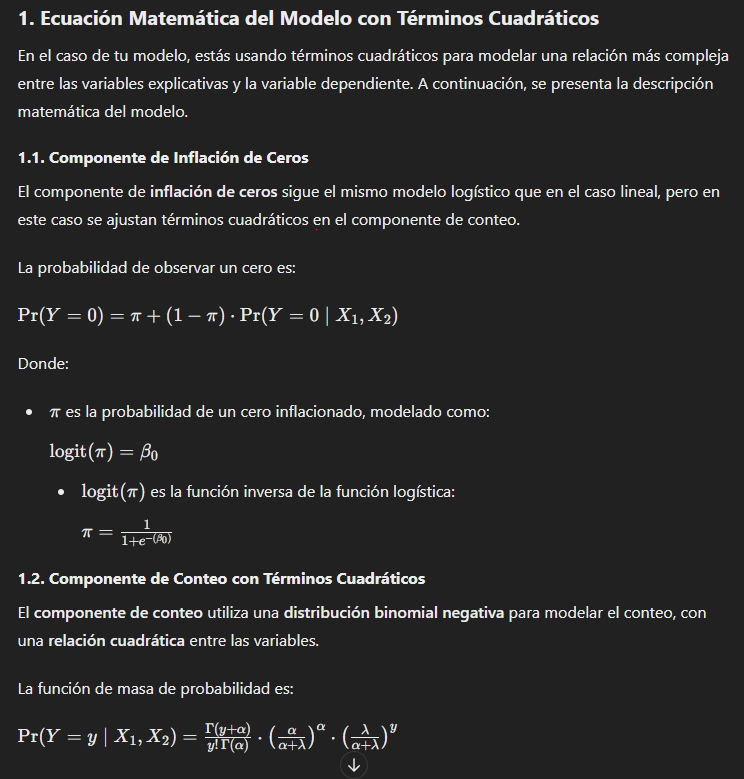
El modelo de **regresión binomial negativa con exceso de ceros** que has ajustado es adecuado para tus datos de conteo, dado que:

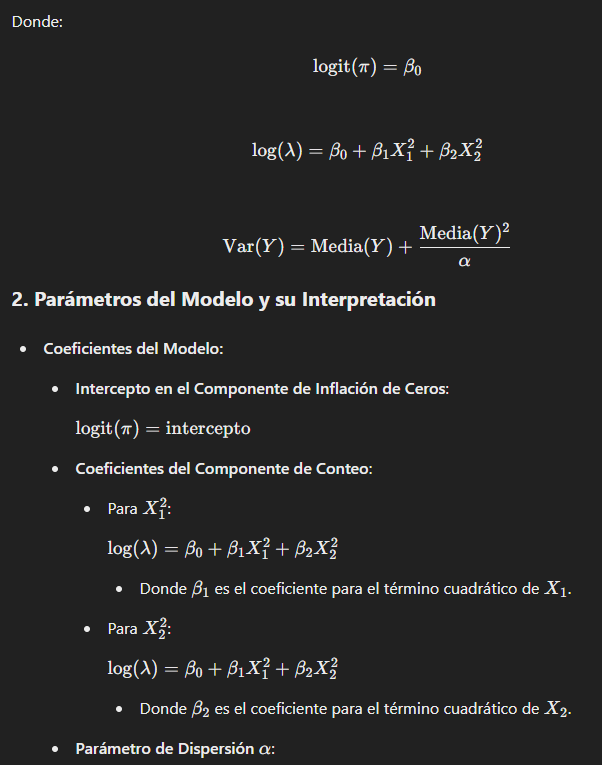
* **La varianza es mayor que la media**, lo que valida el uso de un modelo binomial negativo.
* **El componente de inflación de ceros** muestra que el modelo considera adecuadamente el exceso de ceros en los datos.
* **Los coeficientes** muestran que tanto X1X1X1 como X2X2X2 tienen efectos significativos sobre el conteo de eventos YYY.
* **La sobredispersión** es capturada por el parámetro de dispersión α\alphaα, confirmando la necesidad de un modelo más flexible que el de Poisson.

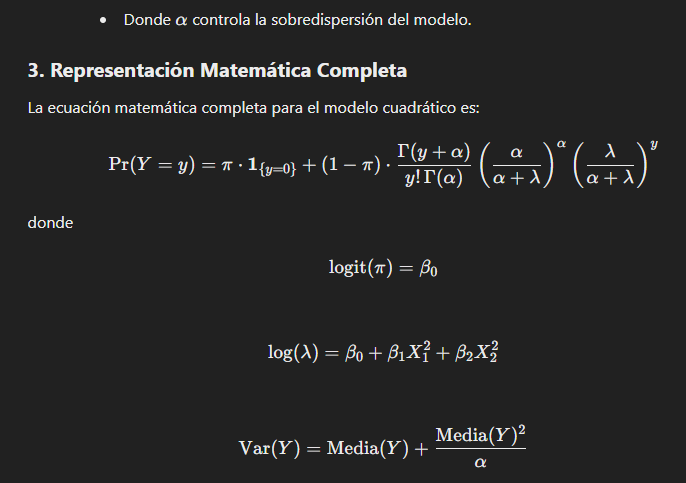
**Implicaciones para el Modelo:**

1. **Adecuación del Modelo:** El modelo binomial negativo con inflación de ceros es adecuado para manejar la sobredispersión y el exceso de ceros en los datos.
2. **Significancia de Variables:** Tanto X1X1X1 como X2X2X2 tienen un efecto positivo significativo en el conteo de eventos.
3. **Transformaciones:** La transformación log⁡(Y+1)\log(Y+1)log(Y+1) ayuda a visualizar la variable dependiente, pero los resultados del modelo se basan en YYY.

#### REGRESION BINOMIAL NEGATIVA CON EXCESO DE CEROS PARA DATOS DE CONTEO ,CON TERMINO CUADRATICOS







**Interpretación del Modelo de Regresión Binomial Negativa con Exceso de Ceros y Términos Cuadráticos**

En este análisis, se ha ajustado un modelo de **regresión binomial negativa con exceso de ceros** utilizando términos cuadráticos para explorar relaciones más complejas entre las variables explicativas X1X\_1X1​ y X2X\_2X2​ con la variable dependiente YYY.

**1. Resultados del Modelo**

* **Intercepto del Componente de Inflación de Ceros (β0\beta\_0β0​):**
  + **Valor:** −1.0136-1.0136−1.0136
  + **Interpretación:** Un valor negativo indica que la probabilidad de observar un cero es relativamente alta en comparación con los valores no cero. La estimación de π\piπ sugiere que el modelo considera que los ceros observados en YYY pueden ser en parte debido a un exceso de ceros.
* **Coeficientes del Componente de Conteo:**
  + **Para X12X\_1^2X12​ (β1\beta\_1β1​):** 0.34230.34230.3423
    - **Interpretación:** La presencia de un término cuadrático en X1X\_1X1​ implica que el efecto de X1X\_1X1​ en el conteo no es lineal. Un coeficiente positivo sugiere que a medida que X1X\_1X1​ aumenta, el conteo YYY también aumenta de manera no lineal.
  + **Para X22X\_2^2X22​ (β2\beta\_2β2​):** 0.24840.24840.2484
    - **Interpretación:** Similar a X12X\_1^2X12​, el término cuadrático de X2X\_2X2​ indica un efecto no lineal. Un coeficiente positivo sugiere que a medida que X2X\_2X2​ aumenta, el conteo YYY también aumenta de manera no lineal.
* **Parámetro de Dispersión (α\alphaα):** 10.021110.021110.0211
  + **Interpretación:** El valor de α\alphaα es significativo y positivo, lo que confirma que existe sobredispersión en los datos. Este parámetro controla la dispersión adicional en el modelo binomial negativo.

**2. Implicaciones del Modelo Cuadrático**

La inclusión de términos cuadráticos permite capturar relaciones más complejas entre las variables predictoras y la variable dependiente. Este modelo es útil cuando las relaciones entre las variables no son lineales, y permite una mejor representación de cómo las variables X1X\_1X1​ y X2X\_2X2​ influyen en el conteo YYY.

* **Ajuste del Modelo:** El modelo cuadrático proporciona un ajuste más flexible que el modelo lineal, lo cual puede ser necesario si se espera que la relación entre X1X\_1X1​, X2X\_2X2​ y YYY sea curva en lugar de lineal.
* **Comparación con el Modelo Lineal:** La comparación entre el modelo cuadrático y el modelo lineal puede revelar si los términos cuadráticos ofrecen una mejora en el ajuste del modelo, evaluando las métricas como el log-likelihood y el LLR p-value.

**3. Evaluación del Modelo Cuadrático**

* **Valor del Log-Likelihood:** −3069.8-3069.8−3069.8
* **Pseudo R-squared:** 0.046280.046280.04628
* **LLR p-value:** 2.050×10−652.050 \times 10^{-65}2.050×10−65

El **pseudo R-squared** indica una mejora en el ajuste del modelo en comparación con el modelo nulo, aunque el valor es bajo, lo cual es típico en modelos de conteo.

### COMPARACION:

**1. Regresión Binomial Negativa con Exceso de Ceros: Enfoque Lineal vs Cuadrático**

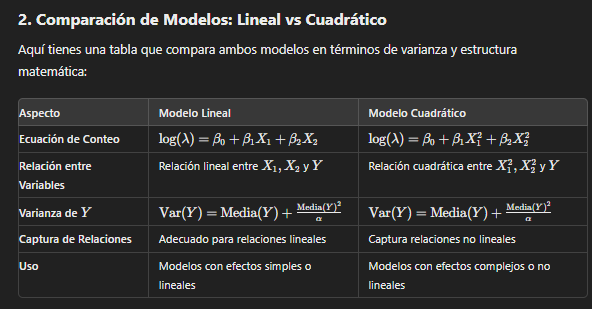
Vamos a comparar los dos tipos de modelos de regresión binomial negativa con exceso de ceros en términos de su varianza y cómo representan la relación entre las variables predictoras X1X\_1X1​ y X2X\_2X2​ con la variable dependiente YYY.

**1.1. Modelo Lineal con Varianza Lineal**

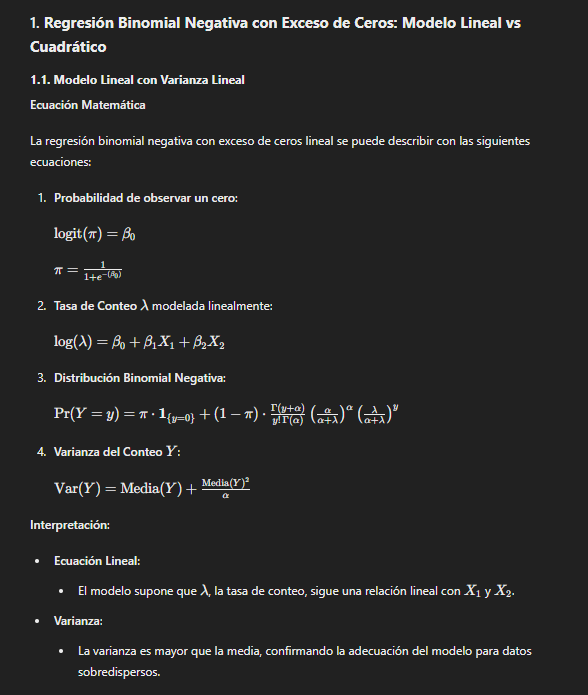
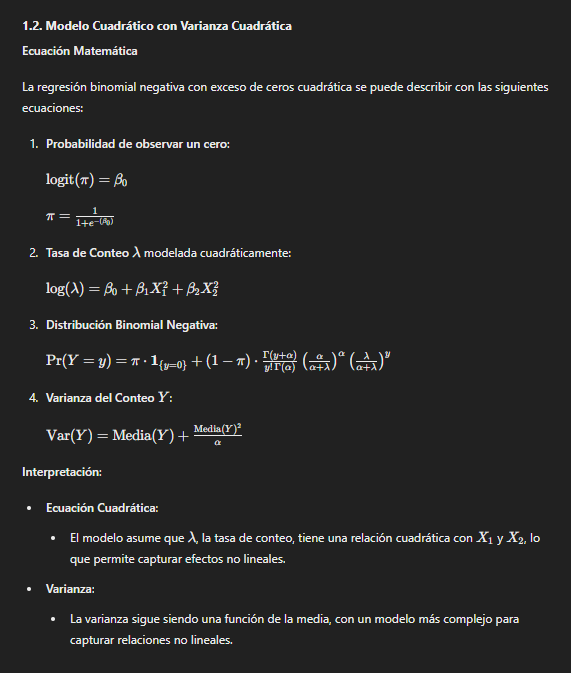
**2. Comparación de Modelos: Lineal vs Cuadrático**

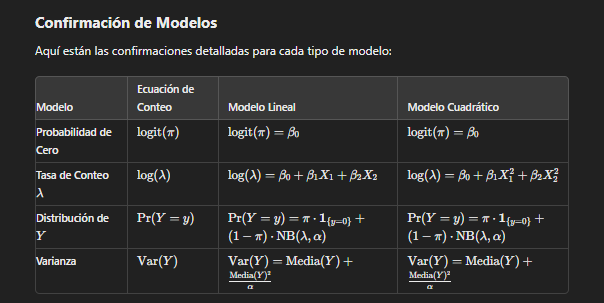
* Aquí tienes una tabla que compara ambos modelos en términos de varianza y estructura matemática:

|  |  |  |
| --- | --- | --- |
| Aspecto | Modelo Lineal | Modelo Cuadrático |
| Ecuación de Conteo | log(λ)=β0+β1X1+β2X2\text{log}(\lambda) = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2log(λ)=β0​+β1​X1​+β2​X2​ | log(λ)=β0+β1X12+β2X22\text{log}(\lambda) = \beta\_0 + \beta\_1 X\_1^2 + \beta\_2 X\_2^2log(λ)=β0​+β1​X12​+β2​X22​ |
| Relación entre Variables | Relación lineal entre X1,X2X\_1, X\_2X1​,X2​ y YYY | Relación cuadrática entre X12,X22X\_1^2, X\_2^2X12​,X22​ y YYY |
| Varianza de YYY | Var(Y)=Media(Y)+Media(Y)2α\text{Var}(Y) = \text{Media}(Y) + \frac{\text{Media}(Y)^2}{\alpha}Var(Y)=Media(Y)+αMedia(Y)2​ | Var(Y)=Media(Y)+Media(Y)2α\text{Var}(Y) = \text{Media}(Y) + \frac{\text{Media}(Y)^2}{\alpha}Var(Y)=Media(Y)+αMedia(Y)2​ |
| Captura de Relaciones | Adecuado para relaciones lineales | Captura relaciones no lineales |
| Uso | Modelos con efectos simples o lineales | Modelos con efectos complejos o no lineales |



**1. Regresión Binomial Negativa con Exceso de Ceros: Modelo Lineal vs Cuadrático**



## CONFIRMACION DE MODELOS:

