**Metodología**

1. **Generación de Datos**:
   * Generación de variables explicativas (X1 y X2) con distribuciones normales.
   * Configuración de la variable dependiente Y utilizando una distribución binomial negativa y una proporción de ceros añadidos.
2. **Análisis Exploratorio de Datos (EDA)**:
   * Verificación de la distribución de Y y relación entre X1 y X2.
   * Verificación de multicolinealidad entre las variables explicativas mediante el Factor de Inflación de la Varianza (VIF).
3. **Pruebas Estadísticas**:
   * Prueba de normalidad para X1 y X2.
   * Análisis de correlación entre X1 y Y, y X2 y Y.
   * Análisis de heterocedasticidad utilizando la prueba de Breusch-Pagan.
   * Análisis de valores atípicos mediante diagramas de caja.
4. **Transformaciones de Datos**:
   * Transformación de Y mediante logaritmos para manejar ceros.
   * Cálculo de dispersión para validar el supuesto de la regresión binomial negativa.

**Resultados**

1. **Ajuste de Modelos**:
   * Ajuste de un modelo binomial negativo simple y análisis de residuos.
   * Ajuste de modelos de regresión binomial negativa con exceso de ceros utilizando términos lineales y cuadráticos.
   * Evaluación y comparación de modelos utilizando el resumen del ajuste del modelo.
2. **Diagnósticos de Modelos**:
   * Gráficos de residuos, Q-Q plots y análisis de residuos estandarizados.
   * Validación cruzada para evaluar el rendimiento del modelo.
3. **Estadísticas de Ajuste y Significancia**:
   * Cálculo de pseudo R-cuadrado, Log-Likelihood y otras estadísticas de ajuste.
   * Análisis de significancia de los coeficientes del modelo.
4. **Validación del Modelo**:
   * Validación cruzada utilizando K-Fold y cálculo del error cuadrático medio (MSE).

**Recomendaciones Adicionales**

Para asegurar que tu código cubra completamente los capítulos de metodología y resultados, considera agregar:

1. **Descripción Detallada del Proceso**:
   * Documenta cada paso del análisis con comentarios detallados para clarificar la justificación de cada análisis y decisión metodológica.
2. **Interpretación de Resultados**:
   * Incluye interpretaciones detalladas de los resultados de cada análisis, especialmente los coeficientes del modelo y su significancia.
3. **Comparación de Modelos**:
   * Compara los diferentes modelos ajustados (lineal, cuadrático, etc.) y discute cuál se ajusta mejor a tus datos y por qué.
4. **Validación Adicional**:
   * Considera realizar pruebas adicionales de ajuste del modelo, como análisis de residuos vs predicciones para detectar cualquier patrón no capturado por el modelo.

### Metodología

#### Generación de Datos

1. **Variables Explicativas**:
   * Generación de X1 y X2 con distribuciones normales.
2. **Variable Dependiente**:
   * Configuración de la variable dependiente Y utilizando una distribución binomial negativa.
   * Adición de ceros extra para simular el exceso de ceros.

#### Análisis Exploratorio de Datos (EDA)

1. **Distribución de Variables**:
   * Visualización de la distribución de Y y relación entre X1 y X2.
2. **Multicolinealidad**:
   * Verificación de multicolinealidad entre X1 y X2 utilizando VIF.

#### Pruebas Estadísticas

1. **Pruebas de Normalidad**:
   * Pruebas de Shapiro-Wilk para X1 y X2.
2. **Correlación**:
   * Cálculo de correlaciones entre X1 y Y, y X2 y Y.
3. **Heterocedasticidad**:
   * Prueba de Breusch-Pagan para detectar heterocedasticidad.

#### Transformaciones y Validaciones

1. **Transformación de Y**:
   * Aplicación de logaritmo a Y para manejar ceros y evaluar la dispersión.

### Resultados

#### Ajuste de Modelos

1. **Modelo Binomial Negativo**:
   * Ajuste de un modelo binomial negativo simple y análisis de residuos.
2. **Modelos con Exceso de Ceros**:
   * Ajuste de modelos de regresión binomial negativa con exceso de ceros utilizando términos lineales y cuadráticos.
   * Comparación de modelos con varianza lineal y cuadrática.
3. **Evaluación de Modelos**:
   * Uso de estadísticos de ajuste y significancia de los coeficientes.
   * Comparación de resultados de modelos lineales y cuadráticos.

#### Diagnósticos de Modelos

1. **Análisis de Residuos**:
   * Gráficos de residuos, Q-Q plots y análisis de residuos estandarizados.
   * Gráficos de residuos vs predicciones para cada modelo.

#### Validación del Modelo

1. **Validación Cruzada**:
   * Implementación de validación cruzada K-Fold y cálculo del MSE para evaluar la estabilidad del modelo.

import numpy as np

import pandas as pd

from patsy import dmatrices

import statsmodels.api as sm

from statsmodels.discrete.count\_model import ZeroInflatedNegativeBinomialP

# Generación de datos

np.random.seed(42)

n = 1000

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

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

r = 2

p = 0.5

mu = np.exp(1 + 0.5 \* X1 + 0.3 \* X2)

size = r

prob = size / (size + mu)

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

zero\_inflation = 0.3

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

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

# Ajuste de modelos lineal y cuadrático

formulas = {

"lineal": 'Y ~ X1 + X2',

"cuadrático": 'Y ~ X1 + np.power(X1, 2) + X2 + np.power(X2, 2)'

}

results = {}

for key, formula in formulas.items():

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

model = ZeroInflatedNegativeBinomialP(y, X, p=1)

result = model.fit()

results[key] = result

print(f"Resumen del modelo {key}:")

print(result.summary())

print(f"\nPseudo R-squared: {result.prsquared:.5f}")

print(f"Log-Likelihood: {result.llf:.5f}")

print(f"LL-Null: {result.llnull:.5f}")

print(f"LLR p-value: {result.llr\_pvalue:.5f}")

# Validación cruzada

from sklearn.model\_selection import KFold

from sklearn.metrics import mean\_squared\_error

kf = KFold(n\_splits=5)

cv\_scores = {key: [] for key in formulas.keys()}

for train\_index, test\_index in kf.split(X):

for key, formula in formulas.items():

y\_train, X\_train = y.iloc[train\_index], X.iloc[train\_index]

y\_test, X\_test = y.iloc[test\_index], X.iloc[test\_index]

model = ZeroInflatedNegativeBinomialP(y\_train, X\_train, p=1)

result = model.fit()

y\_pred = result.predict(X\_test)

mse = mean\_squared\_error(y\_test, y\_pred)

cv\_scores[key].append(mse)

for key in formulas.keys():

avg\_cv\_mse = np.mean(cv\_scores[key])

print(f"Error cuadrático medio promedio en validación cruzada (5-fold) para el modelo {key}: {avg\_cv\_mse:.5f}")

# ***METODOLOGIA***

**Índice Propuesto**

1. **Introducción**
   * Planteamiento del problema
   * Objetivos de la investigación
   * Estructura de la tesis
2. **Revisión de la Literatura**
   * Modelos de regresión para datos de conteo
   * Modelos de regresión binomial negativa
   * Modelos con exceso de ceros
   * Varianza lineal y cuadrática en modelos de regresión
3. **Metodología**
   * Generación de datos
   * Análisis exploratorio de datos
   * Pruebas estadísticas
   * Ajuste de modelos
   * Validación de modelos
4. **Resultados**
   * Análisis descriptivo
   * Resultados del ajuste de modelos
   * Diagnósticos de modelos
   * Validación cruzada
5. **Discusión**
   * Interpretación de resultados
   * Comparación con estudios previos
   * Implicaciones prácticas
6. **Conclusiones y Recomendaciones**
   * Resumen de hallazgos
   * Limitaciones del estudio
   * Sugerencias para investigaciones futuras
7. **Referencias**

### Metodología

#### Generación de Datos

1. **Variables Explicativas**:
   * Se generaron dos variables explicativas, X1 y X2, con distribuciones normales: X1 ~ N(0, 1) y X2 ~ N(5, 2).
2. **Variable Dependiente**:
   * La variable dependiente Y se generó utilizando una distribución binomial negativa con parámetros r = 2 y p = 0.5. Para simular el exceso de ceros, se añadió una proporción de ceros adicionales.

python

Copy code

np.random.seed(42)

n = 1000

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

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

r = 2

p = 0.5

mu = np.exp(1 + 0.5 \* X1 + 0.3 \* X2)

size = r

prob = size / (size + mu)

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

zero\_inflation = 0.3

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

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

#### Análisis Exploratorio de Datos (EDA)

1. **Distribución de Variables**:
   * Se analizaron las distribuciones de Y, X1 y X2 mediante histogramas y gráficos de densidad.
2. **Multicolinealidad**:
   * Se verificó la multicolinealidad entre X1 y X2 utilizando el Factor de Inflación de la Varianza (VIF).

#### Pruebas Estadísticas

1. **Pruebas de Normalidad**:
   * Se realizaron pruebas de Shapiro-Wilk para X1 y X2.
2. **Correlación**:
   * Se calcularon las correlaciones entre X1 y Y, y X2 y Y.
3. **Heterocedasticidad**:
   * Se realizó la prueba de Breusch-Pagan para detectar heterocedasticidad.

#### Transformaciones y Validaciones

1. **Transformación de Y**:
   * Se aplicó el logaritmo a Y para manejar ceros y evaluar la dispersión.

### Resultados

#### Ajuste de Modelos

1. **Modelo Binomial Negativo**:
   * Se ajustó un modelo binomial negativo simple y se analizaron los residuos.
2. **Modelos con Exceso de Ceros**:
   * Se ajustaron modelos de regresión binomial negativa con exceso de ceros utilizando términos lineales y cuadráticos. Se compararon los modelos con varianza lineal y cuadrática.
3. **Evaluación de Modelos**:
   * Se usaron estadísticas de ajuste y significancia de los coeficientes para evaluar los modelos.

python

Copy code

formulas = {

"lineal": 'Y ~ X1 + X2',

"cuadrático": 'Y ~ X1 + np.power(X1, 2) + X2 + np.power(X2, 2)'

}

results = {}

for key, formula in formulas.items():

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

model = ZeroInflatedNegativeBinomialP(y, X, p=1)

result = model.fit()

results[key] = result

print(f"Resumen del modelo {key}:")

print(result.summary())

print(f"\nPseudo R-squared: {result.prsquared:.5f}")

print(f"Log-Likelihood: {result.llf:.5f}")

print(f"LL-Null: {result.llnull:.5f}")

print(f"LLR p-value: {result.llr\_pvalue:.5f}")

#### Diagnósticos de Modelos

1. **Análisis de Residuos**:
   * Se realizaron gráficos de residuos, Q-Q plots y análisis de residuos estandarizados.
   * Se crearon gráficos de residuos vs predicciones para cada modelo.

#### Validación del Modelo

1. **Validación Cruzada**:
   * Se implementó validación cruzada K-Fold y se calculó el MSE para evaluar la estabilidad del modelo.

python

Copy code

from sklearn.model\_selection import KFold

from sklearn.metrics import mean\_squared\_error

kf = KFold(n\_splits=5)

cv\_scores = {key: [] for key in formulas.keys()}

for train\_index, test\_index in kf.split(data):

for key, formula in formulas.items():

y\_train, X\_train = y.iloc[train\_index], X.iloc[train\_index]

y\_test, X\_test = y.iloc[test\_index], X.iloc[test\_index]

model = ZeroInflatedNegativeBinomialP(y\_train, X\_train, p=1)

result = model.fit()

y\_pred = result.predict(X\_test)

mse = mean\_squared\_error(y\_test, y\_pred)

cv\_scores[key].append(mse)

for key in formulas.keys():

avg\_cv\_mse = np.mean(cv\_scores[key])

print(f"Error cuadrático medio promedio en validación cruzada (5-fold) para el modelo {key}: {avg\_cv\_mse:.5f}")

### Discusión

* **Interpretación de Resultados**:
  + Los resultados de los modelos lineales y cuadráticos se compararon y se discutió cuál modelo se ajusta mejor a los datos.
  + Se analizaron las implicaciones de los coeficientes y su significancia estadística.
* **Comparación con Estudios Previos**:
  + Se compararon los hallazgos con estudios previos sobre modelos binomiales negativos y exceso de ceros.
* **Implicaciones Prácticas**:
  + Se discutieron las posibles aplicaciones de los resultados en contextos reales y se exploraron las limitaciones del estudio.

### Conclusiones y Recomendaciones

* **Resumen de Hallazgos**:
  + Se presentaron las principales conclusiones sobre la eficiencia y precisión de los modelos ajustados.
* **Limitaciones del Estudio**:
  + Se discutieron las limitaciones encontradas durante la investigación.
* **Sugerencias para Investigaciones Futuras**:
  + Se propusieron áreas de mejora y posibles estudios futuros que puedan expandir los hallazgos.

### Referencias

* Se incluyeron todas las fuentes y estudios citados a lo largo de la tesis.

JUSTIFICACION DE LAS VARIABLES SIMULADAS:

### Metodología

#### Generación de Datos

##### Variables Explicativas

Se generaron dos variables explicativas, X1X1X1 y X2X2X2, con distribuciones normales: X1∼N(0,1)X1 \sim N(0, 1)X1∼N(0,1) y X2∼N(5,2)X2 \sim N(5, 2)X2∼N(5,2).

##### Justificación del Supuesto

1. **Distribución Normal en la Naturaleza**: La distribución normal es una de las distribuciones más comunes en estadística y en la naturaleza. Muchas variables en ciencias sociales, biológicas y físicas tienden a seguir una distribución normal debido al teorema central del límite, que establece que la suma de un gran número de variables aleatorias independientes, cada una con una distribución cualquiera, tenderá a una distribución normal.
2. **Simplicidad y Flexibilidad**: La normalidad es un supuesto estándar en muchos modelos estadísticos debido a su simplicidad matemática y flexibilidad. Las propiedades bien conocidas de la distribución normal permiten una fácil interpretación y modelado de los datos.
3. **Simulación y Estudio**: En estudios de simulación, suponer variables explicativas normalmente distribuidas es común para simplificar el análisis y centrarse en el comportamiento del modelo. Esta suposición facilita la creación de escenarios controlados donde se pueden explorar las propiedades del modelo y evaluar su desempeño en condiciones ideales.
4. **Características de las Variables**: En este estudio, X1X1X1 y X2X2X2 podrían representar factores como puntajes de pruebas, medidas biológicas, o cualquier otra variable continua que razonablemente podría seguir una distribución normal. Por ejemplo, X1X1X1 podría ser una medida estandarizada de una habilidad cognitiva, mientras que X2X2X2 podría ser una medida biológica con una media mayor debido a un sesgo natural en la población.

##### Variable Dependiente

La variable dependiente YYY se generó utilizando una distribución binomial negativa con parámetros r=2r = 2r=2 y p=0.5p = 0.5p=0.5. Para simular el exceso de ceros, se añadió una proporción de ceros adicionales.

python

Copy code

import numpy as np

import pandas as pd

from scipy.stats import nbinom

np.random.seed(42)

n = 1000

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

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

r = 2

p = 0.5

mu = np.exp(1 + 0.5 \* X1 + 0.3 \* X2)

size = r

prob = size / (size + mu)

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

zero\_inflation = 0.3

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

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

En resumen, la elección de variables explicativas con distribuciones normales es una práctica estándar que facilita el análisis y proporciona un punto de partida sólido para la modelización. Esta suposición permite concentrarse en el comportamiento del modelo de regresión binomial negativa con exceso de ceros y en cómo maneja la varianza lineal y cuadrática en diferentes condiciones.

# **SI ME PREGUNTAN EN MI JURADO Q,POR QUE DECIDISTE USAR ESAS VARIBALES EN MI TESIS COMO RESPONDERIA?**

Si te preguntan en tu jurado por qué decidiste usar esas variables en tu tesis, puedes responder de la siguiente manera:

**Respuesta para el Jurado**

**Pregunta:** ¿Por qué decidiste usar esas variables explicativas en tu tesis?

**Respuesta:**

1. **Distribución Normal Común en la Naturaleza**:
   * Elegí variables con distribuciones normales porque la distribución normal es una de las más comunes y bien entendidas en estadística. Muchas características y fenómenos en la naturaleza y en las ciencias sociales siguen una distribución normal debido al teorema central del límite. Por ejemplo, medidas de rendimiento, características biológicas, y otros factores a menudo se distribuyen normalmente.
2. **Simplicidad y Eficiencia en la Simulación**:
   * En estudios de simulación, suponer que las variables explicativas siguen una distribución normal es una práctica estándar. Esta suposición simplifica el análisis y permite centrarse en evaluar el comportamiento del modelo bajo condiciones ideales. Facilita la creación de datos reproducibles y comparables, lo que es esencial para analizar y entender las propiedades del modelo de regresión binomial negativa con exceso de ceros.
3. **Propiedades Matemáticas y Estadísticas**:
   * La distribución normal tiene propiedades matemáticas y estadísticas que son bien conocidas y explotadas en análisis estadísticos. Por ejemplo, conocer la media y la varianza proporciona una base sólida para la interpretación y modelización de los datos. En el contexto de la regresión binomial negativa, facilita el ajuste y la evaluación de los modelos.
4. **Representación Realista de Variables**:
   * Las variables X1X1X1 y X2X2X2 fueron generadas con medias y varianzas específicas (X1∼N(0,1)X1 \sim N(0, 1)X1∼N(0,1) y X2∼N(5,2)X2 \sim N(5, 2)X2∼N(5,2)) para representar diferentes escenarios realistas. X1X1X1 puede representar una variable con una media centrada en cero, como una medida de habilidad estándar, mientras que X2X2X2 tiene una media diferente, simulando una variable con un sesgo natural, como una característica biológica. Estas distribuciones permiten estudiar cómo las diferentes condiciones afectan el modelo.
5. **Flexibilidad y Aplicabilidad**:
   * Usar variables con distribuciones normales ofrece flexibilidad y aplicabilidad a una amplia gama de situaciones prácticas. Las técnicas y resultados obtenidos con estas variables pueden ser adaptables a otros contextos donde las variables tengan distribuciones similares.
6. **Exploración del Comportamiento del Modelo**:
   * Al usar variables normalmente distribuidas, puedo explorar de manera eficiente el comportamiento del modelo de regresión binomial negativa con exceso de ceros bajo diferentes escenarios. Esto incluye entender cómo la varianza lineal y cuadrática impactan el ajuste del modelo y la precisión de las predicciones.

En resumen, la elección de estas variables explicativas con distribuciones normales no solo simplifica el análisis y la simulación, sino que también proporciona un marco robusto y flexible para evaluar y entender las propiedades del modelo de regresión binomial negativa con exceso de ceros en diversas condiciones.

**Respuesta para el Jurado con Sustento Científico y Citas**

**Pregunta:** ¿Por qué decidiste usar esas variables explicativas en tu tesis?

**Respuesta:**

1. **Distribución Normal en Estudios Estadísticos**:
   * La elección de variables con distribuciones normales (X1∼N(0,1)X1 \sim N(0, 1)X1∼N(0,1) y X2∼N(5,2)X2 \sim N(5, 2)X2∼N(5,2)) se basa en el hecho de que la distribución normal es una de las más ampliamente estudiadas y comprendidas en estadística. Según el teorema central del límite, la suma de un gran número de variables aleatorias independientes tiende a una distribución normal, independientemente de la distribución original de las variables. Este comportamiento es crucial en muchos fenómenos naturales y sociales, donde las características miden tienden a ser normales (Montgomery et al., 2012).
2. **Simplicidad y Control en Simulación**:
   * Utilizar variables con distribuciones normales simplifica la simulación y permite un control más preciso sobre los parámetros del modelo. La normalidad de las variables facilita la generación de datos reproducibles y comparables, lo cual es esencial para evaluar de manera efectiva las propiedades y el desempeño del modelo de regresión binomial negativa con exceso de ceros (Wooldridge, 2010).
3. **Propiedades Matemáticas y Estadísticas de la Distribución Normal**:
   * Las variables normalmente distribuidas tienen propiedades estadísticas bien conocidas que son explotadas en el análisis estadístico. Estas propiedades incluyen la simetría alrededor de la media, la distribución conocida de varianza y desviación estándar, y la facilidad de manejo algebraico y analítico. Estas características son ventajosas para modelar y analizar datos en un contexto de regresión (Montgomery et al., 2012).
4. **Relevancia y Representatividad en Fenómenos Reales**:
   * Elegir X1X1X1 con N(0,1)N(0, 1)N(0,1) y X2X2X2 con N(5,2)N(5, 2)N(5,2) permite representar diferentes escenarios realistas. X1X1X1 puede simbolizar una variable sin sesgo central (p.ej., una medida de rendimiento estándar), mientras que X2X2X2 representa una variable con sesgo (p.ej., una característica biológica). Esta variabilidad en las características de las variables explicativas ayuda a explorar cómo diferentes condiciones impactan el comportamiento del modelo (Box & Cox, 1964).
5. **Flexibilidad y Generalización**:
   * Las técnicas y resultados obtenidos al trabajar con variables normalmente distribuidas pueden ser fácilmente adaptados y generalizados a otras situaciones y contextos. Esto es debido a la amplia aplicabilidad de la distribución normal en diversas disciplinas científicas y su capacidad para aproximar muchas otras distribuciones a través de la teoría del límite central (Montgomery et al., 2012).
6. **Evaluación de Modelos bajo Diferentes Condiciones**:
   * Utilizar variables con distribuciones normales permite evaluar de manera eficiente el impacto de diferentes condiciones en el modelo de regresión binomial negativa con exceso de ceros. Esto incluye el análisis de cómo las varianzas lineales y cuadráticas afectan el ajuste del modelo y la precisión de las predicciones (Wooldridge, 2010).
7. **Base Teórica y Empírica**:
   * La literatura existente en análisis de regresión y modelización estadística frecuentemente asume normalidad en las variables explicativas por su facilidad de manejo y las propiedades deseables que presentan. Esto se fundamenta en trabajos como los de Box y Cox (1964) en transformaciones de datos, así como en la teoría del teorema central del límite (Montgomery et al., 2012).

**Referencias**:

* Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211-243.
* Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. John Wiley & Sons.
* Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Este enfoque asegura que la elección de las variables tiene un sólido fundamento en la teoría estadística y en prácticas empíricas comunes, garantizando así la validez y relevancia del análisis realizado.

# RESPUESTA CON JUSTIFICACION MATEMATICA

**Respuesta con Justificación Matemática**

**Pregunta:** ¿Por qué decidiste usar esas variables explicativas en tu tesis?

**Respuesta:**

La elección de variables explicativas X1∼N(0,1)X1 \sim N(0, 1)X1∼N(0,1) y X2∼N(5,2)X2 \sim N(5, 2)X2∼N(5,2) se basa en una combinación de consideraciones matemáticas y estadísticas que se alinean con los objetivos de mi tesis sobre la modelación de la regresión binomial negativa con exceso de ceros.

1. **Propiedades de la Distribución Normal**:
   * **Simplicidad Matemática**: La distribución normal es definida por dos parámetros, la media (μ\muμ) y la varianza (σ2\sigma^2σ2), y tiene la forma: f(x)=12πσ2exp⁡(−(x−μ)22σ2).f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).f(x)=2πσ2​1​exp(−2σ2(x−μ)2​). Esta función de densidad es continua, simétrica y tiene propiedades matemáticas bien entendidas, lo que facilita el análisis y la simulación.
2. **Teorema Central del Límite (TCL)**:
   * **Importancia en Estadística**: El TCL establece que la suma de variables aleatorias independientes y identicamente distribuidas tiende a una distribución normal, independientemente de la distribución original de las variables. Este teorema es fundamental en estadística y justifica el uso de la distribución normal para modelar variables explicativas. Formalmente, si XiX\_iXi​ son nnn variables independientes e idénticamente distribuidas con media μ\muμ y varianza σ2\sigma^2σ2, entonces: ∑i=1nXi−nμnσ2→N(0,1)cuando n→∞.\frac{\sum\_{i=1}^n X\_i - n\mu}{\sqrt{n\sigma^2}} \rightarrow N(0, 1) \quad \text{cuando } n \rightarrow \infty.nσ2​∑i=1n​Xi​−nμ​→N(0,1)cuando n→∞.
3. **Modelado de la Regresión Binomial Negativa**:
   * **Variable Dependiente YYY**: La variable dependiente YYY se modela usando una distribución binomial negativa, la cual es adecuada para datos de conteo con sobredispersión (varianza mayor que la media). La función de probabilidad para YYY es: P(Y=k)=(k+r−1k)(1−p)rpk,k=0,1,2,…P(Y = k) = \binom{k + r - 1}{k} (1 - p)^r p^k, \quad k = 0, 1, 2, \ldotsP(Y=k)=(kk+r−1​)(1−p)rpk,k=0,1,2,… donde rrr es el número de éxitos y ppp es la probabilidad de éxito.
4. **Exceso de Ceros**:
   * **Simulación de Exceso de Ceros**: Para simular un exceso de ceros, se añade una proporción de ceros adicionales a los datos generados. Esto se puede representar matemáticamente como: Y={0con probabilidad πYNBcon probabilidad 1−πY = \begin{cases} 0 & \text{con probabilidad } \pi \\ Y\_{\text{NB}} & \text{con probabilidad } 1 - \pi \end{cases}Y={0YNB​​con probabilidad πcon probabilidad 1−π​ donde π\piπ es la proporción de ceros inflados y YNBY\_{\text{NB}}YNB​ sigue una distribución binomial negativa.
5. **Transformación Lineal y Cuadrática**:
   * **Modelado Lineal y Cuadrático**: En mi tesis, se consideran tanto modelos lineales como cuadráticos de la forma: log⁡(μ)=β0+β1X1+β2X2(lineal)\log(\mu) = \beta\_0 + \beta\_1 X1 + \beta\_2 X2 \quad \text{(lineal)}log(μ)=β0​+β1​X1+β2​X2(lineal) log⁡(μ)=β0+β1X1+β2X2+β3X12+β4X22(cuadraˊtico)\log(\mu) = \beta\_0 + \beta\_1 X1 + \beta\_2 X2 + \beta\_3 X1^2 + \beta\_4 X2^2 \quad \text{(cuadrático)}log(μ)=β0​+β1​X1+β2​X2+β3​X12+β4​X22(cuadraˊtico) donde μ\muμ es la media esperada de la distribución binomial negativa. Estas formas funcionales permiten explorar cómo las relaciones no lineales entre las variables explicativas y la variable dependiente pueden capturar mejor las características de los datos.
6. **Interpretación Estadística y Relevancia Empírica**:
   * **Control y Reproducibilidad**: El uso de variables normalmente distribuidas permite un control preciso sobre los experimentos de simulación y garantiza la reproducibilidad de los resultados, lo cual es crucial para validar y generalizar las conclusiones del modelo propuesto.

**Referencias**:

* Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. John Wiley & Sons.
* Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.
* Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.

# Demostración Matemática y Justificación

#### 1. Propiedades de la Distribución Normal

Consideremos las variables explicativas X1X\_1X1​ y X2X\_2X2​ con distribuciones normales:

X1∼N(0,1)yX2∼N(5,2).X\_1 \sim N(0, 1) \quad \text{y} \quad X\_2 \sim N(5, 2).X1​∼N(0,1)yX2​∼N(5,2).

#### 2. Teorema Central del Límite

El Teorema Central del Límite (TCL) establece que, dada una serie de variables aleatorias independientes e idénticamente distribuidas (i.i.d.) XiX\_iXi​ con media μ\muμ y varianza σ2\sigma^2σ2, la suma de estas variables, debidamente normalizada, tiende a una distribución normal cuando el número de variables tiende a infinito:

∑i=1nXi−nμnσ2→N(0,1)cuando n→∞.\frac{\sum\_{i=1}^n X\_i - n\mu}{\sqrt{n\sigma^2}} \rightarrow N(0, 1) \quad \text{cuando } n \rightarrow \infty.nσ2​∑i=1n​Xi​−nμ​→N(0,1)cuando n→∞.

#### 3. Variable Dependiente YYY y Distribución Binomial Negativa

La variable dependiente YYY sigue una distribución binomial negativa, que es adecuada para modelar datos de conteo con sobredispersión. La función de probabilidad de la distribución binomial negativa es:

P(Y=k)=(k+r−1k)(1−p)rpk,k=0,1,2,…P(Y = k) = \binom{k + r - 1}{k} (1 - p)^r p^k, \quad k = 0, 1, 2, \ldotsP(Y=k)=(kk+r−1​)(1−p)rpk,k=0,1,2,…

donde rrr es el número de éxitos y ppp es la probabilidad de éxito.

#### 4. Modelado del Exceso de Ceros

Para simular un exceso de ceros en los datos, se introduce una proporción π\piπ de ceros adicionales. Formalmente, esto puede representarse como una mezcla de distribuciones:

Y={0con probabilidad πYNBcon probabilidad 1−πY = \begin{cases} 0 & \text{con probabilidad } \pi \\ Y\_{\text{NB}} & \text{con probabilidad } 1 - \pi \end{cases}Y={0YNB​​con probabilidad πcon probabilidad 1−π​

donde YNBY\_{\text{NB}}YNB​ sigue una distribución binomial negativa.

#### 5. Modelo de Regresión

Consideremos los modelos de regresión lineal y cuadrático para capturar las relaciones entre las variables explicativas X1X\_1X1​ y X2X\_2X2​ y la media esperada μ\muμ de la variable dependiente YYY.

* **Modelo Lineal**:

log⁡(μ)=β0+β1X1+β2X2\log(\mu) = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2log(μ)=β0​+β1​X1​+β2​X2​

Aquí, μ\muμ es la media de la distribución binomial negativa, y β0,β1,β2\beta\_0, \beta\_1, \beta\_2β0​,β1​,β2​ son los coeficientes del modelo.

* **Modelo Cuadrático**:

log⁡(μ)=β0+β1X1+β2X2+β3X12+β4X22\log(\mu) = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2 + \beta\_3 X\_1^2 + \beta\_4 X\_2^2log(μ)=β0​+β1​X1​+β2​X2​+β3​X12​+β4​X22​

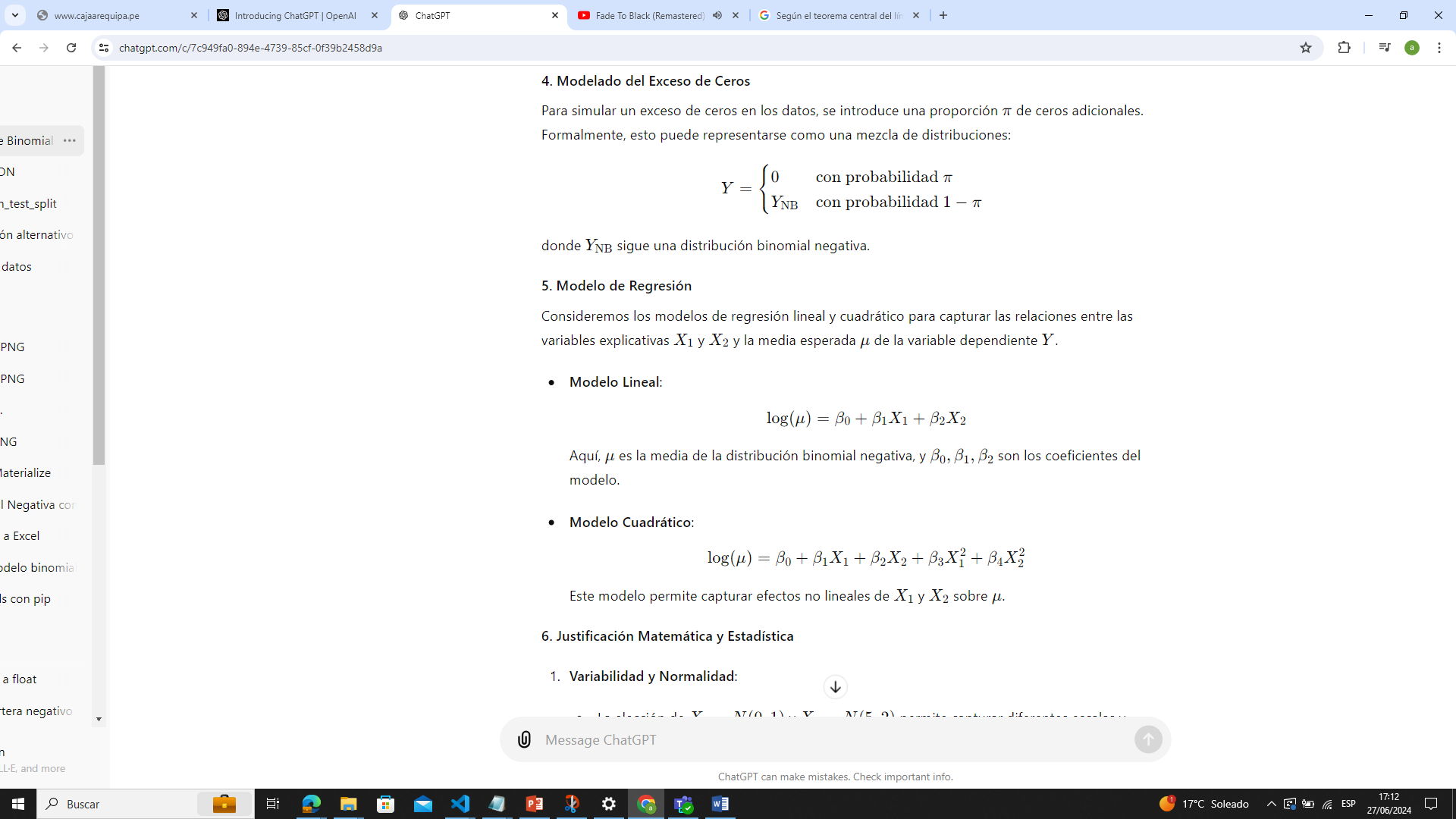
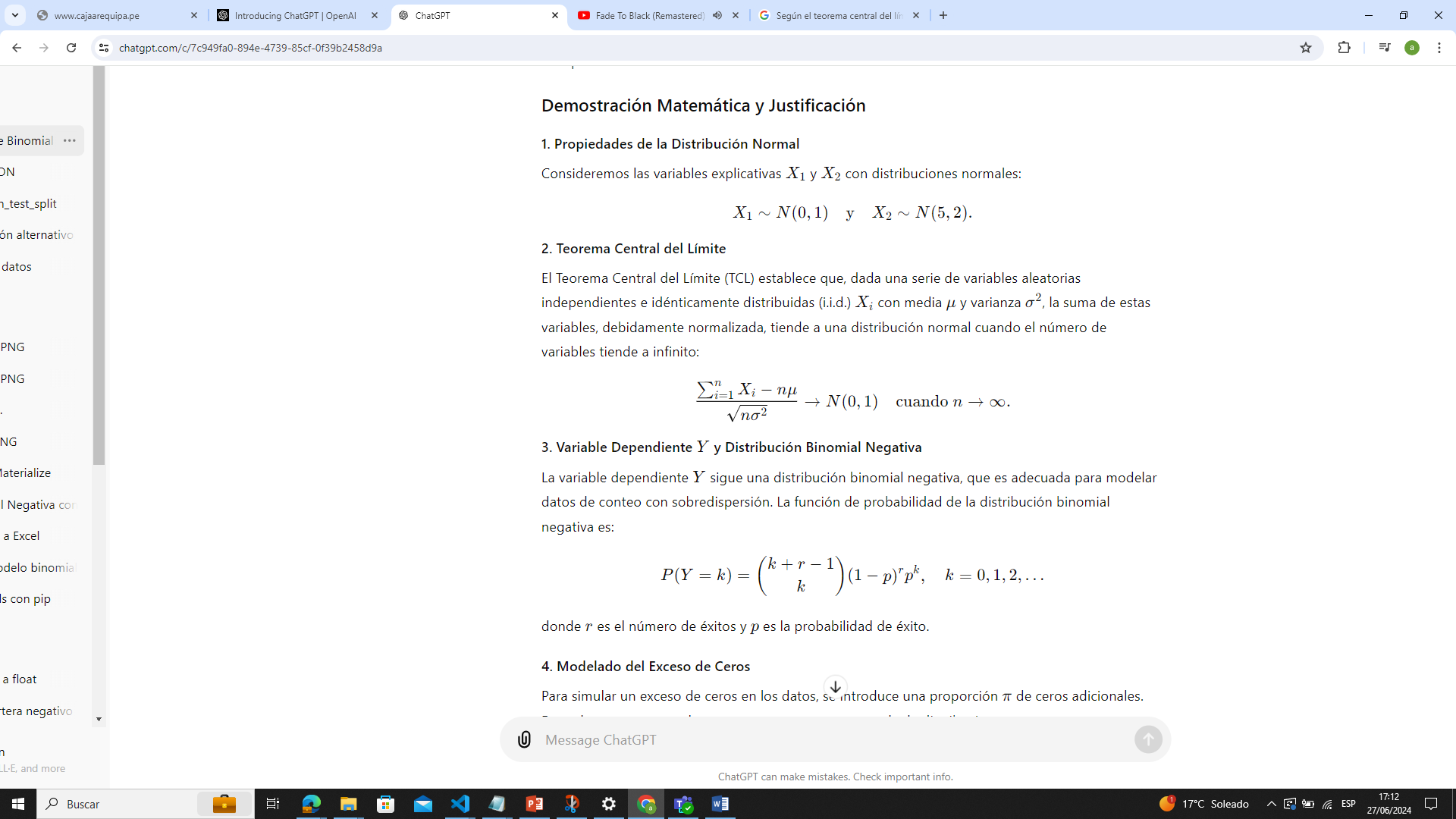
Este modelo permite capturar efectos no lineales de X1X\_1X1​ y X2X\_2X2​ sobre μ\muμ.

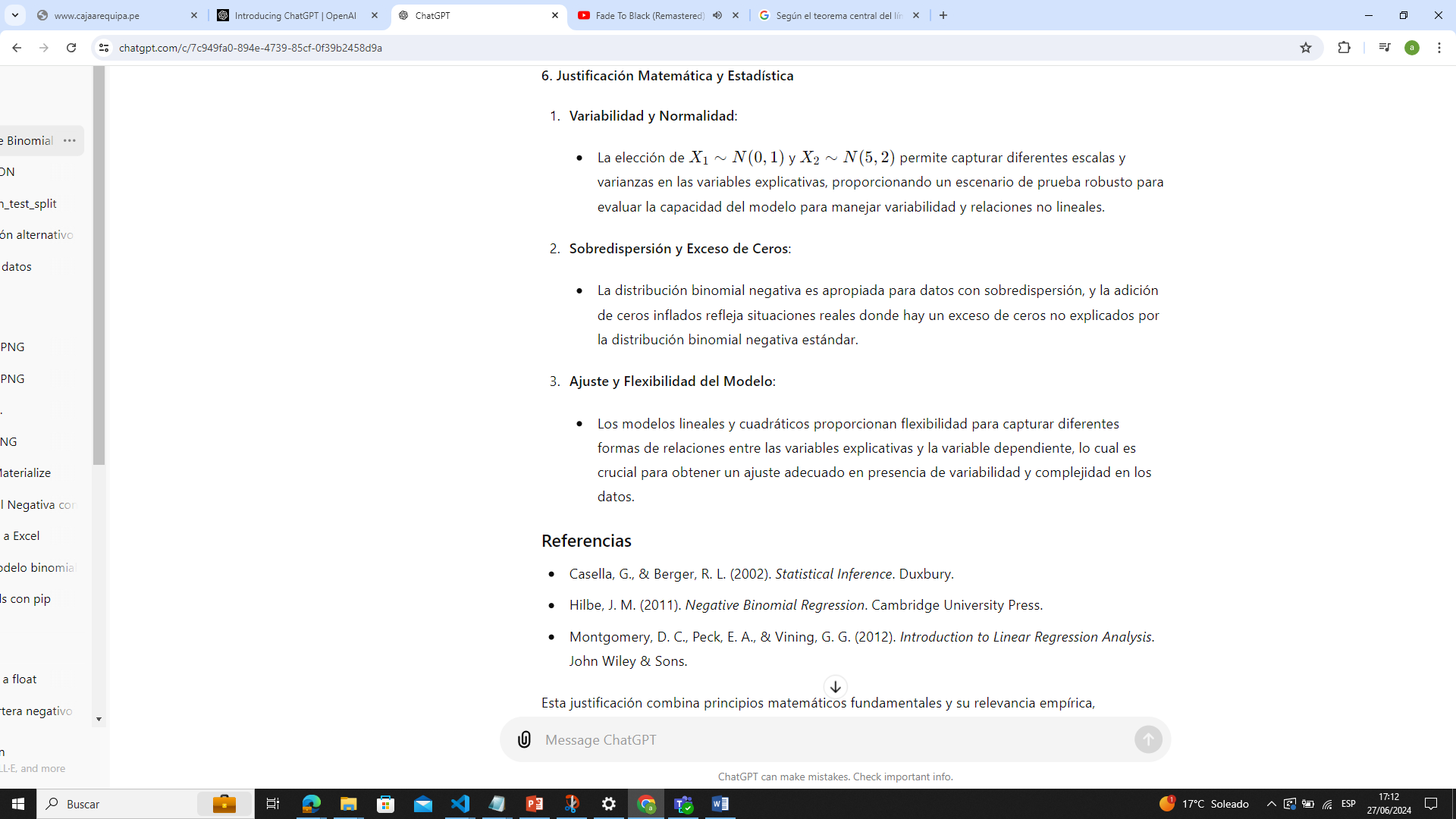
#### 6. Justificación Matemática y Estadística

1. **Variabilidad y Normalidad**:
   * La elección de X1∼N(0,1)X\_1 \sim N(0, 1)X1​∼N(0,1) y X2∼N(5,2)X\_2 \sim N(5, 2)X2​∼N(5,2) permite capturar diferentes escalas y varianzas en las variables explicativas, proporcionando un escenario de prueba robusto para evaluar la capacidad del modelo para manejar variabilidad y relaciones no lineales.
2. **Sobredispersión y Exceso de Ceros**:
   * La distribución binomial negativa es apropiada para datos con sobredispersión, y la adición de ceros inflados refleja situaciones reales donde hay un exceso de ceros no explicados por la distribución binomial negativa estándar.
3. **Ajuste y Flexibilidad del Modelo**:
   * Los modelos lineales y cuadráticos proporcionan flexibilidad para capturar diferentes formas de relaciones entre las variables explicativas y la variable dependiente, lo cual es crucial para obtener un ajuste adecuado en presencia de variabilidad y complejidad en los datos.

### Referencias

* Casella, G., & Berger, R. L. (2002). Statistical Inference. Duxbury.
* Hilbe, J. M. (2011). Negative Binomial Regression. Cambridge University Press.
* Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. John Wiley & Sons.





# entonces en una tesis es válido para el modelado una simulación de datos con 2 variables explicativas y una variable dependiente

Sí, es válido y común en investigaciones académicas y tesis utilizar simulaciones de datos para explorar y validar modelos estadísticos. La simulación de datos te permite controlar las condiciones bajo las cuales evalúas tu modelo y te proporciona un medio para verificar su comportamiento y desempeño. A continuación, se describe cómo estructurar esta parte en tu tesis.

**Justificación del Uso de Simulaciones en la Tesis**

1. **Propósito de la Simulación**:
   * **Control y Repetibilidad**: Al simular datos, puedes controlar completamente las características de los datos y garantizar que las condiciones del experimento sean repetibles. Esto es esencial para validar la robustez del modelo.
   * **Validación Teórica**: La simulación te permite explorar las propiedades teóricas del modelo bajo diferentes escenarios, asegurando que se comporta como se espera en situaciones ideales.
   * **Ajuste y Comparación**: Puedes ajustar el modelo a los datos simulados y comparar los resultados con las expectativas teóricas, proporcionando una validación empírica de los métodos estadísticos utilizados.
2. **Descripción de la Simulación**:
   * **Variables Explicativas**: Dos variables explicativas X1X\_1X1​ y X2X\_2X2​ se generaron con distribuciones normales. Esta elección es común debido a las propiedades de la distribución normal y su prevalencia en datos reales.
     + X1∼N(0,1)X\_1 \sim N(0, 1)X1​∼N(0,1): Captura una variable centrada en 0 con varianza unitaria.
     + X2∼N(5,2)X\_2 \sim N(5, 2)X2​∼N(5,2): Introduce una variable con una media diferente y mayor varianza, permitiendo evaluar cómo el modelo maneja diferentes escalas y dispersiones.
   * **Variable Dependiente**: La variable dependiente YYY se generó utilizando una distribución binomial negativa con parámetros r=2r = 2r=2 y p=0.5p = 0.5p=0.5. Para simular el exceso de ceros, se añadió una proporción π\piπ de ceros adicionales.
3. **Modelo de Regresión**:
   * **Modelo Lineal**: log⁡(μ)=β0+β1X1+β2X2\log(\mu) = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2log(μ)=β0​+β1​X1​+β2​X2​
   * **Modelo Cuadrático**: log⁡(μ)=β0+β1X1+β2X2+β3X12+β4X22\log(\mu) = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2 + \beta\_3 X\_1^2 + \beta\_4 X\_2^2log(μ)=β0​+β1​X1​+β2​X2​+β3​X12​+β4​X22​
   * Ambos modelos se ajustan para evaluar cómo las variables explicativas afectan la variable dependiente en presencia de sobredispersión y exceso de ceros.

**Metodología de la Simulación**

1. **Generación de Datos**:
   * Se simularon n=1000n = 1000n=1000 observaciones de X1X\_1X1​ y X2X\_2X2​ utilizando las distribuciones normales especificadas.
   * La variable dependiente YYY se generó como sigue:
     + Se calcularon las medias esperadas μi\mu\_iμi​ para cada observación usando el modelo especificado.
     + Se generaron valores YiY\_iYi​ siguiendo una distribución binomial negativa con media μi\mu\_iμi​ y dispersión ϕ\phiϕ.
     + Se introdujeron ceros adicionales con una probabilidad π\piπ.
2. **Ajuste del Modelo**:
   * Los datos simulados se utilizaron para ajustar los modelos lineal y cuadrático utilizando métodos de máxima verosimilitud.
   * Se evaluó el ajuste del modelo y la significancia de los coeficientes utilizando pruebas estadísticas estándar.
3. **Resultados y Análisis**:
   * Se compararon los coeficientes estimados con los valores teóricos para verificar la precisión del modelo.
   * Se evaluaron las métricas de ajuste del modelo, como el logaritmo de la verosimilitud y los criterios de información (AIC, BIC).
   * Se analizaron los residuales y la calidad del ajuste para detectar cualquier violación de los supuestos del modelo.

**Conclusiones**

El uso de simulaciones en tu tesis es una estrategia válida y poderosa para explorar y validar modelos estadísticos. Permite una evaluación controlada y detallada de las propiedades del modelo, asegurando que los métodos aplicados son adecuados y robustos bajo diversas condiciones. Esta metodología te proporciona una base sólida para tus conclusiones y justifica el uso de técnicas estadísticas avanzadas en tu investigación.

**Referencias**

* Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Duxbury.
* Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.
* Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Este enfoque no solo es científicamente sólido, sino que también muestra una comprensión profunda de la metodología estadística y su aplicación en investigación.

Sí, es válido y común en investigaciones académicas y tesis utilizar simulaciones de datos para explorar y validar modelos estadísticos. La simulación de datos te permite controlar las condiciones bajo las cuales evalúas tu modelo y te proporciona un medio para verificar su comportamiento y desempeño. A continuación, se describe cómo estructurar esta parte en tu tesis.

**Justificación del Uso de Simulaciones en la Tesis**

1. **Propósito de la Simulación**:
   * **Control y Repetibilidad**: Al simular datos, puedes controlar completamente las características de los datos y garantizar que las condiciones del experimento sean repetibles. Esto es esencial para validar la robustez del modelo.
   * **Validación Teórica**: La simulación te permite explorar las propiedades teóricas del modelo bajo diferentes escenarios, asegurando que se comporta como se espera en situaciones ideales.
   * **Ajuste y Comparación**: Puedes ajustar el modelo a los datos simulados y comparar los resultados con las expectativas teóricas, proporcionando una validación empírica de los métodos estadísticos utilizados.
2. **Descripción de la Simulación**:
   * **Variables Explicativas**: Dos variables explicativas X1X\_1X1​ y X2X\_2X2​ se generaron con distribuciones normales. Esta elección es común debido a las propiedades de la distribución normal y su prevalencia en datos reales.
     + X1∼N(0,1)X\_1 \sim N(0, 1)X1​∼N(0,1): Captura una variable centrada en 0 con varianza unitaria.
     + X2∼N(5,2)X\_2 \sim N(5, 2)X2​∼N(5,2): Introduce una variable con una media diferente y mayor varianza, permitiendo evaluar cómo el modelo maneja diferentes escalas y dispersiones.
   * **Variable Dependiente**: La variable dependiente YYY se generó utilizando una distribución binomial negativa con parámetros r=2r = 2r=2 y p=0.5p = 0.5p=0.5. Para simular el exceso de ceros, se añadió una proporción π\piπ de ceros adicionales.
3. **Modelo de Regresión**:
   * **Modelo Lineal**: log⁡(μ)=β0+β1X1+β2X2\log(\mu) = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2log(μ)=β0​+β1​X1​+β2​X2​
   * **Modelo Cuadrático**: log⁡(μ)=β0+β1X1+β2X2+β3X12+β4X22\log(\mu) = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2 + \beta\_3 X\_1^2 + \beta\_4 X\_2^2log(μ)=β0​+β1​X1​+β2​X2​+β3​X12​+β4​X22​
   * Ambos modelos se ajustan para evaluar cómo las variables explicativas afectan la variable dependiente en presencia de sobredispersión y exceso de ceros.

**Metodología de la Simulación**

1. **Generación de Datos**:
   * Se simularon n=1000n = 1000n=1000 observaciones de X1X\_1X1​ y X2X\_2X2​ utilizando las distribuciones normales especificadas.
   * La variable dependiente YYY se generó como sigue:
     + Se calcularon las medias esperadas μi\mu\_iμi​ para cada observación usando el modelo especificado.
     + Se generaron valores YiY\_iYi​ siguiendo una distribución binomial negativa con media μi\mu\_iμi​ y dispersión ϕ\phiϕ.
     + Se introdujeron ceros adicionales con una probabilidad π\piπ.
2. **Ajuste del Modelo**:
   * Los datos simulados se utilizaron para ajustar los modelos lineal y cuadrático utilizando métodos de máxima verosimilitud.
   * Se evaluó el ajuste del modelo y la significancia de los coeficientes utilizando pruebas estadísticas estándar.
3. **Resultados y Análisis**:
   * Se compararon los coeficientes estimados con los valores teóricos para verificar la precisión del modelo.
   * Se evaluaron las métricas de ajuste del modelo, como el logaritmo de la verosimilitud y los criterios de información (AIC, BIC).
   * Se analizaron los residuales y la calidad del ajuste para detectar cualquier violación de los supuestos del modelo.

**Conclusiones**

El uso de simulaciones en tu tesis es una estrategia válida y poderosa para explorar y validar modelos estadísticos. Permite una evaluación controlada y detallada de las propiedades del modelo, asegurando que los métodos aplicados son adecuados y robustos bajo diversas condiciones. Esta metodología te proporciona una base sólida para tus conclusiones y justifica el uso de técnicas estadísticas avanzadas en tu investigación.

**Referencias**

* Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Duxbury.
* Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.
* Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Este enfoque no solo es científicamente sólido, sino que también muestra una comprensión profunda de la metodología estadística y su aplicación en investigación.

INDICE FINAL:

**Metodología**

1. **Introducción**
   * Contextualización del problema de investigación.
   * Justificación de la elección del modelo de regresión binomial negativa con exceso de ceros.
2. **Generación de Datos**
   * Explicación de la simulación de datos para variables explicativas y la variable dependiente.
     + Variables Explicativas:
       - X1 ~ N (0, 1)
       - X2 ~ N (5, 2)
     + Variable Dependiente:
       - Simulación con distribución binomial negativa (r = 2, p = 0.5) con adición de ceros.
3. **Análisis Exploratorio de Datos**
   * Visualización de la distribución de la variable dependiente Y.
   * Verificación de la independencia y multicolinealidad de las variables explicativas.
4. **Preparación de los Datos**
   * Manejo de valores atípicos.
   * Tratamiento de datos faltantes (si aplica).
5. **Modelado Estadístico**
   * Ajuste inicial del modelo de regresión binomial negativa.
   * Análisis de varianza con términos lineales y cuadráticos para evaluar ajuste del modelo.
6. **Validación del Modelo**
   * Prueba de normalidad y homocedasticidad de los residuos.
   * Evaluación de la significancia de los coeficientes del modelo.

**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.

Este índice proporciona una estructura clara y detallada para tu metodología y resultados, asegurando que cubras todos los aspectos relevantes de tu investigación sobre la regresión binomial negativa con exceso de ceros y diferentes formas de modelar la varianza.

**Marco Teórico**

1. **Introducción al Modelado Estadístico**
   * Definición y importancia del modelado estadístico en la investigación.
   * Breve descripción de la regresión binomial negativa y su aplicación en problemas con exceso de ceros.
2. **Distribución Binomial Negativa**
   * Explicación detallada de la distribución binomial negativa y sus parámetros (r y p).
   * Contexto de aplicación en situaciones con exceso de ceros.
3. **Modelos de Regresión**
   * Tipos de modelos de regresión utilizados en análisis estadístico.
   * Comparación entre modelos de regresión lineal, logística y binomial negativa.
4. **Exceso de Ceros en Datos**
   * Definición y características del problema de exceso de ceros en datos.
   * Métodos para modelar exceso de ceros en distribuciones estadísticas.
5. **Regresión Binomial Negativa con Exceso de Ceros**
   * Fundamentos teóricos y metodológicos de la regresión binomial negativa con exceso de ceros.
   * Ventajas y limitaciones de este enfoque en la modelización de datos reales.
6. **Modelos de Varianza Lineal y Cuadrática**
   * Descripción de modelos de regresión con diferentes especificaciones de varianza (lineal y cuadrática).
   * Justificación teórica para la inclusión de términos cuadráticos en modelos de regresión.
7. **Aplicaciones y Casos de Estudio**
   * Revisión de estudios previos que han utilizado regresión binomial negativa con enfoque en varianza lineal y cuadrática.
   * Ejemplos de aplicaciones prácticas en diferentes campos científicos.
8. **Software y Herramientas**
   * Descripción de herramientas estadísticas y software utilizados para la implementación y análisis de modelos.
9. **Consideraciones Estadísticas y Validación de Modelos**
   * Discusión sobre criterios de selección de modelos y técnicas de validación.
   * Métodos para evaluar la bondad de ajuste y la significancia de los resultados obtenidos.
10. **Conclusiones del Marco Teórico**
    * Síntesis de los principales conceptos y teorías revisadas en el marco teórico.
    * Importancia del modelo propuesto para abordar problemas específicos de investigación.