



# Modelos Lineales Generalizados (GLM): Guía Completa

Los Modelos Lineales Generalizados (GLM) son una extensión flexible de la regresión lineal que permite trabajar con variables de respuesta que siguen diferentes distribuciones. Esta presentación explora los principales tipos de GLM, sus características distintivas y aplicaciones prácticas.

Analizaremos desde el clásico modelo Gaussiano hasta opciones más especializadas como Gamma, Huber, Poisson, Tweedie y modelos de dos etapas, destacando cuándo utilizar cada uno según las características de nuestros datos.



por Manoel Fernando Gadi

# Modelo Gaussiano (OLS)

## Características

El modelo Gaussiano asume que la variable de respuesta sigue una distribución normal y utiliza la función de enlace de identidad. Es la base de la regresión por Mínimos Cuadrados Ordinarios (OLS) y busca minimizar la suma de los errores al cuadrado.

## Aplicaciones

Ideal para predecir valores numéricos continuos y simétricos como precios de viviendas, peso, ingresos o puntuaciones de exámenes. Su principal ventaja es la facilidad de interpretación y la rapidez de cálculo.

## Limitaciones

Es sensible a valores atípicos y a violaciones de supuestos, especialmente la no normalidad y la heterocedasticidad. Puede simplificar demasiado problemas reales donde los datos no son continuos o existen efectos de interacción.

# Modelo Gamma



## Distribución

Se define para números positivos y tiende a ser asimétrica a la derecha y heterocedástica. La media se vincula a los predictores utilizando variables logarítmicas e inversas.



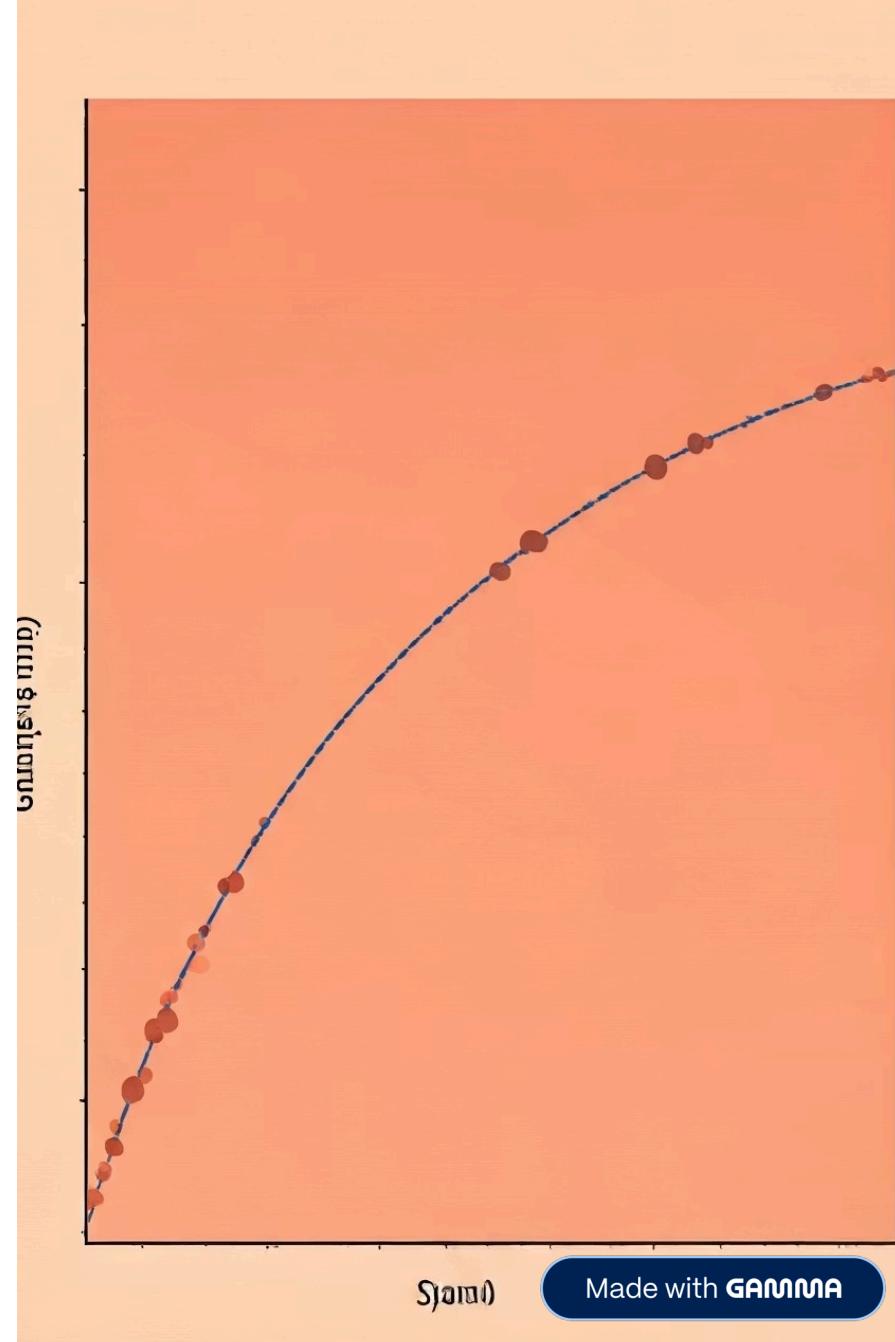
## Casos de Uso

Perfecto para modelar reclamaciones de seguros, montos de transacciones y eventos de tiempo, donde los valores no pueden ser negativos y presentan asimetría positiva.

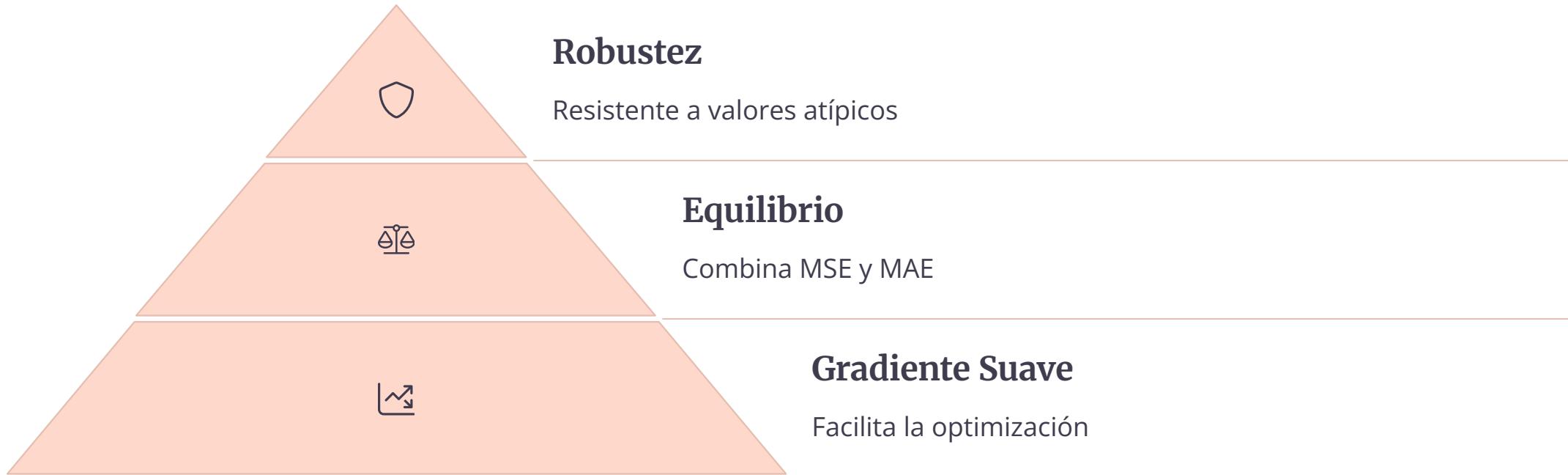


## Limitaciones

La variable de respuesta debe ser estrictamente positiva, no puede manejar valores de 0. Es sensible a valores atípicos debido a la asimetría y presenta dificultad para interpretar los gráficos de residuos.



# Modelo Huber



El modelo Huber utiliza una función de pérdida robusta que equilibra entre el Error Cuadrático Medio (MSE), preciso para residuos pequeños, y el Error Absoluto Medio (MAE), robusto para residuos grandes. Es ideal cuando los datos pueden incluir valores atípicos.

Se aplica en problemas de regresión con ruido como predicción de precios de viviendas o lecturas de sensores. Su principal fortaleza es la mejora de la robustez y generalización, especialmente cuando los datos contienen errores de medición o entradas ruidosas.

# Modelo Poisson

## Características

Modela la probabilidad de un número dado de eventos que ocurren en un intervalo fijo. Asume que los eventos ocurren de forma independiente y que la media y la varianza de la distribución son iguales ( $\lambda$ ). Generalmente presenta una distribución asimétrica y no normal.

## Aplicaciones

Ideal para predecir el número de ocurrencias de un evento cuando el objetivo es un recuento no negativo. Ejemplos típicos incluyen el número de accidentes de tráfico en un lugar o el número de llamadas a un servicio de ayuda.

## Limitaciones

Asume que la media es igual a la varianza, pero los datos reales a menudo muestran sobredispersión. Solo es adecuado para variables objetivo basadas en recuentos y asume independencia de eventos, lo que frecuentemente se viola en la práctica.

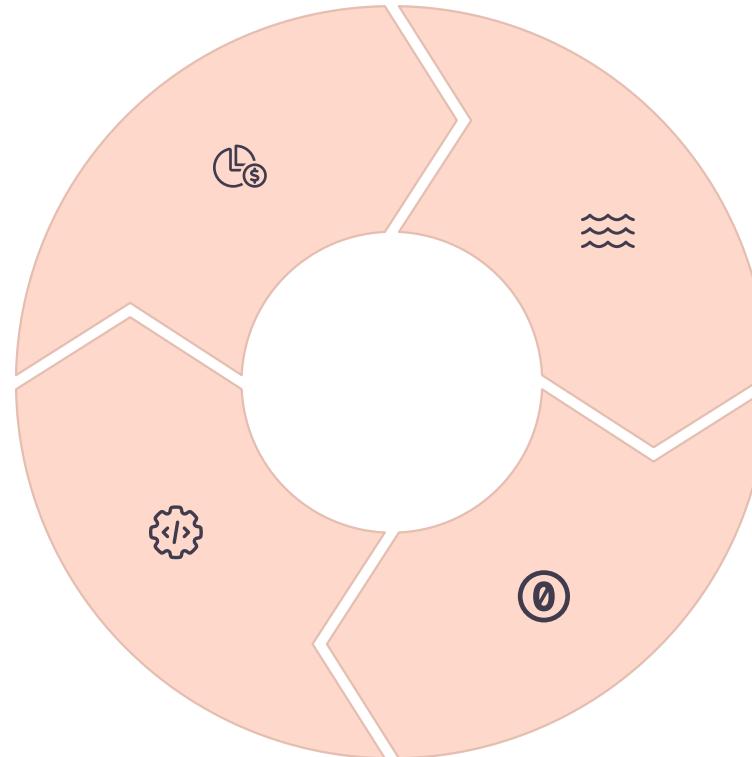
# Modelo Tweedie

## Distribución

Distribución compuesta Poisson-Gamma con parámetro de potencia p

## Desafío

Estimación del parámetro p computacionalmente intensiva



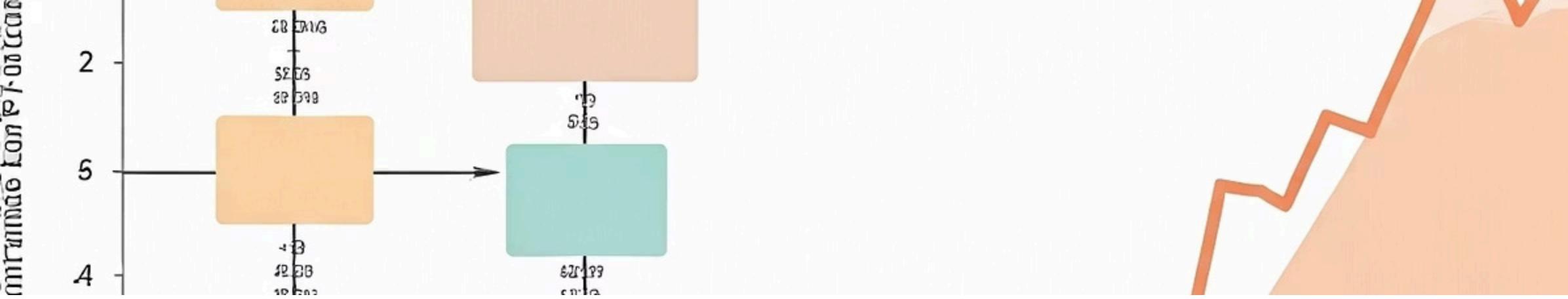
## Aplicaciones

Datos de lluvia, reclamaciones de seguros, costos médicos

## Ventaja

Maneja datos continuos con muchos ceros

El modelo Tweedie es especialmente útil para variables semicontinuas que combinan un gran número de valores cero con valores positivos continuos y asimétricos. Evita la necesidad de crear dos modelos separados para datos inflados en cero, siendo más interpretable y computacionalmente eficiente.

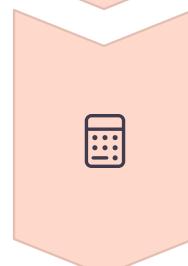


# Modelo de Dos Etapas



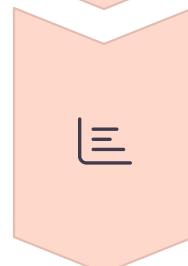
## Primera Etapa: Binomial

Modelo logístico que predice la probabilidad de obtener un valor no nulo ( $P(Y>0)$ ). Determina si habrá algún evento o gasto.



## Segunda Etapa: Gaussiano

Modelo normal que predice el valor esperado dado que es positivo ( $E(Y | Y>0)$ ). Estima la magnitud del evento o gasto.



## Combinación

Las predicciones de ambas etapas se combinan para obtener un valor esperado general, considerando tanto la probabilidad como la magnitud.

Este enfoque es ideal para datos semicontinuos como costos de atención médica, donde muchos sujetos gastan 0 y otros gastan cantidades positivas. Permite que diferentes predictores tengan efectos distintos en la probabilidad y en el nivel del resultado.

**LOBBYING**

Recolección de información  
Identificación de intereses  
Sensibilización de las autoridades  
Sofocación de las críticas



# Comparación de Fortalezas

Modelo	Principal Fortaleza
Gaussiano	Fácil interpretación y cálculo rápido
Gamma	Manejo de datos asimétricos y heterocedasticidad
Huber	Robustez ante valores atípicos con gradiente suave
Poisson	Ideal para modelar resultados basados en recuentos
Tweedie	Combina propiedades discretas y continuas para datos con ceros
Dos Etapas	Flexibilidad para modelar probabilidad y magnitud por separado

# Casos de Uso Prácticos



## Precios de Viviendas

El modelo Gaussiano es ideal para predecir precios de viviendas cuando los datos siguen una distribución aproximadamente normal. Sin embargo, el modelo Huber puede ser más apropiado cuando existen valores atípicos en el mercado inmobiliario.



## Reclamaciones de Seguros

Para modelar reclamaciones de seguros, donde muchos clientes no presentan reclamaciones (ceros) y otros presentan montos variables, los modelos Tweedie o de Dos Etapas son los más adecuados, capturando tanto la frecuencia como la severidad.



## Accidentes de Tráfico

El modelo Poisson es perfecto para predecir el número de accidentes en una intersección, aunque si existe sobredispersión (varianza mayor que la media), podría ser necesario considerar alternativas como la Binomial Negativa.

# Consideraciones Finales



## Análisis de datos

Examinar la distribución y características de la variable objetivo



## Selección del modelo

Elegir el GLM adecuado según la naturaleza de los datos



## Validación

Verificar supuestos y evaluar el rendimiento del modelo

La elección del modelo GLM adecuado depende fundamentalmente de la naturaleza de nuestra variable objetivo. Es esencial analizar si los datos son continuos, discretos o mixtos, si presentan asimetría, valores atípicos o inflación de ceros.

Recordemos que todos los modelos tienen limitaciones. El modelo Gaussiano puede ser demasiado simplista, el Gamma no maneja ceros, el Poisson asume igualdad entre media y varianza, y los modelos más complejos como Tweedie pueden ser computacionalmente intensivos. La validación rigurosa es siempre necesaria.