Modelo de regresión Poisson y Binomial Negativa para la estimación de accidentes de tránsito en la Ciudad de México en el 2023

Carmen Gabriela Angulo Payán
Simulación II, ESFM-IPN, Ciudad de México, México
Teléfono: (55) 3136-0728 E-mail: cangulop1500@alumno.ipn.mx

Resumen - **En la Ciudad de México, los accidentes de tránsito representan un problema significativo de seguridad vial, afectando tanto la movilidad como la seguridad de los habitantes. Este trabajo tiene como objetivo desarrollar un modelo predictivo que estime la frecuencia diaria de accidentes en el año 2023, identificando factores clave como el Tránsito Promedio Diario Anual, el tipo de intersección y la semaforización. Se emplearon los modelos de Regresión de Poisson y Binomial Negativa, ampliamente utilizados en datos de conteo, para abordar la naturaleza discreta de los accidentes y la sobredispersión presente en los datos.**
La base de datos se obtuvo del Portal de Datos Abiertos de la Ciudad de México y comprendió aproximadamente 42,000 registros. Los resultados mostraron que el modelo de Binomial Negativa fue más adecuado, presentando un mejor ajuste, reflejado en una mayor verosimilitud logarítmica y una distribución más uniforme de los residuos. Las variables relacionadas con el TPDA y la semaforización resultaron ser las más influyentes, mientras que otras, como el tipo de intersección, no mostraron significancia estadística.
Se concluye que los modelos predictivos pueden ser herramientas útiles para la planificación de estrategias de seguridad vial basadas en datos y para la toma de decisiones informadas.

Palabras clave - **accidentes de tránsito, Binomial Negativa, modelos de conteo, Poisson, predicción**

Abstract - **In Mexico City, traffic accidents represent a significant road safety problem, affecting both the mobility and safety of the inhabitants. This work aims to develop a predictive model that estimates the daily frequency of accidents in the year 2023, identifying key factors such as Average Daily Annual Traffic, type of intersection, and traffic signalization. Poisson and Negative Binomial Regression models, widely used in count data, were employed to address the discrete nature of accidents and the overdispersion present in the data.**
The database was obtained from the Mexico City Open Data Portal and comprised approximately 42,000 records. The results showed that the Negative Binomial model was more adequate, presenting a better fit, reflected in a higher log likelihood and a more uniform distribution of residuals. The variables related to TPDA and traffic lights proved to be the most influential, while others, such as the type of intersection, did not show statistical significance.
It is concluded that predictive models can be useful tools for data-driven road safety strategy planning and informed decision making.

Key words - **traffic accidents, Negative Binomial, count models, Poisson, prediction**

1. INTRODUCCIÓN

La seguridad vial es un tema prioritario en la Ciudad de México, especialmente ante el incremento en los accidentes de tránsito. Tan solo en 2023, se registraron más de 42,000 incidentes, lo que subraya la urgencia de implementar herramientas predictivas que permitan identificar factores de riesgo y tomar decisiones informadas para reducir su incidencia.

Este artículo tiene como objetivo desarrollar y evaluar el ajuste de modelos de regresión, como Poisson y Binomial Negativa, con el propósito de validar su utilidad en la planificación de estrategias de seguridad vial basadas en evidencia. Para ello, se analizan factores clave como el Tránsito Promedio Diario Anual (TPDA), el tipo de intersección, los fines de semana y la presencia de intersecciones semaforizadas.

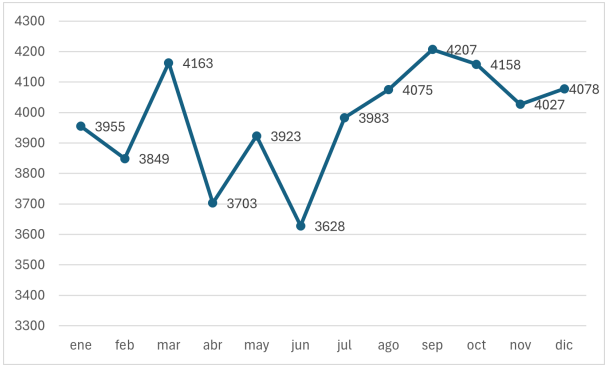

Figura 1: Accidentes del 2023 en la Ciudad de México

2. BASE DE DATOS

Para el desarrollo del modelo, se consultó el Portal de Datos Abiertos de la Ciudad de México, específicamente los hechos registrados por la Secretaría de Seguridad Ciudadana (SSC) (De Innovación Pública, s.f.-b) que comprende información sobre los incidentes ocurridos en toda la Ciudad de México en el 2023. La elección de trabajar con un periodo de un año se debe a que los análisis en seguridad vial frecuentemente abarcan periodos de 1, 2 o 3 años, dependiendo de la naturaleza del estudio y la disponibilidad de datos.

La información contiene aproximadamente 42,000 datos que incluyen variables detalladas relacionadas con los incidentes, como folio, fecha y hora del evento, tipo de evento, tipo de vehículo, ubicación (latitud y longitud), alcaldía, tipo de intersección, entre otras. Estas variables son clave para identificar patrones en los accidentes y modelar su frecuencia. A continuación, se detallan algunas de las variables más relevantes:

Encabezado	**Contenido**
tipo_evento	Atropellado, caída de ciclista, caída de pasajero, choque, derrapado, volcadura
tipo_interseccion	Carriles laterales, cruz, curva, desnivel, gaza, glorieta, ramas multiples, recta, T, Y
interseccion_semaforizada	Si, no
clsificacion_de_la_vialidad	Acceso controlado, eje vial, vac anular, vac radial, vac viaducto, via primaria, via secundaria
tipo_vehiculo	Autobus de pasajeros, automovil, bicicleta, camion de carga, camioneta, metrobús, microbus, monopatín, motocicleta, objeto fijo, taxi, trailer, tren, tren ligero, trolebus

Cuadro 1: Variables principales de la base de datos de la SSC

3. METODOLOGÍA

3.1. *Teoría*

El modelo de regresión de Poisson es un modelo lineal generalizado (GLM) ampliamente utilizado para modelar datos de conteo, como la cantidad diaria de accidentes de tránsito. Este modelo asume que la variable dependiente Y (número de accidentes) sigue una distribución de Poisson. Para transformar la relación no lineal en un modelo lineal, utiliza una función de enlace logarítmica, lo que da lugar a un modelo conocido como log-lineal. La forma matemática general es:

$$log(y) = \alpha + \beta_1 X_1 + \beta_1 X_1 + ... + +\beta_p X_p \tag{1}$$

Donde:

- y: Variable respuesta (acc)
- α: Intercepto del modelo
- $\beta_1, \beta_2, ...$: Coeficientes asociados a las variables predictoras
- $X_1, X_2, ...$: Variables explicativas (e.g., TPDA, tipo de intersección, semaforización)

Los coeficientes (α, β) se estiman utilizando métodos como la Estimación de Máxima Verosimilitud (MLE). En términos prácticos, la regresión de Poisson predice cómo cambia la frecuencia esperada de eventos (y) en función de las variables explicativas, utilizando la fórmula:

$$y = e^{\alpha + \beta x} \tag{2}$$

Una de las características clave de la distribución de Poisson es la equidispersión, es decir, que la media y la varianza de los datos deben ser iguales '$Var(Y) = \mu$'. Sin embargo, en muchos casos reales, como en el presente estudio, la varianza puede exceder la media. Esto se conoce como sobredispersión '$Var(Y) > \mu$' y puede deberse a factores no observados o variabilidad inherente en los datos.

Cuando se detecta sobredispersión en los datos, el modelo de Poisson no es adecuado para realizar predicciones precisas. En estos casos, se utiliza el modelo de Regresión Binomial Negativa, que extiende el modelo de Poisson permitiendo que la varianza condicional de Y sea mayor que la media. Esto es especialmente útil para datos donde los conteos son más variables de lo esperado.

La forma matemática de la distribución Binomial Negativa es:

$$p(Y) = \frac{\Gamma(Y + \alpha^{-1})}{\Gamma(Y+1)\Gamma(\alpha^{-1})} (\frac{\alpha^{-1}}{\alpha^{-1} + \mu})^{\alpha^{-1}} (\frac{\mu}{\alpha^{-1} + \mu}^{Y}), \ \mu > 0, \alpha \geq 0 \tag{3}$$

Donde:

- μ: Media esperada del modelo, que depende de las variables explicativas
- α: Parámetro de dispersión, que controla la relación entre la media y la varianza.
- Γ: Función gamma.

La varianza condicional de Y en este modelo se define como:

$$Var(Y) = \mu + \alpha\mu^2 \tag{4}$$

Esto permite que el modelo capture la variabilidad adicional presente en los datos. En este trabajo, se utilizó la Regresión Binomial Negativa para ajustar los datos de accidentes, ya que se detectó sobredispersión en las frecuencias diarias. Por lo que la ecuación general del modelo de Binomial Negativa es:

$$\begin{aligned} log(acc) = \alpha + \beta_1 \cdot ln(TPDA\_aprox) + \beta_2 \cdot p\_no\_semaforizada + \beta_3 \cdot p\_semaforizada \\ + \beta_4 \cdot dia\_2 + \beta_5 \cdot interseccion\_Y + \beta_6 \cdot interseccion\_T + \beta_7 \cdot interseccion\_X \end{aligned} \tag{5}$$

3.2. *Proceso de análisis*

A partir de la base PRPORCIONADA POR LA ssc, se generó otra para la implementación del modelo, la cual contiene las variables explicativas utilizadas y la variable resultante (acc: número de accidentes). La selección de las variables predictoras se realizó considerando tanto su relevancia teórica como la disponibilidad de datos. Inicialmente, se incluyeron variables relacionadas con el tránsito y las características de las intersecciones, ya que estas influyen directamente en la ocurrencia de accidentes.

Obteniendo un total de 365 datos, correspondientes a un año completo, se consideró esta muestra suficiente para que el modelo trabaje adecuadamente. Las variables explicativas seleccionadas se describen a continuación:

Variable	Descripción
TPDA_aprox	Tránsito Promedio Diario Anual
p_no_semaforizada, p_semaforizada	Proporción de accidetes ocurridos en intersección semaforizada y no semaforizada
interseccion_Y, interseccion_T, interseccion_X	Proporcion de accidentes ocurridos en intersecciones de tipo cruz, T, Y
dia_2	Si los incidentes sucedieron entre semana o fines de semana
acc	Cantidad total de incidentes por dia en la Ciudad de México

Cuadro 2: Variables explicativas

Cabe recalcar que el TPDA fue aproximado debido a la falta de registros oficiales, pero se consideró esencial dada su relación directa con los accidentes. En el análisis exploratorio inicial, se consideraron variables como límites de velocidad y días festivos. Sin embargo, estas fueron descartadas debido a problemas de multicolinealidad reflejados en valores altos de VIF.

Finalmente, las variables seleccionadas se integraron en los modelos de regresión de Poisson y Binomial Negativa, permitiendo evaluar su influencia en la predicción de la frecuencia diaria de accidentes.

3.3. IMPLEMENTACIÓN EN PYTHON

3.3.1. *Biblioteca*

Para este artículo se empleó Python 3.10.12 como lenguaje de programación, utilizando principalmente la biblioteca statsmodels. Esta biblioteca proporciona clases y funciones para la estimación de diversos modelos estadísticos, la realización de pruebas estadísticas y la exploración de datos. Incluye una amplia lista de estadísticas de resultados para cada estimador, las cuales han sido validadas contra paquetes estadísticos existentes para garantizar precisión y consistencia (Daniel, 2023).
Su flexibilidad y fiabilidad lo convierten en una herramienta ideal para el ajuste de modelos de regresión como los de Poisson y Binomial Negativa.

3.3.2. *Implementación*

La implementación de los modelos en Python se llevó a cabo siguiendo los siguientes pasos:

1. Se utilizó la biblioteca pandas para importar la base de datos, transformar las variables categóricas en numéricas y dividir los datos en conjuntos de entrenamiento y prueba (80 % y 20 %, respectivamente).

2. Se establecieron las fórmulas de los modelos de Poisson y Binomial Negativa con las variables seleccionadas.

3. Se ajustaron ambos modelos utilizando la función GLM de la biblioteca statsmodels, evaluando su desempeño mediante métricas como la verosimilitud logarítmica y gráficos de diagnóstico.

4. Los gráficos de residuos y predicciones se generaron con matplotlib para evaluar el comportamiento de los modelos.

4. RESULTADOS

4.1. *Modelos de Poisson y Binomial negativa*

Los modelos ajustados, Poisson y Binomial Negativa, permitieron evaluar la relación entre las variables seleccionadas y la frecuencia diaria de accidentes.

Los resultados obtenidos muestran que en ambos modelos, el TPDA, las intersecciones semaforizadas y no semaforizadas resultaron ser estadísticamente significativas, indicando que estas variables tienen un impacto directo en la ocurrencia de accidentes. Mientras que el tipo de intersección no mostraron relaciones estadísticamente significativas, sugiriendo que el tipo de intersección no influye de manera sustancial en este contexto.

	Poisson		Binomial negativa	
Variable	Coeficiente	P-valor	Coeficiente	P-valor
Intercept	4.3070	0.0	4.3171	0.0
np.log(TPDA_aprox)	0.0650	0.324	0.0687	0.489
p_no_semaforizada	345.8459	0.0	349.4426	0.0
p_semaforizada	340.2321	0.0	343.2924	0.0
dia_2	-0.0014	0.902	-0.0012	0.947
interseccion_Y	-22.0389	0.680	-23.6039	0.775
interseccion_T	5.2268	0.779	4.8127	0.867
interseccion_X	-4.5399	0.858	-2.3692	0.951
Número de observaciones	365		365	
Verosimilitud logarítmica	-985.27		-1102.4	

Cuadro 3: Modelos de predicción de accidentes

Al comparar los dos modelos, la Binomial Negativa mostró un mejor ajuste, reflejado en una mayor verosimilitud logarítmica y un mejor comportamiento en los diagnósticos de residuos.

4.2. GRÁFICOS DE DIAGNÓSTICO

4.2.1. *Gráfico de residuos vs valores ajustados*

El gráfico de residuos frente a valores ajustados del modelo de Binomial Negativa muestra que los residuos están distribuidos de manera más uniforme en torno a la línea cero, aunque todavía se observan ligeras tendencias sistemáticas en valores extremos.

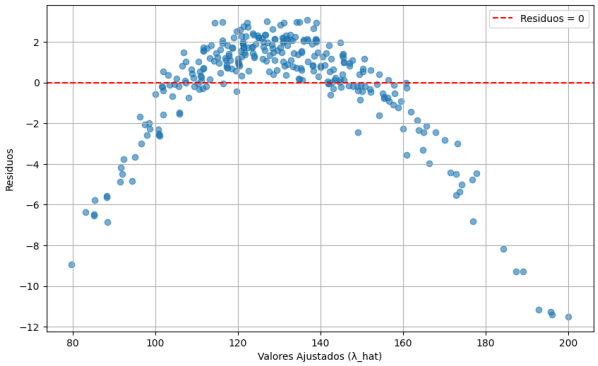

Figura 2: Residuos vs Valores ajustados

Esto sugiere que el modelo captura adecuadamente la mayor parte de la variabilidad de los datos, pero podría mejorarse incluyendo variables adicionales.

4.2.2. *Histograma de Residuos*

El histograma de los residuos para el modelo de Binomial Negativa presenta una distribución centrada en cero, con una ligera asimetría hacia valores positivos.

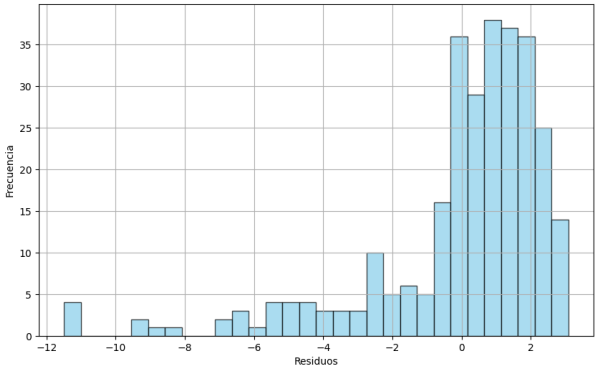

Figura 3: Histograma de residuos

Esto indica que, aunque el modelo ajusta bien los datos, tiende a subestimar algunos valores más altos de frecuencia de accidentes.

4.2.3. *Gráficos de Accidentes Predichos vs Reales*

El gráfico de comparación entre los accidentes reales y las predicciones muestra que el modelo de Binomial Negativa ajusta correctamente los valores promedio. Los puntos están alineados cerca de la línea 1:1, lo que indica que las predicciones son precisas para la mayoría de los casos.

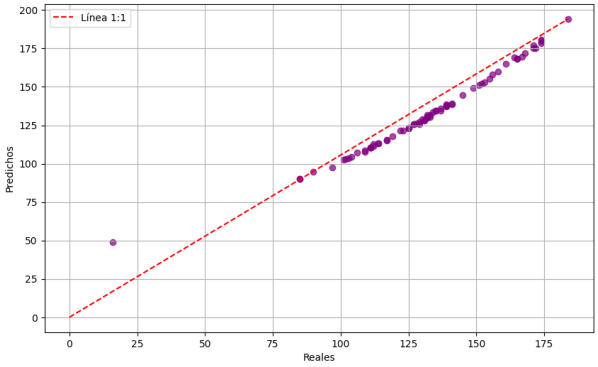

Figura 4: Accidentes predichos vs Reales

Sin embargo, algunos valores extremos presentan discrepancias, lo que podría explicarse por la falta de variables que capturen aspectos específicos, como condiciones climáticas o eventos atípicos.

5. CONCLUSIONES

Este estudio desarrolló modelos predictivos para estimar la frecuencia diaria de accidentes de tránsito en la Ciudad de México para el año 2023, utilizando los modelos de Regresión de Poisson y Binomial Negativa. Los resultados mostraron que el modelo de Binomial Negativa fue más adecuado debido a su capacidad para manejar la sobredispersión en los datos. Este modelo presentó un mejor ajuste, reflejado en una verosimilitud logarítmica más alta y una distribución más uniforme de los residuos.

Entre las variables analizadas, el Tránsito Promedio Diario aproximado y la semaforización demostraron ser factores clave en la ocurrencia de accidentes, mientras que otras, como el tipo de intersección y el día de la semana, no resultaron significativas en este contexto. Esto sugiere que los esfuerzos para reducir accidentes podrían enfocarse en gestionar adecuadamente el volumen vehicular y mejorar la infraestructura de semaforización en intersecciones críticas.

Aunque el modelo logró capturar patrones importantes, su precisión estuvo limitada por la falta de datos específicos, como información climática o condiciones del conductor, así como por la aproximación

del TPDA. Futuras investigaciones podrían abordar estas limitaciones e incorporar más variables para mejorar la capacidad predictiva de los modelos.

En conclusión, los modelos de conteo como la Binomial Negativa son herramientas útiles para el análisis de accidentes y pueden contribuir a la planeación de estrategias de seguridad vial basadas en evidencia, ayudando a reducir el impacto de los accidentes en la Ciudad de México.

Referencias

[1] Código fuente del modelo de predicción de accidentes. Disponible en: `https://github.com/Gabriela2685/Articulo/blob/533bb8fe69bc87cc440374554cec807ca8a71949/C%C3%B3digo_accidentes_2023.ipynb`. Último acceso: 12 de enero de 2025.

[2] Daniel. (2023, 30 octubre). Statsmodels: todo acerca de la biblioteca de Python. Formación En Ciencia de Datos | DataScientest.com. https://datascientest.com/es/statsmodels-todo-acerca

[3] De Innovación Pública, A. D. (s. f.-b). Portal de datos abiertos de la CDMX. https://datos.cdmx.gob.mx/dataset/hechos-de-transito-reportados-por-ssc-base-ampliada-no-comparativa

[4] Elvik, R. (2007). State-of-the-art approaches to road accident black spot management and safety analysis of road networks. TØI Report 883/2007. Oslo, Norway: Institute of Transport Economics. ISBN: 978-82-480-0738-5.

[5] Statgraphics. (2007). Regresión Binomial Negativa. Statgraphics Manual (Revisión del 25 de abril de 2007).

[6] Rivera, J. I. y. G. (2021, 30 abril). Chapter 8 Regresión de Poisson | Modelos lineales generalizados con R. https://bookdown.org/jaimeisaacp/bookglm/regresi%C3%B3n-de-poisson.html