UNIVERSIDAD NACIONAL DE INGENIERIA

FACULTAD DE INGENIERIA ECONOMICA Y
CIENCIAS SOCIALES

"PRONOSTICO DE ACCIDENTES DE TRANSITO: APLICACIÓN DE REGRESION CON DATOS POISSON"

PARA OPTAR EL TITULO PROFESIONAL DE:

LICENCIADO EN ESTADISTICA

POR LA MODALIDAD DE TESIS

ELABORADO POR:

MAGEN DANIELLE INFANTE ROJAS

LIMA – PERU
2004

Índice General

Capítulo 1

Introducción

La mayor parte de las aplicaciones didácticas en Modelos Lineales o temas relacionados a ésta área como Regresión Lineal, hace uso del Modelo Lineal Clásico $Y = X\beta + \epsilon$, también conocido como Modelo de Regresión Lineal Clásico. Para poder desarrollar aplicaciones en el Modelo de Regresión Lineal Clásico, es necesário tener indicios suficientes para aceptar una serie de suposiciones, entre ellas que la variable respuesta Y tiene una distribución normal, esto implica que la variable respuesta es de tipo contínuo. En la práctica, existen muchas aplicaciones que necesitan ser modeladas y donde la variable respuesta Y es de naturaleza distinta a una con distribución normal.

Una situación muy frecuente se presenta cuando la variable respuesta Y que necesita ser estudiada a través de un modelo lineal es de tipo discreto o de conteo. En este caso, la teoría tradicional del Modelo de Regresión Lineal Clásico no es aplicable porque uno de los supuestos de este modelo define a la variable respuesta con distribución normal (variable contínua).

En el presente trabajo de tesis, se estudia un Modelo de Regresión Lineal cuya variable respuesta Y ó variable dependiente será de tipo conteo. Especificamente, observaciones provenientes de una distribución Poisson. Los parámetros desconocidos serán estimados bajo ciertas suposiciones. Se mostrará la utilidad de este modelo con una aplicación, utilizando observaciones de una encuesta sobre accidentes de tránsito en Lima Metropolitana.

Este trabajo, se inicia en el capítulo 2 con la descripción de los conceptos básicos de Regresión Lineal General, la distribución de Poisson como aproximación de la Binomial y sus momentos. Ya en el capítulo 3, se define el Modelo de Regresión Lineal General para observaciones provenientes de una distribución Poisson. También se explican los conceptos de tres formas de

estimación que pueden ser utilizados en la estimación de los parámetros del Modelo de Regresión para Datos Poisson. Con base en éstos métodos, se obtienen los estimadores de los parámetros del modelo utilizando métodos numéricos. Los métodos a desarrollarse son el de Máxima Verosimilitud, Mínimos Cuadrados y Chi Cuadrado Mínimo.

Para obtener un pronóstico de número de accidentes de tránsito, se tomó un particular conjunto de observaciones muestrales obtenidas a través de una encuesta por entrevista a choferes de vehículos de transporte público urbano en Lima Metropolitana. Este tipo de pronóstico para datos de conteo, basados en otras observaciones, muestra la necesidad de diseñar un modelo adecuado como el expuesto anteriormente, y en consecuencia, mostrar la aplicación práctica del mismo. El objetivo del capítulo 4 es realizar un pronóstico sobre Número de Accidentes de Tránsito en Lima Metropolitana. Se definen claramente la población objetivo, las variables, el diseño muestral, se toma la muestra y luego se hace un análisis de los resultados que permiten determinar un modelo adecuado utilizando algunas de las variables explicativas.

El capítulo 5 presenta conclusiones y algunas recomendaciones tanto teóricas como las surgidas de la experiencia obtenida en la elaboración del presente trabajo de tesis. Finalmente, el apéndice contiene todas las tablas y resultados del trabajo.

Capítulo 2

Conceptos Preliminares

Antes de definir el modelo de Regresión General para observaciones con distribución Poisson, este capítulo presenta conceptos básicos de Regresión, propiedades de la Distribución Poisson y algunas de sus propiedades.

2.1 Regresión

Se define el modelo de Regresión General con la siguiente expresión;

$$Y_{ij} = f(X_i, \theta) + \epsilon_{ij}, \qquad i = 1, \cdots, N \quad j = 1, \cdots, n_i \qquad (2.1)$$
$$E(e_{ij}) = 0, \quad E(e_i e_j) = \sigma_{ij}^2$$

donde

Y_{ij}	*variáble aleatória realización particular del experimento, no observáble.*
ϵ_{ij}	*variáble aleatória no observáble*
$X_i = (x_{i1}, ..., x_{im})$	$i-$*ésimo conjunto de variables independientes.*
$\theta = (\theta_1, ..., \theta_p)$	*vector* $p \times 1$ *de parámetros desconocidos.*
n_i	*número de repeticiones de la* $i-$*ésima condición experimental.*

En general, Y_{ij} tiene alguna distribución estadística, la cual es la misma de los errores ϵ_{ij}. La función de regresión $f(X, \theta)$, puede ser lineal o no lineal tanto en las variables independientes X_i, como en los parámetros θ_i desconocidos. La función de regresión, $f(X, \theta)$ relaciona el valor esperado

de la variable dependiente $E(Y_{ij})$ con las variables independientes X_i, y los parámetros θ_i; esto es,

$$E(Y_{ij}) = f(X_i, \theta) \tag{2.2}$$

y, dadas las condiciones experimentales y los datos, se desea estimar los parámetros desconocidos θ.

Si $f(X, \theta)$, es lineal en los parámetros desconocidos, entonces los estimadores de los parámetros se obtendrán usando el análisis tradicional de la Regresión Lineal, lo cual ha sido usado intensamente estudiado en muchos campos de aplicación.

Cuando la función $f(X, \theta)$, es no lineal en al menos uno de los parámetros, entonces se tiene el problema de una regresión no lineal, y los métodos para realizar inferencias estadísticas sobre este modelo son más complicados y por lo general se basan en procedimientos iterativos. Dos procedimientos muy usados son el de máxima verosimilitud, y el de los mínimos cuadrados no lineales. Entre los textos que proporcionan gran cantidad de conceptos sobre éstos procedimientos, se encuentra, Draper and Smith (1989).

2.2 Distribución de Poisson

Esta distribución será explicada considerando primero la Distribución Binomial.

Un experimento puede tener dos resultados mutuamente exclusivos y por lo tanto ser caracterizados por la descomposición simple

$$E = A + \bar{A} \tag{2.3}$$

donde A =Ocurrencia y $\bar{A}$ =No ocurrencia.

Los resultados pueden tener las probabilidades

$$P(A) = p, \qquad P(\bar{A}) = 1 - p = q.$$

El resultado del experimento puede ser representado por una variable aleatoria X_i que toma el valor 1 ó 0 si el evento A ó $\bar{A}$ ocurre, respectivamente. El índice "i"indica los experimentos individuales dentro de la serie.

Si se repite el experimento n veces y se considera la distribución de probabilidad de la variable

$$X = \sum_{i=1}^{n} X_i, \tag{2.4}$$

la probabilidad que k experimentos en particular resulten A y $(n-k)$ $\bar{A}$ es $p^k q^{n-k}$.

También, el evento *k veces A en n experimentos* puede ocurrir en

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{2.5}$$

formas diferentes, de acuerdo a la aparición de la ocurrencia de A y de $\bar{A}$.

Por lo tanto, la probabilidad de este evento es:

$$W_k^n = P(ocurrir\ k\ veces\ A\ de\ n\ experimentos) = \binom{n}{k} p^k q^{n-k} \tag{2.6}$$

Graficamente la distribución W_k^n se muestra en la siguiente figura variando n y p tal que $np = constante$. la figura ayuda a descubrir la similaridad entre la distribución Binomial y la Distribución Poisson.

Figura 2.1

2.2.1 Aproximación de la Distribución Binomial a la Distribución de Poisson

La anterior Figura sugiere que si n tiende a infinito, pero al mismo tiempo $\lambda = np$ permanece constante, la distribución binomial se aproxima a una distribución fija.

Reescribiendo; ($W_k^n = P(ocurrir\ k\ veces\ A\ de\ n\ experimentos)$)

$$\begin{aligned} W_k^n &= \binom{n}{k} p^k q^{n-k} \\ &= \frac{n!}{k!(n-k)!}(\frac{\lambda}{n})^k \frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} \\ &= \frac{\lambda^k}{k!} \frac{n(n-1)(n-2)...(n-k+1)}{n^k} \frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} \\ &= \frac{\lambda^k}{k!}(1-\frac{\lambda}{n})^n \frac{(1-\frac{1}{n})(1-\frac{2}{n})...(1-\frac{k-1}{n})}{(1-\frac{\lambda}{n})^k}. \end{aligned} \tag{2.7}$$

Tomando límite;

$$\lim_{n\to\infty} W_k^n = \frac{\lambda^k}{k!}\exp^{-\lambda} = f(k). \tag{2.8}$$

Siendo esta última la función distribución de probabilidades Poisson.

Definición 2.1 *Sea Y una variáble aleatória que toma los valores posibles: 0,1,..... . Si* $\mathsf{P(Y = k)} = \frac{e^{-\lambda}\lambda^k}{k!}$ $\mathsf{k} = 0, 1, 2, 3, ...$, *entonces se dice que* Y *tiene una distribución de Poisson con parámetro* $\lambda > 0$.

2.2.2 Momentos en la distribución de Poisson

Sea Y_{ij} una variable aleatoria con función de distribución Poisson. Se define los momentos de cualquier función $g(Y)$ con respecto de ' c ' de una variable Y, como los valores medios de las potencias sucesivas de $g(Y)$, es decir, si

$$M^r(g(Y) - c) = Momento - ésimo\ de\ g(Y)\ con respecto\ a\ c,$$

Entonces

$$M^r(g(Y) - c) = E(g(Y) - c)^r$$

En seguida se muestra el cálculo de los cuatro primeros momentos de Y_{ij} con respecto al origen.

Primer Momento (media);

$$\begin{aligned} E(Y_{ij}) &= \sum_{y_{ij}=0}^{\infty} y_{ij} \frac{\lambda^{y_{ij}}}{y_{ij}!} e^{-\lambda} \\ &= \sum_{y_{ij}=1}^{\infty} \lambda \frac{\lambda^{(y_{ij}-1)}}{(y_{ij}-1)!} e^{-\lambda} \\ &= \lambda \sum_{y_{ij}=1}^{\infty} \frac{\lambda^{(y_{ij}-1)}}{(y_{ij}-1)!} e^{-\lambda} \\ &= \lambda \end{aligned} \tag{2.9}$$

que coincide con la esperanza de la variable Y_{ij}.

Segundo Momento;

$$\begin{aligned} E(Y_{ij}^2) &= \sum_{y_{ij}=0}^{\infty} y_{ij}^2 \frac{\lambda^{y_{ij}}}{y_{ij}!} e^{-\lambda} \\ &= \lambda \sum_{y_{ij}=1}^{\infty} y_{ij} \frac{\lambda^{(y_{ij}-1)}}{(y_{ij}-1)!} e^{-\lambda} \\ &= \lambda \sum_{l=0}^{\infty} (l+1) \frac{\lambda^{l}}{l!} e^{-\lambda} \\ &= \lambda (\sum_{l=0}^{\infty} l \frac{\lambda^{l}}{l!} e^{-\lambda} + 1) \\ &= \lambda(\lambda + 1) \end{aligned} \tag{2.10}$$

De este resultado se obtiene la varianza de Y_{ij}

$$Var(Y_{ij}) = E(Y_{ij}^2) - E^2(Y_{ij}) = \lambda(\lambda+1) - \lambda^2 = \lambda.$$

Tercer Momento (asimetría);

$$E(Y_{ij}^3) = \sum_{y_{ij}=0}^{\infty} y_{ij}^3 \frac{\lambda^{y_{ij}}}{y_{ij}!} e^{-\lambda}$$

$$
\begin{aligned}
&= \lambda \sum_{y_{ij}=1}^{\infty} y_{ij}^2 \frac{\lambda^{(y_{ij}-1)}}{(y_{ij}-1)!} e^{-\lambda} \\
&= \lambda \sum_{l=0}^{\infty} (l+1)^2 \frac{\lambda^l}{l!} e^{-\lambda} \\
&= \lambda (\sum_{l=0}^{\infty} l^2 \frac{\lambda^l}{l!} e^{-\lambda} + 2\lambda + 1) \\
&= \lambda(\lambda(\lambda+1) + 2\lambda + 1) \\
&= \lambda(\lambda^2 + 3\lambda + 1) \qquad (2.11)
\end{aligned}
$$

Cuarto Momento (curtosis o apuntalamiento);

$$
\begin{aligned}
E(Y_{ij}^4) &= \sum_{y_{ij}=0}^{\infty} y_{ij}^4 \frac{\lambda^{y_{ij}}}{y_{ij}!} e^{-\lambda} \\
&= \lambda \sum_{y_{ij}=1}^{\infty} y_{ij}^3 \frac{\lambda^{(y_{ij}-1)}}{(y_{ij}-1)!} e^{-\lambda} \\
&= \lambda \sum_{l=0}^{\infty} (l+1)^3 \frac{\lambda^l}{l!} e^{-\lambda} \\
&= \lambda (\sum_{l=0}^{\infty} (l^3 + 3l^2 + 3l + 1) \frac{\lambda^l}{l!} e^{-\lambda} \\
&= \lambda[(\lambda(\lambda^2 + 3\lambda + 1) + 3(\lambda(\lambda+1)) + 3\lambda + 1] \\
&= \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda \qquad (2.12)
\end{aligned}
$$

Capítulo 3

Modelo de Regresión General y Métodos de Estimación para un Modelo con Variable Poisson

En muchas aplicaciones, la naturaleza de los datos pueden haber sido generados por una distribución Poisson. Observaciones de este tipo, podrían ser por ejemplo: número de nacimientos, muertes, matrimonios, divorcios ó accidentes de tránsito, que ocurren en un determinado periodo de tiempo.

3.1 Formulación del Modelo

Sea $Y = (Y_{11}, ..., Y_{N_{n_N}})$, un vector aleatorio con distribución Poisson, para estudiar este vector aleatorio a través de un modelo, se formula el siguiente modelo de regresión general, el cual satisface la ecuación (2.1);

$$Y_{ij} = f(X_i, \theta) + \epsilon_{ij}, \qquad i = 1, 2, ..., N, \quad j = 1, 2, ..., n_i \quad (3.1)$$

Y_{ij}	*variable aleatória de una distribución Poisson con parámetro* $\lambda = E(Y_{ij})$.
$f(X_i, \theta)$	*función de regresión general, lineal o no, diferenciable para* θ.
$X_i = (x_{i1}, ..., x_{ip})$	$(m \leq p)$ $i-$*ésimo conjunto de las* m *variábles no aleatórias, independientes ó explicativas.*
$\theta = (\theta_1, ..., \theta_p)$	*vector* p×1 de parámetros desconocidos a ser *estimados.*
ϵ_{ij}	*perturbaciones aleatórias que deben seguir la ley de distribución de* Y_{ij}.

Un caso particular, y muy usado en las aplicaciones a datos reales, se obtiene cuando la función de regresión general es combinación lineal de los parámetros θ y las variables independientes X_i, esto es;

$$Y_{ij} = f(X_i, \theta) + \epsilon_{ij} = X_i\theta + \epsilon_{ij} = \sum_{k=1}^{p} \theta_k x_{ik} + \epsilon_{ij}, \tag{3.2}$$

Sobre la suposición de que los ϵ_{ij} son ruidos blancos; $\epsilon_{ij} \sim (0, \sigma_\epsilon^2)$, el valor esperado es;

$$E(Y_{ij}) = f(X_i, \theta) = X_i\theta \tag{3.3}$$

y como los Y_{ij} son variables aleatorias Poisson;

$$\lambda = E(Y_{ij}) = f(X_i, \theta) = Var(Y_{ij}) \tag{3.4}$$

donde λ es el parámetro de la distribución Poisson.

En aplicaciones prácticas se buscarán funciones donde $f(X_i, \theta) = X_i\theta$ que expresan la esperanza o función de la esperanza de la variable aleatória observada como una combinación lineal de los parámetros θ y las variables independientes X_i; de la siguiente forma:

$$Y_{ij} = X_i\theta + \epsilon_{ij} \tag{3.5}$$

y consecuentemente

$$E(Y_{ij}) = X_i\theta,$$

la cual se presenta en forma matricial como sigue;

$$E[Y] = \begin{bmatrix} E(Y_{11}) \\ \vdots \\ E(Y_{Nn_N}) \end{bmatrix} = \begin{bmatrix} X_1 \\ \vdots \\ X_{Nn_N} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ & \vdots & \\ X_{Nn_N1} & \cdots & X_{Nn_Np} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$$

donde; $Y_{ij} \sim Poisson(\lambda)$.

La ecuación (3.5) anterior es un Modelo de Regresión Multiple, donde la variable dependiente es un proceso de conteo. Para el proceso de la estimación de los parámetros de este modelo antes, es necesario desarrollar la Función de Verosimilitud de la distribución Poisson, la cual es objeto del presente estudio.

3.2 Función de Verosimilitud de la distribución de Poisson

Usando las propiedades de la distribución de Poisson, descritas en la sección anterior, se desarrolla la función de verosimilitud, a partir del cual se determinarán parámetros máximo verosímiles de $\theta = (\theta_1, ..., \theta_p)$.
Sea $Y = (Y_1', \cdots, Y_N')' = (Y_{11}, \cdots, Y_{1n_1}, \cdots, Y_{N1}, \cdots, Y_{Nn_N})'$, el vector de variables aleatorias observables independientes, provenientes de una distribución Poisson.

La función de verosimilitud es calculada de la densidad conjunta y es considerada como función de los parámetros. Considerar las variables aleatorias anteriores como variables reales $y = (y_1', \cdots, y_N')' = (y_{11}, \cdots, y_{1n_1}, \cdots, y_{N1}, \cdots, y_{Nn_N})'$, así, la función de verosimilitud viene dada por la función de probabilidad conjunta;

$$\begin{aligned} f(y;\lambda) &= f(y_{11}, \cdots, y_{1n_1} \cdots, y_{N1} \cdots, y_{Nn_N}; \lambda) \\ &= \Pi_{i=1}^{N} \Pi_{j=1}^{n_i} f(y_{ij}; \lambda) \\ &= \Pi_{i=1}^{N} \Pi_{j=1}^{n_i} P(Y_{ij} = y_{ij}) \\ &= \Pi_{i=1}^{N} \Pi_{j=1}^{n_i} \frac{e^{-\lambda} \lambda^{y_{ij}}}{y_{ij}!} \\ &= \frac{e^{-\sum_{i=1}^{N} n_i \lambda} \lambda^{\sum_{i=1}^{N} \sum_{j=1}^{n_i} y_{ij}}}{\Pi_{i=1}^{N} \Pi_{j=1}^{n_i} y_{ij}!} \end{aligned} \tag{3.6}$$

haciendo; $\lambda = f(X_i, \theta)$; y $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$ se obtiene;

$$f(y;\lambda) = \frac{e^{-\sum_{i=1}^{N} n_i f(X_i,\theta)} [f(X_i,\theta)]^{\sum_{i=1}^{N} y_{i.}}}{\Pi_{i=1}^{N} \Pi_{j=1}^{n_i} y_{ij}!} \tag{3.7}$$

La razón de esta expresión, substituyendo λ por $f(X_i, \theta)$; es que esta función de versimilitud se ha desarrollado con el fin de estimar los parámetros $\theta = (\theta_1, ..., \theta_p)$ y no el parámetro λ.

Dado que el objetivo es maximizar la función (3.7) en los parámetros $\theta = (\theta_1, ..., \theta_p)$ para hallar sus estimadores, como puede apreciarse, éste no resulta simple de ejecutar debido a la complejidad de la función. Debido a dicha complejidad para la maximización en los parámetros, se utilizarán métodos numéricos, en particular el método de Newton que permitirá hallar estimadores de los parámetros.

3.3 Estimación de los Parámetros

Dado que la anterior expresión no es lineal en los parámetros, se necesitarán procedimientos iterativos para encontrar los estimadores para θ. Entre los métodos alternativos se tienen a tres que son las más usados: el método de *Máxima Verosimilitud*, el de *Mínimos Cuadrados Ponderados* y el método *Chi-Cuadrado Mínimo*, los mismos que pueden proporcionar resultados equivalentes y se describen a continuación.

3.3.1 Máxima Verosimilitud (MV)

Este método es ampliamente usado y se obtiene a partir de la función de verosimilitud. Dicha función de verosimilitud para variables aleatorias de una distribución Poisson ya fue desarrollada en la sección 3.2.

Si se denota por $L(\theta)$ a la función de verosimilitud se obtiene;

$$\begin{aligned} L(\theta) &= \Pi_{i=1}^{N} f(y_i;\theta) \\ &= \Pi_{i=1}^{N}\Pi_{j=1}^{n_i} f(y_{ij};\theta) \end{aligned} \tag{3.8}$$

donde $f(y_{ij};\theta) = \frac{e^{-\lambda}\lambda^{y_{ij}}}{y_{ij}!}$, $\lambda = E(Y_{ij})$ y $Y_{ij} = 0, 1, 2, \cdots$, entonces

$$\begin{aligned} L(\theta) &= \Pi_{i=1}^{N}\Pi_{j=1}^{n_i}\frac{e^{-\lambda}\lambda^{y_{ij}}}{y_{ij}!} \\ &= \frac{\lambda^{\sum_{i=1}^{N}\sum_{j=1}^{n_i} y_{ij}} e^{-\sum_{i=1}^{N} n_i\lambda}}{\Pi_{i=1}^{N}\Pi_{j=1}^{n_i} y_{ij}!} \\ &= \frac{\lambda^{\sum_{i=1}^{N} y_{i.}} e^{-\sum_{i=1}^{N} n_i\lambda}}{\Pi_{i=1}^{N}\Pi_{j=1}^{n_i} y_{ij}!} \\ &= \frac{[f(X_i,\theta)]^{\sum_{i=1}^{N} y_{i.}} e^{-\sum_{i=1}^{N} n_i f(X_i,\theta)}}{\Pi_{i=1}^{N}\Pi_{j=1}^{n_i} y_{ij}!}, \end{aligned} \tag{3.9}$$

donde $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$.

Tomando logaritmo

$$\ln L(\theta) = \sum_{i=1}^{N} y_{i.}\ ln(\lambda) - \lambda\sum_{i=1}^{N} n_i - \sum_{i=1}^{N}\sum_{j=1}^{n_i} ln(y_{ij}!). \tag{3.10}$$

Despreciando la constante que no involucra los parámetros,

$$\ln L(\theta) \propto \sum_{i=1}^{N} y_{i.} ln(\lambda) - \lambda \sum_{i=1}^{N} n_{i.}$$

Substituyendo λ por $f(X_i, \theta)$

$$lnL(\theta) \propto \sum_{i=1}^{N} y_{i.} ln[f(X_i, \theta)] - f(X_i, \theta) \sum_{i=1}^{N} n_i$$

El fin para la anterior contrucción de la función de verosimilitud es maximizar la misma con respecto al vector de parámetros, para ello es necesario derivar la ecuación 3.10 con respecto a cada uno de los parámetros e igualándola a cero.

Luego, como $\theta = (\theta_1, ..., \theta_p)$ es un vector, la maximización en θ resulta en las siguientes p ecuaciones máximo verosímiles:

$$G(\theta) = \begin{bmatrix} \frac{\partial \ln L(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln L(\theta)}{\partial \theta_p} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{3.11}$$

En general, las ecuaciones máximo verosímiles son no lineales con respecto a los parámetros desconocidos, como lo es en este caso. Para encontrar una raíz de la ecuación anterior se utilizará el método de *Scoring* para el desarrollo del algoritmo el cual encontrará una raíz.

Del sistema de ecuaciones anterior,

$$G(\theta) = \begin{bmatrix} \frac{\partial \ln L(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln L(\theta)}{\partial \theta_p} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \left(\frac{\partial f(X_i,\theta)}{\partial \theta_1} \{ \frac{y_i}{f(X_i,\theta)} - n_i \} \right) \\ \vdots \\ \sum_{i=1}^{n} \left(\frac{\partial f(X_i,\theta)}{\partial \theta_p} \{ \frac{y_i}{f(X_i,\theta)} - n_i \} \right) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

Para hallar una raíz del vector de parámetros θ, por el método de NEWTON, se deben expresar las anteriores ecuaciones como el producto de una

matriz de datos y un vector que involucra a los parámetros θ, se consigue el siguiente sistema de ecuaciones:

$$\begin{aligned} C(\theta^{\circ})\,\delta^{\circ} &= G(\theta^{\circ}) \\ C(\theta^{\circ})\,(\theta-\theta^{\circ}) &= G(\theta^{\circ}) \end{aligned} \tag{3.12}$$

donde $\delta^{\circ} = (\theta - \theta^{\circ})$ es un vector $p \times 1$ y $\theta^{\circ} = (\theta_1^{\circ}, ..., \theta_p^{\circ})'$, es un conjunto de valores iniciales.

Utilizando el método SCORING, éste define a $C(\theta^{\circ})$ como una matriz información cuyo $rs - ésimo$ elemento tiene la forma:

$$C_{rs} = E\left(-\frac{\partial^2 lnL(\theta)}{\partial\theta_r \partial\theta_s}\right) = E\left(-\frac{\partial G_r(\theta)}{\partial\theta_s}\right). \tag{3.13}$$

Para hallar la matriz información se estudiará la forma que tiene el $rs - ésimo$ elemento, para ello se analiza primero el término dentro de la esperanza en la ecuación (3.13),

$$\begin{aligned} -\frac{\partial G_r(\theta)}{\partial\theta_s} &= -\frac{\partial}{\partial\theta_s}\left[\sum_{i=1}^{n}\left(\{\frac{\partial f(X_i,\theta)}{\partial\theta_r}\{\frac{y_{i.}}{f(X_i,\theta)} - n_i\}\right)\right] \\ &= -\sum_{i=1}^{n}\left[\left(\frac{\partial^2 f(X_i,\theta)}{\partial\theta_r \partial\theta_s}\{\frac{y_{i.}}{f(X_i,\theta)} - n_i\}\right) - \sum_{i=1}^{n}\frac{\partial f(X_i,\theta)}{\partial\theta_r}\left(\frac{y_{i.}\frac{\partial f(X_i,\theta)}{\partial\theta_s}}{f^2(X_i,\theta)}\right)\right] \\ &= -\sum_{i=1}^{n}\left[\frac{\partial^2 f(X_i,\theta)}{\partial\theta_r \partial\theta_s}\left(\frac{y_{i.}}{f(X_i,\theta)} - n_i\right)\right] - \sum_{i=1}^{n}\left[\frac{y_{i.}}{f^2(X_i,\theta)}\frac{\partial f(X_i,\theta)}{\partial\theta_r}\frac{\partial f(X_i,\theta)}{\partial\theta_s}\right] \end{aligned}$$

Tomando esperanza a la última igualdad se tiene,

$$\begin{aligned} &E\left(-\frac{\partial G_r(\theta)}{\partial\theta_s}\right) = \\ &= E\left(-\sum_{i=1}^{n}\left\{\left[\frac{\partial^2 f(X_i,\theta)}{\partial\theta_r \partial\theta_s}\left(\frac{\sum_{j=1}^{n_i} y_{ij}}{f(X_i,\theta)} - n_i\right)\right] + E\left[-\sum_{i=1}^{n}\frac{\partial f(X_i,\theta)}{\partial\theta_r}\left(\frac{y_{i.}\frac{\partial f(X_i,\theta)}{\partial\theta_s}}{f^2(X_i,\theta)}\right)\right]\right\}\right) \\ &= -\sum_{i=1}^{n}\left[\frac{\partial^2 f(X_i,\theta)}{\partial\theta_r \partial\theta_s}\left(\frac{\sum_{j=1}^{n_i} E(y_{ij})}{f(X_i,\theta)} - n_i\right)\right] - \sum_{i=1}^{n}\left[\frac{\sum_{j=1}^{n_i} E(y_{ij})}{f^2(X_i,\theta)}\frac{\partial f(X_i,\theta)}{\partial\theta_r}\frac{\partial f(X_i,\theta)}{\partial\theta_s}\right]. \end{aligned}$$

Tomando $E(Y_{ij}) = f(X_i, \theta)$, como fue definido inicialmente,

$$E\left(-\frac{\partial G_r(\theta)}{\partial \theta_s}\right) = -\sum_{i=1}^{n}\left[\frac{\partial^2 f(X_i,\theta)}{\partial\theta_r \partial\theta_s}(n_i - n_i)\right] - \sum_{i=1}^{n}\left[\frac{n_i}{f(X_i,\theta)}\frac{\partial f(X_i,\theta)}{\partial\theta_r}\frac{\partial f(X_i,\theta)}{\partial\theta_s}\right]$$

$$= -\sum_{i=1}^{n}\left[\frac{\frac{\partial f(X_i,\theta)}{\partial\theta_r}\frac{\partial f(X_i,\theta)}{\partial\theta_s}n_i}{f(X_i,\theta)}\right]$$

$$= -\sum_{i=1}^{n}\left[\frac{p_{ir}p_{is}n_i}{f(X_i,\theta)}\right]$$

$$= -C_{rs}.$$

Por tanto, la matriz información $C(\theta)$ es

$$C(\theta) = [C_{rs}]_{p\times p} = \left[-\sum_{i=1}^{n}\frac{p_{ir}p_{is}n_i}{f(X_i,\theta)}\right]_{p\times p} \qquad r, s = 1, 2, \ldots, p. \tag{3.14}$$

Dado $\delta^\circ = (\theta - \theta^\circ)$ y el sistema de ecuaciones

$$C(\theta^\circ)\,\delta^\circ = G(\theta^\circ)$$

se tiene

$$(\theta - \theta^\circ) = C^{-1}(\theta^\circ)G(\theta^\circ).$$

Luego, el proceso iterativo estará dado por la ecuación:

$$\theta = \theta^\circ + C^{-1}(\theta^\circ)G(\theta^\circ).$$

Evaluando la ecuación en un conjunto de valores iniciales $\theta^\circ = (\theta_1^\circ, ..., \theta_p^\circ)$, el sistema de ecuaciones en $C(\theta^\circ)\delta^\circ = G(\theta^\circ)$ se resuelve para δ° y se obtienen valores nuevos de los parámetros como $\theta^1 = \theta^\circ + \delta^\circ$.

Una ecuación más apropiada es la siguiente:

$$\theta^{h+1} = \theta^h + C^{-1}(\theta^h)G(\theta^h), \tag{3.15}$$

evaluada en $\theta = \theta^h = (\theta_1^h, ..., \theta_p^h)$, el procedimiento se repite hasta alcanzar una solución estable.

3.3.2 Mínimos Cuadrados (MC)

La estimación por el método máximo verosímil necesita asumir que se conoce la función de distribución para los errores. A diferencia de esto, el método de los Mínimos Cuadrados, además de suponer que los errores tienen media y varianza finita, no presupone ninguna propiedad distribucional adicional.

Aunque la utilización del método de los Mínimo Cuadrados no necesita conocer o asumir la distribución, es importante detallar algunos aspectos de este método antes de utilizarlo para estimar los parámetros del modelo en estudio.

Dado el siguiente modelo de Regresión Lineal;

$$\mathbf{Y} = f(X, \theta) + \epsilon$$

Las suposiciones usuales del modelo de Regresión son

$$E(\epsilon) = 0, \quad y \quad V(\epsilon) = \sigma^2 \mathbf{I}$$

Algunas veces, esas suposiciones usuales no son razonables, por eso será necesario modificar el procedimiento de Mínimos Cuadrados cuando

$$V(\epsilon) = \sigma^2 V \tag{3.16}$$

donde V es una matriz $n \times n$ conocida, es decir, que los errores pueden o no estar correlacionados, y que las varianzas de los errores pueden ser iguales o no.

Si V es diagonal pero con los elementos de la diagonal distintos, entonces las observaciones y_{ij} no están correlacionadas pero tienen varianzas distintas. Si alguno de los elementos fuera de la diagonal de V es distinto de cero, entonces las observaciones están correlacionadas.

Si se presenta cualquiera de los casos anteriores, entonces es necesario hacer una transformación del modelo anterior a un nuevo conjunto de observaciones que satisfacen las suposiciones Mínimos Cuadraticas estandar. Luego, se usa Mínimos Cuadrados Ordinarios en los datos transformados.

Como $\sigma^2 V$ es la matriz de covarianza de los errores, entonces V es definida positiva y no singular, luego existe una matriz "K "simétrica no singular $n \times n$, que satisface;

$$V = K'K = KK = KK' \tag{3.17}$$

la matriz K es a menudo llamada la "raíz cuadrada" de V.

Entonces se hace la transformación definiendo las nuevas variables:

$$Z = K^{-1}Y$$

$$H(X,\theta) = K^{-1}f(X,\theta)$$

$$\mu = K^{-1}\epsilon \tag{3.18}$$

Así, el modelo de regresión anterior resulta ser

$$K^{-1}Y = K^{-1}f(X,\theta) + K^{-1}\epsilon. \tag{3.19}$$

Haciendo el cambio de variables

$$Z = H(X,\theta) + \mu \tag{3.20}$$

Ahora los errores en este modelo modificado tienen esperanza nula, es decir;

$$E(\mu) = E(K^{-1}\epsilon) = K^{-1}E(\epsilon) = 0 \tag{3.21}$$

además, la matriz de covarianzas de "μ "es

$$\begin{aligned} V(\mu) &= E\{[\mu - E(\mu)][\mu - E(\mu)]'\} \\ &= E[\mu\mu'] \\ &= E(K^{-1}\epsilon\epsilon' K'^{-1}) \\ &= K^{-1}E(\epsilon\epsilon')K'^{-1} \\ &= \sigma^2 K^{-1}VK'^{-1} \\ &= \sigma^2 K^{-1}KKK^{-1} \\ &= \sigma^2\mathbf{I}. \end{aligned} \tag{3.22}$$

De este modo, los elementos de "μ "tienen media cero, varianza constante y no estan correlacionados.

Se puede observar que los nuevos errores "μ "del modelo transformado satisfacen las suposiciones usuales, por tanto, ya se tienen las condiciones para aplicar Mínimos Cuadrados Ordinarios. Entonces, la función

de Mínimos Cuadrados es;

$$\begin{aligned} S(\theta) &= \mu'\mu \\ &= (Z - H(X,\theta))'(Z - H(X,\theta)) \\ &= (K^{-1}Y - K'^{-1}f(X,\theta))'(K^{-1}Y - K'^{-1}f(X,\theta)) \\ &= (Y - f(X,\theta))'K'^{-1}K^{-1}(Y - f(X,\theta)) \\ &= (Y - f(X,\theta))'V^{-1}(Y - f(X,\theta)) \\ &= \sigma^2 K^{-1}KKK^{-1} \end{aligned}$$

$$= \begin{bmatrix} Y_1 - f(X_1,\theta) \\ \vdots \\ Y_N - f(X_N,\theta) \end{bmatrix}^t V^{-1} \begin{bmatrix} Y_1 - f(X_1,\theta) \\ \vdots \\ Y_N - f(X_N,\theta) \end{bmatrix}$$

(3.23)

Suponiendo n repeticiones del experimento, tomamos $Y_i = Y_{i.}/n$ y como $V = K'K = KK = KK'$ entonces

$$= \begin{bmatrix} \frac{Y_{1.}}{n_1} - f(X_1,\theta) \\ \vdots \\ \frac{Y_{N.}}{n_N} - f(X_N,\theta) \end{bmatrix}^t (KK')^{-1} \begin{bmatrix} \frac{Y_{1.}}{n_1} - f(X_1,\theta) \\ \vdots \\ \frac{Y_{N.}}{n_N} - f(X_N,\theta) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{Y_{1.}}{n_1} - f(X_1,\theta) \\ \vdots \\ \frac{Y_{N.}}{n_N} - f(X_N,\theta) \end{bmatrix}^t K'^{-1}K^{-1} \begin{bmatrix} \frac{Y_{1.}}{n_1} - f(X_1,\theta) \\ \vdots \\ \frac{Y_{N.}}{n_N} - f(X_N,\theta). \end{bmatrix}$$

(3.24)

Al estimador del parámetro $\theta = (\theta_1, ..., \theta_p)'$ se le denomina "Estimador Mínimo Cuadrado Generalizado".

Un caso especial, el cual se estudiará, es cuando los errores no estan correlacionados pero tienen varianzas distintas tal que la matriz de covarianzas es de la forma

$$\sigma^2 V = \sigma^2 \begin{bmatrix} \frac{1}{W_1} & \cdots & 0 \\ & \vdots & \\ 0 & \cdots & \frac{1}{W_N} \end{bmatrix},$$

(3.25)

es decir, $V^{-1} = W$.

Sea $K-1=W^{\frac{1}{2}}$, de la última igualdad de la función Mínimo Cuadrático se tiene;

$$S(\theta)=\begin{bmatrix} \frac{Y_{1.}}{n_1}-f(X_1,\theta) \\ \vdots \\ \frac{Y_{N.}}{n_N}-f(X_N,\theta) \end{bmatrix}^t W'^{\frac{1}{2}}W^{\frac{1}{2}} \begin{bmatrix} \frac{Y_{1.}}{n_1}-f(X_1,\theta) \\ \vdots \\ \frac{Y_{N.}}{n_N}-f(X_N,\theta) \end{bmatrix}$$

$$=\begin{bmatrix} \frac{Y_{1.}}{n_1}-f(X_1,\theta) \\ \vdots \\ \frac{Y_{N.}}{n_N}-f(X_N,\theta) \end{bmatrix}^t \begin{bmatrix} w_1^{\frac{1}{2}} & & 0 \\ & \ddots & \\ 0 & & w_N^{\frac{1}{2}} \end{bmatrix}^t \begin{bmatrix} w_1^{\frac{1}{2}} & & 0 \\ & \ddots & \\ 0 & & w_N^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \frac{Y_{1.}}{n_1}-f(X_1,\theta) \\ \vdots \\ \frac{Y_{N.}}{n_N}-f(X_N,\theta) \end{bmatrix}$$

$$=\begin{bmatrix} w_1^{\frac{1}{2}}(\frac{Y_{1.}}{n_1}-f(X_1,\theta)) \\ \vdots \\ w_N^{\frac{1}{2}}(\frac{Y_{N.}}{n_N}-f(X_N,\theta)) \end{bmatrix}^t \begin{bmatrix} w_1^{\frac{1}{2}}(\frac{Y_{1.}}{n_1}-f(X_1,\theta)) \\ \vdots \\ w_N^{\frac{1}{2}}(\frac{Y_{N.}}{n_N}-f(X_N,\theta)) \end{bmatrix} \tag{3.26}$$

Finalmente, la función Mínimo Cuadrático tiene la forma;

$$S(\theta)=\sum_{i=1}^{N} w_i(\frac{y_{i.}}{n_i}-f(X_i,\theta))^2 \tag{3.27}$$

Si se hace el cambio $z_i=y_{i.}/n_i$, entonces;

$$S(\theta)=\sum_{i=1}^{N} w_i(z_i-f(X_i,\theta))^2 \tag{3.28}$$

donde w_i es proporcional o es un estimador consistente de $Var(z_i)$. El procedimiento de estimación es el llamado "Mínimos Cuadrados Ponderados".

Generalmente, la función $f(X_i,\theta)$ es no lineal en los parámetros desconocidos, por lo tanto se utilizará el método de la Linealización (o Series de Taylor), este método usa resultados de Mínimos Cuadrados Lineales en una sucesión de etapas.

Si se expande la función $f(X_i, \theta)$ en una Serie de Taylor alrededor del punto $\theta^\circ = (\theta_1^\circ, ..., \theta_p^\circ)'$ y se interrumpe la expansión en la primera derivada, se puede decir que aproximadamente, cuando θ es cerrado a θ°;

$$f(X_i, \theta) = f(X_i, \theta^\circ) + \sum_{i=1}^{p} [\frac{\partial f(X_i, \theta)}{\partial \theta_i}]_{\theta=\theta^\circ} (\theta_i - \theta_i^\circ). \tag{3.29}$$

Sea

$$\delta_i^\circ = (\theta_i - \theta_i^\circ)$$

$$P_i^\circ = P_i(\theta^\circ) = [\frac{\partial f(X_i, \theta)}{\partial \theta_i}]_{\theta=\theta^\circ}. \tag{3.30}$$

Reemplazando en $S(\theta)$ se tiene;

$$S(\theta) = \sum_{i=1}^{N} w_i [z_i - f(X_i, \theta^\circ) - P_i^\circ \delta^\circ]^2 \tag{3.31}$$

donde $\delta_i^\circ = (\delta_1^\circ, \ldots, \delta_p^\circ)$ es desconocido y P_i° es la i-ésima fila de $P(\theta^\circ)$

$$P(\theta) = \begin{bmatrix} \frac{\partial f(X_1,\theta)}{\partial \theta_1} & \cdots & \frac{\partial f(X_1,\theta)}{\partial \theta_p} \\ & \vdots & \\ \frac{\partial f(X_N,\theta)}{\partial \theta_1} & \cdots & \frac{\partial f(X_N,\theta)}{\partial \theta_p} \end{bmatrix}.$$

Para encontrar el estimador Mínimo Cuadrado de $\theta = (\theta_1, ..., \theta_p)$ es necesário derivar la ecuación anterior con respecto a θ (para minimizar la función Mínimo Cuadrática). Luego de esto resultarán "p "ecuaciones normales las cuales deben de resolverse para θ.

Después de derivar la función Mínimo Cuadrática aproximada, las ecuaciones normales tienen la forma;

$$\sum_{i=1}^{N} \frac{P_i^\circ}{[z_i - f(X_i, \theta^\circ) - P_i^\circ \delta^\circ]} = 0. \tag{3.32}$$

En esta igualdad, los únicos parámetros desconocidos son $\theta^\circ = (\theta_1^\circ, ..., \theta_p^\circ)$, para obtener un estimador del parámetro corregido θ°, dado un valor inicial θ° se utiliza el proceso iterativo de Gauss Newton.

El primer valor aproximado del estimador θ será $\theta^1 = \theta^\circ + \theta^\circ$. Así se obtiene el siguiente valor inicial con el que continuará el proceso iterativo. El proceso iterativo continuará hasta que algún criterio de convergencia establecido se cumpla.

3.3.3 Mínimo Chi Cuadrado (MCC)

En la siguiente sección se considera la estimación de los parámetros de un modelo de Regresión con datos Poisson utilizándo uno de los métodos de obtención de "Mejores Estimadores Asintóticos Normales"(M.E.A.N.)

Sea $Y_1, Y_2, \cdots, Y_n$ una secuencia de vectores aleatorios independientes ($n_i = n \ \ \forall i = 1, ..., N$), donde

$$y_j = (y_{1j}, \cdots, y_{Nj})'$$

representa la salida de la j-ésima réplica de las condiciones experimentales X_i, $i = 1, \cdots, N$. La distribución de los Y's depende del parámetro $\theta = (\theta_1, \cdots, \theta_p)$, con vector esperanza

$$E(Y/\theta) = F(\theta) = (f(X_1, \theta), \cdots, f(X_n, \theta))$$

matriz de varianzas covarianzas

$$V(\theta) = E[(Y - F(\theta))(Y - F(\theta))'].$$

y

$$Y = [Y_1, \cdots, Y_n] = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ & & \vdots & \\ y_{N1} & y_{N2} & \cdots & y_{Nn} \end{bmatrix}.$$

Sea Z_n el vector aleatorio N-dimensional definido por

$$Z_n = \bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j = \frac{1}{n}\sum_{j=1}^{n}(Y_{1j}, \cdots, Y_{Nj})', \tag{3.33}$$

entonces

$$E(Z_n) = E(\bar{Y}) = \frac{\sum_{j=1}^{n} Y_j}{n} = \frac{n\,F(\theta)}{n}. \tag{3.34}$$

Para muestras grandes, el estadístico;

$$\frac{\sqrt{n}(Z_n - E(Z_n))}{V(\theta)^{1/2}} \tag{3.35}$$

se distribuye asintóticamente normal, entonces la expresión

$$n[Z_n - F(\theta)]'V(\theta)^{-1}[Z_n - F(\theta)]$$

es llamada una Chi Cuadrado y el valor de θ que la minimiza es llamado estimador *Mínimo Chi Cuadrado* (MCC).

En una forma más general, sea $W(Z_n)$ una matriz $p \times p$ positiva definida que depende de Z_n solamente, entonces la forma

$$n[Z_n - F(\theta)]' W(Z_n)[Z_n - F(\theta)]$$

es llamada una *Chi Cuadrado Modificada*, el estimador $\theta(Z_n)$ que minimiza la Chi Cuadrado Modificada con la función $W(Z_n)$ que depende de Z_n solamente y no de θ o de n será llamado el estimador *Mínimo Chi Cuadrado Modificado* de θ.

Para minimizar la última expresión, se deriva ésta con respecto a cada uno de los parámetros e igualando a cero se obtiene;

$$2n \frac{\partial [Z_n - F(\theta)]}{\partial \theta} W(Z_n)[Z_n - F(\theta)] = 2n \left[\frac{\partial F(\theta)}{\partial \theta} \right]' W(Z_n)[Z_n - F(\theta)] = 0, \tag{3.36}$$

en su forma matricial

$$2n \begin{bmatrix} \frac{\partial f(X_1,\theta)}{\partial \theta_1} & \cdots & \frac{\partial f(X_1,\theta)}{\partial \theta_p} \\ & \ddots & \\ \frac{\partial f(X_N,\theta)}{\partial \theta_1} & \cdots & \frac{\partial f(X_N,\theta)}{\partial \theta_p} \end{bmatrix}' W(Z_n) \begin{bmatrix} \frac{\sum_j Y_{1j}}{n} - f(X_1,\theta) \\ \ddots \\ \frac{\sum_j Y_{1j}}{n} - f(X_1,\theta) \end{bmatrix} = 0,$$

y si denotamos la matriz de derivadas por $P(\theta)$,

$$n[P(\theta)]' W(Z_n)[Z_n - F(\theta)] = 0.$$

Si $W(Z_n)$ converge en probabilidad a $V(\theta)^{-1}$ y ciertas condiciones de regularidad se satisfacen, entonces una raíz de la ecuación es un it Mejor Estimador Asintótico Normal.

Para encontrar θ, se expande $F(\theta)$ en una serie de Taylor de primer orden alrededor de un estimador inicial θ° y sustituyendo la aproximación resultante en esa serie se obtiene

$$[P(\theta)]' W(Z_n)[Z_n - F(\theta^\circ) - P(\theta^\circ)\,\delta^\circ] = 0. \tag{3.37}$$

Como las observaciones se distribuyen Poisson, entonces, los elementos de $W(Z_n)$ serán $[n/f(X_i,\theta)]$.

De la ecuación anterior

$$[P(\theta)]'W(Z_n)P(\theta^\circ)\,\delta^\circ] = [P(\theta)]'W(Z_n)[Z_n - F(\theta^\circ)] = 0. \tag{3.38}$$

Resolviendo para θ° se obtiene;

$$\theta^1 = \theta^\circ + \delta^\circ. \tag{3.39}$$

El procedimiento se repite para θ^1,θ^2, etc. hasta alcanzar un criterio de convergencia. Así se obtiene el estimador de θ.

Si los elementos de los Y_j son mutuamente independientes, entonces $V(\theta)$ será una matriz diagonal. Si se toma W(Zn) como la matriz diagonal cuyos elementos son estimadores *consistentes* de las inversas de las varianzas, entonces el procedimiento ya descrito es idéntico al descrito por el método de los Mínimos Cuadrados Ponderados.

Cuando las observaciones tienen una distribución Poisson, entonces los procedimientos iterativos para encontrar estimadores MV, MCP y MCC son computacionalmente equivalentes cuando se emplean los métodos *Scoring, Gauss Newton* y *Mínimo Chi Cuadrado Modificado.*

Capítulo 4

Aplicación al Pronóstico de Accidentes de Tránsito en Lima Metropolitana

El número de accidentes de tránsito es un fenómeno cuya distribución es Poisson con un cierto parámetro λ. El objetivo del presente capítulo es, primero, especificar un modelo de regresión para explicar la frecuencia promedio anual del número de accidentes de tránsito por unidad de transporte en función de otras variables que puedan influir en ella. Luego, estimar el modelo, el que será utilizado para pronosticar número de accidentes de tránsito por unidad de transporte en similares condiciones. De este modo se demuestra que para casos donde las observaciones tienen un comportamiento de Poisson o de conteo, es posible establecer un modelo para la predicción o pronóstico de observaciones futuras o no observables.

El año 1993, se tomaron observaciones, a través de una encuesta por muestreo probabilístico realizada a la población de vehículos de transporte urbano de pasajeros en Lima Metropólitana. Los resultados de la ejecución de la encuesta, procesamiento de los datos y presentación de los resultados; se encuentran en el apéndice.

4.1 Objetivo

Se busca estimar un Modelo de Regresión Lineal para datos Poisson. Utilizar el modelo estimado para elaborar un pronóstico de la frecuencia del número de accidentes de tránsito en Lima Metropolitana de cualquier año en similares condiciones, teniendo como observaciones la frecuencia de

accidentes ocurridos el año 1993 y los valores de las variables que pueden influir ella.

4.1.1 Variables

La Población consiste del total de las unidades de transporte en Lima Metropolitana. Esta población es estratificada en Buses y Camionetas Rurales a ser definidas posteriormente.

La información se obtuvo midiendo variables en la unidad de muestreo, el chofer del vehiculo, seleccionado en la muestra. Se consideran variables posibles de ser causantes de accidentes de tránsito y a opinión del conductor, opinión considerada como valiosa por la experiencia de los propios conductores.

Variable respuesta o dependiente
Y_i : *Nro. de accidentes que tuvo el chofer "i"en 1993.*

Variables explicativas o independientes	
$E = x_{i1}$:	Edad del chofer.
$X = x_{i2}$:	Experiencia conduciendo vehículos de pasajeros.
$C = x_{i3}$:	Nro. de veces que resultó culpable el chofer.
$U = x_{i4}$:	Nro. de veces que resultó no culpable el chofer.
$V = x_{i5}$:	Nro. de accidentes por exceso de velocidad del chofer.
$P = x_{i6}$:	Nro. de accidentes por imprudencia del pasajero.
$F = x_{i7}$:	Nro. de accidentes por imprudencia del chofer.
$G = x_{i8}$:	Nro. de accidentes debidos a Congestión Vehicular.
$O = x_{i9}$:	Nro. de accidentes debidos a Otras Causas.

La variable dependiente, es decir, el número de accidentes, tiene una distribución de Poisson. Se asume que las variables explicativas se distribuyen independientemente. Se busca medir el efecto lineal de todas las variables independientes al número de accidentes ocurridos durante el año.

Luego de obtener la ecuación de regresión estimada, se utilizará este modelo para pronosticar la frecuencia de accidentes para un año en condiciones similares.

4.2 Modelo

Un caso particular del modelo $Y_{ij} = f(X_i, \theta) + \epsilon_{ij}, \quad i = 1, \cdots, N \quad j = 1, \cdots, n_i, \quad E(e_{ij}) = 0,$ y $E(e_i e_j) = \sigma^2_{ij}$, descrito en la ecuación (3.1) es el Modelo de Regresión Lineal General;

$$Y_{ij} = X_i\theta + \epsilon_{ij} = \sum_{k=1}^{p} \theta_k x_{ik} + \epsilon_{ij} \tag{4.1}$$

lineal en los parámetros θ y lineal en las variables independientes X_i.

Y_{ij}	*variáble aleatória con distribución Poisson, experimento observáble.*
ϵ_{ij}	*error aleatorio no observáble*
$X_i = (x_{i1}, ..., x_{im})$	$i-$*ésimo conjunto de observaciones independientes.*
$\theta = (\theta_1, ..., \theta_p)$	*vector* $p \times 1$ *de parámetros desconocidos.*
n_i	*número de repeticiones de la* $i-$*ésima condición experimental.*

En forma matricial, el vector respuesta es $Y = (Y_1, \cdots, Y_N)$, con vector de parámetros $\theta = (\theta_1, \cdots, \theta_p)$ y matriz de observaciones independientes;

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix} = \begin{bmatrix} x_{11}, & \cdots, & x_{1p} \\ \vdots & \cdots & \vdots \\ x_{N1}, & \cdots, & x_{Np} \end{bmatrix}.$$

El modelo $Y = X\theta + \epsilon$ puede ser escrito entonces, en la forma matricial;

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{bmatrix} = \begin{bmatrix} x_{11}, & \cdots, & x_{1p} \\ \vdots & \cdots & \vdots \\ x_{N1}, & \cdots, & x_{Np} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{bmatrix}.$$

Como vimos en el capítulo 3, al utilizar el Método de Máxima Verosimilitud, la estimación de los parámetros en el modelo descrito, se obtiene luego de un proceso iterativo.

Del sistema de ecuaciones en (3.11); $C(\theta^\circ)\,\delta^\circ = G(\theta^\circ)$, para $\delta^\circ = (\theta - \theta^\circ)$, se tiene

$$(\theta - \theta^\circ) = C^{-1}(\theta^\circ)G(\theta^\circ).$$

Luego, el proceso iterativo esta dado por la ecuación:

$$\theta = \theta^{\circ} + C^{-1}(\theta^{\circ})G(\theta^{\circ}).$$

Evaluar la última expresión en un conjunto de valores iniciales $\theta^{\circ} = (\theta_1^{\circ}, ..., \theta_p^{\circ})$.

Como $\delta^{\circ} = (\theta - \theta^{\circ})$, el sistema de ecuaciones en $C(\theta^{\circ})\delta^{\circ} = G(\theta^{\circ})$ se resuelve para δ° y se obtienen valores nuevos de los parámetros como $\theta^1 = \theta^{\circ} + \delta^{\circ}$.

Una ecuación más apropiada es la siguiente:

$$\theta^{h+1} = \theta^h + C^{-1}(\theta^h)G(\theta^h). \tag{4.2}$$

Evaluar en un valor inicial $\theta = \theta^h = (\theta_1^h, ..., \theta_p^h)$ y el procedimiento se repite hasta alcanzar una solución estable.

De la ecuación (3.14) la matriz información $C(\theta)$ es

$$C(\theta) = [C_{rs}]_{p\times p} = \left[-\sum_{i=1}^{n}\left(\frac{\frac{\partial f(X_i,\theta)}{\partial\theta_r}\frac{\partial f(X_i,\theta)}{\partial\theta_s}n_i}{f(X_i,\theta)}\right)\right]_{p\times p} = \left[-\sum_{i=1}^{n}\frac{p_{ir}p_{is}n_i}{f(X_i,\theta)}\right]_{p\times p} \tag{4.3}$$

$r, s = 1, 2, \ldots, p$.

El proceso iterativo será evaluado con el Software Estadístico GLIM, que permite evaluar las iteraciones hasta encontrar el valor para el cual converge la iteración y por lo tanto los estimadores del modelo en estudio.

Una vez hallados los estimadores del modelo, se dispone del modelo estimado y a partir de éste, se desea pronosticar la frecuencia del número de accidentes de tránsito.

4.3 Metodología

4.3.1 Población

La Población objetivo son todas las unidades de Transporte Urbano de Pasajeros conocidos como **Camionetas Rurales** o **Combis**, **Cousters** o **Micros** y **Omnibus**; definidos como Vehículos de Categoría M2-CLASE III, M3-CLASE I y M3-CLASE II respectivamente. Clasificación según el último Reglamento Nacional de Vehículos DS Nro 058-2003-MTC. Vehículos operando en Lima Metropolitana al momento de la investigación.

4.3.2 Definiciones

Vehículo

Medio capaz de desplazamiento, pudiendo ser motorizado o no, que sirve para transportar personas o mercancías.

(i) Vehículos de Categoría M2-Clase III

Son vehículos automotores de cuatro ruedas o más diseñados y construídos para el transporte de pasajeros con más de ocho asientos sin contar el asiento del conductor. El peso bruto vehícular es de 5 toneladas o menos. Vehículos construídos exclusivamente para el transporte de pasajeros sentados. Conocidos en el medio como **Camionetas Rurales** o**Combis.**

(ii) Vehículos de Categoría M3-Clase II

Son vehículos automotores de cuatro ruedas o más diseñados y construídos para el transporte de pasajeros con más de ocho asientos sin contar el asiento del conductor. El peso bruto vehícular es de más de 5 toneladas. Vehículos construídos principalmente para el transporte de pasajeros sentados y también diseñados para permitir el transporte de pasajeros en pie en el pasadizo y/o en un área que no excede el espacio provisto para dos asientos dobles. Identificados comunmente como **Cousters** o **Micros.**

(iii) Vehículos de Categoría M3-Clase I

Son vehículos automotores de cuatro ruedas o más diseñados y construídos para el transporte de pasajeros con más de ocho asientos sin contar el asiento del conductor. El peso bruto vehícular es de más de 5 toneladas. Vehículos construídos con áreas para pasajeros de pie, permitiendo el desplazamiento frecuente de éstos. Son los llamados **Omnibus.**

4.3.3 Marco muestral y fuente

El Marco Muestral está conformado por el registro de todas las rutas autorizadas a operar en Lima Metropolitana el año 1993. Este registro se obtuvo de la Secretaría Municipal de Transporte Urbano de la Municipalidad de Lima. Para los vehículos con categorías M2 -CLASE III, M3-CLASE II y M3-CLASE I, este registro contiene la siguiente información por ruta.

La información obtenida se enuncia a continuación:

1.- Nro. de Registro en la Municipalidad.
2.- Nro. de Ruta y bifurcación.
3.- Nombre de la Empresa de Transportes.
4.- Tipo de vehículo de transporte (Omnibus, Micro ó Camioneta Rural).
5.- Flota por Empresa de Transporte para cada Ruta y bifurcación.
6.- Cono Origen (Ejem. Este-Oeste).
7.- Distrito origen (Ejem. Vitarte-Callao).
8.- Cono destino (Ejem. Este-Oeste).
9.- Distrito destino (Ejem.: Vitarte-Callao).

Todas las informaciones obtenidas del registro o Marco Muestral fueron utilizadas para la selección de la muestra.

4.3.4 Muestreo y método de recolección de la muestra

El Muestreo es probabilístico. La muestra es seleccionada aplicando Muestreo Estratificado con Afijación proporcional al tamaño de la población. Se seleccionaron dos estratos, el primero conformado por todas las unidades de categorías M3-CLASE II y M3-CLASE I (**Omnibus** y **Micros** y el segundo constituído de la categoría M2 -CLASE III (**Camionetas Rurales**). El método de la recolección de datos es mediante una Encuesta aplicando entrevista personal a través de Cuestionario.

ESTRATO I	Omnibus ó Micro	M3-CLASE II y M3-CLASE I
ESTRATO II	Camioneta Rural	M2 -CLASE III

La Tabla 4.1 muestra las Rutas de vehículos del primer estrato: Omnibus ó Microbuses. La Tabla 4.2 contiene las Rutas de vehículos del segundo estrato: Camionetas Rurales ó Combis. Ambos estratos están agrupados por Límites cardinales.

4.3.5 Cálculo del tamaño de la muestra

El número de encuestas realizadas es 335 choferes, que constituye aproximadamente el 4% del parque automotor de Lima Metropolitana. Se toma la muestra seleccionando número de Ruta, no el número de Línea, porque una misma Línea puede tener diferentes Rutas.

Esta población, ha sido dividida en estratos, cada estrato consiste de todas las rutas que ubican sus paraderos en los mismos extremos cardinales, ejemplos: Norte-Norte, Norte-Sur, Sur-Este, etc.

$N = 10,465$ Población Total.
$N1 = 7,035$ Estrato 1 (Buses y Microbuses).
$N2 = 3,430$ Estrato 2 (Camionetas Rurales).

Se puede ver claramente que la proporción entre Buses y Camionetas Rurales es de 2 a 1, es decir, el número total de Camionetas es aproximadamente 1/3 del total de unidades de transporte, y de otro modo, el número total de Buses ó Microbuses es aproximadamente 2/3 del total de unidades de transporte.

Teniendo esto en cuenta se procedió a calcular el tamaño de muestra de unidades de transporte de la población total;

$$n = \frac{N \times K^2 \times (PQ)}{N \times e^2 + K^2 \times (PQ)} \tag{4.4}$$

$$n = \frac{10,465 \times (1.96)^2 \times (0.5)(0.5)}{10,465 \times (0.0525)^2 + (1.96)^2 \times (0.5)(0.5)}$$

donde

$K = 1.96$ Coeficiente de confiabilidad.
$e = 0.0525$ Error permisible.
$P.Q = 0.5 \times 0.5$.

entonces

$$n = 337.22$$

Teniendo el tamaño de la muestra poblacional, se calculó el tamaño de la muestra de cada estrato por *Afijación Proporcional*;

$$n_h = K \times N_h \tag{4.5}$$

para

$$K = \frac{n}{N}$$

donde

n_h = Tamaño de la muestra del estrato h.
N_h = Tamaño de la población del estrato h.
$h = 1, 2$.

Tamaño de la muestra para el estrato I:

$$n_1 = 0.03 \times 7035 = 211.05212,$$

entonces el tamaño de muestra para el estrato II es :

$$n_2 = 125.$$

Para elegir las unidades que conformaron la muestra de cada estrato, se utilizaron tablas aleatorias. En el apéndice se presenta la muestra seleccionada.

4.3.6 Fuente de Información

Los datos mostrados en las Tablas 4.1 y 4.2, exhiben la población de todas las Rutas representadas por los números de las Líneas de Transporte y las diferentes bifurcaciones de los vehículos de transporte urbano en Lima Metropolitana del año 1993.

Esta información está agrupada por los límites cardinales donde se ubican los paraderos de salida y de llegada de las diferentes Rutas.

4.3.7 Diseño del Cuestionario

Se tuvo especial cuidado en las preguntas de tal forma que sean breves de responder, fáciles de entender para los conductores y sobre todo la inclusión de variables cuyas respuestas sean probables de ser respondidas por los conductores para el presente estudio.

El formulario a continuación de las Tabla 4.1 y 4.2 muestra el diseño del cuestionario cuyo contenido consta de diez preguntas entre abiertas y cerradas.

Tabla 4.1.- Rutas de Omnibus o Micros agrupadas por Límites Cardinales: Origen - Destino

N-N	N-C	N-S	N-E	N-O	S-C	S-S	S-E	S-O	E-C	E-E	E-O	O-C	O-O	C-C
84	33	14	5	59	53	130	6	10	115	17	1	83	56	136
84A	33A	15	19	131	57	147	7	68	137	17A	1A	155	56A	256
134	58	65	55	181	70	197	8	191	154	67	1B	161	69	268
134A	86	66	55A	196	72	199	60	191A	162	67A	1C	189	157	
141	100	75	80	283	91	253	73		162A	74	9	195	157A	
237	133	78	90		104		73A		165	74A	11	414		
250	140	85	93		160		95		185	148	11A			
451	140A	89	93A		169		95A		200	148A	48			
	143	138	97		170		98		215	178	51			
	149	138A	135		171		132		220	178A	54			
	166	142	139		175		150		270	201	71			
	167	164	141A		177		151		281	201A	71A			
	182	176	158		177A		153		285	233	76			
	182A	183	180		190		172		403	234	87			
	206	193	180A		205		172A		437	234A	88			
	221	193A	180B		257		270A		439	270	88A			
	223	259A	187		284		408		439A		92			
	223A	278	188		421						92A			
	247	401	188A		433						94			
	248	413	192		433A						146			
	259		192A		434						146A			
	276		198		440						156			
	282		262								156A			
	426		277								184			
	509		280								203			
			400								204			
			405								228			
			417								232			
			442								269			
			444								445			
											453			
											462			

Tabla 4.2.- Rutas de Camionetas Rurales agrupadas por Límites Cardinales: Origen - Destino

N-N	N-C	N-S	N-E	N-O	S-C	S-S	S-E	S-O	E-C	E-E	E-O	O-C	O-O	C-C
109	20	25	407		24	39	40	45	21	218	48	102		
120	30	46	438		27	49	524	435	21A	275	415	105		
186	31	419	500		29	50		435A	21B		454	410		
207	34	504			101	108		517	22		459	431		
208	36	530			116	110			23		505	443		
209	37				124	112			28		507			
216	122				125	114			28A		511			
219	210				243	119			28B					
261	217				406	121			47					
267	255				469	123			107					
425	411				508	224			107A					
451	432				513	226			212					
506					514	231			214					
					514A	258			214A					
					516	271			225					
					518	282			227					
					519	412			227A					
					522	428			238					
					044	429			241					
					523	430			260					
						441			272					
						512			409					
						503			418					
									418A					
									418B					
									420					
									427					
									449					
									501					

Cuestionario

Encuesta de Accidentes de Tránsito en Lima Metropolitana

Línea Encuestada: .....
Ruta: .....Hora Encuesta: .....

1. ¿Cuanto tiempo conduce vehículos de pasajeros?
   (a) [.....] años (b) [.....] meses
2. ¿Qué otro tipo de vehículos ha conducido:?
   (a) Trayler (b) Omnibus (c) Camión (d) Camioneta Rural (e) Auto
3. ¿Cuál es su edad? : [.....]
4. ¿Tiene Brevete?: (a) NO (b) SI Categoría: [.....]
5. ¿En este año ha tenido algún tipo de accidente?
   (a) NO (b) SI (c) Cuántos: [.....]
6. De esos accidentes: ¿Cuántas veces lo declararon culpable o no?,
   (a) Culpable [.....] (b) No Culpable [.....]
7. De los accidentes que ha tenido, Cuántos eran :
   (a) Graves [.....] (b) Serios[.....] (c) Leves [.....]
8. ¿A que causas atribuye los accidentes:?
   (a) Exceso Velocidad
   (b) Congestión vehicular
   (c) Imprudencia del pasajero
   (d) Imprudencia del chofer
   (d) Otro [.....]
9. Mencione el lugar ó lugares en los cuales tuvo el accidente: (Calles, Av., etc.)
   [.....]
   [.....]
10. Puede precisar que tipos de accidentes ha tenido?
    (a) Choque (b) Atropello (c) Caida (d) Apedreo

4.3.8 Observaciones

Existe una real dificultad para obtener información en cuanto al desarrollo del Transporte Urbano en Lima Metropolitana. En primer lugar porque no existen registros unificados de toda la información en una sola institución, con frecuencia se encuentra la información dispersa en una u otra institución. Muchas de las informaciones importantes, no son registradas y almacenadas para llevar estadísticas en el tiempo, por ejemplo los accidentes de tránsito. Por otro lado, la informalidad del transporte contribuye tambien a este tipo de inconvenientes. El aun deficiente libre acceso a la información, en especial cuando son datos deben ser proporcionados por intituciones del estado ó privadas de gran envergadura. Como consecuencia, la obtención de la información para la obtención del Marco Muestral fue una de las dificultades más grandes que se presentó para el presente trabajo de tesis; antes de conseguirlo, se pasó por varios intentos.

Los primeros intentos se dieron cuando se recurrió a la Comandancia de la Policía de Tránsito, allí, después tres o cuatro citas, en gentiles conversaciones, explicaciones y permisos, se pudo obtener datos registrados de accidentes de tránsito indicando valores valiosos para el estudio, pero no suficientes, de los accidentes de tránsito ocurridos en Lima Metropolitana en el primer medio año de 1993, es decir de enero a julio.

Debido a la insuficiencia de los datos se decidió recurrir a alguna autoridad conocida de la misma secretaría para obtener datos más precisos que los anteriores, pero la sorpresa fue grande al recibir la siguiente información: "los datos que se registran en la secretaría representan aproximadamente el 8

Teniendo esto en cuenta, se decidió realizar una encuesta para recabar la información más apropiada estadísticamente. Información necesaria para un modelamiento y posterio pronóstico de Nro. de Accidentes de Tránsito.

Cabe señalar que, aprovechando una afortunada oportunidad, fue posible obtener el para nosotros Marco Muestral, donde se encuentran los datos necesarios para el cálculo y selección de la muestra. Fuente: Comandancia de la Policía de Tránsito de Lima Metropolitana.

4.4 Procesamiento de la Información y Análisis de Resultados

Se procesaron los datos obtenidos luego de aplicar la encuesta. Primero se describen los tipos de variábles recolectadas. Luego, para establecer las relaciones entre la variable de interes Nro de Accidentes y las variables explicativas, se hace un análisis de correlaciones entre éstas para seguidamente descartar las variables que no tendran suficientes indicios de ser relevantes para ser incluídas en el modelo que explique el Nro de Accidentes.

Tanto el procesamiento como los análisis de datos se hicieron por separados para Omnibus-Micros como para Camionetas Rurales. Igualmente se realizó el análisis para todos ambos estratos en su conjunto.

4.4.1 Descripción de Variables

Inicialmente, se describe las variables que intervienen en el análisis inicial de las observaciones muestrales, previo a la especificación del modelo final. Recordando que el objetivo es obtener la estimación de un modelo que se pueda utilizar para el pronóstico de accidentes de tránsito en Lima Metropolitana.

Y: Nro. de accidentes ocurridos en 1993.
Cantidad de accidentes que sufrió el chofer encuestado, de Enero-Diciembre de 1993.

E: Edad del chofer (en meses).
Edad del chofer en el momento de ser encuestado.

X: Experiencia del chofer.
Tiempo de experiencia que tiene el chofer encuestado, manejando unidades de transporte urbano hasta el momento de la encuesta.

C: Culpabilidad del chofer.
De los accidentes sufridos durante el año 1993, en cuántos de esos, el chofer encuestado fue declarado Culpable.

U: No culpabilidad en la ocurrencia de accidentes.
De la cantidad de accidentes sufridos durante el año 1993, en cuántos de esos accidentes, el chofer encuestado fue declarado No culpable.

V: Exceso de velocidad (binario).
De la cantidad de accidentes sufridos en 1993, número de accidentes que se debieron al Exceso de Velocidad.

P: Imprudencia del pasajero (binario).
De la cantidad de accidentes sufridos en 1993, número de accidentes que se debieron a la Imprudencia del Pasajero.

F: Imprudencia del chofer.
De la cantidad de accidentes sufridos en 1993, número de accidentes que se debieron a la Imprudencia del Chofer.

G: Congestión vehicular (binario).
De la cantidad de accidentes sufridos en 1993, número de accidentes que se debieron a la Congestión Vehicular.

O: Otras causas (binario).
De la cantidad de accidentes sufridos en 1993, número de accidentes que se debieron a Otras Causas.

4.4.2 Análisis de Correlaciones

Se realiza un análisis de las correlaciones entre todas las variables involucradas, tanto la variable dependiente como las variables explicativas. Este análisis permite explicar el grado de asociación entre la variable de interés Nro de Accidentes y las variables explicativas anteriormente mencionadas.

Después de detectar las variables explicativas cuya asociación lineal con la variable respuesta es relevante, se elige este conjunto de variables para ser incluídas en el modelo que explique la variable respuesta Nro de Accidentes.

Tabla 4.3 .- Correlaciones con las variables Y, E y X

		.Y		.E		.X
Y	(1,1)	1.	(1,2)	0.024918	(1,3)	0.0525501
E	(2,1)	0.024918	(2,2)	1.	(2,3)	**0.71419**
X	(3,1)	0.0525501	(3,2)	**0.71419**	(3,3)	1.
C	(4,1)	**0.572041**	(4,2)	0.00756	(4,3)	0.059033
U	(5,1)	**0.913927**	(5,2)	0.0260851	(5,3)	0.0336905
V	(6,1)	**0.451942**	(6,2)	-0.00261	(6,3)	-0.050395
P	(7,1)	**0.475982**	(7,2)	0.0734552	(7,3)	0.0945799
F	(8,1)	**0.490431**	(8,2)	-0.089196	(8,3)	-0.05648
G	(9,1)	0.25544	(9,2)	-0.061567	(9,3)	-0.129755
O	(10,1)	0.25242	(10,2)	0.0940568	(10,3)	0.147925

Tabla 4.4 .- Correlaciones con las variables C, U y V

		.C		.U		.V
Y	(1,4)	0.572041	(1,5)	0.913927	(1,6)	0.451942
E	(2,4)	0.00756	(2,5)	0.0260851	(2,6)	-0.00261
X	(3,4)	0.059033	(3,5)	0.0336905	(3,6)	-0.050395
C	(4,4)	1.	(4,5)	0.190	(4,6)	0.336208
U	(5,4)	0.189892	(5,5)	1.	(5,6)	0.374603
V	(6,4)	**0.336208**	(6,5)	**0.375**	(6,6)	1.
P	(7,4)	0.263798	(7,5)	**0.43921**	(7,6)	0.277235
F	(8,4)	0.245188	(8,5)	**0.465715**	(8,6)	0.077137
G	(9,4)	0.0480553	(9,5)	**0.281981**	(9,6)	-0.00434
O	(10,4)	0.166964	(10,5)	**0.219525**	(10,6)	-0.048892

Tabla 4.5 .- Correlaciones con las variables P, F y G

		P.		F.		G.
Y	(1,7)	0.475982	(1,8)	0.490431	(1,9)	0.25544
E	(2,7)	0.0735	(2,8)	-0.089196	(2,9)	-0.0616
X	(3,7)	0.0945799	(3,8)	-0.05648	(3,9)	**-0.129755**
C	(4,7)	0.263798	(4,8)	0.245	(4,9)	0.0480553
U	(5,7)	0.43921	(5,8)	0.465715	(5,9)	**0.281981**
V	(6,7)	0.277235	(6,8)	0.0771	(6,9)	-0.00434
P	(7,7)	1.	(7,8)	-0.00365	(7,9)	-0.043103
F	(8,7)	-0.00865	(8,8)	1.	(8,9)	**-0.117106**
G	(9,7)	-0.043103	(9,8)	-0.117106	(9,9)	1.
O	(10,7)	0.165334	(10,8)	-0.00741	(10,9)	-0.058659

Tabla 4.6 .- Correlaciones con la variable O

		O.
Y	(1,10)	0.25242
E	(2,10)	0.0941
X	(3,10)	**0.147925**
C	(4,10)	**0.166964**
U	(5,10)	**0.219525**
V	(6,10)	-0.0488918
P	(7,10)	**0.165334**
F	(8,10)	-0.00741
G	(9,10)	-0.0586597
O	(10,10)	1.

La Tabla 4.3 muestra que ***todas*** las variables se correlacionan con la variable respuesta Y, pero más correlacionadas que las otras están las variables C, U, V, P **y** F.

Las variables E y X están altamente correlacionadas, por lo cual sólo una de ellas puede entrar al modelo.

Hasta aquí, los conjuntos posibles para el modelo son:

$$[E,\ C,\ U,\ V,\ P,\ F,\ G,\ O, f(E,\ C,\ U,\ V,\ P,\ F,\ G,\ O)]$$

ó

$$[X,\ C,\ U,\ V,\ P,\ F,\ G,\ O, f(E,\ C,\ U,\ V,\ P,\ F,\ G,\ O)]$$

En la Tabla 4.4 las variables C y V están altamente correlacionadas, entonces sólo una de ellas entrará en el modelo.

Sin embargo, la variable C es una de las variables más altamente correlacionadas con la variable Y, por lo cual no se debe excluir del modelo.

Entonces, los posibles conjuntos de variables que pueden estar incluídos en el modelo son:

$$[E,\ C,\ U,\ P,\ F,\ G,\ O, f(E,\ C,\ U,\ P,\ F,\ G,\ O)]$$

ó

$$[X,\ C,\ U,\ P,\ F,\ G,\ O, f(E,\ C,\ U,\ P,\ F,\ G,\ O)]$$

La segunda columna de correlaciones de la Tabla 4.4 muestra que las variables P, y F están altamente correlacionadas con la variable U, pero U

no se puede excluir del modelo por ser la variable más correlacionada con la variable Y. Entonces los conjuntos posibles ahora pueden ser:

$$[E,\ C,\ U,\ G,\ O, f(E,\ C,\ U,\ G,\ O)]$$

ó

$$[X,\ C,\ U,\ G,\ O, f(E,\ C,\ U,\ G,\ O)]$$

De las Tablas 4.5 y 4.6 se concluye que las variables G y O en general tienen mayor correlación con todas las otras variables que E y X, por lo cual se excluyen del modelo y los modelos finales incluirán a los siguientes conjuntos de variables:

$$[E,\ C,\ U,\ f(E,\ C,\ U,\ )]$$

ó

$$[X,\ C,\ U,\ f(E,\ C,\ U,\ )]$$

4.5 Pronóstico del Nro de Accidentes de Tránsitos y Selección de Modelos

En el modelo que explica Nro de Accidentes, se desea descubrir si existe diferencia en un modelo para todo tipo de unidades de transporte urbano y modelos para Buses-Micros o Combis. Por tal motivo, se analizan los tres modelos por separado.

El primer grupo es el total de las unidades, es decir, los llamados Buses, Micros y Combis. En el segundo grupo se estudian Buses y Micros juntos, por ser considerardos unidades de transporte similares en forma y operatividad. En el tercer grupo sólo se estudian las llamadas Camionetas Rurales, más conocidas como Combis; por ser un caso especial del transporte urbano en Lima Metropolitana.

Como vimos en el capítulo 3, para la aproximación de los modelos por el método de Máxima Verosimilitud, debe resolverse el siguiente proceso iterativo de la ecuación.

$$\theta^{h+1} = \theta^h + C^{-1}.(\theta^h).G(\theta^h).$$

Este proceso iterativo fue resuelto con el software estadístico GLIM (Generalized Linear Model) el cual permitió realizar este procedimiento y consecuentemente obtener los estimadores del modelo.

El paquete estadístico proporcionó los datos requeridos luego de haber aplicado el siguiente programa para cada uno de los tres casos considerados:

Programa: Todas las unidades de transporte en estudio

```
$INPUT 15$
FILE NAME? ACTRA
$UNITS 335
$DATA Y E X C U V P F G O
$READ (** lectura de datos **)
$RETURN
$YVARIATE Y
$ERROR P $LINK L
$CALC EX=E*X:EU=E*U:XU=X*U:E2=E*E:X2=X*X:U2=U*U
$CALC? $FIT :+E:+X:+U:+EX:+EU:+XU:+E2:+X2:+U2
$DIS LMEC
$STOP$
```

Programa: Omnibus y Micros

```
$INPUT 15$
FILE NAME? MICRO
$UNITS 212
$DATA Y E X C U V P F G O
$READ (** lectura de datos **)
$RETURN
$YVARIATE Y
$ERROR P $LINK L
$CALC EC=E*C:CU=C*U:EU=E*U:C2=C*C:E2=E*E:U2=U*U
$CALC? $FIT :+E:+C:+U:+EC:+EU:+CU:+E2:+C2:+U2
$DIS LMERCV
$STOP$
```

Programa: Camionetas Rurales

```
$INPUT 15$
FILE NAME? COMBI
$UNITS 123
$DATA Y E X C U V P F G O
$READ (** lectura de datos **)
$RETURN
$YVARIATE Y
$ERROR P $LINK L
$CALC XC=X*C:XU=X*U:CU=C*U:X2=X*X:C2=C*C:U2=U*U
```

$CALC? $FIT :+X:+C:+U:+XC:+XU:+CU+:X2:+C2+:U2
$DIS LMEC
$STOP$

En la siguiente sección se presentan los resultados de las estimaciones para cada uno de los 3 modelos. Para un mejor estudio de las contribuciones de éstas variables explicativas, se incluye como variables lineales a las interaciones entre las variables explicativas dos a dos y a los términos cuadráticos.

Los estimadores de los coeficientes determinan cuales de las variables resultan significantes o no para determinar el modelo final en cada caso.

4.5.1 Estimación para Todas las Unidades de Transporte Público: Buses, Micros y Camionetas Rurales

Los resultados después de efectuadas 4 iteraciones, se muestran en los Cuadros 4.1 y Cuadro 4.2, muestra la contribución de cada una de las variables en el modelo.

Cuadro 4.1 .- Primer cálculo de estimadores

Nro.	Estimadores	Cuadrado del Error	Parametro
1	-1.756	0.9038	1
2	0.004269	0.04586	E
3	1.529	0.3816	C
4	1.442	0.2115	U
5	0.004902	0.005318	EU
6	-0.001827	0.009733	EC
7	-0.3290	0.06239	CU
8	-0.0001958	0.0005901	E^2
9	-0.1589	0.06016	C^2
0	-0.1813	0.02405	U^2

Cuadro 4.2 .- Modelo I

F.V.	Residual. S.C.R.	Cambio S.C.R.	G.L.	Cambio G.L.	Significancia Significancia
Media	517.20	—	334	—	NS
E	516.89	0.31	333	1	NS
C	413.71	103.18	332	1	**
U	168.41	245.30	331	1	**
EC	167.65	0.766	330	1	NS
EU	167.10	0.545	329	1	NS
CU	150.38	16.724	328	1	**
E^2	150.14	0.234	327	1	NS
C^2	141.11	9.037	326	1	**
U^2	70.272	70.83	325	1	**
Residual	70.272		325		

Las variables E, EC, EU y E^2 resultan ser no significativas para el modelo. Luego se calculan los estimadores para el conjunto de variables restantes, las que se presentan en el Cuadro 4.5.

Cuadro 4.3 .- Segundo cálculo de estimadores

Nro.	Estimadores	Cuadrado del Error	Parametro
1	-1.755	0.2165	1
2	0.0005683	0.001769	X
3	1.465	0.2295	C
4	1.565	0.1362	U
5	0.0001141	0.0007459	XC
6	-0.0003349	0.0004407	XU
7	-0.3324	0.06512	CU
8	-1.643E-06	5.007E-06	X^2
9	-0.1606	0.06070	C^2
0	-0.1778	0.02366	U^2

El análisis del Cuadro nro 4.4 también lleva a los mismos resultados del Cuadro nro 4.2, cuyos términos resultan ser altamente significativos. Cuadro nro 4.6.

Cuadro 4.4 .- Modelo II

F.V.	Residual. S.C.R.	Cambio S.C.R.	G.L.	Cambio G.L.	Significancia Significancia
Media	517.20	—	334	—	—
X	515.84	1.32	333	1	NS
C	413.39	102.45	332	1	**
U	168.27	245.1	331	1	**
XC	167.77	0.494	330	1	NS
XU	167.64	0.137	329	1	NS
CU	151.58	16.057	328	1	**
X^2	150.67	0.909	327	1	NS
C^2	150.67	8.228	326	1	**
U^2	142.44	72.25	325	1	**
Residual	70.192		325		

Las variables X, XC, XU y X^2 resultan ser no significativas para el modelo. Luego se calculan los estimadores para el conjunto de variables restantes.

Cuadro 4.5 .- Tercer cálculo de estimadores

Nro.	Estimadores	Cuadrado del Error	Parametro
1	-1.859	0.1619	1
2	1.475	0.2170	C
3	1.599	0.1278	U
4	-0.3348	0.05701	CU
5	-0.1578	0.06013	C^2
6	-0.1758	0.02313	U^2

Cuadro 4.6 .- Modelo Final

F.V.	Residual. S.C.R.	Cambio S.C.R.	G.L.	Cambio G.L.	Significancia
Media	517.20	—	334	—	—
C	413.99	103.21	333	1	**
U	169.79	244.2	332	1	**
CU	153.83	15.963	331	1	**
C^2	144.26	9.568	330	1	**
U^2	71.295	72.96	329	1	**
Residual	71.295		329		

Todas las variables son significativas, es decir, C, U, CU, C^2 y U^2 para el modelo, por lo tanto, el subconjunto seleccionado queda como sigue

$$logY = -1.859 + 1.475C + 1.599U - 0.3348CU - 0.1578C^2 - 0.1758U^2.$$

Evaluando para diversos valores de las variables:

Valores de C / U	Valor de Log Y	Valor de Y
C=0, U= 1	-0.4358	0.3666
C=0, U= 2	0.6358	4.3231
C=1, U= 0	-0.5418	0.2872
C=2, U=1	1.187	15.3815
C=1, U=4	1.6494	44.6067
C=0, U=5	1.741	55.0808

1. Si un chofer participó de 1 accidente en el año y no se la atribuye la culpa a el, se pronostica que probablemente tenga 1accidente en promedio al año.

2. Si un chofer participó de 2 accidentes en el año y no se la atribuye la culpa alguna, se pronostica que probablemente tenga 4 accidentes en promedio al año.

3. Si un chofer participó de 1 accidente en el año del cual se le declara culpable, se pronostica que probablemente tenga 1accidente en promedio

al año.

4. Si un chofer participó de 3 accidentes en el año y se la atribuyen la responsabilidad de dos de ellos por su imprudencia, se pronostica que probablemente tenga 15 accidentes en promedio al año.

5. Si un chofer participó de 5 accidentes en el año y se la atribuyen la responsabilidad de solo uno de ellos por su imprudencia, se pronostica que probablemente tenga 45 accidentes en promedio al año.

6. Si un chofer participó de 5 accidentes en el año de los cuales no se le atribuye ninguno, se pronostica que probablemente tenga 55 accidentes en promedio al año.

Se distingue el hecho que a mayor participación en accidentes, se pronostica mayor número de accidentes al año, sea o no que el chofer tuvo responsabilidad.

Cabe notar que la no responsabilidad del chofer no lo excluye de la tendencia a participar de accidentes de tránsito. La atribución de responsabilidad por un accidente se asume con algo de subjetividad, debido a la deficiencia del sistema de vigilancia en el Transporte Urbano en Lima Metropolitana.

De esta forma, tenemos la herramienta de predicción y sus correspondientes resultados de pronóstico del Nro de Accidentes de Tránsito promedio por chofer en Lima Metropolitana en un año determinado. Caso de Buses, Microbuses y Camionetas Rurales.

4.5.2 Estimación para Buses y Microbuses

Se tomó por separado los datos para Buses y Microbuses con el fin analizar Nro. de Accidentes en este tipo de transporte.

Cuadro 4.7 .- Primer cálculo de estimadores

Nro.	Estimadores	Cuadrado del Error	Parametro
1	-1.734	1.082	1
2	0.01557	0.05582	E
3	1.337	0.4309	C
4	1.275	0.2548	U
5	0.0002061	0.01080	EC
6	0.004404	0.006460	EU
7	-0.2969	0.06810	CU
8	-0.0003377	0.0007432	E^2
9	-0.1430	0.06422	C^2
0	-0.1505	0.026316	U^2

Cuadro 4.8 .- Modelo I

F.V.	Residual. S.C.R.	Cambio S.C.R.	G.L.	Cambio G.L.	Significancia Significancia
Media	315.62	—	211	—	—
E	315.12	0.51	210	1	NS
C	243.97	71.14	209	1	**
U	95.929	148.00	208	1	**
EC	95.377	0.551	207	1	NS
EU	94.768	0.610	206	1	NS
CU	85.853	8.915	205	1	**
E^2	85.366	0.487	204	1	NS
C^2	80.177	5.189	203	1	**
U^2	41.135	39.04	202	1	**
Residual	41.135		202		

Las variables *E*. *EC*, *EU* y E^2 resultan ser no significativas para el modelo, por lo tanto se prescinde de ellas. Luego se calculan las estimadores para el conjunto de variables restantes.

Cuadro 4.9 .- Segundo cálculo de estimadores

Nro.	Estimadores	Cuadrado del Error	Parametro
1	-1.554	0.2639	1
2	0.0002787	0.002105	X
3	1.315	0.2575	C
4	1.365	0.1599	U
5	0.0004514	0.0008890	XC
6	0.0004838	0.0005464	XU
7	-0.3162	0.07572	CU
8	-5.175E-06	6.332E-06	X^2
9	-0.1402	0.06563	C^2
10	-0.1455	0.02563	U^2

Cuadro 4.10 .- Modelo II

F.V.	Residual. S.C.R.	Cambio S.C.R.	G.L.	Cambio G.L.	Significancia Significancia
Media	315.62	—	211	—	—
X	315.10	1.52	210	1	NS
C	243.82	70.28	209	1	**
U	96.259	147.61	208	1	**
XC	96.023	0.236	207	1	NS
XU	95.787	0.236	206	1	NS
CU	84.937	10.850	205	1	**
X^2	82.459	2.478	204	1	NS
C^2	78.501	3.958	203	1	**
U^2	40.365	38.136	202	1	**
Residual	40.365		202		

Las variables X, XC, XU y X^2 resultan ser no significativas para el modelo, por lo tanto se prescinde de ellas. Luego se calculan los estimadores para el conjunto de variables restantes.

Cuadro 4.11 .- Tercer cálculo de estimadores

Nro.	Estimadores	Cuadrado del Error	Parametro
1	-1.631	0.1927	1
2	1.344	0.2375	C
3	1.416	0.1460	U
4	-0.2937	0.06081	CU
5	-0.1404	0.06402	C^2
6	-0.1455	0.02511	U^2

Cuadro 4.12 .- Modelo Final

F.V.	Residual. S.C.R.	Cambio S.C.R.	G.L.	Cambio G.L.	Significancia Significancia
Media	315.62	—	211	—	—
C	244.37	71.25	210	1	**
U	97.10	147.27	209	1	**
CU	87.723	9.377	208	1	**
C^2	82.202	5.521	207	1	**
U^2	41.780	40.422	206	1	**
Residual	41.780		206		

Todas las variables son significativas, es decir, C, U, CU, y U^2 para el modelo, por lo tanto, el modelo final queda como sigue
$log\, Y = -1.631 + 1.344C + 1.416U - 0.2937CU - 0.1404C^2 - 0.1455U^2$.

Evaluando para diversos valores de las variables:

Valores de C / U	Valor de Log Y	Valor de Y
C=0, U=1	-0.3605	0.4360
C=0, U=2	0.619	4.1591
C=1, U=0	-0.4274	0.3738
C=2, U=1	1.1785	15.0834
C=1, U=4	1.7338	54.1751
C=0, U=5	1.8115	64.7888

1. Si un chofer participó de 1 accidente en el año y no se la atribuye

la culpa a el, se pronostica que probablemente tenga 1accidente en promedio al año.

2. Si un chofer participó de 2 accidentes en el año y no se la atribuye la culpa alguna, se pronostica que probablemente tenga 4 accidentes en promedio al año.

3. Si un chofer participó de 1 accidente en el año del cual se le declara culpable, se pronostica que probablemente tenga 1accidente en promedio al año.

4. Si un chofer participó de 3 accidentes en el año y se la atribuyen la responsabilidad de dos de ellos por su imprudencia, se pronostica que probablemente tenga 15 accidentes en promedio al año.

5. Si un chofer participó de 5 accidentes en el año y se la atribuyen la responsabilidad de solo uno de ellos por su imprudencia, se pronostica que probablemente tenga 54 accidentes en promedio al año.

6. Si un chofer participó de 5 accidentes en el año de los cuales no se le atribuye ninguno, se pronostica que probablemente tenga 64 accidentes en promedio al año.

Se distingue el hecho que a mayor participación en accidentes, se pronostica mayor número de accidentes al año, sea o no que el chofer tuvo responsabilidad. Sin embargo, se nota un leve decrecimiento en el pronóstico del nro de accidentes de tránsito para sólo Buses y Microbuses.

Aquí también, la no responsabilidad del chofer no lo excluye de la tendencia a participar de accidentes de tránsito aunque con menos tendencia que del total de la población de unidades de transporte urbano. Posiblemente debido a que no se incluyen las Camionetas Rurales en este pronóstico. También, la atribución de responsabilidad por un accidente se asume con algo de subjetividad, debido a la deficiencia del sistema de vigilancia en el Transporte Urbano en Lima Metropolitana.

De esta forma, tenemos la herramienta de predicción y sus correspondientes resultados de pronóstico del Nro de Accidentes de Tránsito promedio por chofer en Lima Metropolitana en un año determinado. Caso Buses y Microbuses.

4.5.3 Estimación para Camionetas Rurales

Esta ocasión, el análisis es específicamente para las Camionetas Rurales, por representar un caso particular del Transporte Público.

Cuadro 4.13 .- Primer cálculo de estimadores

Nro.	Estimadores	Cuadrado del Error	Parametro
1	-1.917	2.062	1
2	-0.02957	0.1046	E
3	2.797	1.326	C
4	2.146	0.6169	U
5	0.003102	0.01605	EU
6	-0.02777	0.03819	EC
7	-0.6211	0.2935	CU
8	0.0003145	0.001322	E^2
9	0.000	aliased	C^2
10	-0.2972	0.05805	U^2

Cuadro 4.14 .- Modelo I

F.V.	Residual. S.C.R.	Cambio S.C.R.	G.L.	Cambio G.L.	Significancia Significancia
Media	178.53	—	122	—	—
E	177.91	0.62	121	1	NS
C	152.89	25.016	120	1	**
U	62.278	90.612	119	1	**
EU	61.144	1.134	118	1	NS
EC	59.789	1.355	117	1	NS
CU	52.982	6.807	116	1	**
E^2	52.858	0.124	115	1	NS
C^2	52.858	0.000	115	1	NS
U^2	21.181	31.677	114	1	**
Residual	21.181		114		

Las variables E, EU, EC, E^2 y C^2 resultan ser no significativas para el modelo. Luego se calculan los estimadores para el conjunto de variables restantes.

Cuadro 4.15 .- Segundo cálculo de estimadores

Nro.	Estimadores	Cuadrado del Error	Parametro
1	-2.286	0.4953	1
2	-0.004188	0.004970	X
3	1.940	0.6197	C
4	2.210	0.3157	U
5	-0.003089	0.003924	XC
6	0.001032	0.001504	XU
7	-0.4738	0.2892	CU
8	8.435E-06	0.00001262	X^2
9	0.000	aliased	C^2
10	-0.3211	0.06800	U^2

Cuadro 4.16 .- Modelo II

F.V.	Residual. S.C.R.	Cambio S.C.R.	G.L.	Cambio G.L.	Significancia Significancia
Media	178.53	—	122	—	—
X	177.91	0.62	121	1	NS
C	152.89	25.020	120	1	**
U	61.875	91.015	119	1	**
XC	59.865	2.011	118	1	NS
XU	59.762	0.103	117	1	NS
CU	46.581	13.181	116	1	**
X^2	46.164	0.417	115	1	NS
C^2	46.164	0.000	115	1	NS
U^2	20.218	25.946	114	1	**
Residual	20.218		114		

Las variables X, XC, XU, X^2 y C^2 resultan ser no significativas para el modelo. Luego se calculan los estimadores para el conjunto de variables restantes.

Cuadro 4.17 .- Tercer cálculo de estimadores

Nro.	Estimadores	Cuadrado del Error	Parametro
1	-2.470	0.3374	1
2	1.815	0.4733	C
3	2.174	0.2845	U
4	-0.5788	0.2187	CU
5	-0.000	aliased	C^2
6	-0.2809	0.05403	U^2

Cuadro 4.18 .- Modelo III

F.V.	Residual. S.C.R.	Cambio S.C.R.	G.L.	Cambio G.L.	Significancia Significancia
Media	178.53	—	122	—	—
C	152.90	25.63	121	1	**
U	62.353	90.55	120	1	**
CU	55.872	6.481	119	1	**
C^2	55.872	0.000	119	1	NS
U^2	22.423	33.45	118	1	**
Residual	22.423		118		

Las variables C, U, CU. C^2 y U^2 resultan ser no significativas para el modelo. Luego se calculan los estimadores para el conjunto de variables restantes.

Cuadro 4.19 .- Cuarto cálculo de estimadores

Nro.	Estimadores	Cuadrado del Error	Parametro
1	-2.470	0.3374	1
2	1.815	0.4733	C
3	2.174	0.2845	U
4	-0.5788	0.2187	CU
5	-0.2809	0.05403	U^2

Cuadro 4.20 .- Modelo Final

F.V.	Residual. S.C.R.	Cambio S.C.R.	G.L.	Cambio G.L.	Significancia Significancia
Media	178.53	—	122	—	—
C	152.90	25.63	121	1	**
U	62.353	90.55	120	1	**
CU	55.872	6.481	119	1	**
U^2	22.423	33.45	118	1	**
Residual	22.423		118		

Todas las variables son significativas, es decir, C, U, CU, C^2 y U^2 para el modelo, por lo tanto, el modelo final queda como sigue

$$log\, Y = -2.470 + 1.815C + 2.174U - 0.5788CU - 0.2809U^2.$$

Evaluando para diversos valores de las variables:

Valores de C - U	Valor de Log Y	Valor de Y
C=0, U=1	-0.5769	0.2649
C=0, U=2	0.7544	5.6807
C=1, U=0	-0.655	0.2213
C=2, U=1	1.8955	78.6140
C=1, U=4	1.2314	17.0373
C=0, U=5	1.3775	23.8506

1. Si un chofer participó de 1 accidente en el año y no se la atribuye la culpa a el, se pronostica que probablemente tenga 1accidente en promedio al año.

2. Si un chofer participó de 2 accidentes en el año y no se la atribuye la culpa alguna, se pronostica que probablemente tenga 5 accidentes en promedio al año.

3. Si un chofer participó de 1 accidente en el año del cual se le declara culpable, se pronostica que probablemente tenga 1accidente en promedio al año.

4. Si un chofer participó de 3 accidentes en el año y se la atribuyen la responsabilidad de dos de ellos por su imprudencia, se pronostica que probablemente tenga 78 accidentes en promedio al año.

5. Si un chofer participó de 5 accidentes en el año y se la atribuyen la responsabilidad de solo uno de ellos por su imprudencia, se pronostica que probablemente tenga 17 accidentes en promedio al año.

6. Si un chofer participó de 5 accidentes en el año de los cuales no se le atribuye ninguno, se pronostica que probablemente tenga 23 accidentes en promedio al año.

Se distingue el hecho que a mayor participación en accidentes, se pronostica mayor número de accidentes al año, sea o no que el chofer tuvo responsabilidad. Esta vez se ve incrementado el pronóstico del nro de accidentes de tránsito cuando sólo de Camionetas Rurales se trata.

Aquí también. la no responsabilidad del chofer no lo excluye de la tendencia a participar de accidentes de tránsito y ésta tendencia es alta. La atribución de responsabilidad por el accidente se asume con algo de subjetividad, debido a la deficiencia del sistema de vigilancia en el Transporte Urbano en Lima Metropolitana y sobretodo a la evidente indisciplina de los transportistas en este tipo de vehículos. (Camionetas Rurales).

De esta forma, tenemos la herramienta de predicción y sus correspondientes resultados de pronóstico del Nro de Accidentes de Tránsito promedio por chofer en Lima Metropolitana en un año determinado. Caso de Camionetas Rurales.

Capítulo 5

Conclusiones

5.1 Conclusiones

El presente trabajo de tesis estuvo orientado a utilizar el Modelo de Regresión cuyas variables dependientes tienen comportamiento de la distribución de Poisson, para elaborar pronósticos de una variable observada con un comportamiento similar al de la distribución Poisson.

Se formuló el Modelo de Regresión general, a partir del cual se hizo la suposición que las variable respuesta tiene distribución Poisson. Es decir, se a buscado modelar datos de tipo discreto que con frecuencia se presentan en la vida cotidiana. Este procedimiento de modelar datos de tipo discreto, es uno que no considera los supuestos del Modelo Clásico exigidos en el método clásico de Regresión.

Para lograr el objetivo, hubo una búsqueda de información en diversas instituciones como La Municipalidad de Lima Metropolitana, Ministerio de Transportes y Comunicaciones, Comisarías como Radio Patrulla y otras. Como conclusión se obtuvo que la información disponible consideraba solo accidentes registrados por motivos graves, ignorando los casos no graves, los mismos no fueron registrados en su totalidad ni en lugares suficientes como para representar a toda la ciudad de Lima Metropolitana y por último, existe información sobre accidentes de tránsito pero no se registran las variables mínimas para poder desarrollar un modelo de dependencia como un modelo de regresión.

Debido a la no disponibilidad de datos idóneos, se elaboró un diseño muestral y se procedió a encuestar a choferes seleccionados de la muestra

que incluye unidades de transporte urbano, tanto omnibuses, micros como camionetas rurales. Se midieron suficientes variables como para elaborar el modelo de regresión.

Como la principal variable medida en la encuesta es la variable respuesta de características similares a la de una distribución de Poisson, en esta tesis se ha podido desarrollar una aplicación práctica de Regresión con Datos Poisson. En esta aplicación se ha mostrado que se puede especificar y pronosticar a través de un modelo para el número de Accidentes de Tránsito en Lima Metropolitana al año. Principalmente, utilizar el modelo para hacer pronósticos de Nro de Accidentes de Tránsito en circunstancias similares.

Es importante observar que el estudio representa un pequeño esfuerzo de estudio, para cualquier especialista o profesional que desee utilizar este tipo de metodologías, porque requiere saber en que momento aplicar métodos matemáticos como en el presente trabajo ha sido necesario utilizar, métodos numéricos, para poder culminar el proceso de estimación en modelos no tan simples como lo es el desarrollado en esta tesis.

De esta aplicación, sobresale la importancia de la investigación en observaciones de esta naturaleza y otras donde las observaciones son de características similares y particularmente de los métodos desarrollados por este tema de tesis.

De los resultados del estudio, se distingue el hecho que a mayor participación en accidentes, se pronostica mayor probabilidad del número de accidentes al año, con responsabilidad o no del chofer encuestado. Las Camionetas Rurales muestran alta probabilidad a participar de accidentes de tránsito, mientras que los Buses y Microbuses tiene una probabilidad ligeramente menor.

El pronóstico realizado, se puede hacer para cualquier año con condiciones similares al año de estimación del modelo porque el estudio no depende del tamaño de la población sino del sistema de transporte en similares condiciones. El sistema de transporte actual no ha sufrido cambios desde hace más de 15 años, pues las únicas características que han diferenciado el transporte desde hace más de 15 años, son el tamaño de la población, las imposiciones económicas, la abertura de nuevas rutas, pero el sistema bajo el cual se maneja la ciudad de Lima Metropolitana, no varía, pues las empresas perteneces a los mismos dueños de hace muchos años y los nuevos dueños o empresas cumplen las mismas características.

Finalmente, se vislumbra que hay mucha necesidad de seguir utilizando los conceptos y metodologías de la estadística no tan utilizados, para obtener información valiosa a partir de data apropiada, como lo es en este caso los datos de conteo, específicamente las observaciones de tipo Poisson.

5.2 Recomendaciones

Luego de definir la metodología con la cual se desarrolló esta investigación, la primera necesidad es la base de datos a analizar. La data necesaria o base de datos, como información secundaria, casi nunca fue recolectada, registrada y las variables necesarias no fueron adecuadamente medidas como para la investigación que se llevó a cabo en esta tesis.

Se recomienda poner especial cuidado en la información a ser utilizada. Se hace evidente la necesidad de disponer de Marcos Muestrales adecuados para ser utilizados por muchos usuarios, en particular por investigadores estadísticos, los cuales deben analizar los datos a través de rigurosos métodos estadísticos y cuando no hay la información idónea, les resulta más costoso en todos los sentidos, disponer de tal información para continuar con la investigación.

Como la mayoría de información disponible en nuestro medio, es adecuada a lo más para un análisis estadístico descriptivo; como estadísticos estamos llamados a impulsar la cultura del registro de información mas detallada y este tipo de trabajos de tesis o de investigación puede servir de motivación. Para este trabajo se tuvo gran dificultad para conseguir el Marco Muestral. Por un lado no hay total confianza en los reportes estadísticos de la policía de tránsito, no por la falta de sinceridad o de voluntad de trabajo, sino por un funcionamiento aun inadecuado del sistema estadístico nacional en las instituciones.

En nuestro estudio, no se deja de lado la posibilidad de existencia de sesgo en la opinión de los conductores del transporte urbano. Sin embargo, se tomó especial cuidado en la forma de elaborar las preguntas para obtener una respuesta lo más fidedigna posible. Esta es una recomendación que no debe ignorarse en posteriores estudios similares y más aun, ponerle especial atención, incluso con la ayuda de un psicólogo para elaborar las preguntas, si fuera necesario.

Una posibilidad en trabajos posteriores es utilizar datos de los registros de la institución correspondiente como información previa y posteriormente, conducir una encuesta para recolectar la información muestral necesaria. Consecuentemente, llevar a cabo la versión bayesiana de este modelo. Esto podría ser motivo de otro estudio.

Lo mencionado en las conclusiones, no excluye que se pueda profundizar mucho más en el tema. Para ello, es necesario un considerable apoyo bibliográfico. En nuestro medio el apoyo aun es insuficiente en la obtención de revistas científicas tecnológicas en estadísticas y en la compra de bibliografía especializada. En particular, Modelos de Regresión No Convencionales y urge suficiente apoyo en otras áreas como el factor económico.

Apéndice B

Bibliografía

1. Draper, N.R. and Smith, H., "Applied Regression Analysis", New York : John Wiley Sons, Inc., 1966.

2. Ferguson, T.S., "A Method of Generating Best Asymptotically Normal Estimates with Application to the Estimation of Bacterial Densities", The Annal of Mathematical Statistics.

3. Frome, E. L., Kutner M. H., and Beauchamp J. J., "Regression Analysis of Poisson-Distributed Data" Journal of de American Statistical Association. 1973, vol. 68 344, págs. 935-940. Theory and Methods Section.

4. Jorgenson, D. W., "Multiple Regression Analysis of a Poisson Process" Journal of de American Statistical Association, 56, June 1961, 235-45.

5. Kendall, M.G. and Stuart, A., "The Advance Theory of Statitics", Vol.2, New York : Hafner Publishing Company, 1946.

6. Montgomery, D.C. and Peck, E.A., "Introduction to Linear Regression Analysis", New York : John Wiley Sons, Inc., 1982.

7. Rao, C.R., "Linear Statistical Inference and Its Applications", New York:John Wiley Sons, Inc., 1965.

8. Siegmund, Brandt, "Statistical and Computational Methods in Data Analysis", North-Holland Publishing Company, $2^{o}Ed$, 1976.

9. Harald Cramer, "Métodos Matemáticos de Estadística", Aguilar, S.A. de Ediciones Madrid, $2^o Ed$, 1960.