Matemática: Una publicación del FCNM– ESPOL
2017, Vol.15, No. 2 **Publicado: 2017-10-02**

APLICACIÓN DE TÉCNICAS DE ANÁLISIS MULTIVARIADO EN EL ESTUDIO DEL RIESGO DE SINIESTRO ASOCIADO A LAS CARACTERÍSTICAS DE LOS VEHÍCULOS EN LA CIUDAD DE QUITO

APPLICATION OF MULTIVARIATE ANALYSIS TECHNIQUES IN THE STUDY OF ACCIDENT RISK ASSOCIATED WITH THE CHARACTERISTICS OF VEHICLES IN THE CITY OF QUITO

García Sandra[1], Jiménez Johnny[2], Ordóñez Guillermo[3]

Resumen. *En este trabajo se analiza si las características físicas asociadas a los vehículos de la ciudad de Quito influyen en la ocurrencia de siniestros. Para ello se tomó una muestra de 62,474 vehículos y se utilizaron las técnicas estadísticas multivariantes de Regresión Logística y de Análisis de Correspondencias Múltiples, a fin de estudiar las relaciones entre los factores físicos de los vehículos y la ocurrencia de un siniestro. Con la ayuda del software estadístico IBM SPSS 22, se determinó que la marca, el tipo y el color del vehículo son características que inciden en la probabilidad de ocurrencia de un siniestro.*
Palabras clave: Seguro de vehículos, Análisis de Correspondencias Múltiple (ACM), Análisis de Regresión Logística (ARL).
Abstract. *In this paper we analyze if the physical characteristics associated to the vehicles of the city of Quito influence the occurrence of accidents. For this, a sample of 62,474 insured vehicles was taken and the multivariate statistical techniques of Multiple Correspondence Analysis and Logistic Regression were used, to study the relationships between the physical factors of the vehicles and the occurrence of a casualty. With the help of the statistical software IBM SPSS 22, it was determined that the brand, type and color of the vehicle are characteristics that affect the probability of occurrence of a loss.*
Key words: Vehicle Insurance, Multiple Correspondence Analysis (ACM), Logistic Regression Analysis (ARL).

Recibido: Septiembre 2017
Aceptado: Septiembre 2017

1. INTRODUCCIÓN

Debido a la gran cantidad de accidentes de tránsito que se producen, los propietarios de vehículos buscan protegerse ante un posible siniestro de tránsito. En el presente trabajo se entenderá como "siniestro" a todo tipo de daño físico (choques, golpes, rayones, etc.) incluido el robo de partes o pérdida total que sufra el vehículo asegurado. Las empresas aseguradoras ofrecen coberturas que ayudan a compensar las pérdidas económicas que pudiera sufrir un asegurado, siempre y cuando este haya cancelado previamente la correspondiente prima a la aseguradora, lo que le da derecho a exigir una compensación económica por los daños causados. El seguro le brinda al asegurado la posibilidad de reducir significativamente las pérdidas causadas, de tal manera que estas no generen tanto impacto del que hubiera sufrido, sino disponía de una póliza contratada.

Esta problemática origina en las aseguradoras el interés de encontrar una forma técnica que les permita estimar la probabilidad de que un vehículo sea propenso al riesgo de sufrir un siniestro.

[1]*García Sandra, Ph.D., Centro de Estudios e Investigaciones Estadísticas, Facultad de Ciencias Naturales y Matemáticas, (ESPOL); (e-mail: slgarcia@espol.edu.ec).*
[2]*Jiménez Johnny, Ing., Mg, Docente de la Universidad Politécnica Salesiana (UPS); (e-mail: jjimenez@ups.edu.ec).*
[3]*Ordóñez Guillermo, Ing., Mg, Analista Coord. Política de Inversiones, Ministerio de Comercio Exterior; (e-mail: guillermo.ordonez@comercioexterior.gob.ec).*

Dentro del ámbito de los seguros se han realizado algunas investigaciones a nivel internacional, para estudiar la siniestralidad, como los de Das & Sun (2016) [1]; Paefgen, Staake, & Fleisch (2014) [2]; Hemrit, Arab, & Raissi (2013) [3], pero a nivel nacional no existen estudios de similares características, lo que hace de esta investigación pionera dentro del mercado asegurador ecuatoriano y servirá como base para futuros trabajos en este campo.

Es fundamental tener información relacionada con las características que definen a los vehículos, por lo cual se requieren variables que involucren a la marca, categoría, color y tipo de los vehículos asegurados, además una variable relacionada con el registro de siniestros. Estas variables se obtuvieron de una muestra de 62,474 vehículos proporcionada por una aseguradora de la ciudad de Quito.

A través de técnicas estadísticas multivariantes, como el análisis de correspondencias múltiples y el análisis de regresión logística, se determinará la propensión al riesgo de un vehículo según sus características, pudiendo clasificarlo o discriminarlo dentro del grupo de los que pueden sufrir o no un siniestro, lo cual ayudará a las aseguradoras a decidir si se le otorga o no la cobertura.

2. MARCO TEÓRICO

2.1 Análisis de Correspondencias. Es una técnica multivariada factorial que busca reducir la dimensión de una tabla de datos formada por variables cualitativas, con la finalidad de obtener un número pequeño de factores, cuya interpretación a

posteriori hará que el problema investigado sea más simple de estudiar. [4]

El análisis de correspondencias al trabajar con variables cualitativas, o con variables cuantitativas categorizadas (por ejemplo, si se define la variable edad, pero categorizada en distintos rangos de edad) hace que posea dos características importantes.

Primero, trabaja con frecuencias que son el resultado de cruzar dos o más variables. Segundo, cuando se cruzan dos variables, utiliza como individuos y variables las diferentes categorías existentes. Esto hace posible aplicar el llamado análisis de correspondencias simple (ACS), pero cuando las categorías pertenecen a más de dos variables, el método se generaliza y se obtiene el análisis de correspondencias múltiples (ACM). [5]

Para el análisis de correspondencias simple, los datos de las dos variables se representan a través de una tabla de contingencia, pero en el análisis de correspondencias múltiples, la tabla de contingencia se convierte en una hipertabla que puede tener tres o más dimensiones, la cual no es muy fácil de representar y suele resumirse en la llamada tabla de Burt. [4]

No se podría trabajar con una tabla de contingencia de 2x2, como en el caso del análisis de correspondencias simple, ya que la representación tabular de los datos ahora se complica, por lo cual es necesario realizar un análisis de correspondencias múltiples, porque permite estudiar las relaciones entre las modalidades de todas las variables cualitativas consideradas.

En el análisis de correspondencias múltiples se disponen los datos en una tabla Z que se denomina tabla disyuntiva completa, la cual está formada por un conjunto I de n individuos o filas, un conjunto J de Q variables o caracteres cualitativos y por un conjunto de modalidades o categorías excluyentes $1, \cdots, m_k$ para cada variable cualitativa. [4]

2.2 Regresión Logística. Supongamos que disponemos de un conjunto de variables, de las cuales una es dicotómica, que se considera como la variable dependiente o variable respuesta, y además se tienen dos o más variables cualitativas o cuantitativas, que se consideran como variables independientes. Si se desea estudiar la relación entre esta variable dependiente y las demás variables independientes, se aplica una técnica de análisis multivariante conocida como regresión logística múltiple.

En la regresión logística múltiple, la variable dependiente, se dice que es dicotómica o binaria, porque esta sólo puede tomar dos valores 0 o 1, que definen características opuestas o mutuamente excluyentes, por ejemplo: ser o no propenso a sufrir un accidente de tránsito. Además, en el modelo para cada variable independiente cualitativa, se tendrán que generar variables DUMMY, por ejemplo, si se tiene una variable cualitativa con K categorías, se deberán generar K – 1 variables DUMMY, cada una de las cuales tomará el valor 0 o 1, es decir, que la variable DUMMY tomará el valor de 1 si un individuo pertenece a una determinada categoría de la variable cualitativa y tomará el valor de 0 en caso contrario, de tal manera que todas las categorías de la variable cualitativa puedan ser representadas en el modelo [6]

En la regresión logística múltiple además de encontrar la relación entre la variable dependiente y las demás variables independientes, se busca determinar la medida de dicha relación y estimar la probabilidad de que ocurra o no el evento que define la variable dependiente en función de los valores que tomen las variables independientes. [5].

Sean, y una variable dependiente, que toma valores que sólo pueden ser: 1 con probabilidad p o 0 con probabilidad $1-p$, y $x_1, x_2, \cdots, x_n$ las variables independientes, que pueden ser cuantitativas o cualitativas.

Un modelo de regresión logística múltiple busca estimar la probabilidad de que ocurra un determinado evento, es decir, la probabilidad de que un individuo i elija una respuesta binaria (1 o 0) en función de los valores que tomen las variables independientes, $x_{1i}, x_{2i}, \cdots, x_{ni}$. [5]

La expresión matemática del modelo de regresión logística múltiple está dado por:

$$E(y_i) = P(y_i = 1 \mid x_{1i}, x_{2i}, \cdots, x_{ni}) = p_i = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}}}$$

o simplemente expresado como:

$$p_i = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}}}$$

De manera equivalente, este modelo se puede representar por medio de las siguientes expresiones:

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n}$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}$$

Para poder realizar el ajuste de este modelo y la estimación de los parámetros $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$ se usa el método de estimación por máxima verosimilitud (EMV).

En primer lugar, se construye una función L, que se denomina función de verosimilitud, la cual es usada para expresar la probabilidad de los datos observados como una función de parámetros desconocidos. Los valores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_n$ que

permiten maximizar la función L, son los llamados estimadores maximoverosímiles de los parámetros $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$. Es decir, que este método calcula los valores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_n$ que son estimadores de los parámetros desconocidos $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$, y son tales que maximizan la probabilidad de que con ellos se puedan obtener los valores observados. [5]

Cuando ya se haya ajustado el modelo y encontrado los estimadores máximos verosímiles $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_n$

La estimación de la probabilidad $\hat{p}$ viene dada por:

$$\hat{p}_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+\cdots+\hat{\beta}_n x_{ni}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+\cdots+\hat{\beta}_n x_{ni}}},$$

Que estima la probabilidad de que un individuo elija la respuesta binaria 1 dado un determinado valor de las variables independientes $x_{1i}, x_{2i}, \cdots, x_{ni}$.

Una expresión equivalente a la anterior es:

$$\frac{\hat{p}_i}{1-\hat{p}_i} = e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+\cdots+\hat{\beta}_n x_{ni}},$$

Que se denomina "odds" o "ventaja", y se define como el cociente entre la probabilidad de que ocurra el evento de interés, que en regresión logística es siempre $y_i = 1$, y la probabilidad de que dicho evento no ocurra, es decir, estima la ventaja o preferencia de un individuo por la respuesta 1 de la variable dependiente frente a la respuesta 0 para cada valor de las variables independientes. [5]

Si se toma el logaritmo natural al odds, se obtiene un modelo lineal que se denomina transformación logística o simplemente modelo logit, a diferencia de la probabilidad p_i, que constituye un modelo no lineal o modelo logístico. Este modelo en función del logaritmo del odds, es importante porque permite que los coeficientes del modelo se puedan interpretar sencillamente en términos de independencia o asociación entre variables. [4]

El modelo logit, se puede expresar como:

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_n x_{ni}$$

La prueba de significancia en la regresión logística múltiple es parecida a la prueba de significancia en la regresión múltiple. En primer lugar, se realiza una prueba para probar la significancia global del modelo, por lo cual se deben plantear y contrastar hipótesis sobre los coeficientes de regresión en forma conjunta. Esto permite verificar si las variables independientes que están presentes en el modelo ajustado están relacionadas significativamente con la variable dependiente. [21]

Las hipótesis para probar la significancia global son las siguientes:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_n = 0$

H_1 : Al menos algún $\beta_i \neq 0, \forall i = 1,2,\cdots,n$

Esta prueba de significancia global se basa en el valor del estadístico de prueba G que se define como:

$$G = -2\ln\left[\frac{\text{Verosimilitud del modelo sólo con la constante}(L_0)]}{\text{Verosimilitud del modelo seleccionado}\ (L_p)}\right]$$

El estadístico G se distribuye como una ji-cuadrada χ^2 con p-1 grados de libertad, donde p es el número de parámetros en el modelo bajo estudio.

El estadístico G está basado en la función de verosimilitud de cada modelo y permite comparar la probabilidad de que los valores estimados por cada modelo representen a los valores observados de la variable dependiente. [5]

Para poder realizar el contraste de hipótesis de esta prueba, se aplica el método del valor crítico que genera la siguiente regla de rechazo:

Método del valor crítico:

Rechazar $\boldsymbol{H_0}$ si $G \geq \boldsymbol{\chi^2_{\alpha,(p-1)}}$

Donde α es el nivel de significancia y p-1 son los grados de libertad.

Si se rechaza la hipótesis nula, se concluye que el modelo global es significativo.

Si después de haber realizado la prueba G, esta indicó que sí existe una significancia global, lo siguiente a realizar es determinar si la contribución que hace cada variable independiente al modelo es significativa, para lo cual se utiliza una prueba de significancia individual para cada coeficiente de regresión, que está basada en el estadístico W de Wald. [5]

Las hipótesis para probar la significancia individual son:

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0, \forall i = 1,2,\cdots,n$

El valor del estadístico de prueba W se define como:

$$W = \left(\frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)}\right)^2$$

Este estadístico es igual al cuadrado de la razón entre el estimador del coeficiente de la variable independiente y el estimador de su error estándar, y sigue una distribución una ji-cuadrada χ^2 con 1 grado de libertad si la variable independiente es cuantitativa, pero si la variable independiente es cualitativa, el número de grados de libertad es igual al número de categorías menos uno. [5]

Para poder realizar el contraste de hipótesis de esta prueba, se aplica el método del valor crítico que genera la siguiente regla de rechazo:

Método del valor crítico:

Rechazar $\boldsymbol{H_0}$ si $W \geq \boldsymbol{\chi^2}_{\boldsymbol{\alpha,1}}$ o

Rechazar $\boldsymbol{H_0}$ si $W \geq \boldsymbol{\chi^2}_{\boldsymbol{\alpha,(p-1)}}$

Donde α es el nivel de significancia, $\boldsymbol{p}$ es el número de categorías de la variable cualitativa y $\boldsymbol{(p-1)}$ son los grados de libertad. Si se rechaza la hipótesis nula, se concluye que la variable independiente es estadísticamente significativa.

3. DATOS

A continuación, se describen las variables que se utilizan en los modelos presentados en esta investigación:

Categoría: Variable cualitativa dicotómica que permite determinar la categoría del vehículo, la cual puede tomar dos valores (categorías) posibles, que se codifican como: 0 (Liviano) y 1 (Pesado).

Marca: Variable cualitativa, de carácter nominal que permite determinar la marca del vehículo. Puede tomar los siguientes valores (categorías) posibles, los cuales se los codifica como: 0 (Chevrolet), 1 (Hyundai), 2 (Kia), 3 (Toyota), 4 (Nissan), 5 (Suzuki), 6 (Ford), 7 (Volkswagen), 8 (Mazda), 9 (Renault), 10 (Great Wall), 11 (Mitsubishi), 12 (Honda), 13 (Hino), 14 (Skoda) y 15 (Otra).

Tipo: Variable cualitativa, de carácter nominal que permite determinar el tipo del vehículo, la cual puede tomar los siguientes valores (categorías) posibles, que se codifican como: 0 (Sedán), 1 (Todoterreno), 2 (Camioneta), 3 (Camión), 4 (Moto) y 5 (Otro).

Color: Variable cualitativa, de carácter nominal que permite determinar el color del vehículo. Puede tomar los siguientes valores (categorías) posibles, los cuales se los codifica como: 0 (Plateado), 1 (Blanco), 2 (Plomo), 3 (Negro), 4 (Rojo), 5 (Azul), 6 (Beige), 7 (Dorado), 8 (Verde), 9 (Concho de vino), 10 (Gris), 11 (Celeste) y 12 (Amarillo).

Siniestro: Variable dependiente o respuesta, de naturaleza cualitativa dicotómica, que permite determinar si el vehículo bajo estudio sufrió o no un siniestro, la cual puede tomar dos valores posibles, que se codifican como: 0 (No hubo Siniestro) y 1 (Hubo Siniestro).

Año: Variable cuantitativa discreta, que permite determinar el año de fabricación del vehículo.

4. METODOLOGÍA

Para el Análisis de Correspondencias múltiples se redefinieron las variables Marca, Color y Tipo, dado que tienen muchas categorías se crearon nuevas categorías que agrupen a las originales de acuerdo con su peso, es decir, a la frecuencia de siniestralidad que presentaron estas clases. Las tablas 1, 2 y 3 resumen este proceso.

Tabla 1:
Recategorización de la variable Marca

Codificación	Siniestralidad	Intervalo	Marca	Frecuencia	Frecuencia relativa
GM1	ALTA	f ≥900	CHEVROLET	2,352	28.96%
			HYUNDAI	1,096	13.49%
			KIA	985	12.13%
GM2	MEDIA	500≤ f <900	TOYOTA	655	8.06%
			NISSAN	493	6.07%
GM3	BAJA	200≤ f <500	VOLKSWAGEN	394	4.85%
			FORD	373	4.59%
			OTRA	372	4.58%
			RENAULT	340	4.19%
			MAZDA	330	4.06%
			SUZUKI	320	3.94%
GM4	MUY BAJA	f<200	GREAT WALL	170	2.09%
			MITSUBISHI	89	1.10%
			HONDA	77	0.95%
			SKODA	41	0.50%
			HINO	35	0.43%

Tabla 2:
Recategorización de la variable Color

Codificación	Siniestralidad	Intervalo	Color	Frecuencia	Frecuencia relativa
GC1	ALTA	f ≥900	PLATEADO	1,636	20.14%
			BLANCO	1,403	17.27%
			PLOMO	1,326	16.33%
			NEGRO	926	11.40%
GC2	MEDIA	500≤ f <900	ROJO	783	9.64%
			AZUL	512	6.30%
GC3	BAJA	200≤ f <500	DORADO	354	4.36%
			BEIGE	292	3.60%
			CONCHO DE VINO	260	3.20%
			VERDE	243	2.99%
GC4	MUY BAJA	f<200	GRIS	172	2.12%
			OTRO	104	1.28%
			CELESTE	75	0.92%
			AMARILLO	32	0.39%
			ACERO	4	0.05%

Tabla 3:
Recategorización de la variable Tipo

Codificación	Siniestralidad	Intervalo	TIPO	Frecuencia	Frecuencia relativa
GT1	ALTA	f≥2000	SEDÁN	3,880	47.77%
			TODO TERRENO	2,848	35.07%
GT2	MEDIA	1000≤ f< 2000	CAMIONETA	1,104	13.59%
GT3	BAJA	f<1000	CAMIÓN	147	1.81%
			OTRO	93	1.15%
			MOTO	50	0.62%

Por el contrario, las variables Categoría (LIVIANO, PESADO) y Siniestro (NO HUBO

SINIESTRO, HUBO SINIESTRO) como sólo tienen dos categorías se las analizará sin realizar cambio alguno.

Para la variable Año, se la recodificará, agrupando los años de fabricación de acuerdo con el percentil 25, 50 y 75, ya que para el análisis multivariado se requiere que esta variable sea cualitativa, por lo cual se la renombrará a Año según percentil. Esta variable abarca las categorías 1, 2, 3 y 4, agrupan a los años de fabricación de acuerdo con los intervalos, [1949, 2007), [2007, 2011), [2011, 2013) y [2013, 2015].

Con la finalidad de favorecer la interpretación de los resultados del ACM respecto a la relación entre la variable Siniestro con el resto de las variables se consideraron 7 dimensiones.

Luego, se efectúa el análisis de regresión logística, para lo cual se realizaron algunas pruebas, tratando de encontrar el mejor modelo. Se utilizó el proceso de introducción manual de variables a SPSS, donde se probó primero con todas las variables y se fueron eliminando sucesivamente variables que no fueron significativas, además se probó con interacciones entre las variables, este proceso se fue repitiendo hasta que se encontró un conjunto de variables donde todas fueran significativas, de esa manera se logró determinar el modelo que mejor se ajusta a los datos observados.

La descripción y análisis de datos, se obtuvo en base a la introducción de las variables Color, Marca y Tipo, a través de las cuales se encuentra el modelo de regresión logística. Debido a que las variables mencionadas son categóricas, y para poder aplicar el análisis de regresión logística, se necesita recodificarlas, creando variables DUMMY cuyo número es igual al número de categorías de la variable original pero disminuida en uno. En la Tabla 4 se muestran las codificaciones de las variables categóricas.

Tabla 4:
Codificaciones de variables categóricas

Variables originales y sus categorías		Codificación de variables DUMMY		
		COLOR(1)	COLOR(2)	COLOR(3)
COLOR	GC1	1	0	0
	GC2	0	1	0
	GC3	0	0	1
	GC4	0	0	0
		MARCA(1)	MARCA(2)	MARCA(3)
MARCA	GM1	1	0	0
	GM2	0	1	0
	GM3	0	0	1
	GM4	0	0	0
		TIPO(1)	TIPO(2)	
TIPO	GT1	1	0	
	GT2	0	1	
	GT3	0	0	

Los análisis estadísticos se desarrollarán usando el software IBM SPSS versión 22.

5. RESULTADOS Y ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES (ACM)

Los resultados obtenidos del ACM se describen y analizan como sigue. En la Tabla 5, se muestran las medidas discriminantes por variable, y se pueden identificar las puntuaciones que tiene cada una de ellas en la dimensión correspondiente, las cuales permiten ver cuánto discrimina cada variable en la dimensión respectiva.

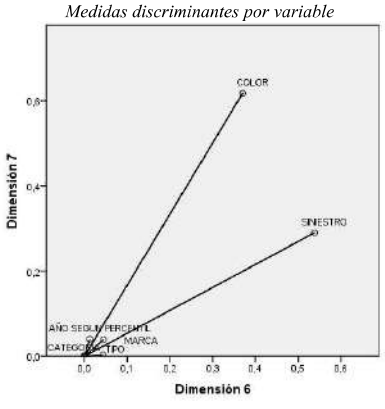
Tabla 5:
Medidas discriminantes por variable

	Dimensión							Media
	1	2	3	4	5	6	7	
CATEGORÍA	0.76	0.03	0.00	0.01	0.00	0.00	0.00	0.11
MARCA	0.18	0.29	0.37	0.32	0.11	0.05	0.04	0.19
TIPO	0.78	0.05	0.36	0.09	0.05	0.04	0.00	0.20
COLOR	0.05	0.45	0.10	0.19	0.34	0.37	0.62	0.30
SINIESTRO	0.01	0.01	0.02	0.01	0.00	0.54	0.29	0.13
AÑO SEGUN PERCENTIL	0.03	0.47	0.27	0.45	0.54	0.01	0.04	0.26
Total activo	1.80	1.29	1.12	1.06	1.05	1.01	0.99	1.19
% de varianza	29.96	21.47	18.63	17.69	17.43	16.85	16.53	19.79

En esta tabla, el ACM permitió encontrar las dimensiones donde la variable SINIESTRO fuera significativa, es decir, las dimensiones que presenten los mayores pesos. Se seleccionaron las dimensiones 6 y 7, por presentar pesos de 0.54 y 0.29 respectivamente, estas dimensiones son las que permitirán conseguir una mejor discriminación entre la variable SINIESTRO y el resto de variables.

En la Figura 1, se grafican las medidas discriminantes correspondientes a las dimensiones 6 y 7, donde se puede apreciar que se forman dos grupos de variables de acuerdo con la cercanía que estas presentan.

Figura 1:
Medidas discriminantes por variable

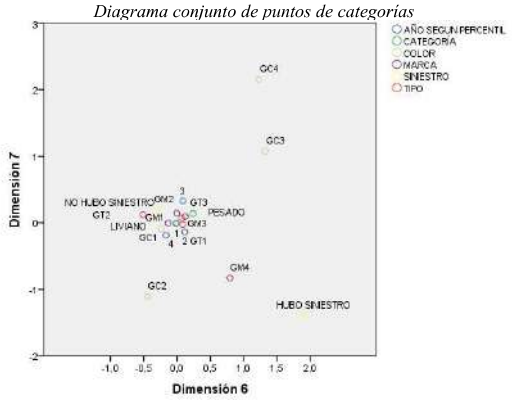


En el primer grupo están las variables COLOR y SINIESTRO, indicando una relación existente entre

ellas; mientras que, en el segundo grupo se distinguen las variables AÑO SEGÚN PERCENTIL, MARCA, CATEGORÍA y TIPO las cuales se encuentran relacionadas entre sí, dada su cercanía en el gráfico.

En la Figura 2, se presenta el diagrama conjunto de puntos de categorías de las variables objeto de estudio, que permite representar las correspondencias entre todas las variables y sus respectivas categorías. En este gráfico se pueden observar las relaciones entre las categorías de todas las variables y las categorías de la variable SINIESTRO.

Figura 2:
Diagrama conjunto de puntos de categorías



En este diagrama, se puede apreciar que la categoría GM4 de la variable MARCA, que agrupa a las marcas GREAT WALL, MITSUBISHI, HONDA, SKODA e HINO; y la categoría GC2 de la variable COLOR, compuesta por los colores ROJO y AZUL; se relacionan con la categoría HUBO SINIESTRO de la variable SINIESTRO, ya que se encuentran cercanas entre ellas. Es decir, que los vehículos de marca GREAT WALL, MITSUBISHI, HONDA, SKODA e HINO, que son de colores ROJO y AZUL, están más relacionados con la existencia de siniestralidad.

6. RESULTADOS Y ANÁLISIS DE REGRESIÓN LOGÍSTICA (ARL)

Los resultados que se obtuvieron con el ARL permitieron identificar cuáles son los factores que influyen significativamente en la probabilidad de que un vehículo sufra un siniestro, cuyo análisis se describe a continuación.

La codificación de las variables categóricas permitió determinar la significancia global del modelo que contiene estas variables, la cual está basada en el valor del estadístico de prueba G.

Las hipótesis para probar la significancia global son las siguientes:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_n = 0$$

$$H_1 : \text{Al menos algún } \beta_i \neq 0, \forall i = 1,2,\cdots,n$$

Este contraste de hipótesis se lo realizó a través de la prueba de Omnibus que provee SPSS, y se obtuvieron los siguientes resultados: El valor del estadístico de prueba G es 269.12, con 8 grados de libertad y el valor-p de la prueba es igual a 0 (menor a 0.05), por lo que existe evidencia estadística para rechazar la hipótesis nula. Por lo tanto, al menos uno de los coeficientes de las variables independientes es diferente de cero.

Como en la prueba de Omnibus se obtuvo que al menos uno de los coeficientes de las variables independientes que están presentes en el modelo, es diferente de cero, hay que determinar cuáles son las variables que tienen coeficientes que son estadísticamente significativos a un nivel de significancia de 0.05, por lo cual se utiliza una prueba de significancia individual para cada coeficiente de las variables independientes que se basa en el estadístico W de Wald. Se realizó el siguiente contraste de hipótesis.

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0, \forall i = 1,2,\cdots,n$$

En la Tabla 6 se pueden observar los resultados de efectuar los contrastes de hipótesis a cada uno de los coeficientes de las variables independientes.

Tabla 6:
Variables en el Modelo de Regresión Logística

Factores	Coef. Var. (β)	Error estándar	Estadístico de Wald (W)	G. L.	Valor p	e^{β} (Odds Ratio)	I. C. (95%) Odds Ratio (e^{β})	
							L. Inf.	L. Sup.
Constante	-2.69	0.09	864.74	1	0.00	0.07		
MARCA(1)	-0.15	0.06	7.32	1	0.01	0.86	0.77	0.96
MARCA(2)	-0.07	0.06	1.13	1	0.29	0.94	0.83	1.06
MARCA(3)	-0.11	0.06	3.53	1	0.06	0.90	0.80	1.01
TIPO(1)	0.83	0.06	172.40	1	0.00	2.29	2.02	2.59
TIPO(2)	0.70	0.07	102.24	1	0.00	2.02	1.76	2.32
COLOR(1)	0.15	0.06	7.49	1	0.01	1.17	1.05	1.30
COLOR(2)	0.07	0.06	1.32	1	0.25	1.07	0.95	1.21
COLOR(3)	0.28	0.06	19.20	1	0.00	1.32	1.17	1.49

En base a esto, se puede afirmar que existe evidencia estadística para no rechazar la hipótesis nula por medio del estadístico de Wald y con valores p asociados ($p>0.05$), para los coeficientes de las variables DUMMY, MARCA(2), MARCA(3) y COLOR(2), lo cual significa que estos coeficientes no son estadísticamente significativos. Por el contrario, los coeficientes de las variables DUMMY, MARCA(1), TIPO(1), TIPO(2), COLOR(1) y COLOR(3), de acuerdo con sus estadísticos de prueba y valores-p respectivos ($p<0.05$), se rechaza la hipótesis nula del contraste de hipótesis planteado, lo cual significa que estos coeficientes son estadísticamente significativos.

Un modelo de regresión logística debe estar formado sólo con variables que sean

estadísticamente significativas, por lo cual deben eliminarse las variables cuyos coeficientes no fueron estadísticamente significativos, es decir, se eliminan las variables DUMMY que globalmente no aportan en la solución del modelo, las cuales son: MARCA (2), MARCA (3) y COLOR (2). En base a esto se obtuvieron los siguientes resultados, los que aparecen en la siguiente tabla.

Tabla 7:
Variables que quedan en el Modelo de Regresión Logística

Factores	Coef. Var. (β)	Error estándar	Estadístico de Wald (W)	G. L.	Valor p	e^{β} (Odds Ratio)	I. C. (95%) Odds Ratio (e^{β})	
							L. Inf.	L. Sup.
Constante	-2.69	0.09	864.74	1	0.00	0.07		
MARCA(1)	-0.15	0.06	7.32	1	0.01	0.86	0.77	0.96
TIPO(1)	0.83	0.06	172.40	1	0.00	2.29	2.02	2.59
TIPO(2)	0.70	0.07	102.24	1	0.00	2.02	1.76	2.32
COLOR(1)	0.15	0.06	7.49	1	0.01	1.17	1.05	1.30
COLOR(3)	0.28	0.06	19.20	1	0.00	1.32	1.17	1.49

En la Tabla 7 se puede distinguir que los coeficientes de las variables DUMMY, MARCA (1), TIPO (1), TIPO (2), COLOR (1) y COLOR (3), son significativos ya que sus respectivos valores-p son menores a 0.05 ($p<0.05$), por lo tanto, estas variables son las que finalmente componen el modelo.

Para evaluar la bondad del ajuste del modelo obtenido, se realizó la prueba de Hosmer-Lemeshow, la cual realiza el siguiente contraste de hipótesis:

H_0 : El modelo seleccionado ajusta bien a los datos

$H_1 : \neg H_0$

El valor del estadístico de prueba chi-cuadrado es 6.95, con 8 grados de libertad y el valor-p de la prueba es igual a 0.33 (mayor a 0.05), por lo cual, existe evidencia estadística para no rechazar la hipótesis nula. Por lo tanto, el modelo seleccionado ajusta bien a los datos observados.

Otra forma que se utilizó para evaluar la bondad de ajuste del modelo fue a través de la tasa global de aciertos que se obtiene de la tabla de clasificación de resultados que permite analizar la capacidad predictiva del modelo obtenido, la cual plantea el siguiente contraste de hipótesis:

H_0 : Número de casos correctamente clasificados por el modelo no difiere de la clasificación esperada sólo por efecto del azar.

H_1: $\neg H_0$

En la Tabla 8 se puede observar la clasificación de resultados, donde el porcentaje global de aciertos fue de 73%, además se presenta el test de Huberty que permite comprobar la significación estadística de esta tasa y de esta manera contrastar las hipótesis anteriores. De este test se obtiene el número esperado de casos correctamente clasificados debidos al azar (e) que es igual a 48,341.82, a partir del cual se calcula el valor del estadístico Z^* que se distribuye normalmente, cuyo valor es igual a 26.17. Si se compara este valor con el valor de 1.96 obtenido de la tabla de una distribución normal para un nivel de significancia de 0.05, se tiene que el valor del estadístico $Z^* = 26.17 > 1.96$, por lo cual, se puede afirmar que existe evidencia estadística para rechazar la hipótesis nula, y se concluye que la tasa global de aciertos del modelo es significativamente mayor que la que se hubiera obtenido debido al azar. En otras palabras, el modelo seleccionado clasifica adecuadamente a los datos observados.

Tabla 8:
Tabla de clasificación de resultados

Observado		Pronosticado		
		SINIESTRO		Corrección de porcentaje
		NO	SÍ	
SINIESTRO	NO	43,778	10,574	80.5
	SÍ	6,295	1,827	22.5
Porcentaje global				73.0
e = 48,341.82			$Z^* = 26.17$	

En base a los resultados obtenidos se encuentra que el modelo de regresión logística permite estimar la probabilidad de que un vehículo sufra un siniestro dados los valores de las variables independientes, es:

$$p = \frac{e^{-2.69-0.15x_1+0.83x_2+0.70x_3+0.15x_4+0.28x_5}}{1+e^{-2.69-0.15x_1+0.83x_2+0.70x_3+0.15x_4+0.28x_5}}$$

Que en forma resumida se puede expresar como:

$$p = \frac{1}{1+e^{2.69+0.15x_1-0.83x_2-0.70x_3-0.15x_4-0.28x_5}}$$

Donde:
X_1: MARCA(1)
X_2: TIPO(1)
X_3: TIPO(2)
X_4: COLOR(1)
X_5: COLOR(3)

Como a los datos observados ya se les ajustó un modelo de regresión logística, lo que ahora sigue es su interpretación. Para esto se considerarán las categorías que tienen el valor de 1 en las variables DUMMY, ya que a partir de estas categorías se realizará la interpretación. En esta investigación, las categorías cuyas variables DUMMY tienen el valor de 1 son las siguientes: GM1, GT1, GT2, GC1 y GC3. Estas categorías están asociadas a las variables que resultaron significativas en el modelo.

En la tabla 7 se presentan los valores del Odds ratio, tal que, si el valor del odds ratio es mayor a 1, indica que es más probable que ocurra el evento de interés, HUBO SINIESTRO, en relación con la no ocurrencia de dicho evento; caso contrario, si este valor es menor a 1, indica que es menos probable que ocurra el evento de interés HUBO SINIESTRO. En base a esto se puede interpretar que:

Si un vehículo es de las marcas del grupo GM1, entonces es menos probable que sufra un siniestro a que sea de las marcas del grupo GM4.

Si un vehículo es del tipo SEDÁN o TODOTERRENO (GT1), entonces es más probable que sufra un siniestro a que sea del tipo CAMIÓN, MOTO u OTRO (GT3, que es la categoría de referencia).

Si un vehículo es del tipo CAMIONETA (GT2), entonces es más probable que sufra un siniestro a que sea del tipo CAMIÓN, MOTO u OTRO (GT3, que es la categoría de referencia).

Si un vehículo es de color PLATEADO, BLANCO, PLOMO o NEGRO (GC1), entonces es más probable que sufra un siniestro a que sea de color GRIS, CELESTE, AMARILLO, ACERO u OTRO (GC4, que es la categoría de referencia).

Si un vehículo es de color DORADO, BEIGE, CONCHO DE VINO o VERDE (GC3), entonces es más probable que sufra un siniestro a que sea de color GRIS, CELESTE, AMARILLO, ACERO u OTRO (GC4, que es la categoría de referencia).

7. CONCLUSIONES

Después de los resultados obtenidos sobre la siniestralidad en el ramo de seguros de vehículos en la ciudad de Quito, es necesario que las compañías aseguradoras efectúen estudios sobre este y los demás ramos que ofrecen, debido a la insuficiencia de los mismos. En las pólizas recogen información valiosa tanto del cliente como del vehículo, pero no cuentan con una herramienta que les ayude a discriminar el riesgo de sufrir un siniestro cuando otorgan una cobertura.

Este trabajo se basó en brindar conocimiento a través de modelos multivariados; en primer lugar, un modelo ARL que fue ajustado a los datos y permitió medir el riesgo asociado a las características de los vehículos; en segundo lugar, a través del ACM se establecieron las relaciones que existen entre las variables seleccionadas, lo cual generaría un enorme beneficio para las aseguradoras ya que pueden disponer de estos modelos para entre otras utilidades, decidir si se otorga una cobertura.

Comparando los resultados obtenidos en los dos modelos multivariados utilizados en el estudio, se determina que ambas técnicas identifican diferentes factores en relación con la presencia de siniestralidad, sin embargo, en el análisis de regresión logística, el modelo obtenido integró además el color y el tipo de vehículo, cuyas variables no fueron asociadas en el análisis de correspondencias múltiples.

8. REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

[1]. **Das, S., Sun, X.** (2016). *Association Knowledge for Fatal Run-Off-Road Crashes by Multiple Correspondence Analysis*. IATSS Research. Volume 39, Issue 2, 1 March 2016, Pages 146-155.

[2]. **Paefgen, J., Staake, T., Fleisch, E.** (2014). *Multivariate exposure modeling of accident risk: Insights from Pay-as-you-drive insurance data.* Transportation Research Part A: Policy and Practice. Volume 61, March 2014, Pages 27-40.

[3]. **Hemrit, W., Arab, M.B., Raissi, N.** (2013). *The correspondence analysis between the key indicators and events of operational risk: a case study of the insurance sector in Tunisia.* International Journal of Risk Assessment and Management. Volume 17, Issue 2, 2013, Pages 107-147.

[4]. **Pérez, C.** (2004). *Técnicas de análisis multivariante de datos*. Madrid: Pearson Prentice Hall.

[5]. **Luque, T.** (2000). *Técnicas de análisis de datos en investigación de mercados*. [Madrid]: Ediciones Pirámide.

[6]. **Álvarez, R.** (2008). *Estadística multivariante y no paramétrica con SPSS*. Madrid: Ediciones Díaz de Santos.