# Trabajo 1 Modelos Lineales

Ana María López Pedro Pablo Villegas

25 de septiembre de 2017

**Resumen**

El presente trabajo los resultados del mazda 2 publicado en la página https://www.tucarro.com.co consultados el 18 de agosto de 2017.
En total se consultaron 631 registros de Mazda 2 nuevos y usados con 28 variables cada uno

## 1. Introducción

Se descarcaron los datos de la página https://www.tucarro.com.co y se obtuvieron

## 2. Objetivos

Construir un modelo para determinar el precio de oferta de un vehículo Mazda 2 con base en información publicada en internet.

## 3. Metodología

## 4. Desarrollo

### 4.1. Recolección de Datos

La página web https://www.tucarro.com.co es hace parte de la plataforma MercadoLibre y es propiedad de la empresa MercadoLibre Colombia limitada; por lo tanto las publicaciones en tucarro.com.co obedecen las mismas reglas de negocio que mercadolibre.com.co. pero especializados en vehículos (motos, automóviles, camionetas, camiones, maquinaria pesada, carros de colección y vehículos náuticos) de todas las marcas, precios y años.

### 4.2. Análisis de Datos

Dado que las publicaciones siguen la lógica de MercadoLibre, son publicaciones libres que no son controladas y verificadas y sólo se solicita que las personas que deseen publicar algún vehículo diligencien ciertos campos predefinidos. Éstos son:

- Precio: Representa el valor de oferta del vehículo. La página permite que se publique en pesos o dólares (variable continua)
- Modelo/Año: Representa el modelo del vehículo (variable discreta ordinal)
- Ubicación: Representa el lugar dónde se ofrece el vehículo. La ubicación está compuesta por tres campos predefinidos:
  - Departamento (variable discreta nominal)
  - Ciudad (variable discreta nominal)
  - Barrio (variable discreta nominal)

- Color: Representa el color del vehículo. Es un campo con 16 opciones de colores preestablecidos (variable discreta nominal)
- Combustible: Representa la clase de combustible que usa el vehículo; hay tres opciones, diesel, gasolina, o gasolina y gas (variable discreta nominal)
- Recorrido: Representa los kilómetros recorridos por los vehículos ofertados, si el valor es cero se entiende que es un vehículo nuevo (variable continua)
- Único dueño: Campo donde se indica si el vehículo ha tenido más de un solo dueño, sólo hay dos opciones, si o no (variable binomial)
- Versión: Representa la versión del vehículo; si es una versión especial, full o sencilla. Es un campo libre que admite cualquier valor (variable discreta nominal)
- Frenos ABS: Representa si el vehículo tiene frenos ABS, sólo hay dos opciones, si o no (variable binomial)
- Aire: Representa si el vehículo tiene aire acondicionado, sólo hay dos opciones, si o no (variable binomial)
- Airbag: Representa si el vehículo cuenta con airbags, sólo hay dos opciones, si o no (variable binomial)
- Asientos: Representa el material con el que están tapizados los asientos del vehículo, hay tres opciones: cuero, semi-cuero y tela (variable discreta nominal)
- Cilindros: Representa el número de cilindros del motor. Es un campo libre que solo acepta valores numéricos (variable continua)
- Financiamiento: Representa el si el vendedor ofrece opciones de financiamiento, sólo hay dos opciones, si o no (variable binomial)
- Motor: Representa el tipo de motor del vehículo, es un campo libre (variable discreta nominal)
- Motor Reparado: Indica si el motor ha sido reparado últimamente, sólo hay dos opciones, si o no (variable binomial)
- Sonido: Representa el tipo de sonido que tiene el vehículo, hay cuatro opciones, CD, MP3, No y R/Rep (variable discreta nominal)
- Tracción: Representa la tracción del vehículo, sólo hay dos opciones, 4x4 o 4x2 (variable binomial)
- Transmisión: Representa el tipo de transmisión del vehículo, sólo hay dos opciones, Automática o Mecánica (variable binomial)
- Vidrios: Representa el tipo de vidrios que tiene el vehículo, sólo hay dos opciones, Eléctricos o Manuales (variable binomial)
- Seguridad: Representa el tipo de seguridad que tiene el vehículo. Es un grupo conformado por tres variables (si o no) cada una:
    - Alarma con Control (variable binomial)
    - Asegurado (variable binomial)
    - Rastreo Satelital (variable binomial)
- Placa: Representa el número de placa del vehículo. Es un campo libre (variable discreta nominal)
- Equipamiento: Representa el tipo de equipamiento adicional con que cuenta el vehículo. Es un grupo conformado por cuatro variables (si o no) cada una:
    - Bloqueo Central (variable binomial)
    - Forro del Volante (variable binomial)

    - Forro de Asientos (variable binomial)
    - Volante Deportivo (variable binomial)
- Sonido:Representa los elementos adicionales de sonido con que cuenta el vehículo. Es un grupo compuesto por cuatro variables (si o no) cada una:
    - Caja de CD's (variable binomial)
    - DVD (variable binomial)
    - Planta (variable binomial)
    - Sub-Buffer (variable binomial)
- Exterior: Representa los elementos decorativos adicionales con que cuenta el vehículo. Es un grupo compuesto por diez variables (si o no) cada una:
    - Estribos (variable binomial)
    - Forro Llanta de Repuesto (variable binomial)
    - Llantas Nuevas (variable binomial)
    - Luces Anti Niebla (variable binomial)
    - Película de Seguridad (variable binomial)
    - Retrovisores Eléctricos (variable binomial)
    - Revisión Tecnicomecánica (variable binomial)
    - Rines de Lujo (variable binomial)
    - Spoiler (variable binomial)
    - Sun Roof (variable binomial)

Como se observa, la base de datos descargada de la página web cuenta con cuarenta y cuatro variables, de las cuales sólo tres son continuas, treinta son binomiales y once son discretas nominales. Adicionalmente, del total de variables, sólo nueve (precio, modelo, ciudad, departamento, barrio, color, recorrido, transmisión y placa) son obligatorias, por lo que se presenta gran cantidad de datos faltantes en las no obligatorias.

Otro paso importante para el análisis de datos, fue cruzar la información descargada de la tucarro.com.co con las fichas técnicas de los diferentes modelos de Mazda 2 que se han comercializado en Colombia y observamos que el vehículo sólo se vende desde el 2008, todos los modelos y referencias son a gasolina, tracción 4x2, vidrios eléctricos, radio con MP3 y que cada modelo tiene la opción de ser automático o mecánico.

Algo que también se identificó fue que han habido tres generaciones de Mazda 2. La primera generación ha sido la única con referencias Sedan y abarca los modelos desde el 2008 hasta 2011 para las referencias Hatchback y desde el 2011 hasta el 2015 para las referencias Sedan. La segunda generación abarca los modelos desde el 2011 hasta el 2015. La tercera y última generación de Mazda 2 comprende las referencias 2016, 2017 y 2018.

### 4.2.1. Selección de variables

Como se mencionó anteriormente, sólo 9 de las 44 variables son obligatorias, por lo que se presentan datos faltantes superiores al 70 % en todas las variables no obligatorias por lo que se decidió no considerarlas para la construcción de modelos. Ésta decisión también está soportada con el estudio de las fichas técnicas de las diferentes generaciones de Mazda 2, ya que hay muchas variables que así sean faltantes tienen valores iguales como es el caso de tipo de combustible, tracción, vidrios, aire, aibag, audio, sun roof, motor y asientos.

Analizando las nueve variables obligatorias se decidió eliminar las variables *ciudad* y *barrio* debido a que las ciudades capitales de los departamentos concentran más del 90 % del total de las ofertas del departamento. También se decidió eliminar la variable *placa* ya que el campo aceptaba cualquier valor y en algunos casos teníamos la placa completa, en otros los últimos tres dígitos y

en otros sólo un dígito.

En conclusión, se seleccionaron 6 variables para analizar, éstas son:

1. Precio (variable continua positiva)
2. Departamento (variable nominal)
3. Modelo (variable discreta ordinal)
4. Color (variable discreta nominal)
5. Recorrido (variable continua)
6. Transmisión (variable binomial)

#### 4.2.2. Analisis descriptivo

- Se eliminó un registro con recorrido 99999999
- Se eliminaron 7 registros con transmisión vacía
- Se eliminó uno con 800.000 km de recorrido del 2011
- Se eliminó uno con 520.000 km de recorrido del 2012
- Se eliminó uno con precio 0

En total 620 datos

## 4.3. Estimación de parámetros

Debido a que el objetivo es construir un modelo para determinar el precio del Mazda 2, procedemos a plantear inicialmente un modelo de regresión que considere el precio como una respuesta a las 5 variables restantes.

$$Y = \beta_0 + \beta_{1,j} X_{1,j} + \beta_2 X_2 + \beta_{3,i} X_{3,i} + \beta_4 X_4 + \beta_5 I + \epsilon \tag{1}$$

Donde:
$X_{1,j}$ representa los departamentos de venta y está dado por:

$$X_{1,j} = \begin{cases} X_{1,1} = 1 & \texttt{si } X_{1,j} = \texttt{antioquia} \\ X_{1,2} = 2 & \texttt{si } X_{1,j} = \texttt{atlantico} \\ X_{1,3} = 3 & \texttt{si } X_{1,j} = \texttt{bogota} \\ X_{1,4} = 4 & \texttt{si } X_{1,j} = \texttt{bolivar} \\ X_{1,5} = 5 & \texttt{si } X_{1,j} = \texttt{boyaca} \\ X_{1,6} = 6 & \texttt{si } X_{1,j} = \texttt{caldas} \\ X_{1,7} = 7 & \texttt{si } X_{1,j} = \texttt{casanare} \\ X_{1,8} = 8 & \texttt{si } X_{1,j} = \texttt{cauca} \\ X_{1,9} = 9 & \texttt{si } X_{1,j} = \texttt{cesar} \\ X_{1,10} = 10 & \texttt{si } X_{1,j} = \texttt{cordoba} \\ X_{1,11} = 11 & \texttt{si } X_{1,j} = \texttt{cundinamarca} \\ X_{1,12} = 12 & \texttt{si } X_{1,j} = \texttt{huila} \\ X_{1,13} = 13 & \texttt{si } X_{1,j} = \texttt{magdalena} \\ X_{1,14} = 14 & \texttt{si } X_{1,j} = \texttt{meta} \\ X_{1,15} = 15 & \texttt{si } X_{1,j} = \texttt{narino} \\ X_{1,16} = 16 & \texttt{si } X_{1,j} = \texttt{norte\_santander} \\ X_{1,17} = 17 & \texttt{si } X_{1,j} = \texttt{quindio} \\ X_{1,18} = 18 & \texttt{si } X_{1,j} = \texttt{risaralda} \\ X_{1,19} = 19 & \texttt{si } X_{1,j} = \texttt{santander} \\ X_{1,20} = 20 & \texttt{si } X_{1,j} = \texttt{tolima} \\ 0 & \texttt{cualquier otro valor} \end{cases} \tag{2}$$

$X_2$ representa el modelo y está dado por:

$$X_2 \in \{2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018\} \tag{3}$$

$X_{3,i}$ representa el color del carro y está dado por:

$$X_{3,i} = \begin{cases} X_{3,1} = 1 & \texttt{si } X_{3,i} = \texttt{azul} \\ X_{3,2} = 2 & \texttt{si } X_{3,i} = \texttt{beige} \\ X_{3,3} = 3 & \texttt{si } X_{3,i} = \texttt{blanco} \\ X_{3,4} = 4 & \texttt{si } X_{3,i} = \texttt{dorado} \\ X_{3,5} = 5 & \texttt{si } X_{3,i} = \texttt{gris} \\ X_{3,6} = 6 & \texttt{si } X_{3,i} = \texttt{marron} \\ X_{3,7} = 7 & \texttt{si } X_{3,i} = \texttt{negro} \\ X_{3,8} = 8 & \texttt{si } X_{3,i} = \texttt{plateado} \\ X_{3,9} = 9 & \texttt{si } X_{3,i} = \texttt{cesar} \\ X_{3,10} = 10 & \texttt{si } X_{3,i} = \texttt{rojo} \\ X_{3,11} = 11 & \texttt{si } X_{3,i} = \texttt{verde} \\ X_{3,12} = 12 & \texttt{si } X_{3,i} = \texttt{vinotinto} \\ 0 & \texttt{cualquier otro valor} \end{cases} \tag{4}$$

$X_4$ representa los kilómetros recorridos y está dado por:

$$X_4 \in \{0 \to \infty\} \tag{5}$$

$I$ representa el tipo de transmisión y está dado por:

$$I = \begin{cases} 1 & \texttt{si } = \texttt{automatica} \\ 0 & \texttt{si } = \texttt{mecanica} \end{cases} \tag{6}$$

y representa los errores del modelo, los cuales distribuyen normal con media cero y varianza conocida e independientes:

$$\epsilon \sim N(0, \sigma^2) \tag{7}$$

De acuerdo al modelo anterior tenemos 36 parámetros a estimar.

# 5. Conclusiones

# Referencias