HERRAMIENTAS INFORMÁTICAS DE PRODUCTIVIDAD APLICADAS A LOS MÉTODOS CUANTITATIVOS: MODELO DE REGRESIÓN LOGÍSTICA.

Autores:

Bernal García, Juan Jesús. juanjesús.bernal@upct.es.

Dpto. de Métodos Cuantitativos e Informáticos. Universidad Politécnica de Cartagena

Escuder Vallés, Roberto. roberto.escuder@uv.es.

Departamento de Economía Aplicada. Universidad de Valencia.

Palacios Sánchez, Mª Angeles mangeles.palacios@upct.es.

Dpto. de Métodos Cuantitativos e Informáticos. Universidad Politécnica de Cartagena

Palabras clave: Regresión logística, Hojas de cálculo.

Resumen:

Es usual la aplicación de "modelos cuantitativos"como apoyo a la toma de decisiones, mediante el empleo de programas informáticos específicos; en nuestra investigación realizamos la búsqueda de innovadoras aplicaciones de las herramientas informáticas de productividad, utilizando las potencialidades de las hojas de cálculo, demostrando su validez para abordar la mayoría de los modelos de planificación empresarial.

En este caso se ha elegido un supuesto empírico que requiere el empleo de la *regresión logística***,** se trata de un parque de vehículos que pueden averiarse o no en función de unas características a considerar en dicha regresión, tendremos que elegir y categorizar las variables independientes, proceder seguidamente a la estimación de los parámetros y a contrastar los resultados obtenidos; ello nos proporcionará la probabilidad de que cada uno de los vehículos pueda averiarse o no durante un periodo determinado; todo ello debe ser programado mediante modelos realizados con una hoja de cálculo estándar.

Finalmente se apuntan posibles aplicaciones de los resultados obtenidos, demostrando así que su elaboración con hoja de cálculo es una opción que permite una gran potencia al tiempo que una mayor flexibilidad.

Introducción:

Siempre hemos propugnado recurrir a la *elaboración propia* de modelos informáticos, debido a la dificultad de que un programa estándar pueda adecuarse suficientemente a nuestra casuística, consecuentemente emplearemos la herramienta ofimática de productividad para el tratamiento de datos por excelencia, las denominadas hojas de cálculo (H.C.), que perfectamente podrían llamarse *"hojas de análisis de datos"* por las nuevas posibilidades que en este sentido se van incorporando en cada actualización, de forma que se posibilite la utilización de los métodos matemático-estadísticos a todas las PYMES, como apoyo a su toma de decisiones. En el presente trabajo nos hemos centrado en un modelo de **regresión logística**, estimación de parámetros y aplicación a un supuesto empírico, que deja patente las tremendas posibilidades que una adecuada programación de las citadas hojas de cálculo pueden ofrecernos, abriendo así nuevos cauces y procedimientos de análisis a acometer con esta herramienta de productividad.

Planteamiento del supuesto empírico:

Se trata de una empresa de alquiler de coches que dispone de una flota de ellos y desea realizar una previsión de las posible averías que puedan producirse, ya que ello la obliga a dejar de prestar un servicio a sus clientes y provoca por un lado una pérdida económica y por otro un deterioro de su imagen de empresa, que debe ofrecer calidad y seriedad. Se aborda el problema de las posibles averías a partir de una *regresión logística* que proporcione la probabilidad de que cada uno de los vehículos considerados pueda averiarse o no durante un periodo determinado. Los objetivos perseguidos, por tanto, con el supuesto práctico que vamos a plantear son fundamentalmente dos:

1.- Analizar que factores influyen en la posible avería de los vehículos.
2.-Conocer la probabilidad de que los vehículos considerados se averíen por mes.

Breve fundamentación teórica de la Regresión Logística:

Entre los *modelos de elección binaria*, que sirven para explicar una variable dependiente binaria (0/1), se encuentra la Regresión Logística, que se utiliza para predecir la probabilidad estimada P(Y) de que la variable dependiente (Y) presente uno de los dos valores posibles (1= sí 0 = no) en función de los diferentes valores que adoptan el conjunto de variables independientes $X_{i.}$ La función logística la podemos presentar por:

$$P(Y=1)=\frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n}} \qquad \text{o bien:} \qquad P=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}}$$

donde $ß_0$ es el término independiente o constante y $ß_i$ son los coeficientes de regresión asociados a cada variable independiente X_i.

Los parámetros de la ecuación de Regresión Logística se estiman por el *método de máxima verosimilitud*, a partir de la expresión matricial siguiente: $\beta_a=\beta+(X'VX)^{-1}X'(Y-\hat{Y})$; donde **V** es una matriz diagonal con términos $p_i(1-p_i)$ y $\hat{Y}$ el vector de valores esperados de **Y**; pudiendo aplicarse el siguiente algoritmo iterativo para obtener el estimador *MV* de ß[1]:

Paso 1º: Fijar un valor arbitrario inicial $ß_0$ (término independiente o constante), para los parámetros y obtener el vector $\hat{Y}_1$ para dicho valor en el modelo *Logit*. Si $ß_o$=0:

$\hat{y}_i=\hat{p}_i=\frac{1}{1+e^{-0}}=\frac{1}{2}$ y el vector $\hat{Y}$ tiene todas sus componentes iguales a 0,5.

Paso 2º: Definir una variable auxiliar z_i de *residuos estandarizados* por:

$z_i=\frac{y_i-\hat{y}_i}{\hat{y}_i(1-\hat{y}_i)}=\frac{y_i-\hat{p}_i}{\hat{p}_i(1-\hat{p}_i)}$ o vectorialmente: $Z=\hat{V}^{-1}(Y-\hat{Y})$

donde $\hat{V}$ es una matriz diagonal de términos: $\hat{y}_i(1-\hat{y}_i)$.

Paso 3º: Estimar por mínimos cuadrados ponderados una regresión con variable dependiente **Z**, regresores **X** y coeficientes de ponderación $\hat{y}_i(1-\hat{y}_i)$. Los parámetros estimados $\hat{b}_1$ vendrán dados por: $\hat{b}_1=(X'\hat{V}X)^{-1}X'\hat{V}Z=(X'\hat{V}X)^{-1}X'(Y-\hat{Y})$, donde se aprecia que b_1 estima el incremento de los parámetros que nos acerca al máximo.

Paso 4º: Obtener un nuevo estimador de los parámetros ß del modelo mediante:

$$\beta_1=\beta_0+\hat{b}_1$$

Paso 5º: Tomar el valor estimado resultante del paso anterior, que llamaremos $ß_h$ y sustituirlo en la ecuación del modelo logístico: $p_i=\frac{1}{1+e^{-\beta' x_i}}$, para obtener el vector de estimadores $\hat{Y}(\beta_h)=\hat{Y}_h$ y utilizando $\hat{Y}_h$ construir la matriz $\hat{V}_h$ y la nueva variable Z_h:

$$Z_h=\hat{V}_h^{-1}(Y-\hat{Y}_h)$$

---

[1] DAMORARN.GUJARATI. "Econometría" . 2ª Edición. McGraw-Hill.1981. Cap. 12 (pp. 367-404) y Cap. 13. (pp. 405-444)

El nuevo valor $ß_{h+1}$ será: $\beta_{h+1} = \beta_h + (X'\hat{V}X)^{-1}X'(Y - \hat{Y})$, donde el término de ajuste se calcula regresando, por $\boldsymbol{Z_h}$ sobre **X** con ponderaciones $\hat{V}_h$. El proceso se repite hasta obtener convergencia ($ß_{h+1} \cong ß_h$).

NOTA: Se han elaborado un organigrama al efecto (*Figura 1-ANEXO*). Aunque se ha elaborado otro para el caso de observaciones repetidas, no lo hemos incluido aquí.

En el modelo realizado hemos empleado el algoritmo matricial anterior, el problema reside en que las hojas de cálculo disponen de sólo 256 columnas, por lo cuál, no es posible operar con matrices de mayor número de columnas; por ello hemos recurrido a las "matrices particionadas por bloques". Veamos cómo sería preciso realizar dicha partición y cómo se transformarían las operaciones:

Matriz de las variable independientes : $(X_i)_{mxn} \rightarrow (X_{1p})_{m/2xn}$ y $(X_{2p})_{m/2xn}$:

$$(X)= \begin{pmatrix} X_{1p} \\ X_{2p} \end{pmatrix}$$

Vector $(Y - \hat{Y})$: $\rightarrow (Y_{1p})_{m/2xn}$ e $(Y_{2p})_{m/2xn}$:

$$(Y - \hat{Y}) = \begin{pmatrix} Y_{1p} \\ Y_{2p} \end{pmatrix}$$

V: Matriz diagonal de términos: $\hat{y}_i(1 - \hat{y}_i)$:

$$V= \begin{pmatrix} V_{1p} & 0 \\ 0 & V_{2p} \end{pmatrix}$$

con $(V_{1p})_{m/2xn/2}$.

De forma que las operaciones de forma secuencial serán:

La transpuesta de (X): $(X') = (X_{t1}\ X_{t2})$ donde $X_{t1}=(X_{1p})^t$ y $X_{t2}=(X_{2p})^t$.

$$PM1= (X')V = \begin{pmatrix} PM11 \\ PM12 \end{pmatrix}$$

Con $PM11= X_{t1}\ V_{1p}$ y $PM12= X_{t2}\ V_{2p}$

$$PM2= X'V\ X = PM21+PM22 = PM11\ X_{1p}+PM12\ X_{2p}$$

De forma que la inversa **:** $MI= (X`V\ X)^{-1}= (PM2)^{-1}$

$$PM3=(X'V\ X)^{-1}\ X'= \begin{pmatrix} PM31 \\ PM32 \end{pmatrix}$$

Con $PM31=MI\ X_{t1}$ y $PM32=MI\ X_{t2.}$

Finalmente:

$$PM4= (X'V\ X)^{-1}\ X'(Y-Y_0)=PM41+PM42=PM31\ Y_{1p}+PM32\ Y_{2p}= (ß)$$

De esta forma podremos operar hasta con 500 datos (250x2) haciendo una partición en dos matrices de 250 observaciones cada una, y reiterando el proceso podríamos seguir hasta la capacidad máxima de la hoja de cálculo (65.536 filas).

La bondad de ajuste del modelo:

La idoneidad del modelo la valoraremos siguiendo los dos criterios siguientes:

1. La medida de la bondad de ajuste.
2. La medida de la eficacia predictiva.

Para medir la bondad los "paquetes" estadísticos estándar[2] suelen analizar los siguientes estadísticos[3]:

1. –2 Log Likelihood (-2LLo)
2. Goodness of fit.
3. Model chi-square.
4. WALD.
5. Tabla de aciertos. Gráficas de clasificación.
6. Razones.
7. Residuos.

Los estadísticos *-2LLo* y *Goodness of fit* contrastan como hipótesis nula, que el modelo es significativo y, como hipótesis alternativa lo contrario. El estadístico *Goodness of fit* viene determinado por: $\sum_{i=1}^{N} \frac{E_i^2}{P_i(1-P_i)}$ donde:

E_i : i-ésimo residuo (diferencia entre la probabilidad observada y la estimada).

P_i : Es la probabilidad estimada del i-ésimo caso. *P*: Probabilidad observada.

---

2 Concretamente los del SPSS.

3 SATOS PEÑA, J., MUÑOZ ALAMILLOS, A. JUEZ MARTEL, P. y GUZMÁN JUSTICIA, L. "Diseño y tratamiento estadístico de encuestas para estudios de mercado". Ed. Dentro de Estudios Ramón Areces, S.A. 1.999. pp. 355-382.

Distribuyéndose como una chi-cuadrado con n-2 grados de libertad.

Para saber si las variables que introducimos en el análisis son o no válidas se usa el test de WALD; su estimación se presenta siempre al lado del valor del coeficiente, y junto a ella la probabilidad asociada a tal valor, si éste es inferior a 0,05 diremos que la variable es significativa, y válida para el modelo.

La fórmula para determinarlo será: $WALD = \frac{\beta_i}{EE(\beta_i)}$ con EE :Error Estándar del coeficiente. NOTA: En el algoritmo matricial hemos obtenido el EE para cada $ß_i$ calculando la raíz cuadrada de los términos de la diagonal principal de la matriz $(X'VX)^{-1}$.

La *Tabla de aciertos* es otra medida de la bondad del ajuste realizado, se mide el porcentaje de elementos de la muestra que eligen la opción predicha por el modelo; para ello consideraremos que predice como suceso 1 aquellos casos en que la función arroja una probabilidad superior a 0,5 y como 0 a aquellos casos en los que la probabilidad sea inferior a 0,5 (punto de corte). A partir de esta tabla se suelen determinar diversos porcentajes, como los de *Porcentaje de Verdaderos Positivos*: PVP, *Porcentaje de Verdaderos Negativos*: PVN, *Porcentaje de Falsos Positivos* : PFP y *Porcentaje de Falsos Negativos:* PFN.

Así mismo está la denominada ODDS RATIO (*OR*), medida del riesgo muy utilizada. La forma práctica de determinar el *Odds Ratio* de cada variable, con respecto de la de referencia, es elevar el número *e* al coeficiente de regresión logística de dicha variable: $ODDS\ RATIO = e^{\beta}$. Este valor nos explicará el grado en el que el aumento de una unidad de la variable (X_i), contribuye a aumentar o disminuir la probabilidad de la variable explicada (Y). Si el *OR* es mayor que la unidad, explicará que el aumento en una unidad de la variable considerada incrementa la probabilidad de ocurrencia del suceso, y la contribución será mayor cuanto más grande sea la cifra en cuestión. Algún programa estadístico determina además los intervalos de confianza (95% CI) inferior y superior mediante la expresión siguiente: $e^{ß \pm 1,95 * EE}$ con *EE* :Error Estándar del coeficiente.

Otro estadístico de diagnóstico es la medida de los valores residuales**,** cuya función es la detección de posibles defectos de predicción producidos en observaciones o sujetos en

los que el modelo no predice bien el valor observado (Y) para esa observación o individuo. A estos patrones atípicos se les suele conocer cómo *outliers*. Denominamos *residuo* a la diferencia entre la probabilidad observada (p_i) y la probabilidad estimada o predicha ($\hat{p}_i$): $Z_i = \frac{R_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$ con $R_i = p_i - \hat{p}_i$ como residuo i-ésimo.

Etapas de elaboración de un modelo de regresión logística:

Para elaborar uno de estos modelos, se aconseja seguir las siguientes etapas:

1ª etapa: Determinar el fenómeno que se desea explicar (variable dependiente).

2ª etapa: Buscar variables que permitan explicar la variable dependiente. Pueden ser de varios tipos: Cualitativas de dos niveles (0/1) o Cualitativas de más de dos niveles y Cuantitativas, estas últimas deben categorizarse (o dicotomizarse) mediante subdivisión en (n-1) variables "indicadoras" o "*dummys*", una variable por categoría menos una que queda como referencia. No obstante, al tratar de categorizar una variable continua se asume el riesgo de perder información o modificarla

3ª etapa: Resolución del modelo y determinación de los coeficientes de las variables.

4ª etapa: Analizar la bondad del modelo hallado.

5ª etapa: Interpretación de los resultados (ALBERT J. JOVELL, nos recomiendan, sobre todo, tener en cuenta la replicabilidad de los mismos).

Existen diversas estrategias con sus correspondientes algoritmos para seleccionar las variables a incluir en el modelo, dependiendo de si se realiza una incorporación progresiva o inclusión secuencial de variables hacia delante (*forward)*, según su nivel de significación (normalmente > 0,05), o por el contrario procediendo a una eliminación progresiva de variables hacia atrás (*backward).*

Tratamiento informático del modelo de Regresión Logística:

Para la resolución de los modelos de regresión logística ha de recurrirse a programas informáticos, bien de propósito general estadístico como puede ser el SPSS el S-PLUS, o el STATA, o bien específicos de econometría tipo EVIEWS. En el supuesto práctico hemos utilizado la salida con los resultados, tanto con SPSS (ver. 9.0) como con EVIEWS (ver. 3.0), con el fin de compararlos con los valores que obtengamos con el modelo que vamos a programar en H.C.

Consecuentes con el resumen de este trabajo, hemos elaborado unos modelos con H.C., ya estas herramientas aún no disponen de la posibilidad de realizar tratamientos de tipo multivariante, por eso hemos querido *afrontar el reto* de acometer la formulación y la programación de estas técnicas, comenzando por el de la regresión logística. Las ventajas son evidentes, por su mayor accesibilidad y menor coste que los programas estadísticos mencionados, amén de su sencillo manejo; pero existen otras razones que podemos resumir en:

1.-Poder disponer de los distintos cálculos intermedios del proceso.

2.-Mayor inmediatez en modificar datos y recalcular el modelo.

3.-Poder realizar otros estudios –analíticos y gráficos- utilizando opciones avanzadas de análisis de datos de las actuales H.C.

4.-Facilitar tanto la entrada y preparación de datos cómo la calidad de presentación de los resultados.

5.-Poder conectar los resultados de este modelo con otros modelos también elaborados con el programa de hoja.

Modelos para la Regresión Logística elaborados con H.C:

El modelo elaborado consta de tres módulos bien diferenciados pero complementarios:

Módulo 1.- Entrada, depuración y preparación de datos para la regresión logística.

Módulo 2.-Estimación de los parámetros ß (por distintos métodos).

Módulo 3.-Evaluación de la bondad del modelo.

Módulo 1: Preparación de datos:

Tras la adquisición o exportación de datos en formato compatible con *Excel*, el primer módulo se encarga de seleccionar los valores deseados (mediante *Autofiltro*) y de realizar las distribuciones de frecuencias para cada posible variable independiente (X_i) para la categorización de los valores de las mismas. Unos gráficos de barras ayudan a visualizar mejor estas distribuciones.

Seguidamente se realiza la dicotomización de las variables a considerar; seguidamente, su incorporación a la variable (Y) (más el vector de unos) nos proporciona la matriz de datos a procesar. Otro modelo se encarga de realizar "el conteo" de casos repetidos para determinar los n_i y las correspondientes p_i, para la posible aplicación del método para el caso de observaciones repetidas.

Módulo 2º: Estimación de parámetros:

Para la estimación de la matriz (**ß**), hemos elaborado distintos tipos de hojas de cálculo:

1.-Modelo de hoja para observaciones no repetibles con iteraciones multihoja.

2.-Modelo de hoja para observaciones no repetibles mediante macros de VBA(1).

3.-Modelo de hoja para observaciones repetibles (2)

(1) El citado método matricial particionado se ha programado en una sola celda mediante fórmulas de Excel de la siguiente forma:

```
=MMULT(MMULT(MINVERSA(MMULT(MMULT(TRANSPONER(MatrizX1);
MatrizV1);MatrizX1)+MMULT(MMULT(TRANSPONER(MatrizX2);MatrizV2);
MatrizX2));TRANSPONER(MatrizX1));MatrizY_Y01)+MMULT(MMULT
(TRANSPONER(MatrizX1);MatrizV1);MatrizX1)+TRANSPONER(MatrizX2)+MatrizV2
```

(2) Para realizar la Regresión Lineal en este caso, se ha empleado la expresión:

$$(X'X)^{-1}X'Y$$

Formulada en Excel mediante:

```
=MMULT(MMULT(MINVERSA(MMULT(TRANSPONER(xo);xo));TRANSPONER(xo));y)
```

Se han construido dos tipos de modelos, uno donde cada iteración de la estimación conecta con la siguiente hoja del mismo "libro", y otro realizado con macros de VBA (*Visual Basic para Aplicaciones*). El primero de ellos tiene la ventaja de no necesitar conocer el lenguaje de los macros y presentar más cálculos intermedios, por el contrario, el segundo cuenta con la gran cualidad de estar bastante más "compactado" y una más sencilla utilización posterior.

El modelo además de servirnos para calcular las probabilidades estimadas $(\hat{p}_i o \hat{Y})$, permite introducir valores concretos para las variables independientes estimadas, y de forma automática informa de la probabilidad de que el suceso ocurra (Y=1). También se han elaborado sendas tablas y gráficos donde se puede visualizar la evolución de la probabilidad de ocurrencia del suceso objeto de estudio, según cambia la variable en cuestión.

Módulo 3ª: Contraste del modelo:

El tercer módulo del programa es el que se encarga de medir la bondad del modelo realizado, haciendo los siguientes cálculos y gráficos:

1. Tabla y gráficos comparativos entre la proporción de "unos" en la muestra y la predicción.
2. Tabla de clasificación o concordancia con el porcentajes de aciertos para (Y), así como su promedio.
3. Cálculo de las razones PVP, PVN, PFP, PFN, VPP, VPN, VPN, PFN y PRN.
4. Tabla con el valor del OR para cada variable (más la constante), con sus intervalos inferior y superior de confianza del 95%.
5. Tabla para las variables y la constante, con los valores de error estándar, el estadístico de WALD y correspondiente significación N(0,1).
6. Estadístico ρ^2 (tipo R^2).
7. Valor del *Goodness of Fit*, con su correspondiente contraste χ^2, que nos informa de modo automático si el modelo supera este test.
8. Valor de la función de máxima verosimilitud para ß=0 (o restricción de *Log. Likelihood*).
9. Nº de casos residuales o frontera (para corte 0,5) con su porcentaje frente al total de observaciones y creación de una tabla con la extracción de dichos casos.
10. Gráfica con la distribución de frecuencia de la probabilidad estimada (intervalo 0-1 y escalones de 0,1 en 0,1) para observar la concentración de los mismos.

Además, se han realizado los siguientes análisis:
a.- Determinación del "punto de corte " que maximiza la proporción de aciertos.
b.-Comprobación de la sensibilidad del modelo al variar el número de observaciones.

Investigación empírica: *Modelo de regresión logística para predicción de averías*:

Planteamiento del problema:

Cuando se utilizan máquinas, o coches (como en nuestro caso), es frecuente que se den una serie de averías de forma aleatoria, en lógica proporción a la antigüedad y desgastes de los mismos; a la hora de estimar estas posibles averías, podemos en función de históricos de ocurrencia de las mismas, determinar con que probabilidad pueden ocurrir por unidad de tiempo.

Concretamente, la empresa dispone de una flota de 50 coches para su alquiler sin conductor, que podrán averiarse o no, en función de sus características (que serán factores a considerar en la regresión logística); primeramente deberemos realizar un estudio, basado en una base de datos amplia, que nos permita elegir y categorizar las variables independientes, para proceder seguidamente a la estimación de los parámetros correspondientes y a contrastar los resultados obtenidos, que serán aplicados a la evaluación de la flota concreta objeto de estudio.

Diseño del proceso:

A fin de responder a los objetivos planteados, deberemos realizar los siguientes pasos:

Paso 1º: Estudio de los datos de partida y depuración de los mismos.

Paso 2º: Preparación de los datos para la regresión logística.

Paso 3º: Estimación de los parámetros.

Paso 4º: Medida de la bondad de los modelos de regresión logística y de su capacidad predictiva.

Paso 5º: Comparación de resultados y elección del modelo más idóneo.

Paso 1º: Datos de partida. Depuración de los datos:

Con el fin de disponer de la información necesaria para poder realizar la estimación logística, se partió de una base de datos sobre las características de los vehículos que acuden al taller de un concesionario de una conocida marca de automóviles. En primer lugar fue preciso realizar una selección de datos, eliminando las visitas al taller que lo fueron para una revisión y no por una avería, ya que las citadas revisiones son de tipo periódico y nos interesan solo las averías de tipo aleatorio.

Se realizaron cálculos intermedios (cómo el tiempo entre averías y antigüedad del vehículo) y se seleccionaron aquellos registros de averías producidas en un mes, (concretamente se seleccionaron las ocurridas entre el 1/2/2000 y el 1/3/2000); también se eliminaron datos extremos y poco significativos (p.e. vehículos con más de 150.000 km., ya que estos acuden con menor frecuencia al taller oficial, pudiéndose dar la paradoja de que coches con tantos kilómetros sean los que menos acuden a él). Para seleccionar y/o eliminar datos, sugerimos cómo procedimiento la opción de "Autofiltro" de Excel. Finalmente se consiguió una población compuesta por 200 observaciones.

Paso 2º: Preparación de los datos para la regresión logística:

A continuación, para elaborar el Modelo de regresión logística, se realizó un estudio de cada uno de los conceptos que se consideraban significativos *–a priori-* a la hora de estar relacionada con la avería (Y=1), y por tanto formar parte como variables dependientes (Xi) de la regresión logística:

1. La antigüedad (en años).
2. Los kilómetros.
3. La gama (Ga: Gasolina/DI: Diesel).

Aunque en un principio también se analizaron otras posibles variables, como "la diferencia" de tiempo desde la última avería y el número de averías producidas con anterioridad en el mes considerado, pero, tras el estudio de frecuencias y categorización realizado a continuación, se descartaron por no presentar una relación coherente con la probabilidad de avería. En la *Figura 2* se muestran los 14 primeros valores, donde se incluye la columna con la variable dependiente (Y=avería), el número de averías producidas en las doscientas observaciones (el 10,5%), así como los valores mínimo, máximo y medio de cada variable independiente.

Con los datos de la tabla anterior, se escogieron distintos intervalos para cada variable y sus consiguientes "conteos" de frecuencia; primeramente para la antigüedad, y posteriormente para las variables kilometraje y Gama; mostrándose el valor porcentual del nº de casos de cada intervalo en sendos gráficos de sectores. (*Figura 3).* Se observó que hay un número mayor de vehículos con menos de un año de antigüedad, la mayoría tienen menos de 50.000 Km, y que es mayor el número de ellos son del tipo diesel.

A continuación se procedió a realizar una hoja de cálculo que categoriza a las variables y realiza un conteo de cuántas averías [P(Y=1)] y de no averías [P(Y=0)] se dan para cada categoría de la variable independiente; así, para la antigüedad, se escogen cuatro posibles categorías, de forma que si hacemos binaria dicha variable, tendremos la tabla que nos muestra la *Figura 4;* si se toma como referencia la última categoría (Ant1, Ant2 y Ant3), podremos tomar las tres primeras para construir la tabla correspondiente a las 200 observaciones (*Figura 5-tabla recortada*). Además de categorizar la antigüedad podemos contar cuántas averías se han producido [Y=1] y elaborar a partir de la misma una matriz con los ceros y unos para cada una de dichas categorías. (*Figura 6*). Se

constata que aunque el número de averías disminuye al amentar el número de años del vehículo, va creciendo –sensiblemente- la probabilidad de acudir al taller de avería según se cambia de una categoría con menos años a otra mayor

Una categorización análoga (de 0 a 20.000 Kms, de 20.000 a 50.000 Kms, de 50.000 a 90.000 Kms. y de 90.000 a 150.000 kms), se ha realizado para la variable independiente Kms., y finalmente se realizó a tabla y gráfica correspondiente a la variable Gama (0/1).

Otra hoja de cálculo realizada permite además, componer todas las variables categorizadas y realizar el conteo de cuantos casos reiterados nos encontramos, con el fin de poder aplicar con posterioridad un "modelo de regresión logística para observaciones repetidas", supuesto que en nuestra investigación empírica no es posible emplear, ya que el número de casos distintos –al combinar las tres variables- es tan elevado (64) que al realizar el conteo en las 200 observaciones, nos encontramos con que alguno de ellos tiene frecuencia cero (*Figuras 7-Recortada*). NOTA: Para realizar dicha cuenta aconsejamos utilizar la función =DBCONTAR de *Excel*.

Paso 3º: Estimación de los parámetros:

Hemos querido, no sólo elaborar un modelo para el caso específico que estamos analizando, sino toda una serie de modelos diversos, de forma que se disponga de las distintas opciones posibles para aplicar en supuestos posteriores. Se analizo el caso de que se disponga solamente de una variable independiente, luego dos y finalmente los tres; ello nos permite por un lado conocer cómo aplicar los distintos modelos en el caso de contar con menos variables que en nuestro caso, y por otro el poder analizar cómo evoluciona la bondad de la regresión al ir añadiendo unas u otras variables a la regresión.

Comenzaremos por tanto, por tratar una sola variable, por ejemplo la "Gama", que puede tomar los valores 1:Diesel y 0:Gasolina. Se realizó un modelo para "observaciones repetidas" (mediante una regresión lineal), más, por brevedad, mostramos aquí el modelo que hemos programado mediante el método matricial (con matrices particionadas), ya que sólo precisa la introducción de los datos de la variable/s independientes a considerar, al que añade la columna de unos (para estimar el término independiente) (*Figura 8*) y mediante la activación del macro "Calcular" realiza las

iteraciones convenientes para que converjan los valores de $ß_{i.,}$ proporcionando el valor de dichos estimadores, el valor de función logit de probabilidad para las 200 observaciones de la muestra y el EE (error cuadrático) correspondiente a cada $ß_{i.}$ En este caso el proceso "se detiene" en la sexta iteración (*Figura 9-recortada).*

No obstante, si queremos conocer los resultados parciales de todas las iteraciones, también se ha elaborado un modelo multihoja - que va realizando cada nueva iteración *i* (una nueva estimación que toma como y_0 el valor y_{i-1} de la anterior) en una nueva hoja del libro de Excel . A partir de la primera iteración (y_0=0,5), se crea la matriz ($\mathbf{V}_{200x200}$) (*Figura 10*) y aplica el método matricial "paso a paso".

Una vez realizados los ajustes con una sola variable independiente, se acometió el de dos variables conjuntamente, realizándose finalmente con las tres variables consideradas, tanto en forma categorizada como sin estarlo. Los resultados obtenidos se recogen –de forma automática- en una hoja denominada RESUMEN (*Figura 11*).

Paso 4º: Medida de la bondad de los modelos y de su capacidad predictiva:

Para contrastar los resultados de la estimación de parámetros de regresión logística, hemos elaborado un modelo denominación CONTRASTE, que consta de dos partes bien diferenciadas, la primera que realiza una serie de cálculos intermedios y la segunda que contiene el cálculo de los estadísticos, tablas y gráficas que vamos a utilizar.

Para explicar el citado modelo se ha tomado un caso concreto, que incluye las variables *Gama* (Ga/DI), *Kms/10.000* (continua) y *Antigüedad* en años (Categorizada), de forma que son "capturados" los resultados obtenidos para dicho caso, la tabla de los coeficientes ($ß_i$), la $\overline{y}_i$ estimada (que denominaremos p_i) para la totalidad de las observaciones (200) y la tabla de EE($ß_i$). A partir de dichos valores la tabla realiza los citados resultados intermedios (*Figura 12*):

p_i*:* Valores de la función logit estimada.

p_i binario: La conversión de los valores anteriores a forma binaria, mediante la comparación con el "punto de corte" elegido, que por omisión es de 0.5 ($p_i \geq 0{,}5 \rightarrow 1$).

y_i*:* Variable Y observada (Y =1 avería).

Con1, Con2, Con.3 y Con 4: Para contar las concordancias que se producen entre la probabilidad observada y la predicha (por ejemplo "C11" proporciona el nº de casos en que $p_i = 1$ *y* $y_i = 1$.)

La *Figura 13* nos muestra otros cálculos intermedios: El valor de los residuos ($y_i - p_i$), valor que elevado al cuadrado y sumado nos proporciona el valor de *SR*, mientras que la siguiente columna realiza la diferencia entre el valor observado y su valor medio, obteniéndose *ST* como suma de las doscientas diferencias de este tipo existentes. Bajo el nombre de G, están los cálculos que sirven para determinar el *Goodness of Fit*,; y finalmente los valores de Z_i, como cociente entre los residuos y la raíz de $p_i (1 - p_i)$; para contar cuántos casos se dan en que el valor de Z_i supera en valor absoluto a 1,5.

A partir de la tabla intermedia expuesta, pasemos a ver las distintas medidas que nos van a permitir medir la bondad del ajuste:

A)*Yi versus Pi:* En primer lugar comparamos, tanto de forma analítica como gráfica (*Figura 14*), el número de "ceros" y "unos" existentes en la muestra y en la predicción.

En principio se ha considerado un punto de corte (PC) de 0,5, pero hemos planteado con ayuda del "Buscar Objetivo", cuánto debería ser el valor del PC que hiciese coincidir el número de ceros y unos en la estimación con los observados, el valor obtenido es de 0,37; dicho valor aseguraría una magnífica capacidad de predicción de averías.

B)*Tabla de Concordancias*: Para medir mejor la bondad predictiva del modelo, realizaremos la tabla de concordancia (o de clasificación), midiendo los porcentajes de aciertos en unos y ceros, y realizando el promedio de ambos casos. Al tiempo, el modelo calcula las distintas probabilidades sobre verdaderos y falsos, así como las distintas ratios de predominio (PRF y PRN). (*Figura 15*). Hemos elaborado una pequeña "macro" ("botón" PC) que calcula el valor de dicho promedio para distintos valores del punto de corte y construye la tabla y gráfica correspondiente. (*Figura 16*).

También se realiza un gráfico que hemos llamado de concentración, dónde se encuentra la mayor cantidad de valores de la predicción y en que punto interesa situar en consecuencia el citado punto de corte.

C)*Bondad de ajuste:* Bajo esta denominación hemos determinado una serie de estadísticos de los expuestos en la fundamentación teórica, vemos cuales (*Figura 17*):

- Tipo ρ^2=1-SR/ST
- El *Goodness of Fit*, con prueba χ^2 y diagnostico automático de su cumplimiento.
- La restricción de máx. verosimilitud (L_0) para ß $_i$=0 .
- Para cada variable independiente : OR e Intervalos ± 95%OR .
- Tabla que contiene los errores estándar para cada estimador, el estadístico de Wald y su significación.

D.-*Valores residuales o frontera:* Obtenidos "filtrando" los valores de Z_i que superan, en valor absoluto, el 1,5 (*Figura 18*). En este caso si eliminamos de las 200 observaciones los seis valores "frontera", crecerá ρ^2 y mejorará el contraste.

Queremos comentar también que medimos la sensibilidad de los resultados obtenidos, al disminuir el tamaño de la muestra, pudiendo informar que realizadas sendas prueba para 150 y 100 observaciones (en lugar de las 200), no se apreció ninguna disminución de los estadísticos del contraste, ni en la capacidad predictiva del modelo realizado.

Con el fin de comprobar que tanto los modelos de estimación de parámetros como los de contrate realizados con la hoja de cálculo funcionan de forma adecuada, se han contrastado los resultados con los obtenidos con otros programas estadísticos, concretamente con el SPSS y el Eviews. Se muestra a continuación parte de la "salida" obtenida con el citado programa SPSS (que coinciden con los del modelo con H.C.):

SPSS Regresión logística

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R
X1	,3871	,1141	11,5120	1	,0007	,2661
X2	2,1267	,7425	8,2045	1	,0042	,2149
X3	-7,8938	2,0905	14,2580	1	,0002	-,3020
X4	-5,7465	1,5310	14,0882	1	,0002	-,2999
X5	-2,7642	1,3441	4,2294	1	,0397	-,1288
Constant	-,9304	1,3404	,4818	1	,4876	

Paso 5º: Comparación de resultados y elección del modelo más idóneo:

Seguidamente se compararon los resultados obtenidos para decidir sobre cuál de ellos seleccionar; se recogieron en una tabla resumen las principales medidas de bondad de ajuste de cada uno de ellos, pudiéndose entresacar las conclusiones que se detallan:

1.- La bondad de ajuste mejora conforme vamos añadiendo cada una de las variables seleccionadas para el estudio, por tanto incluiremos las tres variables.

2.- El modelo que incluye las tres variables y que de forma conjunta presenta mejor contraste además de mayor capacidad de predicción, resultó ser el que toma los kilómetros (divididos por 10.000), la variable dicotómica gama (gasolina/diesel) y la antigüedad (en años) y categorizada en cuatro variables indicadoras.:

MODELO	Promedio de Aciertos	Tipo Ro^2	G. of Fit	Sig(Wald)	% Residual	Descripción Modelo
MGD	50,00%	0,48%	No	No	10,0%	Gama
MKM	64,58%	8,84%	No	Si	8,0%	Kms.
MAN	77,47%	30,46%	No	Si	5,5%	Antigüedad
MANca	82,65%	33,31%	No	Si	5,0%	Antigüedad (Cat.)
MKMG	64,86%	8,29%	No	No	9,0%	Kms+Gama
MAG	83,07%	41,27%	Si	Si	3,0%	Anti+Gama
MAK	72,87%	14,78%	Si	No	6,0%	Anti+Kms
MGAK	83,07%	41,27%	Si	Si	3,0%	Gama+Anti+Km
MKGAc	**85,45%**	**42,85%**	**Si**	**Si**	**6,0%**	**Gama+Anti (cat)+Km**
MKcGAc	87,28%	-15,22%	No	Si	12,0%	Gama+Anti(cat)+ Km(cat)

Otros modelos realizados:

También se elaboró otro modelo con H.C. "PREDICCIONES", que permite determinar la probabilidad de avería para unos datos dados, permitiendo predecir la posibilidad de una avería para un vehículo determinado, pero también puede utilizarse para conocer los valores límites de cada variable que provoca el que "bascule" de avería "sí" a avería "no" de acuerdo con el punto de corte elegido (*Figura 19*).

Aplicación a los datos concretos de la flota a evaluar:

La empresa dispone de **50** coches y de ellos debemos de disponer de las características que se consideran en la regresión logística realizada: *Antigüedad* (en años), *Gama* (gasolina o diesel) y *Kilometraje.*

Para determinar el nº de averías que pueden producirse en estos vehículos, recurriremos a la estimación de la regresión logística antes realizada (concretamente hemos utilizado el modelo seleccionado que categorizaba la antigüedad y punto de corte de 0,39), quedando de la forma en que se presentan en la *Figura 20* . En la siguiente *Figura 21*, se encuentran los resultados, es decir, el número de vehículos que se prevén que se averíen por mes, que supone un total de seis coches, o lo que es lo mismo, una p_i= 12%.

Para estimar las avería que se pueden producir los siguientes meses, hemos previsto dos posibilidades, una de ellas necesita la introducción de los nuevos datos con los nuevos valores de kilometraje que llevan los 50 vehículos, mientras que la antigüedad puede ser incrementada de forma automática por el modelo. La segunda opción más automática, y que puede resultar más conveniente por ser más restrictiva y garantizar mejor los fallos por avería, consiste en considerar que el número de averías real que puede producirse de forma aleatoria, sigue una distribución normal de media el valor estimado anteriormente, procediendo a continuación a realizar tiradas aleatorias según dicha distribución, que en el caso considerado se tratará de una N(6, 2.3) (*Figura 21*).
NOTA: En Excel se programaría de la siguiente forma:

=ABS(REDONDEAR(DISTR.NORM.INV(ALEATORIO(); Media; Desviación Típica);0))

Posible aplicación posterior:

Dadas las buenas cualidades predictivas del modelo realizado, podría servirnos, por ejemplo, para utilizar su resultados en un modelo de Simulación de averías-mantenimiento, de forma que podamos determinar el número óptimo de mecánicos a emplear, de acuerdo con un coste total mínimo, que considere tanto el coste de dicho mantenimiento, como el coste derivado de la no utilización por las averías.

Conclusiones:

Creemos que ha quedando patente que la metodología presentada puede extenderse hasta donde nosotros deseemos, quedando abiertos los caminos, tanto de la realización con hoja de cálculo de las técnicas multivariantes, como de la aplicación de los modelos cuantitativos al análisis de datos en concreto y a la planificación de la empresa en general.

Bibliografía :

AGRESTY, A. "An Introduction to Categorical Data Analysis. Wiley NY 1.996.

ALCAIDE, A y ÁLVAREZ, N. "Econometría. Métodos Determinísticos y Estocásticos". Ed. Centro de Estudios Ramón Areces.

ALDRICH, J.H. y NELSON, F.E. Linear Probability, Logit and Probit Models. Beverley Hills, California, Sage Publications, 1984.

AMENIYA, TAKESHI."Modelos de respuesta cualitativa: Un examen". Cuadernos Económicos del I.C.E. nº 39. 1988/2. pp. 173-245.

B. FOMBY, T, CARTER HILL, R Y STANLEY, R.J. "Advanced Econometric". Ed. Springer-Verlag. 1.984. Cap. 10: Autocorrelación (Maximum Likelihood Estimation:AR). (pp. 205-236).

CARRASCAL, URSICINO. "Significado de los estimadores en un modelo Logit de variables dependientes cualitativas múltiples". Anales de Estudios Económicos y Empresariales". 1997. nº 12. pp 135-144.

CARRASCO, J. L. y HERNÁN MIGUEL, A. "Estadística multivariante en las Ciencias de la Vida". Ed. Ciencia 3. 1.993. pp. 197-245

DAMORARN.GUJARATI. "Econometria" . 2ª Edición. McGraw-Hill.1981. Cap. 12: Regresión con una variable docotómica. (pp. 367-404) y Cap. 13: Regresión en una v.a. dicotómica: Modelos MPL, LIGT y PROBIT (pp. 405-444)

DEMARIS, A. "Logit Modeling". London: Sage. 1.992.

FERRANDO BOLADO, M. y BLASCO RAMOS, F." La previsión del fracaso empresarial en la Comunidad Valenciana: Aplicación de los modelos Discriminante y LOGIT". Revista Española de Financiación y Contabilidad. Vol. XXVII, nº 95. Abril-junio 1998. pp-499-540.

GREENE, W.H. Análisis Econométrico. Cap. 19: Modelos con variables dependientes discretas (pp. 749-815).Prentice Hall. 1998. Madrid.

Guía SPSS "Modelos de Regresión". SPSS Inc. 1999

HOSMER, D.W. y LEMESHOW, S. "Applied Logistic Regression". N. York: John Wiley and Sond. 1.989.

JOHNSTON. Métodos de Econometría". Ed. Vicens Universidad. 1977. Cap. 7: Mínimos Cuadrados Generalizados (pp. 221-255).

JOBSON, J.D. "Applied Multivariate Data Analysis. Vol. 2 .NY Springer-Verlas. 1.992

JOVELL, A.J. “Análisis de Regresión Logística”. Cuadernos metodológicos 15. Ed. Diáz de Santos. Centro de Investigaciones Sociológicas. 1.995.

MADDALA, G.S. “Econometría”. McGraw-Hill. 1985.

McCULLOCH, CHARLES E. ”Maximun Likelihood Algorithms for Generalized Linear Mixed Models”. “Journal of the American Statistical Association”. Marzo 97, Vol 92. Issue. 437.

MILLÁN,J.A. y RUÍZ, P.“Modelos Logit de adopción de innovaciones en invernaderos de Almería”. Investigación Agraria, nº2. Vol. 2.Diciembre. (pp. 115-125). 1.987.

NOVALES, A. “Econometría”. Mc.Graw-Hill. 1998. Cap.14: variables dependientes cualitativas. (pp. 354-374).

PHOEBUS J. DHRYMES. Econometría. Ed. AC. Cap 7: Modelos de elección discreta: Análisis Logit y Probit. (pp. 322-350).

SANTOS PEÑA, J., MUÑOZ ALAMILLOS, A, JUEZ MARTEL, P. y GUZMÁN JUSTICIA, L. “Diseño y tratamiento estadístico de encuestas para estudios de mercado”. Ed. Dentro de Estudios Ramón Areces, S.A. 1.999. pp. 355-382.

SWAIT, JOFFRE y LOUVIERE, JORDAN. “The role of the scale parameter in the estimation and comparison of multinomial *logit* models”. JMR: Journal of Marketing Research. Aug.93. vol. 30.Issue.3. (pp. 305-315).

VIDAL DÍAZ DE RADA. “Técnicas de análisis de datos para investigadores sociales. Aplicaciones prácticas con SPSS para Windows”. Ed. Ra-ma. 1.999. pp. 223-253

XIE, X. y MANSKI, C.H. “The Logit Model and Response-Based Samples”. Sociological Methods Research. Vol. 17, nº 3. Febrero 1.989. pp. 283-302.