Análisis de datos con el Modelo Lineal Generalizado. Una aplicación con R

por Emelina LÓPEZ-GONZÁLEZ y Marcos RUIZ-SOLER
Universidad de Málaga

Desde una perspectiva matemática, el fundamento de gran parte de las pruebas estadísticas se encuentra en el Modelo Lineal (ML) general o clásico. Su importancia radica en que su estructura, suponemos, refleja los elementos explicativos de un fenómeno por medio de relaciones funcionales probabilísticas entre variables. El Modelo Lineal Generalizado (MLG), que tratamos en el presente trabajo, es la extensión natural del Modelo Lineal clásico. Inicialmente expuesto por Nelder y Wedderburn (1972), ha llegado a suponer "una auténtica revolución estadística" (Ato y Vallejo, 2007, 53), convirtiéndose en una solución especialmente adecuada para modelos de dependencia con datos no métricos.

En los estudios en educación es frecuente trabajar atributos, actitudes o conductas que, siendo en su dimensión latente continuos, se miden de forma no métrica (discreta, nominal u ordinal), no ajustándose, en consecuencia, al Modelo Lineal clásico e incumpliendo los supuestos de linealidad y normalidad. Ejemplos son la clasificación binaria de apto-no apto, la medida ordinal del curso académico, el recuento del número de años cursados, etc., situaciones que requieren de modelos que trabajen con datos dicotómicos, ordinales, categóricos o de elecciones discretas, es decir, de modelos de probabilidad de un evento (fundamentalmente modelos logit, probit, modelos de regresión de Poisson y modelos de regresión ordinal). Estos modelos son parte integrante de los Modelos Lineales Generalizados y, junto con la regresión lineal, el análisis de varianza, la regresión logística, los modelos de respuesta multinomial, e incluso ciertos análisis de la supervivencia y de series temporales, son, en última instancia, extensiones del Modelo Lineal clásico.

Por tanto, para abordar aquí el Modelo Lineal Generalizado nos dete-

nemos antes en el Modelo Lineal clásico, justificando así la ubicación del primero. Seguimos con una comparación entre ambos, Modelo Lineal y Modelo Lineal Generalizado, terminando con la presentación de un caso particular de MLG: un ejemplo de regresión de Poisson empleando el software *R*.

Ahora bien, para hablar del Modelo Lineal Generalizado debemos explicar antes el marco natural en el que se desarrolla, el *modelado estadístico*, un entorno que, a diferencia de las aplicaciones más tradicionales centradas en el contraste de hipótesis y en las pruebas de significación (ver López-González, 2003), se establece a partir de la década de los 60 atendiendo a la estimación de parámetros y a la comparación y ajuste de modelos de probabilidad a los datos empíricos (Ato y López-García, 1996, 80; y Ato y Vallejo, 2007, 40).

Modelado estadístico

El empleo de modelos es un proceso consustancial al ser humano. Para comprender lo que sucede a nuestro alrededor, a partir de la observación detallada de los acontecimientos solemos elaborar "modelos" mentales sobre cómo funcionan los fenómenos, pudiendo realizar, incluso, predicciones sobre ellos. En el ámbito científico, un modelo que explica un fenómeno suele expresarse de forma matemática (un modelo que, igualmente, ha sido derivado de descripciones y que probablemente será útil para predecir). Con esa vestimenta formal, la elaboración de modelos a la que estamos acostumbrados en la vida cotidiana puede transformarse aparentemente en algo extraño, pero no es más que una versión elegante del mismo procedimiento. Así lo señalan McCullagh y Nelder (1989) cuando afirman que la construcción de modelos requiere de una mezcla de arte y conocimientos por parte del investigador. Este proceso se conoce en ciencia como *modelado matemático* o *modelización matemática,* y cuando los fenómenos a explicar son probabilísticos, antes que determinísticos (como sucede frecuentemente en las ciencias del comportamiento, sociales y de la educación), hablamos de *modelado estadístico* o *estocástico* [1].

En un sentido amplio, un modelo pretende explicar la variación de una respuesta a partir de la relación conjunta de dos fuentes de variabilidad, una de carácter determinista y otra aleatoria, lo que responde a la expresión:

Respuesta = componente sistemático + componente aleatorio.

Judd y McClelland (1989, 1) toman la expresión anterior como: *DATOS = MODELO + ERROR,* asociando MODELO a la parte sistemática. Así, los *DATOS* corresponderían a las observaciones que se quieren analizar (la variable de respuesta o variable dependiente). *MODELO* es la función que se introduce con objeto de explicar los datos (una función ponderada de una o más variables explicativas o predictores). Y, dado que la variabilidad recogida en *DATOS* no termina de estar explicada, se introduce el término *ERROR,* que contiene la discrepancia o falta de ajuste entre *DATOS* y *MODELO* (entre la realidad empírica y la

explicación teórica o sustantiva). Es deseable que el *MODELO* sea, por tanto, una buena representación de los *DATOS,* de forma que el *ERROR* se reduzca lo máximo posible.

De la construcción, formulación y ajuste de modelos a los datos empíricos se encarga precisamente el modelado estadístico, debiendo responder a tres criterios: (a) criterio estadístico o *principio de bondad de ajuste*: la inclusión de parámetros en el MODELO en beneficio de una mejor representación de los DATOS con la correspondiente disminución del ERROR (ver McClelland, 1997); (b) criterio lógico o *principio de parsimonia*: la selección de los parámetros que formen parte del modelo de tal modo que éste se convierta en una representación simple y sobria de la realidad (Judd y McClelland, 1989, 3; y Ruiz-Soler, Pelegrina y López-González, 2000) y (c) criterio sustantivo o *integración teórica* del modelo en la red conceptual que lo generó (Ato y Vallejo, 2007, 47).

Esta construcción del modelo más parsimonioso que explique la variable de respuesta con el menor error posible se realiza atendiendo a unas etapas:

1. *Especificación del modelo* teórico, determinando qué variables son de interés, así como cuáles son las relaciones entre ellas. Esta situación da de lleno con el dilema entre los principios de parsimonia versus ajuste: que el modelo describa de la forma más simple posible, o bien que la concordancia entre el modelo y los datos sea lo más completa posible, es decir, con el mínimo error.

2. *Estimación de parámetros*, calculando el valor de los coeficientes del modelo examinado a partir del conjunto de datos observados, al objeto de determinar si el modelo teórico propuesto es aceptable como representación aproximada de los datos.

3. *Selección del modelo*, valorando si el nivel de discrepancia entre los datos observados y los datos ajustados es suficientemente bajo como para optar por el modelo o, por el contrario, suficientemente elevado como para rechazarlo.

4. *Evaluación del modelo*, examinando las observaciones individuales (*leverage points*), los datos influyentes (*influentials*) y los datos anómalos (*outliers*), así como comprobando los supuestos de normalidad, linealidad, homoscedasticidad e independencia.

5. *Interpretación del modelo*, comprendiendo sus implicaciones con respecto a la variable de respuesta. Esta fase conlleva una explicación detallada de los parámetros del modelo para comprobar si se cumplen los criterios estadístico, lógico y sustantivo.

Finalmente se acepta o no el modelo y, si es preciso, se reinicia el proceso.

Un software estadístico que reúne las características necesarias para trabajar el modelado estadístico es *R* (una aproximación en el ámbito educativo es el trabajo de Ruiz-Soler y López-González, 2009) [2]. La manera de trabajar del programa *R* con decisiones y pasos sucesivos se adapta fácilmente a la filosofía del

modelado: se van construyendo distintos modelos, calculando, al mismo tiempo, medidas de la *desvianza* o discrepancia entre los valores empíricos y los ajustados para valorar el modelo, aceptándolo tentativamente o rechazándolo, y permitiendo, finalmente, una mejor integración de la solución obtenida en la teoría sustantiva de partida. Por estas razones el ejemplo que aquí presentamos sobre una regresión de Poisson es ejecutado con *R*, especialmente diseñado también para el trabajo con cualquiera de las funciones matemáticas que comprenden los Modelos Lineales Generalizados.

Modelo Lineal

La fórmula general del Modelo Lineal es $Y = f(X)+g(\varepsilon)$, donde toda observación sobre la variable de respuesta es la suma de: (a) los efectos de un grupo de factores o *componentes sistemáticos* $-f(X)-$, que implican un conjunto de parámetros de una población y un conjunto de variables independientes relevantes medidas sobre cada uno de los sujetos con los que se trabaja, y (b) la función $g(\varepsilon)$, que representa el efecto de los *componentes aleatorios* y es resultado de una o más distribuciones de probabilidad dependientes de un pequeño número de parámetros. En esta fórmula general tienen cabida una amplia variedad de modelos lineales representativos de las relaciones estadísticas entre variables explicativas y de respuesta. Ahora bien, interesa resaltar aquellos que cumplen con una serie de restricciones respecto a las variables explicativas.

La primera es que para el caso de una variable de respuesta *Y* debe haber un conjunto de observaciones y_i sobre una o varias variables explicativas, y es necesario establecer ciertos supuestos respecto a la distribución de probabilidad de dichas variables aleatorias, los cuáles, además, varían según la escala de medida utilizada. Una segunda restricción sobre las variables explicativas es que cada una representa una muestra de valores observados seleccionados arbitrariamente por el investigador (componentes fijos) luego, al tratarse de valores prefijados, cualquier transformación de una variable explicativa puede ser considerada también como variable independiente. El modelo debe incluir, además, un conjunto de variables aleatorias no observables pero sí estimables: los parámetros del modelo (su estimación es una etapa fundamental en el ajuste del modelo, como ha quedado dicho). Por último, es preciso que el modelo incluya una o más variables que no son ni observables ni estimables, los componentes aleatorios, siendo el más importante el componente de error aleatorio que recoge la variabilidad debida a las diferencias individuales, a los errores de medida y, en general, a otras variables explicativas no incluidas en el modelo.

Al hablar de Modelo Lineal es conveniente señalar que la *linealidad* puede tener lugar de distintos modos y que, según ellos se obtienen modelos de uno u otro tipo. Cabe considerar como Modelo Lineal, no obstante, todo aquel que lo sea en sus parámetros, con independencia de que sus variables explicativas cumplan esta condición o no. Se habla, entonces, de un *Modelo Lineal de primer orden* para *k* variables explicativas y *k+1* parámetros

si el modelo es lineal en sus parámetros y en sus variables explicativas, respondiendo a la siguiente fórmula general:

$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon .$$

Si el modelo es lineal en sus parámetros pero no en las variables explicativas sería un *Modelo Lineal de m-ésimo orden* (cuadrático, cúbico, etc.) con *km* variables independientes y *km+1* parámetros. Puede incluir componentes de interacción y ser susceptible de ser linealizado transformando sus variables explicativas. Su formulación es:

$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \sum_{j=1}^{k} \beta_{j1} X_{j1}^2 + \cdots + \sum_{j=1}^{k} \beta_{j1} X_{j1}^m + \varepsilon .$$

Si el modelo no es lineal en los parámetros y/o en las variables explicativas puede adoptar múltiples formulaciones. No obstante, al poderse linealizar mediante las transformaciones adecuadas suele ser tratado como *modelo intrínsecamente lineal* (Draper y Smith, 1998, 459).

Los modelos que no se ajustan a ninguna de las anteriores situaciones son, entonces, *modelos no lineales*.

La literatura respecto al Modelo Lineal clásico es amplia. Una presentación sencilla se encuentra en Fox (1997); Kleinbaum, Kupper y Miller (1988); Neter, Kutner, Nachtsheim y Wasserman (1996); Pedhazur (1997); y Weisberg (2005). Con un nivel más avanzado están los trabajos de McCullagh y Nelder (1989); Horton (1985); Seber y Lee (2003); y Tatsuoka (1993). Desde la perspectiva del modelado estadístico pueden consultarse los textos de Judd y McClelland (1989) y Losilla y cols. (2005). El Modelo Lineal explicado con *R* se encuentra en Crawley (2007), Faraway (2004), y en la imprescindible obra de Fox (2002).

Por otro lado, el Modelo Lineal es susceptible de numerosas variaciones para ajustarse a las particularidades de una investigación específica, variaciones que a nivel matemático se ven reflejadas en las relaciones existentes entre la variable de respuesta, las variables explicativas, los parámetros del modelo y el componente de error aleatorio. Sin embargo, por lo general, tres son las principales formas que el Modelo Lineal general puede adoptar dependiendo de la estructura métrica de las variables explicativas: la forma de un modelo de regresión, la forma de un modelo de análisis de varianza (ANOVA) o de diseño experimental, y la forma de un modelo de análisis de covarianza (ANCOVA) o de diseño experimental con variables concomitantes (Ato y Vallejo, 2007; Bock, 1985; Dobson y Barnett, 2008; Hocking, 1985; Tatsuoka, 1993; y Timm, 2002).

En los *modelos de regresión* las variables explicativas son de naturaleza métrica, cuantitativa continua o discreta, cumpliendo con los supuestos básicos del Modelo Lineal, a saber: linealidad, homoscedasticidad, normalidad e independencia de los errores (pueden consultarse, por ejemplo, en López-González, 1994). Los modelos de regresión simple, múltiple, multivariante o la correlación canónica se incluyen en este tipo de modelos.

En los *modelos de análisis de varianza* (ANOVA) o de diseños experimentales las variables explicativas presentan una estructura no métrica de carácter categórico (dicotómico o politómico), utilizando variables *dummy* con la intención de representar la pertenencia a los grupos que configuran las categorías. El interés fundamental de estos modelos es la búsqueda de inferencias válidas acerca de las medias poblacionales a partir de las medias muestrales obtenidas en cada una de las condiciones de tratamiento experimental. Aquí se sitúan los modelos de ANOVA, que según se considere el efecto de tratamiento pueden ser fijos, aleatorios o mixtos, y los modelos de ANOVA factorial, que dependiendo de las relaciones entre los factores pueden ser de clasificación cruzada o anidada.

Por último, en los *modelos de análisis de covarianza* (ANCOVA) o modelos de diseño experimental con variables concomitantes unas variables tienen una estructura métrica cuantitativa continua o discreta (covariables) y otras variables explicativas tienen una estructura no métrica. El objetivo de estos modelos es el mismo que el de los modelos de análisis de varianza, a saber, realizar inferencias sobre las medias de los distintos grupos o condiciones de tratamiento, pero aquí se considera también la posibilidad de reducir la varianza de error. Entre estos modelos se encuentra el modelo de ANCOVA factorial, el modelo de ANCOVA multivariante, los diseños de bloques aleatorios y, en general, el conjunto de diseños con variables concomitantes (cuadrado latino, grecolatino, etc.).

En todos los modelos señalados se ha partido del supuesto de que la variable de respuesta posee una estructura métrica cuantitativa continua (habitualmente normal) o discreta. Sin embargo, es posible flexibilizar este criterio y considerar una formulación más general que permita también contemplar variables dependientes con una estructura no métrica, es decir, variables categóricas (ordinales o nominales), al tiempo que relajar los supuestos del Modelos Lineal clásico, como la linealidad o la homoscedasticidad, no así la independencia de los errores. Los modelos resultantes guardan una estrecha similitud con los modelos citados y dan paso a los *modelos con variables categóricas* [3] y a los *Modelos Lineales Generalizados*. En este gran grupo podemos situar los modelos de regresión logística, modelos logit, probit y modelos loglineales, entre otros.

A modo de síntesis clasificatoria, los principales modelos apuntados hasta ahora se recogen en la Tabla 1 siguiendo las clasificaciones de Ato y cols. (2005) y Losilla y cols. (2005). También puede consultarse la clasificación de Dobson y Barnett (2008, 3).

Modelo Lineal Generalizado

Los primeros trabajos donde se introduce y desarrolla el Modelo Lineal Generalizado son, respectivamente, Nelder y Wedderburgn (1972) y McCullagh y Nelder (1989). Una buena introducción se encuentra en las terceras ediciones de las obras de Draper y Smith (1998) y Weisberg (2005), así como en el imprescindible texto de Dobson y Barnett (2008). Desde el entorno de modelado estadístico, el

TABLA 1: *Principales Modelos Lineales Generalizados*

Naturaleza de la Variable de Respuesta	Componente		Función de enlace	Modelo Lineal	
	Sistemático	Aleatorio			
▪ *Numérica cuantitativa*	· Numérico	· Normal	· Identidad	- Regresión lineal	ML
	· Categórico	· Normal	· Identidad	- ANOVA o de diseño experimental	
	· Mixto	· Normal	· Identidad	- ANCOVA o de diseño experimental con variables concomitantes	
▪ *Categórica binaria*					MLG
– No agrupada	· Mixto	· Binomial (1) - Bernoulli-	· Logit	- Regresión logística	
– Agrupada (frecuencias)	· Categórico	· Binomial (n)	· Logit · Probit	- Análisis logit - Regresión probit	
▪ *Categórica politómica*					
– No agrupada	· Mixto	· Multinomial	· Logit generalizado	- Regresión logística multinomial	
– Agrupada (frecuencias)	· Categórico	· Multinomial	· Logit generalizado	- Análisis logit multinomial	
▪ *Recuento*	· Mixto	· Poisson	· Logarítmica	- Regresión de Poisson	
▪ *Frecuencia*	· Categórico	· Poisson	· Logarítmica	- Análisis loglineal	

MLG se estudia en Hutcheson y Sofroniou (1999). En castellano la literatura es todavía escasa; destacamos Ato y López-García (1996) y Ato y cols. (2005). El empleo de los Modelos Lineales Generalizados con *R* se trabaja en Crawley (2007), Dobson y Barnett (2008), Faraway (2006), Fox (2002) y en Wood (2006), donde incluso se amplía su estudio a los Modelos Generalizados Aditivos.

Como ha quedado planteado, tanto el Modelo Lineal Generalizado (MLG) como el modelado estadístico son herramientas metodológicas que permiten codificar todas las situaciones de análisis dentro de un mismo esquema general. Obviamente, esto facilita el aprendizaje de nuevos modelos de análisis porque se trata simplemente de contemplarlos como casos particulares de un modelo más general ya conocido, el Modelo Lineal (ML). Veamos las relaciones entre ambos.

En la Tabla 1 puede observarse cómo el Modelo Lineal es el caso más elemental del Modelo Lineal Generalizado. Las coincidencias y las diferencias entre uno y otro hacen posible, en el caso del MLG, un tratamiento matemático y estadístico adecuado a los niveles de medida de las variables que contiene.

Siguiendo a Ato y cols. (2005), el MLG tiene componentes empíricos (las variables que se registran) y componentes teóricos que son:

El vector de la respuesta media: $g(\mu_i) = \eta_i$

El vector del predictor lineal: β_j x_{ij} (componente sistemático) + ε_i (componente aleatorio).

Precisamente los términos *componente sistemático* y *componente aleatorio* responden al enfoque del *modelado estadístico* al que nos hemos referido anteriormente.

TABLA 2: *Comparación entre ML y MLG (López-González y cols., 2002a)*

Modelo Lineal (ML)	Modelo Lineal Generalizado (MLG)
$\mathbf{y}_i = \sum_j \boldsymbol{\beta}_j \mathbf{X}_{ij} + \varepsilon_i$	$\mathbf{y}_i = \sum_j \boldsymbol{\beta}_j \mathbf{X}_{ij} + \varepsilon_i$
$\mu_i = E(Y_i)$	$\mu_i = E(Y_i)$
$\boldsymbol{\eta}_i = \sum_j \boldsymbol{\beta}_j \mathbf{X}_{ij}$	$\boldsymbol{\eta}_i = \sum_j \boldsymbol{\beta}_j \mathbf{X}_{ij}$
$\eta_i = \mu_i$	$\eta_i = g(\mu_i)$
$\mathbf{y}_i$: vector de la variable respuesta,	
$\mathbf{X}_{ij}$: matriz de variables predictoras y covariables	
$\boldsymbol{\beta}_j$: vector de parámetros	
$\boldsymbol{\eta}_i$: vector del predictor lineal	

Como puede verse en la Tabla 2 la expresión del modelo en ambos casos (ML y MLG) es la misma, siendo los valores ajustados $\mu_i = E(Y_i)$. El predictor lineal también coincide: $\eta_i = \sum_j \beta_j x_j$. Aparecen diferencias, no obstante, en la relación entre los valores ajustados μ_i y el predictor lineal η_i. Estas diferencias se concretan en la función de enlace y en la distribución que ésta debe seguir, función que cobra un especial significado que pasamos a explicar.

(a) Mientras que en el ML se produce una *relación de identidad* entre los valores ajustados y el predictor lineal, $\mu_i = \eta_i$, en el MLG la linealidad se establece en la escala del predictor lineal pero no en la escala de los valores ajustados. No se da, por tanto, la identidad entre valores ajustados y valores predichos, sino que entre ellos media una función que los relaciona, la *función de enlace*: $g(\mu_i) = \eta_i$. Esto hace que en el MLG ambos lados de la ecuación no se expresen en la misma escala de medida, lo que sí sucede en el ML.

(b) El componente aleatorio del Modelo Lineal debe distribuirse normalmente, y este hecho tiene una importancia considerable: según sea la distribución de los errores serán las distribuciones condicionadas de los valores pronosticados del criterio, que, por tanto, deben ser normales también. Esto es así porque ambas distribuciones están relacionadas a través de una

matriz, la matriz *hat* (ver López-González, 1994). En el MLG sucede que el componente aleatorio no sigue necesariamente una distribución normal sino que utiliza cualquier distribución de la familia exponencial y, en consecuencia, las distribuciones de los valores pronosticados del criterio no serán normales necesariamente (en Wood, 2006, 61 y en Faraway, 2006, 120 pueden consultarse las distintas distribuciones de la familia exponencial).

(c) Las distribuciones condicionadas de los valores pronosticados de la variable de respuesta en el ML deben ser homoscedásticas, y ello es posible siempre que esta condición se cumpla en el componente aleatorio. Como en el MLG los errores pueden seguir cualquier distribución de la familia exponencial, resulta que para la distribución de los errores la homoscedasticidad no es imprescindible.

(d) Las diferencias expresadas hasta ahora obligan a estimar los parámetros de un MLG con un método de ajuste distinto al procedimiento de mínimos cuadrados que se emplea en el ML: el método de máxima verosimilitud (que también puede ser aplicado en el ML).

Destacamos en el Modelo Lineal Generalizado, por tanto, el protagonismo de ese tercer elemento que relaciona los componentes aleatorio y sistemático, es decir, el valor esperado y el valor predicho por el modelo: nos referimos a la función de enlace $g(\mu)$ (un estudio detallado de la naturaleza de la función de enlace se encuentra en Krzanowski, 1998, 168). Así, por ejemplo, para el caso que presentamos en el próximo apartado con una variable de respuesta de recuento, el valor esperado μ sólo puede tomar valores enteros iguales o superiores a cero, mientras que el predictor lineal η_i puede adoptar cualquier valor entre $-\infty$ y $+\infty$. Esto hace que el valor esperado y el predictor lineal tengan diferentes escalas de medida, precisamente por mediar entre ellos dicha función de enlace que termina transformando el valor de recuento esperado a la escala del predictor lineal: $g(\mu_i) = \eta_i$. La *inversa de la función de enlace* (o función de transformación) realiza el proceso contrario, y al ser aplicada al resultado del predictor lineal η_i (que se halla en una escala de $-\infty$ y $+\infty$) se obtiene el valor esperado, μ que se encuentra en la escala de la variable de respuesta (Ato y cols., 2005, 8):

$$\mu_i = g^{-1}(\beta_0 + \beta_0 X_i).$$

Regresión de Poisson con R

La regresión de Poisson es el modelo más básico adecuado para variables de respuesta de recuento (Long, 1997, 217), aunque existen también otras opciones dentro del Modelo Lineal Generalizado que pueden adaptarse bien, como el modelo de regresión binomial negativa (ver Hilbe, 2007), el modelo de regresión truncada y los modelos de conteo modificados a cero.

Inicialmente la distribución de Poisson comenzó a aplicarse en el estudio de conductas criminales, pero ya desde finales del siglo pasado pasó a ser un modelo frecuente en ámbitos como la bioestadística, la econometría y el marketing (Cameron y Tribedi, 1998, 94), así como en diversas áreas aplicadas: criminología,

sociología, ciencia política o relaciones internacionales (o.c., 17). Un estudio pormenorizado de la regresión de Poisson está en Agresti (2002); en el entorno de *R* hay interesantes ejemplos en Dobson y Barnett (2008, 165 y ss.) y Faraway (2006, 55 y ss.).

El caso que aquí mostramos es una aplicación sencilla. A partir de un conjunto de datos tomado de Venables y Ripley (2003, 446) buscamos llamar la atención sobre el análisis de estos datos con el modelado estadístico y con la función probabilística más adecuada. Estas dos inquietudes se resuelven fácilmente empleando el programa *R*. Incorporamos también una solución con el Modelo Lineal (regresión clásica con la función de enlace identidad) al objeto de observar brevemente las ventajas del empleo del Modelo Lineal Generalizado de regresión de Poisson frente al primero.

Método

Suponiendo que se realiza un sondeo en diversos Institutos de Enseñanza Secundaria, se observa el número de *Conflictos* o faltas de disciplina que han quedado registrados en los centros durante un periodo de trece años, desde 1990 a 2003. El modelo que se especifique debe explicar el aumento de conflictos con el paso del tiempo. Se modelan los datos mediante el MLG de regresión de Poisson utilizando la distribución de Poisson como la más adecuada para una variable de respuesta de recuento (no métrica, categórica) y empleando la función de enlace logarítmica propia de este modelo.

Resultados y discusión

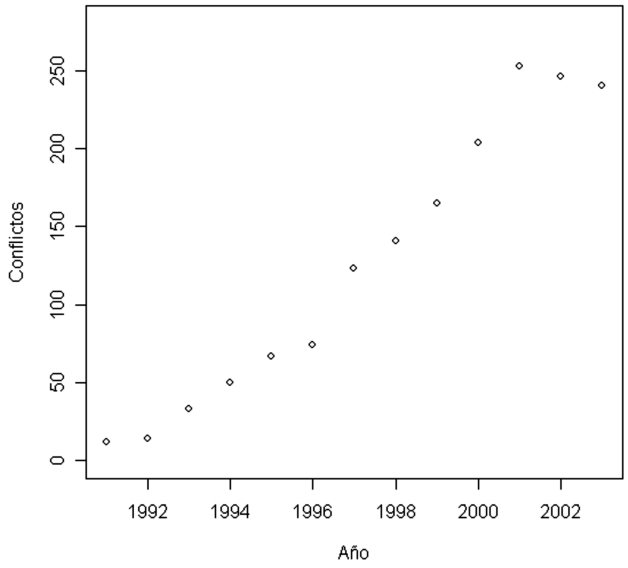
En la Tabla 3 se muestran las instrucciones para que *R* cree los vectores necesarios: *Conflictos*, que comprende los valores de recuento, y *Año*, que contabiliza un vector de valores de 1 a 13. A continuación se pide el diagrama de dispersión de las variables *Año* y *Conflictos*, etiquetando cada uno de los ejes. El resultado se muestra en el Gráfico 1 donde puede apreciarse que no existe una relación lineal entre esta dos variables: el avance de los años parece relacionarse con un aumento suavizado de los conflictos tal que: $\mu_i = \gamma \exp(\delta X_i)$, donde γ y δ son parámetros desconocidos, y X_i es el valor de cada Año. El enlace *log* para esta función se encuentra dentro del MLG, así que podemos especificar el siguiente modelo: $\log(\mu_i) = \log(\gamma) + \delta X_i = \beta_0 + \beta_1 X_i$.

TABLA 3: *Creación de datos y ejecución del diagrama de dispersión en R*

R entrada:

```
> Conflictos <- c(12,14,33,50,67,74,123,141,165,204,253,246,240)
> Año <- 1:3
> plot (Año+1990,Conflictos, xlab="Año", ylab="Conflictos", ylim=c(0,280))
```

GRÁFICO 1: *Diagrama de dispersión Conflictos x Año*



Para dar respuesta al ajuste de estos datos pasamos a estimar los parámetros de dos modelos de regresión de Poisson. El primero, que denominamos modelo nulo (modelo 0 -*m0*-), responde a una regresión de Poisson simple. Después calcularemos un nuevo modelo más completo: el modelo 1 *(m1).*

En la Tabla 4 aparece la sencilla orden a ejecutar en *R*, indicando que es un Modelo Lineal Generalizado (*glm*), la variable de respuesta (*Conflictos*), el predictor (*Año*) y el tipo de modelo: *poisson.*

TABLA 4: *Instrucciones y resultados del modelo m0*

R entrada:
> m0 <- glm (Conflictos~Año, poisson)
> m0
R salida:
Call: glm(formula = Conflictos ~ Año, family = poisson)
Coefficients:
(Intercept) Año
3.1406 0.2021
Degrees of Freedom: 12 Total (i.e. Null); 11 Residual
Null Deviance: 872.2
Residual 80.69 AIC: 166.4
Deviance:

A partir de *m0* hemos asumido que el número de *Conflictos* sigue una distribución de Poisson tal que: y_i = Poi (μ_i), donde y_i es el número observado de nuevos *Conflictos* en el *Año* correspondiente. Una vez obtenidos los coeficientes se pueden construir los modelos propios de la regresión de Poisson:

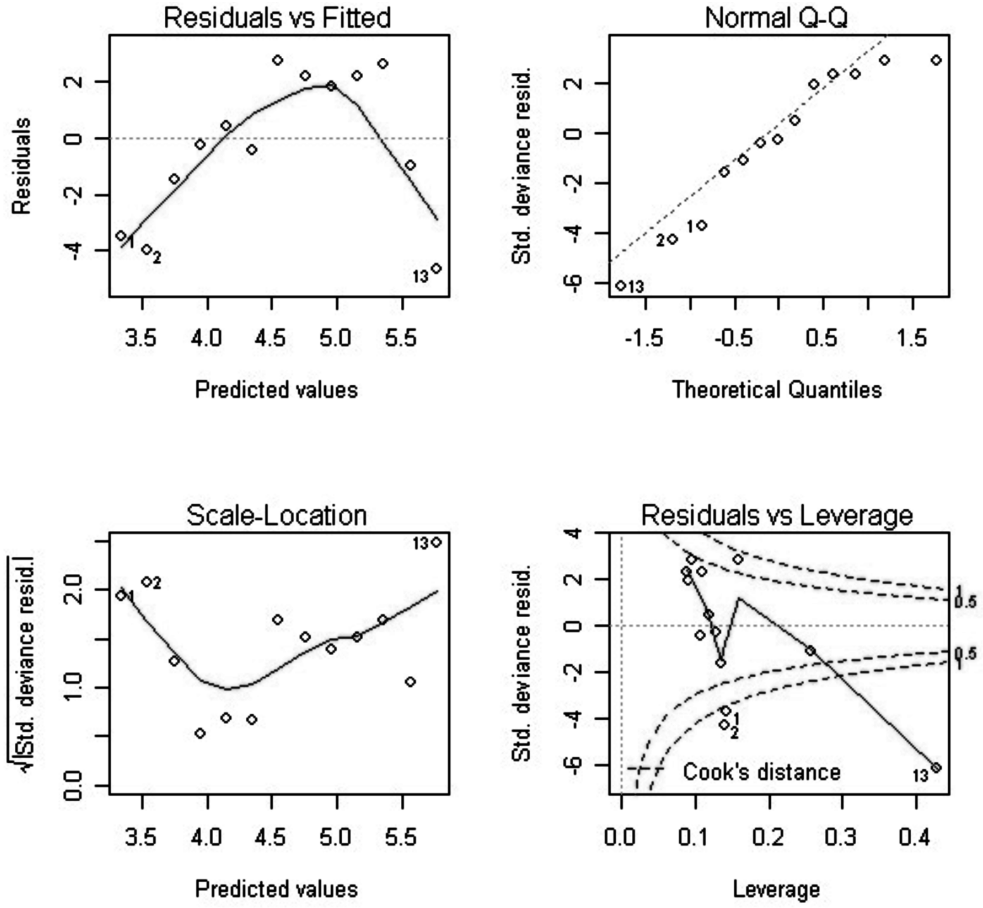
Primero, el *modelo aditivo* $log(\mu_i) = \beta_0 + \beta_1 X_i$, que en términos muestrales será: $log(m_i)$ = 3.1406+0.2021 x *Año*. Es importante tener en cuenta que el modelo aditivo representa la función de enlace y que, tal y como comentamos arriba, las unidades calculadas no se corresponden con la medida de la variable original. Para volver a los valores originales de la variable de respuesta se construye el *modelo multiplicativo*, que representa la inversa de la función de enlace o función de transformación. Es decir, se calcula la exponencial correspondiente, en este caso en base *e*. Así tenemos $m_i = e^{(b_0+b_1X_i)}$, donde sustituimos por los coeficientes estimados, resultando: $m_i = e^{(3,1406)}$ x $e^{(0,2021Xi)}$. Puede expresarse también la función exponencial inversa a la logarítmica, tal que:

$m_i = exp\ (b_0 + b_1X_1) = exp\ (3,1406)\ x\ exp\ (0,2021\ X_i)$.

Para proceder a la selección del modelo retomamos la información de la Tabla 4 donde se observa que la *discrepancia* (*residual deviance*) es extremadamente alta, entendiendo este término como una medida de bondad de ajuste relacionada con las diferencias entre los valores observados y los esperados o generados por la regresión (Faraway, 2006, 29). La discrepancia dividida por los grados de libertad suele usarse para detectar sobre o baja dispersión. En la regresión de Poisson la media y la varianza son iguales, lo que implica que la discrepancia dividida por los grados de libertad debe aproximarse a uno. Valores mayores que uno indican sobredispersión (la verdadera varianza es mayor que la media); valores menores que uno indican baja dispersión (la verdadera varianza es menor que la media) (López-González y cols., 2002). Tanto un caso como otro informan de un ajuste inadecuado, como sucede aquí.

Siguiendo con la evaluación del modelo, resultan preocupantes los gráficos de residuos de diagnóstico del modelo 0 (Gráfico 2). El *gráfico de residuos frente a valores pronosticados* muestra cómo los residuos no presentan una tendencia a la media, lo que refleja el incumplimiento de la condición de independencia de los errores probablemente por la omisión de algún término importante en el modelo. Recordando el diagrama de dispersión inicial (Gráfico 1), los *Conflictos* aumentaban monótonamente con los *Años*; esto permite pensar que quizá la introducción de un término cuadrático en el modelo podría mejorar considerablemente el ajuste (Wood, 2006, 89). Así mismo, en el *gráfico de residuos frente a influencia* se aprecian puntos de influencia muy elevados en los últimos años, especialmente para el caso 13 que corresponde al año 2003.

GRÁFICO 2: *Diagnóstico del modelo m0*



Construimos, por tanto, un nuevo modelo añadiendo un término cuadrático, el modelo 1 -*m1*-, de forma que:

$$log(\mu_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$$

$$\mu_i = exp\ (\beta_0 + \beta_1 X_i + \beta_2 X_i^2).$$

Siguiendo los resultados de la Tabla 5 sustituimos en las expresiones anteriores los coeficientes, quedando ahora los modelos aditivo y multiplicativo del modelo 1 en términos muestrales del siguiente modo:

$$log(m_i) = 1{,}9014 + 0{,}5566\ X_i - 0{,}02135\ X_i^2$$

$$(m_i) = exp\ (1{,}9014 + 0{,}5566\ X_i - 0{,}02135\ X_i^2) = exp\ (1{,}9014)\ x\ exp\ (0{,}5566\ X_i - 0{,}02135 X_i^2)$$

Al evaluar el nuevo modelo *m1* se aprecian cambios importantes en el Gráfico 3: los residuos manifiestan ahora una tendencia a la media en el *gráfico de residuos frente a pronósticos*, luego la condición de independencia de los errores parece cumplirse. Igualmente, la dispersión vertical de los residuos es razonablemente pequeña y la influencia del caso 13 se ha reducido. Los resultados de la Tabla 5 reflejan que el modelo *m1* se ajusta mejor a los datos: la discrepancia residual

TABLA 5: *Instrucciones y resultados para el modelo m1*

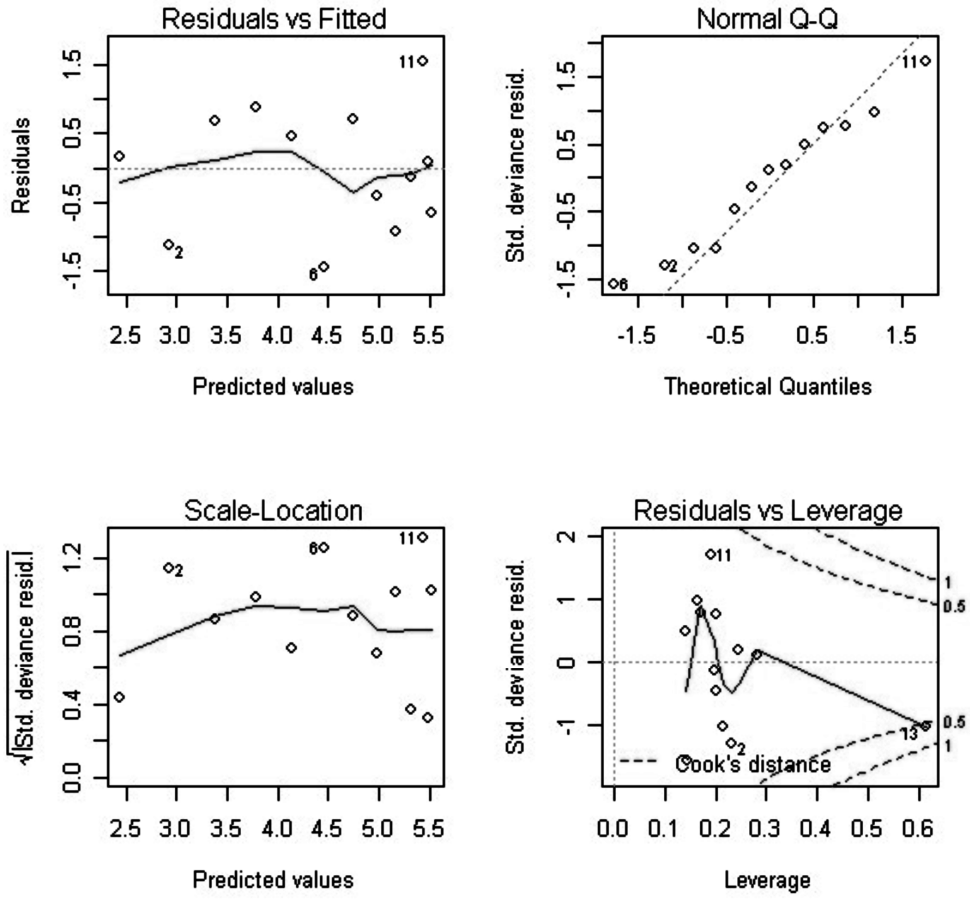
R entrada:

```
> glm (Conflictos~Año + I(Año^2), family = poisson)
> m1 <- glm (Conflictos~Año + I(Año^2), family = poisson)
```

R salida:

```
Call:  glm(formula = Conflictos ~ Año + I(Año^2), family = poisson)
Coefficients:
(Intercept)      Año        I(Año^2)
   1.90146      0.55600   -0.02135
Degrees of Freedom: 12 Total (i.e. Null);  10 Residual
Null Deviance:        872.2
Residual              9.24      AIC:   96.92
Deviance:
```

GRÁFICO 3: *Diagnóstico del modelo m1*



ha disminuido considerablemente, pasando de 80.69 a 9.24.

En la representación de los valores ajustados por ambos modelos (Gráficos 4 y 5) es fácil distinguir que el modelo *m1*

GRÁFICO 4: *Ajuste del número de Conflictos con el modelo m0*

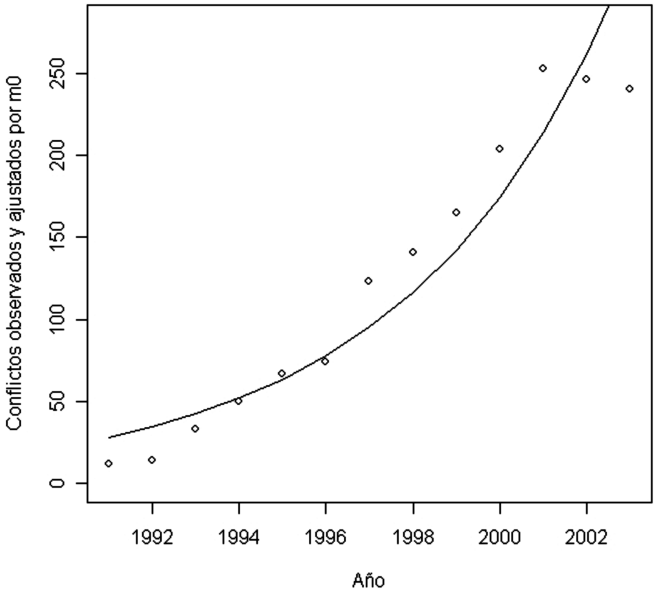
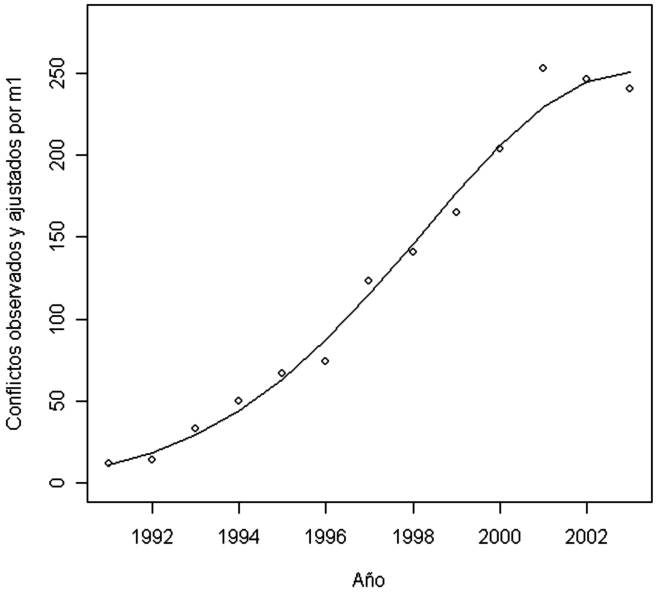


GRÁFICO 5: *Ajuste del número de Conflictos con el modelo m1*



se ajusta mejor al número de *Conflictos* observados que el modelo *m0*.

Esta conclusión puede reforzarse empleando el *test de la razón de verosimilitud generalizado* (Faraway, 2006, 120) que compara las *discrepancias* que se producen en ambos modelos, tal que: $\Delta D = D_{m0} - D_{m1}$ (Tabla 6). ΔD sigue una distribución asintótica Chi cuadrado y evalúa si la inclusión de los términos del modelo *m1* reduce significativamente el valor de discrepancia con respecto al modelo *m0*. Como se obtiene un valor de *p* mínimo (Tabla 6) se puede afirmar que, efectivamente, *m1* tiene una influencia significativa en la disminución de la discrepancia siendo, por tanto, más adecuado.

TABLA 6: *Resultados del test de razón de verosimilitudes entre m0 y m1*

R entrada:

```
> anova (m0, m1, test="Chisq")
```

R salida:

```
Analysis of Deviance Table
Model 1: Conflictos ~ Año
Model 2: Conflictos ~ Año + I(Año^2)
```

Resid.	Df	Resid.	Df	Deviance	P(>|Chi|)
1	11	Dev	1	71.446	2.849e-17
2	10	80.686			
		9.240			

Para finalizar el modelado cabría interpretar el modelo en relación a los valores obtenidos. Por tal motivo, además de transformar los valores de los parámetros para obtener su valor en la escala original de la variable de respuesta, habría que estimar sus intervalos de confianza y ser interpretados en términos de efectos simples (particularmente en caso de existencia de interacción), finalizando con el cálculo del intervalo de predicción de los valores de la variable de respuesta.

Siguiendo con nuestro objetivo de animar al empleo del MLG en las situaciones en que los datos lo requieran, mostramos ahora la salida que se obtiene en este ejemplo con un Modelo Lineal clásico: una regresión lineal simple. A estas alturas ya sabemos que no es la solución adecuada. Tal y como señala Long (1997, 3), las características de una variable de respuesta de recuento son tales que si para su ajuste se emplea un Modelo Lineal con función de enlace identidad, las estimaciones resultantes son ineficientes, inconsistentes y sesgadas, aunque, y esto es lo delicado, pueden ser de magnitud y significación similares a las obtenidas por la regresión de Poisson. No podemos eludir, por tanto, el extendido e incorrecto uso del ML clásico en el ámbito de las ciencias humanas en

numerosas ocasiones en las que se relacionan variables cuantitativas con una función de dependencia, y en las que no se reflexiona suficientemente sobre la métrica de la variable de respuesta que interviene.

TABLA 7: *Instrucciones y resultados de la regresión lineal*

R entrada:

```
> m3<- lm(Conflictos~Año)
> m3
> summary (m3)
```

R salida:

```
Call:
lm(formula = Conflictos ~ Año)
Residuals:
    Min      1Q  Median      3Q     Max
-28.060  -6.643  -1.769   7.687  37.396
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -34.192     10.701  -3.195  0.00853 **
Año           22.709      1.348  16.844 3.34e-09 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 18.19 on 11 degrees of freedom
Multiple R-Squared: 0.9627,     Adjusted R-squared: 0.9593
F-statistic: 283.7 on 1 and 11 DF,  p-value: 3.344e-09
```

Así, apoyando el comentario de Long, en la Tabla 7 puede verse que los coeficientes del modelo de regresión lineal son significativos. Sin embargo, al observar los valores pronosticados obtenidos en la Tabla 8, para el caso 1 (año 1991) resulta un número de conflictos pronosticado negativo, lo que no debiera suceder con una variable de recuento. En el Gráfico 6 se aprecia también cómo el ajuste de la recta a los datos no es adecuado, a diferencia de lo que sucedía con el modelo *m1* de la regresión de Poisson (Gráfico 5). La solución con el ML es muy forzada. La recta aumenta el valor de los Conflictos pronosticados en ocho de los trece Años considerados, de ahí la frecuencia de residuos negativos

(Tabla 8). El *gráfico de residuos frente a valores ajustados* refleja de nuevo una falta de ajuste de los residuos a la media, con lo que no se cumple la independencia de los errores, como también ocurría con el modelo *m0* (Grafico 7).

TABLA 8: *Valores pronosticados y residuos de la regresión lineal*

Diagnósticos por caso[a]

Número de caso	Residuo tip.	Conflictos	Valor pronosticado	Residuo bruto
1	1,291	12	-11,48	23,484
2	,153	14	11,23	2,775
3	-,051	33	33,93	-,934
4	-,365	50	56,64	-6,643
5	-,679	67	79,35	-12,352
6	-1,543	74	102,06	-28,060
7	-,097	123	124,77	-1,769
8	-,356	141	147,48	-6,478
9	-,285	165	170,19	-5,187
10	,611	204	192,90	11,104
11	2,056	253	215,60	37,396
12	,423	246	238,31	7,687
13	-1,156	240	261,02	-21,022

a. Variable dependiente: Conflictos

GRÁFICO 6: *Ajuste con el Modelo Lineal*

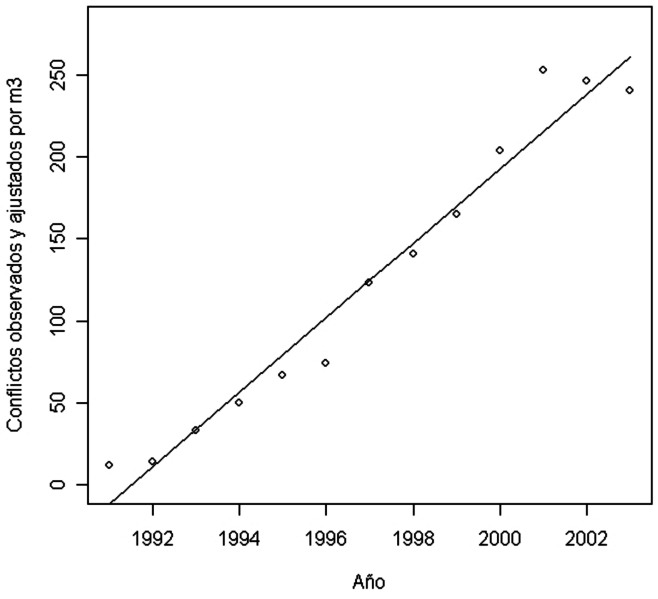
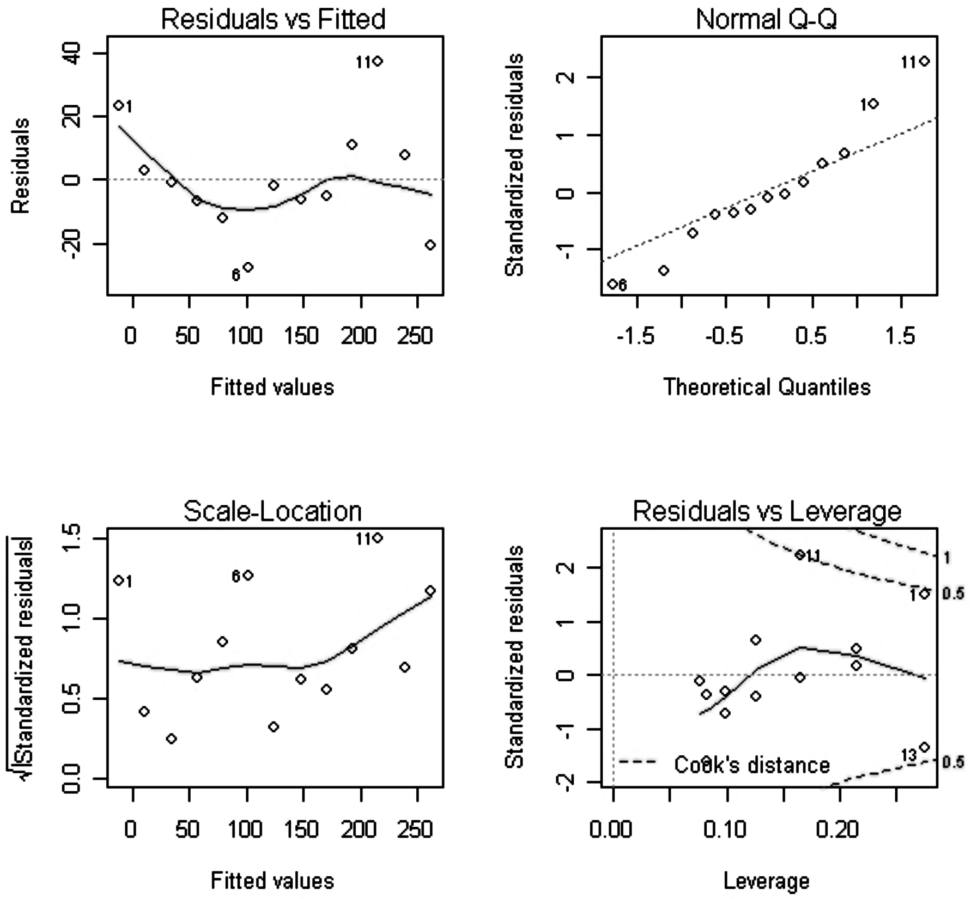


GRÁFICO 7: *Diagnóstico de la regresión lineal*



En suma, creemos que las razones señaladas son importantes para preferir el MLG de regresión de Poisson frente al Modelo Lineal clásico respetando la métrica de los datos del ejemplo que se ha descrito. Esta preferencia del MLG puede hacerse extensible a cualquier situación en la que la variable de respuesta no sea métrica. Los argumentos aquí expuestos pueden ampliarse con la lectura de Ato y cols. (2005), donde se describen con detalle ejemplos modelando diversas soluciones y empleando también transformaciones de los predictores. Son interesantes, igualmente, las extensiones de la regresión de Poisson y de las diversas aplicaciones del Modelo Lineal Generalizado a las soluciones multinivel, así como a los modelos jerárquicos, modelos aditivos y modelos mixtos.

Conclusión

El Modelo Lineal y el Modelo Lineal Generalizado son referentes imprescindibles actualmente en el análisis de datos de investigaciones que pretenden la explicación de fenómenos probabilísticos. Las peculiaridades matemáticas del MLG que aquí se han descrito le confieren una muy interesante adaptabilidad a las caracte-

rísticas métricas de las variables con las que se trabaja, lo cual viene a solventar el tratamiento estadístico inadecuado en el análisis de datos de investigaciones educativas, donde sucede con frecuencia que las variables que se registran no cumplen los presupuestos matemáticos de los modelos estadísticos más tradicionales, como el Modelo Lineal o las pruebas de significación estadística. Por otro lado, si un modelo debe ser, en cualquier caso, una buena "representación" de la realidad, el modelado estadístico ofrece el marco adecuado para que los criterios de ajuste, parsimonia e integración teórica exigibles al modelo puedan irse conformando.

Ahora bien, las ventajas señaladas no pasarían de ser soluciones teóricas si no existiera una herramienta que permitiera desarrollar plenamente estas propiedades. El software *R* reúne las características necesarias. El modo de trabajar de *R* se adapta fácilmente a la filosofía del modelado estadístico, así como a las propiedades de los modelos de dependencia estadísticos adaptados a variables no métricas: los Modelos Lineales Generalizados. Además, desde el ámbito de las Ciencias Sociales es especialmente interesante contribuir al cambio de filosofía que implica el modelado estadístico, así como propiciar el empleo de análisis gráficos por las numerosas ventajas que aportan. *R* cumple con estos requisitos en el análisis y explotación de datos, tanto si se trata de técnicas más clásicas, como si se emplean métodos más novedosos.

Dirección para la correspondencia: Emelina López González. Departamento de Métodos de Investigación e Innovación Educativa. Facultad de Ciencias de la Educación. Campus de Teatinos. 29071 Málaga. E-mail: emelopez@uma.es

Fecha de recepción de la versión definitiva de este artículo: 10.X.2010

Notas

[1] En el ámbito de las ciencias del comportamiento el modelado estadístico es denominado también como *enfoque de la comparación de modelos*. Textos imprescindibles que trabajan esta línea son: Judd y McClelland (1989); Krzanowski (1998); Lunneborg (1994) y Maxwell y Delaney (2004).

[2] Puede decirse que los grandes atractivos de *R* son su modo de trabajar, la participación constante de una comunidad de investigadores en su desarrollo y su libre acceso en internet.

[3] No deben confundirse los modelos de datos categóricos que emplean el procedimiento de estimación por máxima verosimilitud, como es el caso, con las técnicas de análisis de datos categóricos (también llamadas modelos de datos categóricos) correspondientes a los análisis exploratorios multivariantes de interdependencia cuyo procedimiento de estimación es el de mínimos cuadrados alternantes. Estos últimos comprenden, fundamentalmente, las contribuciones procedentes del *sistema GIFI*, tales como el análisis de homogeneidad, el análisis de correspondencias y el análisis de componentes principales no lineal (ver van der Geer, 1993).

Bibliografía

AGRESTI, A. (2002) *Categorical Data Analysis* (2ª ed.) (New York, Wiley).

ATO, M. y LÓPEZ-GARCÍA, J. J. (1996) *Análisis estadístico para datos categóricos* (Madrid, Síntesis).

ATO, M.; LOSILLA, J. L.; NAVARRO, J.; PALMER, A. y RODRIGO, M. (2005) *Modelo lineal generalizado* (Girona, EAP).

ATO, M. y VALLEJO, G. (2007) *Diseños experimentales en Psicología* (Madrid, Pirámide).

BOCK, R. D. (1985) *Multivariate Statistical Methods in Behavioural Research* (Reprinted. New York, Scientific Software).

CAMERON, A. y TRIVEDI, P. (1998) *Regression Analysis of Count Data* (Cambridge, Cambridge University Press).

CRAWLEY, M. J. (2007) *The R Book* (Chichester, Wiley & Sons, Ltd).

DRAPER, N. y SMITH, H. (1998) *Applied Regression Analysis* (3 ed.) (New York, Wiley).

DOBSON, A. J. y BARNETT, A. (2008) *An Introduction to Generalized Linear Models* (3ª ed.) (Boca Raton, FL., Chapman and Hall/CRC).

FARAWAY, J. J. (2004) *Linear Models with R* (Boca Raton, FL., Chapman & Hall/CRC).

FARAWAY, J. J. (2006) *Extending the Linear Model with R. Generalized Linear, Mixed Effects and Nonparametric Regression Models* (Boca Raton, FL., Chapman & Hall/CRC).

FOX, J. (1997) *Applied Regression Analysis, Linear Models and Related Methods* (Thousand Oaks, Sage).

FOX, J. (2002) *An R and S-PLUS Companion to Applied Regression* (Thousand Oaks, Sage).

HILBE, J. M. (2007) *Negative Binomial Regression* (Cambridge, Cambridge University Press).

HOCKING, R. R. (1985) *The Analysis of Linear Models* (Monterey, CA, Brooks/Cole).

HORTON, R. L. (1985) *The General Linear Model: Data Analysis in the Social and Behavioural Sciences*. (Reprinted. Malabar, FL., Robert E. Krieger).

HUTCHESON, G. y SOFRONIOU, N. (1999). *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models* (London, Sage).

JUDD, C. M. y McCLELLAND, G. H. (1989) *Data Analysis. A Model-Comparison Approach* (San Diego, Harcourt Brace Jovanovich).

KLEINBAUM, D. G.; KUPPER, L. L. y MULLER, K. E. (1988) *Applied Regression Analysis and other Multivariable Methods* (2ª ed.) (Boston, Pws-Kent).

KRZANOWSKI, W. J. (1998) *An Introduction to Statistical Modelling* (London, Arnold).

LONG, J. S. (1997) *Regression Models for Categorical and Limited Dependent Variables* (Thousand Oaks, CA, Sage).

LÓPEZ-GONZÁLEZ, E. (1994) La importancia del estudio de residuos para el análisis de las condiciones de aplicación de la regresión, *Bordón,* 46:1, pp. 53-68.

LÓPEZ-GONZÁLEZ, E. (2003) Las pruebas de significación: una polémica abierta, *Bordón*, 55:2, pp. 241-252.

LÓPEZ-GONZÁLEZ, E.; RUIZ-SOLER, M. y PELEGRINA, M. (2002) Estimación de parámetros en el Modelo Lineal General y en los Modelos Lineales Generalizados. Diferencias e interpretación, *Metodología de las Ciencias del Comportamiento*, vol. especial, pp. 341-345.

LOSILLA, J. L.; NAVARRO, J. B.; PALMER, A.; RODRIGO, M. y ATO, M. (2005) *Del contraste de hipótesis al modelado estadístico* (Girona, EAP).

LUNNENBORG, C. E. (1994) *Modelling Experimental and Observational Data* (California, Duxbury).

MAXWELL, S. E. y DELANEY, H. D. (2004) *Designing Experiments and Analyzing Data. A Model Comparison Perspective* (2ª ed.) (Hillsdale, Lawrence Erlbaum Associates).

McCLELLAND, G. H. (1997) Optimal Design in Psychological Research, *Psychological Methods,* 2:1, pp. 3-19.

MCCULLAGH, P. y NELDER, J. (1989) *Generalized Linear Models* (2 ed.) (London, Chapman & Hall).

NELDER, J. y WEDDERBURN, R. (1972) Generalized Linear Models, *Journal of the Royal Statistical Society (A),* 135, pp. 370-384.

NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. J. y WASSERMAN, W. (1996) *Applied Linear Statistical Models* (4ª ed.) (Chicago, Irwin).

PEDHAZUR, E. J. (1997) *Multiple Regression in Behavioural Research: Explanation and Prediction* (3ª ed.) (New York, Holt, Rinehart & Winston).

RUIZ-SOLER, M. y LÓPEZ-GONZÁLEZ, E. (2009) El entorno estadístico *R*: ventajas de su uso en la docencia y la investigación, **revista española de pedagogía**, 67:243, pp. 255-274.

RUIZ-SOLER, M.; PELEGRINA, M. y LÓPEZ-GONZÁLEZ, E. (2000) Modelización matemática y análisis de varianza: el enfoque de la comparación de modelos, en LÓPEZ, A. M.; LÓPEZ, J. y MORENO, R. (coords.) *A.E.M.C.C.O: V Congreso de Metodología de las CC. Humanas y Sociales*, vol. 1 (Sevilla, Kronos), pp. 361-365.

SEBER, G. A. F. y LEE, A. J. (2003) *Linear Regression Analysis* (2ª ed.) (New York, Wiley).

TATSUOKA, M. (1993) Elements of the General Linear Model, en KEREN, G. y LEWIS, C. *A Handbook for Data Analysis in the Behavioural Sciences. Statistical Issues* (London, LEA), pp. 3-42.

TIMM, N. H. (2002) *Applied Multivariate Analysis* (New York, Springer).

VAN DER GEER, J. P. (1993) *Multivariate Analysis of Categorical Data: Applications* (Newbury Park, CA, Sage).

VENABLES, W. N. y RIPLEY, B. D. (2003) *Modern Applied Statistics with S* (4ª ed.) (New York, Springer-Verlag).

WEISBERG, S. (2005) *Applied Linear Regression* (3 ed.) (New York, Wiley).

WOOD, S. N. (2006) *Generalized Additive Models. An Introduction with R* (Boca Raton, FL., Chapman & Hall/CRC).

Resumen:
Análisis de datos desde el Modelo Lineal Generalizado. Una aplicación con *R*

El empleo de modelos matemáticos para la explicación de fenómenos probabilísticos ha sido imprescindible en la investigación científica. No obstante, en el ámbito educativo es frecuente trabajar con variables que no cumplen las características requeridas por el Modelo Lineal (ML), utilizado durante mucho tiempo como única opción para representar datos de dependencia; por el contrario, el Modelo Lineal Generalizado (MLG) responde muy adecuadamente a los problemas generados por la métrica de las variables. En este trabajo se comentan los aspectos particulares del MLG en relación al ML dentro del entorno en el que cobran sentido ambos: el modelado estadístico. Así mismo se anima al uso del software estadístico *R*, poco conocido en el ámbito de los estudios educativos, pero especialmente sensible a las particularidades matemáticas del Modelo Lineal Generalizado y al modo de trabajar con el modelado estadístico.

Descriptores: Modelo Lineal Generalizado, Modelo Lineal, regresión de Poisson, modelado estadístico, *R*.

Summary:
Data analysis from the Generalized Linear Model approach: an application using *R*

To use mathematical models in order to explain probabilistic phenomena have been essential in scientific research. However, in educational settings is common to work with variables which do not satisfy the demanded assumptions in the Linear Model (LM). The Generalized Linear Model (GLM) answers quite well to the problems originated by measurement variable questions. In this work some aspects of the GLM are commented in relation to the LM and this is done from the framework where both of them make sense: statistical modelling. Besides, the use of statistical software *R* is emphasized because this one is not very much known in educational research. However, this software is very appropriate for the GLM and for the working style in statistical modelling.

Key Words: Generalized Linear Model, Linear Model, Poisson regression, statistical modelling, *R*.