ESTUDIO DEL RIESGO CARDIOVASCULAR EN EL MUNICIPIO DE SANTA CLARA UTILIZANDO EL MÉTODO DE REGRESIÓN CATEGÓRICA[1]

Juan Manuel Navarro-Céspedes*, Gladys M. Casas-Cardoso**,
Emilio González-Rodríguez***y. Santiago Cuadrado-Rodríguez***:
* Departamento de Matemática, Facultad MFC, Universidad Central "Marta Abreu" de Las Villas
**Departamento de Computación, Facultad MFC, Universidad Central "Marta Abreu" de Las Villas
***Centro de Desarrollo de la Electrónica, Universidad Central "Marta Abreu" de Las Villas

ABSTRACT
This paper shows an application of the regression analysis method using categorical data. Data have been obtained from five clinics of the Santa Clara City, Cuba. *849* cases were diagnosed by a highly qualified Medical Doctors Committee. *18* categorical predictor variables and a continuous variable RiesgoC (cardiovascular risk) were used in the analysis. Initially, the standard regression analysis was applied. The determination coefficient obtained was $R^2=0.643$. Standardized residuals plot showed that there was not normal distribution. Categorical Regression Procedure of SPSS software was then applied. R^2 was now *0.833* and the standardized residuals plot showed that there was normal distribution. The categorical regression analysis proved to be an important method when the most of the predictor variables are categorical.

KEY WORDS: categorical regression, cardiovascular risk, categorical data analysis

MSC: 62P10

RESUMEN
En el presente trabajo se muestra una aplicación del método de análisis de regresión para datos categóricos. Se utilizaron datos de *849* individuos de cinco policlínicos de la ciudad de Santa Clara tomados por un comité de expertos altamente calificado. En el análisis intervienen 18 variables predictoras categóricas y la variable RiesgoC (riesgo cardiovascular) que es numérica. Inicialmente se aplicó el procedimiento de regresión lineal estándar. Se obtuvo un $R^2=0.643$. El ploteo de los residuos estandarizados mostró un marcado desajuste a la distribución normal. Posteriormente se aplicó el método de análisis de regresión categórica del paquete estadístico SPSS. El coeficiente de determinación fue ahora de $R^2=0.833$ y el ploteo de los residuales estandarizados mostró ajuste a la distribución normal. El análisis de regresión categórica constituye una buena técnica cuando se está en presencia de problemas en los que la mayoría de las variables predictoras son categóricas.

1. INTRODUCCIÓN

El análisis de regresión lineal estándar es una técnica estadística ampliamente utilizada desde la segunda mitad del siglo XIX, cuando el científico británico Francis Galton introdujo dicho término [Stanton (2001)]. El análisis de regresión lineal clásico minimiza las diferencias de la suma de los cuadrados entre la variable respuesta (dependiente) y una combinación ponderada de las variables predictoras (independientes). Las variables son normalmente cuantitativas, con los datos categóricos (nominales) recodificados como variables binarias. Los coeficientes estimados reflejan cómo los cambios en las variables predictoras afectan a la variable respuesta. Puede obtenerse un pronóstico de esta última para cualquier combinación de los valores predictores [Draper-Smith (2002)].

[1] Presented at COMPUMAT 2007

.

En numerosas investigaciones, sobre todo en el campo médico o social, se tienen variables predictoras categóricas. Algunas tienen un orden entre sus valores, otras son simplemente nominales [Agresti (2002)].

En estos casos pudiera pensarse en realizar una regresión de la respuesta con respecto a los propios valores predictores categóricos. Como consecuencia, se estima un coeficiente para cada variable. Sin embargo, para las variables discretas, los valores categóricos son arbitrarios. La codificación de las categorías de diferentes maneras proporciona diferentes coeficientes, dificultando las comparaciones entre los análisis de las mismas variables. De manera general, la aplicación de las técnicas clásicas de regresión se dificulta notablemente.

Por otra parte, no existen dudas de que la regresión lineal múltiple es la técnica estadística más utilizada para predecir el comportamiento de una variable dependiente, a partir de los valores de varias independientes. Lo que ocurre es que no siempre tal relación es lineal. A través de los años se han reportado en la literatura numerosas contribuciones que son en esencia, generalizaciones no lineales de la regresión [Mccullagh-Nelder (1989)], [Berthold (2007)]. Puede mencionarse por ejemplo, el desarrollo reciente de varios métodos de regresión no lineal en el área de minería de datos, conocidas por el nombre inglés de "machine learning" o aprendizaje automatizado. En aras de obtener una definición más cercana al ambiente estadístico, ha surgido el término "statistical learning" (aprendizaje estadístico) para referenciar a métodos como estos [Hastie-Tibshirani-Friedman (2001)]. El método que es objeto de estudio de este artículo: la Regresión Categórica, trabaja bajo el enfoque de la regresión con transformaciones, aplicando la metodología del escalamiento del óptimo desarrollado por el sistema Gifi [Gifi (1990)] para transformar la respuesta y los predictores. En el enfoque de la regresión con transformaciones, los predictores y/o la variable respuesta se transforman de manera no lineal sin considerar ajuste de distribución. Por tanto, la relación entre la respuesta y los predictores se linealizan a través de transformaciones no lineales separadas de las variables, dando lugar a un modelo flexible. La función de transformación puede ser paramétrica o no paramétrica. Un ejemplo de uso de funciones no paramétricas es el MORALS implementado en el TRANSREG [Van der Kooij (2007)]. El SPSS (Statistical Package for the Social Science), desde su versión 11, trae incorporado un módulo para realizar regresiones categóricas. Existen otros paquetes estadísticos que tiene implementado la regresión para variables categóricas como es el caso del S-Plus [Lam (2008)], pero con un enfoque diferente. El ejemplo sobre un estudio para predecir el riesgo cardiovascular en el municipio de Santa Clara, que se desarrollará más adelante, será ejecutado con ayuda del SPSS.

2. FUNDAMENTOS DE LA REGRESIÓN CATEGÓRICA

La regresión categórica cuantifica los datos categóricos mediante la asignación de valores numéricos a las categorías, obteniéndose una ecuación de regresión lineal óptima para las variables transformadas. La regresión categórica se conoce también por el acrónimo CatReg, del inglés Categorical Regression [Haber (2001)].

La regresión categórica extiende la regresión lineal ordinaria, considerando simultáneamente variables continuas, ordinales y nominales. Las variables categóricas se cuantifican de manera que ellas reflejen las características de las categorías originales, utilizando transformaciones no lineales para hallar el modelo que mejor ajuste. Finalmente las variables cuantificadas se tratan de la misma forma que las variables continuas [Van der Kooij-Meulman (1997)].

El objetivo fundamental de la regresión categórica con escalamiento óptimo consiste en describir las relaciones entre una variable respuesta y un conjunto de variables predictoras [De Leeuw (2005)]. El propósito es, en esencia, el mismo que cualquier otro análisis de regresión. Lo interesante en este caso es que la CatReg puede aplicarse para aquellas variables en las que los análisis clásicos o estándares de regresión fallan. De hecho, si se realiza un análisis de regresión lineal sobre las variables transformadas, se obtienen los mismos resultados que con el análisis de regresión categórica.

La regresión categórica constituye una generalización de varias técnicas estadísticas, por ejemplo si la variable dependiente es continua y se tiene sólo una independiente con nivel de medición nominal, entonces la

regresión categórica se convierte en un análisis de varianza clásico (ANOVA). Con un nivel de escalamiento nominal, la cuantificación para cada categoría es la media de los valores de la variable dependiente tomando los casos que pertenecen a esa categoría. La variable transformada coincide entonces con la variable original, en la que los valores de la categoría se sustituyen por los valores de su media y el resultado se estandariza. Si se tiene una variable dependiente nominal, la regresión categórica se convierte entonces en un análisis discriminante clásico [Meulman-Heiser (2004)].
Por otra parte, CatReg es equivalente al análisis de correlación canónica categórico mediante escalamiento óptimo (OVERALS) con dos conjuntos, uno de los cuales contiene sólo una variable. Si se escalan todas las variables a nivel numérico, el análisis se corresponderá con el análisis de regresión múltiple típico.

3. ESTUDIO DE RIESGOS CARDIOVASCULARES

A nivel mundial existen diversos tipos de enfermedades cardiovasculares. Según la Organización Mundial de la Salud, las enfermedades cardiovasculares causan 12 millones de muertes en el mundo cada año. Gracias a muchos estudios y miles de pacientes, los investigadores han descubierto ciertos factores que desempeñan un papel importante en las probabilidades de que una persona padezca de una enfermedad del corazón, a ellos se les denomina factores de riesgo [Wilson et. al. (1998)]. Como ejemplo de estudio se puede citar el Framingham Heart Study cuyo objetivo era identificar los factores comunes de riesgos que contribuían a la enfermedad cardiovascular. Entre los principales factores de riesgo podemos citar: la presión arterial, el colesterol elevado, la diabetes entre otros [Wilson et. al. (1998)].

Tabla 1. Variables consideradas en el análisis

Variable dependiente:

Variable	Etiqueta	Valores
RiesgoC	Riesgo cardiovascular	Real

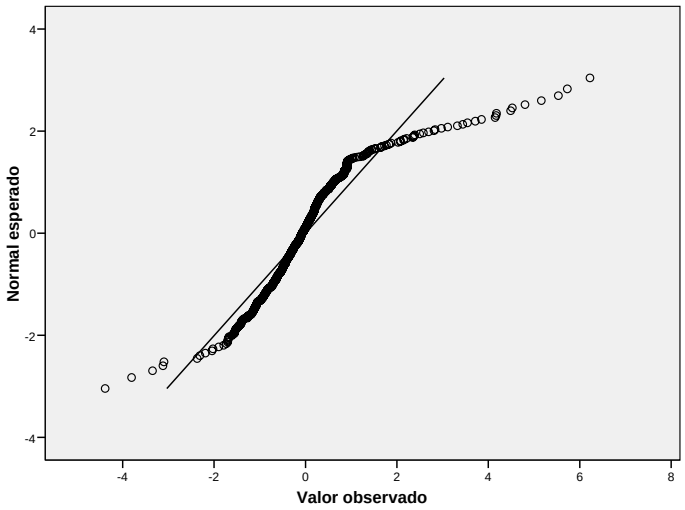
Variables independientes:

Variable	Etiqueta	Valores
PAM	Presión arterial media	Baja, Media, Alta
IMC	Índice de masa corporal	Bajo, Medio, Alto
Sexo	Sexo del paciente	Masculino, Femenino
Edad	Edad del paciente	16- 80 años
Coltot	Colesterol total	Bajo, Medio, Alto
Fuma	Hábito de fumar	Sí, No
Bebe	Habito de tomar	Sí, No
Diabetes	Diabetes mellitas	Sí, No
Raza	Raza del paciente	Blanca, Mestiza
Rsisbas	Presión sistólica basal	Bajo, Medio, Alto
Rsismin1	Presión sistólica basal en el primer minuto	Bajo, Medio, Alto
Rsismin2	Presión sistólica basal en el segundo minuto	Bajo, Medio, Alto
Rdiastbas	Presión diastólica basal	Bajo, Medio, Alto
Rdiastmin1	Presión diastólica basal en el primer minuto	Bajo, Medio, Alto
Rdiastmin2	Presión diastólica basal en el segundo minuto	Bajo, Medio, Alto
ColLDL	Colesterol LDL	Bajo, Medio, Alto
ColHDL	Colesterol HDL	Bajo, Medio, Alto
Diagnóstico	Diagnóstico del paciente dado por el Comité de	Hipertenso, Hiperreactivo

	Expertos	vascular, Normotenso

En el estudio que se presenta en este trabajo, la muestra está constituida por un total de 849 individuos pertenecientes a cinco policlínicos de la ciudad de Santa Clara y pretende predecir el riesgo a padecer enfermedades cardiovasculares a partir de la presencia de varias variables, todas categóricas. En la tabla 1 muestran las variables originales que se midieron a cada uno de los pacientes.

Figura 1. Gráfico Q-Q normal de residuos estandarizados



En la Tabla 2 se muestran sus coeficientes estandarizados.

Tabla 2. Coeficientes

	Standardized Coefficients				
	Beta	Std. Error	df	F	Sig.
Edad	.679	.015	25	1934.493	.000
Sexo	.331	.016	1	402.210	.000
Bebe	.055	.016	1	12.104	.001
Fuma	-.162	.015	1	113.972	.000
Diabetes mellitus	-.037	.015	1	6.188	.013
Diagnóstico	.045	.021	1	4.430	.036
rsistbas	.073	.025	2	8.519	.000
rdiastbas	.075	.021	1	12.274	.000
rdiastmin2	.026	.017	1	2.326	.128
dcolldl	-.032	.015	2	4.542	.011
rpam	.042	.029	2	2.062	.128
rsistmin1	.064	.029	2	4.825	.008
Raza	.009	.015	1	.346	.557
rsistmin2	.003	.020	1	.024	.876
rdiastmin1	-.058	.025	2	5.433	.005
dcoltot	-.015	.015	2	.957	.384
dcolhdl	.029	.015	2	3.721	.025
dimc	-.006	.015	1	.141	.707

Dependent Variable: RiesgoC

El problema que se presenta en esta investigación no puede tratarse adecuadamente por una regresión lineal múltiple, debido a que la variable independiente es numérica, mientras que todas las predictoras son categóricas. No obstante, el primer intento de lograr un modelo se realizó siguiendo esta alternativa. Al aplicar la misma se obtuvo un coeficiente de determinación R^2 *=0.643*, pero al realizar un gráfico de los residuos se obtuvo un desajuste muy marcado de la distribución normal, véase el gráfico *Q-Q* en la figura 1.

Se decidió entonces aplicar la opción de regresión categórica del paquete SPSS.

En la primera corrida de la CatReg se consideraron todas las variables, que aparecen mencionadas en la Tabla 1. El escalamiento óptimo utilizado en el estudio está en total correspondencia con el nivel de medición de las variables. El valor del coeficiente de determinación fue ahora de R^2*=0.835*, lo cual indica que el *83.5%* de la variable RiesgoC está explicado en el modelo. El resultado del análisis de varianza resultó significativo mostrando la validez del modelo

El análisis de regresión categórica no tiene implementado métodos paso a paso, sino el método más sencillo y directo en el que todas las variables consideradas en el análisis pasan a formar parte de la ecuación. En la Tabla 2 se puede apreciar que algunas de las variable analizadas no son significativas.

Se decidió entonces obtener nuevamente el modelo eliminando la variable menos importantes. Este procedimiento se ejecutó repetidamente hasta que se obtuvo una ecuación válida (desde el punto de vista del análisis de varianza), donde todas las variables resultaron significativas. Los resultados hallados se muestran a continuación:

Tabla 3. Resumen del Modelo

Multiple R	R Square	Adjusted R Square
.913	.833	.825

Dependent Variable: RiesgoC
Predictors: Edad Sexo Bebe Fuma Diabetes mellitus Diagnóstico rsistbas rdiastbas rdiastmin2 dcolldl rpam rsistmin1

En este nuevo modelo se aprecia que el *83.3%* de la variable RiesgoC está explicado por los predictores. Téngase en cuenta que se eliminaron seis variables con respecto al modelo original y el valor del R^2 prácticamente se quedó igual. El ANOVA resultó ser significativo nuevamente indicando la validez del modelo.

En la Tabla 4 se muestran los coeficientes estandarizados asociados al modelo final. El valor de cada coeficiente indica el cambio que esa variable produce en la variable dependiente RiesgoC. Como que la tabla muestra los coeficientes estandarizados, las interpretaciones se basan en las desviaciones estándares de cada uno de ellos [Agresti (2002)].

Además se realizó un estudio gráfico de los residuales. En la figura 2 puede notarse una mejoría considerable con respecto al gráfico visto en la ver figura 1, pues la mayoría de los valores observados se encuentran sobre la línea diagonal (que representa los valores esperados si la distribución fuese normal).

Además del examen gráfico, para el análisis de los supuestos se utilizó el test de Kolmogorov Smirnov para comprobar usando un test estadístico, que los residuos estaban normalmente distribuidos. La significación hallada fue de 0.161, lo que indica normalidad. Para verificar la homogeneidad de las varianzas y para comprobar la ausencia de multicolinealidad se realizó una regresión lineal tomado como datos los valores de las variables transformadas [Van der Kooij (2007)] ya que el módulo de regresión categórica implementado aún no realiza este tipo de análisis [Meulman-Heiser (2004)].

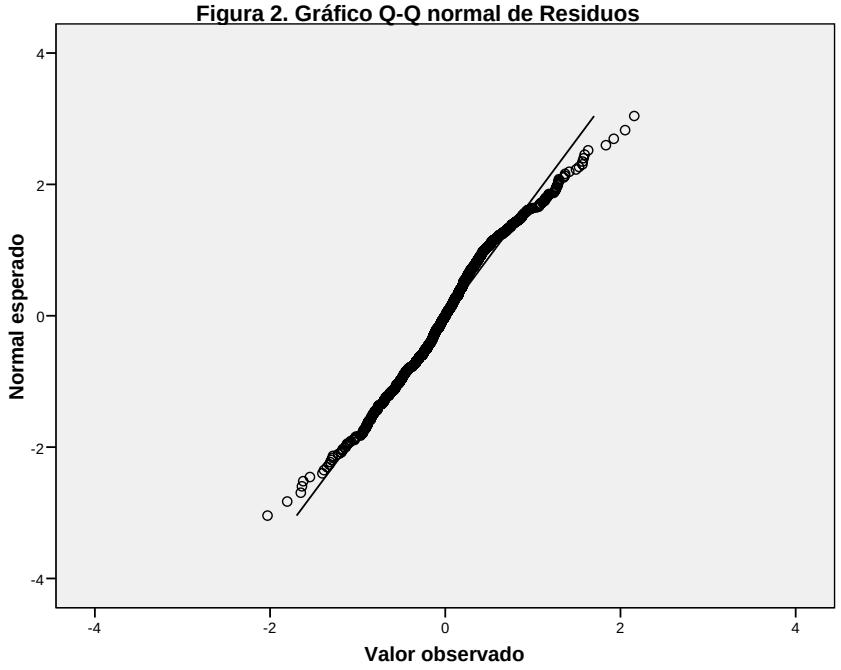
Tabla 4. Coeficientes

	Standardized Coefficients				
	Beta	Std. Error	df	F	Sig.
Edad	.679	.015	24	1965.113	.000
Sexo	.334	.016	1	419.291	.000
Bebe	.057	.016	1	12.895	.000
Fuma	-.161	.015	1	113.957	.000
Diabetes mellitus	-.036	.015	1	6.136	.013
Diagnóstico	.043	.019	1	5.005	.026
rsistbas	.076	.024	2	9.955	.000
rdiastbas	.073	.020	1	13.924	.000
rdiastmin2	-.058	.023	1	6.346	.012
dcolldl	-.030	.015	2	4.281	.014
rpam	.055	.028	2	3.844	.022
rsistmin1	.068	.025	2	7.453	.001

Dependent Variable: RiesgoC

El estadístico de Durbin Watson obtenido fue de 1.534, indicando que no hay autocorrelación [Calero (1998)]. El índice de condición obtenido fue de 5.520 lo que reafirma la ausencia de multicolinealidad.

Figura 2. Gráfico Q-Q normal de Residuos



4. CONCLUSIONES

El análisis de regresión categórica resulta ser una buena opción cuando nos enfrentamos a problemas en los que la mayoría de las variables analizadas son del tipo categóricas. Aplicando esta técnica se puede realizar

un estudio para descubrir relaciones no lineales entre las variables predictoras y obtener modelos adecuados para realizar predicciones.
En el ejemplo que se desarrolla se obtiene un modelo de regresión que cumple con todos los supuestos y por tanto puede utilizarse para predecir el riesgo cardiovascular. Su coeficiente determinación indica que el 83.3% de la variable respuesta es explicado por las predictoras.

Received September 2007
Revised June 2008

REFERENCIAS

[1] AGRESTI, A. (2002): **Categorical Data Analysis**. Second edition, John Wiley & Sons, Inc., Publication.

[2] BERTHOLD M, H. (2007): **Intelligent Data Analysis.** Second Edition. 2007: Springer, Berlin.

[3] CALERO, A. (1998): **Estadística II**. La Habana. Cuba: Pueblo y Educación, La Habana.

[4] DE LEEUW, J. (2005): Multivariate analysis with optimal scaling. **Department of Statistics, UCLA**. [cited 5 May 2006]; Available from: **http://repositories.cdlib.org/uclastat/papers/2005103002/**

[5] DRAPER, N.R. AND H. SMITH (2002): **Applied regression analysis.** Third edition. : Wiley, New York.

[6] GIFI, A. (1990): **Nonlinear multivariate analysis,** Wiley, Chichester.

[7] HABER, L. (2001): Categorical regression analysis of toxicity data. **Comments on toxicology.** 7, 437-452.

[8] HASTIE, T.J., R.J. TIBSHIRANI, and J.H. FRIEDMAN (2001): **The Elements of Statistical Learning**. Springer, New York.

[9] WILSON, P.W.F, D'AGOSTINO, R.B, LEVY, D. BELANGER, A.M, SILBERSHATZ, H, KANNEL W.B. (1998): Prediction of coronary heart disease using risk factor categories. Available from: **http://www.e-psicometria.com/Portals/0/Bibliografia/Framingam_revisi%C3%B3n.pdf**

[9] LAM, L. (2008): **An introduction to S-Plus for Windows**. [cited 5 May 2008]; Available from: **http://www.splusbook.com/**.

[10] MCCULLAGH, P. and J.A. NELDER (1989): **Generalized Linear Models**. Chapman and Hall, London.

[11] MEULMAN, J.J. AND W.J. HEISER (2004): **SPSS Categories** 13.0. SPSS Inc. Chicago.

[12] STANTON, J.M., (2001): GALTON, Pearson and the Peas: A brief history of linear regression for statistics instructors. **Journal of Statistics Education**. **9,** 64-87

[13] VAN der KOOIJ, A.J. and J.J. MEULMAN (1997): MURALS: Multiple regression and optimal scoring using alternating least squares. In: W. Bandilla and F. Faulbaum (Eds.), **Softstat '97 Advances in Statistical Software 6**. Stuttgart: Lucius \& Lucius, pp 99-106

[14] VAN der KOOIJ, A.J. (2007): **Prediction accuracy and stability of regression with optimal scaling transformations**. [cited 24 Jan 2008; Doctoral thesis.]. Available from: **https://www.openaccess.leidenuniv.nl/dspace/handle/1887/12096**.