ELECCIÓN DEL MEJOR MODELO DE REGRESIÓN LOGÍSTICA MEDIANTE EL USO DE LA REGRESIÓN NO PARAMÉTRICA

Eliseo Gabriel Argüelles*, Ernesto Pedro Menéndez Acuña**[1] y Yojan Guevara Pérez***
*Facultad de Estadística e Informática. Universidad Veracruzana. México.
**Facultad de Matemáticas. Universidad Veracruzana. México.
*** Médico Internista. Hospital Básico Guamote-Ministerio de Salud Pública, Chimborazo, República del Ecuador.

ABSTRACT
In this paper, it is presented a study to justify the use of nonparametric regression for obtaining evidence of the linear model to be considered in the application of logistic regression with a single independent variable. By a simulation study, three scenarios were considered, each showing different relationships between the independent variable "x" and the probability "π", when the dependent variable takes the value 1, assuming a Bernoulli probability distribution. The results obtained, with the application of the nonparametric regression, showed the possibility of selecting the most appropriate model. Finally, it is presented an application in a steatosis study.

KEYWORDS: Logistic regression; Nonparametric regression; Scatterplot; Smoothing Splines.

MSC: 62G08; 62J12

RESUMEN
En este trabajo se presenta un estudio para justificar el uso de la regresión no paramétrica para obtener evidencia del modelo lineal a considerar en la aplicación de la regresión logística con una sola variable independiente. Se realizó un estudio por simulación en el que se consideraron tres situaciones, mostrándose en cada una diferentes relaciones entre la variable independiente y la probabilidad "π" que genera el valor 1 para la variable dependiente, bajo el supuesto de una distribución de probabilidad Bernoulli. Los resultados obtenidos, con la aplicación de la regresión no paramétrica, mostraron en cada una de las situaciones consideradas la posibilidad de seleccionar el modelo más adecuado. Finalmente se presenta una aplicación en un estudio de esteatosis.

Palabras claves: Diagrama de dispersión; Regresión logística; Regresión no paramétrica; Suavizamiento Spline.

1. INTRODUCCIÓN

Una actividad importante que se desarrolla dentro de la estadística es la construcción de modelos que se espera reflejen aspectos importantes del objeto de estudio con un cierto grado de realismo (Seber and Lee, 2003). La modelación estadística, basada en el análisis de regresión brinda una herramienta muy útil y potente para realizar el trabajo de modelación. Bajo el nombre de análisis de regresión, se incluyen un conjunto de técnicas estadísticas que tratan de explicar cómo se relaciona una variable respuesta con una o más variables predictoras. Lo que caracteriza en principio a las distintas clases de modelos de regresión es la naturaleza de la variable dependiente; con variables continuas los modelos de regresión lineal son los más utilizados, sin embargo, cuando la distribución de la variable dependiente no sea la Normal, un modelo lineal generalizado (GLM) es adecuado. En el caso particular de que la variable dependiente tenga una distribución Bernoulli, el modelo de regresión logística (LR), como un caso particular de un GLM, es el utilizado (Núñez, Núñez y Steyerberg, 2011). Los GLM son una extensión de los modelos lineales, creados para considerar variables dependientes con distribución no Normal y establecer funciones de modelación de la media (Agresti, 2002). Para establecer un GLM se deben identificar tres componentes: un componente aleatorio en la variable respuesta que fija la distribución de probabilidad, un componente sistemático que especifica la estructura de las variables predictoras utilizadas y una función de enlace que relaciona el componente aleatorio y el componente sistemático a través de una función de $E(y|x)$.

[1] emenendeza@gmail.com

Actualmente el uso de la LR se ha hecho cada vez más popular (Calderón y Alzamora, 2009). Muchas y variadas aplicaciones se realizan en diversas áreas: bioquímica, cardiología, psiquiatría, psicología, turismo, epidemiología, educación, etc. Ejemplos de algunas de estas aplicaciones son consideradas en los trabajos de Gunvig, Hansen y Borggaard (2013); Chen y Wang (2013); Montemagni, Frieri, Villari y Rocca (2012); Teich, Marty, Gollut, Gret-Regamey y Bebi (2012); Gautam (2012); Calderon y Alzamora (2009); Alderete (2006); García, Alvarado y Jiménez (2000).

Desde la perspectiva paramétrica, para aplicar el análisis de regresión lineal es necesario conocer la estructura del modelo que mejor exprese la relación entre la variable dependiente y las variables independientes. En el caso de una sola variable independiente, que es el interés de este trabajo, si no existe un conocimiento previo acerca de la posible relación entre estas variables, un diagrama de dispersión puede ayudar a obtener evidencia de esta posible relación (Sheather, 2009). En el caso de la aplicación de la LR, construir un diagrama de dispersión tiene un valor nulo (Györfi, Kohler, Krgyźak y Walk, 2002). Como es conocido, la variable respuesta sólo puede tomar uno de dos posibles valores 1 o 0 y en consecuencia el diagrama de dispersión siempre mostrará una forma muy peculiar, que consiste en puntos ordenados sobre las líneas rectas paralelas $y = 1$ y $y = 0$, que no proporciona evidencia de la posible relación que presentan las variables. Agresti (2002), Ryan (1997) y Wright y London (2009), mencionan que, si se cuenta con mediciones repetidas en la variable independiente o se agrupan los datos de esta variable en categorías, una manera de proceder para identificar un modelo de LR, es graficar en un diagrama de dispersión la proporción de éxito versus la variable independiente; esto daría una idea de la forma que tendría el modelo. Otra alternativa que pudiera ser utilizada siguiendo esta idea para seleccionar el modelo, es mediante la utilización de una regresión no paramétrica (NPR).

Por esta razón el presente trabajo tiene por objetivo, mostrar como la regresión no paramétrica sustituye al diagrama de dispersión para determinar la forma del modelo que mejor represente la relación entre las variables bajo estudio, en la aplicación de la regresión logística con una sola variable independiente.

2. LA REGRESIÓN NO PARAMÉTRICA Y LA REGRESIÓN LOGÍSTICA

Un enfoque alternativo que se utiliza también para realizar un análisis de regresión, sin imponer un modelo a los datos, es mediante el empleo de la NPR. El uso de la NPR se centra en determinar gráficamente la relación existente entre las variables; no imponiendo un modelo antes de realizar el análisis de regresión, sino, que el modelo es determinado por los datos, de acuerdo a la forma que despliega la relación. (Takezawa, 2006; Eubank, 1999; Ruppert, Wand y Carroll, 2003). La RNP es una colección de técnicas denominadas métodos de suavización, que permite estimar la forma funcional de la función de regresión, donde cualquier suposición a priori de linealidad que se establezca es remplazada con la suposición mucho más débil de una función de regresión suave; por consiguiente, es apropiado utilizar LR cuando no existe conocimiento previo sobre la relación entre las variables de estudio, o cuando el uso de la modelación paramétrica es muy complicada, dada la estructura de la relación entre las variables dependiente e independientes. Esta característica hace muy flexible este tipo de regresión (Eubank, 1999). La RNP no asume un modelo particular. El modelo en este caso es muy general y está dado por

$$E(y|x) = m(x)$$

donde $m(x)$ es una función suave desconocida que expresa la forma funcional de la relación entre y y x. El objetivo es estimar la forma funcional de $m(x)$ de los datos (Keele, 2008). Para lograr esto se puede utilizar algún método estimación de RNP, como el spline cúbico (Takezawa, 2006). El suavizador spline, también llamado suavizamiento spline, es una de las técnicas más importantes que se utilizan para modelar la relación entre variables y esta técnica consiste en minimizar el siguiente argumento:

$$\min_m \frac{1}{n}\sum_{i=1}^{n}\left(y_i - m(x_i)\right)^2 + \lambda\int_a^b \left(m^p(x)\right)^2 dx$$

cuya solución es una función del espacio de Sobolev $W_2^p[0,1]$ (GU, 2002). En particular si $p = 2$ el suavizador que se obtiene es el spline cúbico. Una vez estimada $m(x)$ por $\hat{m}(x)$, para cada valor de x, una estimación de $E(y|x)$ es obtenida y esta información es, en cierto sentido, equivalente a la información proporcionada por un diagrama de dispersión.

Para el caso del estudio de la relación entre una variable dependiente dicotómica y, con distribución Bernoulli con parámetro $\pi = P\{y = 1|x\}$ y una variable independiente cuantitativa x, el modelo más simple a considerar esta dado por

$$E(y|x) = \beta_0 + \beta_1 x = \pi \qquad (1)$$

pero la expresión (1) no tiene sentido, porque una vez estimado los parámetros, el modelo ajustado $\hat{\pi} = \hat{\beta}_0 + \hat{\beta}_1 x$ no garantiza que los valores estén dentro del intervalo $(0,1)$. Una alternativa es formular la relación en términos de una función $g(.)$, esto es

$$\pi = g(\beta_0 + \beta_1 x)$$

Entonces, siempre que $g(.)$ sea una función de distribución, sus valores estarán en el intervalo $(0,1)$. Si se asume que

$$\pi = g(\beta_0 + \beta_1 x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad (2)$$

$g(.)$ es la función de distribución logística con parámetro de localización $\mu = 0$ y parámetro de escala $s = 1$, evaluado en $\beta_0 + \beta_1 x$. De la expresión (2) se obtiene

$$In\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x \quad (3)$$

Otras expresiones que suelen aparecer son aquellas de segundo grado y en general de grado n, como se muestran en (4), y (5)

$$In\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x + \beta_2 x^2 \quad (4)$$

$$In\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n \quad (5)$$

En términos generales, la expresión (3) representa un modelo de regresión logístico, que es un caso especial de los GLM, donde la función de enlace es $In(\pi/1-\pi)$, que es conocida como la función logit. El componente sistemático es el predictor lineal $\beta_0 + \beta_1 x$ y el componente aleatorio es la variable aleatoria con distribución Bernoulli. Entonces, el aplicar una NPR a datos correspondientes a un análisis de LR, lo que se obtiene es la estimación de $E(y|x)$, que es el valor estimado de π, que es el objeto de estimación en un análisis de LR.

3. METODOLOGÍA

Con el propósito de mostrar como el uso de un diagrama de dispersión puede ser sustituido por una NPR para obtener información de la componente sistemática que pueda ser considerada en la LR, se realizó un estudio de simulación. En este estudio se crearon tres situaciones diferentes, cada una con una relación entre la probabilidad de que la variable dependiente tome el valor 1 y la variable independiente x, mostradas por un diagrama de dispersión de $\pi_i \; vs \; x_i$, para $i = 1, \ldots, 150$. En la primera situación, la gráfica de la relación presenta la forma tradicional sigmoidea, para la cual el modelo de la expresión (3) es el adecuado. Las otras dos gráficas que presentan las restantes relaciones muestran patrones de comportamiento diferentes al que hace recomendable el uso de la LR con la componente sistemática fijada por un modelo lineal simple. Estas dos últimas relaciones se muestran en (4) y (5) con $n = 6$. Con las probabilidades dadas para cada una de las tres diferentes relaciones, se generaron 100 muestras aleatorias de tamaño 150, con distribución Bernoulli y parámetro π_i. La consideración de estas muestras tiene como objetivo validar el comportamiento del suavizamiento por Spline, para visualizar el tipo de modelo a ajustar y cuánto puede influir la muestra. Debido a que todas las muestras produjeron resultados similares correspondientes a cada una de las tres situaciones consideradas, en el análisis de los resultados solo se muestran cinco de ellas elegidas aleatoriamente.

Es de resaltar que el proceso que se describe para generar los valores de la variable dependiente es totalmente ficticio, ya que en la práctica los valores π_i son desconocidos. Sin embargo este proceso de generar los valores de las variables respuesta a partir del conocimiento de las verdaderas probabilidades nos dará una idea de qué tanto nos acercamos a éstas, cuando sólo se trabaja con una variable respuesta dicotómica "y" y una sola variable predictora "x". Por último, se realiza una aplicación en un estudio de esteatosis.

El proceso de análisis se realizó en dos etapas. La primera consistió en realizar un suavizamiento spline cúbico para cada una de las muestras, con el propósito de obtener evidencia de la forma en que se relacionan las probabilidades y la variable predictora para hacer la elección del modelo. En la segunda etapa se efectuó un análisis de LR con el modelo sugerido por el resultado de la etapa anterior con el objeto de comprobar gráfica y analíticamente si el modelo elegido es el correcto. Para evaluar los ajustes de los modelos se utilizaron diferentes criterios de selección, tales como la prueba de la Deviance, la prueba de la diferencia de la Deviance, el estadístico de Wald, R^2 y AIC. El nivel de significación considerado fue del 5%, La realización de los análisis se llevó a cabo mediante el empleo del paquete estadístico R Core Team (2017).

3.1. Análisis de resultados

En la figura 1 se presenta un panel de gráficos en el que se muestran varios diagramas de dispersión. Estos diagramas representan el resultado de analizar la primera relación entre la variable independiente y la probabilidad π. En la figura 1, el gráfico (*a*) muestra el diagrama de dispersión donde se graficaron las probabilidades originales junto con los valores obtenidos de la variable respuesta de la primera muestra *versus* la variable independiente. En este gráfico se observa que el patrón de comportamiento que presentaron las probabilidades tiene una forma sigmoidea. En el gráfico (*b*) se muestra el resultado de aplicar un suavizamiento spline Cúbico como análisis inicial, donde el número de nudos a considerar y la cantidad de suavizado fue determinado automáticamente. Se observa que los valores obtenidos se encuentran muy próximos a las probabilidades π_i originales, lo que indica que es adecuado utilizar una LR con el modelo (3).

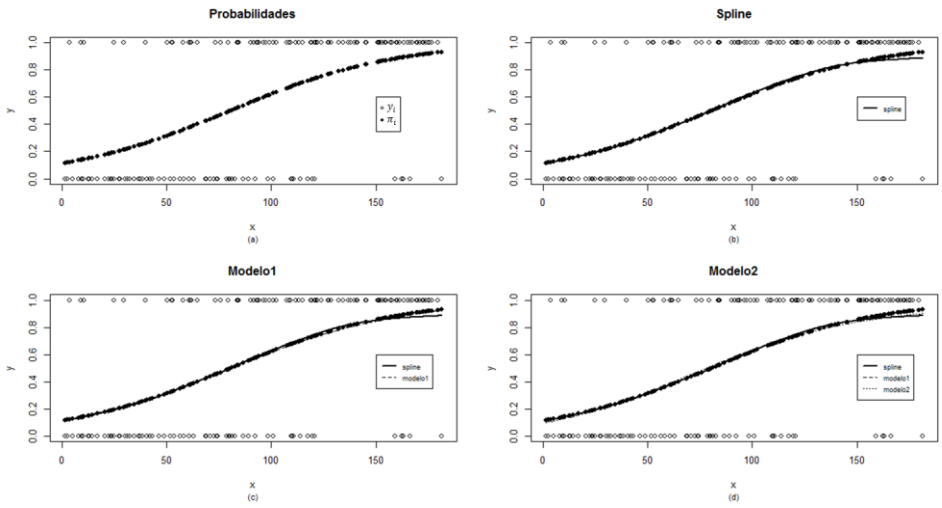

Figura 1: (*a*) Forma de la relación entre x y π; (*b*) Forma de la estimación obtenida por el spline cúbico; (*c*) y (*d*) Formas de los ajustes polinomiales de primero y segundo orden para la componente sistemática de una LR.

En el gráfico (*c*) se procede a establecer una LR bajo el modelo antes señalado, y se observa que el resultado obtenido es adecuado. El gráfico (*d*) se muestra el resultado de aplicar un modelo de LR de segundo grado. En este gráfico se observa que el ajuste obtenido no mejora considerablemente en comparación con el ajuste anterior. Esto nos da evidencia de quedarnos con el modelo con componente sistemático lineal.
En lo referente al estudio analítico que se realizó para apoyar las conclusiones que se obtuvieron en el análisis anterior, se encontró que al aplicar el criterio de la prueba de la Deviance para el modelo de LR simple (ver tabla 1), éste hace un buen ajuste a los datos (p = 0.1051635). El resultado es confirmado al observar el coeficiente de determinación R^2 que resulta ser alto (R^2 = 0.776261).

Tabla 1: Ajustes polinomiales de primero y segundo orden para la componente sistemática de una LR.

Modelo / Criterio	Mod1	Mod2
β_0	-2.01904*	-2.204*
β_1	0.02502*	3.06e-2
β_2		-3.12e-5
Dev.	0.105163	0.070253
Dif. Dev.	1.312e-11	0.7263
R^2	0.776261	0.778339
AIC	42.94965	44.82707

*Significativo al 5%

En lo correspondiente al análisis para el ajuste del modelo de LR cuadrático, el criterio de la prueba de la Deviance, aunque sugiere que el ajuste del modelo es adecuado (p=0.070253), la prueba de la diferencia de la Deviance nos informa que el agregar un predictor cuadrático al modelo, éste resulta ser no significativo (p=0.7263). Este resultado es confirmado al observar que el AIC resulta ser más pequeño para el primer

modelo ajustado, lo cual indica que no hay una ganancia sustancial al adicionarse un término cuadrático al modelo.

En el siguiente panel de gráficos se presentan cuatro nuevas muestras generadas a partir de las probabilidades π_i y en cada una de estas muestras se efectúa una estimación por suavizamiento spline cúbico y se ajusta un modelo de LR con predictor lineal y uno con predictor cuadrático, los cuales son denotados dentro de los gráficos como modelo 1 y modelo 2, respectivamente (ver figura 2). Y se puede observar que con la realización del suavizamiento por spline cúbico nuevamente se recupera la forma distribucional que presentaron las probabilidades. Además, se puede observar, que tanto visual como analíticamente (ver tabla 2), al igual que para el análisis elaborado para la primera muestra, un ajuste con un componente sistemático lineal resultó ser el mejor.

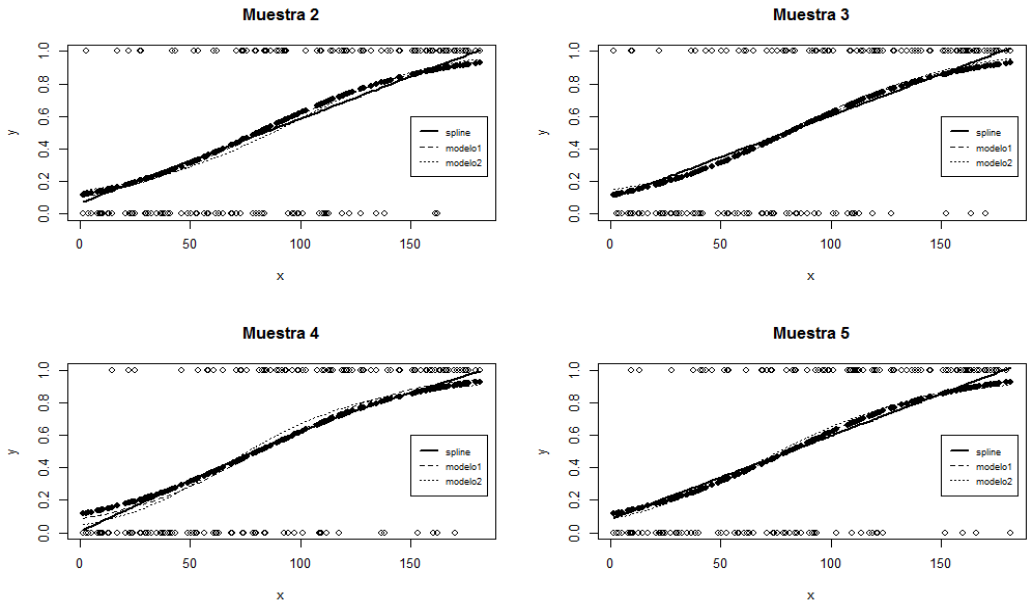

Figura 2: [•••] **F**orma de la relación entre x y π; [—]Forma de la estimación obtenida por el spline cúbico; [– – –] y [·····] Formas de los ajustes polinomiales de primero y segundo orden en la componente sistemática de una LR para las muestras generadas 2,3,4 y 5.

Tabla 2: Ajustes polinomiales de primero y segundo orden en la componente sistemática de una LR para las muestras 2,3,4 y 5.

	Modelo LR	Criterio	Valor p	R^2	AIC
Muestra 2	Modelo 1	Deviance	0.591591	0.883536	36.52287
	Modelo 2	Deviance	0.538087	0.892213	38.03876
		Dif. Dev	0.4866		
Muestra 3	Modelo 1	Deviance	0.986947	0.964242	33.48835
	Modelo 2	Deviance	0.973144	0.965258	35.43767
		Dif. Dev	0.8219		
Muestra 4	Modelo 1	Deviance	0.741108	0.916479	35.80697
	Modelo 2	Deviance	0.758783	0.932244	36.83437
		Dif. Dev	0.324		
Muestra 5	Modelo 1	Deviance	0.209422	0.811035	41.75653
	Modelo 2	Deviance	0.156671	0.815523	43.49844
		Dif. Dev	0.6114		

*Significativo al 5%

En la siguiente fase del estudio se presenta el análisis correspondiente a la segunda relación entre la variable independiente y la probabilidad "π". En la figura 3 se muestra que el patrón de comportamiento que presenta

el diagrama de dispersión donde se graficaron las probabilidades originales y los valores obtenidos de la variable respuesta de la primera muestra *versus* la variable predictora no es sigmoidea, sino la de una parábola abierta hacia arriba

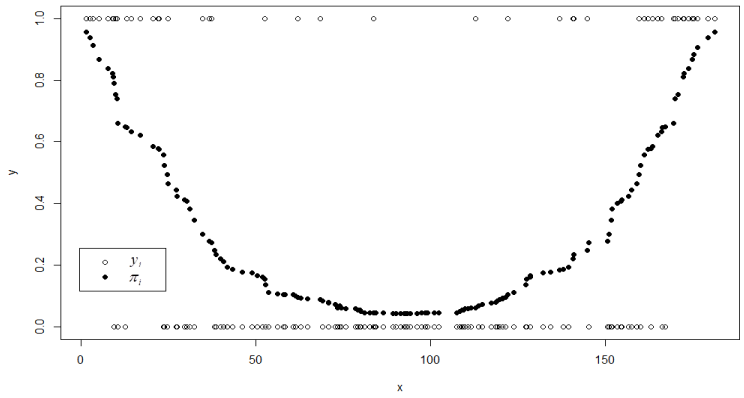

Figura 3: Forma de la relación entre x y π.

Al aplicar un suavizamineto spline cúbico como un análisis inicial (ver figura 4, gráfico (a), se observa que las estimaciones obtenidas por este método están aproximadamente muy cercanas a los valores de las probabilidades originales.

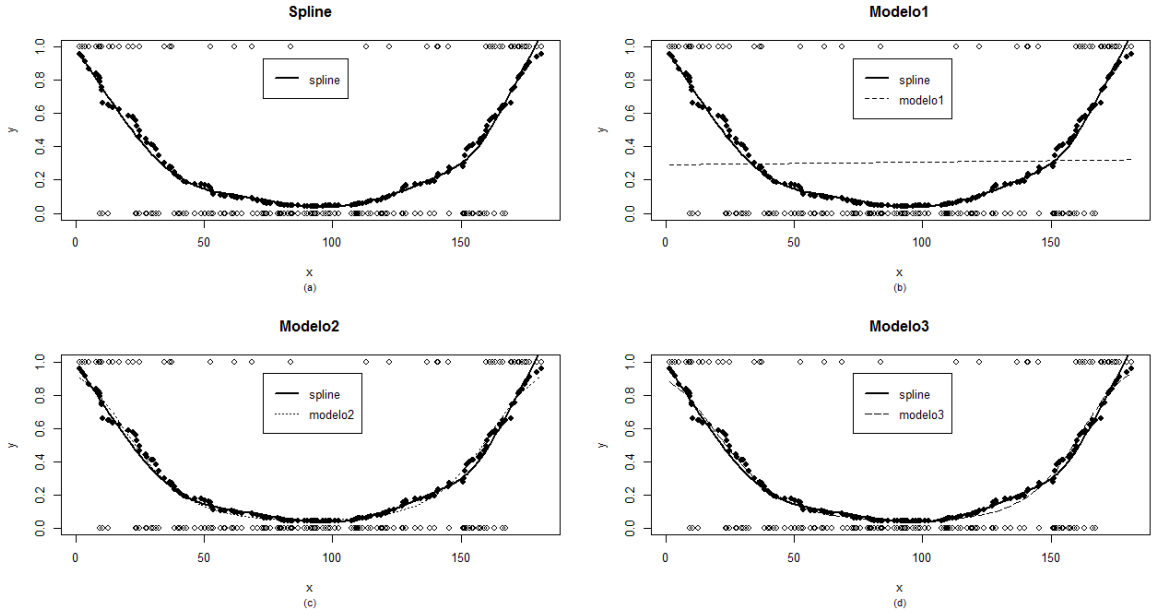

Figura 4: (*a*) Forma de la estimación obtenida por el spline cúbico; (*b*), (*c*) y (*d*) Formas de los ajustes polinomiales de primero, segundo y tercer orden para la componente sistemática de una LR.

Como se señaló antes, la gráfica no despliega una forma sigmoidea, lo cual indica que no es adecuado utilizar LR con el modelo (*3*). Sin embargo, se realiza el análisis de LR bajo este modelo para verificar su mal desempeño y demostrar que otro tipo de modelo si puede brindar un mejor ajuste. En la figura 4, gráfico (*b*) muestra el resultado de aplicar una LR bajo un modelo lineal simple y se observa que el ajuste obtenido no es adecuado. En la figura 4, gráfico (c), se muestra el ajuste de un modelo de LR cuyo componente sistemático corresponde a un polinomio de segundo grado. El gráfico demuestra que el ajuste obtenido mejora considerablemente en comparación con el ajuste anterior. En La figura 4, gráfico (*d*), representa el resultado de aplicar una modelación logística cúbica. El gráfico indica que el ajuste no mejora considerablemente en

comparación con el ajuste anterior. Esto nos da evidencia de quedarnos con el modelo con componente sistemático de segundo orden.
Con respecto al estudio que se realizó para comprobar las adecuaciones de los modelos de LR presentados hasta el momento. La prueba de la Deviance mostró que el modelo de LR simple (ver tabla 3) no hace un buen ajuste a los datos (p = 3.81e-10). Este resultado es confirmado al observar que el coeficiente de determinación R^2 resulta ser muy bajo (R^2 = 0.00017343).

Tabla 3: Ajustes polinomiales de primero, segundo y tercer orden para la componente sistemática de una LR.

Modelo / Criterio	Mod1	Mod2	Mod3
β_0	-0.84751*	2.221*	2.161*
β_1	0.00035	-1.10e-1*	-1.06e-1*
β_2		6.05e-4*	5.49e-4
β_3			2.09e-7
Dev.	3.81e-10	0.404187	0.299931
Dif. Dev.	0.9184	3.006e-13	0.9225
R^2	0.000173	0.880221	0.880377
AIC	88.11139	36.90715	38.89769

*Significativo al 5%

En lo que respecta al análisis del ajuste para el modelo de LR cuadrático encontramos que este resulta ser adecuado (p = 0.4041871). Pues como notamos la cantidad de ajuste (R^2 = 0.880221) resulta ser bueno. Y en lo concerniente al análisis del ajuste del modelo de LR cúbico, encontramos que, aunque la prueba de la Deviance sugiere que el ajuste del modelo es adecuado (p=0.299931), la prueba de la diferencia de la Deviance nos informa que el agregar un término cúbico al modelo, este resulta no ser significativo (p=0.9225). Este resultado es confirmado al observar que el AIC resulta ser más pequeño para el modelo ajustado con predictor lineal cuadrático, lo que revela que el modelo con componente sistemático cuadrático es el que mejor ajusta a los datos.

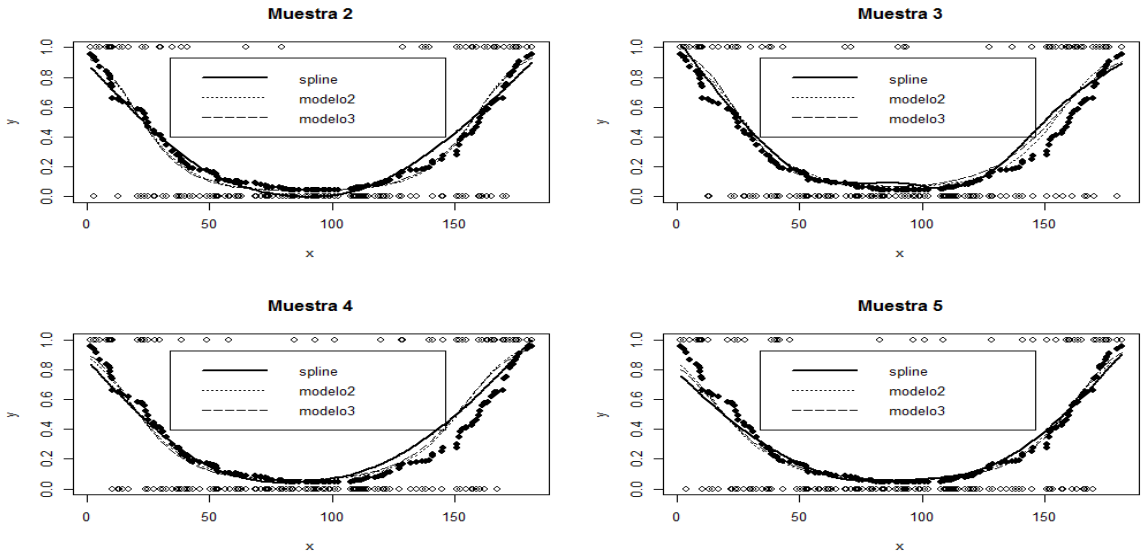

Figura 5: [•••] **F**orma de la relación entre x y π; [—]Forma de la estimación obtenida por el spline cúbico; [·····] *y* [– – –] Formas de los ajustes polinomiales de segundo y tercer orden en la componente sistemática de una LR para las muestras generadas 2,3,4 y 5.

A continuación, se presenta un análisis que se realizó con cuatro muestras que fueron generadas con las probabilidades originales y en cada muestra se realiza un suavizamiento spline cúbico y se ajusta un modelo de LR con predictor cuadrático y uno con predictor cúbico, los cuales son denotados dentro de los gráficos como modelo 2 y modelo 3, respectivamente (ver figura 5). Y se observa que el estimador spline cúbico recupera la forma distribucional que presentaron las probabilidades. Asimismo, se observa que es recomendable llevar a cabo una LR con componente sistemático de segundo orden para cada caso. Por otra

Tabla 4: Ajustes polinomiales de primero, segundo y tercer orden en la componente sistemática de una LR para las muestras 2,3,4 y 5.

parte, en lo referente a los resultados que se obtuvieron al aplicar el análisis de LR para cada muestra, se encontró (ver tabla 4), al igual que en análisis visual, que el ajuste de un modelo de LR con componente sistemático cuadrático resultó ser el mejor.

Tabla 4. Ajustes polinomiales de segundo y tercer orden en la componente sistemática de una LR para las muestras 2,3,4 y 5.

	Modelo LR	Criterio	Valor p	R^2	AIC
Muestra 2	Modelo 1	Deviance	7.4806e-13*	0.004528	98.90503
	Modelo 2	Deviance Dif. Dev	0.628489 2.2e-16*	0.929357	32.06689
	Modelo 3	Deviance Dif. Dev	0.579549 0.4652	0.936522	33.53362
Muestra 3	Modelo 1	Deviance	4.2415e-11*	0.001853	93.47135
	Modelo 2	Deviance Dif. Dev	0.410795 2.482e-14*	0.890284	37.36374
	Modelo 3	Deviance Dif. Dev	0.345275 0.5105	0.896904	38.93076
Muestra 4	Modelo 1	Deviance	3.2427e-09*	0.049956	84.97296
	Modelo 2	Deviance Dif. Dev	0.611879 1.323e-12*	0.907963	36.67884
	Modelo 3	Deviance Dif. Dev	0.499468 0.8366	0.908689	38.63629
Muestra 5	Modelo 1	Deviance	6.1612e-08*	0.014406	76.0034
	Modelo 2	Deviance Dif. Dev	0.110924 9.757e-10*	0.765038	40.63063
	Modelo 3	Deviance Dif. Dev	0.069439 0.8997	0.765357	42.61473

*Significativo al 5%

En la siguiente etapa del estudio se desarrolla el análisis referente a la tercera relación entre la variable independiente y la probabilidad "π". En la figura 6 se observa que el diagrama de dispersión en el que se graficaron las probabilidades originales junto con los valores obtenidos de la variable respuesta de la primera muestra versus la variable predictora, no presenta un patrón de comportamiento sigmoideo.

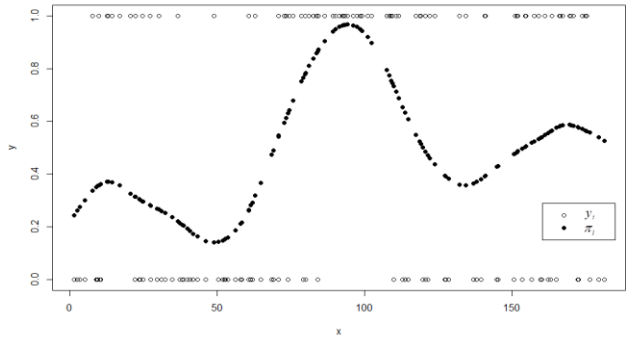

Figura 6: Forma de la relación entre x y π.

Por su parte, al efectuar un suavizamiento spline cúbico, se observa (ver figura 7, gráfico (a)) que las estimaciones que se obtuvieron por este técnica se encuentran muy cercanas a los verdaderos valores de las probabilidades de π_i. No obstante, como se señaló anteriormente, la forma distribucional de las probabilidades no es sigmoidea. Presenta una forma ondulada, en la cual no es recomendable utilizar una LR con el modelo (1) y en general, ni con ajustes polinomiales de segundo, tercero, cuarto y quinto orden. Sin embargo, se procederá a realizar un análisis de LR bajo estos modelos para verificar su mal desempeño y demostrar que otro tipo de modelo si puede brindar un mejor ajuste. En la figura 7, en los gráficos (b), (c), (d), (e) y (f) se observa que, al aplicar una LR bajo los modelos

sugeridos, los resultados obtenidos no son adecuados.

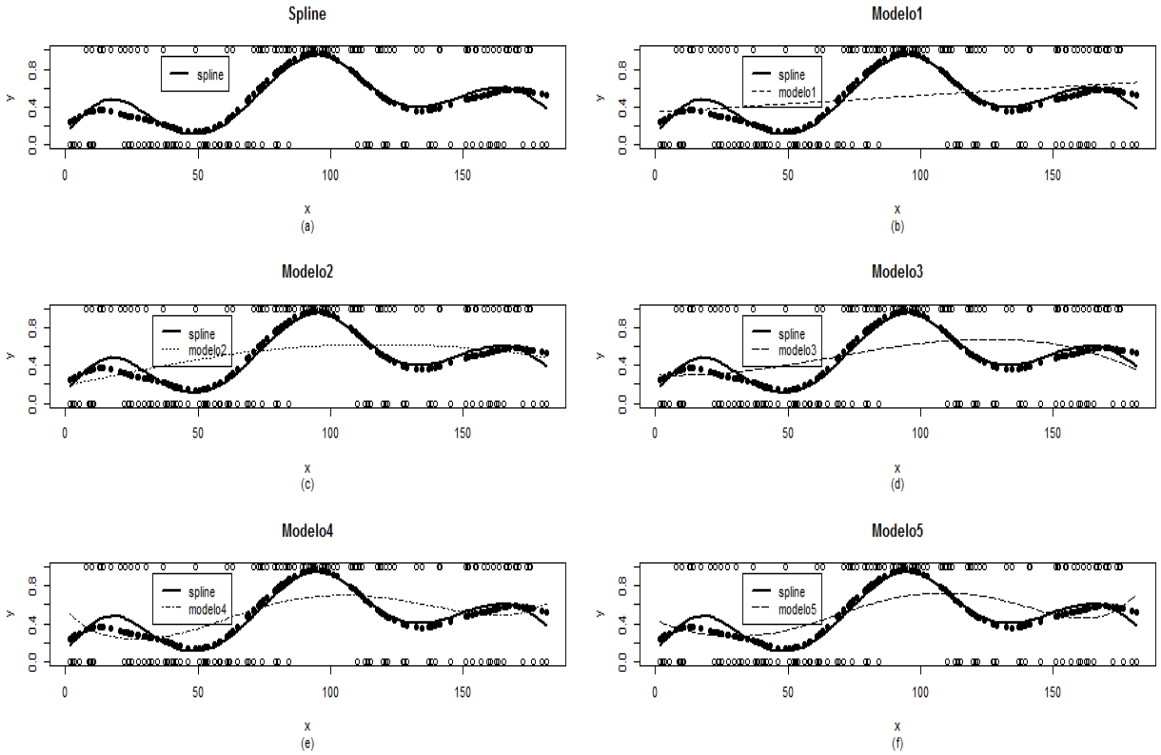

Figura 7: (*a*) Forma de la estimación obtenida por el spline cúbico; (*b*), (*c*), (*d*), (*e*) y (*f*) Formas de los ajustes polinomiales de primero, segundo, tercer, cuarto y quinto orden para la componente sistemática de una LR.

En la figura 8, grafico (*g*) se procede ajustar un modelo de LR cuyo componente sistemático corresponde a un polinomio de sexto grado. Este gráfico indica que el ajuste mejora considerablemente en comparación con los ajustes anteriores.

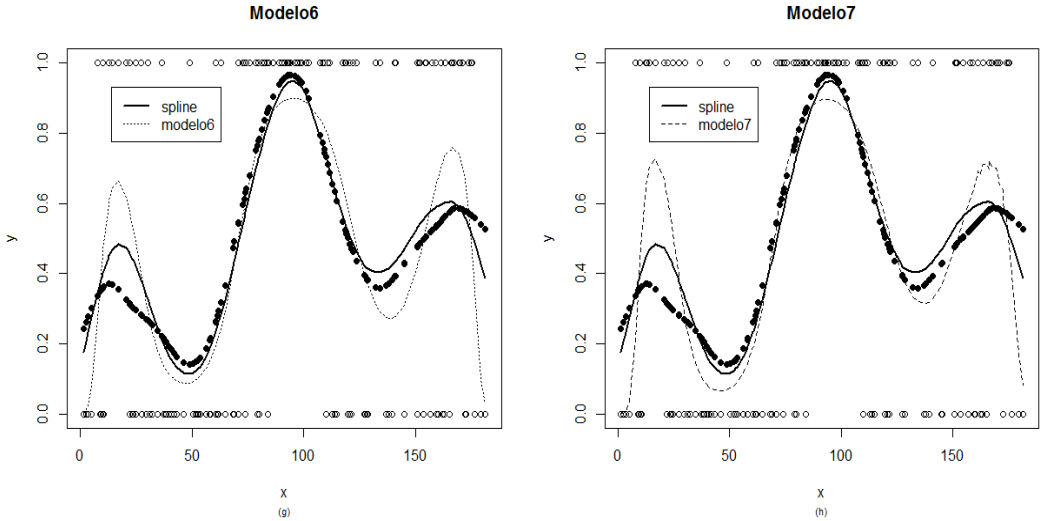

Figura 8: (*g*) y (*h*) Formas de los ajustes polinomiales de sexto y séptimo orden para la componente sistemática de una LR.

La figura 8, gráfico (*h*) muestra el resultado de aplicar una modelación logística de séptimo orden. Como se observa, el ajuste de este último modelo en comparación con el anterior no difiere mucho. Esto nos da evidencia de quedarnos con el modelo con componente sistemático de sexto orden.

En lo que respecta al estudio que se realizó para comprobar las adecuaciones de los modelos de LR, la prueba de la Deviance reveló que los modelos de LR de primer, segundo, tercero, cuarto y quinto orden (ver tabla 5) no hace un buen ajuste a los datos (p=2.25e-6, p=2.31e-6, p=1.41e-6, p=0.00013 y p=5.37e-5).

Tabla 5: Ajustes polinomiales de primero, segundo, tercero, cuarto, quinto, sexto y séptimo orden para la componente sistemática de una LR.

Modelo / Criterio	Mod1	Mod2	Mod3	Mod4	Mod5	Mod6	Mod7
β_0	-0.62442	-1.223*	-6.73e-1	1.632	1.162	-6.597*	-8.988
β_1	0.00715*	2.61e-2	-9.11e-3	-2.59e-1*	-1.89e-1	1.285*	1.811
β_2		-1.03e-4	3.79e-4	6.49e-3*	3.91e-3	-7.09e-2*	-1.04e-1
β_3			-1.77e-6	-5.39e-5*	-1.69e-5	1.56e-3*	2.47e-3
β_4				1.44e-7*	-8.24e-8	-1.59e-5*	-2.88e-5
β_5					4.96e-10	7.59e-8*	1.73e-7
β_6						-1.4e-10*	-5.12e-10
β_7							5.79e-13
Dev.	2.25e-6*	2.31e-6*	1.41e-6*	0.00013*	5.37e-5*	0.20855	0.119491
Dif. Dev.	0.02452*	0.13763	0.28914	0.00043*	0.62932	6.56e-6*	0.293204
R^2	0.110239	0.158285	0.182776	0.453068	0.458146	0.900992	0.907383
AIC	71.48927	71.28504	72.16144	61.76111	63.52813	45.21143	46.91822

*Significativo al 5%

En lo referente al análisis del ajuste del modelo de LR de sexto orden, observamos que este resulta ser adecuado (p=0.20855). Este resultado es confirmado al observar que el coeficiente de determinación R^2 resulta ser bueno (R^2 = 0.900992). Sin embargo, en lo concerniente al análisis del ajuste del modelo de LR de séptimo orden encontramos que, aunque la prueba de la Deviance sugiere que el ajuste del modelo es adecuado (p= 0.119491), la prueba de la diferencia de la Deviance nos informa que el agregar un término de séptimo grado al modelo, este resulta ser no significativo (p=0,293204). Este resultado es confirmado al observar que el AIC resulta ser más pequeño para el modelo ajustado de sexto orden, lo que indica que el modelo de LR con componente sistemático sexto orden es el que mejor ajusta a los datos.

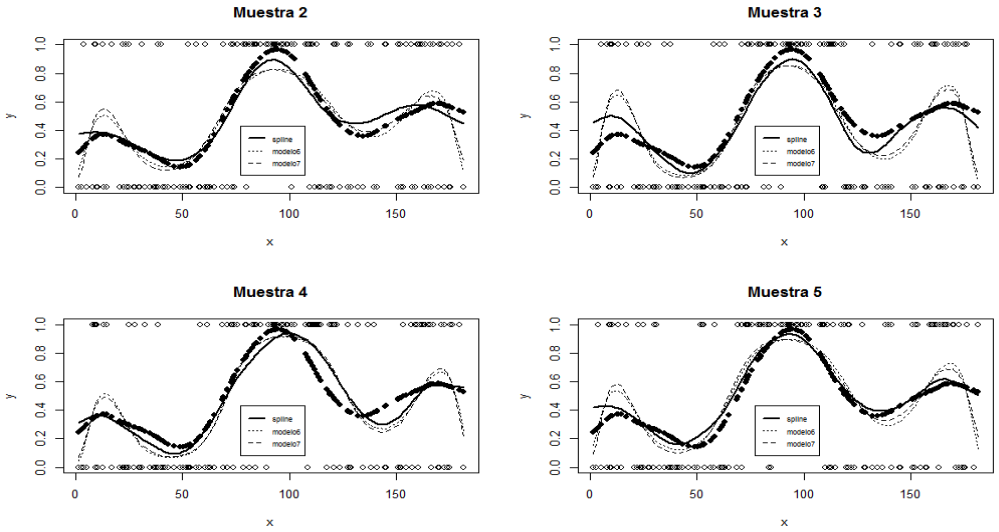

Figura 9: Gráfico de la relación entre x y π [•••]. Gráfico de la estimación obtenida por el spline cúbico [—]. Gráficos de los ajustes polinomiales de sexto y séptimo orden [·····] y [– – –] respectivamente, en la componente sistemática de una LR para las muestras generadas 2,3,4 y 5.

En la figura 9 se presenta los resultados del análisis que se llevó a cabo con cuatro muestras que se generaron a partir de las probabilidades π_i. En cada muestra generada se determinó un suavizamiento spline cúbico y se ajustaron los modelos de LR con predictor de sexto y séptimo orden, los cuales fueron denotados dentro de los gráficos como modelo 6 y modelo 7, respectivamente (ver figura 9), y se observa que al aplicar una estimación por suavizamiento spline cúbico para cada muestra generada, está sigue recuperando eficientemente la forma de la distribución que presentaron probabilidades. De la misma manera, se observa en cada una de las muestras, tanto visual como analíticamente (ver tabla 6), que un modelo de sexto orden el predictor lineal es que mejor ajusta.

Tabla 6: Ajustes polinomiales de sexto y séptimo orden en la componente sistemática de una LR para las muestras 2,3,4 y 5.

	Modelo LR	Criterio	Valor p	R^2	AIC
Muestra 2	Modelo 6	Deviance Dif. Dev	0.097101 0.000446*	0.822971	49.36753
	Modelo 7	Deviance Dif. Dev	0.049807* 0.572024	0.831918	51.04823
Muestra 3	Modelo 6	Deviance Dif. Dev	0.979322 0.000394*	0.994829	43.11944
	Modelo 7	Deviance Dif. Dev	0.959071 0.745246	0.997715	45.01387
Muestra 4	Modelo 6	Deviance Dif. Dev	0.369918 0.002639*	0.936861	44.85619
	Modelo 7	Deviance Dif. Dev	0.433599 0.224879	0.966439	45.38323
Muestra 5	Modelo 6	Deviance Dif. Dev	0.151436 0.000449*	0.874546	46.41824
	Modelo 7	Deviance Dif. Dev	0.076592 0.692552	0.878250	48.2619

*Significativo al 5%

4. UNA APLICACIÓN

La aplicación que a continuación se presenta, utiliza parte de la información contenida en una investigación realizada en el año 2011, por el Médico Internista Dr. Yojan Guevara Pérez, en el Hospital Clínico Quirúrgico Docente "Dr. Salvador Allende", en La Habana, Cuba. En esta investigación, se caracterizó la esteatosis hepática no alcohólica diagnosticada por ultrasonido en pacientes obesos. Todos los pacientes que participaron en el estudio, en total 236, presentaban un grado de esteatosis I, II o III, es decir ninguno no padecía de esteatosis. Del total de pacientes 109 eran de color de piel blanca, 68 de color de piel negra y 59 mestizos; 149 mujeres y 87 hombres. Los resultados que se presentan, no formaron parte de la investigación inicial, sino como un ejercicio posterior adicional, para ilustrar las posibilidades de la regresión logística y la regresión no paramétrica. Muchas variables fueron consideradas en esta investigación, de las cuales fueron seleccionadas para este ejercicio: una variable dependiente que toma el valor 1 si el paciente presenta esteatosis hepática grado II y 0 en otro caso, y una variable independiente que representa el Índice de Masa Corporal (IMC) del paciente medida en kg/m^2. Los hallazgos de esta investigación sugirieron una asociación entre la intensidad de la esteatosis hepática y grado de obesidad, lo cual fue comprobado por el valor p de la prueba Chi cuadrado realizada (<0.000), la cual demostró asociación entre estas dos variables.
En la figura 10, gráfico *(a)* se despliega un diagrama de dispersión donde se graficaron los valores de la variable y versus x. Podemos observar que los puntos se acumulan sobre las paralelas que representa la ausencia ($y = 0$) y la presencia ($y = 1$) de esteatosis en los pacientes. En el gráfico *(b)* se realiza un

suavizamiento spline cúbico y se observa que el posible comportamiento de las probabilidades que dieron origen a estos eventos presenta un patrón de comportamiento que no es sigmoideo. En los gráficos (*c*) y *(d)* se muestra el resultado de aplicar una LR con un modelo de primero y segundo grado y se observa que con un modelo lineal no se logra un buen ajuste. Sin embargo, con un modelo cuadrático el ajuste mejora, pero aún no lo suficiente.

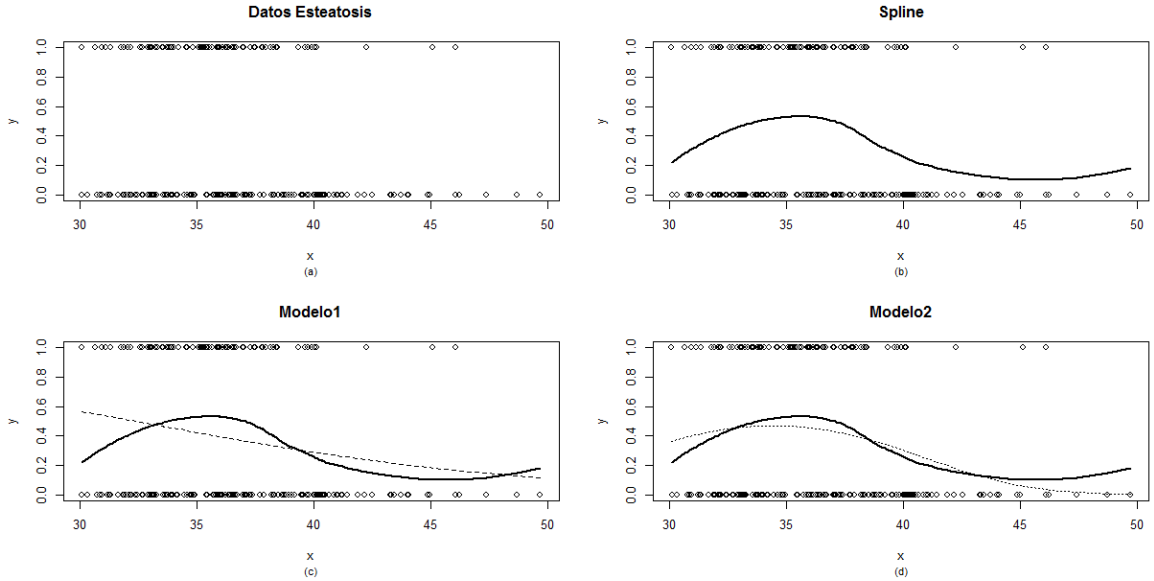

Figura 10: (*a*) Forma de la relación entre x y π; (*b*) Forma de la estimación obtenida por el spline cúbico; (*c*) y (*d*) Formas de los ajustes polinomiales de primero y segundo orden para la componente sistemática de una LR.

En la figura 11, grafico (*a*) se procede ajustar un modelo de LR cuyo componente sistemático corresponde a un polinomio de tercer grado. Este gráfico indica que el ajuste mejora considerablemente en comparación con los ajustes anteriores. En la propia figura 11, gráfico (*b*) muestra el resultado de aplicar una modelación logística de cuarto orden. Y se observa que el ajuste de este último modelo en comparación con el anterior no difiere mucho. Esto nos da evidencia de quedarnos con el modelo con componente sistemático de tercer orden, ya que con este se obtiene el mejor ajuste a los datos.

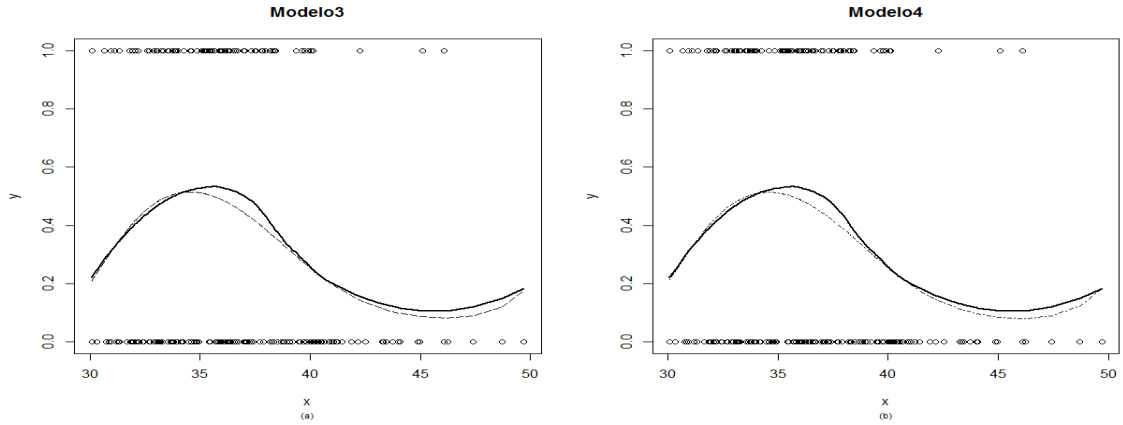

Figura 11: (*a*) y (*b*) Formas de los ajustes polinomiales de tercer y cuarto orden para la componente sistemática de una LR.

En lo que respecta al estudio que se realizó para comprobar las adecuaciones de los modelos de LR, la prueba de la Deviance reveló que el modelo de LR de primer orden (*ver tabla 7*) no hace un buen ajuste a los datos (*p=0.072289*). Sin embargo, con un modelo de segundo grado obtiene un buen ajuste (*p=0.224999)*.

Tabla 7: Ajustes polinomiales de primero, segundo, tercero y cuarto orden para la componente sistemática de una LR.

Modelo / Criterio	Mod1	Mod2	Mod3	Mod4
β_0	3.859*	-28.869	-2.556*	1.648e3
β_1	-0.119*	1.664*	1.99e1*	-1.838e2
β_2		-0.024*	-5.08e-1*	7.609
β_3			4.26e-3*	-1.39e-1
β_4				9.36e-4
β_5				
β_6				
β_7				
Dev.	*0.072289*	*0.224999*	*0.433273*	*0.555323*
Dif.Dev.	*0.00154**	*0.01674**	*0.04069**	*0.124239*
R^2	*0.322771*	*0.507079*	*0.641948*	*0.718036*
AIC	*70.39432*	*66.67031*	*64.11872*	*64.48176*

*Significativo al 5%

En lo referente al análisis del ajuste del modelo de LR de tercer orden, observamos que este resulta ser adecuado (*p=0.433273*). Este resultado es confirmado al observar que el coeficiente de determinación R^2 resulta ser bueno ($R^2 = 0.641948$) en comparación con los ajustes anteriores. Sin embargo, en lo concerniente al análisis del ajuste del modelo de LR de cuarto orden encontramos que, aunque la prueba de la Deviance sugiere que el ajuste del modelo es adecuado (*p= 0.555323*), la prueba de la diferencia de la Deviance nos informa que el agregar un término de cuarto grado al modelo, este resulta ser no significativo (*p=0,124239*). Este resultado es confirmado al observar que el AIC resulta ser más pequeño para el modelo ajustado de tercer orden, lo que indica que el modelo de LR con componente sistemático de tercer orden es el que mejor ajusta a los datos. En general, con el ajuste del modelo obtenido, en términos prácticos, se logra una interpretación coherente con respecto al contexto del problema y es la siguiente. Si el IMC de los pacientes está contenido dentro del intervalo (30,40), entonces la probabilidad de que el paciente padezca esteatosis tipo II tiende a ser mayor en comparación con los pacientes que tiene un IMC mayor a 40 y menor que 30. Desde la clínica médica este resultado es razonable, ya que pacientes con IMC por debajo de 30, es más probable que padezcan de esteatosis de grado I y por encima de 40, que padezcan de esteatosis de grado III.

5. CONCLUSIONES Y RECOMENDACIONES

En este trabajo se ha presentado un estudio en el que se muestra la contribución de la regresión no paramétrica en la determinación del modelo más adecuado a considerar en la componente sistemática para la realización de una regresión logística. Los resultados obtenidos, mostraron tanto visual como analíticamente, lo adecuado que resultó considerar un análisis de regresión no paramétrica, utilizando un suavizamiento spline cúbico, para determinar el modelo que presenta un mejor ajuste en un análisis de regresión logística. Por lo tanto, podemos concluir que la propuesta formulada, como objetivo de este trabajo, es razonable, siendo válida la aplicación de la regresión no paramétrica en lugar de la utilización de un diagrama de dispersión, con el fin de obtener información sobre el modelo que puede ser considerado en la regresión logística. La aplicación con datos reales reafirma esta última sentencia.

Agradecimientos: Los autores desean dejar constancia que, con las sugerencias y señalamientos expresados por los árbitros, ha sido posible presentar un trabajo mejor terminado y con mucha más calidad. Por ello, nuestro agradecimiento.

RECEIVED: JANUARY, 2019.
REVISED: MAY, 2019.

REFERENCIAS

[1] AGRESTI, A. (2002): **Categorical Data Analysis.** 3rd ed. Wiley. N. York.
[2] ALDERETE, A. M. (2006): Fundamentos del Análisis de Regresión Logística en la Investigación Psicológica, **Evaluar** 16, 52 – 67.
[3] CALDERON P. y ALZAMORA L. (2009): Regresión Logística Aplicada a la Epidemiología. **Revista de Salud, Sexualidad y Sociedad.** 1(4), 5-13.
[4] CHEN, L. and WANG, J. (2013): **Testing the fit of the logistic model for matched case-control studies**. Computational Statistics and Data Analysis, 57, 309- 319.
[5] EUBANK, R. L. (1999): **Nonparametric Regression and Spline Smoothing**. 2nd ed. Marcel Dekker. N. N. York.
[6] GARCÍA, M. V., ALVARADO, J. M. y JIMÉNEZ, A. (2000): La predicción del rendimiento académico: regresión lineal versus regresión logística. **Psicothema**, 12, 248 – 252.
[7] GAUTAM, V. (2012): An empirical investigation of consumers' preferences about tourism services in Indian context with special reference to state of Himachal Pradesh. **Tourism management: research, policies, practic**e, 33, 1591 – 1592.
[8] GU, C. (2002): **Smoothing Spline Anova Models.** Springer. New York.
[9] GUNVIG, A., HANSEN, F. y BORGGAARD, C. (2013): A mathematical model for predicting growth/no-growth of psychotropic C. botulinum in meat products with five variables**. Food Control**, 29, 309-317.
[10] GYÖRFI, L., KOHLER, M., KRGYŹAK, A. and WALK, H. (2002): **A Distribution Free Theory of Nonparametric Regression**. Springer. New York.
[11] KEELE, L. J. (2008): **Semiparametric regression for the social sciences**. Wiley. N. YORK.
[12] MONTEMAGNI, C., FRIERI, T., VILLARI, V. and ROCCA, P. (2012): Compulsory admissions of emergency psychiatric inpatients in Turin: the role of diagnosis. **Progress in Neuro-Psychopharmacology Biological Psychiatry.** 39, 288 – 294.
[13] NÚÑEZ J., NÚÑEZ E. y STEYERBERG E. (2011): Estrategias para la elaboración de modelos estadísticos de regresión. **Revista Española de Cardiología,** 64, 501- 507.
[14] R Core Team (2017): **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ULR http://www.R-project.org/.
[15] RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003): **Semiparametric Regression**. Cambridge University Press. N. YORK.
[16] RYAN, T. P. (1997)**: Modern Regression Methods**. 2nd ed. Wiley. N. N. York.
[17] SEBER, G. A. F. and LEE, A. J. (2003): **Linear Regression Anal**ysis. 2nd ed. Wiley. N. N. York.
[18] TAKEZAWA, K. (2006): **Introduction to Nonparametric Regression**. Wiley. N. N. York.
[19] TEICH, M., MARTY, C., GOLLUT, C., GRET-REGAMEY, A. and BEBI, P. (2012): Snow and weather conditions associated with avalanche release in forest: Rare situations with decreasing trends during the last 41 years. **Cold Regions Science and Technology**, 83 – 84, 77 – 88.
[20] SHEATHER S. (2009): **A Modern Approach to Regression with R.** Springer. New York.
[21] WRIGHT, D and LONDON, K. (2009): **Modern Regression Techniques Using R: A Practical Guide for Students and Researchers.** SAGE, Washington DC.