Modelos de regresión binaria: Aplicaciones para Cáncer Cervicouterino en una clínica de atención de la ciudad de Durango, Dgo. México

LARES-BAYONA, Edgar Felipe*†[1], NARANJO-ALBARRÁN, Lizbeth[2] y SÁNCHEZ-ANGUIANO, Luis Francisco

[1]Universidad Juárez del Estado de Durango, Instituto de Investigación Científica, Avenida Universidad y Volantín s/n Zona Centro, 34000 Durango, Dgo. México
[2]Universidad Nacional Autónoma de México, Facultad de Ciencias

Recibido Octubre 30, 2017; Aceptado Diciembre 20, 2017

Resumen

Objetivo. Aplicar y comparar modelos estadísticos de regresión binaria para Cáncer Cervicouterino (CaCu) en una clínica de atención de la ciudad de Durango, Dgo. México. Metodología. Método de investigación deductivo, inductivo y de análisis. Es un diseño de estudio descriptivo para el análisis de regresión binaria. Se analizó una base de datos sobre resultados de papanicolou de una clínica de la ciudad de Durango para los años 2011 a 2014 con un total de 4939 pacientes, de los cuales sólo 30 son positivas para CaCu. Resultados. Se estudiaron modelos de regresión binaria con funciones de enlace simétricas y asimétricas, un modelo de regresión logística para eventos raros y un modelo de regresión binaria Bayesiano semiparamétrico. Los modelos con enlaces simétricos y asimétricos muestran resultados similares en la población estudiada. El modelo logístico corregido para eventos raros obtuvo una mejor evaluación diagnóstica, estimando mayor cantidad de verdaderos positivos. El modelo Bayesiano obtuvo coeficientes más acordes al contexto para CaCu. Un análisis posterior de submuestras en un proceso iterativo de aleatorización, identificó a covariables que se pueden seguir manteniendo dentro del modelo y a otras covariables que no mantienen resultados acordes al contexto son considerados como sesgos no propios del análisis estadístico.

Cáncer Cervicouterino, Eventos Raros, Modelos de Regresión Binaria

Abstract

Objective. Apply and compare statistical models of binary regression for Cervical Cancer (CaCu) in a clinic of the city of Durango, Dgo. Mexico. Methodology. The research method is deductive, inductive and analytical. It is a descriptive study design for the binary regression analysis. We analyzed a database on papanicolou results from a clinic in the city of Durango for the years 2011 to 2014 with a total of 4939 patients, for which only 30 were positive to CaCu. Results. We studied binary regression models with symmetric and asymmetric link functions, a logistic regression model for rare events and a semi-parametric Bayesian binary regression model. The models with symmetric and asymmetric links show similar results in the studied population. The corrected logistic model for rare events obtained a better diagnostic evaluation, estimating a greater number of true positives. The Bayesian model obtained coefficients more in line with the context for CaCu. A posterior analysis of subsamples in an iterative process of randomization, identified covariables that can be maintained within the model and other covariates that do not maintain results according to the context are considered as biases not characteristic of statistical analysis.

Cervical Cancer, Rare Events, Binary regression models

Citación: LARES-BAYONA, Edgar Felipe, NARANJO-ALBARRÁN, Lizbeth y SÁNCHEZ-ANGUIANO, Luis Francisco. Modelos de regresión binaria: Aplicaciones para Cáncer Cervicouterino en una clínica de atención de la ciudad de Durango, Dgo. México. Revista de Ciencias de la Salud. 2017. 4-13: 41-54

*Correspondencia al Autor (edgarlares@ujed.mx)
† Investigador contribuyendo como primer autor.

Introducción

Mundialmente el Cáncer Cervicouterino (CaCu) o Cáncer de Cérvix ocupa los primeros lugares en morbilidad y mortalidad, reconocido como un problema de Salud Pública en los países en desarrollo. Una infección por el Virus del Papiloma Humano (VPH) de alto riesgo es un factor preponderante para desarrollar una lesión cervical, pero sólo una fracción de ellos dan como resultado a un cáncer invasor (OMS, 2013). Esto sugiere conocer de más factores adicionales que aumentan la probabilidad de desarrollar un cáncer de cérvix.

Variables de tipo sexual y reproductivos son factores que están estrechamente relacionados al CaCu invasor y lesiones precursoras. Variables como múltiples compañeros sexuales, edad temprana de la actividad sexual, no uso del preservativo, embarazos y partos, han sido consideradas como factores de riesgo para CaCu. Otros factores son la edad y tabaquismo, puesto que a mayor edad y mayor consumo del tabaco inciden en un mayor riesgo de padecer dicha enfermedad (OMS, 2013).

En el año 2015 la Secretaria de Salud de México declaró ser el primer país de la OCDE con mayor tasa de mortalidad por cáncer de cuello uterino. Desde el 2006 en México el CaCu ocupa la segunda causa de muerte por cáncer en la mujer. Para el año 2014 en México se registraron 3,063 casos nuevos de tumores malignos de cuello uterino con una tasa de incidencia del 6.08 por 100,000 mujeres mayores de 10 años (Secretaría de Salud, 2017).

La modelación estadística es muy utilizada en el ámbito profesional para determinar la relación que hay entre variables explicativas (independientes o regresoras) y la variable respuesta (dependiente). Las variables dependientes de tipo cuantitativo o cualitativo son consideradas como los efectos del estudio, mientras que las variables independientes son las causas que generan el efecto a estudiar (Wayne, 2006).

Los modelos de regresión binaria describen la relación no lineal entre el resultado de un dato binario de la variable dependiente y un valor de las variables independientes. Los modelos de respuesta binaria permiten realizar la investigación para explorar cómo cada variable independiente afecta la probabilidad de que ocurra el evento de interés de la variable dependiente entre sus dos posibles resultados (McCullagh y Nelder, 1989).

El fundamento del presente trabajo se basa en aplicar entre varios modelos de regresión binaria con funciones de enlace simétricas y asimétricas (McCullagh y Nelder, 1989 y Agresti, 2002), un modelo de regresión logístico para casos raros (King y Zeng, 2001, Imai et al., 2007) ajustado para prevalencias bajas y un modelo de regresión binaria semiparamétrico Bayesiano (Jara et al., 2006 y 2012), cuáles entre ellos mejor ajusta a problemas para CaCu en pacientes atendidos en la clínica de atención familiar del Instituto de Investigación Científica (IIC) de la Universidad Juárez del Estado de Durango (UJED) en la ciudad de Durango, Dgo. México.

1.1. Justificación

El impacto del análisis de datos categóricos es de suma importancia en el área de la salud, ya que permite describir y comprender la posible relación de las variables que implican el estado de salud de la población, como ejemplo el CaCu en mujeres de la ciudad de Durango que acuden al centro de atención familiar del IIC de la UJED en la ciudad de Durango, México.

La importancia de los modelos de regresión en contextos de CaCu es utilizado para pronosticar la probabilidad de padecer CaCu dados ciertos factores de riesgo que se cree están relacionados. Factores como el VPH, la edad en años cumplidos, el inicio temprano de vida sexual activa (IVSA), múltiples compañeros sexuales, no uso del preservativo, múltiples embarazos, partos y tabaquismo, son considerados en la literatura medica como factores de riesgo para CaCu.

ISSN 2410-3551

Es entonces que la probabilidad de padecer CaCu es un pronóstico de acertabilidad multifactorial y que se puede tener a la mano el resultado como parte del diagnóstico clínico dada una población de referencia. El impacto que genera esta investigación permitirá comprender la problemática del CaCu bajo diferentes modelos comparativos de regresión que se presentarán en este artículo. Otro impacto son las descripciones de los resultados de los modelos de regresión, que permiten conocer la probabilidad de enfermar para que de esta manera se cuente con un diagnóstico oportuno de predicción para CaCu y así hacerles frente oportunamente a los gastos futuros derivados del estado actual de salud de las pacientes.

La disminución de los gastos familiares sobre los tratamientos oportunos para CaCu son significativos en consideración cuando el grado de enfermedad es avanzada (probabilidad cercana a 1) y, por lo tanto, se tendrá un efecto de impacto positivo en la disminución de los gastos de los recursos públicos destinados para afrontar este padecimiento. La aplicación de diferentes modelos estadísticos para la regresión binaria permite obtener mayor conocimiento acerca del comportamiento de la población modelada para un problema de CaCu y su posible relación con factores de riesgo. Este conocimiento logra comprender el estado de salud de la población referente al CaCu de las pacientes que acuden a la clínica de atención familiar del IIC de la UJED en la ciudad de Durango, Dgo. México.

1.2 Problema

Los modelos de regresión binaria permiten realizar investigación para explorar cómo las variables independientes afectan la probabilidad de la variable dependiente. La principal dificultad encontrada en algunas grandes bases de datos es la detección de una gran cantidad de pacientes con diagnóstico negativo contra una pequeña cantidad de pacientes positivas, a esto se le llama "eventos raros", mencionados por King y Zeng (2001).

Aplicar modelos de regresión binaria tradicionales para relacionar la variable dependiente para casos raros con un conjunto de variables independientes resultan en estimaciones sesgadas, por lo que es necesario buscar modelos que se adecuen al comportamiento de los datos.

Por lo tanto, conocer y aplicar nuevas alternativas de regresión para datos binarios, permitirá contextualizar el problema en diferentes alternativas de análisis siendo una estrategia de selección aquella que contenga mejores argumentos estadísticos. Utilizar diferentes modelos de regresión binaria con funciones de enlace simétricas y asimétricas, así como modelos para eventos raros y modelos de regresión Bayesiana, sería de suma importancia en el área de la Salud, debido al mejor ajuste del modelo que podría discriminar mejor el contexto de la enfermedad, y que no siempre la regresión logística será el mejor modelo que ajuste adecuadamente cierto tipo de datos.

En este sentido, es importante comparar diferentes modelos de regresión usando distintos criterios de evaluación o de pruebas diagnósticas, como son: factores de inflación de la varianza, bondad de ajuste, devianza residual, y el área bajo la curva Característica Operativa del Receptor (ROC), sensibilidad y especificidad, entre otras. Estos criterios permitirían identificar cuál modelo será el más adecuado para el tipo de datos, qué en este caso, el interés es estudiar al CaCu mediante sus factores de riesgo, dado que se tiene una prevalencia observada baja en pacientes que acuden a la clínica de atención familiar del IIC de la UJED. Los diferentes modelos de regresión evaluados permitirán aportar mayor documentación científica acerca del entendimiento de este problema de salud, además de conocer cuáles modelos son los más adecuados para este problema de Salud Pública e identificar con más detalle los factores de riesgo de la enfermedad que afecta a las mujeres atendidas en la clínica de atención familiar del IIC de la UJED en la ciudad de Durango, Dgo. México.

ISSN 2410-3551

1.3 Hipótesis

El modelo de regresión logística corregido para eventos raros (Relogit) y el modelo de regresión binario Bayesiano semiparamétrico (DPbinary), son los modelos que mejor se ajustan a problemas de CaCu en una población de la clínica del IIC de la ciudad de Durango, comparados con otros modelos de regresión binaria con funciones de enlace simétricas y asimétricas.

1.4 Objetivo

Aplicar y comparar modelos estadísticos de regresión binaria para CaCu en una población de la clínica de atención familiar del IIC de la UJED en la ciudad de Durango, Dgo. México.

2. Marco teórico

2.1 Modelos lineales generalizados

Los modelos lineales generalizados (MLG) son modelos de regresión extendidos para utilizar distribuciones de respuestas no normales (McCullagh y Nelder, 1989). Autores como Nelder y Wedderburn en 1972 establecieron los tipos de MLG, tomando algunos modelos como ejemplos dentro de los tipos que ellos mismos definieron (Agresti, 2002).

2.1.1 Componentes de los MLG

Tres son los componentes de los MLG (Agresti, 2002):

- Componente aleatorio.
- Componente sistemático.
- Función liga o de enlace.

El componente aleatorio consiste de una variable respuesta Y con observaciones independientes $(Y_{1,\dots,}Y_n)$ provenientes de una distribución conocida, generalmente pertenecen a la familia exponencial (Agresti, 2002).

El componente sistemático $(\eta_{1,\dots,}\eta_n)$ está conformado por una combinación lineal del conjunto de variables explicativas. El conjunto de todas las variables predictivas en cada sujeto calculará los valores predictivos de la función lineal. Esta combinación lineal de las variables explicativas es llamada predictor lineal (Agresti, 2002). Es la función lineal de las variables explicativas, suponiendo que $X_1 = x_1, \dots, X_k = x_k$ entonces el predictor lineal asociado a esa combinación de valores de las variables explicativas es:

$$\eta(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

El terecer componente es la función liga o función de enlace que conecta la aleatoridad y el componente sistemático (Agresti, 2002). La función de enlace es una función que especifica la relación entre la esperanza condicional $E[Y \mid X_1 = x_1, \dots, X_k = x_k]$ y el predictor lineal (Agresti, 2002). En los modelos lineales, ésta relación es directa siendo:

$$E[Y|X_1 = x_1, \cdots, X_k = x_k] = \eta(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

En esta relación, la media puede tomar valores entre $(-\infty, +\infty)$ lo que no es válido para todos los MLG's, como por ejemplo cuando queremos estimar la media de una variable binaria $\{0,1\}$. Debido a esto, se considera la función de enlace g que relaciona la esperanza condicional que se quiere modelar con el predictor lineal. De manera que la función g es la función de enlace representada por la siguiente manera (Agresti, 2002):

$$g[\mu(x)] = \eta(x)$$

Como ejemplo, para los modelos con funciones de enlace logit o probit, la distribución de la variable Y es binomial y la función de enlace es la función logit o probit, respectivamente (Agresti, 2002).

ISSN 2410-3551

2.2 MLG para datos binarios

Cuando *Y* es una variable de respuesta binaria, por definición, *Y* puede significar una respuesta dicotómica, como ausencia o presencia, vivo o muerto, sano o enfermo, positivo o negativo. Las observaciones muestreadas tienen una de las dos respuestas, especificando a 0 para la ausencia y 1 para la presencia, resultado binomial de un ensayo particular. La media es $E(Y) = P(Y = 1)$. Si se denota a P(Y = 1) por *p(x)*, refleja la dependencia sobre los valores de $X = x = (x_1, \dots, x_k)$ conocida como los predictores. La varianza de Y estará dada por (Agresti, 2002):

$$Var(Y) = p(x)[1 - p(x)]$$

Que representa la varianza de un ensayo binomial (Agresti, 2002).

Cuando se utilizan modelos de regresión para variables aleatorias dicotómicas usualmente se usa la función de enlace logit, probit, cloglog ó loglog (McCullagh y Nelder, 1989), las cuales se describen a continuación.

2.2.1 Regresión logística

El modelo de regresión logística es la relación no lineal entre el resultado de un dato binario *Y* y un valor de *X*. Sea *Y* una variable respuesta binaria, donde se ha codificado como 1 a la categoría de interés y 0 para la otra categoría, y *X* un vector de variables explicativas, entonces la media $E[Y|X = x] = P[Y = 1|X = x] = p(x)$ se puede modelar mediante un modelo de regresión logística, esto es:

$$p(x) = \frac{\exp[\eta(x)]}{1+\exp[\eta(x)]} = \frac{1}{1+\exp[-\eta(x)]} \quad (1)$$

De forma similar la función de la trasformación logit es:

$$logit[p(x)] = log\left[\frac{p(x)}{1-p(x)}\right] = \eta(x) \quad (2)$$

Mientras que p estaría restringido en el rango (0,1), el logit $log\left[\frac{p}{1-p}\right] =$ puede ser cualquier número real (Agresti, 2007). Un cambio fijo en *X* seguido tiene un menor impacto cuando *p(x)* es cercana a 0 o 1 que cuando *p(x)* es cercana a 0.5 (Agresti, 2002).

Los datos binarios son los más comunes de los datos categóricos. La regresión logística es uno de los modelos más utilizados en la modelación para datos binarios.

Los usos frecuentes de los modelos de regresión logística sirven para estimar la probabilidad de que se presente el evento de interés y para evaluar la influencia que cada variable independiente tiene sobre la variable respuesta (Agresti, 2007).

2.2.2 Función de enlace probit

La función de enlace probit consiste en considerar como función de enlace la transformación de la inversa de la función de distribución acumulada de una normal estándar N(0,1):

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$$

Y la expresión del modelo sería:

$$\Phi^{-1}\big(p(x)\big) = \eta(x) \quad (3)$$

Esta transformación también acota *p(x)* entre 0 y 1. La función probit se acerca más rápidamente a probabilidades de 0 y 1 que la función logit (Agresti, 2007).

2.2.3 Función de enlace cloglog

La transformación cloglog (complemento log-log) es de la forma:

$$p(x) = 1 - exp[-\exp(\eta(x))]$$

ISSN 2410-3551

Que en forma lineal es:

$$log[-\log(1-p(x))] = \eta(x) \quad (4)$$

Esta transformación no tiene un comportamiento simétrico, sino que se aleja del valor de probabilidad 1 de forma más rápida de lo que se acerca el valor 0 (Agresti, 2007).

2.2.4 Función de enlace loglog

La transformación loglog es similar a la cloglog, y se define como:

$$p(x) = exp[-\exp(\eta(x))]$$

Que en forma lineal es:

$$log[-\log(p(x))] = \eta(x) \quad (5)$$

La transformación loglog es la inversa de la transformación cloglog, de forma que si la transformación loglog es adecuada para modelar la probabilidad condicionada a *X = x* de un suceso, la transformación cloglog es adecuada para modelar el suceso complementario (Agresti, 2007).

2.3 Modelo de regresión logística para casos raros (Relogit)

Los modelos de regresión logística para eventos raros (Relogit) son considerados en eventos como la guerra, una revolución, depresiones económicas, o una enfermedad muy poco común que también está considerada como evento raro. La ocurrencia en estos eventos es poco frecuente, pero aún así son considerados de gran importancia.

Los eventos raros son caracterizados numéricamente como una variable dependiente con una frecuencia de miles de ceros y pocos unos, y su función es explicar o predecir el evento de interés mediante variables explicativas (independientes) (King y Zeng, 2001, Imai et al., 2016).

El modelo Relogit es un modelo de regresión binaria para casos raros cuando se tiene en la variable respuesta dicotómica una gran cantidad de resultados negativos contra una cantidad pequeña de resultados positivos (King y Zeng, 2001, Imai et al., 2016).

El procedimiento estimado por el Relogit es similar al modelo estándar de la regresión logística, pero a diferencia, éste es estimado con una corrección parcial que ocurre cuando la muestra es pequeña o bien las observaciones del evento de interés son raras.

Este procedimiento también es óptimo usando una cantidad a priori como corrección para diseños de estudios de casos y controles (King y Zeng, 2001, Imai et al., 2016). El método de corrección se ajusta en el término del intercepto. Se considera un τ que debe ser la fracción verdadera del evento de interés, y $\bar{y}$ la fracción de eventos dentros de la muestra, y $\widehat{\beta_0}$ el término del intercepto sin corregir. El intercepto corregido es:

$$\beta = \widehat{\beta_0} - log\left[\frac{(1-\tau)}{\tau}\frac{\bar{y}}{(1-\bar{y})}\right] \quad (6)$$

Para los demás β's de los coeficientes sin corregir, se corrigen mediante la estimación de los sesgos para eventos raros, esto es, si $\hat{\beta}$ son los coeficientes logit sin corregir y el sesgo $(\widehat{\beta})$ es el término del sesgo, los coeficientes corregidos $\tilde{\beta}$ son:

$$\hat{\beta} - sesgo(\hat{\beta}) = \tilde{\beta} \quad (7)$$

Los términos del sesgo son:

$$sesgo(\hat{\beta}) = (X'WX)^{-1}X'W\xi$$

Donde:

$$\xi_i = 0.5Q_{ii}\left((1+w-1)\hat{\pi}_i - w_i\right)$$
$$Q = X(X'WX)^{-1}X'$$
$$w_1 = \frac{\tau}{\bar{y}}$$

ISSN 2410-3551

$w_0 = \frac{(1-\tau)}{(1-\bar{y})}$
$w_i = w_1 Y_i + w_0(1 - Y_i)$
$W = diag\{\hat{\pi}(1-\hat{\pi})w_i\}$

Donde: $Y_i \sim Bernoulli(\pi_i)$ es la variable dependiente binaria, y $\pi_i = \frac{1}{1+\exp(-x_i\beta)}$ es la función de enlace logit. Por lo tanto, lo que hace el modelo Relogit es corregir la estimación de los parámetros de regresión, a través de los sesgos del modelo con eventos raros.

2.4 Modelo de regresión binaria semiparamétrico Bayesiano (DPbinary)

El modelo de regresión binario semiparamétrico Bayesiano se estima usando la función DPbinary. La función DPbinary es una función creada en lenguaje R utilizada para modelos de regresión binaria semiparamétrico. Esta función genera una muestra de densidad posterior para un modelo de regresión binario semiparamétrico (Jara et al., 2012). Esta función genérica ajusta a un modelo semiparamétrico de regresión binaria usando un proceso Dirichlet (Muller et al., 2015).

$y_i = I(V_i \leq X_i\beta), i = 1, \dots, n$
$V_1, \dots, V_n | G \sim G$
$G | \alpha, G_0 \sim DP(\alpha G_0)$

Donde, G_0 = Logística(V |0,1) si la línea base es una logística, G_0 = N(V|0,1) si la línea base es normal, y G_0 = Cauchy(V|0,1) si la línea base es una Cauchy. Para completar las especificaciones del modelo, las siguientes distribuciones se suponen (Jara et al., 2012).

$\alpha | a_0, b_0 \sim Gama(a_0, b_0)$
$\beta | \beta_0, S_{\beta 0} \sim N(\beta_0, S_{\beta 0})$

Lo siguiente son algunos de los significados de términos: Y_i es una variable aleatoria, I es una función indicadora, V_i es un conjunto de vectores, X_i es la transformación lineal o valores predictivos lineales, α es un parámetro constante, y G_0 es una función logística (Mitra y Muller, 2015).

La precisión o parámetros totales de la masa α del Proceso Dirichlet (DP) anterior puede ser considerado como aleatoria, teniendo una distribución $Gama(a_0,b_0)$ o ajustado para un valor particular (Mitra y Muller, 2015). Los coeficientes de regresión se obtienen utilizando el algoritmo Metropolis-Hastings sobre la muestra que ajusta a la distribución condicional (Robert y Casella, 2010).

3. Material y Método

El método de investigación es deductivo, inductiva y de análisis. La tipología de estudio es de campo. Es un diseño de estudio descriptivo para el análisis de regresión binaria. Se obtuvo una base de datos sobre la consulta de Papanicolaou de la clínica del IIC de la ciudad de Durango para los años 2011 a 2014. La información recolectada a través de la base de datos representa la situación actual de las mujeres sobre los resultados para CaCu y sus factores de riesgo.

La condición de inclusión de las pacientes analizadas son que haya tenido al menos una relación sexual y que acudan a la atención de Papanicolaou por primera vez en la clínica de atención familiar del IIC de la UJED.

Se aplicó el diseño de una base de datos mediante el software SPSS (Statistics Program Science Socials), para la captura de información sobre variables extraídas del expediente clínico del IIC de la UJED. El análisis de información y el diseño de gráficos fue llevado a cabo por software R (R Development Core Team, 2008).

ISSN 2410-3551

4. Resultados

Con una cantidad de 4939 pacientes que acudieron a la clínica de atención familiar del IIC de la ciudad de Durango durante los años 2011 al 2014 se encontró una prevalencia del 0.6% de CaCu, considerado ser una prevalencia baja, véase la Tabla 1.

	Frecuencia	**Frecuencia Relativa**
Positivos	30	0.0061
Negativos	4909	0.9939
Total	4939	1.0000

Tabla 1 Prevalencia de CaCu en la clínica del IIC de la UJED
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014

De las variables explitivas utilizadas, cinco de ellas son cuantitativas discretas y las otras tres son cualitativas dicotómicas evaluados para la variable de respuesta categórica dicotómica del CaCu. Sobre la descripción estadística de las variables cuantitativas discretas independientes de la población en general, se encontraron promedios de 41.56 en años de edad, 20.27 de edad al inicio de la vida sexual, 1.95 de compañeros sexuales, 2.6 en embarazos y el 1.6 en partos, véase la Tabla 2.

	Mínimo	**Máximo**	**Media**	**Mediana**	**Desviación estándar**
Edad IVSA	12	54	20.27	19	4.0264
Compañeros Sexuales	1	24	1.95	1	1.5844
Edad	14	83	41.56	42	11.8378
Embarazos	1	14	2.65	3	1.8819
Partos	1	14	1.66	1	1.7549

Tabla 2 Estadísticos descriptivos de las variables cuantitativas discretas independientes
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014

En las variables categóricas independientes, las frecuencias relativas encontradas para la población de estudio fueron del 0.6% de VPH, el 13.8% de tabaquismo y el 87.5% del no uso del preservativo, véase la Tabla 3.

	Sí		**No**	
	Frecuencia	**Frecuencia relativa**	**Frecuencia**	**Frecuencia relativa**
Virus del Papiloma Humano	30	0.0060	4909	0.9939
Tabaquismo	683	0.1382	4256	0.8617
Preservativo de barrera (condón)	617	0.1249	4322	0.8750

Tabla 3 Estadísticos descriptivos de las variables categóricas independientes
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014

Se estimaron los modelos de regresión binaria con funciones de enlace simétricos (Logit y Probit), asimétricos (Cloglog y Loglog), modelo para casos raros (Relogit) y el modelo de regresión binaria Bayesiano semiparamétrico (DPbinary).

Los parámetros estimados con los diferentes modelos de regresión binaria tienen resultados similares. Los coeficientes acordes al contexto según su signo en el modelo Relogit fue para VPH, Edad IVSA y Embarazos. Problemas en los signos y variables regresoras no significativas son mostradas en la Tabla 4 y Tabla 5.

Coeficientes:	**Logit**	**Probit**	**Cloglog**
Intercepto	-4.9725**	-2.1349**	-5.6984***
X1= VPH	**7.4056*** **	**3.4188*** **	**6.8040*** **
X2=Edad IVSA	**-0.0055**	**-0.0113**	0.0145
X3=CompañerosSexuales	-0.2765	-0.0973	-0.2764 .
X4=Edad	-0.0104	-0.0069	-0.0078
X5=Embarazos	**0.2797**	**0.1077**	**0.2786**
X6=Partos	-0.0875	-0.0144	-0.1436
X7=Uso del Condón	-1.4171 *	-0.5425 *	-1.0171 *
X8=Tabaquismo	-1.3263 .	-0.7653 .	-0.8778
Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1			

Tabla 4 Comparación de los coeficientes entre los modelos de regresión binaria.
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014

ISSN 2410-3551

Coeficientes:	**Loglog**	**Relogit**	**Bayes**
Intercepto	-6.2250	-1.2807	-8.7111
X1= VPH	**6.1042*** **	**5.8394*** **	**8.4910*** **
X2=EdadIVSA	0.0192	**-0.0126**	**-1.7591**
X3=Compañeros Sexuales	-0.1961	-0.1511	-4.0502
X4=Edad	-0.0005	-0.0167	**3.7951**
X5=Embarazos	**0.0492**	**0.5055** **	**1.8750**
X6=Partos	-0.0569	-0.2152	-1.2061
X7=UsodelCondón	-0.3695	-1.1040 .	-1.5962 *
X8=Tabaquismo	-0.5412	-0.9229	-1.5481
Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1			

Tabla 5 Comparación de los coeficientes entre los modelos de regresión binaria
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014

De la Tabla 4a y Tabla 4b los parámetros de los coeficientes del modelo de regresión Bayesiano fue el que obtuvo una mayor cantidad de coeficientes con signos acordes al contexto, con 4 parámetros (***β***) adecuados para CaCu.

La variable regresora VPH en todos los modelos de regresión binaria fue estadísticamente significativa y con signo acorde para CaCu, en el modelo Relogit la covariable embarazos fue estadísticamente significativa y con signo acorde al contexto para CaCu.

Cinco modelos de regresión binaria fueron evaluados (Logit, Probit, Cloglog, Loglog y Relogit) a través de pruebas diagnósticas, el mejor evaluado fue el modelo Relogit, con una sensibilidad del 76.7% siendo el más alto entre los demás modelos permitiendo encontrar a 23 verdaderas positivas de entre 30 y con una especificidad del 99.3% véase la Tabla 6.

	Logit	**Probit**	**Cloglog**	**Loglog**	**Relogit**
AIC	168.12	168.08	168.35	169.45	**168.12**
BIC	226.66	226.63	226.89	228.01	___
Devianza Residual:					
Mínimo	-1.86	-1.72	-1.84	-1.55	-2.80
Cuartil 1	-0.05	-0.05	-0.05	-0.05	-0.46
Cuartil 2	-0.04	-0.04	-0.04	-0.05	-0.37
Cuartil 3	-0.03	-0.03	-0.04	-0.04	-0.29
Máximo	3.72	3.75	3.72	3.69	**2.46**
Devianza Nula gl.=4938	366.04	366.04	366.04	366.04	366.04
Devianz. residual gl.=4930	150.12	150.09	150.35	151.46	150.12
Test Residual Devianz	p=1	p=1	p=1	p=1	p=1
Hosmer Lemesh.	p=0.47	p=0.47	p=0.35	p=0.09	P=0.00
Área ROC	0.93	**0.94**	0.92	0.89	**0.93**
IC al 95% de ROC	(0.87, 0.99)	**(0.89, 0.99)**	(0.86, 0.99)	(0.82, 0.97)	**(0.87, 0.99)**
Sensibili.	0.60	0.66	0.60	0.56	**0.76**
Especific.	0.99	**0.99**	0.99	0.99	**0.99**
VPP	0.64	0.66	0.69	0.70	0.41
VPN	0.99	0.99	0.99	0.99	0.99
LR+	294.54	327.26	368.17	397.39	114.01
LR-	0.40	0.33	0.40	0.43	**0.23**
VIF:					
X1	1.68	1.56	1.51	1.21	___
X2	1.52	1.38	1.68	22.96	___
X3	1.17	1.19	1.14	44.50	___
X4	3.61	3.20	3.86	1968.01	___
X5	3.73	3.92	3.54	2094.79	___
X6	2.92	3.24	2.61	848.42	___
X7	1.27	1.24	1.36	2.20	___
X8	1.26	1.32	1.18	1.14	___

AIC= Criterio de Información Akaike. BIC= Criterio de Información Bayesiano. g.l.= Grados de libertad. IC=Intervalos de Confianza. ROC= Curva Característica Operativa del Receptor VIF= Factor de Inflación de la Varianza. VPP=Valor predictivo positivo. VPN=Valor predictivo negativo. LR+ =Cociente de verosimilitudes positivo. LR– = Cociente de verosimilitudes negativo.

Tabla 6 Bondad de ajuste y evaluación general de los modelos de regresión binaria.
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014

El gráfico 1 muestra las funciones de los diferentes modelos de regresión binaria con funciones de enlace simétricas, asimétricas, y el Relogit sobre el predictor lineal η.

ISSN 2410-3551

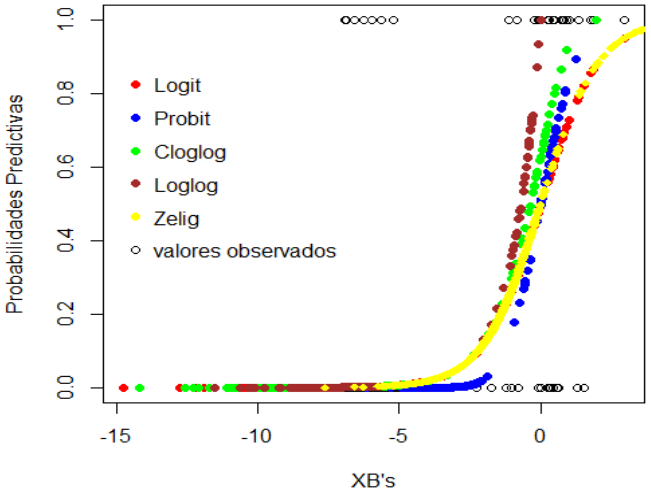

Gráfico 1 Modelos de regresión binaria con funciones de enlace Logit, Probit, Cloglog, Loglog y Relogit para CaCu
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014

4.1 Resultados igualando casos y controles

El objetivo de un análisis de un ciclo repetitivo de aleatorización de submuestras es con el fin de igualar el tamaño de muestras en los grupos de respuestas de la variable dependiente, se diseñó bajo el siguiente proceso:

1. Seleccionar aleatoriamente a 30 pacientes de las respuestas de Y=0 (controles). Se consideran todas las pacientes con Y=1 (casos).

2. Modelar el ajuste de regresión con función de enlace Logit para 60 pacientes (30 casos y 30 controles).

3. Guardar en un vector los resultados de regresión: coeficientes, probabilidades, bondades de ajuste, pruebas diagnósticas.

4. Repetir 1000 veces los pasos 1, 2 y 3, esto es, seleccionar otras 30 pacientes de las *Y=0* de forma aleatoria y compararlas con los 30 casos registrados (*Y=1*).

5. Obtener los estadísticos descriptivos (media, mediana, cuartiles, desviación estándar, percentiles, etc.) sobre las 1000 muestras aleatorias de los coeficientes, probabilidades y evaluaciones del modelo con función de enlace Logit.

Esta estrategia permite igualar el tamaño de las muestras de forma aleatoria equiparándolas en un análisis regresivo mediante la repetición del estudio en 1000 veces de forma aleatoria para la submuestras de *Y=0*. Las Tablas 7 y 8 muestra el resumen mediante una estadística descriptiva de 1000 muestras aleatorias de *Y=0* con resultados descriptivos de los coeficientes del modelo de regresión con función de enlace Logit.

	Coeficientes				
	Constante	**VPH**	**IVSA**	**NumComp. Sexuales**	**Edad**
Media	3.8426	36.0994	-0.1073	0.0506	-0.0596
Mediana	2.5178	37.6462	-0.0850	-0.0011	-0.0581
Máximo	28.5259	67.0351	0.3442	1.7984	0.0926
Mínimo	-5.3775	4.5022	-1.4060	-1.4264	-0.2714
Rango	33.9035	62.5328	1.7503	3.2247	0.3641
Desviación estándar	5.3853	8.9701	0.1632	0.4072	0.0385
Cuartil 1	0.7413	36.2213	-0.1872	-0.1845	-0.0836
Cuartil 3	5.0187	39.2910	-0.0105	0.2725	-0.0326
Percentil 98	22.6033	56.1262	0.1882	1.1036	0.0143
Percentil 99	24.5809	57.7794	0.2272	1.2431	0.0259
Percentil 99.9	27.2986	60.4798	0.3172	1.6510	0.0745
Percentil 99.99	28.4032	66.3796	0.3415	1.7836	0.0908

Tabla 7 Estadísticas descriptivas de los coeficientes del modelo de regresión binaria con función de enlace Logit.
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014.

ISSN 2410-3551

	Coeficientes			
	Gestas	**Partos**	**Condón**	**Tabaquismo**
Media	**0.5697**	-0.0061	-2.7109	-17.6780
Mediana	**0.5001**	**0.0449**	-1.7357	-17.9544
Máximo	3.0646	1.3983	0.4154	0.2325
Mínimo	-0.3424	-3.5830	-22.0545	-46.0997
Rango	3.4071	4.9813	22.4699	46.3322
Desviación estándar	0.4028	0.4786	4.1559	5.6517
Cuartil 1	0.2894	-0.2401	-2.5100	-19.1491
Cuartil 3	**0.7924**	**0.2828**	-1.2105	-16.7794
Percentil 98	1.5548	0.8684	-0.2879	-1.1096
Percentil 99	**1.8497**	**0.9437**	**0.0009**	-0.6719
Percentil 99.9	**2.4408**	**1.0853**	**0.2675**	-0.0182
Percentil 99.99	**3.0022**	**1.3670**	**0.4006**	**0.2074**

Tabla 8 Estadísticas descriptivas de los coeficientes del modelo de regresión binaria con función de enlace Logit
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014

En los resultados obtenidos por medio del modelo de regresión binario con función de enlace Logit, se encontró a cuatro coeficientes con signo acorde al contexto para CaCu por medio de la media y mediana. En el tercer cuartil de la distribución de coeficientes se identificaron a cinco con signos acordes a contexto de CaCu. Sin embargo, covariables como uso del preservativo y tabaquismo tienen una distribución de los coeficientes no acorde al contexto demostrado por los percentiles 98 y 99.9% de la distribución, véase las Tablas 6a y 6b.

Las Tablas 9 y 10 muestran las evaluaciones mediante las pruebas diagnósticas y bondades de ajuste del modelo de regresión binaria con función de enlace Logit en un proceso de aleatorización de submuestras de *Y=0*. Como resultado de las tablas 9 y 10 dentificó que el proceso de igualación de tamaños de muestra entre casos y controles permitió encontrar una mayor cantidad de pacientes con diagnóstico positivo (25) comparados con el proceso anterior del relogit.

	AIC	**BIC**	**RD**	**V+**	**S**	**E**
Media	45.46	64.31	27.46	25.41	0.84	0.97
Mediana	45.32	64.17	27.32	25	0.83	0.96
Máximo	59.96	78.81	41.96	28	0.93	1
Mínimo	35.40	54.25	17.40	23	0.76	0.86
Rango	24.56	24.56	24.56	5	0.16	0.13
Desviación estándar	3.72	3.72	3.72	1.16	0.04	0.02
Cuartil 1	43.22	62.07	25.22	25	0.83	0.96
Cuartil 3	47.30	66.15	29.30	26	0.86	1
Percentil 98	55.42	74.27	37.42	27.02	0.90	1
Percentil 99	57.04	75.89	39.04	28	0.93	1
Percentil 99.9	58.81	77.66	40.81	28	0.93	1
AIC=Criterio de información akaike. BIC=Criterio de información Bayesiano. RD=Devianza residual. V+=Verdaderos positivos. S=Sensibilidad. E=Especificidad.						

Tabla 9 Estadísticos descriptivos de los indicadores para la evaluación del modelo de regresión binaria con función de enlace Logit.
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014.

	VPP	**VPN**	**RV+**	**RV–**
Media	0.96	0.86	22.14	0.15
Mediana	0.96	0.85	25	0.16
Máximo	1	0.93	28	0.25
Mínimo	0.85	0.79	6	0.06
Rango	0.14	0.14	22	0.19
Desviación estándar	0.03	0.03	5.84	0.04
Cuartil 1	0.96	0.84	23	0.13
Cuartil 3	1	0.88	26	0.17
Percentil 98	1	0.91	27	0.24
Percentil 99	1	0.93	28	0.24
Percentil 99.9	1	0.93	28	0.25
VPP=Valor predictivo positivo. VPN=Valor predictivo negativo. RV+=Razón verosimilitud positiva. RV-=Razón verosimilitud negativa				

.

Tabla 10 Estadísticos descriptivos de los indicadores para la evaluación del modelo de regresión binaria con función de enlace Logit. *Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014.*

ISSN 2410-3551

El gráfico 2 muestra los histogramas de los coeficientes estimados para cada variable regresora del modelo de regresión binaria con función de enlace Logit en un proceso de ciclo repetitivo de aleatorización de submuestras.

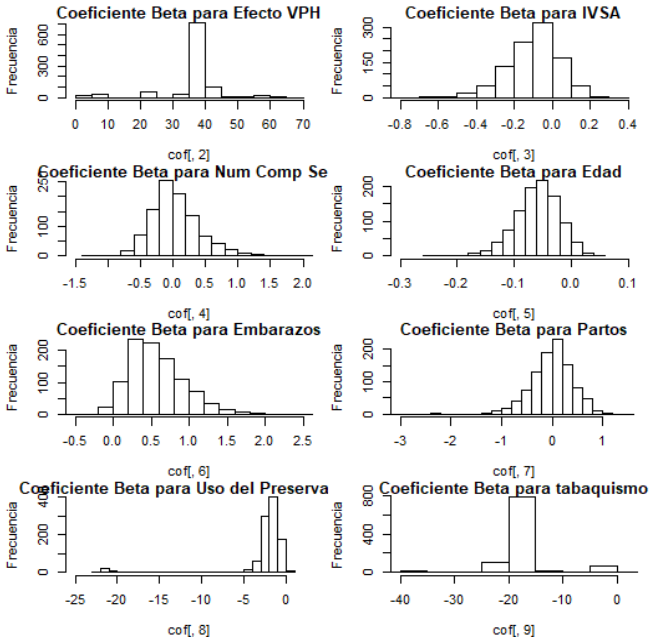

Gráfico 2 Histograma de coeficientes de cada variable regresora mediante el modelo de regresión binaria con función de enlace Logit.
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014.

Las covariables VPH y Embarazos mostrados en el gráfico 2 se identifican claramente como una relación directa entre el CaCu. El 75% de la distribución del coeficiente de IVSA mantiene congruente una relación inversa entre el CaCu. El resto de los coeficientes son descritos por el histograma como una ausencia de relación entre la variable dependiente, por lo tanto, no permiten explicar la relación de obtener una probabilidad de padecer CaCu, dado que los histogramas contienen valores de cero o en el peor de los casos, valores muy negativos considerados como "no acordes al contexto" (sesgos no propios del análisis estadístico), siendo un ejemplo las covariables uso del preservativo y tabaquismo, ésta última covariable no mantienen una relación cercana para el CaCu.

El gráfico 3 muestra los histogramas del predictor lineal ($X\beta's$) y del modelo de regresión con probabilidades predictivas ajustadas $p(X)$ en el proceso de ciclo repetitivo de aleatorización de submuestras.

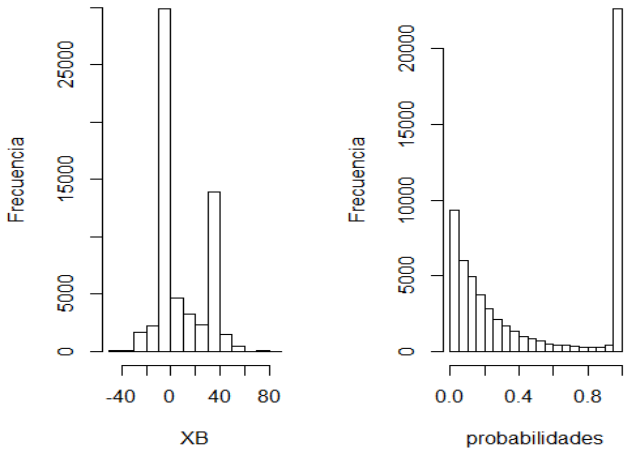

Gráfico 3 Histograma del predictor lineal y probabilidades predictivas del modelo de regresión binario con función de enlace Logit.
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014.

El modelo de regresión binario con función Logit en un proceso repetitivo de aleatorización de submuestras, determinó un modelo ajustado parsimonioso dibujando una curva logística de color morado, como el que se muestra en el gráfico 4. Este gráfico muestra la comparación de los modelos de regresión con funciones de enlace simétricas, asimétricas, función para casos raros y el modelo de regresión binario con función de enlace Logit modelando para 1000 muestras aleatorias siendo un modelo parsimonioso en un proceso repetitivo de aleatorización de submuestras.

ISSN 2410-3551

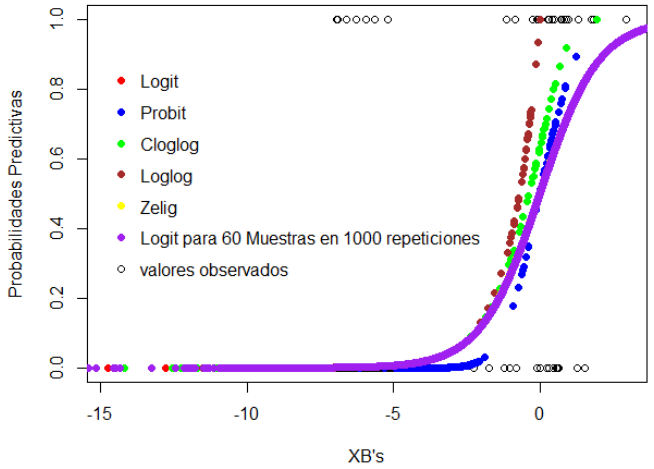

Gráfico 4 Modelos de regresión binaria con funciones de enlace Logit, Probit, Cloglog, Zelig y regresión binaria Logit en un ciclo repetitivo de aleatorización de submuestras
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014.

El gráfico 5 muestra los modelos de regresión Logit, Probit, Cloglog y Loglog para muestras de tamaño 60 en 1000 repeticiones de selección aleatoria de 30 negativas. Los modelos de regresión con funciones de enlace simétricas y asimétricas son parsimoniosos sobre la variable de respuesta para CaCu en pacientes de la clínica de atención familiar del IIC de la ciudad de Durango, durante los años 2011 a 2014.

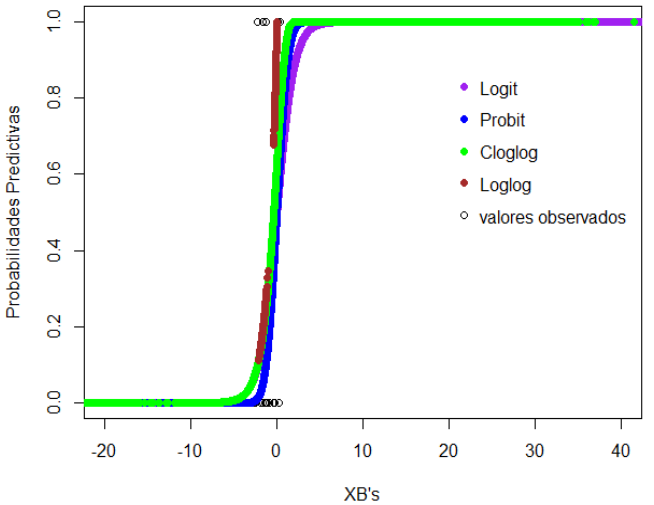

Gráfico 5 Modelos parsimoniosos de regresión binaria con funciones de enlace Logit, Probit, Cloglog y Loglog, en un proceso repetitivo de aleatorización de submuestras
Fuente: Pacientes de la clínica de atención familiar del IIC de la UJED 2011 a 2014.

ISSN 2410-3551

5. Conclusiones

En general los modelos evaluados, aunque muestran un comportamiento similar para el conjunto de datos estudiados, hay entre ellos que los hacen mejores modelos de regresión de acuerdo a las evaluaciones diagnósticas y de bondad de ajuste, como ejemplo el modelo para casos raros (Relogit) fue el mejor evaluado identificando una cantidad mayor de positivos (23 de 30) más que los modelos de regresión con funciones de enlace simétricas y asimétricas.

Bajo el proceso de aleatorización de submuestras (igualando el tamaño de muestra para casos y controles) se identificó que ciertas covariables no tienen el peso suficiente para seguir manteniendo dentro del modelo (sesgos no propios del análisis estadístico), aunque la literatura los menciona como factores de riesgo para CaCu, entonces, es necesario verificar en un futuro inmediato que existe la posibilidad de sesgos que no fueron controlados desde la clasificación de las respuestas de las covariables, hasta los sesgos de información sobre la veracidad de las preguntas recabadas en las entrevistas de las pacientes dentro del consultorio.

Agradecimientos

Este trabajo no hubiera sido posible sin el apoyo de la encargada de la clínica de atención familiar, la Dra. Nadia Velázquez Hernández quien apoyó incondicionalmente para el acceso de información. A la *Universidad Juárez del Estado de Durango* y a la *Facultad de Ciencias Exactas* por su apoyo para publicación a través del *Programa de Fortalecimiento a la Calidad Educativa* P/PFCE-2016-10MSU0010C-06.

Se agradece también a la Asociación de Investigación Pediátrica (AIP) de México por las observaciones vertidas al presente proyecto de investigación.

Referencias

Agresti, A. (2002). Categorical Data Analysis. *John Wiley & Sons*, New Jersey, segunda edición.

Agresti, A. (2007). An Introduction to Categorical Data Analysis. *Wiley - Interscience, A John Wiley & Sons*, Inc., Publication.

Imai, K., King, G., y Lau, O. (2016). Relogit: Rare events logistic regression for dichotomous dependent variables. *Technical report, Harvard*.

Jara, A., Garcia-Zattera, M.J., y Lesaffre, E. (2006). Semiparametric Bayesian Analysis of Misclassified Binary Data. *XXIII International Biometric Conference*, July 16-21, Montréal, Canada.

Jara, A., Hanson, T., Quintana, F., Mueller, P., y Rosner, G. (2012). *Package DPpackage, primera edición.*

King, G. y Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2):137-163.

McCullagh, P. y Nelder, J. A. (1989). Generalized Linear Models. Número 37 en Monographs on Statistics and Applied Probability. *Chapman & Hall, Boca Raton, Florida*, segunda edición.

Mitra, R. y Muller, P. (2015). Nonparametric Bayesian Inference in Bioestatistics. Springer.

Muller, P., Quintana, A., Jara, A., y Hanson, T. (2015). Bayesian Nonparametric Data Analysis. Springer.

OMS (2013). Prevención y control integrales del Cáncer Cervicouterino: un futuro más saludable para niñas y mujeres. Organización Mundial de la Salud. Nota de orientación de la OPS/OMS. *http://apps.who.int/iris/bitstream/10665/85344/1/9789275317471_spa.pdf*

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing,* Vienna, Austria.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25(1):11-163.

Robert, C. y Casella, G. (2010). Introducing Monte Carlo Methods with R. Springer, USA, primera edición.

Secretaria de Salud (2017). Cáncer de la mujer: Información estadística. Technical report, *Centro Nacional de Equidad de Género y Salud Reproductiva (CNEGSR)*, S.S. México.

Wayne, D. (2006). Bioestadística Base para el Análisis de las Ciencias de la Salud. *Limusa Wiley.*

ISSN 2410-3551