Regresión Logística Ordinal Aplicada a la Identificación de Factores de Riesgo para Cáncer de Cuello Uterino

Ordinal Logistic Regression Applied to the Identification of Risk Factors for Cervical Cancer

Evaristo Navarro Manotas*
Aníbal Verbel Castellar**
Delia Robles García***
Kennedy Hurtado Ibarra****

RESUMEN

La identificación de factores de riesgo para cáncer de cuello uterino es determinante a la hora de establecer diagnósticos efectivos que, en un momento dado, pueden ser determinantes para salvar vidas. Desde esta perspectiva se realizó este estudio en una muestra constituida por 105 pacientes, que circunscribió a todas las mujeres que concurrieron a la consulta ginecológica en una campaña desarrollada por la Secretaría de Salud del Departamento del Atlántico. Se utilizaron dos instrumentos para la recolección de los datos: aquellos vinculados a cáncer de cuello uterino se registraron en un formulario elaborado para tales efectos. En el estudio fue considerada como variable dependiente el Cáncer de cuello uterino (CCU) y como variables independientes los factores relacionados con la paridad (Edad (ED), Número de Hijos Nacidos Vivos (NHV), Número de Hijos Nacidos Muertos (NHM), tipo de parto (TP) y tipo de embarazo (TE)). Finalmente, también se incluyeron las características de la conducta sexual [Enfermedades venéreas (EV): sífilis, herpes, gonorrea u otras]. De manera general se observa que el riesgo de tener cáncer de cuello uterino es mayor cuando aumenta el número de hijos en partos por cesárea y se ha perdido un hijo.

Palabras Claves: cáncer, cuello uterino, riesgos, regresión logística, variables aleatorias, muestra, parámetros.

ABSTRACT

Identifying risk factors for cervical cancer is crucial to effectively conduct diagnostics which, in a given time, can be a key issue to save lives. From this perspective, this study was performed on a sample of 105 patients, which circumscribed to all women who went to the gynecologist in a campaign developed by the Health Secretary of the department of Atlántico, Colombia. Two instruments were used for data collection: The data linked to cervical cancer were recorded on a form developed for this purpose. In this study, the Cancer of the cervix (CCU) was considered as the dependent variable and factors related to birth [(Age (ed), number of live born children (OVC), Number of Children Born Dead (NHM), birth rate (tp) and type of pregnancy (I))] were regarded as independent variables. Finally, the characteristics of sexual behavior (venereal disease (VD): syphilis, herpes, gonorrhea or other) were also treated. In a general way, it is observed that the risk of cervical cancer is greater when the number of children increases in cesarean deliveries and there exist the loss of a child.

Key Words: cancer, cervix, risk, logistic regression, random variables, sample, parameters.

****Universidad de la Costa CUC. Docente Tiempo Completo. Correo eletrónico: enavarro3@cuc.edu.co.***
*****Universidad Libre de Colombia. Docente Tiempo Completo. Correo eletrónico: averbel@unilibrebaq.edu.co***
******I. E. Francisco José de Caldas. Docente Tiempo Completo. Correo eletrónico: derobles72@hotmail.com***
******* Universidad de la Costa CUC. Docente Tiempo Completo. Correo eletrónico: Khurtado1@cuc.edu.co***

Fecha de recepción: 11 de Julio del 2014 • Fecha de aceptación: 25 de Agosto del 2014

INTRODUCCIÓN

En Colombia, el cáncer de cuello uterino ha sido considerado como el problema de salud reproductiva femenina más importante. A su vez, se ha determinado que ésta, aunque es una enfermedad prevenible, presenta altas tasas de significación en la salud pública a nivel mundial en las tres últimas décadas. Es así como, de acuerdo a los estudios realizados por la Agencia Internacional para la Investigación del Cáncer (IARC), se estimó que en el año 2003 se produjeron 5.000.000 de nuevos casos de cáncer en mujeres, de los cuales 500.000 corresponden al de cuello uterino; en este estudio resulta preocupante el hecho de que el 80% de estas neoplasias ocurrieron en países en vías de desarrollo. Actualmente, el cáncer, junto con las enfermedades cardiovasculares y la violencia, son las principales causas de mortalidad en Colombia donde, en particular el de cuello uterino, presenta una tasa de incidencia de 35/100.000 mujeres por año, constituyéndose en una de las más altas del mundo [1].

Muy a pesar que esta enfermedad, antes de presentar manifestaciones clínicas, puede detectarse con técnicas sencillas (dentro de las cuales se encuentran aceptadas la citología cervicouterina, la colposcopia y el estudio histopatológico de muestras obtenidas por biopsia), poco invasivas y de bajo costo es poca la población femenina que toma estas medidas preventivas. Esta situación muestra estudios que revelan que en el área urbana de Colombia la cobertura de la población, con estas pruebas, se encuentra sólo alrededor del 69%, a pesar del descenso demostrado sobre la mortalidad y la morbilidad en otros países, no obstante, hasta la fecha, no se han encontrado en el país, estudios que permitan conocer la asociación entre los hallazgos de los exámenes de tamizado y la presencia de esta patología. Por esta razón, se planteó como objetivo la identificación de factores de riesgo para cáncer de cuello uterino y la prevalencia de algunas alteraciones patológicas en la citología, colposcopia y biopsia, y establecer su asociación con el diagnóstico final de carcinoma escamo-celular invasor de cuello uterino, para así determinar cuáles de estas alteraciones están asociados en forma positiva y negativa al diagnóstico de cáncer, para así mejorar el rendimiento de estas pruebas en el diagnóstico con la implementación de dichos indicadores [1].

1. MODELOS LOGÍSTICOS

1.1. Introducción

El Modelo de Regresión Logística ha sido utilizado por muchos años, pero no fue hasta que los autores [2] aplicaron el Modelo de Regresión Logística utilizando los datos de Framingham, el cual trata de un estudio del corazón, donde se pudo apreciar el poder y la aplicación de estos modelos. Entre los pocos textos que incluyen temas sobre regresión logística se encuentran los mencionados en los libros de [3-6]. En cada uno de estos textos, el tema central no es regresión logística. Muchas de las técnicas para aplicar el método e interpretación de los resultados pueden ser solamente los encontrados en la literatura estadística, lo que está fuera de la comprensión de muchos usuarios potenciales [7].

Este modelo es una generalización del modelo de regresión lineal clásico para variables dependientes categóricas dicotómicas [8]. Tiene la ventaja de no requerir supuestos como el de normalidad

multivariable y el de homocedasticidad (igualdad de las varianzas), que son difíciles de verificar [9]. Además, es más potente que el análisis discriminante cuando estos supuestos no se cumplen. Otra ventaja radica en su similitud con la regresión múltiple: permite el uso de variables independientes continuas y categóricas (estas últimas por medio de su codificación a variables ficticias), cuenta con contrastes estadísticos directos, tiene capacidad de incorporar efectos no lineales y es útil para realizar diagnósticos [10]. Tiene una amplia aplicación en estudios observacionales, de encuesta y experimentales, así también como en estudios epidemiológicos [6,8,10-11]. Numerosas investigaciones muestran las ventajas de utilizar el análisis de regresión logística en la evaluación del Funcionamiento Diferencial del Ítem (DIF), especialmente en la detección de DIF cuando es no uniforme y mixto y cuando no se cuenta con muestras grandes [12-14].

1.2 Modelos logísticos binarios

En el estudio se aplicaron los modelos logísticos binarios, los cuales se pasan a definir. Sea Y una variable dependiente binaria, que toma dos valores posibles etiquetados como 0 y 1[15].

Sean $X_1, \dots, X_K$ un conjunto de variables independientes observadas con el fin de explicar y/o predecir el valor de Y.

El objetivo es determinar $P\left[Y = 1 \mid X_1, \dots, X_K\right]$, por lo tanto $P\left[Y = 0 \mid X_1, \dots, X_K\right] = 1 - P\left[Y = 1 \mid X_1, \dots, X_K\right]$

Se construye un modelo de la forma: $P\left[Y = 1 \mid X_1, \dots, X_K\right] = p(X_1, \dots, X_K; \beta)$ (1.1)

donde $p(X_1, \dots, X_K; \beta) \in$ [0,1], es una función que recibe el nombre de función de enlace (función de probabilidad) cuyo valor depende de un vector de parámetros $\beta = (\beta_0, \beta_1, \dots, \beta_K)'$.

Con el fin de estimar β y analizar el comportamiento del modelo considerado, observamos una muestra aleatoria simple de tamaño *n* dada por $\{(x_i', y_i); i = 1, \dots, n\}$ donde $x_i = (1, x_{i1}, \dots, x_{iK})'$ es el valor de las variables independientes e $y_i = \{0,1\}$ es el valor observado de Y en el *i-ésimo* elemento de la muestra.

$$Y \mid (X_1, \dots, X_K) \sim Bernoulli\left(1, p\left(Y = 1 \mid X_1, \dots, X_K; \beta\right)\right)$$

Utilizando el hecho de que la variable dependiente toma sólo dos resultados (*éxito* y *fracaso*), cuando el número de éxitos en *n* repeticiones tiene una distribución binomial $B(n, p)$,

la función de verosimilitud es:

$$L\left(\beta \mid (x_1', y_1), \dots, (x_n', y_n)\right) = \prod_{i=1}^{n} p_i^{yi}(1 - p_i)^{1-y_i}, \qquad (1.2)$$

donde $p_i = p_i(x_i'; \beta) = p(x_{i1}, \dots, x_{iK}; \beta);\ i = 1, \dots, n$

1.2.1. Modelo de Regresión Logística Binaria

Sea

$$p(X_1, \dots, X_k; \beta) = G(\beta_1 X_1, \dots, \beta_k X_k), \quad (1.3)$$

donde $G(x) = \frac{e^x}{1+e^x}$

es la función de densidades acumuladas que es la función logística. El modelo normalmente conocido es:

$$\ln\left(\frac{p(x_1,\dots,x_K;\beta)}{1-p(x_1,\dots,x_K;\beta)}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_K x_K \quad (1.4)$$

Llamado *modelo logit*. Cuando la variable cualitativa toma el valor 1, la expresión:

$$\frac{P(y=1 \mid x_1, \dots, x_K)}{P(y=0 \mid x_1, \dots, x_K)} = \frac{p(x_1, \dots, x_K; \beta)}{1 - p(x_1, \dots, x_K; \beta)},$$

se conoce con el nombre de odds, que es un factor de riesgo en el mundo de la medicina, donde la variable Y indica habitualmente la presencia de una determinada enfermedad, objeto de estudio y en ausencia toma el valor 0.

1.2.2 Función de Verosimilitud

Teniendo en cuenta la forma matricial de:

$$p(x_i'; \beta) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}},$$

según la función de verosimilitud viene dada por:

$$L(\beta|(x_1', y_1), \dots, (x_n', y_n)) = \prod_{i=1}^{n} \left[\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right]^{y_i} \left[\frac{1}{1+e^{x_i'\beta}}\right]^{1-y_i} \quad (1.5)$$

1.3. Modelos logísticos polinómicos

Cuando la variable respuesta tiene más de 2 categorías, el modelo logístico se denomina polinómico. En la implementación del modelo logístico polinómico se debe seleccionar un nivel de la variable respuesta, como el nivel de referencia con el cual se comparan los demás niveles. Si se cambia el nivel de referencia escogido no se altera el modelo pero si cambian las interpretaciones de los parámetros.

Para el modelo logístico polinómico con variable de respuesta $Y = \{0,1,2\}$, K variables explicativas representadas por $X = X_1, \dots, X_K$, y 0 como categoría de referencia, se tienen los modelos logísticos

$$\ln\left(\frac{P(Y=1\mid X)}{P(Y=0\mid X)}\right)=\beta_1+\beta_{11}X_1+\cdots+\beta_{1K}X_K=\beta_1+\sum_{i=1}^{K}\beta_{1i}X_i=Z_1 \quad (1.6)$$

y

$$\ln\left(\frac{P(Y=2\mid X)}{P(Y=0\mid X)}\right)=\beta_2+\beta_{21}X_1+\cdots+\beta_{2K}X_K=\beta_2+\sum_{i=1}^{K}\beta_{2i}X_i=Z_2 \ (1.7),$$

1.4. Estimación de parámetros logísticos

Se considera una variable aleatoria dependiente Y categórica nominal politómica con valores soporte (Y) = {1, 2,3} y con probabilidades $p_1=P(Y=1)$, $p_2=P(Y=2)$ y $p_3=P(Y=3)=1-p_1-p_2$ Suponga que se quiere analizar el efecto que ejercen dos variables explicativas continuas X_1, X_2 sobre las probabilidades p_1 y p_2 que caracterizan a la variable Y. Se puede redefinir a la variable Y mediante un vector (Y_1,Y_2) de variables dummy o ficticias construido de la siguiente forma:

$$(Y_1,Y_2)=\begin{cases}(1,0)\ si\ Y=1\\(0,1)\ si\ Y=2\\(0,0)\ si\ Y=3\end{cases}$$

Las variables $\boldsymbol{Y_1}$ e $\boldsymbol{Y_2}$ tienen una distribución de Bernouilli con E($\boldsymbol{Y_1}$)=$\boldsymbol{p_1}$ y E($\boldsymbol{Y_2}$)=$\boldsymbol{p_2}$, al igual que la variable dependiente en una regresión logística binaria clásica. Obviamente estas dos variables no son independientes ya que Cov($\boldsymbol{Y_1,Y_2}$)=-$\boldsymbol{p_1\,p_2}$ [16].

En consecuencia se verifica que

$$\begin{cases}\mathrm{P(Y=1|X_1,X_2)}=p_1=E(Y_1)=\dfrac{\exp(\mathrm{Z_1})}{1+\exp(\mathrm{Z_1})+\exp(\mathrm{Z_2})}\\ \mathrm{P(Y=2|X_1,X_2)}=p_2=E(Y_2)=\dfrac{\exp(\mathrm{Z_2})}{1+\exp(\mathrm{Z_2})+\exp(\mathrm{Z_1})},\end{cases}$$

donde $\boldsymbol{Z_1=\beta_1+\beta_{11}X_1+\beta_{12}X_2}$ $\boldsymbol{y}$ $\boldsymbol{Z_2=\beta_2+\beta_{21}X_1+\beta_{22}X_2}$, siendo $\boldsymbol{\beta_1,\beta_{11},\beta_{12},\beta_2,\beta_{21},\beta_{22}}$, parámetros que se desean estimar.

Además,

$$\mathrm{P(Y=3|X_1,X_2)}=p_3=1-p_1-p_2=\frac{1}{1+exp(Z_1)+exp(Z_2)} \quad (1.8)$$

Dada una muestra de datos $(Y_{1i}, Y_{2i}, X_{1i}, X_{2i})$ con i=1,2.... , n se puede definir, en función de los parámetros del modelo, las funciones $Z_{1i}, Z_{2i}, p_{1i}, p_{2i}$ y abordar el problema de la estimación de los mismos mediante el método de máxima verosimilitud, como se muestra a continuación.

Con el modelo planteado, la función de verosimilitud de la muestra viene dada por la siguiente expresión:

$$L = \prod_{1=1}^{n}\left(p_{1i}^{Y_{1i}} \cdot p_{2i}^{Y_{2i}} \cdot p_{3i}^{1-Y_{1i}-Y_{2i}}\right) = \prod_{1=1}^{n}\left(\left(\frac{p_{1i}}{p_{3i}}\right)^{Y_{1i}} \cdot \left(\frac{p_{2i}}{p_{3i}}\right)^{Y_{2i}} \cdot p_{3i}\right) \quad (1.9)$$

En vez de trabajar con esta expresión se utiliza la función auxiliar:

$$\Lambda = -2 \cdot \ln(L) = -2 \cdot \sum_{i=1}^{n}\left(Y_{1i} \cdot ln\left(\frac{p_{1i}}{p_{3i}}\right) + Y_{2i} \cdot ln\left(\frac{p_{2i}}{p_{3i}}\right) + ln(p_{3i})\right)$$

$$= 2 \cdot \sum_{i=1}^{n}\left(ln\left(1 + exp(Z_{1i}) + exp(Z_{2i})\right) - Y_{1i} \cdot Z_{1i} - Y_{2i} \cdot Z_{2i}\right) \quad (1.10)$$

El problema de maximizar la verosimilitud equivale al de minimizar la función auxiliar Λ y puede resolverse por métodos numéricos, de forma iterativa partiendo de la estimación inicial $\beta_{11}=\beta_{21}=\beta_{12}=\beta_{22}=0$, $\beta_1=\ln(n_1)-\ln(n-n_1-n_2)$ y $\beta_2=\ln(n_2)-\ln(n-n_1-n_2)$ siendo n_1 y n_2 el número de observaciones en las categorías 1 y 2, respectivamente. Estos estimadores iniciales se obtienen suponiendo que no hay una influencia de las variables regresoras en el modelo planteado y para ellos el valor inicial de la función auxiliar que se debe minimizar es:

$$\Lambda_0 = -2 \cdot \left(n_1 \cdot ln\left(\frac{n_1}{n}\right) + n_2 \cdot ln\left(\frac{n_2}{n}\right) + (n - n_1 - n_2) \cdot ln\left(\frac{n-n_1-n_2}{n}\right)\right) \quad (1.11)$$

Una vez alcanzada la convergencia del método iterativo, se designará por Λ_1 al mínimo obtenido y por $\hat{\beta}_1, \hat{\beta}_{11}, \hat{\beta}_{21}, \hat{\beta}_2, \hat{\beta}_{12}, \hat{\beta}_{22}$ a los valores estimados de los parámetros del modelo.[16]

1.5. Intervalos de confianza y pruebas de hipótesis

Basados en la normalidad asintótica de los estimadores máximos verosímiles se puede construir, utilizando la distribución normal, intervalos de confianza asintóticos para cada uno de los parámetros del modelo y, mediante las transformaciones correspondientes, intervalos de confianza (I.C.) para las OR para el parámetro β_{11}, y utilizando un grado de confianza de 1–α, se tendría:

$$IC(\beta_{11})_{1-\alpha} = \hat{\beta}_{11} \pm z_{\frac{\alpha}{2}} \cdot s.e.\left(\hat{\beta}_{11}\right)$$

$$IC(OR_1(X_1))_{1-\alpha} = \exp(\hat{\beta}_{11} \pm z_{\frac{\alpha}{2}} \cdot s.e.\left(\hat{\beta}_{11}\right)),$$

Siendo $z_{\alpha/2}$ el valor que, en una distribución normal (0,1), verifica $P(Z>z)=\alpha/2$.

Se puede contrastar la hipótesis de no existencia de un efecto significativo global de las variables regresoras, teniendo en cuenta que la diferencia entre el valor inicial y el valor final de la función auxiliar Λ tiene una distribución χ^2 con 4 grados de libertad (en general, número de regresores

multiplicado por número de categorías menos una). El p-valor del test para la hipótesis nula de que no existe efecto de las variables regresoras ($\beta_{11}=\beta_{21}=\beta_{12}=\beta_{22}=0$) vendrá dado por $P(\chi^2_{(4)} > \Lambda_o - \Lambda_f$. [16].

1.5.1. Significado del efecto de cada variable regresora

Si se llama al mínimo de la función auxiliar, que se obtendría eliminando del modelo la variable X_1 ($\beta_{11}=\beta_{12}=0$) se verifica que la diferencia entre los mínimos de la función auxiliar, en el modelo reducido y en el modelo completo, tiene una distribución con 2 grados de libertad (en general, número de regresores menos uno multiplicado por número de categorías menos una). Por tanto, el p-valor del test para la hipótesis nula de que no existe efecto de la variable X_1($\beta_{11}=\beta_{12}=0$) vendrá dado por $P(\chi^2_{(2)} > \Lambda_{-1} - \Lambda_1)$. De modo similar se podría calcular Λ_{-0} (mínimo de la función auxiliar eliminando β_1 y β_2 del modelo) y Λ_{-2} (mínimo de la función auxiliar eliminando del modelo la variable X_2) y construir pruebas de hipótesis para $\beta_1=\beta_2=0$ y $\beta_{21}=\beta_{22}=0$, respectivamente. (V. Pando Fernández y R. San Martín Fernández 2004).

1.5.2. Significado de cada parámetro

Teniendo en cuenta que el cuadrado de cada estimador, dividido por su error estándar, tiene una distribución χ^2 con 1 grado de libertad, se puede construir test de hipótesis para la igualdad de cada parámetro a cero y se puede saber qué estimadores de los parámetros del modelo son significativamente distintos de cero. Por ejemplo, para el test de hipótesis $\beta_{11}=0$ el p-valor sería:

$$p\left(\chi_1^2 > \left(\frac{\hat{\beta}_{11}}{s.e.(\hat{\beta}_{11})}\right)^2\right), \qquad (1.12)$$

Siendo $s.e.\left(\hat{\beta}_{11}\right)$ el valor correspondiente al error estándar del estimador del parámetro β_{11}. [16]

1.6. Riesgos relativos y razones odds en modelos logísticos

Con el propósito de utilizar las estimaciones de los parámetros que aparecen en el modelo, se presentan a continuación las siguientes medidas estadísticas que juegan un papel importante en el análisis de la información suministrada por el modelo.

1. El **Riesgo** de que la variable respuesta Y tome el valor de g, dado que X_1 toma el valor r y X_2 el valor s, está dado por:

$$p_{g\,|\,\mathrm{rs}} = P(Y = g \,|\, \mathrm{X}_1 = \mathrm{r},\ \mathrm{X}_2 = \mathrm{s})$$

2. El **Riesgo Relativo** (RR) de que Y tome el valor g cuando X_1 toma el valor r comparado cuando toma el valor r ‘ y en presencia del valor s de X_2 , está dado por:

$$RR_{g(r,r')(s)} = \frac{p_{g\,|\,\mathrm{rs}}}{p_{g\,|\,r'\mathrm{s}}}$$

3. **Odds** de Y = g vs. G dados X_1 = r y X_2 = S está dado por:

$$O_{[g,g\prime]rs} = \frac{p_{g\,|\,\mathrm{rs}}}{p_{g\prime\,|\,\mathrm{rs}}}$$

4. La **Razón Odds** (OR) de Y = g vs. g' comparado a X_1 = r vs. r' y en presencia dl valor s de X_2 está dada por:

$$OR_{[g,g']\,(r,r')(s)} = \frac{O_{[g,g']rs}}{O_{[g,g']r's}} = \frac{\dfrac{p_{g\,|\,\mathrm{rs}}}{p_{g'\,|\,\mathrm{rs}}}}{\dfrac{p_{g\,|\,\mathrm{r's}}}{p_{g'\,|\,\mathrm{r's}}}} = \frac{p_{g\,|\,\mathrm{rs}}p_{g'\,|\,\mathrm{r's}}}{p_{g'\,|\,\mathrm{rs}}p_{g\,|\,\mathrm{r's}}} = \frac{RR_{g(r.r')(s)}}{RR_{g'(r.r')(s)}}$$

La variable X_1 en los **RR** y las **OR** se denomina variable de exposición.

En el modelo logístico

$$ln\left(\frac{p_{g\,|\,\mathrm{rs}}}{p_{g'\,|\,\mathrm{rs}}}\right) = \beta_g + \beta_{g1}X_1 + \beta_{g2}X_2 \qquad (1.13)$$

se tiene:

$$\frac{p_{g\,|\,\mathrm{rs}}}{p_{g'\,|\,\mathrm{rs}}} = \exp(\beta_g + \beta_{g1}X_1 + \beta_{g2}X_2)$$

Luego, si X_1 es dicotómica, es decir, X_1 = (0; 1); entonces:

$$OR_{[g,g']\,(1,0)(s)} = \frac{\dfrac{p_{g\,|\,1s}}{p_{g'\,|\,1s}}}{\dfrac{p_{g\,|\,0s}}{p_{g'\,|\,0s}}} = \frac{\exp(\beta_g + \beta_{g1}(1) + \beta_{g2}(s))}{\exp(\beta_g + \beta_{g1}(0) + \beta_{g2}(s))} = \exp(\beta_{g1})$$

si la variable de exposición se encuentra ,y también en un término de interacción, como en:

$$ln\left(\frac{p_{g\,|\,\mathrm{rs}}}{p_{g'\,|\,\mathrm{rs}}}\right) = \beta_g + \beta_{g1}X_1 + \beta_{g2}X_2 + \beta_{g3}X_1X_2 \qquad (1.14)$$

siguiendo el procedimiento anterior se tiene que:

$$OR_{[g,g']\,(1,0)(s)} = \exp(\beta_{g1} + \beta_{g3}s)$$

Lo que muestra que las OR dependen de los coeficientes de los términos en los que interviene la variable de exposición y en las variables que aparecen en interacciones con dicha variable.

Si la variable de exposición X_1 es continua y a y b son dos valores de entonces la del modelo (2) se tiene

$$OR_{[g,g']}(a,b)(s) = \frac{\exp(\beta_g + \beta_{g1}a + \beta_{g2}s + \beta_{g3}as)}{\exp(\beta_g + \beta_{g1}b + \beta_{g2}s + \beta_{g3}bs)}$$

$$= \exp(\beta_{g1}(a-b) + \beta_{g3}(a-b)s)$$

Si la variable de exposición X_1 = E es nominal con K categorías, entonces se forman k - 1 variables dummy o indicadoras, E_1,……., E_{k-1}, donde:

$$E_i = \begin{cases} 1, & si\ se\ da\ la\ categoría\ i \\ 0, & de\ otra\ forma \end{cases}$$

El modelo logístico sin interacciones es:

$$ln\left(\frac{p_{g|\mathrm{rs}}}{p_{g'|\mathrm{rs}}}\right) = \beta_g + \sum_{i=1}^{K-1} \beta_{gi}E_i + \beta_{g2}X_2$$

Para establecer las OR, que compara las categorías E^* y E^{**} de la variable de exposición E, se deben definir estas categorías en términos de las k - 1 variables dummy así: $E_1^*, \ldots\ldots, E_{K-1}^*$ y $E_1^{**}, \ldots\ldots, E_{K-1}^{**}$

$$E_r^* = \begin{cases} 1, & si\ se\ da\ la\ categoría\ E^* \\ 0, & de\ otra\ forma \end{cases} \quad y$$

$$E_s^{**} = \begin{cases} 1, & si\ se\ da\ la\ categoría\ E^{**} \\ 0, & de\ otra\ forma \end{cases}$$

con r, s = 1,……, k – 1

Las OR, para comparar las categorías y de la variable nominal de exposición E, en un modelo logístico sin interacciones es

$$OR = \exp[(E_1^* - E_1^{**})\beta_1 + \ldots\ldots\ldots + (E_{K-1}^* - E_{K-1}^{**})\beta_{K-1}]$$

2. DISEÑO METODOLÓGICO

Las estadísticas mundiales y nacionales muestran que el cáncer es la segunda causa de muerte en las mujeres, siendo el cáncer de cuello uterino, el que aparece en este medio como el segundo cáncer incidente más frecuente, después del cáncer de glándula mamaria y el primero en mortalidad. Estos datos obligan a establecer intervenciones en las mujeres susceptibles para lograr con adecuados programas de detección la disminución de las tasas de incidencia y de mortalidad. Como el cuello uterino es un sitio de fácil abordaje, se facilita la aplicación de pruebas de tamizaje para lograr la detección, tanto de las lesiones malignas como de las premalignas que se asocian en grado variable con la progresión a cáncer. Este contraste está dado por todos los factores asociados con la génesis de la neoplasia y los obstáculos, tanto personales como sociales y del sistema de salud para el diagnóstico precoz de la enfermedad. [17]

En virtud de lo expuesto anteriormente, el presente artículo, en aras de contribuir conceptualmente en la búsqueda de mayor comprensión y explicitación de los determinantes que, en materia de salud, mejor describen los factores de riesgos para mujeres que sufren de cáncer de cuello uterino, tiene como propósito establecer una caracterización que permita emitir diagnósticos.

2.1. Materiales y métodos

Se estudió una muestra constituida por 105 pacientes, que incluyó a todas las mujeres que concurrieron a la consulta ginecológica en una campaña desarrollada por la Secretaría de Salud del Departamento del Atlántico en 2006. Se utilizaron dos instrumentos para la recolección de los datos, registrándose los relacionados a cáncer de cuello uterino en un formulario especial.

Definición de Variables

El análisis realizado estableció que la variable dependiente fuese el Cáncer de cuello uterino (CCU) y las independientes la Edad (ED), el Número de Hijos Nacidos Vivos (NHV), Número de Hijos Nacidos Muertos (NHM), tipo de parto (TP) y tipo de embarazo (TE)), así como las enfermedades venéreas (EV) de la paciente (sífilis, herpes, gonorrea u otras enfermedades de transmisión sexual).

Estas variables se detallan a continuación:

- **Variable dependiente**

Cáncer del cuello uterino (ccu = 0: Leve, 1: Moderado, 2: Agudo)

- **Variables independientes**

Edad (ed), número de hijos nacidos vivos (nhv), número de hijos nacidos muertos (nhm), tipo de parto (tp= 1: espontáneo, 2: cesárea), tipo de embarazo (te =1: bajo riesgo, 2: alto riesgo), enfermedades venéreas (ev =1: no, 2 si)

Para en esta investigación se obtuvo una muestra de (105) mujeres.

2.2. Tablas de validacion del modelo

Para la consecución del objetivo final, se desarrolló de un modelo de regresión logística ordinal donde la variable dependiente es el cáncer de cuello uterino. Utilizando el paquete estadística SPSS, como primera media, es necesario saber si el modelo da predicciones:

En la siguiente tabla:

Tabla 1. Información sobre el ajuste de los modelos.

Modelo	-2 log de la verosimilitud	Chi-cuadrado	Gl	Sig.
Sólo intersección	172.789			
Final	135.382	37.407	6	.000

Función de vínculo: Logit.

Fuente: Elaboración de los autores.

Se presenta la prueba de hipótesis del estudio:

$$H_0: \textit{el modelo es adecuado sólo con la constante}$$

$$H_1: \textit{el modelo no es adecuado sólo con la constante}$$

Debido a que el p-valor de la prueba es menor que 0.05, se rechaza la hipótesis nula. Por tanto, el significado estadístico que resulta, indica que el modelo con las variables introducidas mejora el ajuste de forma significativa, respecto al modelo con sólo la constante.

En la siguiente tabla:

Tabla 2. Bondad de ajuste.

	Chi-cuadrado	Gl	Sig.
Pearson	145.264	126	.115
Desviación	115.032	126	.748

Función de vínculo: Logit.

Fuente: Elaboración de los autores.

se presenta la prueba de hipótesis:

$$H_0: \textit{el modelo se ajusta adecuadamente a los datos}$$

$$H_1: \textit{el modelo no se ajusta adecuadamente a los datos}$$

Esta tabla contiene la estadística de chi-cuadrado de Pearson para el modelo y otra estadística de chi-cuadrado sobre la base de la desviación. Estas estadísticas tienen por objeto comprobar si los datos observados son incompatibles con el modelo ajustado.

Estas estadísticas pueden ser muy útiles para los modelos con un pequeño número de variables predictoras categóricas. Lamentablemente, estas estadísticas son sensibles a las celdas vacías. Al estimar los modelos con covariables continuas, hay muchas celdas vacías, a menudo, como en este caso. Por lo tanto, no se puede confiar en cualquiera de estas estadísticas de prueba con esos modelos. Debido a las celdas vacías, no se puede estar seguros que estas estadísticas realmente siguen la distribución chi-cuadrado, y los valores de significación no se precisa.

En la siguiente tabla:

Tabla 3. Pseudo R-cuadrado.

Cox y Snell	.300
Nagelkerke	.353
McFadden	.188

Función de vínculo: Logit.

Fuente: Elaboración de los autores.

Muestra, en este tipo de modelos, medidas equivalentes al coeficiente de determinación, R2, de los modelos lineales, que resumen la proporción de la variabilidad en la variable dependiente (cáncer del cuello uterino) asociada con los factores de predicción (variables independientes).

Estos valores de la pseudo-r cuadrado son respetables muestras de la variabilidad explicada por el modelo, y en ellas se observa que la Nagelkerke estima en un 35.3% tal variabilidad.

Estas medidas tienen su mayor efectividad cuando se comparan modelos, como los que se muestra a continuación en la siguiente tabla:

Tabla 4. Estimaciones de los parámetros.

		Estimación	Error típ.	Wald	gl	Sig.	Intervalo de confianza 95%	
							Límite inferior	Límite superior
Umbral	[ccu = 0]	-1.352	4.452	.092	1	.761	-10.079	7.374
	[ccu = 1]	.604	4.450	.018	1	.892	-8.117	9.325
Ubicación	Ed	-.102	.108	.882	1	.348	-.314	.111
	Nhv	.928	.266	12.172	1	.000	.407	1.450
	Nhm	-1.513	.467	10.487	1	.001	-2.429	-.597
	[tp=1]	-.542	.451	1.444	1	.230	-1.426	.342
	[tp=2]	0[a]	.	.	0	.	.	.
	[te=1]	1.205	.900	1.791	1	.181	-.560	2.969
	[te=2]	0[a]	.	.	0	.	.	.
	[ev=1]	1.352	1.043	1.680	1	.195	-.692	3.396
	[ev=2]	0[a]	.	.	0	.	.	.

Función de vínculo: Logit.

a. Este parámetro se establece en cero porque es redundante.

Fuente: Elaboración de los autores.

Muestra la estimación de los parámetros del modelo, la prueba de significado de cada predictor y

el intervalo de confianza de cada parámetro ($\delta_i y\ \beta_j$), entre otros. Se observa que hay variables que muestran tienen poca significación en el modelo por presentar, sus pruebas de significado, valores p mayores que 0.05 y, por lo tanto, pueden ser objeto de eliminación.

Tabla 5. Prueba de líneas paralelas[c].

Modelo	-2 log de la verosimilitud	Chi-cuadrado	gl	Sig.
Hipótesis nula	135.382			
General	129.209[a]	6.174[b]	6	.404

La hipótesis nula establece que los parámetros de ubicación (los coeficientes para las pendientes) son los mismos para todas las categorías de respuesta.

a. El valor del logaritmo de la verosimilitud ya no se puede incrementar tras un número máximo de subdivisiones.

b. El estadístico de chi cuadrado se calcula basándose en el valor del logaritmo de la verosimilitud de la última iteración del modelo general. La validez de este contraste es incierta.

c. Función de vínculo: Logit.

Fuente: Elaboración de los autores.

La tabla propuesta, muestra a continuación la prueba para validar el procedimiento de regresión ordinal. El supuesto del modelo queda validado con el no rechazo de la hipótesis nula, como en este caso, pero este resultado es incierto por el aviso dado por el SPSS; en consecuencia se procederá con la eliminación de las variables de poco significado chequeando en cada eliminación esta y las tres primeras tablas para el modelo depurado.

Procediendo a la eliminación de las variables *ed*, *ev* y *te*, se tiene la siguiente tabla con la estimación de los parámetros para las variables que quedaron:

Tabla 6. Estimaciones de los parámetros.

		Estimación	Error típ.	Wald	gl	Sig.	Intervalo de confianza 95%	
							Límite inferior	Límite superior
Umbral	[ccu = 0]	.278	.583	.228	1	.633	-.864	1.420
	[ccu = 1]	2.176	.627	12.058	1	.001	.948	3.405
Ubicación	Nhv	.926	.261	12.617	1	.000	.415	1.436
	Nhm	-1.472	.457	10.361	1	.001	-2.368	-.576
	[tp=1]	-.521	.441	1.400	1	.237	-1.385	.342
	[tp=2]	0[a]	.	.	0	.	.	.

Función de vínculo: Logit.

a. Este parámetro se establece en cero porque es redundante.

Fuente: Elaboración de los autores.

El modelo que corresponde a esta tabla anterior es:

$$\ln\left(\frac{P(ccu\leq g)}{1-P(ccu\leq g)}\right) = \delta_g - \beta_1 nhv - \beta_2 nhm - \beta_3(tp = 1),$$

que despejando $P(ccu \leq g)$ da como resultado la ecuación del riesgo acumulativo:

$$P(ccu \leq g) = \frac{1}{1 + \exp[-\left(\delta_g - \beta_1 nhv - \beta_2 nhm - \beta_3(tp = 1)\right)]}$$

La siguiente tabla muestra la Prueba de hipótesis de líneas paralelas, necesaria para validar el procedimiento de regresión ordinal. Las hipótesis son:

$$H_0: Los\ \beta_i\ son\ los\ mismos\ para\ todos\ los\ niveles\ de\ la\ respuesta$$

$$H_1: Los\ \beta_i\ no\ son\ los\ mismos\ para\ todos\ los\ niveles\ de\ la\ respuesta$$

Tabla 7. Prueba de líneas paralelas[a].

Modelo	-2 log de la verosimilitud	Chi-cuadrado	gl	Sig.
Hipótesis nula	64.402			
General	57.024	7.378	3	.061

La hipótesis nula establece que los parámetros de ubicación (los coeficientes para las pendientes) son los mismos para todas las categorías de respuesta.
a. Función de vínculo: Logit.

Fuente: Elaboración de los autores.

El no rechazo de la hipótesis nula, por ser el p-valor mayor que 0.05, indica que el procedimiento ordinal es viable, ya que no se rechaza la igualdad de las pendientes (βi).

Otros resultados que muestran la validez del nuevo modelo son:

Tabla 8. Información sobre el ajuste de los modelos.

Modelo	-2 log de la verosimilitud	Chi-cuadrado	gl	Sig.
Sólo intersección	96.882			
Final	64.402	32.480	3	.000

Función de vínculo: Logit.

Fuente: Elaboración de los autores.

Tabla 9. Bondad de ajuste.

	Chi-cuadrado	Gl	Sig.
Pearson	25.084	25	.458
Desviación	28.345	25	.292

Función de vínculo: Logit.

Fuente: Elaboración de los autores.

Tabla 10. Pseudo R-cuadrado.

Cox y Snell	.266
Nagelkerke	.313
McFadden	.163

Función de vínculo: Logit.

Fuente: Elaboración de los autores.

2.3. Estimaciones de riesgos, riesgos relativos y razones ODDS

De la ecuación de los Riesgos

$$P(ccu \leq g) = \frac{1}{1 + \exp[-\left(\delta_g - \beta_1 nhv - \beta_2 nhm - \beta_3 (tp = 1)\right)]}$$

al utilizar las estimaciones de la tabla 3.6, para el caso ccu ≤ 0 dados, nhv=1, nhm=1 y tp = 1, se tiene:

$$\hat{P}(ccu \leq 0) = \frac{1}{1+\exp[-(0.278-0.926+1.472+0.521)]} = 0{,}79331$$

También:

$$\hat{P}(ccu \leq 1) = \frac{1}{1 + \exp[-(2.176 - 0.926 + 1.472 + 0.521)]} = 0{,}96242$$

Luego:

$$\hat{P}(ccu = 0) = \hat{P}(ccu \leq 0) = 0{,}79331$$
$$\hat{P}(ccu = 1) = \hat{P}(ccu \leq 1) - \hat{P}(ccu \leq 0) = 0{,}16911$$

Y:

$$\hat{P}(ccu = 2) = 1 - \hat{P}(ccu = 0) - \hat{P}(ccu = 1) = 0{,}03758$$

Estos riesgos, y los correspondientes a cada grupo de valores de las variables independientes. se encuentran en el editor de datos, después de haber implementado el procedimiento que los origina:

2.3.1. Riesgos relativos estimados

$$\widehat{R}_{ccu=1\,|\,(tp=2,1)(nhv=0)(nhm=0)} = \frac{p_{1\,|\,2(0)(0)}}{p_{1\,|\,1(0)(0)}} = \frac{0.32}{0.25} = 1.32$$

Este RR indica que el riesgo (la probabilidad) de que la paciente presente cáncer del cuello uterino moderado cuando el parto es espontáneo, es 1.32 veces mayor que cuando el parto es por cesárea, en mujeres que no han tenido hijos nacidos vivos y/o muertos.

2.3.2. Razones odds estimadas

$$\widehat{OR}_{[ccu,ccu']\,(tp,tp')} = \frac{R_{ccu(tp,tp')(nhv)(nhm)}}{R_{ccu'\,(tp,tp')(nhv)(nhm)}}$$

El riesgo relativo de tener un cáncer del cuello uterino moderado es similar, veces más, a tener cáncer del cuello uterino agudo, cuando se ha tenido parto espontáneo, en comparación con cesárea y en mujeres sin hijos nacidos vivos o muertos.

3. ANÁLISIS Y CONCLUSIONES

Al parecer el riego de cáncer de cuello uterino leve es mayor cuando el único hijo nació muerto y el parto fue espontáneo. Este riesgo disminuye con el nacimiento de hijos vivos y el parto es con cesárea. El parto con cesárea y el aumento del número de hijos vivos son factores más influyen en el mayor riesgo de contraer cáncer de cuello uterino moderado. Este riego disminuye al disminuir el número de hijos nacidos vivos o muertos.

El riesgo de tener cáncer de cuello uterino agudo, al parecer, es mayor en partos por cesárea, con aumento del número de hijos vivos. Este riesgo muestra valores pequeños cuando el parto es espontáneo y solo se ha tenido un hijo que nació muerto.

En síntesis, los riesgos de tener cáncer de cuello uterino leve o moderado, está asociado al parto espontaneó. En el caso de los riesgos para cáncer de cuello uterino agudo estos registran un alta probabilidad para los partos con cesaría.

Por lo general, el riesgo de contraer el cáncer de cuello uterino moderado es un poco mayor cuando el parto es con cesárea que cuando es espontáneo, alcanzando su mayor valor (1.5 veces) en mujeres con un sólo parto y el hijo nació muerto, pero es menor (0.91 y 0.75 veces) si ha se ha tenido más de un parto con sólo nacidos vivos.

También se observa que el riesgo de contraer el cáncer agudo siempre es mayor en partos por cesárea que espontáneo, mostrando un valor máximo (1.67 veces) cuando la mujer no ha tenido hijos, y el valor mínimo (1.25 veces) cuando ha tenido 3 partos y todos nacidos vivos.

De otra parte, el riesgo de presentar cáncer del cuello uterino moderado, cuando tiene tres hijos nacidos vivos, es 5.25 veces mayor que cuando no tiene hijos nacidos vivos, en mujeres que han tenido parto espontaneo y un hijo nacido muerto, pero es menor (0.65 y 0.68 veces), cuando el parto es por cesárea y no tiene hijo muerto.

Cuando se comparan el número de hijos nacidos vivos, el riesgo de adquirir el cáncer mencionado aumenta con el aumento de éstos. La diferencia de esta relación es mayor (11.63) en mujeres con 3 hijos nacidos vivos que en las que no los han tenido, cuando el parto es con cesárea y tuvieron un hijo nacido muerto. La relación se reduce con la disminución de la diferencia entre el número de hijos nacidos vivos, llegando a sus valores menores (1 veces) cuando esta diferencia es 1 y en mujeres que tuvieron parto con cesárea y no han tenido hijos nacidos muertos.

Cuando se compara el número de hijos nacidos muertos, el riesgo de adquirir el cáncer mencionado aumenta con la diminución de éstos. La diferencia de esta relación es mayor (2.29) en mujeres con ningún hijo nacido muerto que en las que no los han tenido cuando el parto es espontáneo y tienen un hijo nacido vivo. La relación se reduce con el aumento el número de hijos nacidos muertos, llegando a sus valores menores (0.64 veces), cuando esta diferencia es 1 y en mujeres que tuvieron parto con cesárea y han tenido 3 hijos nacidos vivos.

Cuando se comparan el número de hijos nacidos muertos, el riesgo de adquirir el cáncer mencionado aumenta con la diminución de éstos. La diferencia de esta relación es mayor (3.75) en mujeres con ningún hijo nacido muerto que en las que no los han tenido cuando el parto es espontáneo y tienen un hijo nacido vivo. La relación se reduce con el aumento el número de hijos nacidos muertos, llegando a sus valores menores (2.24 veces) cuando esta diferencia es 1 y en mujeres que tuvieron parto con cesárea y han tenido 3 hijos nacidos vivos.

Se observa que el riesgo relativo de tener un cáncer del cuello uterino agudo es (1.6 veces) más que tener cáncer del cuello uterino moderado, cuando se ha tenido parto por cesárea, en comparación con parto espontáneo y en mujeres con tres hijos nacidos vivos, ninguno muerto, esta comparación se reduce (0.99 veces) cuando no tiene hijos nacidos vivos, pero presenta un hijo nacido muerto.

Se observa que el riesgo relativo de tener un cáncer del cuello uterino agudo es (3.5 veces) más que tener cáncer del cuello uterino moderado, cuando no se ha tenido ningún hijo nacido muerto, en comparación con un nacido muerto y en mujeres con parto por cesárea, con tres hijos nacidos vivos, esta comparación se reduce (0.96 veces) cuando no tiene hijos nacidos vivos, en mujeres con parto espontáneo.

Se observa que el riesgo relativo de tener un cáncer del cuello uterino moderado es (7.5 veces) más que tener cáncer del cuello uterino agudo, cuando se han tenido tres hijos nacidos vivos, en comparación con ningún hijo nacido vivo y en mujeres con parto por cesárea, ningún hijo nacido

muerto, esta comparación se reduce (0.71 veces) cuando se tiene un hijo nacido muerto, en mujeres con parto por cesárea.

En conclusión, a manera general, se observa que el riesgo de tener cáncer de cuello uterino es mayor cuando, aumenta el número de hijos en partos por cesárea y se ha perdido de un hijo.

REFERENCIAS BIBLIOGRÁFICAS

[1] E. García Ayala, J. Díaz Pérez, M. Melo, F. Parra Fuentes, L.Vera y J.Latorre, "Factores asociados a la identificación del cáncer de cuello uterino en la citología, colposcopia y biopsia en la liga santandereana de lucha contra el cáncer de 2002 a 2003", *Revista española de patología*, vol. 40, No. 1., 2007.

[2] J. Truett, J. Cornfield y W. Kannel, "A multivariate analysis of the risk of coronary heart disease in Framingham", *J Chronic Dis;* vol. 20, n. 7, pp. 511-524., 1967.

[3] Breslow y Day, Statistical Methods in Cancer Research, *Volume I - The analysis of case-control studies*, 1980.

[4] D. R. Cox, *The Analysis of Binary Data*, 1970.

[5] D. G. Kleinbaum, L. L. Kupper and H. Morgenstern, "Epidemiologic Research". *Belmont, Calif: Lifetime Learning Publications*, 1982.

[6] J. J. Schlesslman, "Case control studies". *Desig, Conduct, Análisis*. Nueva York: Oxford University Press, 1982.

[7] L. Flores, *Análisis Estadístico de Factores de riesgo que influyen en la enfermedad Angina de Pecho*, Oficina General del Sistema de Bibliotecas y Biblioteca Central UNMSM., 2002.

[8] M. Ato García, y J. J. López García, *Análisis estadístico para datos categóricos*. Madrid: Editorial Síntesis, 1996.

[9] A. Alderet. "Fundamentos del Análisis de Regresión Logística en la Investigación Psicológica", *Revista Evaluar*, vol. 6, pp. 52-67., 2006.

[10] J. F. Hair, R.E. Anderson, R. L. Tatham y W.C. Black, *Análisis Multivariante*. 5° Edición. Madrid: Prentice Hall, 1999.

[11] M. V. García Jiménez, J. M. Alvarado Izquierdo, y J. Jiménez Blanco, "La predicción del rendimiento académico: regresión lineal versus regresión logística", en: *Psicothema*, vol. 12, n°. 2, pp. 248-252., 2000.

[12] N. Cortada de Cohan, "Teoría de Respuesta al Ítem", *Evaluar*, n°. 4,pp. 95-110, 2004.

[13] D. Ferreres Traver, A. M. Hidalgo Aliste y J. Muñiz, "Detección del Funcionamiento Diferencial de los ítems no uniforme: comparación de los métodos Mantel-Haenszel y regresión logística", *Psicothema*, vol. 12, n°. 2, pp. 220-225., 2000.

[14] M. Hidalgo Montesinos y J. A. López Pina, "Comparación entre las medias de área, el estadístico de Lord y el análisis de regresión logística en la evaluación del funcionamiento diferencial de los ítems", *Psicothema*, vol. 9, n°. 2, pp. 417-431., 1997.

[15] L. Flores Manrique, *Análisis Estadístico de los Factores de Riesgo que Influyen en la Enfermedad Angina de pecho*., 2002.

[16] V. Pando Fernández y R. San Martín Fernández, *Regresión Logística Multinomial*. Madrid: Universidad de Valladolid, 2004.

[17] R. Ortiz serrano, C. Uribe Pérez, "Factores de Riesgo para cáncer de Cuello Uterino", *Colombiana de Obstetricia y Ginecología*, vol. 55, n°. 2, pp. 146-160, 2004.