Modelando la detección del riesgo de niveles altos de *Cryptosporidium* en fuentes de aguas(*)

Carrera, Elena*; Vaira, Stella*; Abramovich, Beatriz**;
Walz, Florencia*; Contini, Liliana*

*- Dpto. de Matemática, Area Estadística.
**- Sección Aguas, Dpto. de Ciencias Biológicas. Facultad de Bioquímica y Ciencias Biológicas.
Universidad Nacional del Litoral. CC. 242. Ciudad Universitaria, Paraje El Pozo. (3000) Santa Fe.
Tel. 0342-4575210. ecarrera@fbcb.unl.edu.ar

RESUMEN: La presencia de *Cryptosporidium* ha producido numerosos brotes de origen hídrico. El análisis de enteroparásitos en agua requiere procedimientos operativos laboriosos y costosos. En este trabajo, se propone un modelo de regresión logística binaria para la determinación de posibles factores de riesgo, que aumenten la probabilidad de la presencia de altas concentraciones de ooquistes en fuentes de agua. Se obtuvieron dos modelos predictores donde intervinieron variables como nivel hidrométrico, materia orgánica y bacterias indicadoras de contaminación.

Palabras claves: modelos, regresión logística, factores de riesgo, aguas.

SUMMARY: Modelling the detection of risk posed by high levels of *cryptosporidium* in water supplies. Carrera, Elena*; Vaira, Stella*; Abramovich, Beatriz, Walz, Florencia*; Contini, Liliana*.** *Cryptosporidium* have been responsible for numerous waterbone outbreaks. The analysis of enteroparasites in ater involves expensive and time-consuming procedures. In this work, a binary logistic regression model for detection risk factors wich could increase the probability of finding high oocyst concentrations in water supplies has been devised. Two predictive models were obtained. Variables such as hydrometric level, organic matter and bacteria indicating contamination were taken into account.

Key words: models, logistic regression, risk factors, water.

Introducción

Los modelos matemáticos han sido usados en las ciencias biológicas profusamente en años recientes. Su empleo en Epidemiología data de 1960 aproximadamente, aunque ya en 1760 Daniel Bernoulli propuso un modelo para comprender los mecanismos de dispersión de la viruela de los pollos (1). Estos fueron incorporados más tarde para entender la dinámica de la transmisión de una determinada enfermedad y predecir los efectos de las diferentes intervenciones del hombre en el desarrollo de la misma, tales como: pruebas de vacunas, medidas profilácticas, relación con la aparición de otras enfermedades y medidas de control.

Son numerosos los trabajos sobre SIDA, tuberculosis y viruela entre otros, aunque actualmente además son de aplicación común para el entendimiento de las particularidades del crecimiento de los tumores y otros mecanismos de la carcinogénesis (1-8).

(*)Este trabajo fue desarrollado en el marco del proyecto CAI+D '2002-15-104: Modelos epidemiológicos: matemáticos y estadísticos avanzados, Facultad de Bioquímica y Ciencias Biológicas de la UNL.

Recibido: 30-06-03
Aceptado: 18-09-03

Algunos de estos modelos epidemiológicos son netamente matemáticos y otros de naturaleza estadística. Los matemáticos, pueden ser determinísticos o estocásticos. Los primeros se expresan, generalmente, a través de ecuaciones diferenciales o sistemas de ecuaciones diferenciales, y los segundos se sustentan en la teoría de probabilidades concurriendo en su formulación: cadenas de Markov, procesos de Poisson, ecuaciones diferenciales probabilísticas y procesos Brownianos, para mencionar sólo algunos (4, 9, 10). Los estadísticos, entre los cuales los regresores son los más conocidos, auxilian en la investigación de los efectos de variables explicatorias en aquellas elegidas como respuesta. La forma estructural del mismo describe las relaciones, asociaciones e interacciones entre las variables. La cantidad de parámetros estimados en el modelo, determinan la fuerza e importancia de los efectos. Finalmente, permiten predecir valores a partir de los datos y proveen una mejor estimación de la respuesta media y su probable distribución. Su valor estriba en que, a veces, sugieren un simple resumen de los datos en términos de los efectos sistemáticos mayores, junto con un resumen de la naturaleza y magnitud de la variación aleatoria o no explicada (11-13).

Antes de elegir un modelo se requiere previamente mirar inteligentemente los datos en demanda de patrones de comportamiento. Formularlos es importante, ya que pueden ser utilizados en mediciones similares recolectados por otro investigador en otro tiempo y lugar. Una importante propiedad del modelo es fijar su alcance o región de influencia, esto es el rango de condiciones sobre los cuales predice bien, pues el realizar predicciones fuera de ese rango conduce a conclusiones casi siempre falsas (12).

Los modelos de regresión logística para respuesta binaria (RL), son en esencia un método multivariado que está diseñado para variables de salida o dependientes con sólo dos valores posibles, siendo las predictoras o covariables de cualquier naturaleza: continuas, discretas, dicotómicas u ordinales, sin realizar ninguna suposición acerca de la distribución de las mismas. Este modelo regresor, transforma la variable dependiente **y** en una variable logit, que es el logaritmo natural del cociente: $\frac{y}{1-y}$

Esta expresión, que deberá leerse: *ocurrencia de un evento dividido no ocurrencia del mismo* es conocida en epidemiología como «odds ratio» o simplemente «odds». Luego se estima la probabilidad de que cierto suceso ocurra a través del cálculo del $\ln \frac{y}{1-y}$ llamado también logaritmo del odds (13-15).

Una de las características que hacen interesante a la regresión logística es la relación que los coeficientes del modelo guardan con un parámetro de cuantificación de riesgo, muy usado en epidemiología, denominado riesgo relativo (12).

El estudio de la polución hídrica se realiza mediante la observación y cuantificación de variables fisicoquímicas y microbiológicas que se consideran posibles predictoras del nivel de contaminación. La laboriosidad y costo elevado de los métodos para determinar la presencia de enteroparásitos ha llevado a la búsqueda de modelos, tanto matemáticos como estadísticos, con los que se trata de predecir no sólo la concentración de protozoos en función de otras variables, sino también establecer si esta concentración supera o no algún valor preestablecido.

Dentro de los enteroparásitos, *Cryptosporidium*, es causante de severas gastroenteritis sobre todo en niños e individuos inmunodeprimidos y presenta una gran dispersión geográfica en aguas superficiales de distintos países, tranformándolo en un problema epidemiológico de interés internacional (16).

El objetivo de este artículo es mostrar la ductilidad de los modelos, matemáticos o estadísticos, en particular el de regresión logística binaria, para la determinación de posibles factores de riesgo que aumenten la probabilidad de la presencia de altas concentraciones de *Cryptosporidium* en fuentes de aguas superficiales.

Materiales y Métodos

Para el análisis de la calidad del agua se determinaron turbiedad y materia orgánica (MO: Oxígeno consumido del permanganato de potasio) como parámetros fisicoquímicos. En el grupo de las variables microbiológicas se consideraron: coliformes totales, coliformes termotolerantes, *Escherichia coli*

(*E.coli*), estreptococos fecales, *Enterococcus spp* y *Pseudomonas aeruginosa* (*P.aeruginosa*) y *Cryptosporidium*; además se incluyó el nivel hidrométrico (NH) correspondiente al día en que se realizaron las mediciones. El muestreo y análisis de las diferentes variables observadas se describieron en detalle en Carrera *et al.*, 2001 (17). En el presente trabajo la variable parasitológica (concentración de ooquistes) fue codificada en dos categorías según superara o no los 300 ooquistes/100 l. Este valor fue seleccionado teniendo en cuenta los tratamientos adicionales en los procesos de potabilización requeridos para su remoción cuando la concentración supera el valor señalado (18).

Análisis estadístico

Se empleó Regresión Logística, para desarrollar un modelo predictor de la variable dicotómica concentración de valores de *Cryptosporidium* mayores o iguales a 300 unidades por cada 100 litro de agua a la que se denominó *«alto riesgo»*, siendo *«bajo riesgo»* aquella donde los valores no superaron al señalado en las 45 muestras de aguas superficiales. Los predictores potenciales considerados fueron: *Enterococcus spp.*, materia orgánica, nivel hidrométrico, *Escherichia coli*, estreptococos fecales y *Pseudomonas aeruginosa*. Las covariables fueron incluidas en el modelo sólo si mejoraban el ajuste, empleándose la prueba chi-cuadrado de Wald para analizar su significancia.

El test de Hosmer-Lemeshow proporcionó una medida global de exactitud predictiva y fue utilizado como una prueba de calidad de ajuste. Para analizar con mayor eficacia la bondad del ajuste se utilizó la deviance (D), que es la medida más directa de su calidad (12, 15, 19-21).

Se utilizó el estadístico Kappa (*K*) para medir el grado o la fuerza de la concordancia entre los valores observados y los que se obtienen por aplicación de los modelos (13, 21).

El modelo de regresión logística binaria propuesto para las observaciones fue:

$$y = \frac{e^{z}}{1 + e^{z}} \quad \text{donde}$$

$$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Se estimaron los coeficientes (β) y el riesgo relativo (e^{β}), así como se construyeron sus intervalos del 95% de confianza (22).

Para medir la capacidad predictora global del resultado obtenido y comparar las salidas se construyeron las curvas ROC (Receiver Operating Characteristic) correspondientes a cada uno de los modelos obtenidos además de calcular el área bajo la curva (23).

Resultados y discusión

Para proponer un conjunto inicial de variables predictoras no asociadas, se calcularon las correlaciones de Spearman y se identificaron aquellas que estaban relacionadas (Tabla 1).

Se propusieron los siguientes conjuntos de covariables para lograr los mejores modelos de regresión logística:

Conjunto A: NH, MO, coliformes totales y alternativamente *Enterococcus* y turbiedad.

Conjunto B: Nivel hidrométrico, Materia Orgánica, coliformes totales, *Escherichia coli* o en su lugar estreptococos fecales.

Con la introducción de los conjuntos A y B secuencialmente en el modelo inicial, con el método de eliminación hacia atrás (backward), con una significancia menor a 0,10 en el estadístico de Wald para la permanencia de una variable y mayor para su remoción, se obtuvieron como modelos finales los que figuran en las tablas 2 y 3. Esto luego de rechazar aquellos que no resultaban óptimos.

El modelo ajustado correspondiente a los datos de la Tabla 2 fue:

$$y_{(ajustado)} = \frac{e^{z}}{1+e^{z}} \quad \text{donde } z = 10{,}450 - 2{,}880*NH + 0{,}155*MO + 0{,}012*Enterococcus$$

(Modelo A) y el correspondiente a los datos de la Tabla 3, la ecuación ajustada es:

$$y_{(ajustado)} = \frac{e^{z}}{1+e^{z}}, \text{ donde } z = 9{,}740 - 2{,}878*NH + 0{,}170*MO + 0{,}006*E.coli \qquad \text{(Modelo B)}$$

Análisis de los modelos

A: Se observa que la variable NH tiene un coeficiente negativo y su riesgo es menor que uno, aún el extremo superior (0,416) de su intervalo de confian-

za. Esto diminuye el valor de z o sea la probabilidad de presencia de niveles de *Cryptosporidium* superiores o iguales a 300/100 l.

El valor de $e^{\beta} = 1,170$ para MO y el análisis de su intervalo de confianza asignó a esta variable un riesgo real, es decir, el aumento de MO tiene asociado un aumento en la probabilidad de niveles de *Cryptosporidium* mayores o iguales a 300/ 100 l de un máximo de 38,3%.

Si bien estadísticamente la presencia de *Enterococcus* fue significativa pero con un riesgo prácticamente inexistente, como máximo de 2,6%. Su presencia fue importante no solo porque mejora el ajuste sino porque figura en la lista de los análisis bacteriológicos de calidad de agua y daría indicios de niveles probables de *Cryptosporidium* mayores al límite prefijado.

B: La variable NH tuvo un comportamiento similar al ya descrito para el **A**, es decir el coeficiente estimado es negativo y ambos extremos del intervalo de confianza para e^{β} son menores a uno. Esto lo transforma en un *factor de protección*, dado que un aumento de su valor, produce una disminución en la probabilidad de presencia de *Cryptosporidium* mayores al valor establecido.

Por el contrario, MO tuvo asociado un riesgo mayor que uno ($e^{\beta} = 1,190$). De la observación del extremo superior de su intervalo de confianza, surgió que cuando está presente aumenta la probabilidad de valores de alto riesgo de *Cryptosporidium* en un 43%. Por otra parte, la presencia de *E.coli* en el agua analizada aumentó sólo en 1,1% el riesgo de encontrar probables concentraciones de riesgo del parásito. Su consideración en el modelo estuvo justificada desde el punto de vista estadístico ya que fue significativa, y al igual que los *Enterococcus*, en el modelo **A**, su detección fue operativa y económicamente más accesible que la de parásitos y de rutina en los laboratorios de control de calidad de agua.

Los dos modelos evidenciaron un buen ajuste ya que en ambos la prueba de Hosmer-Lemeshow arrojó un $p \geq 0,834$. Además la especificidad (probabilidad de detectar como 0: "bajo riesgo" los verdaderos 0) fue igual en los dos, 75%, y la sensibilidad (probabilidad de asignar un 1: "alto riesgo" a los 1) fue de 87,5% en el A y de 92% para el B, de donde surgieron los promedios de clasificación correcta mayores al 81% (figura 1, tablas 2 y 3). El segundo modelo mejoró la clasificación de muestras donde las concentraciones probables de *Cryptosporidium* eran de alto riesgo, es decir fue más sensible, aunque ambos fueron igualmente parsimoniosos, ya que poseen la misma cantidad de variables predictoras. Por otra parte son satisfactorios como predictivos, debido a que el área bajo las curvas ROC, construidas a partir de los valores de sensibilidad y especificidad son de 0,915 para el Modelo A y 0,940 para el B (figura 2). Se obtuvieron modelos en los que se combinaron variables para lograr el mayor valor predictivo. La selección de las mismas, no sólo se basó en identificar aquellas que producen probabilidades cercanas a 1 de altas concentraciones de ooquistes sino que, como en el caso de las microbiológicas que mostraron valores bajos de riesgo, debieron tenerse en cuenta porque contribuyeron significativamente a la Deviance ($D = 5,045$ para el Modelo A y $D = 3,420$ para el B), estadístico de referencia en cuanto a la calidad del ajuste.

El valor del estadístico *K* obtenido para ambos modelos fue 0,78 que puede considerarse bueno según la clasificación de concordancia descripta en Altman (1997).

Tabla 1: Correlaciones bivariadas no paramétricas y valor P exacto.

Variables		Correlación estimada de Spearman	Valor p
Enterococcus	*E.coli*	0,606*	< 0,001
	MO	0,272	0,075
	NH	-0,029	0,852
	Estreptococos fecales	0,480*	0,001
	P. aeruginosa	0,548*	< 0,001
E.coli	*MO*	0,210	0,167
	NH	0,041	0,788
	Estreptococos fecales	0,383*	0,009
	P. aeruginosa	0,463*	0,002
MO	NH	0,097	0,525
	Estreptococos fecales	0,242	0,110
	P. aeruginosa	0,178	0,249
NH	Estreptococos fecales	0,009	0,953
	P. aeruginosa	-0,202	0,189
Estreptococos fecales	*P. aeruginosa*	0,386*	0,010

E.coli: *Escherichia coli*, NH: nivel hidrométrico, MO: materia orgánica, *P.aeruginosa*: *Pseudomona aeruginosa*. * Correlaciones significativas.

Tabla 2: Resultados de la regresión logística binaria en el Conjunto A. (n = 45).

Covariables	Estimaciones de β (p Wald)	Riesgo Relativo: e^{β}	Intervalo del 95% de confianza para e^{β}
Constante	10,450	—	—
NH (x_1)	-2,880 ; (0,005)	0,056	0,008 - 0,416
M O (x_2)	0,155 ; (0,074)	1,170	0,985 - 1,383
Enterococcus (x_3)	0,012 : (0,078)	1,010	0,999 - 1,026
Prueba de Hosmer – Lemeshow p = 0,953		Clasificación correcta (promedio): 81,8%	

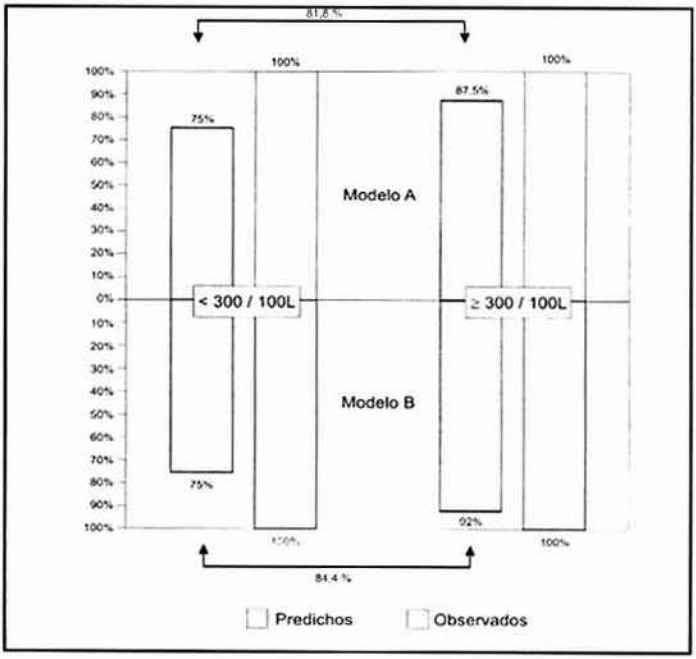
NH: Nivel Hidrométrico; MO: Materia Orgánica. Estimación de los coeficientes β y su significancia (p de Wald). Promedio del poder predictor de clasificación correcta del modelo.

Tabla 3: Resultados de la regresión logística binaria en el Conjunto B. (n= 45).

Covariables	Estimaciones de β (p Wald)	Riesgo Relativo: $_{e}b$	Intervalo del 95% de confianza para $_{e}\beta$
Constante	9,740	—	—
NH (x_1)	-2,878, (0,005)	0,056	0,007 - 0,429
M O (x_2)	0,170, (0,073)	1,190	0,984 - 1,427
E. Coli (x_3)	0,006, (0,031)	1,006	1,001 - 1,011
Prueba de Hosmer - Lemeshow p = 0,834		Clasificación correcta (promedio): 84,4%	

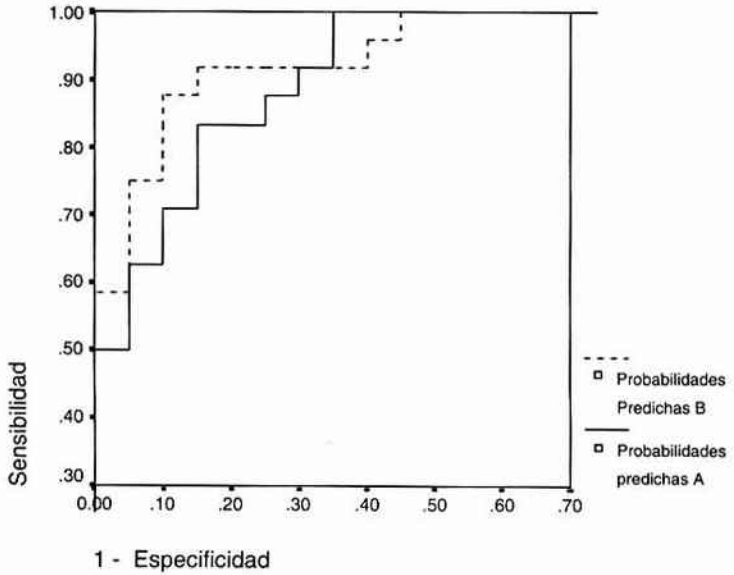
NH: Nivel Hidrométrico; MO: Materia Orgánica; *E. Coli: Escherichia Coli.* Estimación de los coeficientes β y su significancia (p de Wald). Promedio del poder predictor de clasificación correcta del modelo.

Figura 1: Porcentaje de clasificación correcta para los dos modelos obtenidos



Barras en blanco: representan el 100% de los valores observados. Barras sombreadas: representan el porcentaje de valores predichos por ambos modelos. Corresponden las de la izquierda a la especificidad y las de la derecha a la sensibilidad.

Figura 2: Evaluación de los modelos finales de concentración de *Cryptosporidium* (Curvas ROC empíricas)



Las áreas debajo de ambas curvas fueron similares, siendo mayor la del modelo B (0,915 para A y 0,940 para B)

Conclusiones

El beneficio de poder contar con un modelo como los obtenidos se funda en su empleo como predictor de concentraciones de ooquistes. El empleo de técnicas propias de la estadística hacen posible explorar y analizar los probables factores de riesgo asociados con concentraciones de *Cryptosporidium* elevadas en aguas que se emplearán como fuentes para procesos de potabilización.

El nivel hidrométrico, en el caso estudiado, obra como factor de protección. Esto es debido a que en este sistema hidrográfico la concentración de *Cryptosporidium* es menor al aumentar dicho nivel, pues se produciría un efecto de dilución, que coincide con los resultados de Abramovich *et al.* (24).

El bajo riesgo detectado de las variables bacterianas en los modelos finales coincide con el hecho de que estas son deficientes indicadoras de la presencia de *Cryptosporidium* y en general de contaminación parasitológica del agua, debido a la mayor resistencia de los ooquistes a sobrevivir en ella (25).

El factor de riesgo más fuetemente asociado, en ambos modelos finales, a la ocurrencia de más de 300 *Cryptosporidium* /100 l, fue Materia Orgánica.

La importancia del modelo radica en poder ejercer sobre ellos simulaciones que permiten probar acciones tendientes a evaluar los riesgos de fuentes de agua en la transmición de enfermedaes de origen, previo al experimento biológico en sí.

Bibliografía

1. Murray, C; Salomon, J., 1998. Modeling the impact of global tuberculosis control strategies. Proc. Natl.Acad. SCI. USA. **95**: 13881-13886.

2. Becker, N., 1993. Parametric inference for epidemic models. Mathematical Biosciences. **117**: 239-251.

3. Blower, S.; Small, P.; Hopewell, P., 1996. Control strategies for tuberculosis epidemics: New models for old problems.

Science. **273**: 497-500.

4. José, M.; Ruiz, A.; Robadilla, J., 1997. Prevalence of infection mean worm burden and degree of worm aggregation as determinant of prevalence of disease due to intestinal Helminths. Archives of Medical Research. **32**; 1: 121-127.

5. Sattenspiel, L., 1988. Spread and maintenance of a disease in a structured population. American Journal of Physical Anthropology. **77**: 497-504.

6. Billard, L.; Zhao, Z., 1999. Three-Stage stochastic epidemic model: an application to AIDS. Mathematical Biosciences. **107**, 2: 431–449.

7. Hernández-Suarez, C.; Castillo-Chavez, C., 2000. Urn models and vaccini effficacy estimation. Statistics in Medicine. **19**: 827-835.

8. Hethcote, H., 2000. The mathematics of infectious disease. SIAM Review. **42**, 4: 599-653.

9. Robertson, C.; Boyle, P., 1998. Age-Period-Cohort analysis of chronic disease rates. I: Modeling approach. Statistics in Medicine. **17**: 1305-1323.

10. Robertson, C.; Boyle, P., 1998. Age-Period-Cohort analysis of chronic disease rates. II: Graphica approaches. Statistics in Medicine. **17**: 1325-1340.

11. Black, F.; Singer, B., 1987. Elaboration versus simplification in referring mathematical models of infectious disease. Ann. Rev. Microbiol. **41**: 677-701.

12. MacCullag, P.; Nelder, J.A., 1989. "Generalized Linerar Models". Chapman & Hall/CRC. 2nd Ed. (London). 98 – 148.

13. Agresti, A., 1996. "An introduction to categorical data analysis". John Willey & Sons. (New York). 71 – 145.

14. Valsecchi, M.G., 1992. Modelling the relative risk of esophageal cancer in a case-control study. Journal of Clinical Epidemiology, **45**: 347-355.

15. Bagley, S.; White, H.t, Golomb, B., 2001. Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. Journal of Clinical Epidemiology, **54**: 979-985

16. Shaw, K.; Walker, S.; Koopman, B., 2000. Improving filtration of *Cryptosporidium*. J. AWWA. 103 – 111.

17. Carrera, E.; Abramovich, B.; Contini, L.; Vaira, S.; Lurá, M.C., 2001. Parásitos en agua. Modelos estadísticos de predicción. FABICIB. **5**: 77-85.

18. Scharfenaker, M., 2000. Water suppliers assess new rulemaking agreement. J. AWWA. 22-33.

19. Hair J.; Anderson R.; Tatham R.; Black W., 1999. "Análisis Multivariante". Prentice Hall. (Madrid). 249-344.

20. Myers, R. H., 2000. "Classical and modern regression with application". PWS Kent Publishing Company. (Boston). 425 – 449.

21. Altman, Douglas G., 1997. "Practical statistics for medical research". Chapman and Hall. (London). 7ma reimpresión. 339 – 358, 403 - 411.

22. Diaz, M.P.; Demétrio, C.G., 1998. «Introducción a los Modelos lineales generalizados. Su aplicación en las Ciencias Biológicas». SCREEN Editorial. (Córdoba). 7-71.

23. Zweig, M.; Campbell, G., 1993. Receiver Operating Characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry. **39**, 4: 561-577.

24. Abramovich, B.L.; Gilli, M.I.; Haye, M.A.; Carrera, E.; Lurá, M.C.; Nepote, A.; Gomez, P.A.; Vaira, S.; Contini, L., 2001. *Cryptosporidium* y *Giardia* en aguas superficiales. Revista Argentina de Microbiologia. **33**: 167 – 176.

25. Gofti – Laroche, L.; Demanse, D.; Goret, J.C.; Zmirou, D., 2003. Health risks and parasitical quality of water. Journal AWWA. 162 – 172.