UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2011, Vol. 4, No. 2

Carta al Editor

Letter to the Editor

Jorge Iván Velez[a]
jorgeivanvelez@gmail.com

Juan Carlos Correa[b]
jccorrea@unal.edu.co

En el último número de la revista *Comunicaciones en Estadística*, Camacho Quiroga (2011) presenta los resultados de un estudio realizado en Montería, departamento de Córdoba, Colombia, en el que se explora, a través de un modelo de regresión logística, la asociación de un polimorfismo de nucleótido simple (SNP, por sus siglas en inglés) en el gen *MYF5* "*myogenic factor 5 (Bos taurus)*" con el peso vivo al sacrificio en 57 bovinos de las razas Cebú y Romosinuano. En este estudio, el autor encuentra que (*i*) los bovinos Romosinuano homocigotos para el alelo 2 poseen un mayor peso al sacrificio que los bovinos Cebú homocigotos para el mismo alelo; (*ii*) los animales Cebú heterocigotos presentan, en promedio, un menor peso vivo al sacrificio; (*iii*) el genotipo 1/1 no segrega en esta población y (*iv*) las dos variables consideradas en el modelo (raza y presencia del alelo 2) tienen un efecto sobre el peso al sacrificio y lo explican en un 70 % aproximadamente. Si bien estos resultados son muy importantes, consideramos que el artículo presenta algunas inconsistencias técnicas en el modelo propuesto, y otras a lo largo del texto que podrían confundir y/o desinformar al lector. El motivo de esta carta es resaltar estas inconsistencias.

1. Estimación, interpretación y validación del modelo

Consideremos datos simulados[1] correspondientes al peso (75.4 % con peso > 430 kg), raza (Cebú $= 36.84\,\%$) y genotipo ($2/2 = 54.38\,\%$) de 57 animales. En las figuras 1(a) y 1(b) se presentan los pesos por raza y genotipo, respectivamente. Observe que los animales de la raza Romosinuano y los heterocigotos para el alelo 2 parecen tener mejores pesos promedio al sacrificio, lo cual corresponde a la hipótesis de Camacho Quiroga (2011).

[a]Grupo de Investigación en Estadística. Universidad Nacional de Colombia. Sede Medellín.
[b]Profesor Asociado. Escuela de Estadística. Universidad Nacional de Colombia. Sede Medellín.
[1]El código en R (R Development Core Team, 2011) para generar los datos, gráficos y tablas aquí presentados se encuentra disponible a petición del lector.

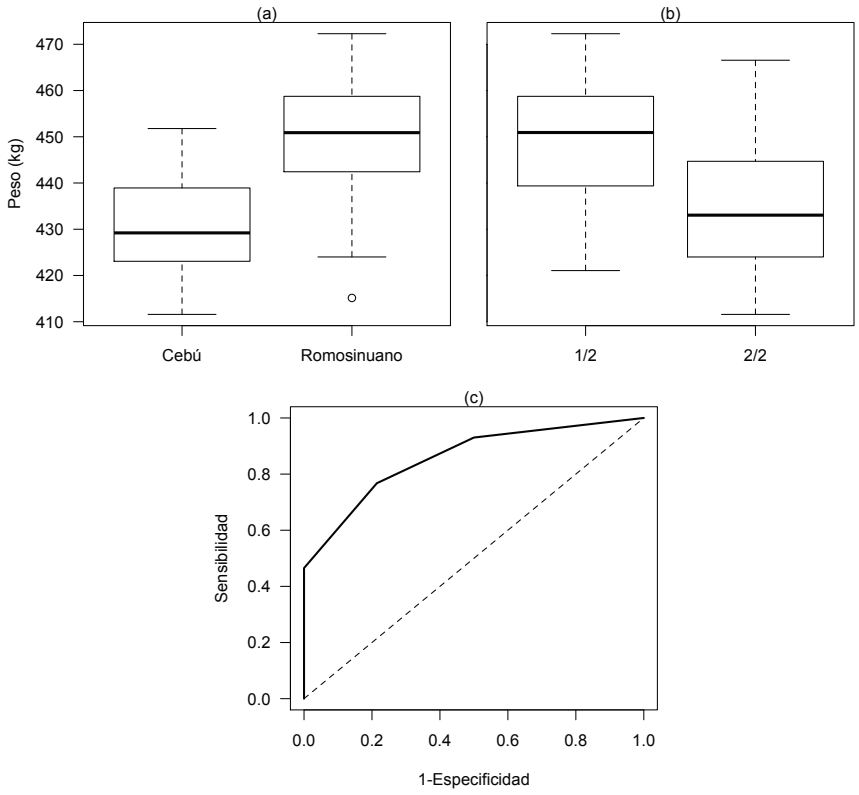

Figura 1: *Gráficos boxplot para los pesos simulados de 57 animales como función de (a) la raza y (b) el genotipo en el SNP localizado en MYF5. En (c) presentamos la curva ROC para el modelo (1).*

El autor propone el modelo de regresión:

$$Y = \beta_0 + \beta_1 \text{ Raza} + \beta_2 \text{ Genotipo} + \epsilon \tag{1}$$

donde

$$Y = \begin{cases} 1, \text{ si Peso} > 430 \text{ kg} \\ 0, \text{ en otro caso} \end{cases},$$

Raza (Cebú = 0; Romosinuano = 1) y Genotipo (1/2 = 0; 2/2 = 1) son variables binomiales, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ corresponde a los parámetros del modelo y ϵ es el error aleatorio.

Es importante resaltar que por las características del modelo de regresión logística (Hosmer & Lemeshow 1989), no es posible hablar de «mejores pesos promedio al sacrificio» por raza y/o genotipo, pero sí de la *probabilidad* de que el peso de un animal sea mayor a 430 kg *dados* su raza y genotipo. Por esta razón, consideramos

que la afirmación "se llegó a corroborar los datos de la revisión bibliográfica donde se afirma que animales de la raza Romosinuano homocigotos para el gen MYF-5 poseen mejores pesos al sacrificio que los animales Cebú homocigotos y también se observó que el grupo de animales de menor peso eran los heterocigotos de la raza Cebú", no es correcta a la luz de los resultados del modelo. Si el objetivo principal fuese comparar los pesos promedio, entonces un modelo de regresión lineal múltiple sería más apropiado (*véase* numeral 2.4).

En la tabla 1 se reportan los coeficientes estimados para el modelo (1) *con* y *sin* intercepto, el primero con la raza Cebú y el genotipo 1/2 como niveles de referencia. Nótese que, al igual que en Camacho Quiroga (2011), en el modelo *con* intercepto resultan ser significativos sólo la Raza y el Genotipo. Adicionalmente, ambos modelos resultan ser *globalmente* significativos al emplear una prueba de razón de verosimilitud (modelo *con* intercepto: $G^2_{(2)} = 19.995, p < 0.0001$; modelo *sin* intercepto: $G^2_{(3)} = 35.463, p < 0.0001$), y tienen el mismo índice de información de Akaike (AIC, por sus siglas en inglés) y el mismo C_{index} o área bajo la curva característica de operación (ROC, por sus siglas en inglés). Llama la atención que en Camacho Quiroga (2011) el valor del AIC sea diferente para ambos modelos.

Tabla 1: *Parámetros estimados para el modelo de regresión logística (a) con y (b) sin intercepto.*

(a)

Variable	$\hat{\beta}$ (SE)	z	p
Intercepto	0.776 (0.607)	1.279	0.201
Raza (Romosinuano)	2.819 (0.842)	3.347	0.001
Genotipo (2/2)	-1.893 (0.828)	-2.286	0.022

$G^2_{(2)} = 19.995, p < 0.0001$; AIC $= 49.556$; $C_{\text{index}} = 0.857$

(b)

Variable	$\hat{\beta}$ (SE)	z	p
Raza (Cebú)	0.776 (0.607)	1.279	0.201
Raza (Romosinuano)	3.596 (0.907)	3.963	$< 10^{-4}$
Genotipo (2/2)	-1.8933 (0.828)	-2.286	0.022

$G^2_{(3)} = 35.463, p < 0.0001$; AIC $= 49.556$, $C_{\text{index}} = 0.857$

Al comparar la tabla 1 con los modelos presentados en Camacho Quiroga (2011, Sección 3.1), vemos con preocupación que en estos últimos no están presentes los niveles de referencia. Al respecto tenemos dos hipótesis: (*i*) el autor ajustó el modelo (1) con los niveles de referencia correctos (raza Cebú y genotipo 1/2) pero desafortunadamente olvidó incluirlos en la versión final del artículo y (*ii*) las variables Raza y Genotipo fueron incluidas en el modelo (1) como variables *numéricas*. Aunque creemos firmemente que la primera hipótesis es la correcta, por como se presentan los resultados en el artículo, desafortunadamente no es posible descartar la segunda.

Otro hecho que llama particularmente la atención es que las variables Raza y Genotipo expliquen «...en un 70 % la respuesta». A pesar de que esta última afirmación podría corresponder al valor del R^2 para el modelo ajustado, este no se presenta en el texto. En el artículo lo más cercano a este valor es el AIC de los modelos *con* y *sin* intercepto (70.164 y 68.218, respectivamente). ¿Es este el 70 % al que se refiere el autor? Esperamos que no.

Algunas interpretaciones basadas en la tabla 1(a) son las siguientes: (*i*) animales de la raza Romosinuano tienen $e^{2.819} \approx 16$ veces más probabilidad de presentar un peso superior a 430 kg que los animales de la raza Cebú y (*ii*) animales de genotipo 1/2 tienen una probabilidad $e^{1.893} \approx 6$ veces mayor de presentar un peso superior a 430 kg que animales de genotipo 2/2 y (*iii*) la probabilidad de que un animal con genotipo 2/2 tenga un peso mayor a 430 kg es $e^{0.776-1.893}/(1+e^{0.776-1.893}) \approx 0.247$ para la raza Cebú y $e^{0.776+2.819-1.893}/(1+e^{0.776+2.819-1.893}) \approx 0.846$ para la raza Romosinuano. Con el modelo *sin* intercepto se obtienen resultados ligeramente distintos. Para otras interpretaciones véase Hosmer & Lemeshow (1989) y Harrell (2001).

La validación de modelos de regresión logística es fundamental, especialmente si estos van a ser utilizados para clasificar *nuevos* individuos. Harrell (2001, Capítulos 10-12) presenta varios ejemplos, medidas y estrategias para ello, todas implementadas en la librería `rms` (Harrell 2011) de R. Otras alternativas incluyen el análisis de curvas ROC (véase Metz (1978) para una introducción) a través de las librerías `pROC` (Robin et al. 2011), `ROCR` (Sing et al. 2009) y `epicalc` (Chongsuvivatwong 2011), o técnicas de validación cruzada como las implementadas en las librerías `caret` (Kuhn 2011) y `boot` (Davison & Hinkley 1997, Canty & Ripley 2011). Si bien el autor presenta el estadístico χ^2 de Pearson como una alternativa para determinar si el modelo es válido o no, creemos que no profundiza lo suficiente en su cálculo e interpretación. La curva ROC para el modelo (1) se presenta en la figura 1(c); el valor del C_{index} es 0.857 (95 %CI = 0.844-0.976). Estos resultados indican que las variables Raza y Genotipo *discriminan* adecuadamente animales con pesos mayores a 430 kg de aquellos con un peso menor.

2. Otros detalles técnicos

2.1. Nomenclatura y conceptos biológicos

Los codones C**A**A y C**G**A difieren en un nucleótido (A→G), por lo que el SNP de interés (alelo 1: A; alelo 2: G), representado como c.578-44A¿G, está ubicado en el intrón 2 del gen *MYF5* «*myogenic factor 5 (Bos taurus)*» (NM_174116) (Li et al. 2004).

Este gen, localizado en el cromosoma 5 del *Bos taurus* y con una longitud de 3235 nucleótidos[2], está constituido por tres exones y ha sido previamente asociado con

[2]Ensamble Nov. 2009, `http://genome.ucsc.edu/cgi-bin/hgGateway`

características de crecimiento en animales de la raza Romosinuano, incluyendo el peso al nacer, el área de ojo del lomo a los 12 y 16 meses y la ganancia diaria antes del destete (Ríos et al. 2010).

De acuerdo con lo anterior y con los estándares actuales[3], la afirmación "...alelo '1' CAA y alelo '2' CGA..." y la nomenclatura "MYF-5" del gen en humanos (Wain et al. 2002) y bovinos no es técnicamente correcta. Además, teniendo en cuenta que un gen *codifica* para una proteína, que a su vez realiza una función específica en el organismo (Alberts et al. 2008), la afirmación "...MYF-5, que es el gen que codifica para el crecimiento muscular..." no es correcta.

2.2. Unidades de medida y tipo y codificación de variables

En la sección de Materiales y Métodos, el autor menciona la adición de "500 ml de una solución amortiguadora" y "...500 ml de tampón de digestión, 22ml de DTT 0.1M y 27 ml...", entre otras soluciones, para extraer ADN de sangre. Evidentemente, estos volúmenes son muy grandes comparados con los que se utilizan normalmente a escala de laboratorio de investigación. Consideramos que ni siquiera cambiando ml por μl, asumiendo que haya sido un error tipográfico, estos volúmenes tienen sentido. De manera similar, en "fue colocada a -20°C por 30 toda la noche" e "incubando a 37 C por 3 horas", creemos que el autor se refiere a "...-20°C toda la noche" e "incubando a 37°C por 3 horas", respectivamente. Por otro lado, al afirmar que para "...el análisis de los datos se usaron variables de tipo binomial, donde la variable dependiente fue el peso vivo de los animales...", el autor no expone con claridad si con "variables de tipo binomial" se refiere a las variables raza y genotipo, o al peso en kilogramos. Aunque las dos primeras pueden considerarse binomiales, el peso es una variable *continua* que fue posteriormente *discretizada*. En nuestra opinión, esta imprecisión genera confusión.

2.3. Cálculos adicionales

Al calcular el estadístico de Wald, el autor comete dos imprecisiones cuando afirma que "Para la variable RAZA, el valor de Wald es $2.6992 = 7.284$" y "Para la variable MYF-5, el valor de Wald es (-3.293)2 = 10.8439 ...". Primero, consideramos que el autor se refiere al *cuadrado* del estadístico de Wald. Segundo, las expresiones correctas para estos cálculos son 2.699^2 y $(-3.293)^2$, respectivamente.

[3] www.genenames.org/guidelines.html

2.4. Un modelo alternativo

Puesto que la categorización de variables continuas (respuesta o independientes) puede generar varias inconsistencias y pérdida de información en análisis de regresión (Cohen 1983, Taylor et al. 2006, Van Walraven & Hart 2008, Fedorov et al. 2009, Naggara et al. 2011), el autor podría utilizar un modelo de regresión lineal múltiple de la forma

$$\text{Peso} = \gamma_0 + \gamma_1 \text{ Raza} + \gamma_2 \text{ Genotipo} + \zeta \tag{2}$$

en lugar del modelo (1).

En este nuevo modelo la variable dependiente corresponde al peso del animal (en kilogramos), Raza y Genotipo a las variables definidas en el modelo (1), $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)$ al vector de parámetros y ζ al error aleatorio. Entre las ventajas del modelo (2) se encuentran su fácil interpretación y la comparación *directa* de los pesos promedio de los animales en función de la raza y genotipo.

En la tabla 2 presentamos los parámetros estimados para el modelo (2) utilizando los mismos niveles de referencia que en el modelo (1). Nuestros resultados indican que, en promedio, los animales de raza Romosinuano con genotipo 1/2 pesan ≈ 18 kg más que aquellos heterocigotos de la raza Cebú. El modelo ajustado es globalmente significativo ($F_{2,54} = 34294.41, p < 0.0001$) y posee un buen coeficiente de determinación ($R^2_{\text{adj}} = 0.542$). El análisis de residuales del modelo indica que estos tienen una distribución normal ($W = 0.993, p = 0.982$) y presentan un buen comportamiento (datos no presentados).

Tabla 2: *Parámetros estimados para el modelo de regresión lineal múltiple.*

Variable	$\hat{\gamma}$ (SE)	z	p
Intercepto	436.429 (2.627)	166.120	$< 10^{-16}$
Raza (Romosinuano)	18.765 (2.860)	6.561	$< 10^{-7}$
Genotipo (2/2)	-13.341 (2.770)	-4.816	$< 10^{-4}$

$F_{2,54} = 34294.41, p < 0.0001$; $R^2_{\text{adj}} = 0.542$, AIC = 433.771

3. Agradecimientos

Agradecemos a Ariel Martínez, Daniel Pineda-Álvarez y Alejandro Álvarez-Prats por sus valiosos comentarios y sugerencias.

Recibido: 2 de noviembre de 2011
Aceptado: 10 de noviembre de 2011

Referencias

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2008), *Molecular Biology of the Cell*, 5th edn, New York: Garland Science.

Camacho Quiroga, R. (2011), 'Asociación de polimorfismos genéticos de las razas Cebú y Romosinuano con el desarrollo muscular y peso vivo final', *Comunicaciones en Estadística* **4**(1), 63–71.

Canty, A. & Ripley, B. (2011), *boot: Bootstrap R (S+) Functions. R package version 1.3-2.*

Chongsuvivatwong, V. (2011), *epicalc: Epidemiological calculator.* R package version 2.13.2.2.
*http://CRAN.R-project.org/package=epicalc

Cohen, J. (1983), 'The Cost of Dichotomization', *Applied Psychological Measurement* **7**(3), 249–253.

Davison, A. C. & Hinkley, D. V. (1997), *Bootstrap Methods and Their Applications*, Cambridge University Press, Cambridge.

Fedorov, V., Mannino, F. & Zhang, R. (2009), 'Consequences of Dichotomization', *Pharmaceutical Statistics* **8**(1), 50–61.

Harrell, F. E. (2001), *Regression Modeling Strategies, with Applications to Linear Models, Survival Analysis and Logistic Regression*, Springer, New York.

Harrell, F. E. (2011), *rms: Regression Modeling Strategies.* R package version 3.3-1.
*http://CRAN.R-project.org/package=rms

Hosmer, D. & Lemeshow, S. (1989), *Applied Logistic Regression*, Jhon Wiley & Sons, New York.

Kuhn, M. (2011), *caret: Classification and Regression Training.* R package version 5.07-001.
*http://CRAN.R-project.org/package=caret

Li, C., Basarab, J., Snelling, W., Benkel, B., Murdoch, B., Hansen, C. & Moore, S. S. (2004), 'Assessment of positional candidate genes $myf5$ and $igf1$ for growth on bovine chromosome 5 in commercial lines of *Bos taurus*', *J. Anim. Sci.* **82**, 1–7.

Metz, C. E. (1978), 'Basic principles of ROC analysis', *Semin. Nucl. Med.* **8**(4), 283–98.

Naggara, O., Raymond, J., Guilbert, F., Roy, D., Weill, A. & Altman, D. (2011), 'Analysis by Categorizing or Dichotomizing Continuous Variables Is Inadvisable: An Example from the Natural History of Unruptured Aneurysms', *American Journal of Neuroradiology* **32**(3), 437–440.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org/

Ríos, R. M., Castro, S. L., Moreno, D. J., Moncaleano, J. S., Santana, M. O. & Ariza, B. F. (2010), 'Regiones del cromosoma 5 asociadas a características de crecimiento en ganado criollo Romosinuano', *Rev. Med. Vet. Zoot.* **57**, 35–47.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. (2011), 'pROC: an open-source package for R and S+ to analyze and compare ROC curves', *BMC Bioinformatics* **12**, 77.

Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. (2009), *ROCR: Visualizing the performance of scoring classifiers.* R package version 1.0-4.
*http://CRAN.R-project.org/package=ROCR

Taylor, A. B., West, S. G. & Aiken, L. S. (2006), 'Loss of Power in Logistic, Ordinal Logistic, and Probit Regression When an Outcome Variable Is Coarsely Categorized', *Educational and Psychological Measurement* **66**(2), 228–239.

Van Walraven, C. & Hart, R. (2008), 'Leave'em Alone − Why Continuous Variables Should Be Analyzed as Such', *Neuroepidemiology* **30**, 138–139.

Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J., Wright, M. W. & Povey, S. (2002), 'Guidelines for Human Gene Nomenclature', *Genomics* **79**(4), 464–470.