

MODELOS MULTINOMIALES: UN ANÁLISIS DE SUS PROPIEDADES

Arlen Guarín*
Andrés Ramírez**
Felipe Torres***

Recibido: 22/03/2011

Aceptado: 06/06/2012

RESUMEN

En este artículo de investigación se desarrolla un análisis de las propiedades de los modelos multinomiales a través de distintos procesos de simulación. Lo anterior se hizo asumiendo tanto el cumplimiento de los supuestos subyacentes de los mecanismos de estimación como el incumplimiento de los mismos. Igualmente se analizó el comportamiento de los estimadores bajo diferentes escenarios de tamaño muestral. Se encontró que bajo un modelo correctamente especificado y tamaños muestrales superiores a 200 observaciones se cumplen las propiedades de insesgadez y consistencia, mientras que la incorrecta especificación de la distribución del proceso lleva a estimaciones sesgadas e inconsistentes; de igual forma se encontró que en tamaños muestrales pequeños y bajo modelos condicionales se pierden las propiedades que una buena especificación del proceso suele generar, y se halla aún más inestabilidad cuando la estimación es llevada a cabo con la metodología Probit.

Palabras clave: modelos multinomiales, modelos condicionales, simulación, propiedades estadísticas.

* Economista de la Universidad EAFIT. Profesional Investigador, Banco de la República. Teléfono: 5767400 ext. 4469. e-mail: aguariga@banrep.gov.co.

** Doctor Estadística de la Universidad Nacional de Colombia, magíster en Finanzas de la Universidad EAFIT, magíster en Economía de la Universidad Nacional de Colombia, Economista de la Universidad Nacional. Profesor de tiempo completo, Escuela de Economía y Finanzas. Programa de Economía. Universidad EAFIT. E-mail: aramir21@eafit.edu.co.

*** Ingeniero Matemático de la Universidad EAFIT. Coordinador de desarrollo, Flores el tragal. E-mail: jtorres9@eafit.edu.co.

MULTINOMIAL MODELS: AN ANALYSIS OF THEIR PROPERTIES

ABSTRACT

In this research paper, we made an analysis of Multinomial models through simulation; this was done under correct and incorrect assumptions on the data generating process. Also, it was analyzed the performance of the models under different sample sizes. It was found that a correct specified model with samples of 200 or more observations achieves estimators which are unbiased and consistent, while incorrect assumptions about the data generating process causes biased and inconsistent estimators. On the other hand, conditional models with small sample sizes imply bad statistical properties, especially when Probit models are estimated.

Key words: multinomial models, conditional models, simulation, statistical properties.

INTRODUCCIÓN

Uno de los enfoques teóricos del campo económico que ha tomado más fuerza en las últimas décadas es la simplificación de la realidad a través de los modelos econométricos, los cuales surgen como una combinación entre la teoría económica y distintos desarrollos estadísticos y matemáticos. Dichos modelos buscan encontrar relaciones de causalidad entre un conjunto de variables específicas.

Dentro de los primeros modelos aplicados al estudio estadístico se encuentran los modelos lineales, utilizados en el análisis de causalidad entre una variable dependiente, la cual presenta soporte continuo en los reales, y una serie de variables independientes. Este enfoque es de gran aplicación en el campo macroeconómico, pero impone serias restricciones cuando del campo microeconómico se trata, pues muchas de las variables que se estudian allí tienen que ver con decisiones del individuo que están restringidas a una serie de opciones, lo cual implica que el soporte de la variable respuesta es un conjunto acotado de enteros. Específicamente, los modelos que surgen en este contexto son los denominados modelos multinomiales, los cuales toman su nombre debido a la función de densidad que se encuentra implícita en el proceso generador de datos.

El mecanismo de estimación usado por excelencia en este escenario es máxima verosimilitud, el cual fue definido por Fischer en 1922 [1], y que finalmente se traduce en encontrar una serie de parámetros, los cuales son la solución a un conjunto de ecuaciones que normalmente son no lineales. La intuición que respalda el mecanismo está definida a través de encontrar los parámetros de la función de verosimilitud, los cuales maximizan la probabilidad de replicar las observaciones.

En el caso de los modelos multinomiales, el proceso generador de datos es multinomial, y la relación entre los regresores y la variable respuesta se establece en términos de las probabilidades asociadas a la elección de las diferentes alternativas a las cuales se ven enfrentados los individuos, y la

parametrización de las mismas en términos de las variables control.

Previa a la utilización de dichos modelos, la principal herramienta empleada consistía en el análisis discriminante, introducido por Fisher en 1936 [2], quien, basado en las características conocidas de los individuos y sus diferencias, los clasifica dentro de grupos tratando de maximizar la homogeneidad de estos. Los modelos multinomiales son claramente métodos mucho más sofisticados pues brindan mayor información acerca de las relaciones entre las variables, y probablemente en estos se cometan menos errores al ser menos drásticos al momento de clasificar los individuos.

Al utilizar los modelos multinomiales, el analista suele encontrarse con ciertas restricciones que pueden afectar los resultados que su investigación arroje. Dentro de estas limitaciones se encuentra como una de las más comunes el desconocimiento del proceso generador de datos que subyace la variable latente implícita en la probabilidad de elección en un contexto multinomial. Este factor es determinante al momento de aplicar los métodos de máxima verosimilitud.

Es por eso que el análisis y conocimiento de las propiedades estadísticas del mecanismo de estimación fundamentado en máxima verosimilitud cobra vital importancia durante la utilización de cualquiera de estos modelos, y constituye el objetivo del presente trabajo. En particular nos concentraremos en análisis de las propiedades de insesgadez y consistencia de los modelos Probit y Logit, tanto multinomiales como condicionales. Esto se hará a través de la implementación de diferentes algoritmos de simulación.

1 MODELOS MULTINOMIALES

Los modelos multinomiales introducidos por McFadden en 1974 [3] son herramientas de análisis que siguen una distribución multinomial y utilizan el método de máxima verosimilitud para estimar las probabilidades asociadas a cada elección, considerando las características particulares

de los individuos o los atributos de las elecciones resumidas en los regresores.

Los distintos tipos de relación entre la variable dependiente y los regresores dan lugar a diferentes modelos tales como el modelo condicional, aplicado a variables independientes que varían entre opciones; el modelo multinomial, en donde los regresores varían con los individuos, y el modelo mixto, el cual es una combinación de variables con las dos características mencionadas anteriormente.

1.1 Especificación de los modelos

Dada una variable dependiente con un conjunto de m opciones, la probabilidad de que el individuo escoja la opción j estará definida por:

$$P_j = P(Y = j) = P(Y_j = 1) \quad (1)$$

Donde,

$$Y_j = \begin{cases} 1 & \text{si } Y = j \\ 0 & \text{si } Y \neq j \end{cases} \quad (2)$$

Asumiendo que los individuos seleccionan la alternativa que más utilidad les brinda, se puede expresar lo anterior como la probabilidad de que la utilidad de la alternativa j (U_j), sea mayor que las de todas las demás alternativas, donde dicha utilidad está dada por la suma de un componente determinístico (V_j) y otro estocástico (ε_j).

Cuando la estructura del componente V_j sigue la forma:

$$V_j = X_i B_j \quad (3)$$

Donde X_i es un vector de regresores que varía para cada individuo i y B_j es un vector de coeficientes que varía para cada opción j . Este es el modelo multinomial.

Por otra parte, si el componente determinístico se comporta de la siguiente forma:

$$V_j = X_{ij} B \quad (4)$$

Donde X_{ij} es un vector de regresores que varía para cada individuo i y para cada opción j , y B es un vector de coeficientes fijo para todas las opciones. Este es el modelo condicional.

Con respecto al componente estocástico ε_j , los supuestos que se hacen sobre su distribución llevan a que los métodos de estimación aplicables cambien, siendo el modelo Probit la herramienta recomendada para la estimación en presencia de perturbaciones estocásticas con distribución normal, y el modelo Logit, el cual aplica en presencia de perturbaciones estocásticas que siguen una distribución valor extremo.

2 METODOLOGÍA

Para analizar las distintas propiedades de los modelos condicional y multinomial bajo las tipologías Probit y Logit, en la presente investigación se llevó a cabo un proceso de simulación de datos basados en el modelo de utilidad aditiva aleatoria (ARUM por sus siglas en inglés) como lo referencian Cameron y Trivedi [4]. El tratamiento se fundamenta en el siguiente desarrollo.

$$\begin{aligned} P[Y = j] &= \Pr[U_j \geq U_k, \forall k \neq j] \\ &= \Pr[U_k - U_j \leq 0, \forall k \neq j] \\ &= \Pr[\varepsilon_k - \varepsilon_j \leq V_j - V_k, \forall k \neq j] \\ &= \Pr[\widetilde{\varepsilon}_{kj} \leq \widetilde{V}_{kj}, \forall k \neq j] \end{aligned} \quad (5)$$

En este sentido se hace posible la simulación de las distintas probabilidades a través del cálculo de la función de densidad acumulada de la distribución asumida para la diferencia de los errores, evaluada en cada valor \widetilde{V}_{kj} . Debe ser claro en este punto que la serie de diferencias con j 's fijos y k 's variables está correlacionada entre sí, por lo que al momento de calcular la densidad acumulada se debe tomar en consideración dicha correlación.

Para el caso de este trabajo, los V_j serán calculados a través de unos regresores generados aleatoriamente y unos coeficientes fijados arbitrariamente con la estructura específica para cada tipo de modelo condicional y multinomial; luego, se calcularán las funciones de densidad acumuladas para la perturbación estocástica compuesta $\widetilde{\varepsilon}_{kj}$ y será evaluada en \widetilde{V}_{kj} . En particular se utiliza una distribución normal, que cumple las propiedades del modelo Probit, y de una distribución logística, que cumple las propiedades del Logit, pues la diferencia de variables distribuidas valor extremo tipo I, da lugar a una variable que sigue una distribución logística.

Adicionalmente, se calculará la densidad acumulada de una distribución t student como herramienta de análisis del comportamiento de los modelos bajo errores de especificación en el proceso generador de datos. Para todas las simulaciones y estimaciones desarrolladas en este trabajo se utilizó el software *STATA 10*.

Una vez calculadas dichas probabilidades a través de los distintos procesos señalados anteriormente, se pasa a simular distintas variables generadas a través de una distribución multinomial con las probabilidades encontradas. Simuladas estas variables es posible entonces realizar el análisis de las propiedades de los estimadores ante el cumplimiento o incumplimiento de los supuestos, a través de técnicas de estimación con muestras crecientes, como una aproximación en muestras finitas de la propiedad de consistencia, y de replicación de muestras, dado un tamaño muestral para el análisis de la propiedad de insesgadez similares a los llevados a cabo por Cameron y Trivedi [5]. Por su parte Griffiths et al. [6] y Peyong et al. [7] analizaron dichas propiedades en muestras finitas a través de la simulación Monte Carlo para los modelos Probit Binomial y Logit Binomial, respectivamente.

Al momento de evaluar la bondad de un estimador en particular, son varias las propiedades estadísticas deseadas; sin embargo, el objetivo del

presente artículo es concentrarse en la insesgadez y la consistencia, puesto que el incumplimiento de estas propiedades implica la obtención de un mal estimador, aunque el cumplimiento de las mismas no garantiza que el estimador en cuestión sea óptimo. En particular, un estimador insesgado se caracteriza porque el valor esperado del mismo, ante muestreo repetido, es igual al valor poblacional. Por otra parte, un estimador consistente se caracteriza porque converge en probabilidad al valor poblacional, lo cual implica que el estimador es asintóticamente insesgado y la varianza del mismo converge a cero, lo cual significa que la distribución asintótica del parámetro estimado es degenerada, es decir, concentra toda su masa probabilística en un solo punto: el valor poblacional.

2.1 Características específicas de los datos

Para este trabajo se contará con una estructura de datos similar para todos los modelos a analizar; esta variará solo cuando la especificación de estos así lo precise. Se asumirán tres (3) regresores, los cuales se originan a partir de distribuciones normales con diferentes parámetros, y una variable dependiente que tomará valores dentro de tres (3) categorías. Bajo este marco, se hace necesaria la imputación de 9 coeficientes para el caso de los modelos multinomiales (3 categorías por 3 regresores) y 3 para el caso de los modelos condicionales. En el caso de los primeros es sabido que para garantizar la unicidad de la suma de las probabilidades (condición de identificación de los parámetros estimados), las estimaciones se hacen con respecto a una opción base, cuyos coeficientes se asumen iguales a cero, siendo los parámetros encontrados una aproximación a la diferencia entre estos y los de la categoría base.

Para la simplificación del análisis en esta investigación, los coeficientes asociados a la categoría 3 (B_{31} , B_{32} , y B_{33}), que tomaremos como base, serán creados desde el inicio con un valor de cero, para así poder realizar una interpretación más directa de los coeficientes estimados. Igualmente por sim-

plicidad, el análisis dentro de cada modelo girará en torno a solo uno de los coeficientes estimados; estos serán B_1 dentro de los modelos condicionales y B_{11} en los multinomiales, ambos con un valor fijo de 0.5 y 0.4, respectivamente.

Para el análisis de insesgadez se llevaron a cabo 1,000 simulaciones por cada modelo y cada tamaño de muestra, el cual a su vez va desde 50 a 10,000. Las distribuciones enseñadas hacen alusión a la densidad estimada bajo la aproximación de Epanechnikov [8] (azul) y una distribución normal con los momentos de los datos (rojo) como medida de comparación. Adicionalmente, las líneas punteadas hacen referencia al intervalo de confianza del 95% estimado; la línea roja indica el valor poblacional del parámetro estimado, y las líneas azules y verdes son introducidas cuando

las especificaciones del modelo simulado y sus respectivas simulaciones no convergen a los valores deseados.

3 RESULTADOS

3.1 Datos con distribución logística

Las siguientes figuras hacen referencia al modelo con ϵ_j distribuido error tipo I. Específicamente, la figura 1 da información sobre el modelo estimado a través de la especificación Logit, el cual a pesar de ser la especificación teóricamente correcta, sufre de sesgo en muestras pequeñas. Esto es corroborado por el hecho de que el coeficiente poblacional solo pasa a estar incluido en el intervalo de confianza del 95% a partir de muestras con más de 200 observaciones.

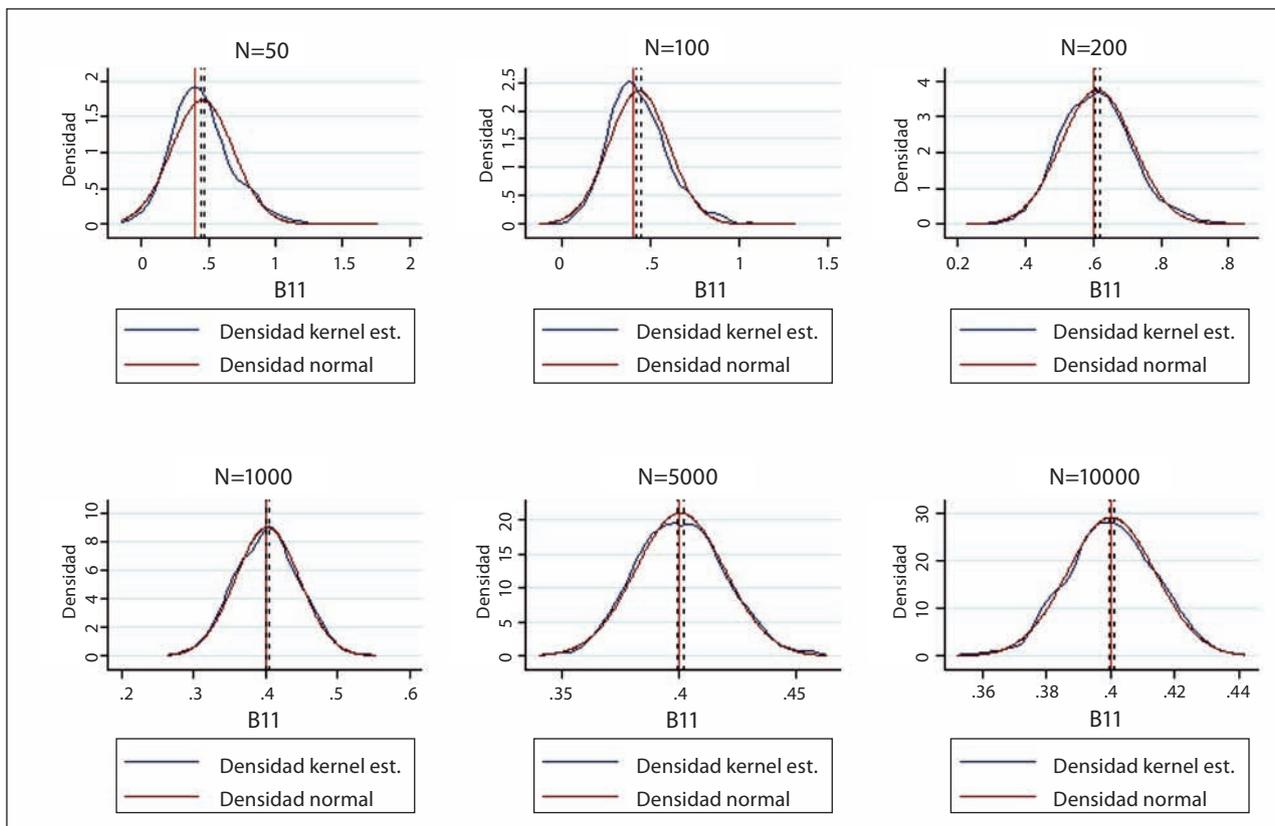


Figura 1. Modelo multinomial estimado por Logit (proceso generador de datos normal)

Fuente: elaboración propia

Es evidente cómo el aumento en el tamaño de la muestra hace las estimaciones cada vez más eficientes, reduciendo el ancho de banda de la distribución alrededor del valor poblacional y pareciéndose cada vez más a una distribución normal. Por el lado de la estimación con el modelo Probit (figura 2), el parámetro poblacional no cae nunca dentro del intervalo de confianza de las estimaciones tanto para muestras pequeñas como para muestras grandes. Cabe anotar que a medida que esta aumenta, la distribución de los coeficientes tiende a una normal cuya media tiende aproximadamente, a su vez, a 0.75 veces el parámetro poblacional. Este valor es razonable dada la relación teórica que se evidencia entre los parámetros estimados en un modelo Logit y un modelo Probit (Cameron y Trivedi [5]).

Un análisis similar al primero de los dos casos anteriores se puede extraer a partir de la figura 3, que expone el caso del modelo condicional estimado de nuevo a través de un Logit y evidencia de nuevo una convergencia hacia una distribución normal, pero en este caso con media $B_1 = 0.5$. Se observa que el coeficiente estimado se encuentra por fuera del intervalo de confianza en las estimaciones con muestras inferiores a 200 observaciones.

Las estimaciones del Probit, por el contrario, exhibieron un comportamiento disímil al llevado a cabo en el modelo multinomial, en especial en las muestras pequeñas, que evidencia un desempeño mucho más irregular, y presenta mucha más dispersión de los datos con valores superiores incluso a 2,000 veces el poblacional. En el caso de los tamaños de muestra 50 y 100, las figuras no representan

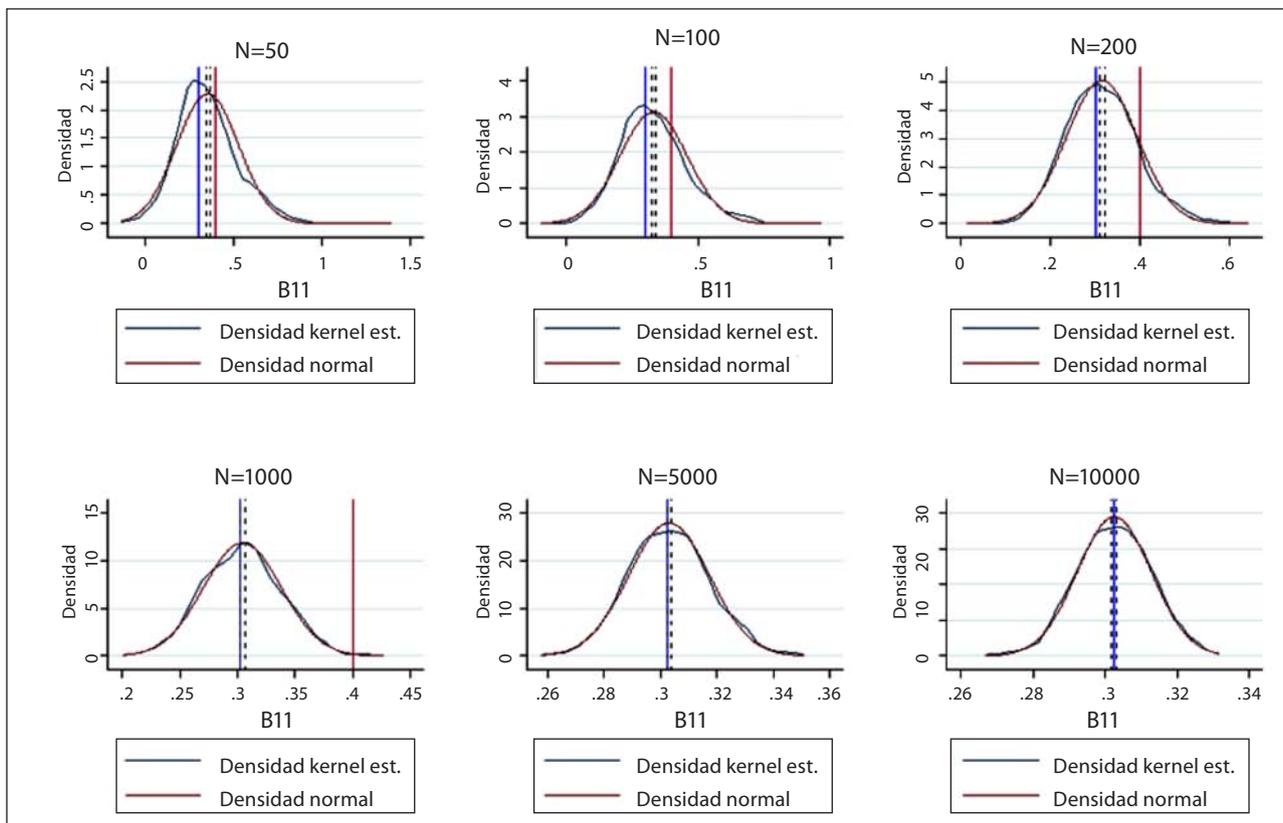


Figura 2. Modelo multinomial estimado por Probit (proceso generador de datos normal)

Fuente: elaboración propia.

el total de los valores de las estimaciones reales, sino las que, dentro de ellas, asumieran un valor inferior a uno, que corresponden al 87.9% y 97.7%, respectivamente. Las estimaciones marginadas aquí fueron aquellas que asumían valores desproporcionados (8.1% y 1.5%) o aquellas para las cuales no se encontró región de convergencia (4% y 0.8%). De nuevo, las estimaciones no convergen al valor poblacional sino a un valor que es 0.75 veces este, aproximadamente. Ver figura 4.

Análogamente, las figuras 5 y 6 facilitan un análisis aproximado a la propiedad de convergencia de los modelos condicional y multinomial generados con diferencias de errores distribuidos con valor extremo tipo I, bajo el aumento progresivo de la muestra, donde t es una variable que ayuda a indexar el número de observaciones con las que

fue realizada cada estimación siendo estas: $N = 2^{t+3}$. De esta forma, la estimación de cada punto de estas figuras habrá necesariamente utilizado el doble de observaciones que su inmediatamente anterior, excepto para el último de ellos, que está acotado a cien mil observaciones. Los valores de t para los que no haya estimaciones son aquellos para los cuales el algoritmo numérico no pudo encontrar una solución.

Para las figuras presentadas anteriormente, vemos cómo en la 5, que representa el caso de los modelos multinomiales, todas las estimaciones del modelo Probit guardan una proporción directa con respecto a las del modelos Logit, y las estimaciones por ambos métodos se estabilizaron a partir de 256 observaciones, presentándose una aproximación a la propiedad de consistencia solo en las del Logit,

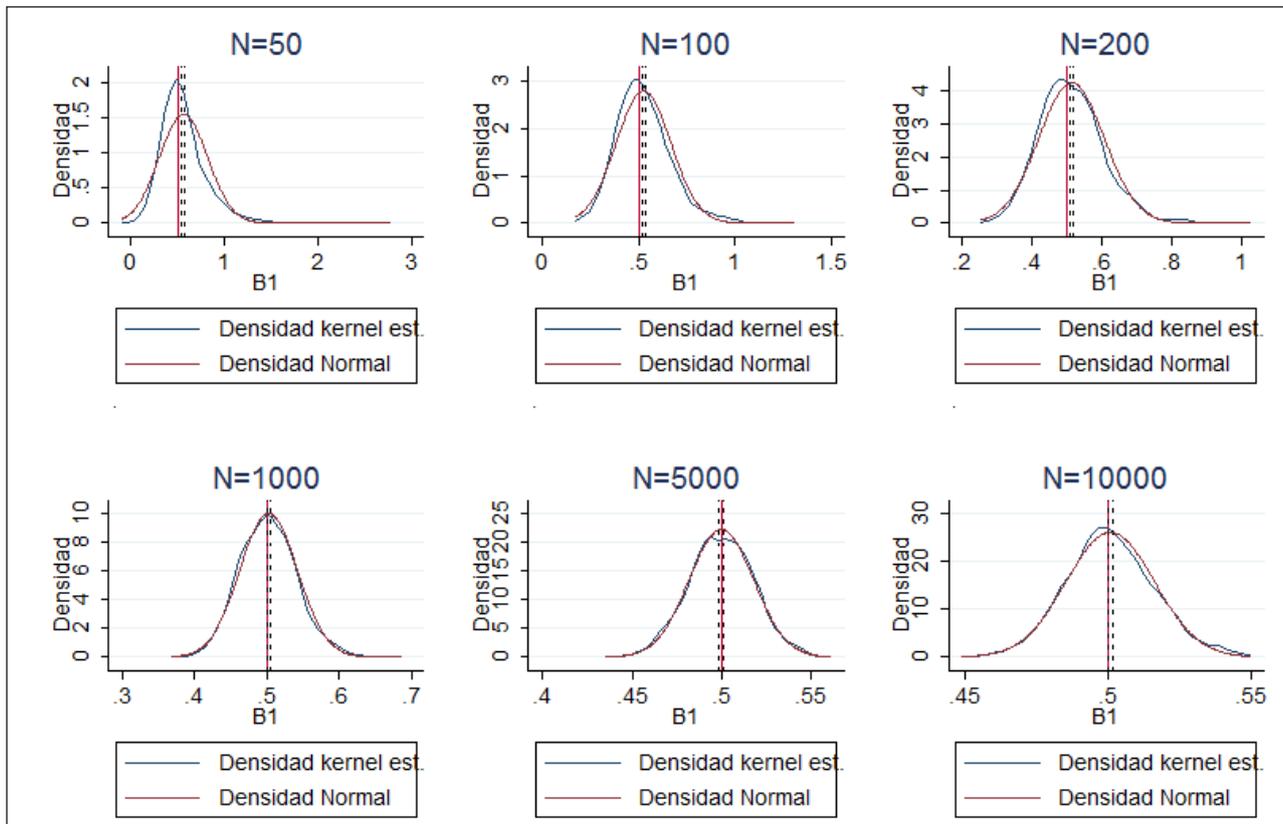


Figura 3. Modelo condicional estimado por Logit (proceso generador de datos normal)

Fuente: elaboración propia.

mientras que el parámetro al que debería converger el Probit fue 1.355 veces el valor al que convergió realmente.

Dentro de las únicas diferencias que tendría el análisis del modelo condicional estimado a través de las dos metodologías, se encuentra el hecho de que aquí la proporción entre los dos coeficientes

no es observable para tamaños de muestras pequeños, siendo el modelo Probit bastante inestable en ese tipo de situación. El Logit es aparentemente consistente mientras que el Probit converge a un valor que es 0.75 veces el poblacional.

En general se puede observar que la varianza estimada de los parámetros converge a cero, lo cual,

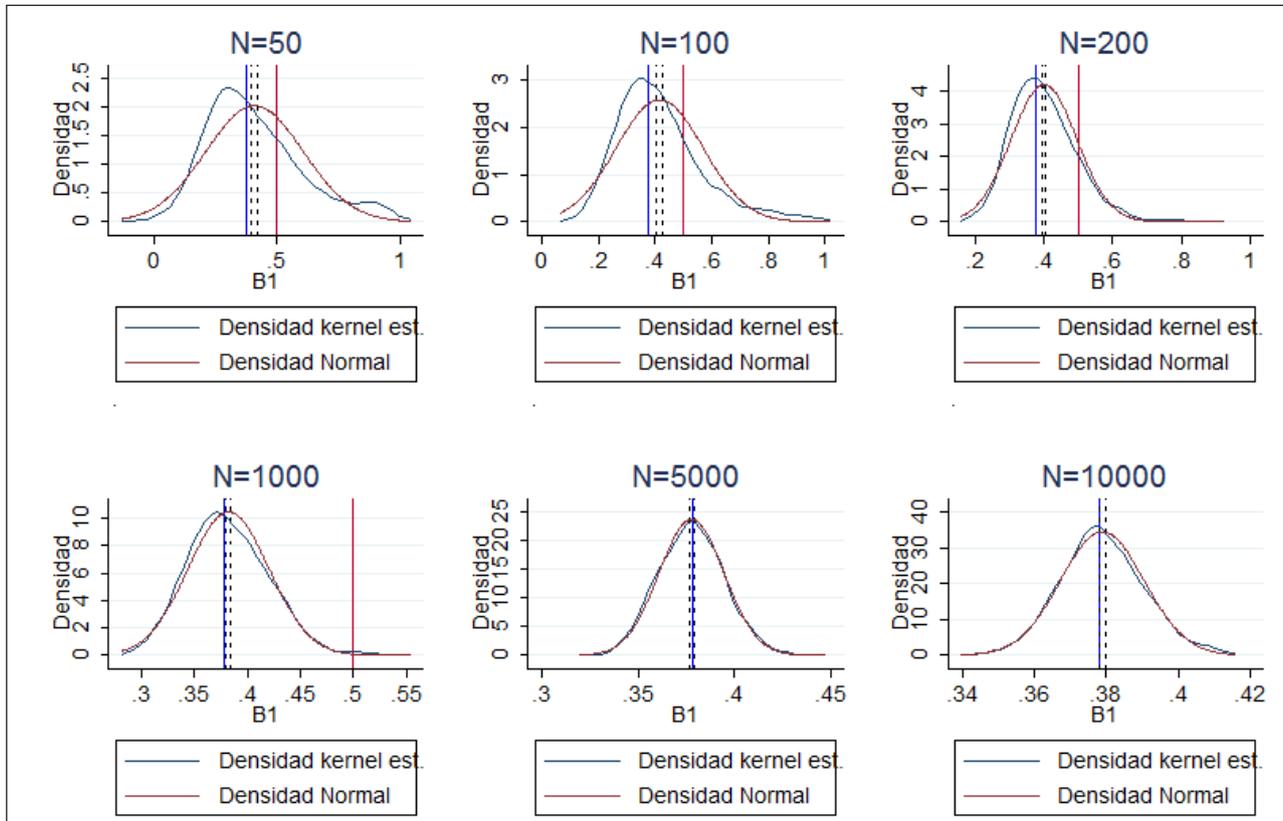


Figura 4. Modelo condicional estimado por Probit (proceso generador de datos logístico)

Fuente: elaboración propia.

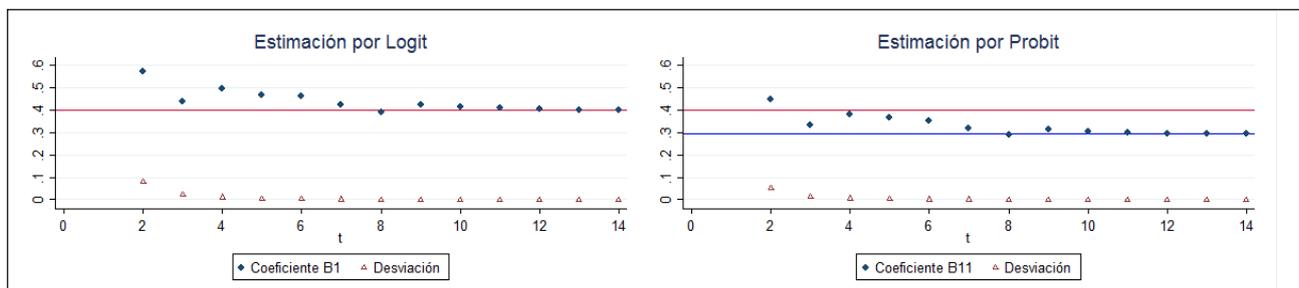


Figura 5. Aproximación a la propiedad de consistencia modelo multinomial (proceso generador de datos normal)

Fuente: elaboración propia.

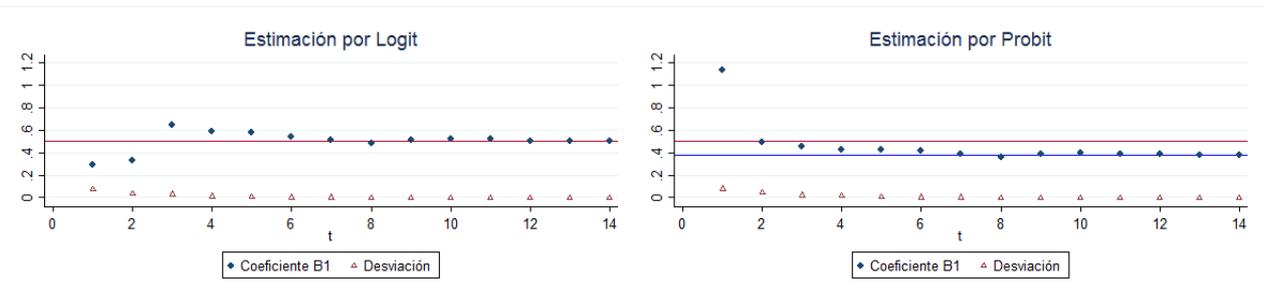


Figura 6. Aproximación a la propiedad de consistencia modelo Condicional (proceso generador de datos normal)
Fuente: elaboración propia.

unido a la insesgadez del modelo Logit, implica la aparente consistencia del parámetro estimado.

3.2 Datos con distribución normal

La figura 7 nos enseña ahora el caso de datos con errores distribuidos normal, pero estimados a

través del modelo Logit, donde eran de esperar los resultados obtenidos, dado que la especificación teóricamente recomendada sería la encontrada a través de los modelos Probit, donde el valor poblacional nunca cae dentro del intervalo de confianza del 95%. Esto pese a que las estimaciones de este

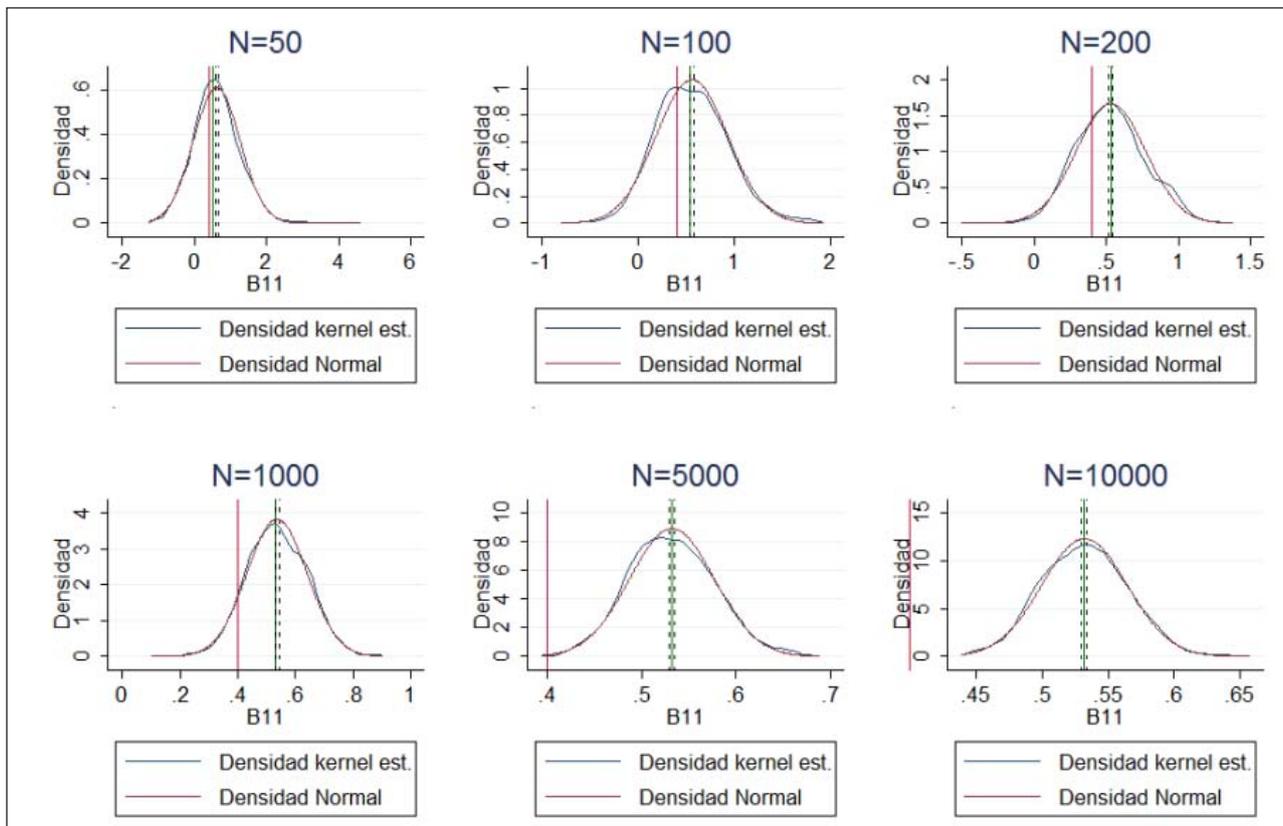


Figura 7. Modelo multinomial estimado por Logit (proceso generador de datos normal)

Fuente: elaboración propia.

modelo tienden a concentrarse a medida que aumenta el tamaño de la muestra, alrededor de un valor que es 1.33 veces aproximadamente el del coeficiente poblacional, valor que está representado por la línea verde; además, las estimaciones tienden a una distribución normal. Por otro lado, el modelo estimado a través de una especificación Probit tiene un mejor comportamiento encontrándose el valor poblacional muy cerca al intervalo de confianza en muestras iguales o inferiores a 100, y siempre dentro de este intervalo para muestras con mayor tamaño. De forma análoga, la distribución de los parámetros estimados tiene cada vez menos dispersión alrededor del valor poblacional, y tiende a una distribución normal. Este caso se puede observar en la figura 8.

Para el caso de los modelos condicionales, el estimado por Logit (figura 9) presenta caracterís-

ticas similares al multinomial: también es sesgado a la derecha, converge en muestras con tamaños grandes a una distribución normal con una media que es 1.311 veces el parámetro poblacional pero con un comportamiento muy irregular en muestras pequeñas. De esta forma se acotó la figura con 50 observaciones a los coeficientes inferiores a 2, y contiene ella el 98.1% de todos los coeficientes estimados. Este último problema también lo presenta el modelo Probit (figura 10), que aunque es insesgado y tiende a una distribución normal para muestras grandes, para las que no lo son, e incluso para las de 1,000 observaciones, el intervalo de confianza no contiene al B_1 poblacional, es más sesgado, presenta más valores atípicos y se distancia bastante de una distribución normal a medida que el tamaño muestral disminuye. Los gráficos con 50 y 100 observaciones están acotadas por 5, y

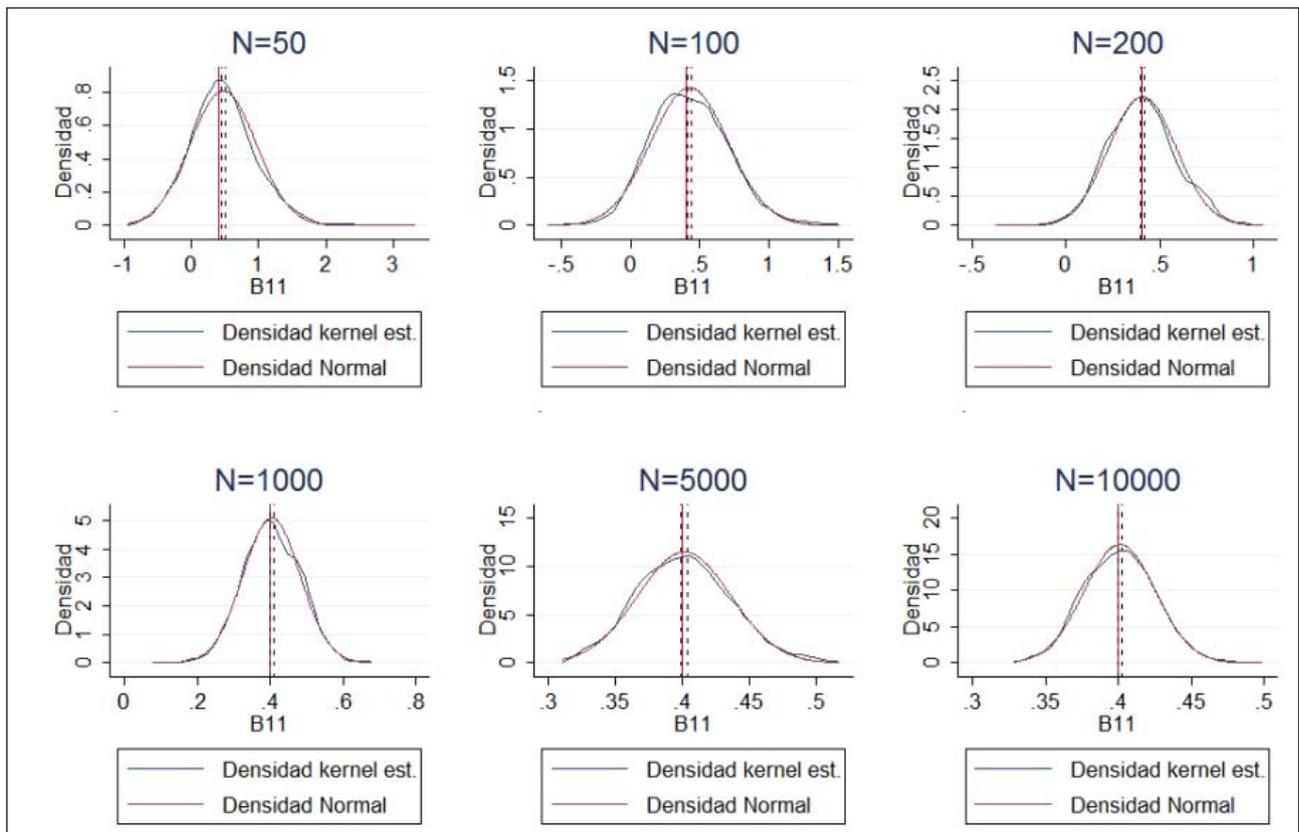


Figura 8. Modelo multinomial estimado por Probit (proceso generador de datos normal)

Fuente: elaboración propia.

contienen el 87.4% y 99.2% de los mil coeficientes estimados.

En cuanto a la aproximación de la propiedad de consistencia de este modelo vemos cómo, por el lado del modelo multinomial (figura 11), los resultados de uno son una proporción directa del otro, y se estabilizan en su punto de convergencia solo a partir de muestras con 16,000 observaciones, en las que converge la estimación Probit al valor poblacional y la estimación Logit a 1.306 veces dicho valor. Igualmente se observa que la varianza del modelo converge a cero, lo cual arroja indicios de consistencia.

Dicha proporcionalidad también se guardó para el caso del modelo condicional (figura 12) con la diferencia de que para el modelo Probit, el cual es aparentemente consistente, el algoritmo numérico no encontró solución para la primera

estimación. Aquí también, el valor de convergencia del modelo Logit es de 1.306 veces el poblacional, y las estimaciones se estabilizaron de nuevo solo a partir de 16,000 observaciones.

3.3 Datos con distribución t-student

Para el caso de la distribución t-student se incumplen los supuestos de distribución tanto del modelo Logit como del modelo Probit por lo que el análisis del comportamiento presenta varias particularidades.

Como se observa en las figuras 13 y 14, los resultados obtenidos, tanto por el Logit como el Probit Multinomial, son sesgados y las formas de sus distribuciones en cada tamaño muestral son muy similares, y ambas convergen a una normal. La media del Logit es 1.13 veces el coeficiente poblacional, y para el Probit 0.75.

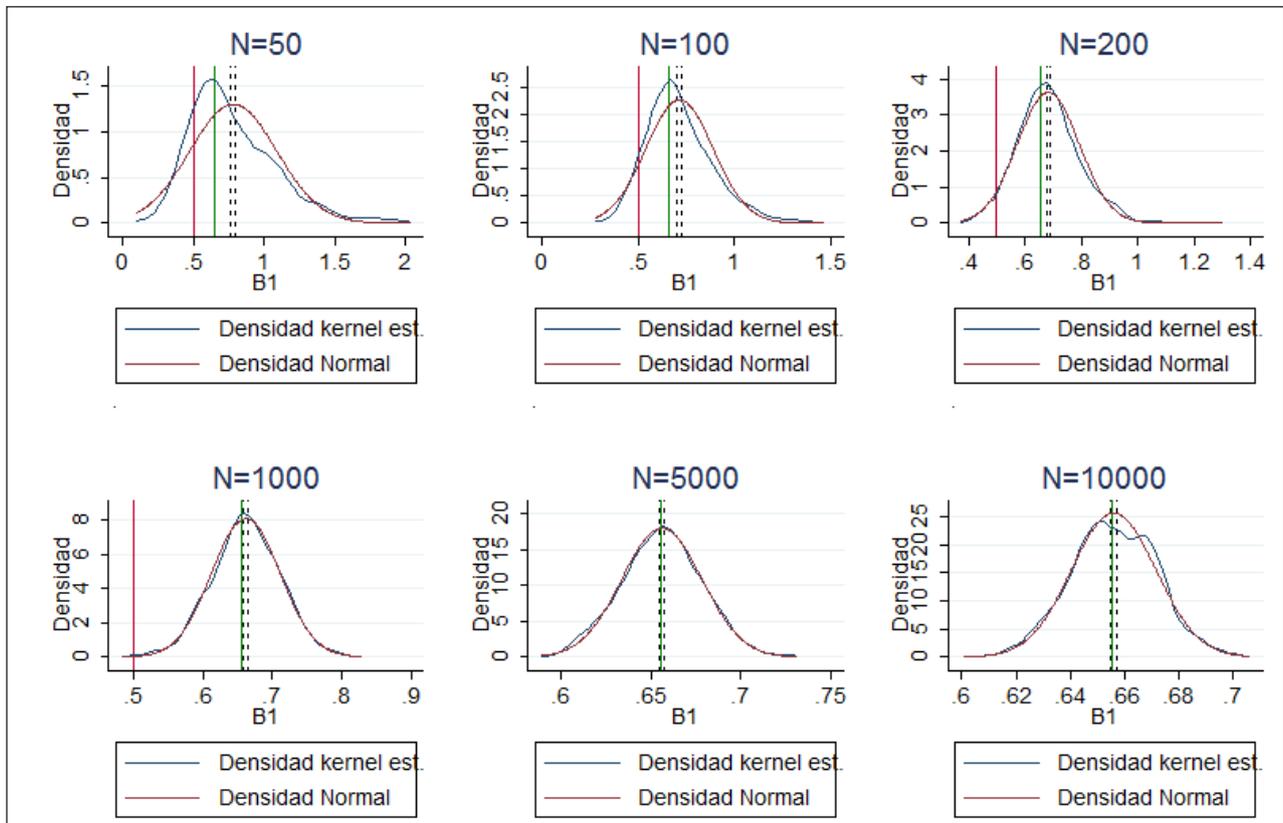


Figura 9. Modelo condicional estimado por Logit (proceso generador de datos normal)

Fuente: elaboración propia.

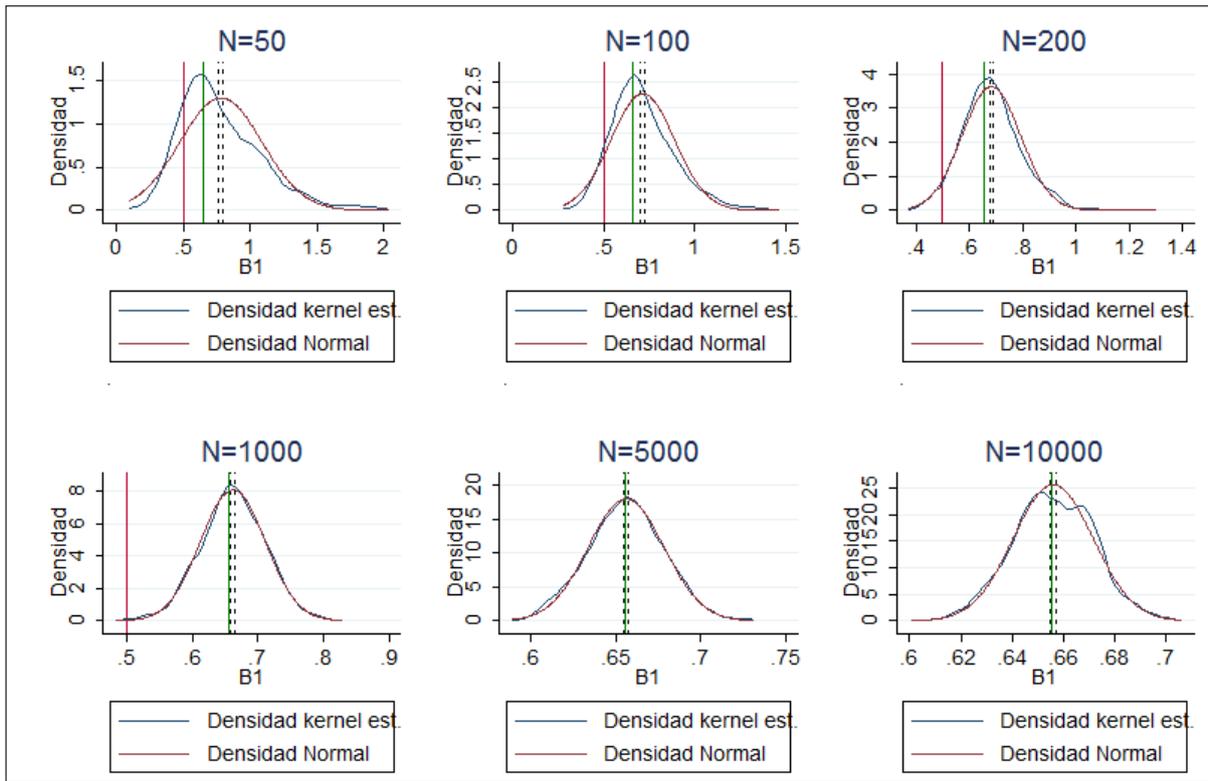


Figura 9. Modelo condicional estimado por Logit (proceso generador de datos normal)

Fuente: elaboración propia.

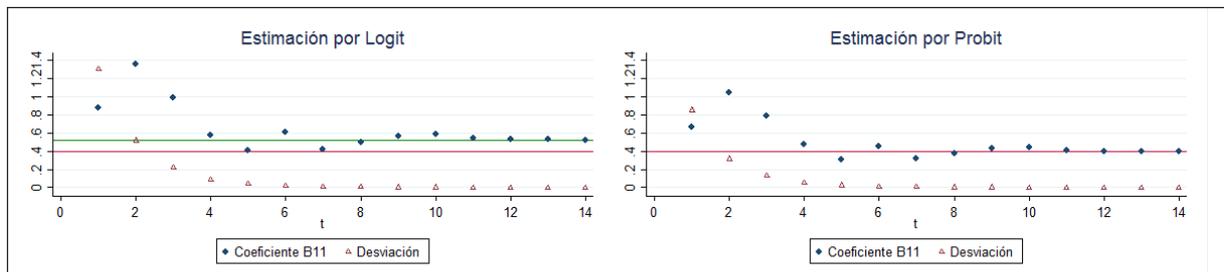


Figura 11. Consistencia modelo multinomial (proceso generador de datos normal)

Fuente: elaboración propia.

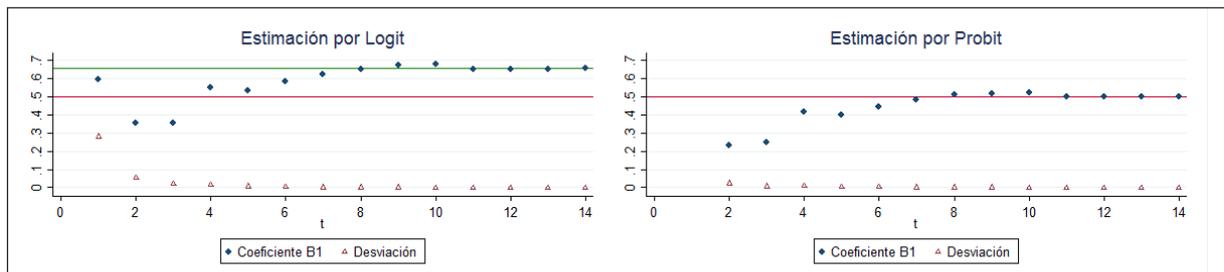


Figura 12. Consistencia modelo condicional (proceso generador de datos normal)

Fuente: elaboración propia

El anterior análisis también aplica para el caso condicional pero se presentan ciertas diferencias para el caso del Probit (figura 16) con tamaños de muestra 50 y 100. Sus estimaciones son mucho más inestables y los gráficos solo representan el 90.4% y el 97.3% de los 1,000 coeficientes respectivamente. Esto obedece a que el 2.8% y 2.3% de los resultados de las estimaciones no convergieron, y el restante 6.8% y 0.4%, fueron datos mayores que 5, los cuales no fueron tenidos en cuenta para facilitar el análisis gráfico. Aquí, el modelo Logit (figura 15) tiende en promedio un valor que es 1.27 el poblacional, y 1.37 veces el promedio que se evidencia bajo la especificación Probit (figura 16).

Finalmente, las figuras 17 y 18 muestran cómo ambas estimaciones, Logit y Probit, tanto para el modelo condicional como para el multinomial, son aparentemente inconsistentes; se presenta de nuevo

una proporcionalidad entre las estimaciones por la dos metodologías, en la cual el Logit es 1.46 veces el B_{11} poblacional y 1.36 veces el promedio estimado a través de la metodología Probit; esto para el caso del modelo multinomial. Para el modelo condicional, se encuentra que el parámetro estimado a través del supuesto logístico es 1.27 veces el B_{11} poblacional y 1.38 veces el obtenido con el modelo Probit. Se debe anotar también cómo en el caso del Probit en el modelo multinomial, pese a que converge a un valor muy cercano al poblacional (1.07 veces B_{11}), no es exactamente ese valor, por lo que, al menos para tamaños de muestra inferiores a 100,000 observaciones, el modelo sigue siendo aparentemente inconsistente. Para facilitar el análisis gráfico, los coeficientes asociados a muestras de tamaño 16 para el multinomial, y 16 y 32 para el condicional, no fueron tenidos en cuenta, pues su valor era de

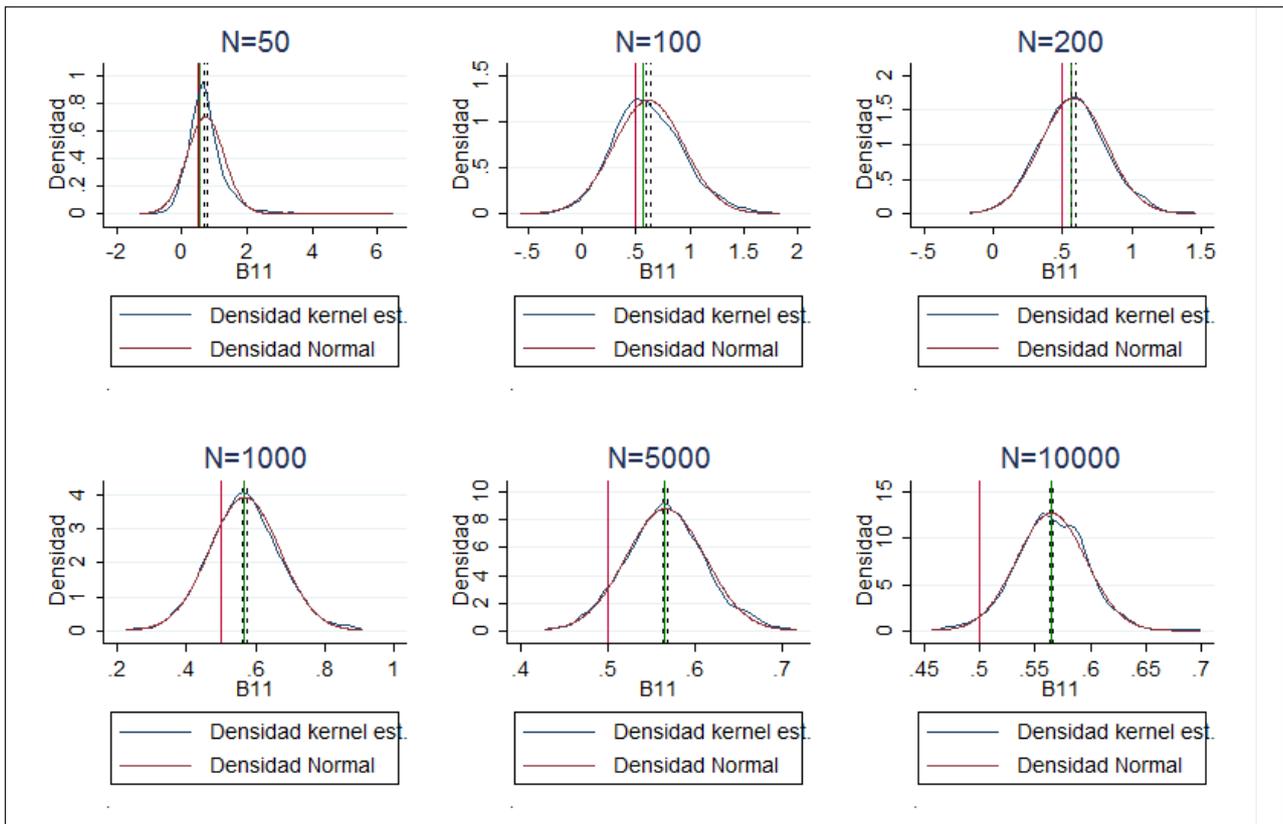


Figura 13. Modelo multinomial estimado por Logit (proceso generador de datos t-student)

Fuente: elaboración propia.

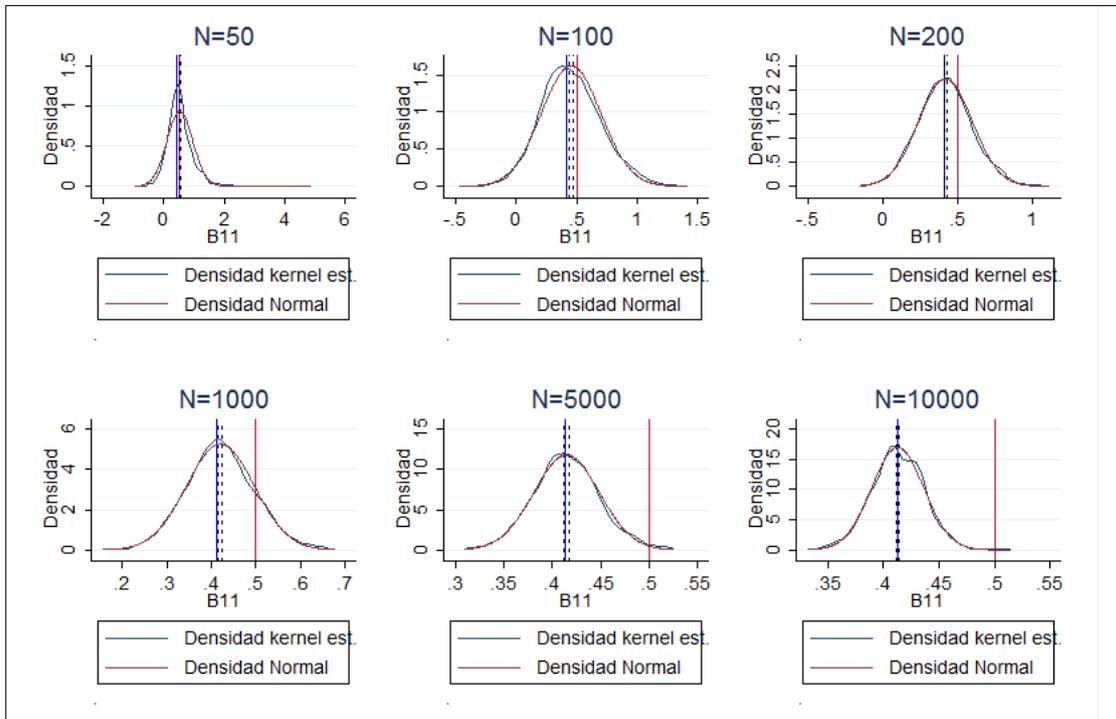


Figura 14. Modelo multinomial estimado por Probit (proceso generador de datos t-student)

Fuente: elaboración propia.

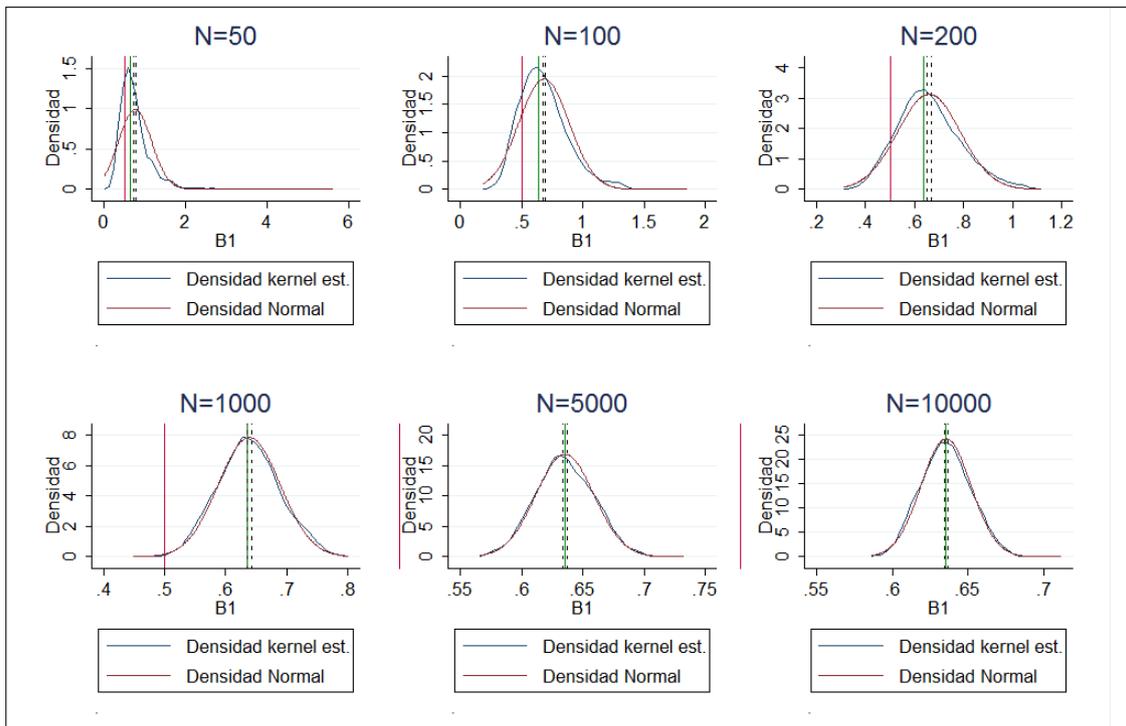


Figura 15. Modelo condicional estimado por Logit (proceso generador de datos t-student)

Fuente: elaboración propia.

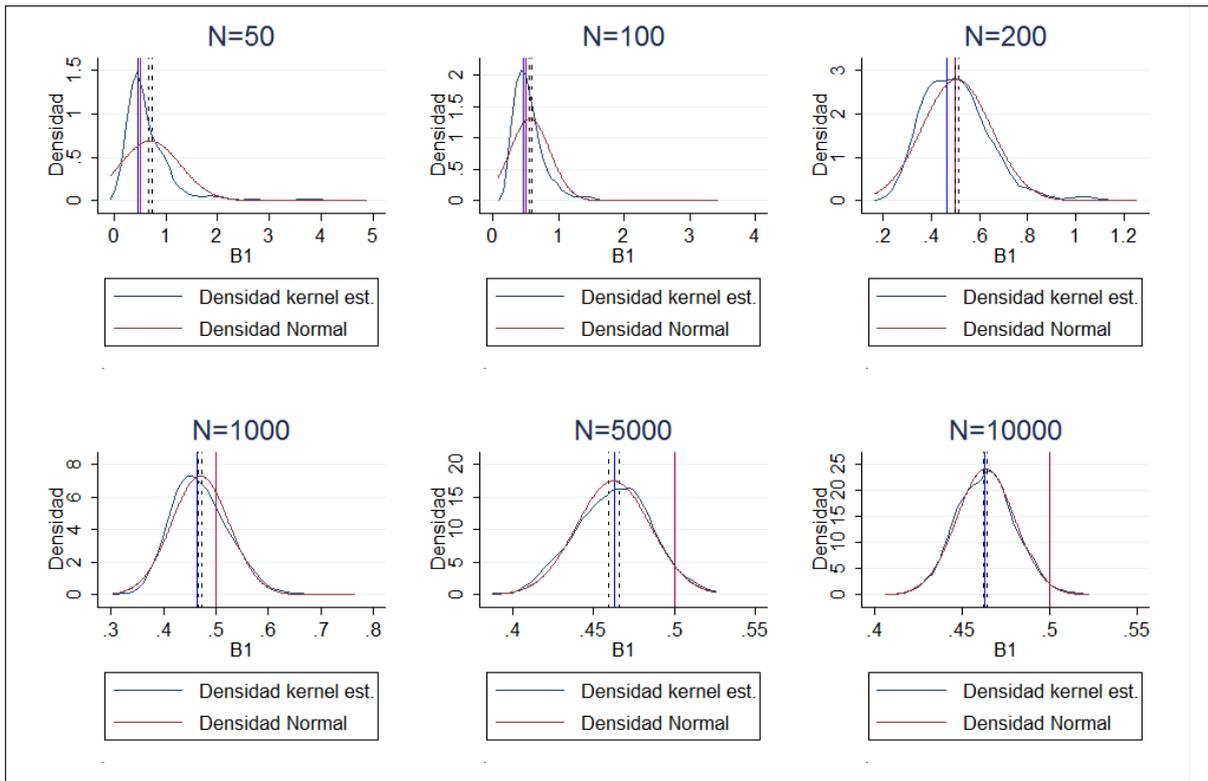


Figura 16. Modelo condicional estimado por Probit (proceso generador de datos t-student)

Fuente: elaboración propia.

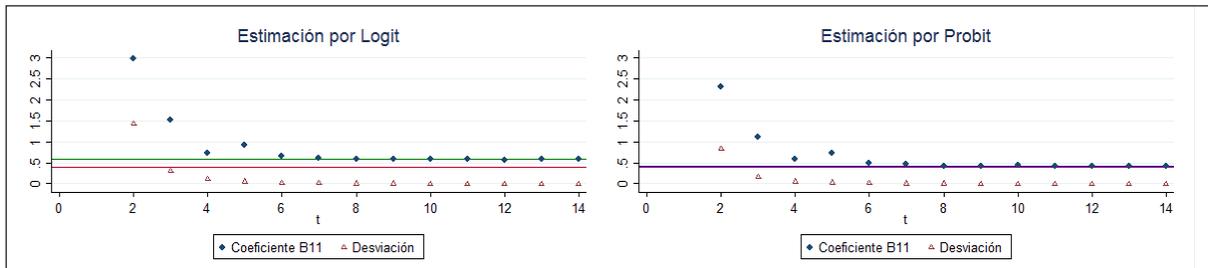


Figura 17. Consistencia modelo multinomial (proceso generador de datos t-student)

Fuente: elaboración propia.

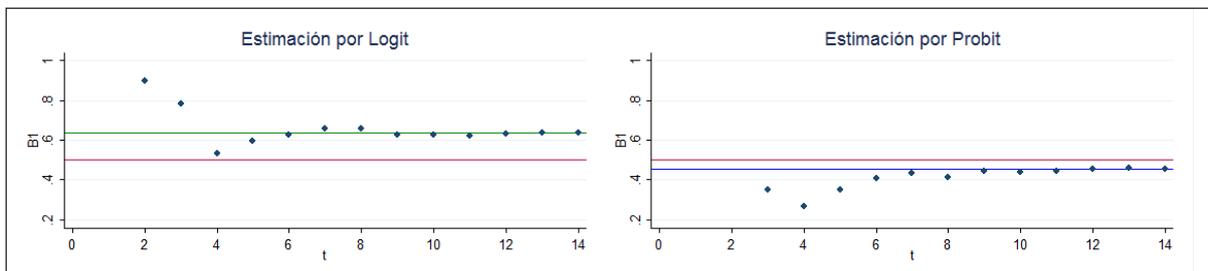


Figura 18. Consistencia del modelo condicional (proceso generador de datos t-student)

Fuente: elaboración propia.

masiado alto o el algoritmo no encontró solución (caso condicional para 16 observaciones).

4 CONCLUSIONES

En la utilización de los modelos multinomiales y condicionales, resultó ser completamente determinante la correcta especificación de la función de distribución que sigue la diferencia de los errores; la utilización de una distribución incorrecta implicará obtener estimaciones de los parámetros sesgadas y aparentemente inconsistentes.

El problema de la limitación en la cantidad de datos disponibles no tiene grandes efectos en los modelos multinomiales, y menos aún, cuando se trata de estimaciones a través del método Logit; la utilización de muestras con un tamaño superior a 200 observaciones, bajo una especificación acertada de los errores, lleva en general a una estimación insesgada de los coeficientes. Esta insesgadez para muestras de 100 o menos observaciones se cumple pero con niveles de confianza más bajos.

Para los modelos condicionales, las restricciones de tamaño muestral sí cobran mucha importancia, en especial cuando se trata de estimaciones bajo la metodología Probit, para la cual se encontró que con muestras cercanas o inferiores a 100 observaciones, se llega a estimaciones sesgadas de los coeficientes, incluso cuando hay una correcta especificación de la distribución de los errores, es decir, cuando estos se distribuyen normalmente. La metodología Logit en muestras pequeñas obtiene mejores resultados, pues es menos sesgada, y hay menor probabilidad de valores atípicos y de no convergencia de los procesos numéricos de estimación.

Bajo la presencia de diferencia de errores distribuidos t-student, aunque ambas metodologías Logit y Probit obtienen estimadores sesgados y aparentemente inconsistentes, en los modelos

multinomiales, el modelo Probit logra aproximarse mucho más a una estimación consistente de los coeficientes.

Para muestras superiores a 200 observaciones, los coeficientes estimados por la metodología Logit resultaron ser en general entre 1.3 y 1.4 veces los estimados por el Probit.

REFERENCIAS

- [1] R. A. Fisher, "On the mathematical foundations of the theoretical statistics," *Philosophical Transactions of the Royal Society of London, Series A*, Vol. 222, pp. 309-368, 1922 .
- [2] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of ugenics*, Vol. 7, pp. 179-188, 1936.
- [3] D. McFadden, "Conditional Logit analysis of qualitative choice behavior," en P. Zarembka (Ed.), *Frontiers in Econometrics*, pp. 105-142, 1974.
- [4] A. C. Cameron y P. K. Trivedi, "Multinomial Models," en *Microeconometrics, Methods and Applications*, Estados Unidos de América: Stata Press, 2009, pp. 113-146.
- [5] A. C. Cameron y P. K. Trivedi, "Multinomial Models," en *Microeconometrics using Stata*, New York: Cambridge University Press, 2005, pp. 490-528.
- [6] W. Griffiths, R. Carter y P. Pope, "Small Sample Properties of Probit Model Estimators," *Journal of the American Statistical Association*, Vol. 82, No. 399, pp. 929-937, 1987.
- [7] K. Peyong, K. Jong y C. Joong, "Small sample properties of generalized Logit model estimators with bootstrap," *Journal of Applied Mathematics and Computing*, Vol. 3, pp. 253-263, 1996.
- [8] V. A. Epanechnikov, "Nonparametric estimation of a multivariate probability density", *Theor. Prob. Appl.*, Vol.14, pp. 153-158, 1969.