UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de grado
Diciembre 2019

COMPARACIÓN DE EFICIENCIA ENTRE MODELOS LOGÍSTICOS MULTINOMIALES A PARTIR DE LA APLICACIÓN DEL TEST DE VUONG

EFFICIENCY COMPARISON BETWEEN MULTINOMIAL LOGISTIC MODELS FROM THE APPLICATION OF VUONG TEST

Diana Cristina Benavides Fonseca[a]
dianabenavidesf@usantotomas.edu.co

Wilmer Pineda Rios[b]
wilmerpineda@usantotomas.edu.co

Resumen

El presente trabajo propone una comparación de eficiencias entre diferentes modelos de regresión logística, teniendo como base la aplicación del test de Vuong. La idea es encontrar dentro de la oferta existente, el modelo que presente el mejor ajuste, partiendo del trabajo realizado sobre datos no anidados y realizando la implementación por medio de simulaciones.

***Palabras clave*:** Test de Vuong, ajuste, simulación.

Abstract

This paper proposes a comparison of efficiencies between different logistic regression models, based on the application of the Vuong Test. The idea is to find within the existing offer, the model that presents the best fit, based on the work done on non-nested data and carrying out the implementation through simulations.

***Keywords*:** Vuong test, adjustment, simulation.

1. Introducción

La teoría de Modelos Lineales Generalizados cuenta con una amplia gama de distribuciones asociadas a la familia exponencial, la cuales fueron unificados de acuerdo a la similitud de sus características y como una extensión al tradicional Modelo de Regresión Lineal. Cuando Nelder y Wedderburn en 1972 [1] propusieron la primera aproximación de los Modelos Lineales Generalizados, lo realizaron con base en la metodología de mínimos cuadrados ponderados para la estimación de máxima verosimilitud, que involucran una distribución de probabilidad asociada, un predictor lineal y una función de enlace como instrumentos de trabajo. En el modelo lineal original, se pretende realizar una predicción sobre una variable respuesta que viene atada a datos de tipo continuo; en la generalización, se puede decir que modelos como el Poisson y el Exponencial comparten estas características, sin embargo, la distribución de los datos es diferente con relación al primero y obliga a utilizar otras técnicas para evaluar su bondad de ajuste, su poder predictivo y su potencia frente a otros modelos.

[a]Estudiante Facultad de Estadística, Universidad Santo Tomás
[b]Docente Facultad de Estadística, Universidad Santo Tomás

[1]Moutinho, G.; Demétrio, C. (2008). Modelos Lineares Generalizados e Extensões. Departamento de Estatística e Informática, UFRPE.

Así como existen modelos con base en variables continuas, también se pueden encontrar variables respuesta de tipo cualitativo. La representación más simple se encuentra en variables dicotómicas, enmarcadas en ejemplos típicos como diagnóticos médicos (cuando se evalúa la presencia o ausencia de una enfermedad) o temas de tipo financiero (cuando se identifica un cliente moroso o no). Por otra parte, se pueden encontrar variables que involucran varias categorías y que por su naturaleza pueden ser de tipo nominal u ordinal, como por ejemplo la afinidad a un movimiento político (izquierda, centro o derecha), el estrato socioeconómico o religión profesada (católico, cristiano, judío). Todo lo anterior tiene como base la distribución Binomial y esta hace parte del conjunto de distribuciones que se enmarcan en los Modelos Lineales Generalizados a través de la representación de Regresiones Logísticas que se ajustan por medio de la función de enlace *Logit*.

El presente trabajo se centrará en la Regresión Logística Multinomial, dada la naturaleza de los datos, se realizarán comparativos entre ellos en los cuales se pretende encontrar el "mejor modeloz finalmente identificar si las diferencias entre variables de tipo nominal u ordinal generan alteraciones en los resultados y en las interpretaciones.

Para este trabajo se utilizará como herramienta principal el **Test de Vuong**, el cual parte del Criterio de Información de **Kullback Leibler** y se utilizarán modelos no anidados. Al final del documento se presentarán resultados a partir de simulaciones de datos.

Algunas aproximaciones al caso de estudio que se propone fueron realizadas con diferentes tipos de distribuciones y aplicadas a varios problemas. Por ejemplo, el estudio hecho por Marquez, Díaz y Ortiz en 2016 [2] en donde se realizó una comparación de tasas de viaje entre algunas cuidades colombianas, involucró la utilización de modelos de regresión lineal como parámetro de comparación, los cuales se construyeron a partir de los datos recolectados en encuestas domiciliarias y el Departamento Administrativo Nacional de Estadística. Este estudio tuvo como finalidad identificar la igualdad en las tasas de viaje entre las ciudades analizadas y las variables involucradas, por medio de la construcción de regresiones lineales que permitieron realizar una validación con respecto a la bondad de ajuste y una prueba t que les sirvió para comparar si las tasas de viaje eran transferibles entre ciudades o no. Pese al ejercicio realizado, el centro del estudio se genera con base en los resultados obtenidos en términos de variables, lo que difiere frente al presente trabajo, pues no se evaluó la eficiencia de los modelos construidos, estos solamente fueron utilizados como mecanismos de generación de resultados.

Por otra parte, Acuña, Dominguez y Toro en 2012 [3] proponen una comparación entre las técnicas tradicionales para la selección de modelos de regresión, las cuales van desde la selección de variables, hasta la minimización del error cuadrático medio, el cumplimiento de los supuestos sobre los residuales y el principio de parsimonia, contra el algoritmo genético Chu-Beasly. El objetivo de este trabajo era demostrar que métodos de selección basados en algoritmos genéticos presentaban un mejor desempeño que los tradicionales (AIC, BIC, SIC), a partir de la comparación de las regresiones construidas. La metodología implementada basó sus resultados en la medición de la eficiencia de las técnicas de selección de variables para conseguir un óptimo rendimiento del modelo y por ende mejorar su potencia y su capacidad predictiva.

En una aproximación a las aplicaciones de la regresión logística, Ladino en 2014 [4] presenta una comparación entre este modelo y las Redes Neuronales; el objetivo del estudio consitió en determinar cual de los dos presentaba un mayor nivel de discriminiación entre los clientes "buenosz "malos"de una entidad financiera para el producto tarjeta de crédito, de esta manera se tomarían decisiones sobre

[2] "Transferibilidad geográfica de modelos de generación de viajes urbanos: Comparación de modelos de regresión y tasas de viaje para algunas ciudades colombianas"
[3] "Una comparación entre modelos estadísticos clásicos y técnicas metaheurísticas en el modelamiento estadístico".
[4] "Comparación de Modelos de Riesgo de Crédito: Modelos Logísticos y Redes Neuronales".

posibles incrementos en el cupo de endeudamiento otorgado. La comparación de modelos se hace evidente nuevamente en este artículo, sin embargo, se parte de tipos diferentes, donde las pruebas de bondad de ajuste y predicción se hacen por separado y de acuerdo a la parametrización de cada uno de ellos. Un caso similar expone Lizares (2017), al realizar la comparación entre la eficiencia de la regresión logística binaria y los árboles de clasificación enfocados a evaluar el rendimiento académico de los estudiantes.

Por último, el estudio adelantado por Hachuel, Boggio, Wojdyla y Servy en 2005 [5], realiza una comparaci ón de diferentes regresiones logísticas binarias para determinar la tasa de ocupación (o desocupación) en "El Gran Rosario". En el, se realiza la selección de variables a incluir en todos los modelos a partir de los Test de significancia individuales y globales. La comparación de modelos se hace a partir de la construcción de diferentes réplicas del original sobre datos de distintos años y finalmente se realizan las conclusiones de las observaciones obtenidas. Este enfoque no mide la robustez ni la eficiencia de cada modelo, se limita a realizar comparativos de resultados a partir de la regresión logística como instrumento.

La utilización de Modelos Lineales Generalizados involucra la aplicación de pruebas de bondad de ajuste frente a los datos y cumplimiento de supuestos de acuerdo con el tipo de distribución que presenten, la construcción generada de acuerdo al modelo escogido y los diferentes tipos de variables dependientes asociadas en cada caso. Los modelos de regresión logística han tenido una gran acogida en diferentes escenarios, debido a la facilidad en el análisis de resultados, el ajuste que tienen a situaciones reales y la flexibilidad en cuanto al cumplimiento de algunos supuestos. El caso más utilizado actualmente es el simple, sin embargo, los modelos multinomiales han permitido diversificar de forma positiva el análisis de resultados. Teniendo en cuenta lo anterior se hace necesaria la comparación entre los diferentes tipos de modelos existentes para validar su potencia y eventualmente tomar decisiones frente la prueba utilizar, es aquí donde el test de Vuong ayudará a verificar lo anteriormente mencionado.

2. Datos Categóricos

Corresponden a variables que se miden con un número limitado de categorías y se encuentran enmarcadas dentro del tipo cualitativo. Teniendo en cuenta la naturaleza de la variable, esta puede presentar clasificaciones de orden dicotómico, nominal, ordinal y de recuento, y tipo independiente o dependiente.

Con respecto a la primera clasificación, las variables de respuesta dicotómica son sencillamente las que tienen dos y solo dos posibles resultados. Cuando se presenta un orden natural de las categorías contenidas en la variable, esta suele ser llamada de tipo ordinal, en caso contrario, cuando no existe un orden específico o no es relevante que se presente, se denominan nominales. En teoría y como lo manifiesta Agresti (2012) el análisis estadístico no debería depender del ordenamiento que se tenga en la variable, sin embargo, se puede presentar pérdida de poder sobre la predicción de los modelos o se pueden generar interpretaciones erradas en los resultados que se generen.

3. Modelos Logísticos de respuesta binaria

Dentro de la familia de Modelos Lineales Generalizados, la Regresión Logística es la encargada de realizar el modelado de variables con respuesta categórica, bien sea de caso especial como el tipo dicotómico, o multinomial cuando se tienen más de dos categorías. El primer caso es el más popular dentro de este tipo de modelos y el que más aplicaciones ha tenido en diferentes campos, basando

[5] "Interpretación y comparación de modelos de regresión logística para el estudio de la desocupación"

su teoría en la búsqueda de ocurrencias de un evento. El modelo logístico opera de manera similar a una Regresión Lineal, donde se tiene una variable dependiente y se busca la relación de esta con un conjunto de variables explicativas, sin embargo, en el primer caso se establece una relación de probabilidad entre la ocurrencia o no de un evento dado que el individuo está representado por determinados valores para cada una de las variables independientes:

$$Pr(Y = 1|x_1, x_2, ..., x_p) = \frac{1}{1 + exp(-\alpha - \beta_1 x_1 - \beta_2 x_2 - ... - \beta_p x_p)} \tag{1}$$

Donde α representa el intercepto del modelo y β el factor que acompaña a la variable independiente y que determina la probabilidad de ocurrencia en función de la ausencia o presencia de la variable. En otras palabras, se puede indicar que la probabilidad aumenta de acuerdo con el valor de β por cada unidad que se tenga de la variable independiente asociada.

La probabilidad asociada a la ocurrencia de dicho evento cuenta con una distribución Binomial, por lo cual, y de acuerdo con lo mencionado por Agresti (2012), el Modelo de Regresión Logística tiene forma lineal para el Logit de esta probabilidad

$$Logit[\pi(x)] = log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \beta x \tag{2}$$

Partiendo de la forma del Odds Ratio ($odds = n/(1 - n)$), la ecuación anterior implica que la probabilidad de x aumenta o disminuye en función de x. Al utilizar la función exponencial [6] $exp(\alpha + \beta x) = exp(\alpha + \beta x)$, esta probabilidad se puede expresar como:

$$\pi(x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)} \tag{3}$$

Las variables explicativas en la regresión logística pueden ser continuas o categóricas sin que esto genere algún tipo de problema al momento de realizar la ejecución del modelo, pero si debe considerarse la correcta interpretación de la ausencia o presencia de la variable; así mismo, debe tenerse en cuenta este efecto al momento de trabajar con la función de regresión.

4. Modelos Logísticos de respuesta multi categórica

Como fue mencionado anteriormente, los modelos derivados del tipo logístico tienen por objetivo realizar el modelado de poblaciones cuya variable de interés es categórica. El modelo dicotómico fue expuesto en la sección anterior y describió aspectos básicos a tener en cuenta, donde se parte del

[6] El concepto del Odds Ratio parte de la razón de probabilidad que se tiene en un conjunto de datos y su base es la distribución binomial, Agresti (2012) aclara que este concepto de Odds Ratio tiene un problema grande de sesgo, el cual corrige con la aplicación del logaritmo natural como medida alternativa. Para la construcción de intervalos de confianza es necesario realizar la inversa del logaritmo natural construido, por esta razón aparece el exponencial que es el que finalmente se utiliza dentro de la función de probabilidad de la regresión logística.

principio de cálculos de probabilidad asociados a la ocurrencia de un evento que en este caso está representado en la variable respuesta y cuyo fundamento proviene de la distribución binomial. Teniendo en cuenta la naturaleza de la variable de interés, en la cual se encuentran diferentes categorías dentro de las cuales un sujeto puede estar clasificado, se puede hablar de un modelo multi categórico. En este sentido, vale la pena profundizar en las diferentes presentaciones que se dan en estos modelos, partiendo de las dos principales, que son nominales y ordinales.

4.1 Modelos Logísticos multi categóricos de respuesta nominal

En los modelos de respuesta nominal, el orden de las categorías no es relevante y su principal objetivo es determinar una correcta clasificación de la población o muestra de estudio. Dicho lo anterior, es necesario mencionar que todos los modelos nominales deben contar con una categoría base, de la cual saldrá la probabilidad de que la respuesta obtenida en la variable dependiente se encuentre en alguna de las categorías y adicionalmente se comparará con cada una de ellas. Una forma de expresar lo anterior es:

$$log\left(\frac{\pi_j}{\pi_J}\right), \qquad j = 1, ..., J-1 \tag{4}$$

Donde J denota el número total de categorías presentes en y y j_i representa cada una de las categorías presentes.

Al igual que en el modelo logístico de respuesta binaria, la materialización del Log en la ecuación característica de regresión (teniendo en cuenta la aclaración realizada del log a partir de la razón de odds) es:

$$log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j x, \qquad j = 1, ..., J-1 \tag{5}$$

Por lo tanto, el modelo siempre estará compuesto por $J-1$ ecuaciones acompañadas de los parámetros tradicionales α como intercepto y β como coeficiente que acompaña a cada una de las variables independientes utilizadas. Se establecerán tantas ecuaciones como categorías existentes. Agresti (2012) ejemplifica esta relación utilizando dos categorías de la forma:

$$log\left(\frac{\pi_a}{\pi_b}\right) = log\left(\frac{\pi_a/\pi_J}{\pi_b/\pi_J}\right) = log\left(\frac{\pi_a}{\pi_J}\right) - log\left(\frac{\pi_b}{\pi_J}\right)$$

$$(\alpha_a + \beta_a x) - (\alpha_b + \beta_b x)$$

$$(\alpha_a - \alpha_b) + (\beta_a + \beta_b)x \tag{6}$$

Note que la ecuación resultante cumple la forma mencionada anteriormente, donde se tiene un conjunto de interceptos que resultan de cada categoría (α) y un conjunto de parámetros que

acompañan a la variable dependiente en cada caso. La elección de la categoría base puede ser realizada a criterio del experto y no interferirá en la construcción del modelo. Por defecto, los parámetros también son calculados para esta categoría base.

La prueba de hipótesis de independencia en el comportamiento de Y frente a las variables independientes sigue siendo la misma que tradicionalmente se ha venido manejando en los diferentes tipos de modelos, donde $H_0 : \beta_1 = \beta_2 = ... = \beta_n = 0$

Otra manera de expresar las probabilidades dentro del modelo multi categórico es:

$$\pi_j = \frac{e^{\alpha_j + \beta_j x}}{\sum_h e^{\alpha_h + \beta_h x}}, \qquad j = 1, ..., J-1 \tag{7}$$

De esta manera, el denominador es la sumatoria de las probabilidades para cada una de las categorías presentes en el modelo, mientras que el numerador representará la ecuación para cada categoría, excepto para la categoría base, la cual se encuentra representada por un uno (1).

Cuando se presentan diferentes categorías dentro de la variable Y como en cada una de las variables independientes, nos referimos a un modelo generalizado de elección discreta.

4.2 Modelos Logísticos multi categóricos de respuesta ordinal

Cuando las categorías de la variable dependiente cuentan con un orden natural, los modelos logísticos multi categóricos pueden ser utilizados para realizar predicciones sobre los individuos objeto de estudio; en este caso, el tratamiento de los datos, las conclusiones sobre los resultados obtenidos y en general el comportamiento del modelo difieren ampliamente frente al escenario de datos nominales revisado en la sección anterior. Adicionalmente, los modelos ordinales presentan subsegmentos de acuerdo con el tipo de respuesta que se quiera encontrar. En tal caso es necesario revisar los posibles escenarios disponibles y elegir la mejor opción en el momento de modelar. Algunos de los más importantes son los modelos de categoría adyacente, la relación de continuación y los modelos de probabilidades proporcionales. [7]

Así como fue mencionado en la descripción del modelo nominal, la comparación es relevante en la composición y en la muestra de los resultados, al tener una categoría base es posible verificar cual de los Logit construídos presenta un mejor ajuste al modelo propuesto. De la misma manera, el modelo ordinal utiliza la metodología de comparación para saber cual es el modelo más razonable para el Logit.

4.2.1 Modelos de respuesta adyacente

De acuerdo con lo mencionado por Hosmer, Lemeshow y Sturdivant (2013), los modelos de respuesta adyacente comparan cada respuesta con la siguiente más grande , de acuerdo al orden de la variable de interés que se está evaluando. Si suponemos que los log-odds no dependen de la respuesta y son lineales en los coeficientes, entonces los logits de categorías adyacentes son los siguientes:

[7] Hosmer, D. , Lemeshow, S. , Sturdivant, R. (2013). “Applied Logistic Regression".

$$a_k(x) = \alpha_k + x'\beta \quad para \quad k = 1, 2, ..., K \tag{8}$$

Así mismo, los autores también mencionan que los logit de categorías adyacentes son una versión restringida de los de categoría base que fueron mencionados en el caso nominal. Una forma de ver esta situación es através de la forma:

$$(\alpha_1 + \alpha_2 + ... + \alpha_k) + k\mathbf{x}'\beta \tag{9}$$

Nótese que la ecuación tiene un intercepto $\beta_{k0} = (\alpha_1 + \alpha_2 + ... + \alpha_k)$ y coeficientes $\beta_k = k\beta$

4.2.2 Modelos de relación de continuación

Es una versión contraria a los modelos adyacentes, donde en vez de realizar la comparación sobre las categorías más altas, se realiza sobre las más bajas de tal manera que $Y < k$ para $k = 1, 2, ..., K$. El logit del modelo se define como:

$$r_k(x) = \theta_k + x'\beta_k \quad para \quad k = 1, 2, ..., K \tag{10}$$

A diferencia de las relaciones adyacentes, las cuales trabajan bajo un modelo restringido en los logits, los modelos de relación de continuacíon tienen diferentes pendientes en cada uno de los logits que se construyen.

4.2.3 Modelos de probabilidades proporcionales

Este modelo cumple una función comparativa entre la probabilidad de una respuesta igual o menor ($Y <= k$) y la probabilidad de una respuesta mayor ($Y > k$), de la forma:

$$c_k(x) = ln\left[\frac{Pr(Y \leq k|X)}{Pr(Y > k|X)}\right]$$

$$\tau_k - x'\beta \quad para \quad k = 0, 1, ..., K-1 \tag{11}$$

En este caso K=1

4.2.4 Modelos de probabilidad acumulada

De acuerdo con lo mencionado por Agresti (2012), una probabilidad acumulativa para Y se da cuando es necesario que esta caiga en un punto determinado de las categorías disponibles. Por lo tanto,

$$P(Y <= j) = \pi_1 + ... + \pi_j, j = 1, ..., J \tag{12}$$

Las cuales se van sumando de acuerdo con el orden de las categorías que se tengan. Los logits de este tipo de probabilidades presentan la misma estructura de los que se encuentran en los modelos tradicionales, la variación en el numerador se da por la suma de las probabilidades hasta $Y <= j$ que es la categoría sobre la cual se espera que la variable dependiente caiga.

Otra manera de ver este tipo de modelos es a través de los Logit acumulados con probabilidades proporcionales, donde se mantiene fijo el valor de β que genera el cambio en la probabilidad frente a la variable independiente. Los cálculos diferenciales pueden observarse en cada una de las $J-1$ categorías de la variable dependiente, generándose un intercepto para cada caso. Al revisar detallamente esta situación, se puede observar que esta metodología logra generar modelos logit independientes en cada caso, por lo cual, la representación de esta interacción se da por

$$P(Y <= j) = exp(\alpha_j + \beta_x)/[1 + exp(\alpha_j + \beta_x)] \tag{13}$$

5. Divergencia de KullBack-Leibler

La divergencia de Kullback-Leibler parte de la medición de la distancia entre dos distribuciones de probabilidad P y Q, las cuales se presentan sobre la misma variable *(x)* y Q es absolutamente continua con respecto a P, representándola de la forma[8]:

$$D_{KL}(P||Q)$$

$$D_{KL}(P||Q) = \sum_i P_i ln \frac{P_{(i)}}{Q_{(i)}} \tag{14}$$

Para distribuciones continuas, la divergencia es representada por las funciones de densidad de P y Q:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) ln \frac{p(x)}{q(x)} dx, \tag{15}$$

En otras palabras, se puede decir que este indicador mide la similitud que tienen dos distribuciones, a partir del promedio ponderado de la diferencia logarítmica de las probabilidades de P con respecto a Q y solamente se define si P y Q suman uno, así, es de esperarse que se cumpla que $Q_i > 0$ y $P_i > 0$. El término divergencia se utiliza dado que no se cumple asimetría, por lo tanto:

$$D(P, Q) \neq D(Q, P) \tag{16}$$

La divergencia de Kullback-Leibler es un indicador que nunca toma valores negativos, por lo tanto *D(P, Q)* es mayor a cero y asume igualdad entre las distribuciones cuando se presenta que P=Q. Adicionalmente, es asociada con el método de distribuciones de ajuste por máxima verosimilitud dado que el principio de este, pretende maximizar el parámetro de estimación para la probabilidad, lo cual es análogo a minimizar la distancia entre las distribuciones que se presentan en un conjunto de datos, por lo cual es posible decir que:

[8] Joyce J.M. (2011) Kullback-Leibler Divergence. In: Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg

$$D(P,Q) = L(P,P) - L(P,Q), \tag{17}$$

También tiene un papel muy importante en la teoría de estadística inferencial (Eguchi y Copas, 2006) y en la aplicación de pruebas de bondad de ajuste en modelos generalizados por medio del criterio de información de Akaike's (AIC).

5.1 **Razones de Probabilidad:**

Asumiendo que P y Q son posibles distribuciones de probabilidad para un conjunto de datos *(x)*, se puede realizar una prueba de hipótesis que tenga H como nula y A como alterna y que pretenda probar cual de las dos distribuciones presenta el mejor ajuste sobre los datos. Al realizar la razón de verosimilitud partiendo de las funciones de densidad de H (el modelo incorrecto se ajusta mejor a los datos que el modelo correcto) y A (el modelo correcto se ajusta mejor a los datos que el modelo incorrecto) encontramos:

$$\begin{aligned} \lambda = \lambda(x) &= log\frac{f_A(X)}{f_H(X)} \\ E_A\{\lambda(x)\} &= D(P_A, P_H), \\ E_A\{\lambda(x)\} &= D(P_H, P_A), \end{aligned} \tag{18}$$

Cuando se valida la hipótesis alternativa, se encuentra la forma de la divergencia de Kullback-Leibler, cuanto mayor es la razón de probabilidad, mayor evidencia se tiene de la hipótesis alternativa.

6. **Test de Vuong**

El test de Vuong es una prueba utilizada para generar la selección del mejor modelo, entre un conjunto finito de posibilidades. Utiliza el criterio de información de Kullback-Leibler (KLIC), cuya base es la medición de la distancia entre un modelo propuesto y uno llamado verdadero, donde es natural tratar de encontrar aquella propuesta que minimice esta distancia, lo que indicará que dicho modelo tendrá la mayor aproximación a la distribución de los datos. Es utilizado en modelos anidados, no anidados, superpuestos o que presentan problemas en la especificación.

Vuong plantea un enfoque que coincide con las pruebas cásicas aplicadas a modelos anidados. Su hipótesis está basada en probar si todos los modelos se aproximan a la verdadera distribución de los datos o por el contrario existe uno que se aproxime más que otro. Dicho en términos de contraste de pruebas de hipótesis, es posible expresar lo anterior como:

Dados dos modelos condicionales F_θ y G_γ $\;g(y|z;\gamma); \gamma \in \Gamma$,

$$H_0 : E^0[\log f(y|z;\theta)] = E^0[\log g(y|z;\gamma)]$$

$$H_1 : E^0[\log f(y|z;\theta)] > E^0[\log g(y|z;\gamma)]. \tag{19}$$

Dado que la base de test es el criterio de información de Kullback-Leibler (KLIC), todas las pruebas realizadas para la elección partirán del estimador de máxima verosimilitud. Para un modelo condicional F_θ, el criterio de información puede verse de la forma:

$$KLIC(H^0_{y|z}; F_\theta) = E^0[logh^0(Y_t|Z_t)] - E^0[logf(Y_t|Z_t;\theta_*)] \tag{20}$$

Donde $h^0(\cdot|\cdot)$ es la verdadera densidad condicional de Y_t dado Z_t y θ_* son los p-seudo valores verdaderos de θ. Una expansión de las hipótesis expuestas anteriormente es:

$$H_0 : E^0\left[log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] = 0, \tag{21}$$

lo que significa que F_θ y G_γ son equivalentes,

$$H_f : E^0\left[log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] > 0, \tag{22}$$

lo que significa que F_θ es mejor que G_γ

$$H_g : E^0\left[log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] < 0, \tag{23}$$

lo que significa que F_θ es peor que G_γ.

Las propiedades que se pueden destacar de estos escenarios son:

- Un modelo especificado correctamente debe ser al menos tan bueno como cualquier otro modelo.
- La hipótesis nula H_o no requiere que ninguno de los modelos competidores se especifique correctamente.
- El log verosomilitud es una estadística natural para discriminar entre dos modelos.

La verdadera densidad $h^0(y|z)$ no se encuentra restringida a pertenecer a ninguno de los dos modelos (F y G), así mismo, dado que la comparación de modelos se puede presentar en los escenarios anidados, no anidados, superpuestos o mal especificados, se debe obtener la distribución asintótica del indicador de máxima verosimilitud y trabajar los contrates de hipótesis bajo el mismo parámetro. La distribución asintótica y la velocidad de la convergencia dependerá de la anidación de los modelos o de la correcta especificación [9].

6.1 **Modelos estrictamente no anidados:**

Dos modelos F_θ y G_γ son estricamente no anidados si no existe intesercción en las variables que componen cada uno de estos. En otros términos:

[9] Vuong H.Quang. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica, Journal of the Econometric Society, Vol 57, No. 2, pp 307-333

$$F_\theta \cap G_\gamma = 0 \tag{24}$$

Esto quiere decir que los modelos F y G no tienen distribuciones condicionales en común, esto se puede presentar cuando son modelos de regresión lineal con diferentes supuestos de distribución o cuando tienen diferentes formas funcionales. Basándonos en la prueba de razón de verosimilitud (LR), la cual es utilizada por el test entendiendo la definición de distancias anteriormente mencionada y teniendo en cuenta lo mencionado por Vuong (1989), si F_θ y G_γ son estrictamente no anidados, entonces:

$$BajoH_0 : n^{(-1/2)} LR_n(\hat{\theta_n}, \hat{\gamma_n})/\hat{\omega_n} \underrightarrow{D} N(0,1),$$

$$BajoH_f : n^{(-1/2)} LR_n(\hat{\theta_n}, \hat{\gamma_n})/\hat{\omega_n} \underrightarrow{a.s.} +\infty,$$

$$BajoH_f : n^{(-1/2)} LR_n(\hat{\theta_n}, \hat{\gamma_n})/\hat{\omega_n} \underrightarrow{a.s.} -\infty, \tag{25}$$

7. Metodología

La parametrización del Test de Vuong se encuentra disponible en muchos software estadísticos en la actualidad, el principio de este, parte de todos los conjuntos de datos para los cuales es posible ajustar un modelo que tenga como base una distribución apropiada para trabajar con conteos. Algunos ejemplos de estos, pueden ser variables de interés atadas a una distribución Poisson, Beta, modelos para trabajar con repeticiones de ceros continuas y algunas otras. Al igual que el criterio de información de Akaike, el BIC, las pruebas de Anova, entre otras, la finalidad del Test es realizar la comparación de diferentes tipos de modelos y determinar cual presenta un mejor ajuste a los datos sobre los cuales se está trabajando. Una de las mayores ventajas del Test de Vuong, es la flexibilidad que tienen frente al cumplimiento de algunos supuestos que son básicos en otro tipo de pruebas, de manera que permite abarcar en mayor medida, muchas distribuciones que no son posibles de comparar mediante las técnicas tradicionales.

Entendiendo lo anterior, fueron realizados ajustes dentro de la parametrización propuesta por Vuong (sin violar ningún componente de la escencia del Test), para que de esta forma pueda ser utilizado como medio de comparación en escenarios donde se utilizan modelos logísticos, teniendo en cuenta que el principio de estos modelos es la distribución binomial, la cual, a diferencia de los anteriormente mencionados, no modelan escenarios de conteo, sino más bien eventos de ocurrencia asociados a las posibles opciones presentadas por la variable de interés. Las corecciones propuestas en el test original, las cuales incluyen las pruebas AIC y BIC como alternativas de elección, no serán tenidas en cuenta en el presente estudio.

7.1 Estudio de simulación

Consideramos varios escenarios de simulación para evaluar el rendimiento de los cambios generados en la estructura del Test de Vuong. Para llevar a cabo las simulaciones, se utilizó el diseño de varios modelos con diferente número de categorías en la variable de interés (y) y diferentes cantidades y tipos de variables independientes (x). En cada escenario se fueron añadiendo una cantidad finita de categorías en la variable respuesta, acompañadas de un incremento progresivo en el número de variables independientes provenientes de una distribución uniforme; algunos de

los escenarios fueron simulados además con una variable independiente cualitativa del tipo dicotómica proveniente de una distribución binomial. El número total de escenarios presentados en este documento, comprende doce diferentes combinaciones con las tipologías anteriormente mencionadas, las cuales serán descritas al detalle y se realizarán las conclusiones y recomendaciones a lugar.

Específicamente, se seguirán los siguientes pasos:

1. Se definieron tamaños de muestra a partir de cien observaciones, hasta nueve mil, divididos en conjuntos de trescientos datos con los cuales se realizaron las simulaciones de las variables dependientes e indpendientes, la construcción de los modelos nominales y ordinales, el testeo de la prueba de Vuong y los resultados presentados en este estudio.

2. Para cada tamaño de muestra, se simularon hasta tres conjuntos de variables independientes, tomando como base la distribución uniforme y un último conjunto procedente de una simulación de valores nominales que representarían una variable dicotómica.

3. De la misma manera, para cada tamaño de muestra fueron construídas las variables dependientes de tres, cuato y cinco categorías a través de un muestreo aleatorio sin reemplazo ajustado a la cantidad de categorías que se deseaban incluir en cada caso. De esta forma, se construyeron doce escenarios sobre los cuales se ejecutarán las pruebas correspondentes.

4. Para cada una de las doce combinaciones presentadas, fueron construídos modelos de regresión logística nominales y ordinales que fueron corridos en cien repeticiones para todos los tamaños de muestra construídos. Únicamente fue considerado el modelo logístico nominal general para el estudio presentado.

5. A partir de lo anterior, se ejecutó la prueba de Vuong. La hipótesis planteada inicialmente en este trabajo pretendía comprobar la superioridad de los modelos nominales sobre los ordinales, por lo cual, se esperaba que esto fuera demostrado a través de los multiples escenarios creados. A partir del conteo generado con las repeticiones propuestas en cada tamaño de muestra y cada escenario, se fueron detectando las veces en las cuales el modelo nonimal era superior al modelo ordinal.

Los resultados del procedimiento descrito anteriormente se muestran a continuación:

7.2 Observaciones generadas

Escenario 1

Se presenta una variable dependiente (y) de tres categorías, atada a modelos de una, dos, tres y cuatro variables independientes (x). Se realizó el conteo de la cantidad de ocasiones en las cuales el Test de Vuong indica que el modelo nominal es mejor que el ordinal para los datos sobre los que se está trabajando, este conteo es expresado en un porcentaje sobre el total de las iteraciones realizadas que fueron mencionadas en el punto cuatro del numeral anterior. La tabla número uno resume los resultados obtenidos y posteriormente se encontrará la gráfica que permite analizar en forma más sencilla las tendencias encontradas:

Tabla 1: Resultados obtenidos con variable dependiente de tres categorías

Tamaño de muestra	Modelo 1	Modelo 2	Modelo 3	Modelo 4
100	0	0.01	0.02	0.09
300	0	0.02	0.02	0.02
600	0	0	0.03	0.02
900	0	0	0.02	0.07
1.200	0	0.01	0.05	0.05
1.500	0	0	0.04	0.03
1.800	0	0.02	0.07	0.04
2.100	0.02	0.02	0.05	0.07
2.400	0	0.01	0.03	0.06
2.700	0.01	0.04	0.07	0.07
3.000	0.01	0.01	0.05	0.07
3.300	0.03	0.02	0.07	0.09
3.600	0.02	0.06	0.11	0.09
3.900	0.01	0.03	0.07	0.08
4.200	0.04	0.02	0.04	0.11
4.500	0.04	0.05	0.06	0.08
4.800	0.02	0.02	0.1	0.07
5.100	0.03	0.06	0.07	0.13
5.400	0.01	0.08	0.07	0.14
5.700	0.07	0.03	0.11	0.14
6.000	0.1	0.11	0.1	0.13
6.300	0.08	0.1	0.11	0.13
6.600	0.02	0.04	0.07	0.21
6.900	0.07	0.08	0.07	0.14
7.200	0.08	0.16	0.14	0.17
7.500	0.09	0.09	0.07	0.16
7.800	0.06	0.08	0.11	0.2
8.100	0.07	0.1	0.19	0.19
8.400	0.05	0.1	0.11	0.22
8.700	0.07	0.09	0.22	0.2
9.000	0.09	0.12	0.16	0.22

Al observar los resultados obtenidos, es posible evidenciar un incremento en el porcentaje de veces que el modelo nominal es escogido sobre el modelo ordinal a medida que se incrementa el tamaño de la muestra; de la misma manera, al aumentar la cantidad de variables independientes ingresadas en el modelo, el porcentaje mencionado incrementa a una mayor velocidad y con mayor estabilidad. Gráficamente, es posible observar lo expuesto en la tabla 1, por lo cual, de manera individual, se encuentra una mayor dispersión en los resultados obtenidos cuando se trabajó con máximo dos variables independientes, donde a pesar de que se encuentra una tendencia creciente frente al porcentaje de casos en los cuales el Test de Vuong escoje el modelo nominal, no alcanza niveles superiores al 20 %. De otra parte, los escenarios tres y cuatro presentan un mejor comportamiento, siendo el úlltimo mencionado, el que mayor estabilidad presenta frente a la tendencia creciente de escogencia. En este punto es posible notar que el nivel de casos donde el modelo nominal es superior al ordinal llega al rededor del 25 %. Para todos los escenarios es claro que al realizar una simulación con un tamaño de muestra superior, se puede llegar a obtener un mayor nivel de aceptación del modelo nominal, encontrando que los escenarios tres y cuatro alcanzarían esta situacón de forma más rápida que los dos primeros.

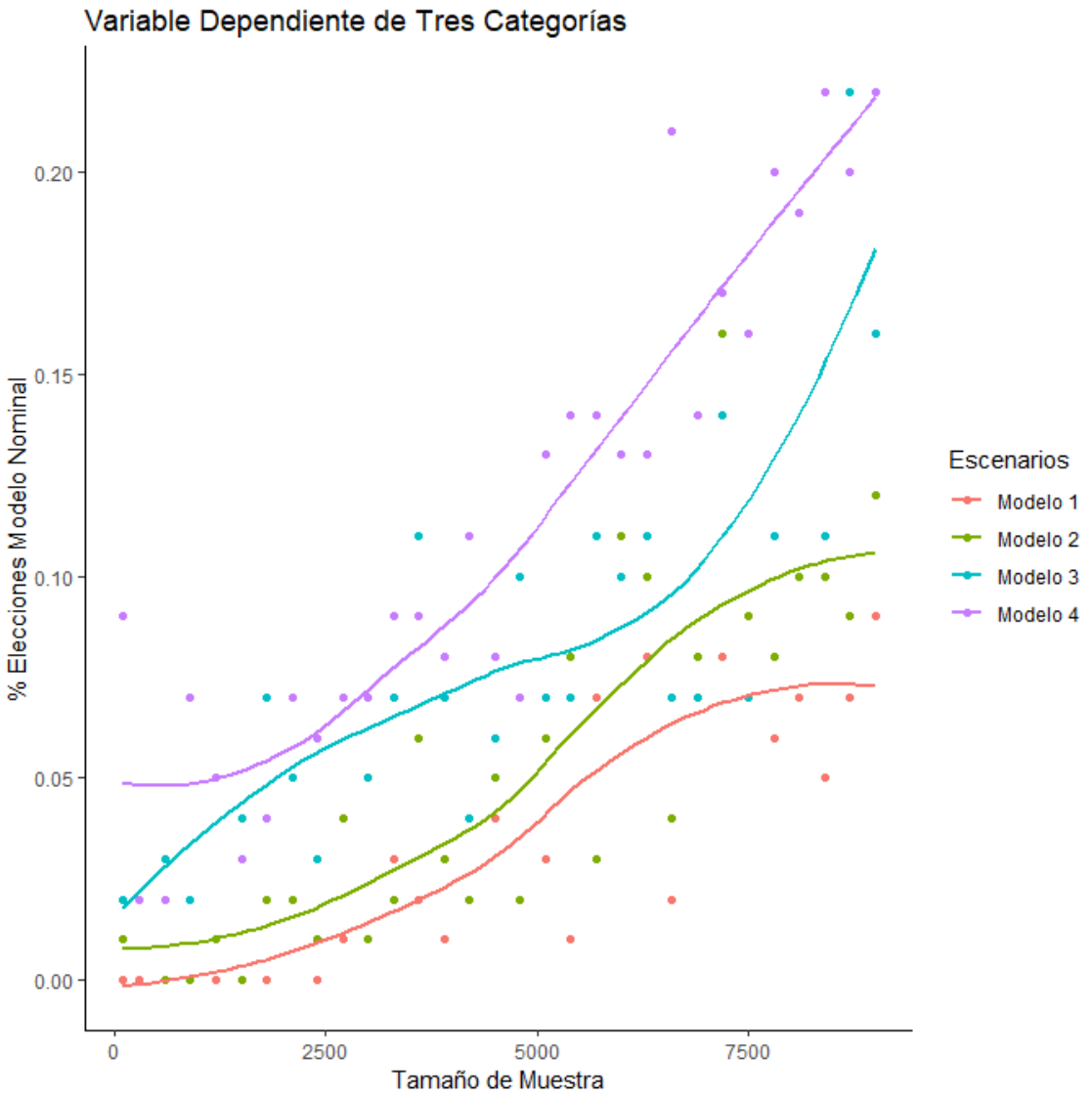

Figura 1: Porcentaje de elecciones de los modelos nominales sobre los ordinales. De acuerdo a la combinación: Modelo 1, una variable independiente. Modelo 2, dos variables independientes. Modelo 3, tres variables independientes. Modelo 4, cuatro variables independientes.

Escenario 2

Se presenta una variable dependiente (y) de cuatro categorías, atada a modelos de una, dos, tres y cuatro variables independientes (x). Se realizó el conteo de la cantidad de ocasiones en las cuales el Test de Vuong indica que el modelo nominal es mejor que el ordinal para los datos sobre los que se está trabajando, este conteo es expresado en un porcentaje sobre el total de las iteraciones realizadas. La tabla número dos resume los resultados obtenidos y posteriormente se encontrará la gráfica que permite analizar en forma más sencilla las tendencias encontradas:

Tabla 2: Resultados obtenidos con variable dependiente de cuatro categorías

Tamaño de muestra	Modelo 1	Modelo 2	Modelo 3	Modelo 4
100	0.02	0.13	0.23	0.38
300	0.06	0.1	0.22	0.37
600	0.12	0.17	0.23	0.41
900	0.19	0.23	0.4	0.49
1.200	0.21	0.27	0.39	0.61
1.500	0.29	0.4	0.51	0.66
1.800	0.47	0.5	0.56	0.75
2.100	0.37	0.54	0.7	0.76
2.400	0.55	0.61	0.71	0.71
2.700	0.51	0.57	0.69	0.74
3.000	0.64	0.66	0.84	0.83
3.300	0.76	0.79	0.88	0.84
3.600	0.68	0.81	0.9	0.93
3.900	0.8	0.8	0.87	0.95
4.200	0.8	0.83	0.89	0.95
4.500	0.87	0.91	0.95	0.98
4.800	0.88	0.92	0.93	0.98
5.100	0.93	0.92	0.96	0.99
5.400	0.94	0.94	0.99	0.98
5.700	0.98	0.99	0.98	1
6.000	0.92	0.99	1	1
6.300	0.95	0.95	0.96	1
6.600	0.98	0.96	0.99	0.99
6.900	0.98	1	1	1
7.200	0.95	0.98	1	1
7.500	0.99	0.99	0.98	1
7.800	0.99	1	1	1
8.100	1	1	1	1
8.400	1	0.99	1	1
8.700	1	0.99	1	1
9.000	1	1	1	1

Al igual que los resultados obtenidos en el escenario uno, un incremento en el tamaño de la muestra para todos los casos presentados, genera un mayor número de ocasiones en las cuales el modelo nominal es escogido por el Test de Vuong sobre el ordinal. En todos los casos y a diferencia de las simulaciones que fueron construídas sobre una variable dependiente de tres categorías, este escenario presenta un mayor nivel de estabilidad y una clara tendencia creciente sobre la elección generada por la prueba, en esta ocasión es fácilmente evidenciable que aunque se presentan diferentes velocidades, las cuales dependen de la cantidad de variables independientes asociadas a los modelos, todas tienden a elegir de forma absoluta un modelo nominal por encima de un ordinal. En el escenario anterior, se mencionaba que probablemente al realizar un número mayor de simulaciones incrementando el tamaño de la muestra, se podría llegar a una convergencia igual a uno, para este caso, la inclusión de una categoría más en la variable dependiente generó una convergencia mucho mayor y más rápida que la anterior. Esta situación es fácilmente evidenciable en el modelo de cuatro variables independientes, en el cual a partir de la primera simulación, el porcentaje de elección del modelo nominal está alrededor del 30 %, mucho mayor al 22 % encontrado como el máximo alcanzado por una variable dependiente de tres categoías y cuatro independientes del escenario uno.

El gráfico de comparación de todos los escenarios resume en mejor medida lo expuesto anteriormente, al realizar una aproximación de todos los casos en una sola vista y partiendo de la misma escala, las diferencias, los alcances y los estimativos mencionados son notados con mayor facilidad. De los principales puntos a resaltar de este parte encontramos :

- Todos los casos tienden a elegir en su totalidad modelos nominales a medida que el tamaño de muestra incrementa
- Los modelos con tres y cusatro variables independientes asociadas, arrancan desde tamaños de muestra pequeños, con alto porcentajes de elección de modelos nominales.

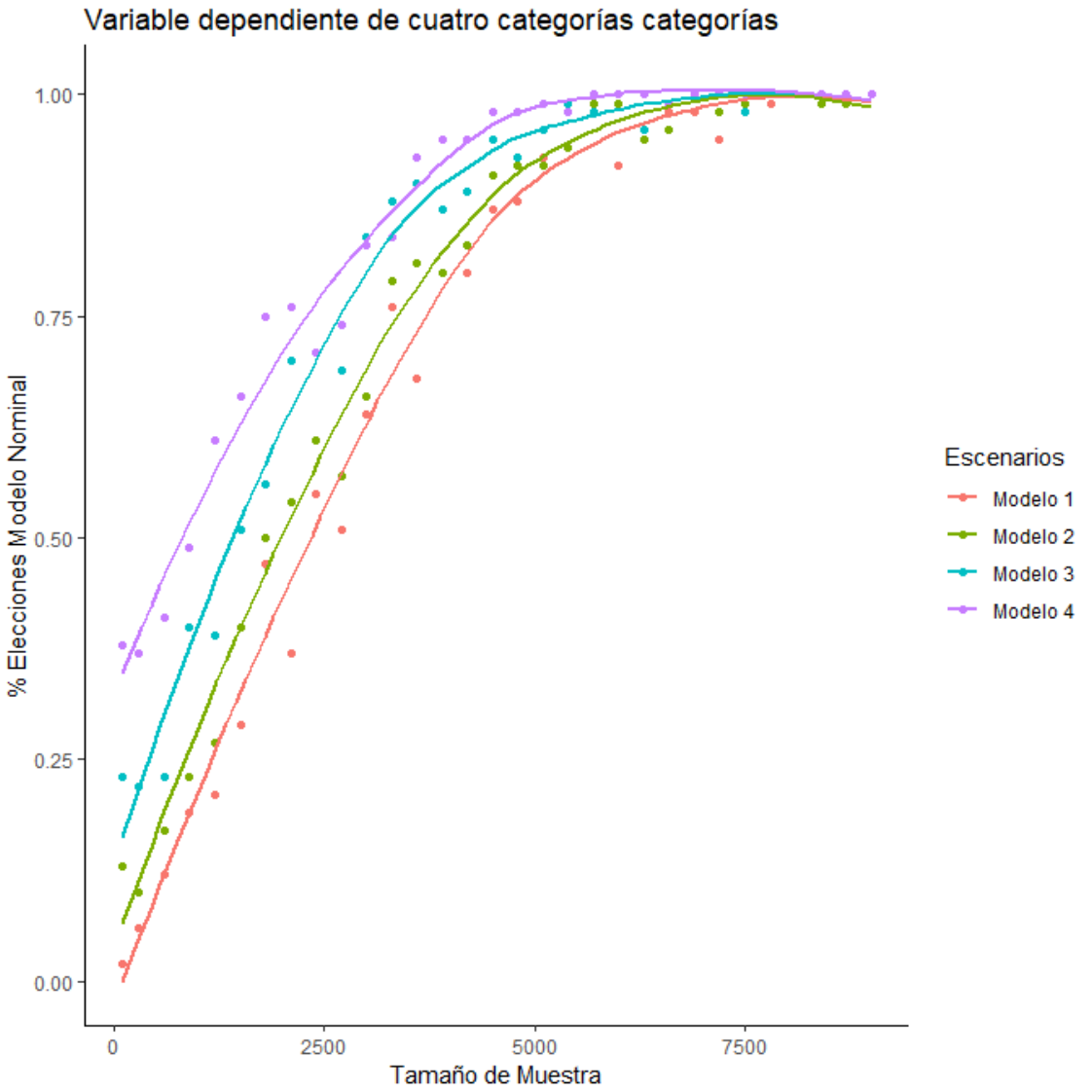

Figura 2: Porcentaje de elecciones de los modelos nominales sobre los ordinales. De acuerdo a la combinación: Modelo 1, una variable independiente. Modelo 2, dos variables independientes. Modelo 3, tres variables independientes. Modelo 4, cuatro variables independientes.

Escenario 3

Se presenta una variable dependiente (y) de cinco categorías, atada a modelos de una, dos, tres y cuatro variables independientes (x). Se realizó el conteo de la cantidad de ocasiones en las cuales el Test de Vuong indica que el modelo nominal es mejor que el ordinal para los datos sobre los que se está trabajando, este conteo es expresado en un porcentaje sobre el total de las iteraciones realizadas que fueron mencionadas en el punto cuatro del numeral anterior. La tabla número dos resume los resultados obtenidos y posteriormente se encontrará la gráfica que permite analizar en forma más sencilla las tendencias encontradas:

Tabla 3: Resultados obtenidos con variable dependiente de cinco categorías

Tamaño de muestra	Modelo 1	Modelo 2	Modelo 3	Modelo 4
100	0.07	0.23	0.4	0.75
300	0.03	0.27	0.5	0.77
600	0.13	0.33	0.51	0.68
900	0.22	0.43	0.59	0.82
1.200	0.27	0.45	0.59	0.85
1.500	0.44	0.6	0.75	0.86
1.800	0.44	0.61	0.81	0.91
2.100	0.55	0.81	0.85	0.91
2.400	0.56	0.78	0.83	0.95
2.700	0.68	0.88	0.94	0.97
3.000	0.72	0.87	0.91	0.95
3.300	0.83	0.88	0.96	1
3.600	0.84	0.93	0.94	0.97
3.900	0.91	0.95	0.98	1
4.200	0.89	0.93	0.99	1
4.500	0.93	0.98	1	1
4.800	0.97	0.97	1	1
5.100	0.96	0.97	0.99	1
5.400	0.99	0.97	0.98	1
5.700	0.99	0.98	1	1
6.000	0.96	1	1	1
6.300	1	1	1	1
6.600	0.99	0.99	1	1
6.900	0.99	0.99	1	1
7.200	1	1	0.99	1
7.500	1	1	1	1
7.800	1	1	1	1
8.100	1	1	1	1
8.400	1	1	1	1
8.700	1	1	1	1
9.000	1	1	1	1

Como ocurrió en los escenarios uno y dos, el incremento en el tamaño de muestra genera un mayor número de ocasiones en las cuales el modelo nominal es escogido por el Test de Vuong sobre el ordinal, sin embargo, para este último caso, la adición de una categoría adicional a la variable dependiente generó un importante incremento en la velocidad sobre la cual se escoge en la totalidad de los casos el modelo nominal sobre el ordinal, donde aproximadamente para un tamaño de muestra cercano a los 2.500 registros, el porcentaje de escogencia es del 100 % . Para el

último escenario construído , en el cual fueron incluídas cuatro variables independientes al modelo, a partir de la primera simulación, el porcentaje de escogencia es cercano al 75 %, por lo cual es importante mencionar que la cantidad de categorías manejadas en la variable dependiente genera grandes cambios en la elección generada por el Test de Vuong, lo que repercute en la utilización de tamanõs de muestra relativamente pequeños para que la prueba se incline por la elección de un modelo nominal.

El gráfico de comparación de todos los escenarios resume en mejor medida lo expuesto anteriormente, al realizar una aproximación de todos los casos en una sola vista y partiendo de la misma escala, las diferencias, los alcances y los estimativos mencionados son notados con mayor facilidad. De los principales puntos a resaltar de este parte encontramos:

- El incremento de categorías en variables dependientes aumenta el nivel de lección sobre modelos nominales.
- El modelo con cuatro variables independientes asociadas, arrancan desde tamaños de muestra pequeños, con alto porcentajes de elección de modelos nominales.

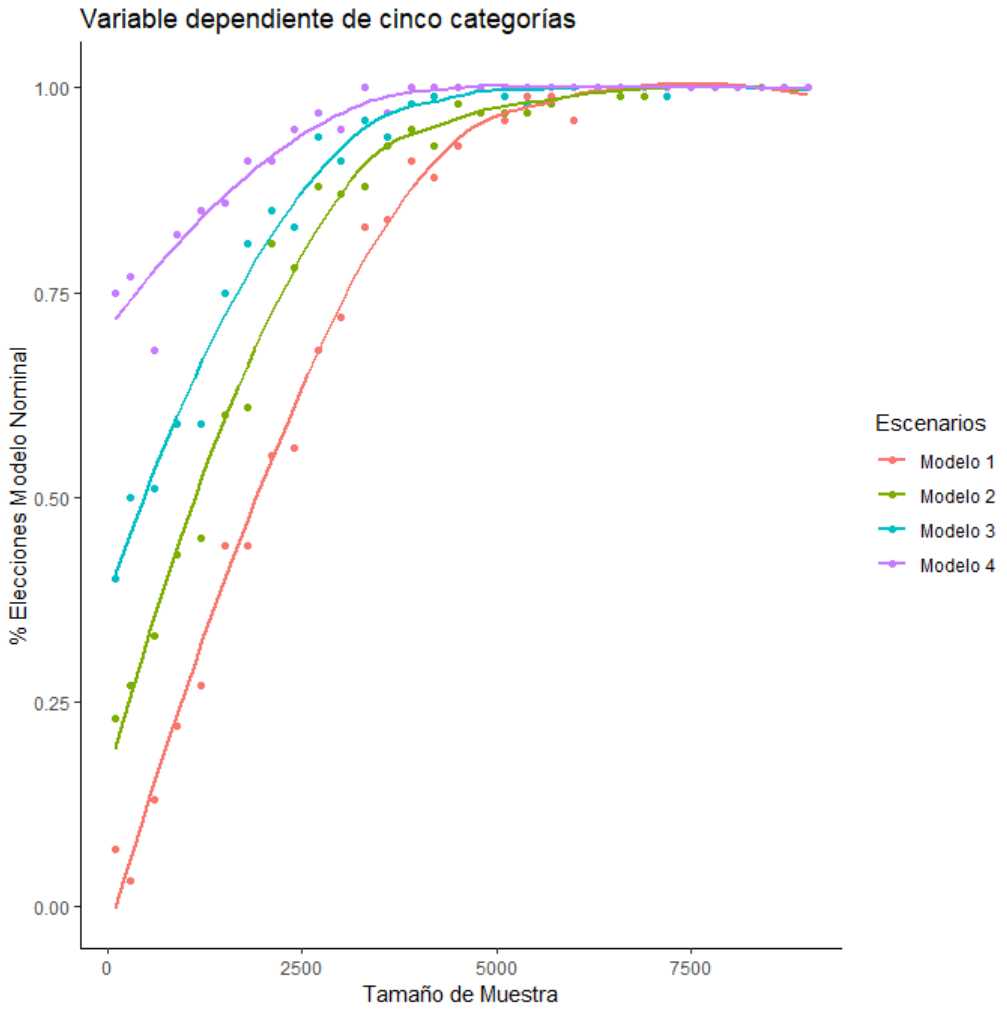

Figura 3: Porcentaje de elecciones de los modelos nominales sobre los ordinales. De acuerdo a la combinación: Modelo 1, una variable independiente. Modelo 2, dos variables independientes. Modelo 3, tres variables independientes. Modelo 4, cuatro variables independientes.

8. Conclusiones

Este trabajo realizó la comparación de eficiencia entre un modelos logístico nominal y un ordinal a partir del ajuste del Test de Vuong para la evaluación y elección del mejor modelo. Para esto fueron generadas simulaciones con diferentes escenarios, los cuales pretendían probar la efectividad de la prueba en múltiples situaciones y comprobar la hipótesis inicial planteada sobre la cual se indicaba que la elección debía inclinarse por los modelos del tipo nominal.

Los resultados obtenidos con variables dependientes de pocas categorías no evidenciaron un mayor porentaje de elección del modelo nominal sobre el ordinal; la adición de variables independientes generó un incremento en este porcentaje pero no fue superior frente al modelo ordinal. Al evaluar la tendencia en la medida en la que el tamanõ de muestra de las simulaciones incrementaba, se identificó que era posible obtener el resultado deseado pero con una mayor cantidad de observaciones.

El incremento de categorías en la variable dependiente vuelve sensible la elección de los modelos nominales sobre los ordinales, esto fue evidenciado en los escenarios dos y tres, siendo mas fuerte el último, donde en todos los casos, el modelo nominal fue superior.

A partir de la cantidad de categorías de la variable dependiente, la adición de variables independientes en los modelos genera una diminución en el tamaño de muestra ideal para el porcentaje de elección del modelo nominal sea mayor.

La inclusión de variables cualitativas en la construcción de los modelos no generó alteraciones significativas en los resultados presentados.

La validación a través de las corecciones por medio de los criterios de información AIC y BIC, las cuales fueron propuestas por el autor de la prueba, no se tuvieron en cuenta en este ejercicio. De la misma manera se trabajó únicamente bajo el supuesto de modelos estrictamente no anidados.

Este ejercicio se realizó a partir de simulaciones de conjuntos de datos y con un fin académico, sin embargo, para futuros trabajos se propone utilizar bases de datos reales y la comprobación de los puntos mencionados en el apartado anterior.

9. Agradecimientos

A todas las personas que me dieron su apoyo, amistad, conocimiento y dedicación durante todos estos años.

A mis papás y mis hermanos, que son el motor más importante de mi vida y la razón de levantarme con una sonrisa todas las mañanas.

A mi tutor, Wilmer, quien ha demostrado el amor por lo que hace diariamente; sus consejos, paciencia, guia y profesionalismo nos han ayudado a nosotros ,como estudiantes, a cuestionarnos sobre la calidad de profesionales que queremos llegar a ser en un futuro. Para él, todo mi respeto y admiración.

A mis grandes amigos Ana María Pinzón, Camila Moreno, Johan Corpus y Gabriel Moreno, por aguantarme durante estos años, por enseñarme y acompañarme en infinitos momentos, horas de estudio, de café y de consejos, por dejarme hacer parte de sus vidas y compartir conmigo su conocimiento. Para todos ellos, espero que nunca olviden lo mucho que los aprecio.

.

Referencias

[1] Agresti, A. (2012). *An Introduction to Categorical Data Analisys.* University of Florida.

[2] Domínguez, A.; Acuña, J.; Toro, E. (2012). "Una comparación entre métodos estadísticos clásicos y técnicas metaheurísticas en el modelamiento estadístico". Recuperado de http://revistas.utp.edu.co/index.php/revistaciencia/article/view/1647

[3] Hachuel, L.; Boggio, G.; Wojdyla, D.; Servy, E.; (2005). "Interpretación y Comparación de Modelos de Regresión Logística Para el Estudio de la Desocupación". Recuperado de https://www.fcecon.unr.edu.ar/web-nueva/sites/default/files/u16/Decimocuartas/hachuel %20y %20boggio %20interpretacion %20y %20comparacion.pdf.

[4] Hosmer, D. , Lemeshow, S. , Sturdivant, R. (2013). *Applied Logistic Regression.* John Wiley & Sons, Inc.

[5] Joyce J.M. (2011).*Kullback-Leibler Divergence. In: Lovric M. (eds).* International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg

[6] Ladino, A. (2014). "Comparación de Modelos de Riesgo de Crédito: Modelos Logísticos y Redes Neuronales". Recuperado de https://repository.javeriana.edu.co/bitstream/handle/10554/14857/LadinoBecerraIvanCamilo2014.pdf?sequence=1& isAllowed=y.

[7] Lizares, M. (2017). "Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico". Recuperado de http://cybertesis.unmsm.edu.pe/bitstream/handle/cybertesis/7122/Lizares_cm.pdf?sequence=1& isAllowed=y.

[8] Llaugel, F.; Fernández, A (2012). "Evaluación del uso de modelos de regresión logística para el diagnóstico de instituciones financieras". Recuperado de https://repositoriobiblioteca.intec.edu.do/handle/123456789/1376.

[9] Márquez, L.; Díaz, M.; Ortiz, D. (2016). "Transferibilidad geográfica de modelos de generación de viajes urbanos: comparación de modelos de regresión y tasas de viajes para algunas ciudades colombianas". Recuperado de http://rcientificas.uninorte.edu.co/index.php/ingenieria/article/viewArticle/6772.

[10] Moutinho, G.; Demétrio, C. (2008).*Modelos Lineares Generalizados e Extensões.* Departamento de Estatística e Informática, UFRPE.

[11] Pineda, W.; Giraldo, R.; Porcu, E. (2019). "Functional SAR models: With application to spatial econometrics". Spatial Statistics, Volume 29 (p 145- 159). Recuperado de https://www.sciencedirect.com/science/article/abs/pii/S2211675318300617#sec4.

[12] Vuong, Q. (1989). *Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses.* Recuperado de https://pdfs.semanticscholar.org/af77/7130dcd810eac080e86e25f5f47b88854e31.pdf