Estimación de Modelos Logit Multinomiales con Variables Endógenas: Dos Nuevos Enfoques

Louis de Grange y Felipe González
Escuela de Ingeniería Civil Industrial, Universidad Diego Portales, Santiago de Chile.
Phone: (56-2) 2676 0469; e-mail: **Louis.degrange@udp.cl , Felipe.gonzalezr@udp.cl**

Matthieu Marechal
Instituto de Ciencias Básicas, Universidad Diego Portales, Santiago de Chile.
e-mail: **Matthieu.marechal@udp.cl**

Rodrigo Troncoso
Facultad de Gobierno, Universidad del Desarrollo, Santiago de Chile.
e-mail: **Rtroncoso@udd.cl**

RESUMEN

En este trabajo presentamos dos nuevos enfoques que permiten obtener estimadores con propiedades de consistencia para los parámetros de modelos Logit Multinomiales que incluyan variables explicativas endógenas. Ambos enfoques se basan en el uso de variables instrumentales. El primer enfoque corresponde a condiciones de momento que incluyen variables instrumentales. El segundo enfoque considera combinar parámetros obtenidos durante dos etapas diferentes de la estimación, y también el uso de variables instrumentales. Ambos nuevos enfoques los implementamos usando datos simulados, y los comparamos con el clásico método de Función de Control. Como resultado, obtenemos estimadores similares entre los tres métodos para los parámetros de variables explicativas; sin embargo, al estimar los términos constantes del modelo (constantes modales), nuestros nuevos enfoques proporcionaron estimaciones mucho más precisas. Esto último tiene consecuencias en la capacidad predictiva de los modelos, y también en la estimación de efectos marginales, elasticidades y beneficios sociales (excedente del consumidor).

Key words: logit multinomial; endogeneidad; variables instrumentales; momentos; dos etapas; función de control; simulación; predicción; evaluación de proyectos.

1. INTRODUCCIÓN

La presencia de variables endógenas en modelos econométricos tiene graves consecuencias tanto en la estimación de parámetros como en la construcción de contrastes estadísticos. En este trabajo exponemos dos nuevos enfoques que permiten obtener estimadores con propiedades de consistencia para los parámetros de modelos Logit Multinomiales que incluyan variables explicativas endógenas.

El primer enfoque se basa en la formulación de condiciones de momentos que incorporan el uso de variables instrumentales. El segundo enfoque se basa en el uso de parámetros estimados en dos etapas diferentes, los cuales, al combinarlos, permiten obtener también estimadores con propiedades de consistencia para los parámetros de modelos Logit Multinomiales. Ambos enfoques los implementamos usando datos simulados para un caso con tres alternativas de transporte (auto, metro y caminata). Primero consideramos solo una única variable explicativa endógena (tiempo de viaje en auto), y luego consideramos el caso con dos variables explicativas (una endógena (tiempo de viaje en auto) y una exógena (costo de viaje)). Comparamos los resultados con el método de Función de Control.

Como resultado, obtenemos estimadores similares entre los tres métodos para los parámetros de variables explicativas (e.g. tiempo de viaje y costo de viaje); sin embargo, al estimar los términos constantes del modelo (o constantes modales), nuestros nuevos enfoques proporcionaron estimaciones significativamente mejores. Esto último puede tener importantes consecuencias al usar estos modelos en fase predictiva, por ejemplo, para estimar demandas frente a proyectos o políticas de transporte que modifiquen los niveles de servicio, y también en la estimación de beneficios sociales (excedente del consumidor) de dichos proyectos o políticas. También las estimaciones de efectos marginales y elasticidades (propias y cruzadas) presentaron diferencias. Los resultados obtenidos sugieren que nuestros nuevos enfoques propuestos serían preferibles frente al clásico método de función de control.

En la sección 2 presentamos una revisión bibliográfica con los principales y más recientes trabajos desarrollados en el tema, enfatizando el método de función de control, que es el más utilizado. En la sección 3 exponemos los dos nuevos enfoques que proponemos para estimar modelos MNL con variables endógenas. En la sección 4 reportamos, con datos simulados, los resultados de nuestros dos nuevos enfoques respecto a la estimación tradicional con Máxima Verosimilitud (ML) que no corrige endogeneidad, y con respecto al método de función de control que sí corrige la endogeneidad. En la sección 5 presentamos un análisis complementario en contexto de usos predictivos de estos modelos, a fin de cuantificar las diferencias entre los enfoques expuestos en la sección 4. Finalmente, en la sección 6 presentamos las principales conclusiones y recomendaciones del trabajo desarrollado.

2. REVISIÓN DE LA LITERATURA

La formulación de modelos de elección discreta de tipo Logit se puede obtener mediante dos enfoques alternativos: el primero es basado en la teoría de la utilidad aleatoria (McFadden, 1974; Williams, 1977; Train, 2003; Ortúzar y Willumsen, 2011), y el segundo es basado en la formulación de problemas de optimización de máxima entropía (Anas, 1983; De Cea et al., 2008; Donoso y De Grange, 2010; Donoso et al. 2011).

En los modelos de utilidad aleatoria (RUM) el individuo *i* enfrenta un conjunto de alternativas, y escoge la alternativa que le entregue la mayor utilidad. De esta forma, el individuo escogerá la alternativa *m* cuando $U_i^m > U_i^{m'} \; \forall m' \neq m$. La función de utilidad U_i^m se suele descomponer en forma aditiva: $U_i^m = V_i^m + \varepsilon_i^m$, donde V_i^m representa la parte determinística de la utilidad, que depende de variables observables, y ε_i^m representa la parte aleatoria, no observable.

El modelador no observa la función de utilidad U_i^m del individuo, pero sí puede observar las elecciones efectuadas por los individuos, así como los atributos de cada una de las alternativas que enfrenta y que definen V_i^m; el *k*-ésimo atributo o variable explicativa que enfrenta el individuo *i* en la alternativa *m* lo podemos definir como $x_{ki}^m \; \forall i, m, k$. Normalmente se define una función lineal en los atributos para el componente determinístico de la utilidad, es decir, $V_i^m = \sum_k \beta_k^m x_{ki}^m$, donde β_k^m son parámetros a estimar y representan los pesos relativos de cada atributo.

El modelo Logit Multinomial (MNL) se obtiene suponiendo que el componente aleatorio de las funciones de utilidad distribuye Gumbel, y que son además independientes e idénticamente distribuidos (McFadden, 1974; Ben-Akiva y Lerman, 1985; Train, 1986, 2003; Ortúzar y Willumsen, 2011). Por lo tanto, modelos Logit Multinomiales basados en la teoría de la utilidad aleatoria parten de la siguiente premisa:

$$U_i^m \geq U_i^{m'} \quad \forall \; m \neq m' \tag{1}$$

$$U_i^m = \sum_k \beta_k^m x_{ki}^m + \varepsilon_i^m \tag{2}$$

Si las variables x_{ki}^m son exógenas y, además, los ε_i^m se suponen independientes e idénticamente distribuidos Gumbel, entonces se obtiene el clásico modelo MNL, donde P_i^m es la probabilidad de que el individuo *i* selecciones la alternativa *m*:

$$P_i^m = \frac{e^{\sum_k \beta_k^m x_{ki}^m}}{\sum_{m'} e^{\sum_k \beta_k^{m'} x_{ki}^{m'}}} \tag{3}$$

Sin embargo, cuando algunas de las variables x_{ki}^m presentan endogeneidad, es decir, $corr\left(\varepsilon_i^m; x_{ki}^m\right) \neq 0$, la estimación de los parámetros β_k^m del modelo (3) es inconsistente, entregando parámetros incorrectos, y distorsionando los contrastes estadísticos habituales (Berry et al., 1995; Louviere et al., 2005; Guevara y Ben-Akiva, 2009; Walker et al., 2011).

La endogeneidad puede presentarse por diferentes razones, lo que puede depender del fenómeno estudiado, de los datos disponibles, del enfoque de modelado, entre otros. Para una revisión reciente del fenómeno de endogeneidad en modelos MNL se puede consultar a Guevara (2023).

En la literatura se han propuesto distintas maneras de abordar la endogeneidad en modelos de elección discreta. Blundell y Powell (2004) proponen un enfoque semiparamétrico para testear la exogeneidad de variables explicativas continuas en modelos de elección binaria. La estimación por Máxima Verosimilitud, comúnmente usada en estos modelos, requiere de una especificación paramétrica explícita de la forma en que cada variable endógena depende de un conjunto de instrumentos y de los errores. Además, en Máxima Verosimilitud también se requiere de una especificación de la distribución conjunta del componente aleatorio de las funciones de utilidad, así como del componente de error en la relación entre la variable endógena y los instrumentos (Lewbel, 2007). Esto representa un inconveniente del método de Maxíma Verosimilitud, ya que especificar correctamente estas relaciones puede resultar difícil. También hay trabajos (Zou y Cirillo, 2021).) en que la variable endógena es simplemente reemplazada por una estimación exógena de la misma, tal como sugiere Train (1986), aunque esto equivale a que la nueva variable explicativa estimada exógenamente sea medida con error.

Entre los distintos enfoques que se han propuesto para corregir la endogeneidad, el más utilizado corresponde al enfoque de Función de Control. Este método está descrito en Heckman (1976), Hausman (1978), Villas-Boas y Winer (1999), Blundell y Powell (2004), Guevara y Ben-Akiva (2006), Petrin y Train (2010), Guevara (2023). El método considera básicamente 2 etapas: primero, la variable endógena es regresionada contra instrumentos exógenos, y el residuo de esta regresión (o una función de dicho residuo) es incorporada como una nueva variable explicativa en la función de utilidad; esta nueva variable es denominada Función de Control (Louviere et al, 2005). Luego, incorporando esta función de control en la especificación original, el problema de endogeneidad puede ser corregido (Guevara y Ben-Akiva, 2009). Al igual que máxima verosimilitud, la estimación mediante función de control requiere especificar correctamente la relación entre los regresores endógenos y sus respectivos instrumentos.

Los dos nuevos enfoques basados en condiciones de momentos y en dos etapas, que proponemos en este trabajo, son alternativas que compiten con el método de Función de Control, que correspone hasta hoy al enfoque más utilizado que busca corregir sesgos por presencia de variables explicativas endógenas.

3. MÉTODOS DE ESTIMACIÓN

Para corregir los problemas de endogeneidad en modelos econométricos, incluyendo modelos MNL, es habitual el uso de variables instrumentales. Una variable instrumental se caracteriza por ser una variable exógena pero altamente correlacionada con la variable explicativa endógena.

Cuando tenemos instrumentos disponibles, en términos generales, podemos expresar las variables endógenas en función de las variables exógenas. Suponiendo una relación lineal, para una variable x_{qi}^{m} endógena, podemos escribir la siguiente expresión:

$$x_{qi}^{m} = \alpha_{0q}^{m} + \sum_{p} \alpha_{qp}^{m} z_{qpi}^{m} + \eta_{qi}^{m} \tag{4}$$

donde z_{qpi}^{m} son variables exógenas, que habitualmente incluyen a las variables exógenas del modelo y a los instrumentos. Los coeficientes α's son parámetros, y η_{qi}^{m} es una variable aleatoria con media 0 y varianza $\sigma_{\eta q}^{2}$, que se correlaciona con la variable ε_{i}^{m} de la ecuación (2), tal que $corr\left(\varepsilon_{i}^{m};\eta_{qi}^{m}\right)=\theta_{q}$. Nótese que x_{qi}^{m} y z_{qi}^{m} son variables observables, por lo que mediante una regresión lineal podemos estimar $\hat{\alpha}'s$, $\hat{\eta}_{iq}^{m}$ y $\hat{\sigma}_{\eta q}^{2}$ de la ecuación (5).

A partir de la estimación de la regresión (5) podemos construir un instrumento $\hat{x}_{qi}^{m}$ para la variable endógena x_{qi}^{m}, instrumento que es el que consideramos en los dos nuevos enfoques de estimación que proponemos. Esto es similar a la primera etapa de la estimación por Two Stages Least Squares (2SLS) para modelos lineales (Boonekamp et al., 2018). Nótese que para las variables x_{ki}^{m} exógenas, su instrumento $\hat{x}_{ki}^{m}$ de acuerdo con la ecuación (4) es ella misma.

Brevemente, en la sección 3.1 presentamos el clásico método basado en la Función de Control, en la sección 3.2 presentamos nuestro enfoque basado en nuevas condiciones de momento, y en la sección 3.3 presentamos un método basado en 2 etapas.

3.1 Método de Función de Control (CF)

Para el caso de modelos MNL, las variables instrumentales son usadas para construir una Función de Control que se adiciona como parte de la función de utilidad del individuo. Específicamente, la Función de Control utiliza el estimador $\hat{\eta}_{i}^{m}$ obtenido de la regresión (5) para luego incorporarlo en la función de utilidad del individuo. De esta forma, la función de probabilidad (3) se modifica de la siguiente manera:

$$P_{i}^{m}=\frac{e^{\sum_{k}\beta_{k}^{m}x_{ki}^{m}+\sum_{q}\gamma_{q}^{m}\hat{\eta}_{qi}^{m}}}{\sum_{m'}e^{\sum_{k}\beta_{k}^{m'}x_{ki}^{m'}+\sum_{q}\gamma_{q}^{m'}\hat{\eta}_{qi}^{m'}}} \tag{6}$$

donde $\sum_{q}\gamma_{k}^{m}\eta_{ki}^{m}$ es la Función de Control. Notar que podrían incluirse términos no lineales de los residuos $\hat{\eta}_{qi}^{m}$. Luego, a partir de máxima verosimilitud, se obtienen los parámetros $\hat{\beta}_{k}^{m}$ que presentan propiedades de consistencia.

3.2 Método de Momentos (MM)

El enfoque de momentos se basa en explotar condiciones poblacionales de momentos, cuyas contrapartes muestrales puedan expresarse en función de los datos y los parámetros del modelo. Cuando las variables explicativas son exógenas, las condiciones de momento o de ortogonalidad usadas para identificar los parámetros del MNL corresponden a:

$$E\left(\left[\delta_{i}^{m}-P_{i}^{m}\right]x_{ki}^{m}\right)=0,\quad\forall k,m \tag{6}$$

De las cuales se obtienen las condiciones de primer orden del problema. Cuando hay endogeneidad, las condiciones (6) no se cumplen, por lo que su uso producirá estimadores sesgados e inconsistentes. Sin embargo, usando los instrumentos exógenos construidos en (4) es posible plantear condiciones de momento que sí se cumplen:

$$E\left(\left[\delta_i^m - P_i^m\right]\hat{x}_{ki}^m\right) = 0, \quad \forall k, m \tag{7}$$

La contraparte muestral de la condición (7) es:

$$\sum_{i=1}^{n}\left[\delta_i^m - P_i^m\right]\hat{x}_{ki}^m = 0, \quad \forall k, m \tag{8}$$

Resolviendo el sistema de ecuaciones definido por (8), es posible obtener los estimadores consistentes de los parámetros $\hat{\beta}_k^m$.

3.3 Método de 2 Etapas (2E)

Este método es similar al método de Función de Control, pero a diferencia de este último, nuestro método de 2 etapas (2S) usa directamente el instrumento $\hat{x}_{qi}^m$ construido en (4) como variable explicativa en el modelo MNL en lugar de las variables x_{qi}^m endógenas.

Para entender la intuición detrás de este reemplazo, consideremos un modelo simple con una sola variable explicativa que es endógena. Las funciones de utilidad en este caso son las siguientes:

$$U_i^m = \beta_0^m + \beta_1^m x_i^m + \varepsilon_i^m \tag{9}$$

Asumiendo que la variable x_i^m es endógena, y que disponemos de algún instrumento z_i^m para dicha variable, podemos escribir la siguiente expresión:

$$x_i^m = \alpha_0^m + \alpha_1^m z_i^m + \eta_i^m \tag{10}$$

Luego, reemplazando (10) dentro de (9) obtenemos:

$$U_i^m = \beta_0^m + \beta_1^m x_i^m + \varepsilon_i^m = \beta_0^m + \beta_1^m\left(\alpha_0^m + \alpha_1^m z_i^m + \eta_i^m\right) + \varepsilon_i^m \tag{11}$$

$$U_i^m = \beta_0^m + \beta_1^m x_i^m + \varepsilon_i^m = \underbrace{\beta_0^m + \alpha_0^m}_{b_0^m} + \underbrace{\beta_1^m \alpha_1^m}_{b_1^m} z_i^m + \underbrace{\varepsilon_i^m + \beta_1^m \eta_i^m}_{v_i^m} \tag{12}$$

$$U_i^m = b_0^m + b_1^m z_i^m + v_i^m \tag{13}$$

Bajo el supuesto de que los v_i^m son independientes e idénticamente distribuidos Gumbel (de manera análoga a lo que se asume en el método de Función de Control cuando se usa Máxima Verosimilitud), obtenemos el modelo siguiente para las probabilidades:

$$\tilde{P}_i^m = \frac{e^{b_0^m + b_1^m z_i^m}}{\sum_{m'} e^{b_0^{m'} + b_1^{m'} z_i^{m'}}} \tag{14}$$

La etapa 1 de este método corresponde a la estimación de los parámetros $\left(\hat{b}_0^m ; \hat{b}_1^m\right)$ del modelo (14), los que pueden ser obtenidos directamente por máxima verosimilitud.

En una segunda etapa, estima el modelo de regresión (10), podemos obtener los parámetros α 's:

$$x_i^m = \alpha_0^m + \alpha_1^m z_i^m + \eta_i^m \rightarrow \left(\hat{\alpha}_0^m ; \hat{\alpha}_1^m\right) \tag{15}$$

Como sabemos que $b_0^m = \beta_0^m + \alpha_0^m$ y que $b_1^m = \beta_1^m \alpha_1^m$, podemos expresar los parámetros del modelo β_0^m y β_1^m en función de los b's y α 's de la siguiente manera:

$$\beta_0^m = b_0^m - \alpha_0^m \qquad ; \qquad \beta_1^m = b_1^m / \alpha_1^m \tag{16}$$

Por último, se reemplazan los estimadores $\left(\hat{b}_0^m ; \hat{b}_1^m\right)$ y $\left(\hat{\alpha}_0^m ; \hat{\alpha}_1^m\right)$ en las expresiones de (16) para obtener los estimadores de $\left(\hat{\beta}_0^m ; \hat{\beta}_1^m\right)$:

$$\hat{\beta}_0^m = \hat{b}_0^m - \hat{\alpha}_0^m \qquad ; \qquad \hat{\beta}_1^m = \hat{b}_1^m / \hat{\alpha}_1^m \tag{17}$$

Las desviaciones estándar o varianzas de $\hat{\beta}_0^m$ y $\hat{\beta}_1^m$ las podemos obtener aplicando el método delta. La generalización de este método de 2 etapas para el caso con múltiples variables endógenas es simple y se resume a continuación:

Etapa 1: plantear el modelo de MNL usando los instrumentos en reemplazo de las variables endógenas (recordar que, para las variables exógenas, el instrumento respectivo son ellas mismas):

$$\tilde{P}_i^m = \frac{e^{\sum_k b_k^m z_{ki}^m}}{\sum_{m'} e^{\sum_k b_k^{m'} z_{ki}^{m'}}} \qquad \text{ó} \qquad \tilde{P}_i^m = \frac{e^{\sum_k b_k^m \hat{x}_{ki}^m}}{\sum_{m'} e^{\sum_k b_k^{m'} \hat{x}_{ki}^{m'}}} \tag{18}$$

Obtener mediante máxima verosimilitud los estimadores $\hat{b} = \left[\hat{b}_k^m\right]$ de (18).

Etapa 2: A partir de la relación entre variables endógenas x_{qi}^m y sus respectivos instrumentos z_{qi}^m ó $\hat{x}_{qi}^m$, obtener los estimadores $\hat{\alpha} = \left[\hat{\alpha}_q^m\right]$ mediante regresión lineal. Finalmente, usando los estmadores de $\hat{b}$ y $\hat{\alpha}$, es posible recuperar los estimadores $\hat{\beta} = \left[\hat{\beta}_k^m\right]$ y sus respectivas desviaciones estándar mediante el método delta.

4. ANÁLISIS DE RESULTADOS MEDIANTE SIMULACIÓN

Para analizar y comparar las propiedades de estimación de nuestros dos nuevos enfoques, y evaluarlas respecto del método de Función de Control, consideramos dos escenarios de simulación: un primer escenario con solo una variable explicativa (tiempo de viaje), y un segundo escenario con dos variables explicativas (tiempo de viaje y costo de viaje). Para ambos escenarios, consideramos tres alternativas o modos de transporte: Auto, Metro y Caminata. Si bien se trata de datos simulados y no de situaciones reales, el contexto de la elección, así como los órdenes de magnitud de los parámetros del modelo generador de datos hacen más sencillo exponer y evaluar los efectos de la endogeneidad, así como el desempeño de los estimadores alternativos. En ambos escenarios, la variable endógena fue el tiempo de viaje en auto (debido a que, al aumentar la demanda o uso del auto, aumenta el tiempo de viaje producto de congestión; por lo tanto, al aparecer un shock de demanda en el uso del auto, también afectará a la variable explicativa tiempo de viaje). Para el Metro y la Caminata, el tiempo de viaje se supone exógeno. En la Tabla 1 se resumen las características de las variables y modos considerados en la simulación:

Table 1
Variables Explicativas para los 2 Escenarios de Simulación

	Escenario 1	**Escenario 2**	
Modo	**Tiempo de Viaje**	**Tiempo de Viaje**	**Costo de Viaje**
Auto	**Endógeno**	**Endógeno**	Exógeno
Metro	Exógeno	Exógeno	Exógeno
Caminata	Exógeno	Exógeno	No Aplica

En la sección 4.1 exponemos los parámetros y supuestos de la simulación para el caso univariado (solo tiempos de viaje). En la sección 4.2 exponemos los parámetros y supuestos de la simulación para el caso bivariado (con tiempos de viaje y costos de viaje). La estimación de los parámetros la realizamos mediante cuatro métodos: Máxima Verosimilitud sin corregir endogeneidad (ML), método de Función de Control (CF, que sí corrige la endogeneidad), y nuestros dos nuevos métodos de Momentos (MM) y Dos Etapas (2E), que también corrigen la endogeneidad.

4.1 Simulación para el Modelo Simple Univariado (Escenario 1)

La función de utilidad para cada modo es la siguiente:

$$U_i^{auto} = \beta_0^{auto} + \beta_1 \cdot tiempo_i^{auto} + \varepsilon_i^{auto} \tag{19}$$

$$U_i^{metro} = \beta_0^{metro} + \beta_1 \cdot tiempo_i^{metro} + \varepsilon_i^{metro} \tag{20}$$

$$U_i^{cam} = \beta_0^{cam} + \beta_1 \cdot tiempo_i^{cam} + \varepsilon_i^{cam} \tag{21}$$

Consideramos que $\beta_0^{auto} = 0$, $\beta_0^{metro} = 0,7$, $\beta_0^{cam} = 0,4$ y $\beta_1 = -0,02$. Para simular los tiempos de viaje de cada modo, supusimos que la velocidad del auto (sin congestión) es de 25 km/h, la del metro es 35 km/h y para la caminata 3 km/h. Las distancias de viaje D_i

varían de manera uniforme entre 2 y 30 kilómetros. Luego, para cada individuo/modo, las distancias de viaje las simulamos de la siguiente manera:

$$D_i^{metro} = 2 + 28 \cdot u_i \tag{22}$$

$$D_i^{auto} = D_i^{metro} + w_i^{auto} \tag{23}$$

$$D_i^{cam} = D_i^{metro} + w_i^{cam} \tag{24}$$

donde u_i es una distribución uniforme entre 0 y 1. Además, w_i^{auto} y w_i^{cam} son errores con distribuciones normales estándar independientes. Una vez simuladas las distancias de viaje, las dejamos fijas y exógenas para cada individuo. A partir de las distancias de viaje definidas en (22), (23) y (24), simulamos los tiempos de viaje de la siguiente manera:

$$tiempo_i^{auto} = \frac{D_i^{auto}}{25} + v_i^{auto} + \eta_i^{auto} \tag{25}$$

$$tiempo_i^{metro} = \frac{D_i^{metro}}{35} + v_i^{metro} \tag{26}$$

$$tiempo_i^{cam} = \frac{D_i^{cam}}{3} + v_i^{cam} \tag{27}$$

donde v_i^{metro}, v_i^{auto} y v_i^{cam} son errores con distribuciones normales estándar independientes. Además, consideramos que $\eta_i^{auto} \sim N\left(0; \sigma_\eta^2\right)$ y que $corr\left(\eta_i^{auto}; \varepsilon_i^{auto}\right) = \theta = 0,7$. Las variables ε_i^{metro} y ε_i^{cam} consideramos que distribuyen gumbel con parámetros 0 y 1. También consideramos que $\left(\beta_1 \eta_i^{auto} + \varepsilon_i^{auto}\right)$ también distribuye gumbel con parámetros 0 y 1. Realizamos 2.000 simulaciones con 1.000 datos en cada simulación. El instrumento para el tiempo de viaje en auto es la distancia de viaje en auto.

En la Tabla 2 se muestra los valores promedios de los tres parámetros estimados $\left(\beta_0^{metro}, \beta_0^{cam}, \beta_1\right)$ y sus desviaciones estándar, considerando el enfoque de estimación de Máxima Verosimilitud clásico que no corrige la endogeneidad (ML), el método de Función de Control (CF), y nuestros dos nuevos enfoques: Momentos (MM) y 2 Etapas (2E). En la Tabla 3 se reportan, para los mismos enfoques de estimación, el sesgo y error cuadrático medio (ECM) para cada uno de los parámetros estimados.

De la Tabla 2 se observa claramente que los enfoques CF, MM y 2S reproducen de manera prácticamente exacta los valores simulados del parámetro que acompaña a la variable endógena “tiempo de viaje” (sesgos casi nulos), confirmando así propiedades de consistencia en el resultado (a diferencia de los valores obtenidos mediante ML, donde el sesgo es evidente, incluso obteniendo signo positivo en el parámetro).

Table 2
Resultados para Media y Desviación Estándar, ML vs CF vs 2E vs MM, Escenario 1

	Media				Desv. Estándar			
Parámetro	**ML**	**CF**	**MM**	**2E**	**ML**	**CF**	**MM**	**2E**
β_0^{metro}	0,70870	0,75651	0,69892	0,69849	0,08381	0,09431	0,08438	0,08431
β_0^{cam}	0,14926	0,45525	0,39700	0,39654	0,14757	0,15289	0,14757	0,14711
β_1	0,03357	-0,01926	-0,01912	-0,01912	0,02332	0,02563	0,02555	0,02555
γ	-	0,43725	-	-	-	0,07472	-	-

Table 3
Resultados para Sesgo y ECM, ML vs CF vs 2E vs MM, Escenario 1

	Sesgo				ECM			
Parámetro	**ML**	**CF**	**MM**	**2E**	**ML**	**CF**	**MM**	**2E**
β_0^{metro}	0,00870	0,05651	-0,00108	-0,00151	0,00710	0,01209	0,00712	0,00711
β_0^{cam}	-0,25074	0,05525	-0,00300	-0,00346	0,08465	0,02643	0,02179	0,02165
β_1	0,05357	0,00074	0,00088	0,00088	0,00341	0,00066	0,00065	0,00065

De la Tabla 3 se aprecia que los resultados asociados al parámetro β_1 son similares entre los tres enfoques que corrigen endogeneidad (CF, MM y 2S). Sin embargo, se observan diferencias en la estimación de las constantes modales $\left(\beta_0^{metro}, \beta_0^{cam}\right)$. Dejando fuera del análisis los resultados obtenidos por ML, observamos que las constantes modales estimadas por CF son diferentes (mayores) que las estimadas por MM y por 2E (estas últimas resultaron ser prácticamente iguales a los parámetros simulados: $\beta_0^{metro} = 0,7$ y $\beta_0^{cam} = 0,4$).

Esta diferencia se explica principalmente por la presencia del término adicional $\left(\gamma\hat{\eta}_i^m\right)$ que se incluye en el enfoque de Función de Control, el que distorsiona la estimación de los términos constantes. Esto se debe a que el término adicional, cuando existe endogeneidad, sí tiene significancia estadística, por lo que el estimador de $\hat{\gamma}$ es distinto de cero. Luego, al ser el MNL un modelo no lineal, la presencia del término $\gamma\hat{\eta}_i^m$ inducirá cambios en las medias de los estimadores de β_0^m, incluso si el promedio de los $\hat{\eta}_i^m$ es igual a cero (e.g. $E\left(\eta_i^m\right) = 0$).

Adicionalmente, la varianza en la estimación de las constantes modales fue mayor bajo el enfoque de CF que bajo los enfoques MM y 2S. Por lo tanto, nuestros dos nuevos enfoques estimaron de mejor manera, tanto en términos de sesgo y de eficiencia, las constantes modales.

4.2 Simulación para el Modelo Simple Bivariado (Escenario 2)

La función de utilidad para cada modo es la siguiente:

$$U_i^{auto} = \beta_0^{auto} + \beta_1 \cdot tiempo_i^{auto} + \beta_2 \cdot costo_i^{auto} + \varepsilon_i^{auto} \quad (28)$$

$$U_i^{metro} = \beta_0^{metro} + \beta_1 \cdot tiempo_i^{metro} + \beta_2 \cdot costo_i^{metro} + \varepsilon_i^{metro} \quad (29)$$

$$U_i^{cam} = \beta_0^{cam} + \beta_1 \cdot tiempo_i^{cam} + \varepsilon_i^{cam} \quad (30)$$

Consideramos que $\beta_0^{auto} = 0$, $\beta_0^{metro} = 0,7$, $\beta_0^{cam} = 0,4$, $\beta_1 = -0,02$, $\beta_2 = -0,04$ y $\theta = 0,7$.

La simulación de las distancias y tiempos de viaje fue la misma a la expuesta en la sección 4.1 para el caso univariado.

La construcción de la variable costo de viaje para el auto y el metro fue la siguiente: el costo del auto lo definimos en \$0,2 por cada kilómetro; el costo del metro se define de \$1 para los primeros 10 km y luego \$0,5 adicional por cada 10 km adicionales. De esta forma, existe correlación entre las variables tiempo y costo (como es habitual en modelos de transporte), pero no colinealidad perfecta, por lo que ambos parámetros se pueden identificar.

Realizamos 2.000 simulaciones con 1.000 datos a cada simulación. El instrumento para el tiempo de viaje en auto es la distancia de viaje en auto.

En la Tabla 4 se muestra los valores promedios de los tres parámetros estimados $\left(\beta_0^{metro}, \beta_0^{cam}, \beta_1, \beta_2\right)$ y sus desviaciones estándar. En la Tabla 5 se reportan el sesgo y ECM para cada uno de los parámetros estimados.

Table 4
Resultados para Media y Desviación Estándar, ML vs CF vs 2E vs MM, Escenario 2

	Media				Desv. Estándar			
Parámetro	**ML**	**CF**	**MM**	**2E**	**ML**	**CF**	**MM**	**2E**
β_0^{metro}	0,83711	0,76826	0,70640	0,70602	0,13647	0,14439	0,13963	0,13949
β_0^{cam}	0,32700	0,47117	0,41010	0,40972	0,20622	0,20435	0,19976	0,19933
β_1	0,04565	-0,02076	-0,02060	-0,02060	0,02827	0,03070	0,03057	0,03056
β_2	0,03233	-0,04028	-0,03927	-0,03925	0,06317	0,06644	0,06422	0,06420
γ	-	0,45012	-	-	-	0,08051	-	-

Table 5
Resultados para Sesgo y ECM, ML vs CF vs 2E vs MM, Escenario 2

	Sesgo				ECM			
Parámetro	**ML**	**CF**	**MM**	**2E**	**ML**	**CF**	**MM**	**2E**
β_0^{metro}	0,13711	0,06826	0,00640	0,00602	0,03742	0,02551	0,01954	0,01949
β_0^{cam}	-0,07300	0,07117	0,01010	0,00972	0,04786	0,04683	0,04000	0,03983
β_1	0,06565	-0,00076	-0,00060	-0,00060	0,00511	0,00094	0,00093	0,00093
β_2	0,07233	-0,00028	0,00073	0,00075	0,00922	0,00441	0,00413	0,00412

Los resultados de las Tablas 4 y 5 son consistentes con los obtenidos en las Tablas 2 y 3: la estimación de los parámetros que acompañan a las variables explicativas son sesgados al usar ML, mientras que bajo los enfoques CF, MM y 2E presentan buenas propiedades de consistencia.

Adicionalmente, la estimación de las constantes modales, tal como ocurrió para el caso con una sola variable explicativa (ver Tabla 2), presenta sesgo, aspecto que es superado por nuestros nuevos enfoques MM y 2E. Adicionalmente, y tal como se observó en el Escenario 1, nuestros dos nuevos enfoques estimaron mejor tanto en términos de sesgo y de eficiencia, las constantes modales. En las

5. ANÁLISIS EN CONTEXTOS PREDICTIVOS Y DE EVALUACIÓN DE PROYECTOS

Los resultados expuestos en la sección 4 mostraron que los valores estimados para los parámetros que acompañan a las variables explicativas (e.g. tiempo de viaje y costo de viaje) resultaron ser muy similares entre los tres métodos comparados. Sin embargo, los estimadores obtenidos para los términos constantes o constantes modales $\left(\beta_0^m\right)$, fueron distintos a los definidos en las simulaciones, especialmente los obtenidos mediante el enfoque de Función de Control (dejando fuera del análisis el enfoque ML).

Como se adelantó en la sección 4.1, esta diferencia se explica principalmente por la presencia del término $\gamma\hat{\eta}_i^m$ que se incluye en el enfoque de Función de Control.

Por lo tanto, si se pretende usar el modelo MNL en fase predictiva, los modelos estimados bajo el enfoque de Función de Control pueden entregar resultados incorrectos. Algo similar puede ocurrir al estimar los efectos marginals y elasticidades.

Supongamos, por ejemplo, un proyecto de infraestructura o una política de gestion operacional que reduce en un 50% el tiempo de viaje en auto, y se require estimar el impacto sobre la demanda usando los modelos calibrados en la sección 4 anterior.

Para este simple ejemplo, consideramos un tiempo de viaje en auto, previo a la reducción del 50%, de 40 minutos. Posterior al proyecto implementado, este tiempo de viaje sería de 20 minutos. Para los modos Metro y Caminata, el tiempo de viaje de considera fijo en 25 y 40 minutos respectivamente.

Este cambio en el tiempo de viaje generará cambios en las proporciones de mercado, así como también cambios en el excedente de los consumidores. El excedente de los consumidores puede estimarse usando la Expected Maximum Utility (*EMU*) para las situaciones antes y después del proyecto (40 min vs 20 min). Analíticamente, el *EMU* para un determinado individuo *i* lo podemos escribir de la siguiente manera (Williams, 1977; Ortúzar y Willumnsen, 2011):

$$EMU_i = \ln\left(\sum_m e^{\sum_k \beta_k^m x_{ki}^m}\right) \tag{31}$$

La expresión (31) puede ser evaluada con los valores de las variables explicativas (e.g. tiempo y costo de viaje) antes del proyecto, y también con los valores después del proyecto. De esta forma, la diferencia de *EMU* representa el cambio en el excedente de los consumidores asociados a la variación en las variables explicativas. Esto es una medida de beneficio social como resultado de implementar un proyecto o política de transporte.

Adicionalmente, también es posible identificar las diferencias en los efectos marginales y elasticidades (propias y cruzadas) entre los modelos estimados bajo los diferentes enfoques (CF, MM y 2E). Las expresiones analíticas para Efectos Marginales y Elasticidades (propias y cruzadas) son las siguientes (Ortúzar y Willumnsen, 2011):

$$\frac{\partial P_i^m}{\partial x_{ki}^m} = \beta_k^m P_i^m \left(1 - P_i^m\right) \tag{32}$$

$$\xi_{ki}^m = \frac{\partial P_i^m}{\partial x_{ki}^m} \frac{x_{ki}^m}{P_i^m} = \beta_k^m x_{ki}^m \left(1 - P_i^m\right) \tag{33}$$

$$\xi_{ki}^{m'} = \frac{\partial P_i^{m'}}{\partial x_{ki}^m} \frac{x_{ki}^m}{P_i^{m'}} = -\beta_k^m x_{ki}^m P_i^m \tag{34}$$

donde (32) es el efecto marginal, (33) es la elasticidad propia y (34) es la elasticidad cruzada para modelos MNL.

En la Tabla 6 se presenta, bajo los enfoques FC, MM y 2E, la proporción de mercado considerando el tiempo de viaje en auto original de 40 minutos, y la nueva proporción de mercado considerando la reducción a 20 minutos. También se reporta el cambio en el excedente de los consumidores, los efectos marginales, y las elasticidades propias y cruzadas:

Table 6
Resultados de Modelos en Fase Predictiva

	MM		2S		CF	
Indicador	**40 min**	**20 min**	**40 min**	**20 min**	**40 min**	**20 min**
P_i^{auto}	19,35%	26,02%	19,36%	26,03%	18,44%	24,95%
EMU	0,8775	0,9638	0,8772	0,9636	0,9200	1,0031
EM	-0,002984	-0,003680	-0,002984	-0,003680	-0,002896	-0,003605
Elasticidad	-0,61682	-0,28291	-0,61661	-0,28281	-0,62815	-0,28902
Elast. Cruzada	0,14802	0,09951	0,14804	0,09952	0,14208	0,09609

De la Tabla 6 se observa que la demanda predicha por el enfoque de Función de Control difiere a la demanda predicha por los otros dos enfoques (MM y 2E), que son prácticamente idénticos entre sí. En términos relativos o porcentuales, la diferencia entre los resultados obtenidos por el modelo estimado mediante Función de Control, difieren en cerca de un 5%.

6. CONCLUSIONES

En este trabajo presentamos dos nuevos enfoques que permiten obtener estimadores con propiedades de consistencia para los parámetros de modelos Logit Multinomiales (MNL) que incluyan variables explicativas endógenas.

El primer enfoque se basa en la formulación de condiciones de momentos, y el segundo enfoque combina parámetros estimados en dos etapas diferentes. Ambos enfoques fueron comparados con el clásico método de Función de Control, usando datos simulados en diferentes escenarios.

La principal conclusión que obtenemos es que la estimación de los parámetros que acompañan a las variables explicativas (ya sean endógenas y exógenas) se realizó de forma satisfactoria usando los dos nuevos enfoques. Estos resultados fueron bastante similares a los obtenidos por el método de función de control. Una segunda conclusión es que el método de Función de Control estima incorrectamente los parámetros correspondientes a las constantes modales. Esto puede tener consecuencias importantes cuando se pretende usar este tipo de modelos (MNL con variables endógenas) para predecir cambios en las demandas como consecuencia de variaciones en los atributos o variables explicativas del modelo. Adicionalmente, y como consecuencia de las incorrectas constantes modales obtenidas en el enfoque CF, también se generan distorsiones en la estimación de excedentes del consumidor. Para las simulaciones realizadas, al reducir un 50% el tiempo de viaje en automóvil, se detectaron diferencias de un 5% en la predicción de demanda cuando usamos nuestros enfoques respecto al método de función de control. Otra manifestación de esta diferencia en las constantes modales, se aprecia en la estimación imprecisa de los efectos marginales y elasticidades (propias y cruzadas). Este resultado es consistente con las diferencias detectadas en la predicción.

Pese a que las diferencias detectadas no parecieran ser demasiado grandes, igual es recomendable usar los dos nuevos enfoques propuestos en lugar del método de función de control al estimar modelos MNL con variables endógenas.

AGRADECIMIENTOS

Los autores agradecen los aportes y discusión con el Profesor Angelo Guevara.

REFERENCIAS

Anas, A. (1983). Discrete Choice Theory, Information Theory and the Multinomial Logit and Gravity Models. Transportation Research, 17B, 13-23.

Ben-Akiva, M. and Lerman, S. (1985). Discrete Choice Analysis: Theory and Application to Travel Demand. Cambridge, MA: The MIT Press.

Berry, S., Levinsohn, J. and A. Pakes. (1995). Automobile Prices in Market Equilibrium. Econometrica 63, 841-889.

Boonekamp, T.; Zuidberg, J. and Burghouwt, G. (2018). Determinants of air travel demand: The role of low-cost carriers, ethnic links and aviation-dependent employment. Transportation Research Part A: Policy and Practice, 112, 18-28. https://doi.org/10.1016/j.tra.2018.01.004.

De Cea, J.; Fernandez, J.E. and De Grange, L. (2008). Combined models with hierarchical demand choices: A multi-objective entropy optimization approach. Transport Reviews, 28, 415-438.

Donoso, P. and De Grange, L. (2010). A Microeconomic Interpretation of the Maximum Entropy Estimator of Multinomial Logit Models and Its Equivalence to the Maximum Likelihood Estimator. Entropy, 12, 2077-2084.

Donoso, P.; De Grange, L. and González, F. (2011). A Maximum Entropy Estimator for the Aggregate Hierarchical Logit Model. Entropy, 13, 1425-1445.

Guevara, C. A., and Ben-Akiva, M. (2006). Endogeneity in residential location choice models. Transportation research record, 1977(1), 60-66.

Guevara, C.A. and Ben-Akiva, M. (2009). Addressing Endogeneity in Discrete Choice Models: Assessing Control-Function and Latent-Variable Methods. Working Paper Series, MIT Portugal, TSI-SOTUR-09-03.

Guevara, C.A. (2023). Endogeneity in Discrete Choice Models. In Handbook of choice modelling (forthcoming). Edward Elgar Publishing.

Hausman, J. (1978). Specification Tests in Econometrics. Econometrica, 46, 1251-1272.

Heckman, J. J. (1976). Simultaneous Equation Models with both Continuous and Discrete Endogenous Variables With and Without Structural Shift in the Equations, in Steven Goldfeld and Richard Quandt (Eds.), Studies in Nonlinear Estimation, Ballinger.

Lewbel, A. (2007). Endogenous Selection or Treatment Model Estimation. Journal of Econometrics, 141, 777-806.

Louviere, J.; Train, K.; Ben-Akiva, M.; Bhat, C.; Brownstone, D.; Cameron, T.; Carson, C.; Deshazo, J.; Fiebig, D.; Greene, W.; Hensher, D. and Waldman, D. (2005). Recent Progress on Endogeneity in Choice Modeling. Marketing Letters, 16, 255-265.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. Frontiers in Econometrics, Zarembka P., Ed.; Academic Press: New York, NY, USA.

Petrin, A. and Train, K. (2010). A control function approach to endogeneity in consumer choice models. Journal of Marketing Research, 47, 370-379.

Train, K. (1986). Qualitative Choice Analysis: Theory Econometrics, and an Application to Automobile Demand. Cambridge: The MIT Press.

Train, K. (2003). Discrete Choice Methods with Simulation. Cambridge University Press.

Villas-Boas, J. and R. Winer. (1999). Endogeneity in Brand Choice Models. Management Science, 45, 1324–1338.

Walker, J.; Ehlers, E.; Banerjee, I. and Dugundji, R. (2011). Correcting for endogeneity in behavioral choice models with social influence variables. Transportation Research, 45A, 362–374.

Williams, H.C.W.L. (1977) On the formation of travel demand models and economic evaluation measures of user benefit. Environment and Planning, 9A, 285-344.

Zou, Z. and Cirillo, C. (2021). Does ridesourcing impact driving decisions: A survey weighted regression analysis. Transportation Research Part A: Policy and Practice, 146, 1-12, https://doi.org/10.1016/j.tra.2021.02.006.