EL USO DE DATOS AGREGADOS E INDIVIDUALES PARA ESTIMAR MODELOS ECONOMETRICOS

por

Sergio L. Gonzalez

Departamento de Ingenieria Civil
Universidad de Puerto Rico

Resumen

Hay un sinnúmero de aplicaciones en el campo del transporte que requieren estimar modelos econométricos. Ejemplo de estos son los modelos de generación de viajes y los de accidentes de tránsito.

Las técnicas existentes para estimar estos modelos incluyen: regresión simple, regresión con restricciones, y el estimador mixto (Theil, 1971). Las dos últimas permiten la incorporación de información adicional sobre un subconjunto de los parámetros a estimarse: determinística en el primer caso, y estocástica en el segundo. Sin embargo, esta información adicional, que puede consistir en elasticidades u otras funciones de los parámetros, tiene que corresponder a la misma unidad de análisis que la muestra usada en la estimación. Por lo tanto, con estas técnicas, no es posible combinar datos agregados con datos desagregados.

Debido a esta limitación de los estimadores existentes, muchas fuentes de datos usualmente disponibles u obtenibles a bajo costo no pueden ser utilizadas en el proceso de estimar modelos econométricos. Por ejemplo, en el caso de modelos de generación de viajes, usualmente tenemos información disponible sobre el número de vehículos que entra y sale del área de estudio (conteos de tráfico). Observe que aunque estos datos proveen información sobre un conjunto de los parámetros a estimarse, no pueden ser utilizados con los estimadores existentes, pues corresponden a un nivel más agregado que la muestra (más específicamente, corresponden a observaciones agregadas de la variable dependiente).

En este trabajo, demostraremos la inhabilidad de los estimadores existentes para incorporar el tipo de datos agregados discutido previamente y desarrollaremos un estimador aplicable a esta situación. Este estimador tiene el potencial de reducir el costo de recopilación de datos, pues, al incorporar los datos agregados, se espera una mayor eficiencia del mismo. El estimador desarrollado en este trabajo es una generalización de los métodos desarrollados por Willumsen (1978), Hendrickson y McNeil (1984), y otros, para estimar matrices de origen-destino al usar conteos de tráfico.

1. Introducción

Un sinnúmero de aplicaciones en el campo del transporte requieren la estimación de modelos econométricos como lo son los modelos de generación de viajes y los de accidentes de tránsito (Fleet, et al, 1968; Morlok, 1978; y Weber, 1971). Estos modelos usualmente se representan mediante la ecuación $y = f(x, \theta)$, en donde y es la variable dependiente del modelo, x el vector

de variables independientes, θ el vector de parámetros a estimarse, y f representa una función.

Basado en la representación y los ejemplos de modelos econométricos presentados anteriormente, se puede observar que estos modelos pueden postularse a diferentes niveles de agregación de la unidad de análisis; definida aquí como la unidad de observación de los variables dependientes e independientes usada en la formulación del modelo. Por ejemplo, en el caso de los modelos de demanda del transporte, las unidades de análisis más aceptadas hoy en día son el individuo y la familia (Domencich y McFadden, 1975 y Ben-Akiva y Lerman, 1985). Estos modelos postulados al nivel micro se denominan modelos desagregados. El término modelos agregados se utiliza al referirse a modelos econométricos cuya unidad de análisis es un grupo de individuos como lo son los residentes de una zona de tráfico o censal.

Indistintamente del nivel de agregación de la unidad de análisis, las técnicas existentes para estimar los parámetros de los mismos (el vector θ solamente se pueden utilizar cuando todas las observaciones en la muestra corresponden al mismo nivel de agregación utilizado en la formulación del modelo. Por ejemplo, para estimar lc parámetros de un modelo desagregado al usar la técnica de regresión lineal simple (Theil, 1971 y Judge et al, 1980), necesitamos una muestra con los valores de las variables dependiente e independientes para cada individuo en la muestra. Denominaremos este tipo de muestra muestra desagregada.

Un grupo de técnicas de estimación conocidas con el nombre de estimadores combinados (Judge et al, 1980), permiten que además de la muestra desagregada, se incorpore en la estimación información adicional en la forma de restricciones en un subconjunto de θ. Esta información adicional, que puede incluir elasticidades conocidas o estimadas, y otras funciones de θ, tiene que corresponder también al nivel desagregado (o en términos más generales, al mismo nivel de agregación de la unidad de análisis). El problema de estimar matrices de origen-destino con datos combinados satisface esta restricción, por lo que, los métodos existentes se pueden utilizar para este problema de estimación (Refiérase a McNeil, 1983 y Hsu, 1985 para más detalles sobre este asunto).

Sin embargo, esta limitación de los estimadores existentes resulta en que, múltiples fuentes de datos comúnmente disponible u obtenibles a bajo costo no pueden utilizarse para el problema más general de estimar modelos econométricos. Por ejemplo, en el caso de modelos de generación de viajes hay usualmente disponibles conteos de tráfico del número de vehículos que entra y sale del área de estudio (Pignataro, 1973). Conteos del número de pasajeros que abordan y se apean en una estación de autobús o metro son recopilados frecuentemente por las agencias que operan estos medios de transporte colectivo (Attanucci et al, 1981).

Preste atención al hecho que tanto los conteos de tráfico como los d transporte colectivo consisten en observaciones de la variable dependient del modelo de generación de viajes para todos los individuos en la población que viajan entre las diferentes zonas de tráfico o estaciones del metro autobús. Por esta razón, denominaremos estas muestras conteos agregados También observe que, aunque los conteos agregados se pueden considerar con restricciones en un subconjunto de θ, éstos no pueden incorporarse en e

proceso de estimación al utilizar los estimadores combinados disponibles, que corresponden a un nivel de agregación diferente al de la muestra desagregada.

El propósito principal de este trabajo es desarrollar estimadores combinados para modelos econométricos que permitan la utilización de una muestra desagregada y conteos agregados en la estimación de los parámetros de interés. La incorporación de los conteos agregados puede aumentar la eficiencia del proceso de estimación y/o predicción; por esta razón, este trabajo ampliará el espacio muestral del problema de muestreo óptimo para la estimación de modelos econométricos. Este trabajo está subdividido en cuatro secciones. En la segunda, repasaremos la literatura de las técnicas de estimación para modelos econométricos y demostraremos que las técnicas disponibles no se pueden utilizar en el problema de estimación de interés. A ésta la seguirá la tercera sección, en la que desarrollaremos varios estimadores consistentes para el problema de interés y discutiremos sus propiedades. Finalmente, en la cuarta sección, presentaremos un breve resumen de este trabajo y discutiremos nuestras principales conclusiones.

2. Repaso de Literatura

En la literatura de econometría se ha discutido extensamente el principal problema de interés de este trabajo: el tópico de estimadores combinados. Esta discusión, sin embargo, ha estado restringida al modelo de regresión lineal y a muestras de un mismo nivel de agregación.

En esta literatura, el desarrollo de estimadores combinados se formula como el problema de incorporar al proceso de estimación información adicional a la muestra. Usualmente, la información adicional se incorpora en la forma de restricciones a los parámetros a estimarse. A continuación discutiremos los dos estimadores más importantes de este tipo presentados en la literatura: regresión lineal con restricciones y el estimador mixto.

2.1. Regresión lineal con restricciones

Los estimadores combinados desarrollados en la literatura son extensiones del modelo de regresión lineal simple. Este modelo se puede representar con la siguiente ecuación:

$$y = X b + \varepsilon \qquad (1)$$

Donde $b_{(Kx1)}$ (léase, el vector b cuyo orden es Kx1) es el vector de parámetros a estimarse; $y_{(Nx1)}$, es el vector de observaciones de la variable dependiente; $X_{(NxK)}$, es la matriz de observaciones de las variables independientes; y $\varepsilon_{(Nx1)}$, el vector de discrepancias o errores. Para más detalles sobre las suposiciones de este modelo refiérase a Theil (1971) y Judge et al (1980).

La extensión más simple del modelo de regresión lineal que puede utilizarse para combinar datos en el proceso de estimación se conoce como

regresión lineal con restricciones (RLCR), (Theil, 1971 y Judge et al, 1980). Este estimador se aplica en la situación cuando $r_{(Mx1)}$, el vector de información adicional a incorporarse en la estimación, se puede representar en la forma de restricciones lineales determinísticas en un subconjunto de b. En este caso, las restricciones se pueden representar con la siguiente ecuación:

$$R\,b = r \qquad (2)$$

En donde $R_{(MxK)}$ es una matriz determinística cuyo elemento R_{mk} es el coeficiente de b_k en la emésima restricción lineal.

Enfatizamos que los estimadores incluídos en la literatura requieren que R sea una matriz determinística, pues como demostraremos más adelante, esta suposición no se satisface en el problema de interés en este trabajo.

2.2. Estimador mixto

El estimador RLCR discutido en la sección anterior puede generalizarse para la situación en donde las restricciones lineales observadas son estocásticas. En esta situación, seguimos el desarrollo de Theil y Goldberger (1961) y representamos las restricciones mediante la siguiente ecuación:

$$r = R\,b + \gamma \qquad (3)$$

En donde $\gamma_{(Mx1)}$ es el vector de discrepancias o errores de r.

Con este modelo, las suposiciones del modelo de regresión lineal simple, si $[\gamma\gamma^T] = \sigma^2 W$ y si ε y γ son independientes, podemos expresar el modelo estadístico de datos combinados mediante la siguiente ecuación:

$$\begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} X \\ R \end{bmatrix} b + \begin{bmatrix} \varepsilon \\ \gamma \end{bmatrix} \qquad (4)$$

El estimador de Aitken de este modelo se denomina estimador mixto y se puede expresar mediante la siguiente ecuación (Theil, 1971):

$$b = x^T x + (R^T W^{-1} R)^{-1} (x^T y + R^T W^{-1} r) \qquad (5)$$

Al igual que en el caso del estimador RLCR, para poder aplicar el estimador mixto, es necesario que la matriz R sea determinística. Esta es una limitación seria de todos los estimadores combinados discutidos en la literatura, pues es debido a ésta que los datos a utilizarse tienen que corresponder a un mismo nivel de agregación. En la próxima sección probaremos este punto en el caso del estimador mixto.

2.3. Limitaciones del estimador mixto

En esta sección probaremos que el estimador mixto no es aplicable a la situación de interés en este trabajo, en donde los datos a combinarse

en la estimación corresponden a diferentes niveles de agregación. Esta demostración se logrará mediante un ejemplo de la aplicación del estimador mixto al problema general de interés en este trabajo. Para la demostración utilizaremos el modelo estadístico presentado a continuación:

$$y_t = X\,b + \varepsilon_t \qquad (6)$$

En donde el subscrito t indica el intérvalo durante el cual las observaciones fueron obtenidas.

Siguimos la notación introducida en la sección anterior, y representamos la información adicional disponible (las observaciones del conteo agregado) por $r_{t'}$; con elementos $r_{mt'}$; m = 1, 2, ..., M . Utilizamos el subscrito t' para indicar que este conteo agregado fue obtenido durante un intérvalo t' sin solape con t (el estimador mixto requiere esta suposición, para más información sobre este punto refiérase a González, 1985).

Para demostrar que el estimador mixto no es aplicable en esta situación, demostraremos, que en este caso, R no es determinística. Para desarrollar una expresión general para R, observamos que la siguiente ecuación relacionaría las observaciones agregadas y desagregadas de la variable dependiente si ambas fueran obtenidas durante el mismo intérvalo t':

$$\begin{aligned} r_{mt'} &= \sum_{n \in C_m} y_{nt'} \\ &= \sum_{n \in C_m} (x_n^T b + \epsilon_{nt'}); \ \forall m \end{aligned} \qquad (7)$$

En donde C_m representa el conjunto de individuos en la población pertenecientes al emésimo grupo y x_n^T es la enésima fila de la matriz X.

Si denominamos R_m como la enésima columna de R, p(x) como la distribución de las variables independientes en la población, y N_m como el número de individuos del emésimo grupo en la población, los elementos de R se pueden expresar de la siguiente forma:

$$R_m = \sum_{n \in C_m} x_n = N_m \int_{C_m} x\, p(x)\, dx \qquad (8)$$

Observe que la expresión en la parte derecha de esta ecuación se obtiene al expresar la sumatoria de todos los individuos en cada grupo de las variables dependientes, como el producto de la media de las variables dependientes en el grupo, y el número de individuos en ese grupo. En términos generales, p(x) no se conoce <u>a priori</u> y esto resulta en una matriz R estocástica. Por lo tanto, cuando r corresponde a observaciones de conteos agregados, no se satisfacen las suposiciones del estimador mixto.

Observamos también que de no haber ningún sesgo en la muestra desagregada, N_m se puede obtener directamente de ésta como la razón del número de observaciones en dicha muestra y la razón de muestreo de cada grupo. En la próxima sección discutiremos las complicaciones adicionales que resultan de una muestra desagregada sesgada y un procedimiento para estimar N_m en esta situación.

2.4. Efecto en la estimación del sesgo de la muestra parcial m

Brög y Meyburg (1980) y Ben-Akiva et al (1983), entre otros, demuestran que es muy poco común no encontrar sesgo en una muestra desagregada. Lo usual en este tipo de muestra es que un número significativo de los individuos en la muestra de diseño no se encuentren en la muestra final, pues puede haber errores de omisión y, en el caso de cuestionarios diseñados para devolverse por correo, muchos individuos no devuelven el cuestionario. Estas situaciones producen un sesgo en la muestra desagregada que denominaremos sesgo de la muestra parcial (Cochran, 1977).

Observe que este sesgo ocasiona que N_m no se pueda obtener directamente de la muestra desagregada. En este caso, seguimos el desarrollo de Ben-Akiva et al (1983) y postulamos el siguiente modelo, basado en la distribución binomial para describir el proceso de la decisión de contestar o no a una entrevista o cuestionario.

$$P[n_m] = \frac{N_m!}{N_m!(N_m-n_m)!} \rho_m^{N_m}(1-\rho_m)^{N_m-n_m} \qquad (9)$$

En donde n_m es el número de individuos del emésimo grupo en la muestra y ρ_m es la probabilidad de que los individuos de este grupo contesten el cuestionario.

Como veremos en el próximo capítulo, este modelo nos permitirá estimar N_m y ρ_m basado en las muestras desagregadas y agregadas y por lo tanto podremos obtener estimadores consistentes para esta situación más general.

3. Desarrollo de Estimadores Combinados para Muestras Desagregadas y Conteos Agregados

En este capítulo desarrollaremos los estimadores combinados que permitan que las diferentes muestras utilizadas correspondan a un nivel de agregación diferente.

En la primera sección, formularemos el problema de estimación de interés e incluiremos: una descripción breve de las muestras supuestas en el desarrollo y una presentación de la notación que utilizaremos en el capítulo. En la segunda sección discutiremos las relaciones que existen entre las distribuciones muestrales de los datos desagregados y las observaciones correspondientes a los conteos agregados. Luego utilizaremos estas relaciones para asegurarnos que las suposiciones de las distribuciones de las diferentes muestras sean consistentes.

En la última sección del capítulo, desarrollaremos la distribución muestral conjunta y la función de máxima verosimilitud para el problema de estimación de interés. Esta función depende de $p(x)$ y N_d, que representan respectivamente la distribución de las variables independientes en las población y el número de individuos en el grupo d. Debido a esto, la función de máxima verosimilitud de nuestro problema es similar a la función no-clásica descrita por Cosslett (1978, 1981a, 1981b) en el contexto de modelos desagregados de elección discreta.

3.1. Caracterización de las muestras y notación

3.1.1 Caracterización de las muestras

La muestra desagregada supuesta en el desarrollo de los estimadores corresponde al mismo nivel de agregación que la unidad de análisis e incluye observaciones de las variables dependientes e independientes del modelo a estimarse.

Observe que en el párrafo anterior hemos descrito con suficiente detalle la muestra desagregada. Sin embargo, existen un sinnúmero de tipos de conteos agregados que están disponibles para el desarrollo de estimadores combinados. Por ejemplo, para la estimación de modelos desagregados de generación de viajes, estos conteos podrían estar disponibles a nivel del par origen-destino. Este tipo de conteo es representativo de aquéllos que se obtienen mediante procesos de agregación basados en grupos mutuamente excluyentes y colectivamente exhaustivos de la población de interés. Utilizaremos el término conteo agregado independiente para referirnos a este tipo de muestra, pues en este caso, las observaciones de los grupos diferentes son probabilísticamente independientes.

Para desarrollar una notación adecuada para este tipo de conteo, vemos que las observaciones de cada grupo de la muestra pueden incluirse en una celda de una tabla de clasificación (Bishop et al, 1975). Por ejemplo, si subdividimos la población de un área urbana en I grupos de ingreso y J grupos de formación educativa, podemos dividir la población en los I x J grupos mutuamente excluyentes y colectivamente exhaustivos definidos por cada combinación específica de ingreso y formación educativa.

En la literatura de análisis de variables discretas (Bishop et al, 1975), en el cual se discuten las tablas de clasificación, se ha desarrollado una notación para representar este tipo de situación. Por esta razón, utilizaremos esa notación ampliamente conocida para representar los observamos de los conteos agregados independientes.

Hay un sinnúmero de conteos agregados en los cuales las observaciones de los diferentes grupos no son probabilísticamente independientes. Por ejemplo, en el caso de los modelos de generación de viajes, los conteos en el cordón y los de transporte colectivo mencionados anteriormente no satisfacen la suposición de independencia probabilítica. Observe, sin embargo, que este tipo de muestra se puede representar mediante los totales de las filas y columnas de una tabla de clasificación, pues están constituidas por todos los viajes que entran (salen) de una estación de conteo, irrespectivamente de cual sea su destino (origen). Observe también que estos datos no son independientes, pues cada par de los totales de filas y columnas tiene en común las observaciones de una celda. Por esta razón, denominamos este tipo de muestra conteo agregado dependiente.

Aunque el tipo de dependencia entre los datos agregados descutida en el párrafo anterior es sólo un caso especial, en este trabajo nos restringiremos al mismo. Hacemos la observación, sin embargo, que las derivaciones presentadas en este capítulo se pueden efectuar también en situaciones más generales de correlación entre los grupos. Esta restricción en las deriva-

ciones se utiliza para simplificar la derivación de los estimadores a prese tarse y no por ser una limitación del método de estimación.

En la próxima sección presentaremos la notación que utilizaremos el resto del capítulo.

3.1.2. Notación

i - subscrito utilizado para identificar la variable de las filas una tabla de clasificación de IxJ. $i \in (1,2,\ldots,I)$.

j - similar a i pero para la variable de las columnas. $j \in (1,2,\ldots,J)$

N_{ij}- número de individuos en la población que pertenecen al grupo identif cado por la fila i y la columna j de la tabla de clasificación.

n - subscrito utilizado para identificar los individuos en la población. $n \in (C_{11}U..UC_{ij}U...UC_{IJ})$; $C_{ij} \in (1,2,\ldots,N_{ij})$, $\forall$ i,j.

Y_{ijn}- variable aleatoria que representa la variable independiente d enésimo individuo. Los subscritos i,j nos indican el grupo al cu este individuo pertenece.

x_{ijn}- vector de atributos o variables independientes para el enésimo ind viduo.

F_{ijn}- inverso de la proporción con la cual se muestreó al grupo del enési individuo. En el caso de muestra aleatoria es constante e igual a F.

$P[Y_{ijn}|x_{ijn}, \theta]$ - probabilidad condicionada de Y_{ijn}. Representa el mode probabilítico supuesto en el desarrollo del estimador.

r_{ij+}- variable aleatoria que se utiliza para representar los conteos agregad independientes. El tercer subscrito (+), indica que este conteo basa en los totales sobre todos los individuos en la población.

r_{i++} y r_{+j+}- variables aleatoreas utilizados para representar los conte agregados dependientes

En la próxima sección discutiremos las relaciones que existen ent las distribuciones de la muestra desagregada y las de los conteos agregados.

3.2. Distribuciones derivadas de los conteos agregados

Un aspecto esencial en el desarrollo de estimadores combinados es los parámetros y las distribuciones de las diferentes muestras están relac nados. Esto implica que los parámetros de la población y las distribucio supuestas para las diferentes muestras deben basarse en las relaciones exist

tes entre las mismas. En esta sección desarrollaremos los parámetros y distribuciones de las muestras agregadas a partir de las suposiciones utilizadas para la muestra desagregada.

Para simplificar estas derivaciones supondremos que tanto las observaciones agregadas como los desagregadas corresponden a un mismo intérvalo t. Sin embargo, nuestro desarrollo puede extenderse fácilmente al caso de observaciones hechas durante intérvalos diferentes.

Observe que las siguientes relaciones existen entre las variables aleatorias de interés en este trabajo.

$$r_{i++} = \sum_{Vj} r_{ij+}; \forall i \qquad (10)$$

$$r_{+j+} = \sum_{vi} r_{ij+}; \forall j \qquad (11)$$

$$r_{ij+} = \sum_{n \in C_{ij}} Y_{ijn}; \forall i,j \qquad (12)$$

Estas relaciones implican que una vez hayamos supuesto la distribución probabilística de la variable Y_{ijn}, las distribuciones de r_{ij+}, r_{i++}, y r_{+j+} se pueden obtener mediante la convolución de las distribuciones desagregadas. Al hacer esto obtenemos la siguiente distribución derivada de r_{ij+} para toda i,j:

$$Q\,[r_{ij+} | p(x_{ij}), \theta, N_{ij}] = \sum_{n=2}^{N_{ij}} \sum_{V Y_{ijn}} \left(P\left[Y_{ij1} = r_{ij+} - \sum_{Vn'=1} Y_{ijn}\right| x_{ij1}, \theta\right] \prod_{n=2}^{N_{ij}} P[Y_{ijn} | x_{ijn}, \theta]\right) \qquad (13)$$

En donde hemos supuesto que la distribución de cada inviduo es la misma y está dada por P y que las distribuciones de los diferentes individuos son independientes.

En esta ecuación utilizamos $p(x_{ij})$ para representar la distribución de x_{ij} en la población. Observe que la función de probabilidad de r_{ij+}, representada por Q, depende de x_{ijn} para todos los individuos en la población. Por esta razón es que Q depende de la distribución de estas variables. Observe también que al sumar sobre Y_{ijn} suponemos que esta variable es discreta. En el caso de que Y_{ijn} sea continua, la sumatoria debe reemplazarse por el integral sobre la región en donde esta variable esté definida.

Observamos que en el caso general de cualquier distribución P no es posible obtener una expresión analítica de la distribución Q. Sin embargo, cuando la distribución P es regenerativa (Benjamín y Cornell, 1970), dicha forma si se puede obtener, pues Q y P corresponden a la misma distribución. En lo subsiguiente, nos limitaremos a dos casos especiales de la distribución

P: la Poisson y la Normal. Estos casos son de particular relevancia para nuestro estudio, pues son los utilizados más frecuentemente en las aplicaciones de Transporte. Representamos estas dos distribuciones mediante las siguientes ecuaciones respectivamente:

$$P[Y_{ijn}|x_{ijn}, \theta] = \text{Poisson}\,[f(x_{ijn}, \theta)], \forall i,j,n \qquad (14)$$

$$P[Y_{ijn}|x_{ijn}, \theta] = N[f(x_{ijn}, \theta), \sigma^2], \forall i,j,n \qquad (15)$$

En donde $f(x_{ijn}, \theta)$, una función, representa la media de las distribuciones y σ^2 la varianza de la distribución normal.

Cuando suponemos que la distribución P es Poisson, r_{ij+} también será Poisson con parámetro dado por la siguiente ecuación para toda i, j.

$$E[r_{ij+}|p(x_{ij}), \theta, N_{ij}] = \sum_{n \in C_{ij}} f(x_{ijn}, \theta) \qquad (16)$$

$$= N_{ij} \int_{X_{ij}} f(x_{ij}, \theta) p(x_{ij}) dx_{ij}$$

En donde X_{ij} es la región en la cual X_{ij} está definida.

Este resultado indica que la media de r_{ij+} se puede representar como la suma de las medias de todos los individuos en la población pertenecientes al grupo (i,j). Fíjese que la representación alterna de esta media se obtiene como el producto de la media de f y el número de individuos en el grupo. Esta segunda representación será la que utilizaremos en lo restante de este trabajo, pues será mucho más conveniente en nuestro desarrollo.

Bajo la suposición Normal, la media de r_{ij+} también se puede representar mediante la ecuación (15). En este caso, la varianza de r_{ij+} se obtiene mediante la siguiente ecuación:

$$\text{Var}\,[r_{ij+}\ \sigma, \theta, N_{ij}] = \sigma^2 N_{ij} \qquad (17)$$

Las ecuaciones (10), (11), y (12) se pueden utilizar para desarrollar la distribución derivada conjunta de los conteos agregados dependientes. En este caso, bajo la suposición Normal obtenemos la distribución Normal multivariable. La distribución resultante bajo la suposición Poisson es la Poisson multivariable (Krishnamoorthy, 1951; Cambell, 1934; y Teicher 1954) para la cual no existe una expresión analítica. Sin embargo, esta distribución puede aproximarse mediante la Normal multivariable, por esta razón, utilizaremos esta última distribución para los dos casos de interés. Esta suposición la representaremos mediante la siguiente ecuación.

$$P[r|p(x), \theta, N] = MVN(f(p(x), \theta, N)); \Sigma] \qquad (18)$$

En donde:

N es la matriz de N_{ij} $\forall i \leq I$, $\forall j \leq J-1$.

r es el vector de observaciones del conteo agregado dependiente de orden (I+J-1)

$f(p(x), \theta, N)$ es el vector de medias de orden $(I+J-1)$

Σ es la matriz de varianza covarianza de orden $(I+J-1)x(I+J-1)$.

Observe que hemos excluído el último elemento de todas estas matrices y vectores por ser linealmente dependiente de los demás elementos.

Encontramos conveniente para nuestra presentación representar estas matrices mediante el uso de submatrices. Al usar esta nueva notación representaremos el vector $r^T = [r_1^T, r_2^T]$ en donde los subvectores se representan como $r_1^T = [r_{1++}, \ldots, r_{I++}]$ y $r_2^T = [r_{+1+}, \ldots, r_{+(J-1)+}]$.

De manera análoga, representaremos el vector de las medias como $f^T(p(x), \theta, N) = [f_1^T(p(x), \theta, N)\ f_2^T(p(x), \theta, N)]$. En cuyo caso representaremos los elementos de cada subvector con la omisión de los argumentos de la función vectorial f como:

$$f^T_1 = \sum_{V_j} N_{ij} \int_{X_{1j}} f_{1j} p(x_{1j}) dx_{1j}, \ldots, \sum_{V_j} N_{IJ} \int_{X_{Ij}} f_{IJ} p(x_{IJ}) dx_{IJ}] \qquad (19)$$

$$f^T_2 = [\sum_{V_i} N_{i1} \int_{X_{i1}} f_{i1} p(x_{i1}) dx_{i1}, \ldots, \sum_{V_i} N_{iJ-1} \int_{X_{iJ-1}} f_{iJ-1} p(x_{iJ-1}) dx_{iJ-1}] \qquad (20)$$

Finalmente, representaremos la matriz de varianza covarianza como

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \qquad (21)$$

Con elementos de las submatrices $\Sigma_{11(IxI)}$, $\Sigma_{12(IxJ-1)}$, $\Sigma_{21(J-1xI)}$, y $\Sigma_{22(J-1xJ-1)}$ expresados por:

$$\Sigma_{11ij} = \begin{cases} 0, \ \forall\ i = j \leq I \\ \\ Var[r_{i++}] = \sum_{k \leq J} \sum_{Vn} \sigma^2_{ikn}, \ \forall\ i = j \leq I \end{cases} \qquad (22)$$

$$\Sigma_{22ij} = \begin{cases} 0, \ \forall\ i = j \leq J-1 \\ \\ Var[r_{+j+}] = \sum_{k \leq I} \sum_{Vn} \sigma^2_{ijn}, \ \forall\ i = j \leq J-1 \end{cases} \qquad (23)$$

$$\Sigma_{12ij} = \Sigma_{21ji} = Var[r_{ij+}] = \sum_{Vn} \sigma^2_{ijn}, \ i \leq I, \ j \leq J-1 \qquad (24)$$

En estas ecuaciones hemos expresado las varianzas individuales como σ^2_{ijn} para así poder permitir tanto la suposición Poisson como la Normal. Para el segundo caso, $\sigma^2_{ijn} = \sigma^2$ V i, j, n; por lo tanto, la sumatoria sobre todo n resulta en $N_{ij}\sigma^2$. En el primer caso, las varianzas individuales son igual a $f(x_{ijn},\theta)$ y la sumatoria sobre n puede expresarse como la parte derecha de la ecuación (16).

En esta sección hemos desarrollado las distribuciones derivadas de los diferentes tipos de muestras que utilizará nuestro estimador. En la próxima sección utilizaremos esta información para derivar las funciones de máxima verosimildad de estas muestras.

3.3. Desarrollo de los estimadores para muestras aleatorias

En la sección anterior observamos que las distribuciones de las muestras de interés en este trabajo dependen de p(x) y N. Manski y McFadden (1981), Cosslett (1978, 1981a, 1981b), y Hsieh et al (1983), en el contexto de estimadores de modelos de elección discreta basados en muestras no aleatorias, encuentran que sus distribuciones muestrales también dependen de p(x). Estos autores indican que esta dependencia en p(x) ocasiona complicaciones significativas en el desarrollo de los estimadores. Por esta razón, en nuestro desarrollo del estimador de interés, seguiremos la estructura utilizada por Manski y McFadden (1981), y presentaremos estimadores para las siguientes situaciones definidas por el conocimiento o desconocimientos de p(x) y N.

1. p(x) conocida a priori

   a- N conocida a priori

   b- N desconocida a priori

2. p(x) desconocida a priori

   a- N desconocida a priori

   b- N desconocida a priori

Observamos que las situaciones en donde p(x) y/o N se conocen a priori puden ocurrir en situaciones prácticas cuando tenemos disponible información sobre las variables independientes de los grupos de interés provenientes de un censo o una muestra grande.

De conocer p(x), nuestro desarrollo del estimador se basará en el método de máxima verosimilitud, pues, en este caso, al igual que los estimadores discutidos por Manski y McFadden (1981), Cosslett (1978, 1981a, 1981b) Hsieh et al (1983), este método mantiene todas sus propiedades asintóticas.

En el caso que p(x) se desconozca, nuestra función de máxima verosimilitud es del tipo no-clásico discutido por Cosslett (1981a). Por está razón, el estimador de máxima velosimilitud no poseerá sus propiedades asintóticas

En este caso, sin embargo, utilizaremos el procedimiento presentados por Cosslett (1981a) para desarrollar un estimador consistente.

Nuestro desarrollo en este trabajo se limitará a la situación de muestras aleatoreas. Sin embargo, en otro trabajo más general (González, 1985) hemos desarrollado estimadores para muestras no aleatorias.

Hemos estructurado muestra discusión en esta sección como sigue. Primero discutiremos la notación que utilizaremos al representar las diferentes muestras. Luego desarrollaremos los estimadores clásicos y finalmente los no-clásicos.

3.3.1. Notación para representar las muestras

Para el desarrollo de los estimadores combinados en este trabajo supondremos tres tipos de muestras diferentes. La primera muestra será la muestra desagregada, la cual supondremos que incluirá sesgos de muestra parcial. Por esta razón, denominaremos $\rho_{ij} = a_i b_j$; $\forall i,j$, como el modelo multiplicativo que representa la probabilidad de que un individuo del grupo i,j no conteste el cuestionario. Aunque en este trabajo nos restringiremos a esta parametrización de la probabilidad, nuestra técnica también aplicará a modelos más generales.

Utilizaremos el sobrescrito (1) para denominar las observaciones provenientes de esta muestra que incluirán:

$N_{ij}^{(1)}$ - el número de individuos del grupo i,j en la muestra

$y_{ijn}^{(1)}$- variable dependiente del enésimo individuo en la muestra

$x_{ij}^{(1)}$ - vector de variables independientes del enésimo individuo.

La segunda muestra consistirá del conteo agregado independiente. Denominaremos las observaciones de este conteo con el sobrescrito (2). La tercer muestra, consistirá del conteo agregado dependiente y sus observaciones se indicarán con el sobrescrito (3).

Utilizaremos la siguiente notación para representar las observaciones de estos conteos:

l - subscrito que distingue los diferentes conteos agregados independientes disponibles, $l \in 1,2,\ldots,L$.

m - similar a l, pero los conteos agregados dependientes $m \in 1,2,\ldots,M$

$r_{ij+l}^{(2)}$- observación del elésimo conteo agregado dependiente correspondiente al grupo i,j.

$r_{i+\cdot+m}^{(3)}$-observación de los totales de la fila i; $\forall i$, del emésimo conteo agregado dependiente.

$r_{ij+m}^{(3)}$-similar al anterior para los totales de la columna j.; $j \leq J-1$.

Para representar las observaciones de todas las filas y columnas en forma más compacta, extenderemos la notación vectorial introducida anteriormente, por lo que incluiremos todas las observaciones del emésimo conteo en el vector $r_m^{(3)}$.

Hacemos la observación que hemos excluído las observaciones r_{+J+m} en nuestra notación. Esto es necesario pues, como la suma de las filas y las columnas son iguales, una de las observaciones de estos conteos depende linealmente de las otras y por lo tanto no provee información adicional.

En las próximas secciones utilizaremos esta notación para desarrollar los estimadores de interés de este trabajo.

3.3.2. Estimadores para muestras aleatorias cuando p(x) se conoce

Como indicamos en la introducción a este capítulo, cuando p(x) se conoce a priori, el estimador de máxima verosimilitud de nuestro problema tiene todas las propiedades clásicas. Por lo tanto, una vez obtengamos las distribuciones muestrales, el desarrollo de este estimador es trivial.

Cuando N es conocida a priori, el logaritmo natural de la función de verosimilitud se puede expresar mediante la siguiente ecuación general:

$$\begin{aligned} \ln L(\theta) = & \sum_{V_{ijn}} \ln P[Y_{ijn} \mid x_{ijn}, \theta] \\ & + \sum_{V_{ijl}} \ln P[r_{ij+1} \mid p(x_{ij}), \theta, N_{ij}] \\ & + \sum_{V_m} \ln P[r_m \mid p(x), \theta, N] \end{aligned} \qquad (25)$$

Observamos que esta función consiste de tres términos principales: el primero es el logaritmo de la función de verosimilitud de la muestra desagregada y el segundo y tercero las de los conteos agregados independientes y dependientes respectivamente. Podemos obtener la función combinada mediante la suma de estos tres términos porque las muestras son independientes.

Observamos además, que en las funciones de verosimilitud, el índice n indica una sumatoria sobre observaciones en la muestra, y no las de la población como lo era en las ecuaciones anteriores.

En el caso en que N es desconocida, podemos obtener también un estimador de máxima verosimilitud clásico. En este caso, para poder estimar N, el logaritmo de la función de verosimilitud incluye la versomilitud de $N_{ij}^{(1)}$ en la muestra desagregada (refiérase a Hsieh et al, 1983 para más detalles). Esta función se expresa mediante la siguiente ecuación:

$$\ln L(\theta, N, a, b) = \sum_{V_{ijn}} \ln P[Y_{ijn} \mid x_{ijn}, \theta]$$

$$+ \sum_{V_{ijn}} \ln P[r_{ij+1} \quad p(x_{ij}), \quad , N_{ij}]$$

$$+ \sum_{V_m} \ln P[r_m \quad p(x), \theta, N] \qquad (26)$$

$$+ \sum_{V_{ij}} \ln P[N_{ij}^{(1)} | N_{ij}, F, a_i, b_j]$$

En donde **a** es el vector con elementos $a_i, \forall i \leq I$; **b** aquel con elementos b_j, $j \leq J$, y F el inverso de la razón de muestreo.

En la próxima sección desarrollaremos estimadores para el caso más general en donde p(x) no se conoce a priori.

3.3.3. Estimador con p(x) desconocida

Cuando p(x) es desconocida, la función de verosimilitud es del tipo no-clásico discutido en Cosslett (1981a). Por esta razón, para el desarrollo de estos estimadores, seguimos el procedimiento desarrollado por Cosslett (1978, 1981a, 1981b) para derivar estimadores no-clásicos. Para hacer esto, comenzamos con la función de versomilitud de las muestras combinadas presentada a continuación. (Observe que desarrollamos en función para **N** desconocida, el caso de **N** conocida se puede obtener excluyendo la verosimilitud de $N_{ij}^{(1)}$).

$$\ln L(\theta, \mathbf{N}, \mathbf{a}, \mathbf{b}, \mathbf{w}) = \sum_{V_{ijn}} \ln P[Y_{ijn} | X_{ijn}, \theta]$$

$$+ \sum_{V_{ijn}} \ln w_{ijn}$$

$$+ \sum_{V_{ijl}} \ln P[r_{ij+1} | \mathbf{w}, \theta, N_{ij}] \qquad (27)$$

$$+ \sum_{V_m} \ln P[r_m | \mathbf{w}, \theta, N]$$

$$+ \sum_{V_{ij}} \ln P[N_{ij}^{(1)} | N_{ij}, F, a_i, b_j]$$

en donde:

$$\sum_{n=1}^{N_{ij}^{(1)}} w_{ijn} = 1 \qquad \forall i,j \qquad (28)$$

y

$$W_{ijn} \geq 0, \ \forall i,j,n \qquad (29)$$

En donde hemos representado las funciones de densidad desconocidas, $p(x_{ij})$, mediante un peso W_{ijn} por cada individuo en la muestra. El vector **w** incluirá todos estos pesos. También observe que incluimos las restricciones (28) y (29) para que estos pesos representen una función de densidad adecuada.

Observamos que esta función incluye los parámetros **w**, cuyo número aumenta con el número de observaciones. Esta es la razón principal por la cual las propiedades clásicas de los estimadores de máxima verosimilitud no aplican a este caso. Para resolver esta situación, que obviamente no puede resultar en estimadores consistentes, Cosslett (1981a) desarrolla la función de verosimilitud concentrada. Esta función se obtiene mediante la solución del problema de optimización representado por (27), (28) y (29) suponiendo **N**, **a**, **b**, θ constantes.

Hasta este momento no hemos podido resolver este problema de optimización para el caso general, en el cual incluimos como muestras la desagregada y ambos tipos de conteo agregado. Sin embargo, en González (1985), demostramos que para los casos en que utilizamos la muestra desagregada con cualquiera de los conteos agregados, la solución a este problema se puede expresar como:

$$W_{ijn} = 1/N_{ij}^{(1)}; \; V_{i,j,n} \qquad (30)$$

El logaritmo de la función de verosimilitud concentrada se obtiene sustituyendo la ecuación (30) en la (27). Observe que luego de esto, el número de parámetros a estimarse no aumenta con el tamaño de la muestra.

Con esta expresión general de la función concentrada de verosimilitud podemos obtener un estimador consistente; el cual denominamos estimador combinado con modelo de sesgo (ECMS). Para lograr esto, primero sustituimos en la ecuación (29) las distribuciones muestrales discutidas en la sección 3.1 y 3.2, y luego maximizamos la función en el espacio de los parámetros de interés.

Los detalles técnicos de estas derivaciones para los modelos Normal y Poisson y una discusión y demostración de las propiedades de este estimador se presentan en González (1985).

Hasta este momento, hemos presentado un desarrollo teórico que nos permite incorporar conteos agregados en la estimación de modelos econométricos. La pregunta relevante es ahora: ¿Qué ganamos al incorporar los conteos agregados en la estimación? En la próxima sección contestaremos esta pregunta mediante la discusión de algunas propiedades del estimador ECMS y del estimador desagregado basado en los resultados de un estudio de simulación.

3.4. Propiedades de los estimadores

Al implementar la simulación, nos percatamos que la eficiencia de los estimadores de θ de la muestra desagregada y el ECMS para muestras aleatorias eran iguales numericamente. En González (1985) demostramos esta propiedad analíticamente.

Esta propiedad implica que, en este caso, al incorporar los conteos agregados en el proceso de estimar modelos econométricos, no aumentamos la eficiencia del estimador de θ . Observe que esto resulta, pues al incorporar el conteo agregado al proceso de estimación, también aumentamos el número de parámetros a estimar (**N, a, b**) y nuestro estimador utiliza toda la información adicional al estimar estos parámetros. En el caso de estimadores para muestras no aleatorias, existe el potencial de aumentar la eficiencia de con el estimador combinado (refiérase a González, 1985).

El resultado presentado en el párrafo anterior no implica que el estimador desarrollado en este trabajo no es útil. Observamos que nuestro estimador nos permite estimar los parámetros **a, b,** y **N** bajo la situación en donde haya sesgo de muestra parcial. El estimador desagregado, sin embargo, solamente nos permite estimar **N** dado **a** y **b**. El estimar **a, b,** y **N** eficientemente es de suma importancia en el caso de modelos desagregados pues estos parámetros son necesarios para expandir o agregar la muestra a la población (refiérase a Koppelman, 1976). Además, en el caso de predicción incremental ("pivot--point") (refiérase a Manheim, 1979 y Ben-Akiva y Lerman, 1985), estos parámetros entran directamente en la función de predicción.

En González (1985) presentamos los resultados de un estudio de simulación diseñado para comparar la eficiencia relativa del estimador desagregado y el ECMS al estimar **N.** (Observe que no podemos incluir **a, b** en nuestra comparación pues el estimador desagregado no puede estimar estos parámetros simultaneamente con **N**). Los resultados de este estudio de simulación indican que la eficiencia del ECMS relativa al estimador desagregado varía entre 0.86 y 3.13. El resultado positivo del 3.13 ocurre cuando la $\rho = 0.50$ y para muestras de tamaño mediano. Este resultado implica que la incorporación de los conteos agregados puede tener resultados sumamente positivos en la estimación de modelos econométricos.

En el próximo capítulo presentaremos un breve resumen de este trabajo.

4. Resumen

En este trabajo hemos demostrado que los estimadores combinados disponibles en la literatura no pueden utilizarse con muestras de diferentes niveles de agregación. Como resultado de ésto, un sinnúmero de conteos agregados usualmente disponibles u obtenibles a bajo costo no pueden utilizarse al estimar modelos econométricos. Esta situación motivó la investigación presentada en este trabajo, en la cual desarrollaremos estimadores combinados on modelo de sesgo (ECMS) que pueden utilizarse con muestras desagregadas y conteos agregados.

La dificultad principal que encontramos en el desarrollo de estos estimadores es la presencia de $p(x)$, la función de densidad de las variables independientes, en la función de estimación. Debido a esto, seguiremos el desarrollo de Manski y McFadden (1981), y desarrollamos estimadores de máxima verosimilitud de muestras aleatorias para la situación en dónde $p(x)$ se conoce a priori; usualmente basada en un censo de la porblación. En el caso en que $p(x)$ se desconoce, seguimos el desarrollo de Cosslett (1981a) y derivamos un estimador consistente para muestras aleatorias.

El estimador no-clásico presentado en este trabajo no resulta en una mayor eficiencia en la estimación del vector ; sin embargo, este estimador nos permite estimar a, b, y N; los parámetros del modelo de sesgo de muestra parcial, más eficientemente. Esto último tiene el potencial de mejorar las predicciones de los modelos desagregados, pues estos parámetros están incluidos en la función de predicción.

Referencias

ATTANUCCI, J. P., BURNS, I. y WILSON, N. (1981) Bus Transit Monitoring Manual: Vol. 1: Data Collection Program Design. NTIS Report PB-82-122227, EE.UU.

BENJAMIN, J. R. y CORNELL, C.A. (1970) Probability Statistics, and Decision for Civil Engineers. McGraw-Hill, Nueva York.

BEN-AKIVA, M.E. y LERMAN, S.R. (1985) Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, Cambridge (en imprenta).

BEN-AKIVA, M.E., GUNN, H. y POL, H. (1983) Expansion of data from mixed random and choice-based survey designs. International Conference on New Survey Methods in Transport, Sidney, 12-16 Septiembre 1983, Australia.

BISHOP, Y., FIENBERG, S., y HOLLAND, P. (1975) Discrete Multivariate Analysis. MIT Press, Cambridge.

BROG, W. y MEYBURG, A.H. (1980) The non-responde problem in travel surveys -an empirical investigation. 59th Annual Meeting of the Transportation Research Board. Washington, D.C., 11-14 Enero 1980, EE.UU.

CAMPELL, J. T. (1934) The Poisson correlation function. Proceedings of the Edinburg Mathematical Society, Vol. 4, 18-26.

COCHRAN, W.G. (1977) Sampling Techniques. John Wiley & Sons, Nueva York.

COSSLETT, S. (1978) Efficient Estimation of Discrete Choice Models from Choice-Based Samples. Ph. D. Dissertation, Department of Economics, University of California at Berkeley, EE.UU.

COSSLETT, S. (1981a) Efficient estimation of discrete choice models. En C. Manski y D. McFadden (eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge.

COSSLETT, S. (1981b) Maximum likelihood estimator for choice based samples. Econometrica, Vol. 49, 1289-1316.

DOMENCICH, T.A. y McFADDEN, D. (1975) Urban Travel Demand: A Behavioural Analysis . North-Holland/Elsevier, Amsterdam.

FLEET, C.R. y ROBERTSON, S.R. (1968) Trip generation in the transportation planning process. Highway Research Record 240, 11-31

GONZALEZ, S. (1985) Combining Survey and Aggregate Data for Model Estimation. Ph. D. Dissertation, Department of Civil Engineering, MIT, EE.UU.

HENDRICKSON, C., y McNEIL, S. (1984) Matrix entry estimation errors Ninth International Symposium or Transportation and Traffic Theory, Delft, 11-13 Julio 1984, Holanda.

HESIEH, D., MANSKI, C. y McFADDEN, D. (1983) Estimation of response probabilities from augmented restrospective observation. Department of Economics, MIT, EE.UU.

HSU, P. (185) Estimation of Parameters for Multiple and Temporally Distributed Populations. Ph. D. Dissertation, Department of Civil Engineering, MIT, EE.UU.

JUDGE, G. G., GRIFFITHS, W.E., CARTER HILL, R. y LEE, T. (1980) The Theory and Practice of Econometrics. John Wiley & Sons, Nueva York.

KOPPELMAN, F. (1976) Guidelines for aggregate travel predictions using disaggrega choice models. Transportation Research Record 610, 19-24

KRISHNAMOORTHY, A. S. (1951) Multivariate binomial and Poisson distributions. The Indian Journal of Statistics, Vol. 11, N°2, 117-124.

MANHEIM, M.L. (1979) Fundamentals of Transportation Systems Analysis-Volume I: Basic Concepts. MIT Press, Cambridge.

MANSKI, C.F. y McFADDEN, D.(1981) Alternative estimators and sample designs for discrete choice analysis. En C.Manski y D. McFadden (eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge

McCARTHY, G.M. (1969) Multiple regression analysis of household trip generation: a critique. Transportation Research Record 297, 31-43.

McNEIL, S. (1983). Quadratic Matrix Entry Estimation Methods. Ph.D. Dissertatio Department of Civil Engineering, Carnegie-Mellon University, EE.UU.

MORLOK, E. K. (1978) Introduction to Transportation Engineering and Planning. McGraw Hill, Nueva York.

NUMERICAL ALGORITHMS GROUP (1984) NAG Fortran Mini Manual-Mark 11: Introductory Guide to the NAG Fortran Manual. Numerical Algorithms Group, Oxford.

PIGNATARO, L. J. (1973) Traffic Engineering: Theory and Practice. Prentice Hall Nueva Jersey.

THEIL, H. (1971) Principles of Econometrics. John Wiley & Sons, Santa Bárbara.

THEIL, H. y GOLDBERGER, A.S.(1961) Pure and mixed statistical estimation in econometrics. International Economic Review, Vol. 2, 65-78.

WEBER, D.C. (1971) Accident rate potential: an application of multiple regression analysis of a Poisson process. Journal of the American Statistical Association, Vol. 66, 285-288.

WILLUMSEN, L. G. (1978) O-D matrices from network data: a comparison of alternative methods for their estimation. Proceedings PTRC Summer Annual Meeting, PTRC Education and Research Services Limited, Londres.