Centro de Investigación en Matemáticas, A.C.

Ordenamiento de variables auxiliares en muestreo balanceado

T E S I S

Que para obtener el grado de

Maestro en Ciencias en Estadística Oficial

P r e s e n t a

José de Jesús Suárez Hernández

Director de Tesis:

Guanajuato, Gto. Diciembre de 2008

CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS, A.C.

Ordenamiento de variables auxiliares en muestreo balanceado

TESIS

Que para obtener el grado de

Maestro en Ciencias en Estadística Oficial

PRESENTA:

José de Jesús Suárez Hernández

Comité de Evaluación:

_____
Dr. Rogelio Ramos Quiroga
(Presidente)

_____
Dr. Abdón Sánchez Arroyo
(Secretario)

_____
Dra. José Elías Rodríguez Muñoz
(Vocal y Director de Tesis)

Guanajuato, Gto. Diciembre de 2008

DEDICATORIA

Dedico esta obra con amor y agradecimiento a mis padres José de Jesús Suárez Marcial† y Ma. Martha Hernández de Luna† quienes, cuando juntos decidieron tomar el riesgo de darme la vida, me dieron la oportunidad de vivir este sueño, de saborear este instante. Dedicarles esta obra es una pequeña muestra de que estaré eternamente agradecido por su apoyo y estímulo incondicionales.

AGRADECIMIENTOS

He recibido ayuda y palabras de aliento de muchas personas durante la realización de este trabajo. En particular expreso mi más sincero agradecimiento a mis hermanos Luis, Javier, Jorge, Ma. Concepción, María Angélica, Ma. Elena, Eberto, Francisco y María por su compañía, paciencia y comprensión.

A mi amigo y compañero de la maestría M. en C. Abel Alejandro Coronado Iruegas, primero en tomar este "vuelo" maravilloso, por aceptar mi colaboración en varias investigaciones previas y en un ejercicio con información de la Encuesta Nacional Ejidal de Mercados de Tierra 2005 a cargo de la Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación (SAGARPA).

A mi amigo, compañero de la maestría y compañero de área de trabajo, M. en C. Juan Martínez Rodríguez por explicarme pacientemente algunos diseños y algoritmos de muestreo, enfocándose en el cálculo de la varianza, coeficiente de variación y efecto de diseño.

A mi profesor de muestreo Dr. Ignacio Méndez Ramírez por compartir sus valiosas experiencias adquiridas en proyectos y estudios realizados a nivel nacional en México.

Un especial agradecimiento a mi asesor y profesor de muestreo Dr. José Elías Rodríguez Muñoz por su paciencia, enseñanza, asistencia y sus sabios comentarios.

Al Dr. Gilberto Calvillo Vives y a la Dra. Graciela González Farías porque con la formalización de sus acuerdos nos brindaron a mis compañeros y a mí, empleados del Instituto Nacional de Estadística y Geografía (INEGI), la oportunidad de adquirir herramientas científicas y tecnológicas para mejorar la calidad de la información que el instituto entrega a los distintos sectores de la sociedad.

Agradezco al Dr. Yves Tillé sus invaluables consejos, uno vía telefónica y los demás por correo electrónico, los cuales fueron relevantes en la consecución de una mayor calidad en el proyecto.

RESUMEN

El Método del Cubo es un método general que selecciona muestras aproximadamente balanceadas para cualquier número de variables auxiliares. Si el número es demasiado grande, más de diez según los autores, para que el problema de programación lineal, utilizado en el método, sea resuelto por un algoritmo simplex, entonces al final de la fase de vuelo debe eliminarse una variable, relajando así una restricción y permitiendo regresar a la fase de vuelo hasta que no sea posible moverse más dentro del hiperplano de restricciones. Las restricciones son así relajadas sucesivamente. Además, la reducción en la varianza del estimador Horvitz-Thompson del total de la variable de interés, en un diseño de muestreo balanceado, depende de su correlación con las variables auxiliares. Por esta razón, es necesario ordenar las variables controladas considerando como criterio de ordenamiento la importancia de sus correlaciones con la variable de interés, de tal forma que las restricciones menos importantes sean relajadas primero.

En este trabajo, la técnica Componentes Principales Supervisadas es aplicada tanto a 9 variables de los datos MU284 como a 104 variables de los datos ENIGH2002, en el ámbito de ajuste por regresión gaussiana, para calcular las correlaciones entre la variable de interés y las variables auxiliares. Luego con estos valores, llamados scores, se asigna un orden estadístico de correlación a cada variable. Después por medio de validación cruzada del cociente de verosimilitudes se determina el número óptimo de variables controladas. Finalmente se aplica el Método del Cubo, a este conjunto de información auxiliar reducido, para obtener muestras aproximadamente balanceadas del estimador Horvitz-Thompson, usando el método de eliminación de variables en su fase de aterrizaje.

CONTENIDO

LISTA DE TABLAS

LISTA DE ILUSTRACIONES

LISTA DE ALGORITMOS

C a p í t u l o I

INTRODUCCIÓN

Quienes nunca olvidan que no saben nada y están dispuestos a aprender todo, lo aprenderán.

Helen Schucman y William Thetford
Meditaciones de un curso de milagros

En este capítulo introductorio se aborda el origen de la motivación del trabajo, los antecedentes, luego se hace una descripción previa del problema del ordenamiento de las variables auxiliares, después se mencionan la solución propuesta, los objetivos de la investigación, su delimitación y finalmente se hace una descripción de la estructura del documento.

1.1 Origen de la motivación

El origen de la motivación de este trabajo es el segundo Seminario sobre Métodos Alternativos para Censos Demográficos, efectuado en oficinas centrales del INEGI del 4 al 6 de julio del 2005, con la participación de 20 países, entre ellos Francia. Los representantes de este país hablaron sobre la metodología utilizada en el censo rodante francés, puesto en marcha a partir del 2004. En la explicación de la metodología utilizada se comentó que uno de los orígenes del censo rodante francés fue el reporte técnico de Jean-Claude Deville y Michael Jacod presentado al INSEE en 1995. Este reporte hace referencia al muestreo balanceado que a su vez llevó al desarrollo y publicación del Método del Cubo por Jean-Claude Deville e Yves Tillé en el 2004.

Los trabajos presentados en este seminario promueven el cambio de las metodologías tradicionales, estáticas y onerosas, por métodos alternativos actuales, dinámicos y económicos.

Una descripción de la nueva situación de la "sociedad de la información" y sus principales demandas se presenta en la siguiente sección, la cual aborda al concepto de "información auxiliar", un término que es producto de las tendencias actuales de los investigadores en estadística.

1.2 Información auxiliar

El dinamismo actual de la generación y uso de información estadística ha llegado a modificar las formas tradicionales de su explotación. Ha pasado de una generación de información estadística burda, pesada y esporádica, a una generación especializada, ligera y frecuente. Las sociedades actuales demandan información estadística constantemente, por lo que grupos de investigadores recolectan datos con una mayor regularidad. Las encuestas por muestreo son usadas cuando no se pueden medir todas las unidades para satisfacer dicha demanda.

Tal demanda constante de información ha creado en los investigadores la necesidad de construir herramientas estadísticas especializadas y por ende el manejo de términos técnicos apropiados. A continuación se revisan los antecedentes de los términos "información auxiliar" y "variable auxiliar" y su relación con los estimadores de regresión dentro de un estudio de muestreo.

En Tillé (2006) se comenta respecto a la disponibilidad de un marco muestral o de un registro administrativo que "es posible mejorar las conclusiones obtenidas de la muestra apoyándose en la información auxiliar de las unidades de interés", y que "en la mayoría de los casos, es posible aprovechar esta información para aumentar la exactitud". Además, se debe considerar que recientemente varios países han recopilado información auxiliar de sus habitantes o de sus viviendas o de sus establecimientos económicos en diversos registros administrativos. Para los países que no disponen de tales registros, Tillé (2006) comenta que "si no fuera el caso, en el diseño de muestreo de dos etapas o bietápico, las unidades primarias son generalmente áreas geográficas, para las cuales hay disponible un gran número de variables auxiliares".

Una de las ventajas principales de hacer uso de información auxiliar es obtener estimadores óptimos bajo modelos de regresión. El Método del Cubo obtiene muestras balanceadas o aproximadamente balanceadas haciendo uso de la correlación de esa información con la variable de interés.

En *Model Assisted Survey Sampling* se menciona respecto al uso de información auxiliar que:

> El énfasis puesto en el uso de información auxiliar para mejorar las precisiones de las estimaciones es característico en teoría del muestreo. El estimador de regresión es un tipo de estimador que intenta hacer eficiente el uso de la información auxiliar relacionada con la población. (Särndal *et al.*, 1992, p. 219)

Respecto a las características de las variables auxiliares, comenta que "hablando generalmente, una variable auxiliar es cualquier variable acerca de la cual disponemos información previa al muestreo". Se supone que dicha información está completa, es decir que "el valor de la variable digamos x, es conocido para cada uno de los N elementos de la población". Agregan que "una variable auxiliar asiste en la estimación de la variable de estudio. El objetivo es obtener un estimador que incremente la exactitud". Respecto a la manipulación o construcción de información auxiliar comentan que "algunos marcos están equipados con una o más variables auxiliares desde el principio, o bien se puede transferir por medio de manipulaciones numéricas simples".

A partir de registros administrativos o de otras fuentes se puede transferir esa información auxiliar al marco por medio de empates.

La información auxiliar se puede usar en la etapa de diseño de una muestra para crear un diseño de muestreo que incremente la precisión del estimador, por ejemplo el de Horvitz-Thompson. También puede usarse en la construcción de estratos de tal forma que dicho estimador, para un diseño de muestreo aleatorio simple estratificado, obtenga una varianza pequeña.

Sin embargo la información auxiliar puede usarse en la etapa de estimación. Aquí las variables auxiliares entran explícitamente en la fórmula del estimador, no únicamente a través de las probabilidades de inclusión como en el caso del diseño.

El supuesto básico es que las variables auxiliares covarían con la variable de estudio, conteniendo información acerca de ésta, unas más y otras menos. Tal covarianza es usada ventajosamente tanto en estimadores de regresión como en muestreo aproximadamente balanceado.

1.3 Antecedentes

A continuación se revisa el desarrollo histórico del muestreo balanceado y del Método del Cubo, considerando su relación con estimadores de regresión.

Neyman (1934) describió varios métodos de muestreo balanceado limitados a una variable y a probabilidades de inclusión iguales.

Royal y Herson (1973 a, b) y Scott *et al.* (1978) comentaron la importancia del muestreo balanceado para proteger la inferencia contra un modelo mal especificado. Ellos propusieron estimadores óptimos bajo un modelo de regresión. En esta aproximación basada en modelos, concibieron a la optimalidad sólo respecto al modelo de regresión, sin tomar en cuenta el diseño de muestreo. No obstante, estos autores llegaron a la conclusión de que la muestra debe ser balanceada pero no necesariamente aleatoria. Cuando ellos desarrollaron esta teoría no existía un método general para seleccionar muestras balanceadas.

Deville y Särndal (1992) investigaron la estimación de totales de poblaciones finitas en presencia de información auxiliar multivariada. Ellos mencionan que "el estimador de regresión general (GREG) fue concebido teniendo en mente información auxiliar multivariada". También comentan que "este estimador es justificado por una relación de regresión entre la variable de estudio y el vector de variables auxiliares".

Sin embargo Deville y Särndal derivaron los GREG por una ruta distinta, se enfocaron en los pesos, mostraron que los pesos implicados por los GREG son tan cercanos como sea posible a los factores de expansión mientras sean respetadas las condiciones llamadas ecuaciones de calibración. Ellos establecen que la suma muestral de las variables auxiliares ponderadas debe ser igual a los totales poblacionales conocidos para esas variables auxiliares. Es decir, los pesos calibrados deben dar estimaciones perfectas cuando las ecuaciones de calibración son aplicadas a cada variable auxiliar.

Por lo tanto, una fuerte correlación entre las variables auxiliares y las variables de estudio significa que los pesos que funcionan bien para las variables auxiliares deben funcionar bien para la variable de estudio.

Valliant *et al.* (2000, p. 49-50) realizaron un estudio de los métodos existentes abordándolos desde tres perspectivas: muestreo balanceado simple, muestreo balanceado ponderado, y una combinación del muestreo balanceado simple con el ponderado.

Por otro lado vieron la necesidad de "establecer una regla para estimación de áreas pequeñas", comentando que "a pesar de la gran cantidad de literatura sobre la materia, estimación de áreas pequeñas es generalmente considerado con inquietud por los que practican el muestreo. No es de sorprender si el investigador asume un paradigma basado en diseño, en el cual los modelos juegan un papel incidental. Así que con frecuencia, son los que están fuera de la profesión del muestreo quienes están más impacientes de hacer uso de métodos de áreas pequeñas".

Valliant *et al.* (2000) también mencionan que "desde un punto de vista de las presunciones de este texto, la inquietud está garantizada, no por el uso de modelos, sino por el mal uso". Y continúan diciendo que "un tema que corre a través de este texto es la necesidad de proteger la inferencia contra un modelo erróneo, ya sea por el uso diestro de métodos de modelo robusto tales como muestreo balanceado, o por la verificación cuidadosa y adecuada del modelo. Debe ser conocido, sin embargo, que la verificación del modelo en el contexto del muestreo es un tema que está en su primera infancia".

En la estimación de áreas pequeñas, recomiendan el uso de validación cruzada para la verificación del estimador de regresión empleado:

> En ningún otro lugar la verificación es más necesaria, y más difícil, que en la circunstancia de estimación de áreas pequeñas. Un componente importante de tal verificación es validación cruzada en dominios pequeños para los cuales los datos están disponibles. Esto involucra la estimación de los parámetros del modelo usando datos fuera del dominio y comparando los resultados predichos en el dominio con los valores reales de la muestra. El grado y tipo de validación cruzada dependerá de la cantidad y del ordenamiento de los datos disponibles dentro del dominio. En el caso de un dominio totalmente vacío de datos, necesitamos investigar la validez de justificaciones a priori para aplicar el modelo, aún cuando esté bien verificado en los datos disponibles. Tales justificaciones deben ser publicadas junto a cualquier estimación. Necesitamos también acompañar las estimaciones con estimaciones de varianza bien fundamentadas. Se requiere hacer mucho trabajo para desarrollar un canon aceptado en forma general de la verificación del modelo y la valoración de la estimación de la varianza para estimación de áreas pequeñas. (Valliant *et al.* 2000 p. 406-407).

Recientemente, Deville y Tillé (2004) propusieron un método general, el Método del Cubo, que permite la selección de muestras aleatorias aproximadamente balanceadas, para una "gran" cantidad de variables auxiliares cualitativas y cuantitativas, con probabilidades de inclusión igual o desigual, en el sentido de que el estimador Horvitz-Thompson de las variables auxiliares es casi igual al total poblacional.

En ese trabajo, algunas simulaciones se efectuan con el archivo de datos MU284 por conglomerados. Este archivo se obtiene del archivo MU284, quitándole los cuatro municipios más grandes y agrupando el resto en cincuenta conglomerados con la variable CL. Las muestras balanceadas fueron seleccionadas con sólo cuatro variables de balanceo: P75, RMT85, SOC82 y ME84. En la sección 5.1 hay una descripción de estas variables.

El Método del Cubo tiene la peculiaridad de que para diez o más variables auxiliares es probable aplicar el método de eliminación de variables en la fase de aterrizaje, cuando no se haya obtenido una muestra en la fase de vuelo, por el riesgo de que el método simplex entre a un ciclo sin fin en el caso de emplear programación lineal.

Por otro lado, la mayoría de las oficinas de estadística oficial generan información con diez o más variables. En el caso de los censos en México a partir del censo de 1921 se han captado más de diez variables y su número sigue en aumento (ver Tabla 1.1).

Tabla 1.1 Censos en México. Número de variables captadas[1]

Año censal	**1895**	**1900**	**1910**	**1921**	**1930**	**1940**	**1950**	**1960**	**1970**
Variables captadas	9	11	9	14	16	22	26	26	31

Ha terminado la fase de generación de grandes cantidades de información a nivel mundial. Ahora los grupos de investigadores en estadística han vuelto la mirada hacia el aprovechamiento de esos conjuntos de datos, para optimizar costos y mejorar sus estimaciones.

[1] http://www.inegi.org.mx/prod_serv/contenidos/espanol/bvinegi/productos/integracion/pais/historicas2/cienanos/EUMCIENI.pdf

Una forma de aprovechar tal información es considerar todas las variables recolectadas en los censos más recientes y emplearlas en los diseños de muestreo. Entonces para estos conjuntos de datos es imprescindible emplear el método de eliminación de variables en la fase de aterrizaje.

De ahí la importancia de proponer un método para el ordenamiento de tales variables auxiliares. Un método que seleccione sólo a las variables más informativas y que sean pocas en número. A continuación se revisa brevemente el surgimiento de la técnica Componentes Principales Supervisadas, la cual reduce la dimensionalidad de un conjunto de datos seleccionando la información más importante.

Componentes Principales Supervisadas tiene su origen en el análisis de datos de pacientes. Donde el número de variables predictoras supera con mucho al número de observaciones. Es una técnica para aplicarla a problemas de análisis de regresión generalizada como el de análisis de supervivencia, en el que ayuda a identificar cuáles variables predictoras son las más importantes. Esta técnica también se propone como una herramienta para el análisis de datos de microarreglo de ADN, donde puede diagnosticarse y tratarse el cáncer con una mayor exactitud.

Componentes Principales Supervisadas surgió de la necesidad de extraer información de células "tipo" a partir tanto de la relación entre la variable de interés y las variables auxiliares como de la correlación entre las variables predictoras en sí mismas. En un mapa de calor (ver la Figura 2 en Bair *et al.* 2006) puede representarse a cada variable en un renglón y a cada columna conteniendo datos de un paciente en un microarreglo. En un estudio de microarreglo con frecuencia a las primeras dos componentes principales se les llama los "eigengenes" (Alter *et al.* 2000).

Supongamos que el microarreglo se divide en dos grupos. Si las variables del grupo A están fuertemente relacionadas a la variable de interés Y, entonces Y estará altamente relacionada con la primera componente principal, a la que los autores la denotan con $\boldsymbol{u_1}$. En este caso se espera que un modelo que usa la componente $\boldsymbol{u_1}$ para predecir Y fuera muy efectivo. Sin embargo, la variación en las variables de A puede reflejar algunos procesos biológicos que no están relacionados a la salida Y. En este caso, Y puede estar más altamente correlacionada con la segunda componente principal $\boldsymbol{u_2}$ o

alguna componente principal de orden más alto. La técnica Componentes Principales Supervisadas está diseñada para descubrir tales estructuras automáticamente. Ésta técnica fue descrita en un ajuste biológico por Bair y Tibshirani (2004) en el contexto de un método relacionado llamado "conglomeración supervisada".

La idea de Componentes Principales Supervisadas es simple, en lugar de aplicar análisis de componentes principales usando todas las variables en un conjunto de datos, se usan sólo aquellas variables con la correlación estimada más fuerte con Y, la variable de interés.

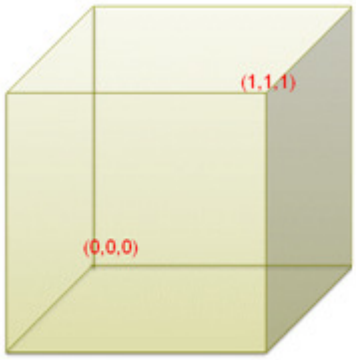
Bair *et al.* (2006) desarrollaron la técnica Componentes Principales Supervisadas y la consideran como un método estándar para modelar correlación. Además utilizan validación cruzada del cociente de verosimilitudes para seleccionar sólo las variables "más importantes".

Una vez revisados algunos antecedentes del uso de información auxiliar para mejorar las estimaciones usando muestras balanceadas, y después de haber revisado el surgimiento de la técnica Componentes Principales Supervisadas para la reducción de la dimensionalidad de los datos, ahora se hace una descripción del problema del ordenamiento estadístico en las variables auxiliares que serán el insumo del Método del Cubo para obtener muestras balanceadas.

1.4 El problema del ordenamiento

El Método del Cubo está basado en una representación geométrica del soporte muestral. Para una población de tamaño $N = 3$, el conjunto de vértices del cubo unitario del primer octante para el cual uno de sus vértices coincide con el origen y cuya magnitud de arista es 1, constituyen el soporte muestral para esa población.

Figura 1.1 Representación geométrica del soporte muestral para $N = 3$.



En esta representación, un algoritmo de muestreo puede ser visto como una forma estocástica de alcanzar un vértice del N-cubo $[0,1]^N$ a partir de un vector de probabilidades de inclusión π, $\pi\epsilon[0,1]^N$, de tal forma que los totales de la variables auxiliares $\boldsymbol{x} = \sum_{k\epsilon U} \boldsymbol{x}_k$ satisfagan las ecuaciones de balanceo:

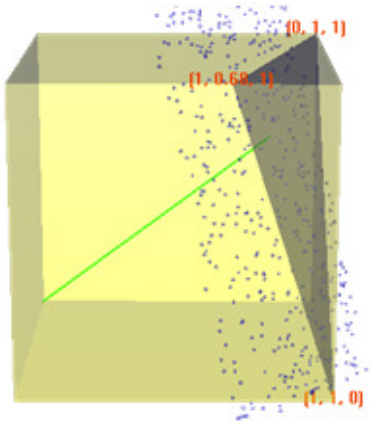
$$\sum_{k\epsilon U} \frac{\boldsymbol{x}_k S_k}{\pi_k} = \sum_{k\epsilon U} \boldsymbol{x}_k \tag{1.1}$$

Parece difícil encontrar un método que detecte cuándo estas ecuaciones se satisfacen exactamente. Esto depende de patrones complejos en $\mathbb{R}^N$. Esos problemas son ya muy intrincados en $\mathbb{R}^3$.

En la fase de vuelo las restricciones son satisfechas en forma exacta y el método utilizado para completarla es por medio de una "martingala balanceada". La fase de aterrizaje es necesaria solo si el vértice alcanzado en el polítopo[2] K no es un vértice del hipercubo C y consiste en lidiar lo mejor posible con el hecho de que las ecuaciones de balanceo no siempre pueden ser satisfechas en forma exacta.

[2] Espacio geométrico resultado de la intersección del "manojo" de hiperplanos de restricción con el hipercubo.

Figura 1.2. El Método del Cubo en $\mathbb{R}^3$ para una sola variable auxiliar.



La función de costo medio condicional, construida con la varianza de las variables auxiliares, es minimizada al resolver el problema de programación lineal correspondiente:

$$min \sum_{s \in S} p(s|\pi^*) \sum_j \frac{\{X_j(s) - X_j\}^2}{X_j^2} \qquad (1.2)$$

El procedimiento de selección de muestras aproximadamente balanceadas se lleva a cabo en dos fases: de vuelo y de aterrizaje. La fase de aterrizaje es necesaria sólo si, al final de la fase de vuelo, el vértice del polítopo no coincide con algún vértice del hipercubo; es decir, si no se ha llegado a la selección de alguna muestra. Esta fase de aterrizaje puede llevarse a cabo por dos métodos: por programación lineal, aplicando un algoritmo simplex para obtener una aproximación, o por eliminación de variables, dejando de considerar la última variable auxiliar (la que está más a la derecha en la matriz de datos auxiliares) y regresando a la fase de vuelo. En la sección 3.3 se describe con más detalle la situación que se presenta en esta fase.

Para más de diez variables auxiliares es muy probable hacer uso del método de eliminación de variables en la fase de aterrizaje:

> Si el número de variables auxiliares es demasiado grande para que el problema de programación lineal sea resuelto por un algoritmo simplex, q>10 en nuestra experiencia, entonces al final de la fase de vuelo una variable auxiliar puede ser descartada. Una restricción es así relajada, permitiendo regresar a la fase de vuelo hasta que no sea posible 'moverse' más dentro del hiperplano de restricciones. Las restricciones son así relajadas sucesivamente. Por esta razón, es necesario ordenar las variables de acuerdo con su importancia de tal forma que las restricciones

> menos importantes sean relajadas primero. Esto naturalmente depende del contexto de la encuesta. (Deville y Tillé 2004. p. 901)

En caso de disponer de un gran número de variables, como en la información recabada por las oficinas de estadística oficial cuya información auxiliar supera las diez variables, surge la necesidad de aplicar el método de eliminación de variables en la fase de aterrizaje. Por ejemplo si se desea aplicar el método a la información obtenida en la Encuesta Nacional de Ingresos y Gastos de los Hogares 2002 (ENIGH2002), la cual arrojó información de más de cien variables para la información de hogares[3], entonces es necesario emplear la opción de eliminación de variables en la fase de aterrizaje, como lo sugieren los autores, y con esto asegurar que el método llegue a una solución. De otro modo, si se emplea el método de programación lineal existe el riesgo de que el algoritmo simplex entre a un ciclo sin fin, como cuando existen soluciones óptimas cubriendo una arista o cara del polítopo, ya que el método simplex (publicado por George Dantzig en 1947) parte de un vértice cualquiera y va cambiando a vértices adyacentes que mejoran el valor que arroja la función objetivo en el vértice anterior.

Por tanto para emplear la opción de eliminación de variables de manera eficiente deben ordenarse previamente las variables de acuerdo con la importancia de su correlación a la variable de interés. Así, las preguntas por resolver son dos:

- ¿Cómo asignar un orden de importancia a ese conjunto de variables para que el método de eliminación las vaya descartando sucesivamente a partir de la menos importante?

De tal forma que, en caso necesario, las variables menos informativas sean eliminadas primero y las más informativas sean eliminadas hasta el final. Y como resultado adicional dar respuesta a la pregunta:

- ¿Cuál es el número óptimo de variables auxiliares a considerar como insumo en el Método del Cubo?

La respuesta a la segunda pregunta ayudaría al investigador a ahorrar recursos al momento de efectuar simulaciones en conjuntos de datos que cuentan con un número

[3] http://www.inegi.org.mx/est/contenidos/espanol/sistemas/enigh/bd/default.asp (hogares.zip)

grande de variables. Para dar respuesta a esta pregunta se emplea validación cruzada del estadístico de prueba cociente de verosimilitudes o "Likelihood Ratio" (LR). A continuación se describe brevemente la solución propuesta.

1.5 Solución propuesta

Para las encuestas en las que hay información auxiliar disponible, el objetivo es obtener muestras balanceadas o aproximadamente balanceadas, empleando las variables auxiliares más correlacionadas con la variable de interés. Los autores del Método del Cubo mencionan que "la reducción en las varianzas de los estimadores de los totales de las variables de interés depende de sus correlaciones con las variables controladas". Entonces para asignar un orden de importancia a las variables auxiliares es imprescindible un método que modele correlación. La técnica Componentes Principales Supervisadas, Bair *et al.* (2006), es considerada por sus autores como un método estándar para modelar correlación. Este método proporciona los scores univariados S_j para cada observación en la variable de interés, que en el caso de regresión gaussiana corresponden a los coeficientes de regresión estandarizados. La forma de calcular estos coeficientes se muestra en la expresión 1.3, en la cual se puede apreciar cómo está implícita la correlación en dicho cálculo.

$$S_j = \frac{x_j^t y}{\sqrt{x_j^t x_j}} = \frac{x_j^t y}{\sqrt{x_j^t x_j}} \cdot \frac{\sqrt{y^t y}}{\sqrt{y^t y}} = \rho_{x_j y}\sqrt{y^t y} \qquad (1.3)$$

La magnitud en valor absoluto de estos coeficientes constituye el criterio de ordenamiento de las variables auxiliares. Los métodos del estadístico de prueba cociente de verosimilitudes y de validación cruzada, ambos para determinar cuántas y cuáles variables deben elegirse, son descritos con detalle en la sección 5.3. Este método puede usarse también en problemas de regresión generalizados como el análisis de supervivencia. La librería "superpc", desarrollada para el lenguaje R, maneja los dos casos: regresión estándar y datos de supervivencia. En este trabajo se utiliza únicamente regresión estándar para dar respuesta a la primera de las dos preguntas. Para responder a la segunda, se emplea esta misma herramienta, la cual forma una matriz de datos reducida, consistente sólo en aquellas variables cuyo coeficiente

univariado excede un umbral θ en valor absoluto. El umbral es estimado por validación cruzada. Además esta herramienta proporciona un gráfico con los resultados de la validación cruzada para darle oportunidad al investigador de seleccionar algún otro umbral que considere pertinente en vez de quedarse con el valor sugerido.

1.6 Objetivos

La finalidad de este trabajo es analizar el comportamiento del Método del Cubo al seleccionar muestras balanceadas considerando sólo las variables auxiliares más informativas y con el orden de importancia determinado por la técnica Componentes Principales Supervisadas.

Objetivos generales:

El objetivo general es evaluar cuantitativamente la ganancia obtenida en la reducción de la varianza de las estimaciones obtenidas con muestras aproximadamente balanceadas, al aplicar la técnica Componentes Principales Supervisadas para ordenar estadísticamente las variables auxiliares y determinar el número óptimo.

- Asignar un orden estadístico a las variables auxiliares seleccionadas, de tal forma que las variables menos importantes sean removidas primero.

- Determinar el número óptimo de variables auxiliares para que el problema de programación lineal, usado en el Método del Cubo, sea resuelto por un algoritmo simplex.

Por medio de la comparación de varios conjuntos de variables auxiliares evaluar cuantitativamente las estimaciones obtenidas con el ordenamiento proporcionado por la técnica Componentes Principales Supervisadas a las variables auxiliares.

Objetivos específicos:

- Evaluar la pérdida en la precisión al considerar sólo las variables auxiliares sugeridas como suficientes por la técnica Componentes Principales Supervisadas comparándola contra el resultado obtenido al considerarlas todas.

- Evaluar cuantitativamente el resultado obtenido con el método de eliminación de variables comparándolo con el método de programación lineal para conjuntos de datos no mayores a diez variables.

- Evaluar por medio de simulaciones la eficiencia en las estimaciones de las variables más informativas obtenidas por la técnica Componentes Principales Supervisadas.

- Evaluar cuantitativamente el comportamiento de las estimaciones al hacer simulaciones en varias regiones del país, considerando una cantidad y un orden de variables auxiliares predeterminados.

Meta:

La meta es hacer tres mil simulaciones en cada uno de los primeros dos ejercicios, correspondientes al conjunto de datos MU284; y hacer cuatro y cinco mil simulaciones en los ejercicios tres y cuatro respectivamente, correspondientes a los datos de ENIGH2002.

Hipótesis:

Lo que se espera, al proponer este método de ordenamiento, es que el error de las estimaciones obtenidas por el Método del Cubo, considerando el ordenamiento estadístico de las variables auxiliares, sea menor al error obtenido sin ordenarlas.

1.7 Delimitación

En este trabajo las simulaciones se hacen sobre muestras balanceadas obtenidas por el Método del Cubo. Se emplea el algoritmo rápido. El diseño de selección es con probabilidades de inclusión desiguales, es decir, con probabilidad proporcional al tamaño (PPT) sin reemplazo. Se estima el total de la variable de interés. Se emplea el estimador Horvitz-Thompson.

Aunque la información de ENIGH2002 corresponde a una muestra, en este trabajo dicha información se considera como la correspondiente a una determinada población finita. Se quitan, del conjunto ENIGH2002, diez variables que contienen una gran cantidad de valores omitidos. Y para que el conjunto de entrenamiento no

cause problemas en la técnica Componentes Principales Supervisadas, se quitan aquellos registros del archivo con valor cero en la variable de interés "ingreso corriente monetario" (ingmon). Además se emplea el supuesto de normalidad en las variables para obtener los coeficientes de regresión.

1.8 Estructura del documento

Este trabajo está estructurado de la siguiente manera. Primero, siguiendo la ya común regla número uno, se efectúa una revisión de los métodos existentes. Esto se hace en el capítulo II. Luego se hace una breve descripción, en el capítulo III, de las dos fases del Método del Cubo, del problema del ciclado en la fase de vuelo y de la técnica Componentes Principales Supervisadas. Se analiza el enlace entre ellos, cómo apoya este último al primero, y sus principales características.

Descripciones de los algoritmos empleados tanto en el cálculo de coeficientes de regresión estandarizados, como en el Método del Cubo, y en la técnica Componentes Principales Supervisadas son efectuadas en el capítulo IV. En el capítulo V se describen las características de la información sobre la que se efectúan las simulaciones en los datos MU284 y ENIGH2002, el procedimiento empleado para efectuar las simulaciones, el resultado de las simulaciones, la evaluación y los comentarios.

Finalmente en el capítulo VI se plantean las conclusiones obtenidas al efectuar las simulaciones en cada uno de los conjuntos de datos MU284 y ENIGH2002, y se abordan los desarrollos posteriores que se vislumbrar a partir de este ejercicio.

C a p í t u l o I I

MÉTODOS EXISTENTES

[Los milagros] Son la opción cuando no sirve soñar, lo que elige el soñador en vez de negar el papel activo al imaginar el sueño.

Helen Schucman y William Thetford
Meditaciones de un curso de milagros

Después de haber descrito brevemente el problema planteado y la solución propuesta en este apartado se revisan los cuatro métodos considerados para darle solución al problema: en la sección 2.2 Análisis de Componentes Principales, Mínimos Cuadrados Parciales en la sección 2.3, Componentes Principales Supervisadas en la sección 2.4 y finalmente en la sección 2.5 Análisis de Discriminante No Paramétrico.

Antes de revisar cada uno de los métodos considerados, en la sección 2.1 se citan dos puntos de vista relacionados con el trabajo de todo investigador en estadística. El primero trata de la existencia de dos culturas en el modelado estadístico según Leo Breiman. En el segundo, Gerard Dallal aborda el problema del significado de la pregunta: ¿cuáles predictores son los más importantes? Estos dos puntos de vista ayudan a clarificar las propiedades del método como herramienta estadística.

2.1 Consideraciones previas

En su trabajo, Breiman (2001), además de proporcionar un punto de vista meta-estadístico acerca del trabajo del investigador, da una sugerencia para clasificar las herramientas estadísticas mencionando que:

> Hay dos culturas en el uso de modelación estadística para obtener conclusiones a partir de los datos. Una supone que los datos son generados por un modelo estocástico de datos predeterminado. La otra usa modelos algorítmicos y trata al mecanismo que generó los datos como desconocido. (Breiman 2001, p. 199)

También menciona que:

> La modelación algorítmica, tanto en teoría como en la práctica, se ha desarrollado rápidamente en campos fuera de la estadística. Esta puede ser usada en grandes conjuntos de datos complejos y como una alternativa más exacta e informativa para modelar conjuntos de datos más pequeños. Si nuestro objetivo, como un campo, es usar datos para resolver problemas, entonces

> necesitamos salirnos de la dependencia exclusiva en modelos de datos y adoptar un conjunto de herramientas más diverso. (Breiman 2001, p. 199)

Esta nueva forma de afrontar los problemas conlleva a considerar como estrategias de solución dos opciones: modelación estocástica o modelación algorítmica. Dentro de la modelación algorítmica se revisa la estadística no paramétrica. Esta estrategia ha tomado fuerza recientemente, sobre todo con la proliferación de grandes volúmenes de información, un ejemplo claro es la minería de datos. De acuerdo con Breiman, un trabajo estadístico que considere únicamente métodos paramétricos como posible solución se queda corto en su propuesta, al grado de proponer soluciones inútiles o ineficientes.

Por otro lado el problema planteado en esta investigación está basado en el uso de la correlación de la variable de interés con las variables auxiliares, por tanto una pregunta interesante es ¿cuáles predictores son los más importantes? Al respecto hay una gran variedad de trabajos de los cuales se rescata uno que aborda de manera general la respuesta a esta pregunta. En su Manual de Estadística, Dallal (2001)[4] comenta respecto a la "importancia estadística" que:

> Cuando se lleva a cabo una regresión múltiple, no es raro que alguien pregunte cuáles predictores son los más importantes.
>
> Una posibilidad es medir la importancia de una variable usando la magnitud de su coeficiente de regresión. Este acercamiento falla porque los coeficientes de regresión dependen de la escala subyacente de medidas.
>
> Otra posibilidad es medir la importancia de una variable por su nivel de significancia observado (valor P). Sin embargo la distinción entre importancia estadística e importancia práctica aplica aquí también.

Dallal se opone a los defensores del uso de coeficientes de regresión estandarizados diciendo que:

> No hay ninguna razón por la cual un cambio de una D.E. en un predictor debería ser equivalente a un cambio de una D.E. de otro predictor. Algunas variables son fáciles de cambiar: la cantidad de tiempo mirando la televisión, por ejemplo. Otras son más difíciles: el nivel de colesterol o el peso. Otras son imposibles: la altura o la edad.

Y continúa con su oposición enfocándose a la relación causa-efecto que existe en los fenómenos sociales argumentando que:

[4] http://www.tufts.edu/~gdallal/importnt.htm

> El investigador y el analista deberían considerar los cambios específicos de cada predictor y el efecto que ellos tendrían sobre la respuesta. Algunos predictores no podrían ser cambiados, independientemente de sus coeficientes. No se trata de preguntar cuál es el que mayormente determina la respuesta, sino cuál es crítico, si el objetivo del ejercicio es desarrollar políticas públicas para efectuar un cambio en la respuesta. Cuando los predictores pueden ser modificados, los investigadores tendrán que decidir cuáles cambios son factibles y cuáles son comparables.

Finalmente considera el factor costo al momento de intentar un cambio en la variable de interés, inclusive cuando el cambio es pequeño:

> El costo también entrará en la discusión. Por ejemplo, suponga que un cambio en la respuesta puede obtenerse ya sea por un cambio grande en un predictor o por un cambio pequeño en otro. De acuerdo con las circunstancias puede proveer un mayor costo efectivo intentar un cambio grande que uno pequeño.

La relación entre las variables auxiliares y la variable de interés puede ser acentuada, pasando desde la dependencia total o dependencia funcional hasta la independencia. En los fenómenos de la vida cotidiana existen diversas asociaciones o relaciones entre ellos. Asociación con dependencia, asociación con interdependencia, relación causa-efecto, relación con un factor común y ausencia de relación. El análisis de Dallal tiende a enfocarse en la relación causa-efecto.

La opinión de Dallal explota puntos de vista que no habían sido abordados de la manera como él lo hace, revisando varias herramientas estadísticas para responder a la pregunta, pero como nuestro objetivo no es desarrollar políticas públicas para efectuar un cambio en la respuesta, como él lo enfatiza en el último párrafo citado, sino encontrar correlaciones en los datos que auxilien en la realización de una inferencia de mayor precisión, su opinión es irrelevante en el aspecto aquí considerado. El objetivo del presente trabajo es detectar cuáles predictores explican en mayor medida la varianza de la variable de respuesta o de interés.

Los métodos para detectar los predictores más explicativos que se revisan en este trabajo están clasificados en dos grupos, de acuerdo con la propuesta de Breiman. En **modelación estocástica** se agrupan Análisis de Componentes Principales, Mínimos Cuadrados Parciales y Componentes Principales Supervisadas. El único método de **modelación algorítmica** que se revisa es Análisis de Discriminante No Paramétrico.

2.2 Análisis de Componentes Principales

Análisis de Componentes Principales (ACP) es una técnica estadística utilizada para simplificar un conjunto de datos. Reduce un conjunto multidimensional a dimensiones más bajas para su análisis.

Hacer análisis de componente principales de un conjunto de datos significa encontrar una transformación lineal que lleve los datos a un nuevo sistema de coordenadas, de tal forma que la varianza más grande de cualquier proyección de los datos quede en la primera coordenada, llamada primera componente principal, la segunda varianza más grande en la segunda coordenada, y así sucesivamente. Este método retiene aquellas características del conjunto de datos que contribuyen más a su varianza, quedándose con componentes principales de bajo orden e ignorando aquellas de orden más alto. Tales componentes de bajo orden con frecuencia contienen los aspectos "más importantes" de los datos.

La técnica ACP se debe a Hotelling (1933) aunque sus orígenes se encuentran en los ajustes ortogonales por mínimos cuadrados introducidos por Karl Pearson (1901) (Peña 2002, p. 134).

Esta herramienta tiene la distinción de ser la transformación lineal óptima que mantiene el subespacio de la varianza más grande. Esta ventaja tiene el costo de requerimientos computacionales más grandes si se compara, por ejemplo, con la transformación coseno discreto. Como sucede con otras transformaciones lineales, ACP no tiene un conjunto fijo de vectores base. Sus vectores base dependen del conjunto de datos. Este método es equivalente a encontrar la descomposición en valor singular de la matriz de datos $\mathbf{X} = \mathbf{W\Sigma V^t}$, y luego obtener la matriz de datos en el espacio reducido $\mathbf{Y}$ proyectando $\mathbf{X}$ en el espacio reducido definido sólo por los primeros L vectores singulares $\mathbf{W}_L$:

$$\mathbf{Y} = \mathbf{W}_L^t\mathbf{X} = \mathbf{\Sigma}_L\mathbf{V}_L^t \tag{2.1}$$

La matriz $\mathbf{W}$ de vectores singulares de $\mathbf{X}$ es equivalente a la matriz de vectores propios de la matriz de covarianzas observadas $\mathbf{C} = \mathbf{XX^t} = \mathbf{W\Sigma\Sigma^tW^t}$. Los vectores

propios con los valores propios más grandes corresponden a las dimensiones que tienen la correlación más fuerte en el conjunto de los datos. Esta técnica es popular en reconocimiento de patrones, ya que minimiza de manera óptima el error de reconstrucción bajo la norma L_2. Pero no es óptima para separar clases. Una alternativa es el análisis de discriminante, el cual toma esto en cuenta.

Una vez revisado este método que considera la varianza de las variables para reducir la dimensión del conjunto, ahora corresponde revisar las propiedades de otro método que además considere la correlación entre las variables auxiliares y la variable de interés. Esto se hace de manera breve en la siguiente sección.

2.3 Mínimos Cuadrados Parciales

Una técnica que en la estimación del modelo de regresión, además de considerar la varianza como en componentes principales, considera la correlación es mínimos cuadrados parciales. Hastie *et al.* (2001) mencionan al respecto que:

> Esta técnica también construye un conjunto de combinaciones lineales de las entradas para la regresión, pero a diferencia de regresión por componentes principales esta usa $\mathbf{y}$ (además de $\mathbf{X}$) para su construcción. Se supone que $\mathbf{y}$ está centrada y que cada $\mathbf{x_j}$ se estandariza para tener media $\mathbf{0}$ y varianza $\mathbf{1}$. La técnica PLS comienza por calcular los coeficientes de regresión univariada $\hat{\varphi}_{1j}$ de $\mathbf{y}$ en cada $\mathbf{x_j}$, es decir $\hat{\varphi}_{1j} = \langle \mathbf{x_j}, \mathbf{y} \rangle$. A partir de esto se construye la entrada derivada $\mathbf{z} = \sum \hat{\varphi}_{1j}\mathbf{x_j}$ la cual es la primera dirección de mínimos cuadrados parciales. De ahí que en la construcción de cada $\mathbf{z_m}$, las entradas son ponderadas por la contracción de su efecto univariado sobre $\mathbf{y}$. A la salida $\mathbf{y}$ se le obtiene su regresión sobre $\mathbf{z_1}$ dando el coeficiente $\hat{\theta}_1$ y luego se ortogalizan $\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_p}$ con respecto a $\mathbf{z_1}$. Se continúa con este proceso hasta que $\mathbf{M} \leq \mathbf{p}$ direcciones hayan sido obtenidas. De esta manera, mínimos cuadrados parciales produce una secuencia de entradas derivadas o direcciones $\mathbf{z_1}, \mathbf{z_2}, \ldots \mathbf{z_M}$. Así como en regresión por componentes principales, si se fueran a construir todas las $\mathbf{M} = \mathbf{p}$ direcciones regresaríamos una solución equivalente a las estimaciones usuales de mínimos cuadrados. Usando $\mathbf{M} < p$ direcciones produce una regresión reducida. El procedimiento se describe completamente en el Algoritmo 2.1.

Enfatizan que a diferencia de componentes principales supervisados la dirección de varianza grande elegida debe tener una alta correlación con la variable respuesta.

> ¿Cuál problema de optimización está resolviendo mínimos cuadrados parciales? Ya que usa la respuesta $\mathbf{y}$ para construir sus direcciones, su solución es una función no lineal de $\mathbf{y}$. Se puede demostrar que mínimos cuadrados parciales busca direcciones que tienen varianza grande y que tienen una alta correlación con la respuesta, en contraste con regresión por componentes

principales (Stone y Brooks, 1990; Frank y Friedman, 1993). En particular, la m-ésima dirección de componente principal v_m resuelve:

$$\max_{\|\alpha\|=1,\ v_\ell^T S\alpha=0,\ \ell=1,\ldots,m-1} \mathrm{Var}(X\alpha), \tag{2.2}$$

Donde S es la matriz de covarianza muestral de x_j. Las condiciones $v_\ell^T S\alpha = 0$ aseguran que $z_m = X\alpha$ no está correlacionada con todas las combinaciones lineales previas $z_\ell = Xv_\ell$. La m-ésima dirección PLS $\hat{\varphi}_m$ resuelve:

$$\max_{\|\alpha\|=1,\ \hat{\varphi}_\ell^T S\alpha = 0,\ \ell = 1,\ldots,m-1} \mathrm{Corr}^2(y, X\alpha)\mathrm{Var}(X\alpha), \tag{2.3}$$

A continuación se presenta el algoritmo para calcular los coeficientes de regresión por medio del método de mínimos cuadrados parciales.

ALGORITMO 2.1. Mínimos cuadrados parciales.

---

1. Estandarizar cada x_j para que tengan media 0 y varianza 1. Inicializar $\hat{y}^{(0)} = \mathbf{1}\bar{y}$, y $x_j^{(0)} = x_j$, $j = 1,2,\ldots,p$.

2. Para $m = 1,2,\ldots,p$

   - $z_m = \sum_{j=1}^{p} \hat{\varphi}_{mj} x_j^{(m-1)}$, donde $\hat{\varphi}_{mj} = \langle x_j^{(m-1)}, y \rangle$.
   - $\hat{\theta}_m = \langle z_m, y \rangle / \langle z_m, z_m \rangle$.
   - $\hat{y}^{(m)} = \hat{y}^{(m-1)} + \hat{\theta}_m z_m$.
   - Ortogonalizar cada $x_j^{(m-1)}$ con respecto a z_m:

   $$x_j^{(m)} = x_j^{(m-1)} - \left[\langle z_m, x_j^{(m-1)} \rangle / \langle z_m, z_m \rangle\right] z_m,\ j = 1,2,\ldots,p.$$

3. Dar como salida la secuencia de vectores ajustados $\{\hat{y}^{(m)}\}_1^p$. Debido a que los $\{z_\ell\}_1^p$ son lineales en el x original, entonces $\hat{y}^{(m)} = X\hat{\beta}^{pls}(m)$. Esos coeficientes lineales se pueden recuperar de la secuencia de transformaciones PLS.

Una desventaja de este método, desde el punto de vista del uso de la correlación, lo revela el siguiente comentario de Hastie *et al.* (2001).

Análisis más profundos revelan que el aspecto de la varianza tiende a dominar, y así mínimos cuadrados parciales se comporta muy parecido a ridge regression y a regresión por componentes principales.

Finalmente Hastie *et al.* (2001) mencionan que si la matriz de información auxiliar es ortogonal entonces el método encuentra el modelo en el primer paso. Porque el método ortogonaliza todas las variables auxiliares a partir de la primera.

Si la matriz de entradas $\mathbf{X}$ es ortogonal, entonces mínimos cuadrados parciales encuentra las estimaciones mínimas cuadradas después de $m = 1$ pasos. Los pasos subsecuentes no tienen efecto ya que las $\hat{\varphi}_{mj}$ son cero para $m > 1$. También se puede demostrar que la secuencia de coeficientes PLS para $m = 1, 2, \ldots, p$ representa la secuencia del gradiente conjugado para calcular las soluciones mínimas cuadradas.

Hasta aquí termina la revisión de las características esenciales de este método. Ahora corresponde revisar una herramienta estadística, de reciente construcción, que combina de manera apropiada el considerar tanto las correlaciones altas como la varianza máxima para encontrar el modelo de regresión.

2.4 Componentes Principales Supervisadas

La técnica Componentes Principales Supervisadas es similar a la de Análisis de Componentes Principales convencional, excepto que usa sólo un subconjunto de las variables auxiliares, seleccionadas de acuerdo con sus asociaciones a la variable de interés. Considera los efectos de otras covariables y ayuda a identificar cuáles variables predictoras son las más importantes. Un resumen muy general del procedimiento de esta técnica en cuatro pasos es el siguiente.

1. Calcular los coeficientes de regresión estándar (univariados) para cada variable.
2. Formar una matriz de datos reducida $\mathbf{X}_\theta$ con aquellas variables cuyos coeficientes exceden un umbral θ en valor absoluto (θ es estimado por validación cruzada del logaritmo del cociente de verosimilitudes de las correlaciones entre las variables auxiliares y la variable de salida).
3. Calcular la(s) primera componente principal de la matriz de datos reducida.
4. Utilizar esa(s) componente principal en un modelo de regresión para predecir la salida.

Puede iterarse este procedimiento, pero hacerlo repetidamente converge al método de componentes principales usuales. Hacer iteraciones de este procedimiento tendría sentido sólo si estuviera basado en un criterio que involucrara tanto la varianza de las variables auxiliares como la bondad de ajuste de la variable de interés.

Más adelante se ve que para el desarrollo de este trabajo, únicamente se utilizan los primeros dos pasos de este procedimiento. En la sección 3.4 se hace una descripción a detalle de estos dos pasos.

Una explicación de este método puede resumirse de la siguiente manera. Suponiendo que se tiene una variable respuesta Y que está relacionada a una variable latente subyacente U por un modelo lineal,

$$Y = \beta_0 + \beta_1 U + \varepsilon \tag{2.4}$$

Además se tiene la expresión de las medidas de un conjunto de variables independientes $X_j = \alpha_{0j} + \alpha_{1j} U + \epsilon_j$ con subíndices $j \in \mathcal{P}$, para las cuales nos gustaría identificar el conjunto $\mathcal{P}$, estimar la variable U y luego ajustar el modelo de predicción. Este es un caso especial de modelo de estructura latente o modelo de Análisis de Factores de componente simple.

El modelo de variable latente puede extenderse fácilmente para acomodar múltiples componentes $U_1, \ldots, U_m$. Una forma de hacer esto es suponer que

$$Y = \beta_0 + \sum_{m=1}^{M} \beta_m U_m + \varepsilon \tag{2.5}$$

y $$X_j = \alpha_{0j} + \alpha_{1j} U + \epsilon_j, \qquad j \in \mathcal{P} \tag{2.6}$$

Para ajustar este modelo se procede como antes, excepto que ahora se extraen M componentes principales de la matriz de datos reducida $\mathbf{X_0}$ en vez de solo una.

2.5 Análisis de discriminante no paramétrico

En esta sección se revisa el único método correspondiente a modelación algorítmica. Para una mejor comprensión del por qué se considera esta herramienta como una

posibilidad para darle solución al problema planteado, en su lectura debe tenerse presente que cuando se hable de "clase" o "dirección" en modelación algorítmica debe pensarse en el concepto de variable auxiliar como un análogo en modelación estadística.

Los tres métodos anteriores corresponden a la clase de modelación estadística. Dentro de los métodos correspondientes a la clase de modelación algorítmica, el análisis de discriminante no paramétrico (NPDA) es el único que se revisa.

En Zhu y Hastie (2003) se menciona respecto a esta herramienta que:

> Análisis de discriminante no paramétrico es un método general para encontrar direcciones discriminantes importantes sin suponer que las densidades de las clases pertenecen a una familia paramétrica en particular. Puede integrarse con el método de estimación de densidades por 'projection pursuit' (búsqueda de proyección) para producir un procedimiento poderoso para el análisis de discriminante no paramétrico de rango reducido. Zhu y Hastie 2003 (p. 101)

Comienzan su análisis citando las características de la herramienta Análisis de Discriminante Lineal.

> Para una mejor comprensión del criterio de Fisher (LDA), considerémoslo desde el punto de vista de la verosimilitud. Supongamos que el vector predictor x de la clase k tiene función de densidad $p_k(x)$. Consideremos:
>
> $$H_0\colon\ p_k = p \text{ for all } k = 1,2,\dots,K$$
> $$H_A\colon\ p_k \neq p \text{ for some } k = 1,2,\dots,K$$
>
> Zhu y Hastie 2003 (p. 101,104)

Es decir se considera como hipótesis nula que todas las clases provienen de una misma población y por lo tanto se pueden modelar con una misma función de densidad de probabilidad. Contra la hipótesis alternativa de que existe al menos un elemento que no pertenece a la única población supuesta. Y continúa mencionando que:

> En este contexto, un candidato natural para medir diferencias de clases en una dirección fija α es el estadístico (marginal) generalizado de la razón de log-verosimilitudes:
>
> $$LR(\alpha) = log\frac{\max_{P_k} \prod_{k=1}^{K} \prod_{x_j \in C_k} P_k^{(\alpha)}(\alpha^T x_j)}{\max_{P_k = P} \prod_{k=1}^{K} \prod_{x_j \in C_k} P^{(\alpha)}(\alpha^T x_j)}$$

> En el numerador tenemos expresado el supuesto en el cual cada una de las k clases comparte la misma función de densidad marginal $p_k^{(\alpha)}(\cdot)$ a lo largo de la proyección definida por α para la clase k; esto es contrastado vía razón de verosimilitudes contra el supuesto en el cual todas las clases pertenecen a una misma función de densidad. Puede demostrarse por medio de cálculos algebraicos directos (Zhu 2001) que el criterio usado en LDA es un caso especial de $LR(\alpha)$. Zhu y Hastie 2003 (p. 104)

Como dato curioso, las estrategias utilizadas por el Método del Cubo en la fase de aterrizaje son análogas con los métodos empleados en este trabajo de Zhu y Hastie, donde se menciona que:

> Dos de las estrategias para encontrar direcciones discriminantes múltiples son la ortogonalización y la eliminación de características. Zhu y Hastie 2003 (p. 106).

Después describen con mayor detalle cada uno de los métodos, mencionando el uso de modelos de programación matemática en la estrategia de ortogonalización:

> En la estrategia de ortogonalización suponemos que hemos encontrado $m-1$ direcciones discriminantes y entonces el problema de encontrar la m-ésima dirección discriminante se reduce a un problema de programación matemática (MPL) sujeto a una métrica donde x y y son ortogonales. En estadística multivariada uno supone que los datos siguen una distribución Gaussiana multidimensional $N(\mu, \Sigma)$, así que hay un sistema coordenado natural (es decir, $\Phi = \Sigma$) en el cual es interesante y significativo enfocarse en las características que son ortogonales unas con otras. La ortogonalidad $\alpha^T \Sigma \alpha_j$ implica que $U = X\alpha$ y así los U_j no están correlacionados. Para datos no-gaussianos, no existe tal métrica natural. De hecho, para datos separados en K clases, aún si suponemos que los datos en cada clase son Gaussianos, no es claro cuál es la métrica apropiada, a menos que uno suponga, como en LDA, que todas las clases comparten una matriz de covarianza COMÚN. En la práctica uno 'esferiza' los datos antes de analizarlos. Esto es lo mismo que usar la matriz de covarianza total como una métrica específicamente para ortogonalización. Otra selección razonable para $\sum$ es la matriz de covarianza intra-clase, la cual, cuando p_k es Gaussiana, es la misma que LDA. En general, no hay razón para decir que la matriz de covarianza total o intra-clase es la métrica apropiada. A pesar de ser poco específica, "ortogonal features" es útil para algunas aplicaciones, tales como visualización de datos.
>
> En la estrategia de eliminación de características (feature removal) suponemos que una dirección α discriminante es encontrada, simplemente transformamos los datos de tal forma que no haya diferencia de clase en la dirección α, mientras se mantienen las otras direcciones sin cambio, y se busca la próxima dirección. Zhu y Hastie 2003 (p. 104-108).

Este método tiene alguna relación con Análisis de Componentes Principales. En Análisis de Discriminante No Paramétrico si en vez de buscar en cualquier dirección, se busca en las direcciones de las variables auxiliares se llega al caso particular de componentes principales. Si se considera cada variable auxiliar como una clase

entonces análisis de discriminante no paramétrico proporciona la dirección más informativa, sin redundancia. Comúnmente la redundancia existe en las variables auxiliares sobre todo en las muy correlacionadas.

Para finalizar con este método se insiste en el hecho de que las dos estrategias, "programación matemática" y "eliminación de características" utilizadas para enfrentar un problema de NPDA (publicado en 2003), son similares a las de "programación lineal" y "eliminación de variables" utilizadas en la fase de aterrizaje del Método del Cubo (publicado en 2004) para obtener muestras balanceadas o aproximadamente balanceadas.

C a p í t u l o I I I

MÉTODO DE ORDENAMIENTO

Un buen maestro debe creer en las ideas que enseña, pero también […] debe creer en los estudiantes a quienes ofrece sus ideas.

Helen Schucman y William Thetford
Meditaciones de un curso de milagros

Después de haber hecho un repaso de los métodos considerados como candidatos para la solución del problema, ahora corresponde describir con mayor detalle el método elegido y la técnica empleada para el ordenamiento de las variables auxiliares.

En la sección 3.1 se delimita el concepto de variables auxiliares fijas. Como no todas las variables auxiliares disponibles están correlacionadas, en forma independiente, con la variable de interés, sobre todo las correspondientes al diseño, entonces se hace una clasificación de ellas para distinguirlas y evitar el manejo de información redundante.

En la sección 3.2 se describen las dos fases del Método del Cubo, sus principales características y bajo cuáles condiciones puede haber problemas para seleccionar la muestra.

En la sección 3.3 se explica con detalle el problema en el que se puede ver envuelto el Método del Cubo al emplear programación lineal en la fase de aterrizaje en un conjunto de datos con más de diez variables auxiliares. Y la necesidad de emplear el método de eliminación de variables en esta fase para evitar el ciclado del algoritmo simplex.

En la sección 3.4 se explica, con detalle también, la técnica que se está utilizando para ordenar las variables auxiliares y con este ordenamiento hacer eficiente el método de eliminación de variables, tan igual o muy poco diferente a emplear programación lineal.

3.1 Variables auxiliares fijas

Antes de que el método de ordenamiento sea descrito, debe delimitarse el concepto de información auxiliar. Este concepto es utilizado tanto en estadística paramétrica como no paramétrica. En estadística no paramétrica constituye tanto el conjunto de entrenamiento como el de prueba.

En estadística paramétrica, específicamente en teoría del muestreo se pone énfasis en el uso de información auxiliar para mejorar las precisiones de las estimaciones.

En el ámbito de muestreo constituye básicamente el grupo de variables que ayudarán al investigador a obtener estimaciones con menor sesgo y a obtener varianzas más pequeñas.

Para el método aquí descrito conviene clasificar a las variables auxiliares en fijas y no fijas. Se considera como variable fija a aquella variable que se utiliza para el diseño o plan de muestreo. Dentro de las fijas se pueden encontrar tres tipos de variables:

- Las correspondientes al plan de muestreo. Aquí entra el "tamaño de muestra" y las que identifican a los estratos y conglomerados.
- Las que clasifican a los dominios geográficos; por ejemplo regiones, estados, municipios y localidades.
- Las que clasifican a los dominios no geográficos o dominios estadísticos. En este caso debe considerarse a las variables que clasifiquen cualquier tipo de división estadística, por ejemplo la "clase de actividad" para establecimientos económicos, y los "grupo de edad" y "sexo" para hogares.

Ahora bien, dentro de las variables no fijas se considera al resto de variables disponibles, obtenidas de investigaciones estadísticas anteriores, ya sean censos, encuestas o registros administrativos. Y también comprenden a todas aquellas variables obtenidas de fuentes de información ajenas a las estadísticas oficiales.

3.2 Descripción breve del Método del Cubo

El Método del Cubo emplea información auxiliar para mejorar las estimaciones por medio de su correlación con la variable de interés. El método consta de dos fases, llamadas por sus autores como "de vuelo" y "de aterrizaje".

Las ecuaciones de balanceo (1) también se pueden escribir como:

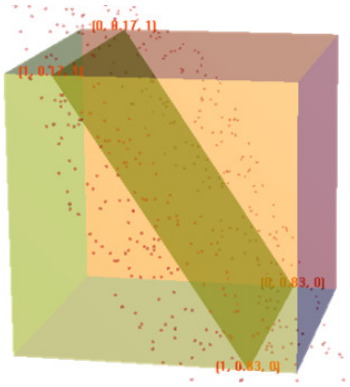
$$\sum_{k\epsilon U} a_k s_k = \sum_{k\epsilon U} a_k \pi_k, \qquad s_k \epsilon \{0,1\}, \qquad k\epsilon U, \qquad (3)$$

donde $a_k = x_k/\pi_k \ (k\epsilon U)$ y s_k es igual a 1 si la unidad k está en la muestra o 0 en otro caso. La primera ecuación de (3) con a_k dado y con coordenadas s_k define un hiperplano Q en $\mathbb{R}^N$ de dimensión $N - p$. Note que $Q = \pi + \ker A$, donde $\ker A$ es el kernel o espacio nulo de la matriz A de pxN dada por $A = (a_1 \dots a_k \dots a_N)$. La idea principal al obtener una muestra balanceada es seleccionar un vértice del N-cubo que permanezca en el hiperplano Q o cerca de Q si no fuera posible.

Si $C = [0,1]^N$ denota al N-cubo en $\mathbb{R}^N$ cuyos vértices son las muestras de U, entonces la intersección entre C y Q es no vacía, porque π está en el interior de C y pertenece a Q. La intersección entre un N-cubo y un hiperplano define un polítopo $K = C \cap Q$, el cual es de dimensión $N - p$ porque es la intersección de un N-cubo y un plano, de dimensión $N - p$, que tiene un punto en el interior de C. Deville y Tillé (2004).

Al final de la fase de vuelo, un vértice del polítopo $K = C \cap Q$ es seleccionado de tal forma que las probabilidades de inclusión $\pi_k \ (k \in U)$, y las ecuaciones de balanceo $\sum x_{kj} s_k / k_\pi = \sum x_{kj}$, $j = 1, \dots, p$, correspondientes a p variables auxiliares, se satisfacen en forma exacta.

La fase de aterrizaje es necesaria sólo si el vértice alcanzado en el polítopo K no es un vértice del hipercubo C, es decir cuando es imposible satisfacer las ecuaciones de balanceo. Consultar el ejemplo 6 en Deville y Tillé (2004).



Esta fase puede efectuarse por dos procedimientos: programación lineal o eliminación de variables.

El método de programación lineal es adecuado para conjuntos de datos con a lo más diez variables auxiliares según experiencias de Deville y Tillé. Para conjuntos mayores aumenta la posibilidad de que el método simplex entre a un ciclo sin fin, como se describe en la siguiente sección.

El procedimiento de eliminación de variables, consiste en ir relajando una restricción a la vez hasta que una muestra sea seleccionada.

3.3 El problema del ciclado

En el método de programación lineal las restricciones lineales, correspondientes a cada una de las variables auxiliares, geométricamente definen un poliedro convexo llamado región factible (Winston 2005, p. 138). Ya que la función objetivo también es lineal (modelo de programación lineal), y por tanto una función convexa, entonces por el teorema de Kuhn-Tucker (KT) todos los óptimos locales son óptimos globales automáticamente (Winston 2005, p. 670-671). La linealidad de la función objetivo también implica que una solución óptima sólo puede ocurrir en un punto frontera de la región factible, a no ser que la función objetivo sea constante, en cuyo caso cualquier punto es un máximo global.

Hay dos situaciones en las cuales no se puede encontrar ninguna solución óptima. Primero, cuando las restricciones se contradicen entre ellas, por ejemplo $x \geq 2$ y $x \leq 1$, entonces la región factible está vacía y no puede haber ninguna solución óptima, debido a que no hay ninguna solución en absoluto. En este caso, el problema de programación lineal, se dice que NO es factible. La segunda situación es cuando el poliedro es ilimitado en la dirección de la función objetivo, por ejemplo al maximizar $x_1 + 3x_2$ sujeto a $x_1 \geq 0$, $x_2 \geq 0$ y $x_1 + x_2 \geq 10$; en este caso no hay ninguna solución óptima, ya que pueden construirse soluciones con valores arbitrariamente altos de la función objetivo (Winston 2005, p. 67).

Salvo estas dos condiciones patológicas, que a menudo son excluidas por restricciones de recursos integrales al problema que está siendo representado, el óptimo siempre es alcanzado en un vértice del poliedro.

El modelo de programación lineal para el Método del Cubo es el siguiente:

$$\underset{p(\cdot\,|\pi^*)}{min} \sum_{s\in\mathcal{C}(\pi^*)} \left[p(s|\pi^*) \sum_j \frac{\left\{\hat{X}_j(s) - X_j\right\}^2}{X_j^2} \right] \tag{1.2}$$

Sujeto a

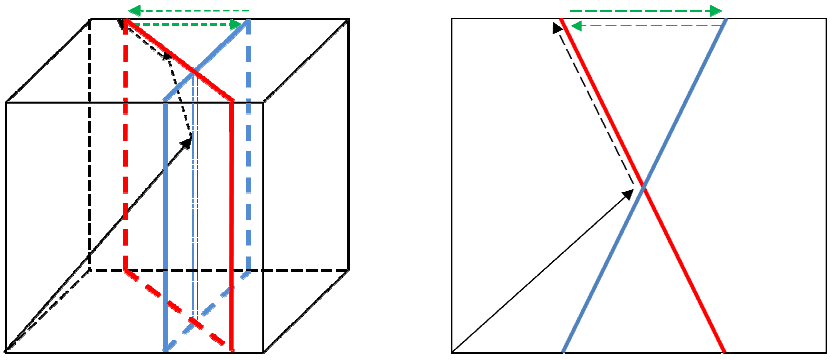
$$\sum_{s\epsilon\mathcal{C}(\pi^*)} p(s|\pi^*) = 1\,, \quad \sum_{s\epsilon\mathcal{C}(\pi^*)|s\ni k} p(s|\pi^*) = \pi^* \quad \left(k\epsilon U, 0 \le p(s|\pi^*) \le 1, s\epsilon\mathcal{C}(\pi^*)\right)$$

Sin embargo, el óptimo no necesariamente es único. Es posible tener un conjunto de soluciones óptimas que cubren un borde (arista) o una cara del poliedro, o aún más el poliedro completo (Winston 2005, p. 64-65). Esta última situación ocurre si la función objetivo es una constante. Es aquí donde existe el riesgo de que el algoritmo entre a un ciclo sin fin.

El método simplex es muy eficiente en la práctica y puede garantizarse el hallazgo del óptimo global si se toman ciertas precauciones contra el ciclado, como en el caso anteriormente descrito. Tiene un comportamiento pobre en el peor de los casos: es posible construir un problema de programación lineal para el cual el método requiera un número de pasos exponencial al tamaño del problema. En efecto, hubo un tiempo en el que se desconocía si el problema de programación lineal tenía solución en tiempo polinomial. Un ejemplo de esto es el "Procedimiento balanceado general" que se describe en la sección 4.2. Sin embargo el Método del Cubo en su versión "Algoritmo rápido", presentado en la sección 4.2 también, requiere de un número de pasos que depende linealmente del tamaño de la población.

La Figura 3.1 muestra un ejemplo geométrico del ciclado en el algoritmo. Aquí la población es de tamaño $N = 3$ y el número de variables auxiliares es $p = 2$. La flecha en línea negra continua corresponde al vector de probabilidades de inclusión, las flechas en línea punteada negra a la fase de vuelo, y las flechas en línea punteada verde a la fase de aterrizaje. Las líneas en rojo y azúl corresponden a los hiperplanos de restricción. En este ejemplo la solución óptima cubre un borde del poliedro.

Figura 3.1. Ejemplo geométrico del ciclo sin fin en la fase de aterrizaje.



En ambos casos presentados en la Figura 3.1 se está suponiendo que la función es constante en toda la cara superior. Cada hiperplano colocado al interior del hipercubo corresponde a una variable auxiliar. El ciclado puede apreciarse en las direcciones opuestas de las flechas paralelas horizontales localizadas en la parte superior. Tales flechas horizontales describen un vaivén debido a los valores idénticos en la función objetivo al pasar de una variable auxiliar a otra. Además, debido a la cercanía de ambos hiperplanos a la parte media de la arista, la probabilidad de seleccionar aleatoriamente la dirección hacia al hiperplano compañero es muy alta.

3.4 El problema de las variables auxiliares sin orden

El problema de las variables auxiliares sin orden, consiste básicamente en que al aplicar el método de eliminación de variables se corre el riesgo de eliminar primero algunas de las "más informativas". Provocando con esto que se incremente la varianza y el sesgo del estimador.

Otra desventaja es que no permite optimizar recursos. No permite obtener beneficios del hecho de que unas variables son "más informativas" que otras. Por lo que se corre el riesgo de trabajar con variables demasiado "redundantes".

3.5 Ordenamiento con Componentes Principales Supervisadas

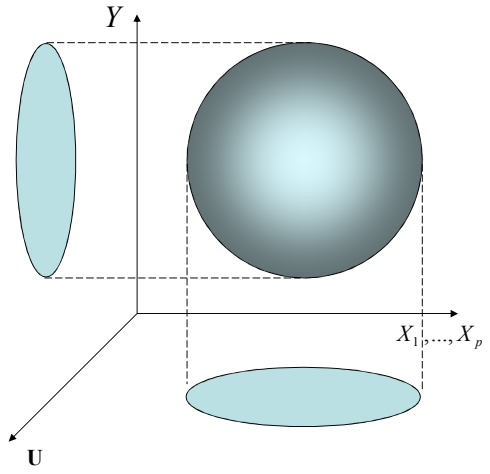
En este método, en vez de hacer un análisis de componentes principales usando "todas" las variables de un conjunto de datos, se usan solo aquellas variables cuyas

"correlaciones" con la variable de interés sean las "más fuertes". Por tanto, primero deben calcularse las correlaciones con la variable de interés y después ordenarlas respecto a algúna medida estadística que mida correlación, sin tomar en cuenta magnitudes en los valores de las variables.

Los supuestos de este método son:

- Como resultado de un estudio previo[5] se posee información de p variables para N individuos de una población. La información está concentrada en una matriz X_{Nxp}.
- Para esos mismos N individuos se posee información de la variable de interés Y.
- Existe alguna relación de la variable de interés Y con una variable latente, no observable, U: $Y = \beta_0 + \beta_1 U + \varepsilon$.
- Cada una de las p variables auxiliares está relacionada con la variable latente U, y se conocen dichas relaciones $X = \alpha_{0j} + \alpha_{1j} U + \varepsilon_j \;\; j = 1, \dots, p$.

Figura 3.2. Supuestos en Componentes Principales Supervisadas.



Los pasos de este procedimiento son:

1. Calcular los coeficientes de regresión estandarizados (estadísticos de log-verosimilitud gaussiana) de cada una de las variables auxiliares (o predictoras) con respecto a la variable de interés.

[5] Censo, encuesta o registro administrativo

$$S_j = \frac{x_j^T y}{\sqrt{x_j^T x_j}};\ j = 1, \dots, p \tag{3.1}$$

Es decir, se mide el efecto univariado de x_j sobre y.

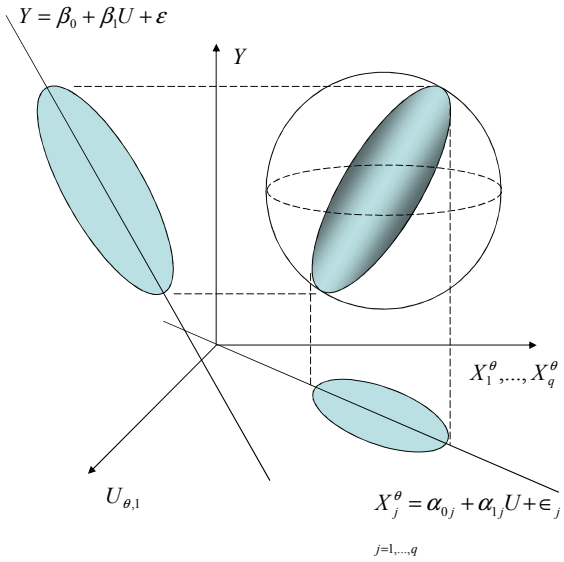
2. Determinar el conjunto de índices (etiquetas) correspondientes a las variables de mayor correlación con la variable de interés.

$$C_\theta = \{j \quad | \quad |S_j| > \theta; j = 1, \dots, p\} \tag{3.2}$$

Para encontrar el valor de θ se utiliza validación cruzada del cociente de verosimilitudes.

3. Definir la matriz reducida X_θ, con las variables de mayor correlación para luego, por medio de descomposición en valores singulares, determinar la matriz de componentes U_θ.
   La componente principal $U_{\theta 1}$ más grande estima la variable latente U.
4. Determinar el modelo de regresión lineal. $\hat{y}^{spc,\theta} = \bar{y} + U_{\theta,1}^T y U_{\theta,1}$
   O bien $\hat{y}^{spc,\theta} = \bar{y} + \hat{\gamma} U_{\theta,1};$ con $\hat{\gamma} = U_{\theta,1}^T y$
   El cual estima a: $Y = \beta_0 + \beta_1 U + \varepsilon$

Figura 3.3. Procedimiento en Componentes Principales Supervisadas.



Para encontrar el criterio de selección θ se utiliza validación cruzada del cociente de verosimilitudes (LR) para estimar el mejor valor. En todas las simulaciones de este trabajo se considera exclusivamente la primera componente, sin embargo existe la posibilidad de usar más de una componente. Para la log-verosimilitud gaussiana, el score del cociente de log-verosimilitudes es equivalente a los coeficientes de regresión estandarizados.

La ecuación de regresión de cada una de las q variables auxiliares respecto a la variable latente U está dada por:

$$X_j = \alpha_{0j} + \alpha_{1j}U + \epsilon_j \tag{3.3}$$

Despejando la variable latente U se obtiene :

$$U = \sum_{j=1}^{q}\left(X_j - \alpha_{0j} - \epsilon_j\right) / \sum_{j=1}^{q}\alpha_{1j} \tag{3.4}$$

Sustituyendo en la expresión de regresión de la variable de interés

$$Y = \beta_0 + \beta_1 U + \varepsilon \tag{3.5}$$

Resulta

$$Y = \beta_0 + \beta_1 \sum_{j=1}^{q}\left(X_j - \alpha_{0j} - \epsilon_j\right) / \sum_{j=1}^{q}\alpha_{1j} + \varepsilon \tag{3.6}$$

Más adelante se ve que los valores de las variables auxiliares, sobre todo las más informativas correspondientes a los aspectos de ingresos y gastos en los hogares, son mucho mayores que los valores absolutos de los coeficientes de regresión. Por lo que al restarles o dividirlos por cantidades pequeñas son dominantes.

Bair *et al.* (2006) comentan que "uno podría iterar el procedimiento de Componentes Principales Supervisadas" y encontrar rasgos que ayuden a encontrar nuevas componentes principales, pero advierten que hacerlo repetidamente converge al método de componentes principales usuales. Por lo que recomiendan hacer iteraciones de este procedimiento sólo si estuviera basado en un criterio que involucrara tanto la varianza de las variables auxiliares como la bondad de ajuste de la variable de interés.

C a p í t u l o I V

ALGORITMOS Y PROCEDIMIENTOS

La realidad es inmutable, no engaña en absoluto y si no ves más allá de las apariencias el engañado eres tú.

Helen Schucman y William Thetford
Meditaciones de un curso de milagros

Como se comenta en la introducción, después de revisar a detalle el método de ordenamiento lo que sigue es hacer un análisis de los algoritmos y procedimientos empleados en el ordenamiento de las variables auxiliares. Los algoritmos y procedimientos aquí explicados son la base para el desarrollo de las simulaciones.

En este capítulo son descritos el procedimiento de la técnica Componentes Principales Supervisadas y el algoritmo para el Método del Cubo. En la sección 4.1 se describe el procedimiento para obtener los coeficientes de regresión estandarizados para el ordenamiento de las variables auxiliares. En la sección 4.2 se describen respecto a la fase de vuelo del Método del Cubo: el procedimiento balanceado general y el algoritmo rápido. Y finalmente en la sección 4.3 se analiza el procedimiento empleado en la técnica Componentes Principales Supervisadas para determinar las variables más informativas.

4.1 Coeficientes de regresión estandarizados

El procedimiento consta de tres pasos:

PROCEDIMIENTO 4.1. Cálculo de los coeficientes de regresión estandarizados

1. Calcular los coeficientes de regresión estándar univariados, llamados scores, para cada variable.
2. Ordenar las variables en forma descendente de acuerdo con los valores absolutos de los scores obtenidos.
3. Formar una matriz cuyas columnas contiene estas variables ordenadas.

A continuación se describe con detalle el primer paso.

a) Estandarizar y centrar los datos de la matriz X

b) Construir la matriz de entrenamiento X_TRAIN para estimar las correlaciones. Construir la matriz de prueba X_TEST para evaluar las estimaciones.
c) Llenar la matriz de prueba con una muestra del 10% de los datos, y la matriz de entrenamiento con el resto de los datos. (ENIGH2002)
d) Obtener los coeficientes de regresión del conjunto de entrenamiento (90% de los datos). Al estar estandarizada y centrada la matriz de datos, estos coeficientes automáticamente son estandarizados. (ENIGH2002)

En el método de validación cruzada, cada modelo es entrenado con el conjunto de entrenamiento (TRAIN) y después el estadístico de prueba cociente de verosimilitudes (LR) es calculado en los datos restantes o conjunto de prueba (TEST). Según Bair *et al.* (2006) "el estadístico cociente de verosimilitudes (LR) del conjunto de prueba es fuertemente significativo".

4.2 Método del Cubo

El siguiente algoritmo permite alcanzar un vértice del polítopo K en al menos N pasos.

ALGORITMO 4.2. Procedimiento balanceado general: fase de vuelo.

INICIALIZAR $\pi(0) = \pi$.
FOR $t = 0, \dots, T$, y hasta que no sea posible realizar el PASO 1 DO

1. Generar cualquier vector, aleatorio o no $\boldsymbol{u}(t) = [u_k(t)] \neq 0$, tal que $\boldsymbol{u}(t)$ está en el kernel de la matriz $\boldsymbol{A}$, y $u_k(t) \neq 0$ si $\pi_k(t)$ es un número entero.
2. Calcular $\lambda_1^*(t)$ y $\lambda_2^*(t)$, los valores más grandes de $\lambda_1(t)$ y $\lambda_2(t)$ tales que $0 \leq \pi(t) + \lambda_1(t)u(t) \leq 1$ y $0 \leq \pi(t) - \lambda_2(t)u(t) \leq 1$
   Note que $\lambda_1(t) > 0$ y $\lambda_2(t) > 0$.
3. Seleccionar

$$\pi(t+1) = \begin{cases} \pi(t) + \lambda_1^*(t)u(t) \text{ con probabilidad } q_1(t) \\ \pi(t) - \lambda_2^*(t)u(t) \text{ con probabilidad } q_2(t) \end{cases}$$

donde $q_1(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$
y $q_2(t) = \lambda_1^*(t)/[\lambda_1^*(t) + \lambda_2^*(t)]$

ENDFOR
FIN

En cada paso, al menos una componente del proceso es redondeada a cero o uno. Así, $\pi(1)$ está en una arista de un cubo de dimensión a lo más $N-1$, $\pi(2)$ está en una arista de un cubo de dimensión a lo más $N-2$ y así sucesivamente.

ALGORITMO 4.3. Algoritmo rápido para la fase de vuelo.

1. *Inicialización*
   a) Eliminar de la población las unidades con probabilidades de inclusión igual a 0 o 1, antes de aplicar el algoritmo, de tal forma que las unidades restantes tengan la propiedad $0 < \pi_k < 1$.
   b) Cargar las probabilidades de inclusión dentro del vector π.
   c) Construir el vector ψ con los primeros $p + 1$ elementos de π.
   d) Crear el vector de rangos $r = (1,2, \dots, p, p + 1)'$.
   e) Construir la matriz $\boldsymbol{B}$ con las primeras $p + 1$ columnas de $\boldsymbol{A}$.
   f) Inicializar $k = p + 2$.
2. *Ciclo básico*
   a) Tomar un vector $\boldsymbol{u}(t)$ del kernel de $\boldsymbol{B}$.
   b) Solamente modificar a ψ (y no al vector π) de acuerdo con la técnica básica. Calcular λ_1^* y λ_2^* los valores más grandes de λ_1 y λ_2 tales que $0 \leq \psi + \lambda_1 u \leq 1$ y $0 \leq \psi - \lambda_2 u \leq 1$
   Note que $\lambda_1^* > 0$ y $\lambda_2^* > 0$.
   c) Seleccionar
   $$\psi = \begin{cases} \psi + \lambda_1^* u \text{ con probabilidad } & q \\ \psi - \lambda_2^* u \text{ con probabilidad } & 1 - q \end{cases}$$
   Donde $q = \lambda_2^*/(\lambda_1^* + \lambda_2^*)$
   d) (Las componentes del vector $\psi(i)$ que corresponden a números enteros son eliminados de $\boldsymbol{B}$ y son reemplazados por nuevas unidades. El algoritmo se detiene al final del archivo.)
   FOR $i = 1,2, \dots, p, p + 1$ do
   IF $\psi(i) = 0$ or $\psi(i) = 1$ THEN
   IF $k \leq N$ THEN
   $\pi\big(r(i)\big) = \psi(i)$;
   $r(i) = k$;
   $\psi(i) = k$;
   FOR $j = 1, \dots, p$,
   DO $\boldsymbol{B}(i,j) = \boldsymbol{A}(k,j)$;
   ENDFOR;
   $k = k + 1$;
   ELSE GOTO STEP 3(a)
   ENDIF;
   ENDIF;
   ENDFOR.
   e) GOTO STEP 2(a)
3. *Fin de la primera parte de la fase de vuelo*
   a) FOR $j = 1, \dots, p$, DO $\boldsymbol{B}(i,j) = \boldsymbol{A}(k,j)$; ENDFOR;

En el algoritmo rápido para la fase de vuelo, la matriz $\boldsymbol{A}$ nunca tiene que ser completamente cargada en memoria, quedando en un archivo que puede ser leído secuencialmente. Por esta razón, no existe ninguna restricción en el tamaño de la

población debido a que el tiempo de ejecución depende linealmente del tamaño de la población.

4.3 Componentes Principales Supervisadas

El procedimiento consta de cinco pasos:

PROCEDIMIENTO 4.4. Determinación de las variables más informativas

---

1. Calcular los coeficientes de regresión estándar univariados, llamados scores, para cada variable.
2. Ordenar las variables en forma descendente de acuerdo con los valores absolutos de los scores obtenidos.
3. Formar una matriz de datos consistente de solo aquellas variables cuyos coeficientes univariados exceden un umbral θ en valor absoluto (θ es estimado por validación cruzada).
4. Calcular la primera componente de la matriz de datos reducida.
5. Usar esa componente principal en un modelo de regresión para predecir la salida.

A continuación se describe con detalle el primer paso.

a) Estandarizar y centrar los datos de la matriz **X.**
b) Construcción de la matriz de prueba X_TEST, y la de entrenamiento X_TRAIN.

$$\mathbf{X} = \begin{bmatrix} X\_TEST \\ X\_TRAIN \end{bmatrix}$$

c) Llenado de las matrices, la de prueba con una muestra del 10% de los datos, llamado conjunto de prueba, y la de entrenamiento con el resto de los datos.
d) Obtener los coeficientes de regresión del conjunto de entrenamiento (90% de los datos), el complemento al de prueba. Al estar estandarizada y centrada la matriz de datos, los coeficientes de regresión obtenidos, automáticamente son estandarizados.

C a p í t u l o V

RESULTADOS DE LAS SIMULACIONES

Todos los caminos que te alejen de lo que eres te llevarán a la confusión y a la desesperación.

Helen Schucman y William Thetford
Meditaciones de un curso de milagros

Como se menciona en la introducción después de haber descrito los algoritmos que se emplean en esta investigación, lo que ahora corresponde es revisar los resultados de las simulaciones al aplicar tales algoritmos a los conjuntos de datos predeterminados para el análisis. Tales conjuntos de datos fueron obtenidos mediante un estudio estadístico. Los resultados que se obtienen de las simulaciones sirven para evaluar el desempeño del método. Son la base de las conclusiones.

Las simulaciones constan de cuatro ejercicios. En el primero se utilizan los métodos tanto de **programación lineal** como **eliminación de variables** ambos en la fase de aterrizaje del Método del Cubo. En los dos primeros (utilizando los datos MU284) la estimación calculada es el total de la variable "Total de ingresos municipales en 1985" (RMT85) usando el estimador Horvitz-Thompson. En los tres últimos se emplea, para seleccionar la muestra balanceada, el método de **eliminación de variables**. En los dos últimos (utilizando los datos ENIGH2002) la estimación calculada es el total de la variable "Ingreso corriente monetario" (INGMON) en el hogar, usando el estimador Horvitz-Thompson.

Para hacer las simulaciones con el Método del Cubo en el conjunto de datos MU284, primero y segundo ejercicios, se seleccionan 50 unidades en cada extracción, que corresponde a un 20 por ciento de muestra aproximadamente.

Una de las preguntas que podría formular el investigador estadístico es la siguiente: para una muestra balanceada con menos de 10 variables auxiliares ¿cuál método es mejor en la fase de aterrizaje, programación lineal o eliminación de variables? Responder esta pregunta es el objetivo del primer ejercicio de simulación.

En el **primer** ejercicio se presenta el resultado de comparar el método de **Programación Lineal contra Eliminación de Variables**, ambos aplicados en la fase de aterrizaje del Método del Cubo a las nueve variables de los datos MU284.

Para el ordenamiento de las variables se emplean los coeficientes de regresión estandarizados obtenidos con la técnica Componentes Principales Supervisadas.

Estas comparaciones se hacen con el objeto de evaluar la pérdida o ganancia en la precisión de las estimaciones al emplear eliminación de variables en vez de programación lineal cuando se consideran diez o menos variables auxiliares. Esto permite comprobar el supuesto que para el caso de menos de diez variables auxiliares siempre es mejor emplear programación lineal.

En el **segundo** ejercicio se presentan las simulaciones efectuadas para los datos MU284 también. Son comparados **varios escenarios de orden**, con el objeto de evaluar el comportamiento del estimador.

En este archivo se utiliza la técnica de coeficientes de regresión estandarizados para ordenar las variables y después se aplica el Método del Cubo a este conjunto ordenado de variables auxiliares, para obtener las muestras y hacer el cálculo del estimador Horvitz-Thompson de la variable de interés. En este caso no es necesario eliminar variables ya que se tienen nueve variables, y la recomendación de aplicar eliminación de variables es para conjuntos de datos con más de diez variables. La diferencia de un escenario a otro es básicamente en cuanto a si las variables más informativas se colocan al principio, en medio o al final del ordenamiento. La otra diferencia es, por un lado, dejar las variables fijas al principio del ordenamiento, para ser las últimas en ser eliminadas o bien, por otro lado, someterlas al criterio de ordenamiento.

En el **tercer** ejercicio, de mayor complejidad y aplicado sobre los datos ENIGH2002, el objetivo es determinar el **número óptimo de variables auxiliares** a ser empleado en el método de eliminación de variables en la fase de aterrizaje del Método del Cubo.

En este ejercicio se emplea la técnica de Componentes Principales Supervisadas. Con esta herramienta se buscan patrones en un subconjunto, llamado "conjunto de

entrenamiento", de los datos de la encuesta ENIGH2002 para hacer pruebas de predicciones con los datos restantes del mismo subconjunto, llamado "conjunto de prueba". Estos patrones son empleados para reducir el número de variables auxiliares, dejando sólo las más informativas (7, 8, 9, 10,11) de acuerdo con el estadístico cociente de verosimilitudes o Likelihood Ratio (LR). El número óptimo de variables auxiliares es determinado haciendo uso de validación cruzada de este estadístico. El rango de variables, de 7 a 11, corresponde a los umbrales en la "meseta" presentada en el estadístico LR (ver Figura 5.5 y Tabla 5.4).

También se hacen simulaciones, en un **cuarto** ejercicio, para los datos de la ENIGH2002, del comportamiento de estas estructuras de correlación en **varias regiones geográficas**. Por simplicidad en este ejercicio se considera a cada una de las 32 entidades de México como una región.

En este caso sólo entran al análisis 9 variables, para evaluar por región el comportamiento, del ordenamiento de variables auxiliares. Las entidades con las cuales se hace la simulación son Nuevo León, Distrito Federal, Jalisco, Puebla y Chiapas.

A continuación se presenta la descripción de los datos, el procedimiento y los resultados de la primera simulación.

5.1 Programación Lineal contra Eliminación de Variables

El conjunto de datos a emplearse es MU284, presentado en el *Apéndice A* al final de este documento. Los datos corresponden a información socioeconómica de Suecia. El archivo contiene 284 observaciones de once variables, más un renglón con los nombres de las variables.

En Särndal *et al.* (1992) se menciona que Suecia estaba dividida en 284 municipalidades por cuestiones administrativas. Algunas variables fueron seleccionadas de tal forma que describieran a las municipalidades en diferentes formas. El conjunto de datos da la oportunidad al lector de llevar a cabo sus propios experimentos en muestreo y estimación. Este archivo incluye los tres municipios más grandes de acuerdo con el valor de la variable P75, correspondientes a Estocolmo, Gotemburgo y Malmö, lo que provoca que la distribución de las variables llegue a tener una asimetría

mayor y mayor varianza. Los datos también pueden obtenerse del sitio de internet StatLib (http://lib.stat.cmu.edu/mu284/).

En este sitio se menciona que han sido corregidas cuatro cifras de la primera impresión del libro:

Registro 107, en ME84 dice 1110, debe decir 1100
Registro 141, en RMT85 dice 369, debe decir 396
Registro 220, en ME84 dice 491, debe decir 461
Registro 229, en ME84 dice 1238, debe decir 1239

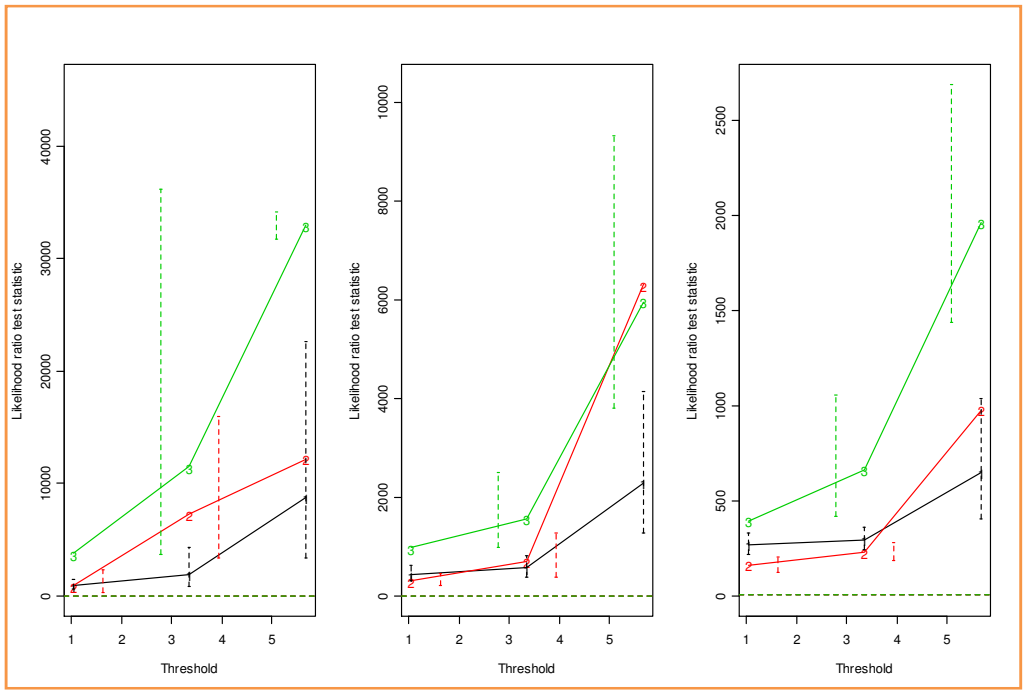
La descripción de las once variables es la siguiente:

1	LABEL	Número de registro
2	P85	Población en 1985
3	P75	Población en 1975
4	RMT85	Total de ingresos municipales en 1985
5	CS82	Número de diputados conservadores en 1982
6	SS82	Número de diputados demócratas en 1982
7	S82	Número de diputados en el Consejo Municipal en 1982
8	ME84	Número de empleados en el municipio en 1984
9	REV84	Valores del estado de acuerdo con el avalúo de 1984
10	REG	Indicador de la región geográfica
11	CL	Conglomerados (grupos de municipios vecinos)

La variable a estimar es RMT85, y como la variable LABEL corresponde al número de registro, entonces estas dos variables salen de la tabla de información auxiliar, la cual se reduce a sólo nueve. De estas nueve variables, CL y REG son fijas, y las otras siete son no fijas.

En un primer ejercicio se compara el método de programación lineal con eliminación de variables para el caso de menos de diez variables auxiliares. Para el cual, una muestra del 20% es seleccionada del conjunto de datos, es decir 50 registros aproximadamente. Aunque no se ocupa en este primer ejercicio, se presentan las gráficas del cociente de verosimilitudes para analizar su comportamiento. Una validación cruzada de 10 dobleces es aplicada al cociente de verosimilitudes como lo recomiendan Hastie *et al.* (2001) p. 215. El cociente de verosimilitudes para dobleces: 2, 5, 10, y para un umbral fijo de 3 se encuentra representado en la Figura 5.1.

Figura 5.1. Cociente de verosimilitudes para dobleces 2, 5 y 10, y un umbral fijo de 3 en todos lo casos.

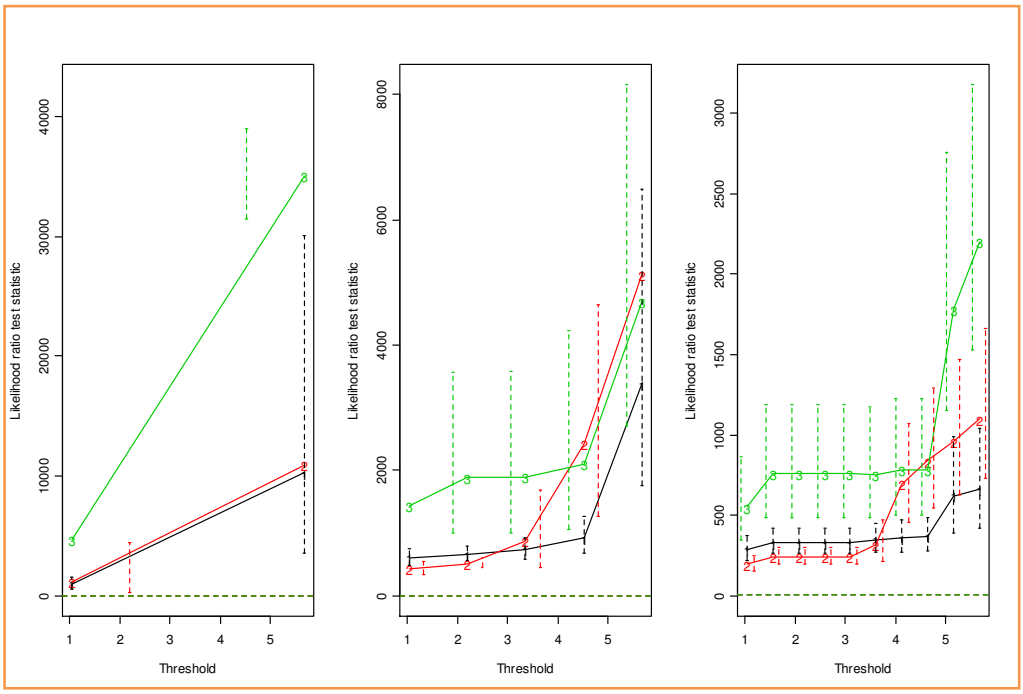


En todas las gráficas del cociente de verosimilitudes, sólo se grafican las primeras tres componentes principales, presentando los números 1, 2 o 3 según corresponda. De éstas, la primera componente es la que mayor varianza de los datos explica.

En la Figura 5.1 puede apreciarse poca diferencia entre una cantidad de dobleces y otra. En las tres opciones la componente 3 es la que mejor explica la variabilidad de los datos.

El cociente de verosimilitudes para dobleces 2, 5, 10 y para umbrales 2, 5, 10 respectivamente proporciona mayor 'resolución' como se muestra en la Figura 5.2.

Figura 5.2. Cociente de verosimilitudes para dobleces 2, 5 y 10, y para umbrales 2, 5 y 10 respectivamente.



Una mayor finura en el comportamiento del cociente de verosimilitudes se puede apreciar usando la opción de 10 umbrales y 10 dobleces. Por esta razón se optó por emplear 10 dobleces y 30 umbrales como puede apreciarse en el ejercicio de simulación correspondiente a "varios escenarios de ordenamiento".

En este primer análisis se hacen tres mil simulaciones. En el cálculo de las correlaciones entre las variables auxiliares y la variable de interés se supone que los datos se distribuyen como una normal. Para evaluar la distribución de los errores de estimación se emplea un diseño de muestreo balanceado para la obtención de cada muestra. El método de selección de elementos se hace con probabilidad proporcional al tamaño (PPT) respecto a la variable P75. El tamaño de muestra en cada extracción es de cincuenta registros, que corresponde a un veinte por ciento de la población aproximadamente, y se estima el total de la variable de interés RMT85 utilizando el estimador Horvitz-Thompson. Con las estimaciones resultantes se calcula la diferencia relativa respecto al total de la variable de interés y estos errores relativos son graficados para cada una de las seis configuraciones obtenidas al considerar: las variables en orden o desorden, combinadas con el número de variables 9 o 6 y a su vez combinadas con los métodos programación lineal o eliminación de variables.

En este ejercicio de simulación se puede comparar la magnitud del sesgo, introducido al aplicar el método de eliminación de variables auxiliares en la fase de aterrizaje, con el obtenido al aplicar el método de programación lineal. También se puede establecer la pérdida en precisión al usar este método. En todos los demás ejercicios se proporciona el sesgo relativo para comparar y evaluar la magnitud del sesgo de las estimaciones en cada escenario o estrategia.

Para comparar cuantitativamente la eficiencia de las distintas estrategias de balanceo, en este y en los demás ejercicios, son calculados algunos estadísticos del error de las estimaciones: la raíz cuadrada del Error Cuadrático Medio relativo (ECM_{rel}), el Sesgo relativo (B_{rel}) y la Desviación Estándar relativa ($D.E._{rel}$).

$$\sqrt{ECM_{rel}(\hat{t})} = \sqrt{\frac{ECM(\hat{t})}{t^2}} = \sqrt{\frac{\frac{1}{M}\sum_{j=1}^{M}\left(\hat{t}_j - t\right)^2}{t^2}} \tag{5.1}$$

$$B_{rel}(\hat{t}) = \frac{B(\hat{t})}{t} = \frac{\bar{\hat{t}} - t}{t} = \frac{\frac{1}{M}\sum_{j=1}^{M}\hat{t}_j - \frac{1}{M}\sum_{j=1}^{M}t}{t} = \frac{1}{M}\sum_{j=1}^{M}\frac{\hat{t}_j - t}{t} \tag{5.2}$$

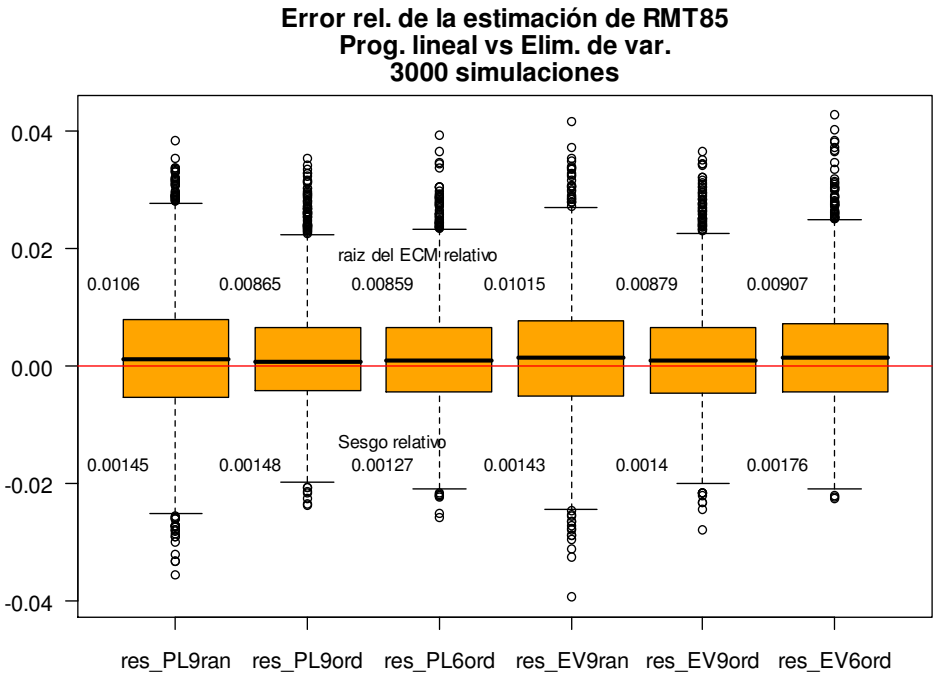
$$D.E._{rel} = \frac{D.E.\left(\hat{t}_j\right)}{t} = \sqrt{\frac{Var\left(\hat{t}_j\right)}{t^2}} = \sqrt{\frac{\frac{1}{M}\sum_{j=1}^{M}\left(\hat{t}_j - \bar{\hat{t}}\right)^2}{t^2}} \tag{5.3}$$

$$ECM(\hat{t}) = Var(\hat{t}) + B^2(\hat{t}) \tag{5.4}$$

De la Figura 5.3 se puede apreciar que la raíz del ECM relativo tiene un rango de variación entre $8.6x10^{-03}$ y $10.6x10^{-03}$. La D.E.[6] relativa es similar a este estadístico, debido a que el sesgo relativo es muy cercano a cero y oscila entre $12.7x10^{-04}$ y $17.6x10^{-04}$.

[6] La desviación estándar (D.E.) también es conocida como Error Estándar o Error Típico.

Figura 5.3. Programación lineal contra Eliminación de variables.



Que el sesgo relativo tenga un rango de variación muy cercano a cero implica que para el caso de menos de diez variables auxiliares, tanto el método empleado como la cantidad de variables utilizadas sólo influyen en la varianza de las estimaciones y no así en el sesgo, conservándolo prácticamente imperceptible. De acuerdo a la gráfica se puede obtener como conclusión que las estrategias que mayor precisión ofrecen en las estimaciones, listadas en orden de mejor a peor y que tuvieron el menor sesgo relativo, son las dos siguientes:

- *EV9ord*. Eliminación de Variables aplicada a las 9 variables ordenadas descendentemente de acuerdo al valor del score.
- *PL6ord*. Programación Lineal aplicada a las 6 variables auxiliares con los scores más altos, ordenadas descendentemente.

Tabla 5.1. Estadísticos del error para cada estrategia.

Estadísticos del error	*Estrategia*					
	PL9ran	*PL9ord*	*PL6ord*	*EV9ran*	*EV9ord*	*EV6ord*
Raíz del ECM relativo	106	87	86	102	88	91
Sesgo relativo	14.5	14.8	12.7	14.3	14.0	17.6
D.E. relativa	106	87	86	102	88	91

Los valores están expresados en magnitudes de $1x10^{-4}$.

Aún para el método de Programación Lineal (PL) es recomendable el ordenamiento de las variables auxiliares como se puede apreciar en la comparación de las dos estrategias presentada en la Figura 5.3.

El método de EV se ve afectado por la limitación en el número de variables auxiliares. Por ejemplo al aplicar la estrategia de eliminación de variables a un conjunto de solo 6 variables resultó ser la de mayor sesgo relativo. Por otro lado, aplicar el método PL a sólo 6 variables ordenadas, produjo el menor sesgo. Por lo que se concluye que incluso para conjuntos con menos de 10 variables auxiliares conviene ordenarlas de acuerdo a su score, aunque no se utilice el método de eliminación de variables.

Ahora se analiza la ventaja de esta técnica de ordenamiento al hacer simulaciones con información de nuestro país.

Para el caso de México existe la posibilidad de aplicar el método a los datos obtenidos del II Conteo de Población y Vivienda 2005 o los que se tiene planeado obtener en el evento censal del año 2010, aplicar el método específicamente en el diseño que considera como dominio a la Entidad, como estrato al Tamaño de Localidad y como conglomerado al Municipio. En tales unidades geográficas hay una mayor estabilidad de las variables "eje" independientes: Habitantes, Viviendas y

Grupos de Edad. En el desglose geográfico a nivel AGEB[7] o UPM[8], esto ya no podría aplicarse para obtener resultados informativos, debido al gran dinamismo de este último en cuanto a unidades habitacionales desaparecidas y de nueva creación se refiere. A este nivel de desglose las manzanas se crean, se fusionan o cambian de unidad político administrativa con mayor dinamismo e incluso llegan a desaparecer provocando un cambio abrupto en los valores de las variables "eje" independientes.

5.2 Varios escenarios de orden

En el análisis de varios escenarios de orden también se hacen tres mil simulaciones. En el cálculo de las correlaciones entre las variables auxiliares y la variable de interés se supone que los datos se distribuyen como una normal. Para evaluar la distribución de los errores de estimación se emplea un diseño de muestreo balanceado para la obtención de cada muestra. La selección de elementos se hace con probabilidad proporcional al tamaño (PPT) respecto a la variable P75. El tamaño de muestra, en cada extracción, es de cincuenta registros, que corresponde a un veinte por ciento de la población aproximadamente, y se estima el total la variable de interés RMT85 utilizando el estimador Horvitz-Thompson. Con los resultados de las estimaciones se calcula la diferencia relativa respecto al total de la variable de interés y estos errores relativos son graficados para cada una de los cinco escenarios de orden: Särndal, peor, regular, mejor y descendente.

Tomando en cuenta que para evaluar las ventajas de ordenar las variables en la fase de aterrizaje se deben efectuar pruebas con varios escenarios, en este ejercicio se realizan pruebas con cinco configuraciones de orden. Primero tal y como se presentan las variables en el libro de Särndal *et al.* (1992). Segundo, el peor escenario, en el cual los scores son ordenados ascendentemente, para que las variables "más importantes" queden al final y sean las primeras en ser eliminadas. Tercero, un escenario intermedio o regular, en el cual las variables "más importantes" son colocadas a la mitad del

[7] Área Geoestadística Básica. Es la unidad básica del Marco Geoestadístico Nacional. Generalmente consta de menos de 50 manzanas en el área urbana (AGEB urbano), y es una porción geográfica de aproximadamente 10,000 hectáreas en el área rural (AGEB rural). El AGEB rural puede incluir localidades no urbanas que van desde caseríos dispersos hasta localidades con menos de 2,500 habitantes.

[8] Unidad Primaria de Muestreo. Es un área geográfica formada por una o más manzanas, una AGEB o la unión de dos o más AGEB, generalmente deben ser colindantes con un peso mínimo de 240 viviendas y un máximo de 480 viviendas.

ordenamiento. Cuarto escenario, en este caso los scores son ordenados en forma descendente. En este escenario, las variables más importantes son eliminadas hasta el final del procedimiento, por ejemplo cuando el problema del ciclado persiste. Y finalmente en el quinto escenario, el ordenamiento es efectuado en forma descendente, respetando todos los scores de las variables auxiliares (ver Tabla 5.2).

En los escenarios del segundo al cuarto ("regular", "mejor" y "descendente") las variables CL y REG aparecen en los primeros lugares y son eliminadas al final del procedimiento en caso necesario, esto debido a que para las variables indicadoras el balanceo puede ser exacto. Es decir, con las variables indicadoras o categóricas no existe el problema del redondeo.

Tabla 5.2. Scores para los escenarios de orden en los datos MU284.

Escenario de orden	*Variables y scores*								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Särndal	P85	P75	CS82	SS82	S82	ME84	REV84	REG	CL
	14.78	15.18	5.27	3.89	5.99	19.58	13.47	1.04	1.11
Peor	CL	REG	SS82	CS82	S82	REV84	P85	P75	ME84
	1.11	1.04	3.89	5.27	5.99	13.47	14.78	15.18	19.58
Regular	CL	REG	SS82	S82	P85	ME84	P75	REV84	CS82
	1.11	1.04	3.89	5.99	14.78	19.58	15.18	13.47	5.27
Mejor	CL	REG	ME84	P75	P85	REV84	S82	CS82	SS82
	1.11	1.04	19.58	15.18	14.78	13.47	5.99	5.27	3.89
Descendente	ME84	P75	P85	REV84	S82	CS82	SS82	CL	REG
	19.58	15.18	14.78	13.47	5.99	5.27	3.89	1.11	1.04

Nota: Los scores fueron obtenidos considerando un conjunto de entrenamiento del 90% de los datos.

Los scores obtenidos reflejan que las variables CL y REG no son tan "informativas" como era de esperarse. Por lo que pueden dejarse al final del archivo para ser las primeras en eliminarse.

Un ejemplo donde las variables menos informativas son eliminadas primero es el escenario denominado "Särndal", el cual resultó ser el mejor de todos, tanto en Error Cuadrático Medio relativo (ECM_{rel}) como en Sesgo relativo (B_{rel}).

Figura 5.4. Varios escenarios de ordenamiento para los datos MU284.

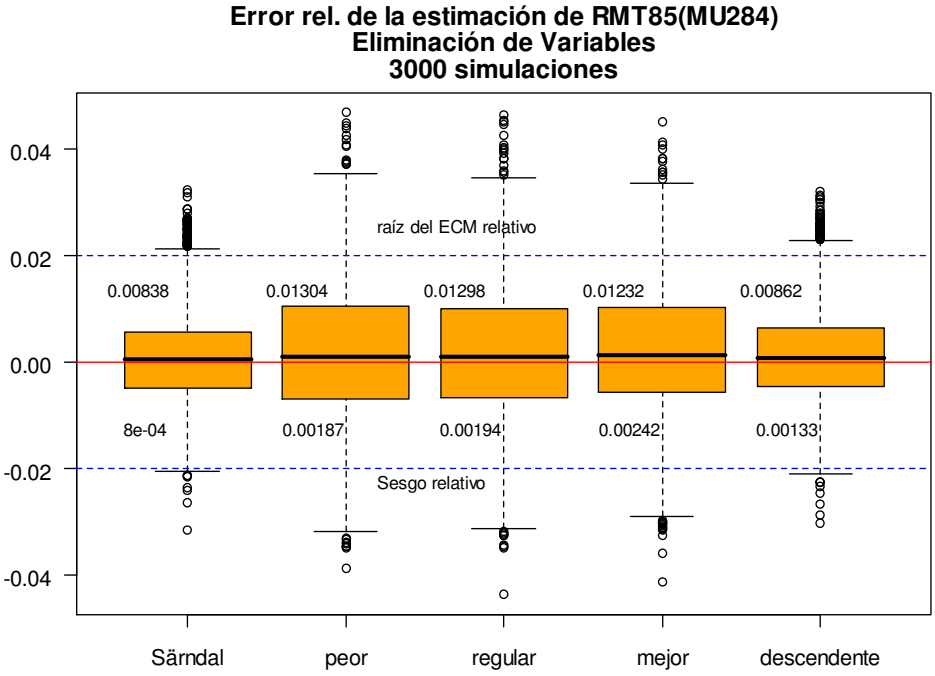


Tabla 5.3. Estadísticos del error en cada escenario de orden.

Estadísticos del error	*Escenario de orden*				
	Särndal	*peor*	*regular*	*mejor*	*descendente*
Raíz del ECM relativo	84	130	130	123	86
Sesgo relativo	8	19	19	24	13
D. E. relativa	84	130	130	123	86

Los valores están expresados en magnitudes de $1x10^{-4}$.

Contrario a lo esperado, los escenarios en los cuales las variables indicadores aparecen al final del ordenamiento, para ser las primeras en eliminarse, fueron los mejores en las estimaciones. Es decir, los mejores escenarios fueron dos, primero el ordenamiento tal y como están presentes las variables en la publicación de Särndal *et al.* (1992), y segundo, el ordenamiento completamente descendente, tal y como lo establece la hipótesis planteada en este trabajo.

El ECM varía de $84x10^{-4}$ para el escenario "Särndal" hasta $130x10^{-4}$ para el escenario "peor". El sesgo relativo tiene un rango de variación entre $8x10^{-4}$ y $24x10^{-4}$. La D.E. relativa va desde $84x10^{-4}$ para el escenario "Särndal" hasta el escenario "peor" con un valor de $130x10^{-4}$.

Respecto a la interpretación del ECM se debe tener en cuenta que entre los estimadores no sesgados, el ECM mínimo es equivalente a minimizar la varianza. También se debe considerar que un estimador con **sesgo** puede tener un ECM más bajo. Además un modelo no sesgado con un ECM más pequeño generalmente se interpreta como el que mejor explica la variabilidad de las observaciones.

5.3 Determinación del número óptimo de variables auxiliares

Los datos de la Encuesta Nacional de Ingresos y Gastos de los Hogares 2002 (ENIGH2002) forma parte de un grupo de encuestas efectuadas en los años 2000, 2002, 2004, 2005 2006, 2007 y 2008. Es decir cada dos años en los primeros tres estudios y cada año a partir del tercer evento. Al momento de desarrollar el presente trabajo las encuestas armonizadas de acuerdo a la conciliación demográfica, fueron las primeras cuatro, una razón de peso al momento de elegir los datos sobre los cuales hacer las simulaciones.

El Instituto Nacional de Estadística y Geografía (INEGI), en un esfuerzo por fortalecer el servicio público de información estadística, presentó estas tablas de datos, "Armonizadas de acuerdo con la Conciliación Demográfica", con la intención de dar respuesta a los requerimientos de los usuarios especializados, con un interés particular en el análisis de los microdatos, que permiten un conocimiento más detallado del monto, la estructura y la distribución de los ingresos de los hogares y del destino de los gastos del hogar en bienes de consumo duradero y no duradero. En esta encuesta

también se obtiene información sobre la infraestructura de las viviendas, la composición familiar de los hogares, así como de la actividad económica de cada uno de sus miembros.

La encuesta proporciona información a nivel nacional con desglose urbano-rural, es decir también proporciona información para el conjunto de localidades de 2 500 y más habitantes, y para aquellas con menos de 2 500 habitantes.

Las cifras de las encuestas de los años 2000, 2002, 2004 y 2005 se sometieron a un proceso de armonización acorde con las cifras de la Conciliación Demográfica realizada conjuntamente por el Consejo Nacional de Población (CONAPO), El Colegio de México y el INEGI.

A la par del ejercicio de armonización, se llevó a cabo una revisión de la información captada por estas cuatro encuestas, con el propósito de uniformarlas. Este ejercicio permitió homologar las bases de datos en conjunto para hacerla comparable, y no limitar el análisis de alguna de ellas en forma aislada y sin conexión. Abriendo la posibilidad de analizar la evolución de los ingresos y de los gastos de los hogares mexicanos en ese periodo.

Las bases de datos están conformadas por diez archivos. Siete archivos de datos en formato DBF: "hogares", "población", "ingresos", "gastos", "eroga", "nomon" y "concen". Y tres de catálogos en formato PDF: "catalogo", "describe" y "variable".

Para acceder a los datos de la encuesta ENIGH, sólo tienen que descargarse estos archivos en alguna carpeta de su equipo y desempacarlo.

En los dos ejercicios siguientes los hogares contenidos en el conjunto de información no se consideran como provenientes de una muestra sino como una población completa.

En el Apéndice C se presenta la descripción de 114 variables correspondientes al archivo 'hogares.dbf' para la ENIGH2002. Este archivo consta de 17,167 registros y se puede obtener fácilmente del sitio web del INEGI[9].

Las diferencias de los dos ejercicios siguientes con los anteriores, respecto a la información utilizada, son principalmente que la cantidad de variables es mucho mayor y que la cantidad de datos también es mucho mayor, y por ende la diversidad de correlaciones entre las variables auxiliares de la población es más completa. Esta última diferencia es un factor preponderante para definir estrategias de investigación estadística en conjuntos de información auxiliar de mediana magnitud.

Tabla 5.4. El cociente de verosimilitudes (LR) y sus scores.

Umbral	*Estadístico LR*	*Score*	*Umbral*	*Estadístico LR*	*Score*
1	621.9786	0.1009125	**16**	676.5450	9.1267274
2	634.3279	0.7026335	**17**	658.4860	9.7284484
3	635.6534	1.3043545	18	683.8066	10.3301694
4	661.7233	1.9060755	**19**	746.6461	10.9318904
5	685.9612	2.5077965	**20**	861.0961	11.5336114
6	696.7138	3.1095175	**21**	943.0255	12.1353324
7	714.4670	3.7112385	**22**	996.8644	12.7370534
8	704.0117	4.3129595	**23**	986.9414	13.3387744
9	720.7627	4.9146805	**24**	1058.3436	13.9404954
10	718.5627	5.5164015	**25**	1004.6325	14.5422163
11	724.3825	6.1181225	**26**	886.3695	15.1439373
12	728.3857	6.7198435	**27**	863.8200	15.7456583
13	713.5608	7.3215644	**28**	889.1919	16.3473793
14	704.0304	7.9232854	**29**	934.6093	16.9491003
15	700.3450	8.5250064	**30**	934.6093	17.5508213

La Tabla 5.4 muestra el umbral óptimo a través del estadístico cociente de verosimilitudes (LR) y su score correspondiente. El score óptimo es 13.94 correspondiente al cociente de verosimilitudes 1,058.34. En otras palabras el método recomienda seleccionar, para construir el conjunto reducido, aquellas variables cuyo score supere el umbral θ=13.94

Hacer la validación cruzada con una cantidad de umbrales mayor es con el objeto de lograr mayor resolución en la detección del score correspondiente al estadístico LR óptimo. En la Tabla 5.4 se puede apreciar que el estadístico LR óptimo (el de máximo

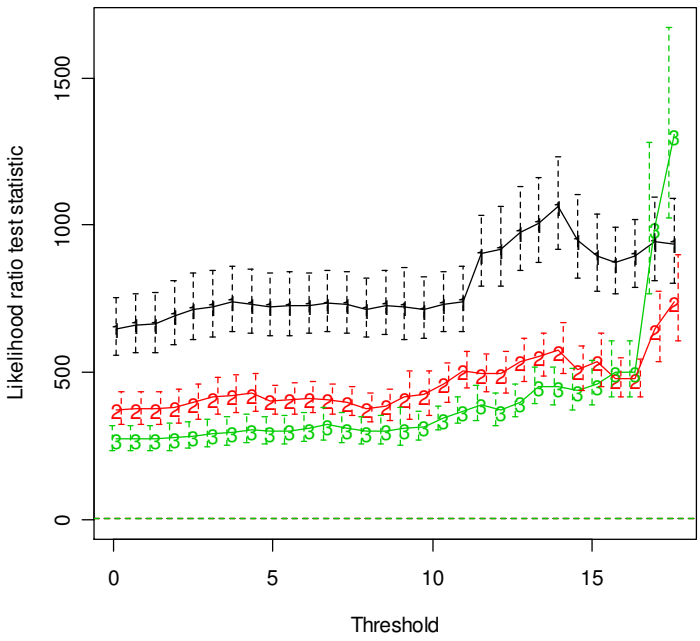
[9] http://www.inegi.org.mx/est/contenidos/espanol/sistemas/enigh/bd/default.asp

valor) se encuentra entre los umbrales 23 y 25, y una estimación de ese óptimo es el obtenido en el umbral 24 que corresponde a un score de 13.9405, por lo que la cantidad óptima de variables auxiliares recomendadas por la técnica componentes principales supervisadas son todas aquellas cuyo score supere ese valor.

La Figura 5.5 muestra la curva de validación cruzada para estimar el mejor umbral. Cada modelo es entrenado y se calcula el cociente de verosimilitudes en el conjunto de datos que quedó fuera del entrenamiento, es decir en el conjunto de prueba.

En la gráfica de la primera ("1") componente se puede apreciar una "meseta" situada entre los umbrales 10 y 15. En las gráficas de las componentes segunda ("2") y tercera ("3") no es posible distinguir claramente dicha característica por ser muy tenue la elevación. Al considerar, numéricamente, el óptimo de esa primera componente damos respuesta a la pregunta ¿cuál es el número óptimo de variables auxiliares? Dicha respuesta está limitada a la consideración de sólo la primera componente.

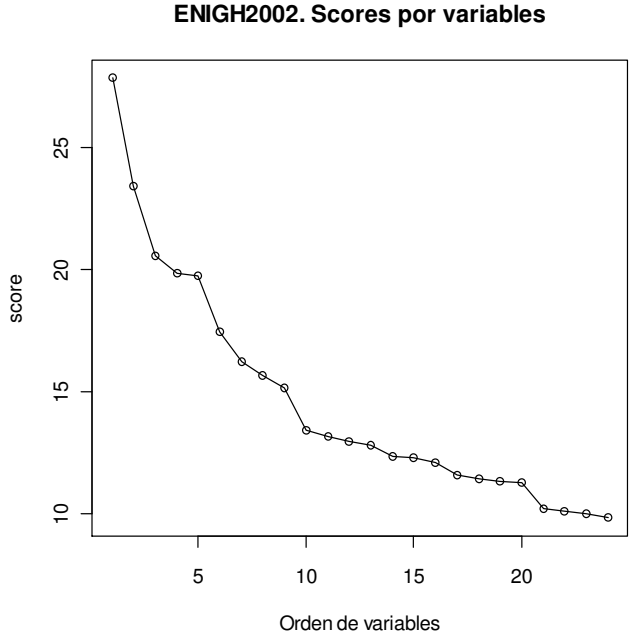
Figura 5.5. Curva de validación cruzada para las componentes primera ("1"), segunda ("2") y tercera ("3")



El cociente de verosimilitudes del conjunto de prueba es fuertemente significativo. Este ejemplo también ilustra que el procedimiento puede ser sensible al valor del umbral.

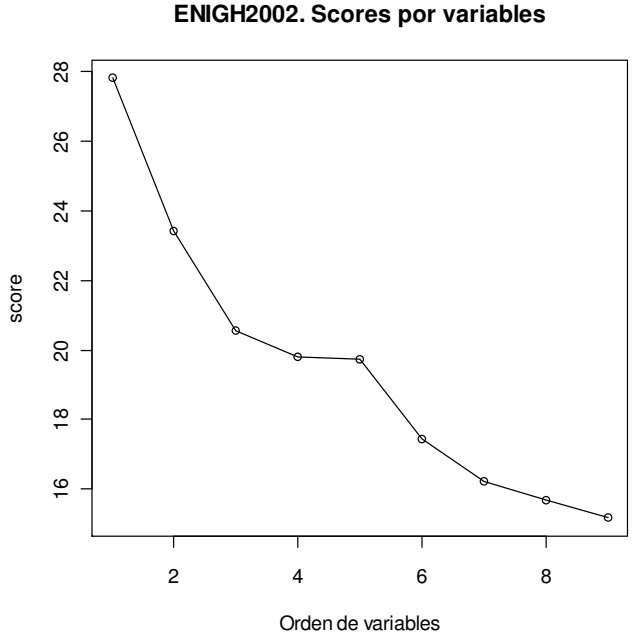
Para las 24 variables más informativas se prueba con 30 umbrales y se grafica el cociente de verosimilitudes para las primeras tres componentes, sin embargo en este trabajo únicamente se toma en cuenta sólo la primera componente para determinar el modelo de regresión.

Figura 5.6. Gráfica de los 24 scores ordenados descendentemente.



El comportamiento de los scores, al graficarlos en forma descendente, coincide con el comportamiento de los scores para las componentes principales. En las cuales regularmente se forma un "codo" en los cambios de los scores, primero los que disminuyen en forma rápida y luego los que lo hacen lentamente.

Figura 5.7. Gráfica de los 9 scores más altos.



De la Tabla 5.5 se puede apreciar que aquellas variables que superan el score de 13.9405 son nueve, que citadas junto con su nombre corto en orden decreciente son: "Ingreso corriente total" (ingcor), "Ingreso total" (ingtot), "Gasto corriente monetario" (gasmon), "Gasto corriente total" (gascor), "Gasto total" (gastot), "Artículos y servicios para la limpieza y cuidados de la casa, enseres domésticos, muebles, cristalería, utensilios domésticos y blancos" (limpieza), "Artículos y servicios para la limpieza y cuidados de la casa" (cuidados), "Remuneraciones al trabajo asalariado" (trabajo) y "Sueldos y horas extras" (sueldos).

Tabla 5.5. Orden de las 24 variables "más importantes".

Orden	*Variable*	*Score*	*Orden*	*Variable*	*Score*
1	ingcor	27.82	**13**	cuidado	12.81
2	ingtot	23.43	**14**	comunica	12.38
3	gasmon	20.57	**15**	energia	12.30
4	gascor	19.82	**16**	esparci	12.10
5	gastot	19.75	**17**	vestido_c	11.61
6	limpieza	17.44	**18**	ves_3ymas	11.42
7	cuidados	16.23	**19**	vestido	11.32
8	trabajo	15.69	**20**	transporte	11.31
9	sueldos	15.17	**21**	educa	10.23
10	alimentos	13.43	**22**	transfe	10.12
11	personal	13.15	**23**	pago_tar	10.03
12	Educacion	12.99	**24**	negocio	9.84

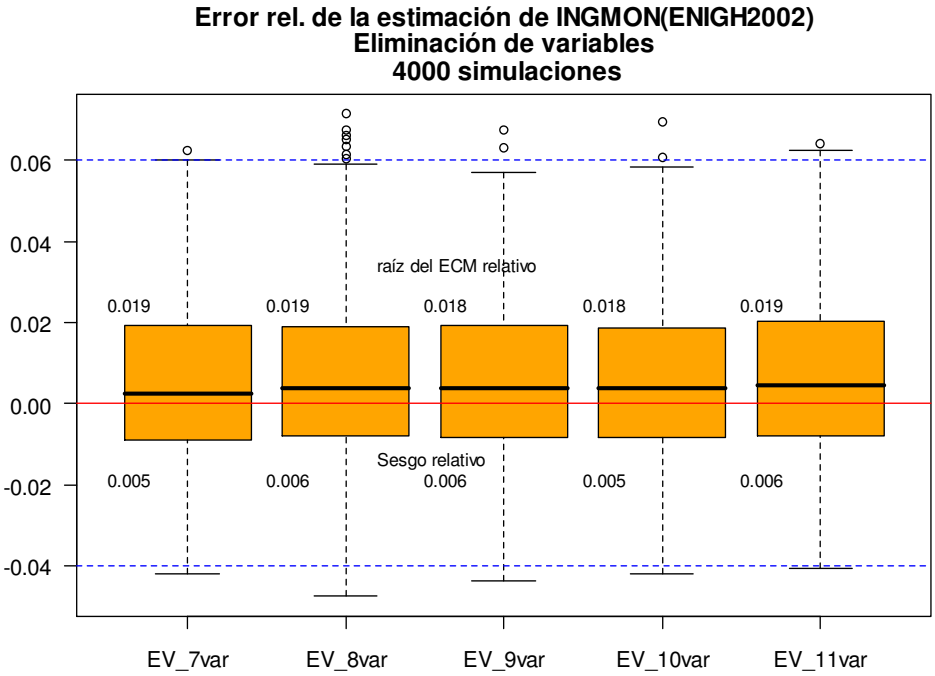
Para una descripción de esta y otras variables ver la Tabla C2 del Apéndice C. Recordar que en las simulaciones la variable de interés es "Ingreso corriente monetario" (ingmon).

Después de haber aplicado el método de ordenamiento estadístico al conjunto de variables de la ENIGH2002, ahora corresponde describir las características de las simulaciones que permiten evaluar el número óptimo de variables auxiliares.

Para evaluar la determinación del número óptimo de variables auxiliares son efectuadas cuatro mil simulaciones. En el cálculo de las correlaciones entre las variables auxiliares y la variable de interés se supone que los datos se distribuyen como una normal. Para evaluar la distribución de los errores de estimación se emplea un diseño de muestreo balanceado en la obtención de cada muestra. La selección de elementos se hace con probabilidad proporcional al tamaño (PPT) respecto a la variable "Gasto total" (gastot). El tamaño de muestra en cada extracción es de cincuenta hogares. Y se estima el total la variable de interés "Ingreso corriente monetario" (ingmon) utilizando el estimador Horvitz-Thompson. Con los resultados de las estimaciones se calcula la diferencia relativa respecto al total de la variable de interés y estos errores relativos son graficados para cada uno de los cinco conjuntos de variables auxiliares, desde el que contiene siete hasta el que contiene once.

Los resultados obtenidos con estas simulaciones se presentan en la Figura 5.8. En esa gráfica se puede apreciar que es casi imperceptible la diferencia en el error de estimación para los conjuntos con diferente cantidad de variables.

Figura 5.8. Número óptimo de variables auxiliares.



Aunque se aumente el número de variables en el conjunto de información auxiliar el método las elimina en caso necesario, y deja sólo la cantidad suficiente para obtener la muestra. En la gráfica se observa que el sesgo es prácticamente el mismo para los conjuntos que constan de 7 a 11 variables.

Tabla 5.6. Estadísticos del error para diferentes conjuntos.

Estadísticos del error	*Cantidad de variables*				
	7	*8*	*9*	*10*	*11*
Raíz del ECM relativo	19	19	18	18	19
Sesgo relativo	5	6	6	5	6
D. E. Relativa	19	19	18	18	19

Los valores están expresados en magnitudes de $1x10^{-3}$.

Los resultados de estas simulaciones muestran que para el caso del conjunto de datos ENIGH2002, el menor valor de la raíz del ECM relativo se obtiene con 9 y 10

variables, así mismo el sesgo mayor se obtiene con 8, 9 y 11 variables auxiliares. Es decir, a mayor número de variables auxiliares, después del óptimo, se presenta una mayor redundancia en las variables auxiliares, como es el caso del conjunto con once variables, el cual arroja un sesgo y un error mayores que los presentados por el conjunto con diez variables. Además en todos los cinco casos el sesgo relativo es menor a uno por ciento. El sesgo es prácticamente igual en todos los casos; sin embargo el ECM relativo es ligeramente menor para el caso de 9 y 10 variables auxiliares.

La raíz del ECM relativo tiene un rango de variación entre $18x10^{-3}$ y $19x10^{-3}$. En cambio el sesgo relativo tiene mucha estabilidad y varía entre $5x10^{-3}$ y $6x10^{-3}$. La D.E. relativa presenta un comportamiento igual al de la raíz del ECM relativo.

5.4 Estimaciones en varias regiones geográficas

Después de haber analizado el comportamiento del método de ordenamiento en varios aspectos, por ejemplo al comparar el método de Programación Lineal contra Eliminación de Variables, al comparar varios escenarios de orden y al determinar el número óptimo de variables, ahora es conveniente analizar si tal ordenamiento y tal cantidad de variables auxiliares funcionan de manera estable para diferentes regiones geográficas de un país.

Para evaluar el comportamiento estable en varias regiones geográficas del número óptimo de variables auxiliares ordenadas estadísticamente se efectúan cinco mil simulaciones. Los cálculos de las correlaciones entre las variables auxiliares y la variable de interés son bajo el supuesto de que los datos se distribuyen como una normal. Para evaluar la distribución de los errores de estimación se emplea un diseño de muestreo balanceado para la obtención de cada muestra. La selección de elementos se hace con probabilidad proporcional al tamaño (PPT) respecto a la variable "Gasto total" (gastot). El tamaño de muestra en cada extracción es de cincuenta registros, no importando que los tamaños de las poblaciones por región sean diferentes, y se estima el total de la variable de interés INGMON utilizando el estimador Horvitz-Thompson. Con los resultados de las estimaciones se calcula la diferencia relativa respecto al total

de la variable de interés y estos errores relativos son graficados para cada una de las cinco regiones: Chiapas, Distrito Federal, Jalisco, Nuevo León y Puebla.

Tabla 5.7. Tamaños de muestra para las entidades.

Clave	*Abreviación*	*Nombre*	*Registros*
07	CHS	Chiapas	640
09	DF	Disitrito Federal	1,232
14	JAL	Jalisco	541
19	NLN	Nuevo León	420
21	PUE	Puebla	580

El mejor resultado se obtiene siempre con las variables ordenadas. Al aumentar el número de variables auxiliares, la ganancia en precisión es poca, en cambio el sesgo aumenta, como se puede apreciar con la media de las estimaciones.

Con el método de eliminación de variables auxiliares pareciera que es inevitable el sesgo, como lo muestra claramente la gráfica de los cuatro escenarios donde todos tienen un sesgo apreciable, a diferencia del método de programación lineal donde el sesgo es prácticamente cero como puede apreciarse en la gráfica del resultado de las simulaciones del primer ejercicio.

Figura 5.9. Errores de estimación en regiones geográficas distintas.

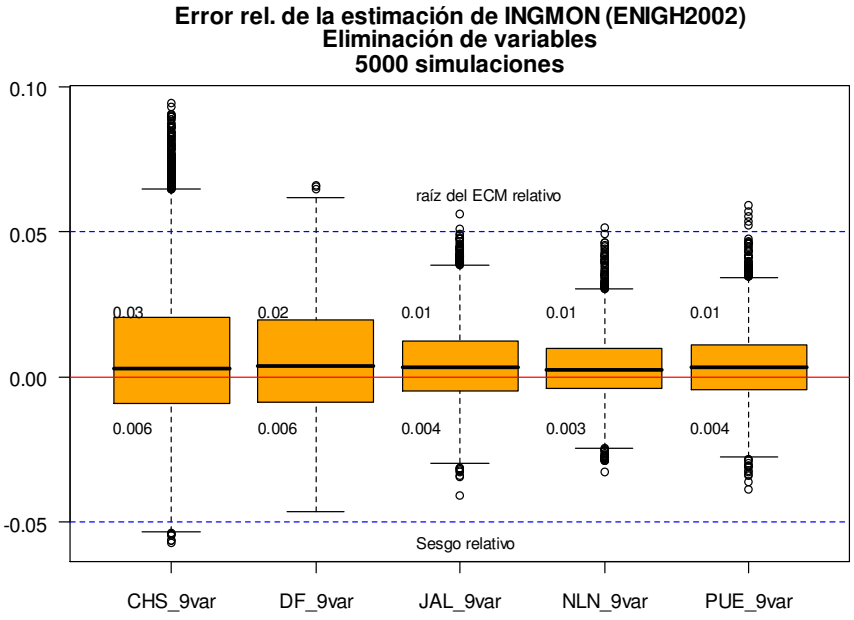


Tabla 5.8. Estadísticos del error en cada región.

Estadísticos del error	*Entidad Federativa*				
	Chiapas (CHS)	*Distrito Federal (DF)*	*Jalisco (JAL)*	*Nuevo León (NLN)*	*Puebla (PUE)*
Raíz del ECM relativo	3	2	1	1	1
Sesgo relativo	0.6	0.6	0.4	0.3	0.4
D. E. Relativa	3	2	1	1	1

Los valores están expresados en magnitudes de $1x10^{-2}$.

Los resultados de las simulaciones indican que las estimaciones Horvitz-Thompson para la variable INGMON en el estado de Chiapas (CHS) presentan una mayor varianza relativa en comparación con el Distrito Federal (DF). La entidad de menor varianza resultó ser Nuevo León (NLN). Al graficar el error absoluto el DF es la de mayor error, pero este estadístico no proporciona alguna posibilidad para comparar los datos.

El ECM tiene un rango de variación desde 2e-04 para CHS, hasta 115e-05 para DF. El sesgo relativo varía entre 0.003 y 0.006, siendo CHS y DF los más altos. La D.E. relativa es de 0.03 para CHS, 0.02 para el DF y del 0.01 para el resto.

5.5 Comentarios

El resultado de las simulaciones muestran que para el caso de menos de diez variables auxiliares es prácticamente igual, tanto desde el punto de vista del error estándar como del sesgo relativo, hacer estimaciones utilizando programación lineal o eliminación de variables en la fase de aterrizaje del Método del Cubo. Inclusive al comparar variables auxiliares ordenadas contra las de un orden aleatorio, tampoco hay alguna diferencia apreciable.

C a p í t u l o V I

CONCLUSIONES Y RECOMENDACIONES

El instante sagrado no es un instante de creación, sino de reconocimiento. Porque el reconocimiento proviene de la visión y no de emitir juicios.

Helen Schucman y William Thetford
Meditaciones de un curso de milagros

Después de obtener los resultados de las simulaciones se está en condiciones de mencionar algunas conclusiones importantes. Las conclusiones que aquí se establecen permiten aplicar la técnica con mayor seguridad a una gama más amplia de problemas que los considerados antes de su aparición.

Con los resultados de las simulaciones, efectuadas con los datos MU284, se concluye que el cálculo de los coeficientes de regresión estandarizados es el indicado para establecer el orden de variables en cualquier conjunto menor a diez. Como el Método del Cubo obtiene la muestra sin tener que eliminar variables entonces no interesa la cantidad. Por orto lado, los resultados obtenidos con los datos ENIGH2002 permiten ver la necesidad de usar el estadístico cociente de verosimilitudes (LR) para determinar la cantidad de variables auxiliares apropiada.

La modelación del error estándar y del sesgo relativo de las estimaciones permite evaluar varios aspectos del método. Uno de ellos es el sesgo introducido al utilizar eliminación de variables. Otro es el cambio en la varianza del estimador. Uno más es el comportamiento de un determinado ordenamiento al cambiar de aplicación de una región geográfica a otra dentro del país.

El sesgo relativo que aparece en aquellas simulaciones en las que se empleó el método de eliminación de variables en la fase de aterrizaje es provocado porque al eliminar variables, el vector de probabilidades de inclusión se ve afectado en sus valores originales.

6.1 Datos MU284

De los resultados obtenidos se puede concluir que el haber hecho un ordenamiento de las variables auxiliares en la opción de eliminación de variables trae como resultado una reducción en el error de estimación. Y en el peor de los casos es igual que hacerlo sin ordenar las variables.

En un primer ejercicio con un archivo de menos de diez variables (MU284) se comparan en la fase de aterrizaje, el método de programación lineal contra el método de eliminación de variables. La conclusión obtenida es que el método de eliminación de variables aplicado a un conjunto de variables auxiliares ordenadas puede igualar o mejorar las precisiones obtenidas con el método de programación lineal considerando a las variables auxiliares sin ordenar.

En un segundo ejercicio se hace la comparación en varios escenarios de ordenamiento (datos MU284) del método de eliminación de variables en la fase de aterrizaje. En el resultado de las simulaciones se aprecia una clara mejora, producto del ordenamiento de las variables. Y en mayor medida si se hace el ordenamiento sin hacer distinción de si las variables auxiliares pertenecen al diseño de muestreo o no, es decir, sin hacer distinción entre las fijas y no fijas. Como dato curioso, el resultado de las simulaciones con el ordenamiento tal como aparece en el libro de Särndal es el que menos sesgo tiene incluso que el obtenido con un ordenamiento completamente descendente.

6.2 Datos ENIGH2002

En los datos ENIGH2002 permitió una reducción del número de variables auxiliares a considerar en el Método del Cubo, la técnica Componentes Principales Supervisadas permitió reducir de 104 a 24 variables el conjunto de información auxiliar.

Llama la atención que en este grupo de 24 variables auxiliares más informativas, aparezcan en los primeros lugares las variables 'limpieza' y 'cuidados', inmediatamente después de las variables de los grupos de ingresos ('ingcor' e 'ingtot') y gastos ('gasmon', 'gascor' y 'gastot'). Esto puede deberse a que la información corresponde a la ciudad más grande del país: el Distrito Federal. Esto se puede corroborar haciendo

simulaciones sobre otras entidades, por ejemplo una del norte y otra del sur, Nuevo León y Chiapas por mencionar algunas.

Al hacer un comparativo de la ganancia o pérdida en la precisión de las estimaciones al pasar de 24 variables auxiliares a sólo seis, se observa que cuando se tienen las variables ordenadas, prácticamente permanecen sin cambio las estimaciones, al reducir el número de variables de 24 a sólo seis.

En cambio el sesgo sí se ve afectado al hacer ese cambio, presentando las estimaciones un mayor sesgo conforme aumenta el número de variables auxiliares, aunque estén ordenadas.

6.3 Desarrollos posteriores

La primera aplicación real del Método del Cubo la hizo Jean Dumais al poner a prueba el programa para la selección de muestras balanceadas de municipios para el censo continuo del INSEE. La segunda aplicación fue llevada a cabo por Benoît Merlat, quien probó el Método del Cubo para seleccionar los distritos de la muestra maestra del INSEE. En resumen el INSEE ha adoptado el Método del Cubo para sus proyectos estadísticos más importantes. Sin embargo no se hace mención del ordenamiento de las variables auxiliares para mejorar el método.

Esta herramienta del Método del Cubo con la mejora sugerida en este trabajo tiene las siguientes aplicaciones inmediatas.

Muestra Censal 2010. Para que la muestra censal reproduzca la estructura demográfica y otras características de la población de México es necesario balancear la muestra en las variables que describen la estructura de la población como Grupos de edad y Sexo.

Marco Maestro. En la determinación de la muestra maestra balanceada al nivel de desglose conveniente y con las variables auxiliares adecuadas. Por ejemplo construir un Marco Maestro balanceado en las variables número de viviendas, número de habitantes y número de hogares.

Reducción del Sesgo. Para reducir el sesgo en las estimaciones, en los intervalos de confianza y en las pruebas de hipótesis.

Imputación. En este caso se utilizan los registros en los que se posee información para detectar patrones de comportamiento y se hacen estimaciones de los datos faltantes aplicando tales patrones o modelos.

Datos geográficos. Calcular la distribución de muestras espacialmente balanceadas.

Estratificación. Herramienta para detectar conglomerados o grupos de unidades muestrales ponderados de manera equilibrada en cuanto a la estructura socio-económico-demográfica de la población.

Apéndice A

Tabla A1. Los datos MU284

LABEL	P85	P75	RMT85	CS82	SS82	S82	ME84	REV84	REG	CL
1	33	27	288	13	24	49	2 135	2 836	1	1
2	19	15	139	14	12	41	957	2 035	1	1
3	26	20	196	12	14	41	1 530	6 030	1	1
4	19	15	159	12	19	41	1 059	4 704	1	1
5	56	52	536	20	27	61	3 951	5 183	1	1
6	16	15	134	16	12	41	918	2 157	1	2
7	70	62	623	18	27	61	4 367	7 072	1	2
8	66	54	517	15	32	61	4 345	5 246	1	2
9	12	12	96	10	12	31	754	951	1	2
10	60	50	467	14	29	61	3 902	6 067	1	2
11	32	29	277	14	20	45	1 993	3 264	1	3
12	20	14	155	10	21	41	1 312	1 899	1	3
13	53	40	386	24	13	51	2 780	5 931	1	3
14	28	27	241	24	8	45	1 649	3 877	1	3
15	48	43	422	19	18	51	2 983	4 968	1	3
16	653	671	6 263	34	41	101	45 324	59 877	1	4
17	79	78	612	14	31	61	5 331	7 027	1	4
18	59	54	532	23	23	61	3 994	6 529	1	4
19	27	28	250	9	22	41	1 616	2 208	1	4
20	49	55	412	20	27	61	3 240	3 976	1	4
21	38	36	339	21	11	51	2 055	4 438	1	5
22	6	6	55	11	12	31	304	960	1	5
23	42	39	290	12	25	57	2 294	7 990	1	5
24	29	27	249	13	23	49	1 899	2 719	1	5
25	21	19	164	8	25	45	1 217	2 389	1	5
26	14	9	97	12	15	35	679	1 462	2	6
27	9	10	74	3	20	31	490	1 751	2	6
28	20	21	144	5	27	49	1 109	2 259	2	6
29	153	138	1 277	21	36	81	7 910	13 205	2	6
30	33	32	240	10	24	51	1 837	3 281	2	6
31	21	19	163	7	23	49	1 176	7 082	2	7
32	10	10	63	4	18	35	454	952	2	7
33	65	62	488	14	33	61	3 254	6 389	2	7
34	13	14	111	4	19	31	759	1 830	2	7
35	17	18	128	6	22	45	871	1 833	2	7
36	32	33	230	9	30	51	1 788	2 914	2	7
37	89	92	720	15	46	79	5 495	6 772	2	7
38	25	22	179	11	22	49	1 286	2 628	2	7
39	6	6	37	8	17	41	257	655	2	8
40	4	4	24	5	11	35	177	637	2	8
41	10	11	65	9	19	49	470	1 387	2	8
42	6	6	37	5	23	41	255	742	2	8
43	13	13	89	6	24	45	591	1 307	2	8
44	24	25	187	8	33	55	1 313	2 165	2	9
45	9	9	60	10	21	49	415	1 136	2	9

LABEL	P85	P75	RMT85	CS82	SS82	S82	ME84	REV84	REG	CL
46	116	108	939	21	36	79	6 313	10 879	2	9
47	118	119	1 008	22	45	85	7 619	12 112	2	9
48	12	10	81	11	17	45	555	1 370	2	9
49	41	42	310	10	31	57	2 402	3 863	2	9
50	8	8	58	7	16	35	369	831	2	9
51	26	25	169	10	26	51	1 311	2 409	2	9
52	7	6	38	6	11	35	280	767	3	10
53	9	8	68	9	13	41	387	790	3	10
54	28	27	190	10	20	49	1 327	2 546	3	10
55	12	11	78	7	14	41	536	1 254	3	10
56	107	108	807	19	38	81	6 107	9 343	3	10
57	31	33	220	10	27	57	1 597	2 653	3	11
58	30	30	226	10	18	49	1 713	2 834	3	11
59	12	12	69	9	14	45	493	1 134	3	11
60	28	29	179	9	18	49	1 349	2 903	3	11
61	18	18	129	8	17	49	970	1 652	3	11
62	18	19	119	9	19	41	963	1 470	3	11
63	11	12	73	6	18	41	450	1 220	3	12
64	9	9	59	6	24	41	458	790	3	12
65	14	15	92	11	17	49	659	1 626	3	12
66	20	19	127	9	20	49	956	1 983	3	12
67	16	15	104	10	21	49	762	1 768	3	12
68	11	12	75	8	22	49	506	1 169	3	12
69	66	62	505	15	23	61	3 789	6 850	3	12
70	27	26	183	9	18	49	1 366	3 053	3	12
71	8	8	49	6	21	41	361	807	3	13
72	8	8	41	7	15	41	349	765	3	13
73	13	11	76	12	19	49	520	1 457	3	13
74	17	18	113	8	20	49	784	1 733	3	13
75	13	13	90	8	24	49	538	2 042	3	13
76	11	12	79	5	21	41	522	1 060	3	14
77	54	52	408	16	29	61	3 095	5 302	3	14
78	21	22	140	7	24	49	1 110	2 010	3	14
79	28	28	200	10	31	59	1 499	5 717	3	14
80	40	42	284	12	38	75	2 087	3 773	3	14
81	16	16	103	10	19	49	736	1 696	3	14
82	11	11	58	10	12	49	442	1 743	3	14
83	56	54	654	13	30	71	5 434	6 050	3	15
84	15	18	118	6	28	49	813	1 309	4	15
85	60	60	431	15	39	75	3 285	4 954	4	15
86	30	30	192	8	27	49	1 575	2 553	4	15
87	32	32	233	9	27	51	1 593	4 006	4	15
88	16	16	102	11	24	49	753	1 515	4	15
89	15	15	97	6	24	41	735	1 321	4	16
90	9	9	56	13	14	45	380	894	4	16
91	12	13	69	9	15	41	487	1 343	4	16
92	12	11	73	5	28	41	492	1 222	4	16
93	14	14	86	6	17	41	574	1 306	4	16
94	7	7	53	9	16	35	339	779	4	17
95	16	16	105	13	22	49	852	1 508	4	17

LABEL	P85	P75	RMT85	CS82	SS82	S82	ME84	REV84	REG	CL
96	13	11	81	9	21	41	598	1 026	4	17
97	12	11	77	16	10	49	481	2 027	4	17
98	70	67	472	17	35	71	3 613	6 317	4	17
99	20	20	129	11	19	49	868	2 216	4	17
100	31	28	194	17	16	49	1 343	3 073	4	17
101	49	48	299	15	25	61	2 212	4 055	4	17
102	13	12	81	7	18	41	528	1 277	4	18
103	17	15	122	14	17	41	894	1 670	4	18
104	14	14	113	10	24	41	729	1 276	4	18
105	25	21	186	24	14	49	1 142	3 252	4	18
106	14	14	91	8	25	41	664	1 380	4	18
107	21	19	145	12	25	49	1 100	4 117	4	19
108	17	15	133	16	17	45	882	1 610	4	19
109	16	13	114	10	22	45	753	1 413	4	19
110	13	11	78	8	15	41	521	1 232	4	19
111	15	14	78	10	17	49	543	1 622	4	19
112	13	12	72	8	12	41	532	1 276	4	20
113	11	10	70	11	13	41	502	1 201	4	20
114	229	247	3 471	20	32	61	24 694	17 949	4	20
115	81	75	641	19	23	65	4 807	6 382	4	20
116	35	38	275	13	30	51	2 112	3 096	4	20
117	105	102	815	20	31	65	6 323	9 371	4	21
118	22	21	149	15	17	41	1 031	2 872	4	21
119	26	26	185	10	23	49	1 495	2 540	4	21
120	24	24	172	13	23	49	1 299	2 355	4	21
121	34	35	240	12	30	51	2 089	3 475	4	21
122	11	11	67	6	20	49	469	1 726	5	22
123	77	74	526	17	32	71	4 245	7 533	5	22
124	21	19	122	10	15	49	917	2 784	5	22
125	36	33	229	10	19	51	1 560	4 014	5	22
126	46	43	298	10	19	51	2 439	10 691	5	22
127	48	37	317	17	15	51	2 115	6 087	5	22
128	25	20	179	11	16	41	1 089	2 412	5	23
129	29	27	233	12	16	41	1 377	2 489	5	23
130	10	9	66	14	12	41	485	880	5	23
131	17	14	129	10	16	41	940	3 690	5	23
132	12	10	87	10	10	41	543	1 828	5	23
133	13	10	82	9	15	41	576	2 092	5	24
134	9	9	63	7	19	41	385	1 438	5	24
135	11	10	65	6	17	41	418	1 135	5	24
136	12	11	72	8	12	41	524	1 736	5	24
137	424	446	6 720	21	35	81	47 074	38 945	5	24
138	49	47	381	14	25	61	2 655	4 660	5	25
139	31	28	240	12	19	49	1 677	3 307	5	25
140	15	15	118	7	23	41	781	3 397	5	25
141	46	47	396	12	30	61	2 610	4 623	5	25
142	10	9	66	5	16	39	463	1 190	5	25
143	5	5	35	6	12	41	253	538	5	26
144	7	7	43	6	13	35	344	725	5	26
145	23	22	166	9	26	49	1 122	2 025	5	26

LABEL	P85	P75	RMT85	CS82	SS82	S82	ME84	REV84	REG	CL
146	31	28	216	12	14	41	1 470	3 065	5	26
147	9	8	53	10	11	41	393	895	5	26
148	12	12	82	9	19	45	539	1 162	5	27
149	12	12	90	6	19	45	534	1 243	5	27
150	11	10	62	8	14	41	414	1 119	5	27
151	12	11	77	6	21	41	570	1 178	5	27
152	32	30	207	10	23	51	1 537	2 955	5	27
153	11	10	68	10	14	41	474	1 224	5	28
154	9	9	52	9	13	41	382	784	5	28
155	36	34	277	10	23	51	1 918	3 376	5	28
156	49	50	396	11	34	61	2 881	4 798	5	28
157	31	28	220	10	18	49	1 579	2 742	5	28
158	100	106	751	20	38	79	5 742	7 710	5	28
159	22	21	146	12	15	49	1 053	2 066	5	28
160	13	13	95	7	20	41	731	1 101	5	28
161	6	5	37	8	9	39	256	658	5	29
162	6	6	35	11	10	41	280	521	5	29
163	7	5	39	9	13	35	329	637	5	29
164	9	6	46	8	12	35	387	813	5	29
165	8	8	59	6	18	35	401	762	5	29
166	6	7	44	8	19	41	311	731	5	30
167	17	17	105	12	12	49	782	1 887	5	30
168	13	12	87	8	17	41	567	1 324	5	30
169	11	11	66	6	18	41	597	1 007	5	30
170	10	10	62	8	16	41	473	971	5	30
171	24	25	175	10	21	49	1 386	2 269	5	31
172	35	35	252	10	20	51	1 890	3 335	5	31
173	18	17	128	11	19	49	1 058	1 592	5	31
174	46	45	330	12	27	59	2 656	4 041	5	31
175	9	8	57	8	14	33	389	818	5	31
176	13	13	83	5	19	41	623	1 192	5	31
177	32	33	208	11	19	51	1 559	2 681	5	31
178	11	9	74	8	17	39	533	977	6	32
179	9	10	63	6	22	41	411	869	6	32
180	15	16	109	7	25	49	803	2 266	6	32
181	5	5	37	5	21	35	235	513	6	32
182	13	11	105	6	18	31	702	1 366	6	32
183	5	6	38	3	27	39	245	433	6	33
184	12	11	82	6	25	41	646	935	6	33
185	10	11	76	5	27	45	501	1 392	6	33
186	10	10	63	8	13	41	419	1 011	6	33
187	13	14	86	7	15	41	552	1 444	6	33
188	74	73	603	17	28	61	4 270	6 635	6	34
189	26	28	209	9	26	49	1 398	2 291	6	34
190	14	16	100	8	25	41	753	1 509	6	34
191	17	19	121	4	32	49	885	1 868	6	34
192	27	27	191	9	23	49	1 441	2 557	6	34
193	18	20	130	9	22	49	839	1 719	6	34
194	8	9	63	3	23	41	411	950	2	35
195	17	17	128	6	25	45	1 048	1 606	2	35

LABEL	P85	P75	RMT85	CS82	SS82	S82	ME84	REV84	REG	CL
146	31	28	216	12	14	41	1 470	3 065	5	26
147	9	8	53	10	11	41	393	895	5	26
148	12	12	82	9	19	45	539	1 162	5	27
149	12	12	90	6	19	45	534	1 243	5	27
150	11	10	62	8	14	41	414	1 119	5	27
151	12	11	77	6	21	41	570	1 178	5	27
152	32	30	207	10	23	51	1 537	2 955	5	27
153	11	10	68	10	14	41	474	1 224	5	28
154	9	9	52	9	13	41	382	784	5	28
155	36	34	277	10	23	51	1 918	3 376	5	28
156	49	50	396	11	34	61	2 881	4 798	5	28
157	31	28	220	10	18	49	1 579	2 742	5	28
158	100	106	751	20	38	79	5 742	7 710	5	28
159	22	21	146	12	15	49	1 053	2 066	5	28
160	13	13	95	7	20	41	731	1 101	5	28
161	6	5	37	8	9	39	256	658	5	29
162	6	6	35	11	10	41	280	521	5	29
163	7	5	39	9	13	35	329	637	5	29
164	9	6	46	8	12	35	387	813	5	29
165	8	8	59	6	18	35	401	762	5	29
166	6	7	44	8	19	41	311	731	5	30
167	17	17	105	12	12	49	782	1 887	5	30
168	13	12	87	8	17	41	567	1 324	5	30
169	11	11	66	6	18	41	597	1 007	5	30
170	10	10	62	8	16	41	473	971	5	30
171	24	25	175	10	21	49	1 386	2 269	5	31
172	35	35	252	10	20	51	1 890	3 335	5	31
173	18	17	128	11	19	49	1 058	1 592	5	31
174	46	45	330	12	27	59	2 656	4 041	5	31
175	9	8	57	8	14	33	389	818	5	31
176	13	13	83	5	19	41	623	1 192	5	31
177	32	33	208	11	19	51	1 559	2 681	5	31
178	11	9	74	8	17	39	533	977	6	32
179	9	10	63	6	22	41	411	869	6	32
180	15	16	109	7	25	49	803	2 266	6	32
181	5	5	37	5	21	35	235	513	6	32
182	13	11	105	6	18	31	702	1 366	6	32
183	5	6	38	3	27	39	245	433	6	33
184	12	11	82	6	25	41	646	935	6	33
185	10	11	76	5	27	45	501	1 392	6	33
186	10	10	63	8	13	41	419	1 011	6	33
187	13	14	86	7	15	41	552	1 444	6	33
188	74	73	603	17	28	61	4 270	6 635	6	34
189	26	28	209	9	26	49	1 398	2 291	6	34
190	14	16	100	8	25	41	753	1 509	6	34
191	17	19	121	4	32	49	885	1 868	6	34
192	27	27	191	9	23	49	1 441	2 557	6	34
193	18	20	130	9	22	49	839	1 719	6	34
194	8	9	63	3	23	41	411	950	2	35
195	17	17	128	6	25	45	1 048	1 606	2	35

LABEL	P85	P75	RMT85	CS82	SS82	S82	ME84	REV84	REG	CL
196	12	12	90	3	26	41	607	1 347	2	35
197	10	11	79	4	22	39	486	1 102	2	35
198	7	7	50	1	18	35	325	688	2	35
199	118	118	1 025	15	32	65	7 700	11 126	2	36
200	18	17	121	5	21	41	938	1 645	2	36
201	12	11	87	7	23	45	552	1 408	2	36
202	35	38	309	10	29	51	2 173	3 635	2	36
203	10	9	75	4	20	35	600	1 064	2	36
204	25	25	189	7	25	49	1 437	2 898	2	36
205	5	5	36	4	24	39	255	692	2	37
206	11	11	82	4	28	39	612	913	2	37
207	13	13	84	4	18	41	564	1 334	2	37
208	8	8	60	6	19	39	390	723	2	37
209	17	19	132	7	31	49	989	1 197	2	37
210	7	7	46	4	22	35	296	605	2	38
211	118	118	965	15	35	65	6 856	10 702	2	38
212	21	20	155	8	21	49	1 150	1 988	2	38
213	14	16	132	6	24	41	769	1 460	2	38
214	27	28	219	6	28	49	1 544	2 441	2	38
215	14	15	112	7	22	41	761	1 238	2	38
216	8	9	53	6	24	49	311	844	6	39
217	12	12	90	6	24	49	543	2 098	6	39
218	10	9	66	6	19	41	452	1 113	6	39
219	14	13	100	8	18	49	700	1 841	6	39
220	11	11	77	9	18	49	461	1 416	6	39
221	7	7	52	7	15	31	356	826	6	40
222	8	9	60	5	23	49	422	2 100	6	40
233	13	13	100	5	32	49	665	1 500	6	40
224	20	18	153	4	19	49	1 039	2 223	6	40
225	51	47	418	14	26	61	2 771	5 137	6	40
226	46	46	392	9	36	61	2 711	4 852	6	41
227	11	10	76	7	18	41	503	1 220	6	41
228	17	17	130	6	20	41	922	1 630	6	41
229	25	27	197	5	28	49	1 239	2 461	6	41
230	30	33	262	6	32	51	1 741	2 395	6	41
231	7	6	49	3	23	41	321	694	6	42
232	12	15	105	4	27	41	672	1 008	6	42
233	14	13	87	3	17	41	661	1 447	6	42
234	12	12	70	3	18	41	579	1 150	6	42
235	21	22	147	5	22	49	1 242	2 459	6	42
236	88	85	720	14	44	75	4 758	8 760	6	43
237	41	43	342	8	32	51	2 182	4 146	6	43
238	31	32	233	5	32	51	1 579	3 298	6	43
239	28	28	203	6	25	49	1 524	2 903	6	43
240	38	37	249	5	24	51	2 073	3 882	6	43
241	13	14	92	3	23	41	659	1 989	7	44
242	18	18	134	3	29	49	945	2 616	7	44
243	28	27	231	8	22	49	1 560	2 414	7	44
244	94	93	782	12	46	81	5 779	9 828	7	44
245	26	28	184	5	35	61	1 455	2 067	7	44

LABEL	P85	P75	RMT85	CS82	SS82	S82	ME84	REV84	REG	CL
246	26	26	182	6	37	61	1 494	6 928	7	44
247	60	60	432	6	33	61	3 070	6 502	7	44
248	7	8	49	3	27	45	427	3 832	7	45
249	9	10	63	5	30	49	502	970	7	45
250	14	13	94	5	23	49	813	2 486	7	45
251	17	18	124	4	29	49	1 061	2 881	7	45
252	10	9	69	7	23	49	593	2 399	7	45
253	9	9	53	7	21	45	484	1 491	7	45
254	13	13	91	5	28	49	736	3 647	7	45
255	56	53	451	12	38	75	3 430	4 677	7	45
256	8	8	53	3	14	31	428	727	8	46
257	3	4	21	5	11	31	173	347	8	46
258	7	7	42	5	14	35	323	903	8	46
259	8	7	50	3	15	41	408	706	8	46
260	5	6	40	3	15	31	287	1 063	8	46
261	4	4	32	2	19	31	236	704	8	47
262	8	8	58	6	20	41	431	2 246	8	47
263	4	4	26	2	14	31	202	623	8	47
264	4	4	28	2	19	31	199	422	8	47
265	8	8	61	4	17	35	477	1 321	8	47
266	9	9	60	2	20	35	518	1 313	8	48
267	4	5	35	2	19	31	252	687	8	48
268	84	74	764	10	30	65	5 292	10 827	8	48
269	14	15	105	4	24	45	942	1 911	8	48
270	74	72	592	7	36	65	4 777	7 624	8	48
271	8	8	55	3	25	41	462	690	8	49
272	4	4	28	1	15	31	214	888	8	49
273	7	8	46	4	21	35	414	9 052	8	49
274	5	6	34	1	20	31	280	359	8	49
275	19	18	123	4	27	41	1 175	1 544	8	49
276	6	7	41	3	18	39	302	430	8	50
277	9	10	54	4	19	41	528	551	8	50
278	24	26	207	5	20	41	1 582	3 703	8	50
279	10	9	64	2	24	39	480	689	8	50
280	67	64	562	9	34	61	3 948	6 583	8	50
281	39	35	295	5	32	51	2 227	4 033	8	50
282	29	27	226	7	28	49	1 682	2 898	8	50
283	10	9	63	5	19	41	604	594	8	50
284	27	31	233	5	27	45	1 788	2 366	8	50

Apéndice B

CÓDIGO EN LENGUAJE R

El siguiente es el código en R usado para generar y evaluar las simulaciones de datos para el **primer** ejercicio con los datos MU284, correspondiente a la evaluación de los métodos **Programación Lineal** y **Eliminación de Variables** cuando el conjunto de información auxiliar posee menos de diez variables auxiliares.

```
##
#   Maestría en Ciencias en Estadística Oficial
#
#            I N E G I  -  C I M A T
##


######
#
#                T  E  S  I  S
#    ORDENAMIENTO DE VARIABLES AUXILIARES
#           EN MUESTREO BALANCEADO
#
#        José de Jesús Suárez Hernández
#                 Junio 2007
#
######

###
#             METODOLOGÍA EMPLEADA
#
#                Método del Cubo
#              Fase de aterrizaje
# Programación Lineal versus Eliminación de Variables
#
#    Componentes Principales Supervisadas
# Orden estadístico de las variables auxiliares
#   Coeficientes de regresión estandarizados
#
###

###
#           A  R  T  I  C  U  L  O    1
#                 Biometrika
#
#  "Efficient balanced sampling: The cube method"
#              Jean-Claude Deville
#                   Yves Tillé
#
#                   March 2004
###

###
#           A  R  T  I  C  U  L  O    2
# Journal of the American Statistical Association
#
# "Prediction by Supervised Principal Components"
#          Eric Bair, Trevor Hastie,
#      Debashis Paul, Robert Tibshirani
#
#                   March 2006
###

##
```

```
#                 Ejercicio 1
#         Simulación de datos de MU284
#           284 municipios de Suecia
#         Sarndal et al.(1992, pp 252-9)
#
# Programación Lineal contra Eliminación de Variables
#
#
#           Descripción de Variables
#
#   LABEL   Identifier running from 1 to 284.
#   P85     1985 population (in thousands).
#   P75     1975 population (in thousands).
#   RMT85   Revenues from the 1985 municipal taxation (in millions of
#           kronor).
#   CS82    Number of Conservative seats in municipal council.
#   SS82    Number of Social-Democratic seats in municipal council.
#   S82     Total number of seats in municipal council.
#   ME84    Number of municipal employees in 1984.
#   REV84   Real estate values according to 1984 assessment (in millions of
#           kronor).
#   REG     Geographic region indicator.
#   CL      Cluster indicator (a cluster consists of a set of neighboring
#           municipalities).
##

###
# Los casos que se simulan son los siguientes:
#
# 1) Programacion Lineal con 9 Variables en Orden Aleatorio
# 2) Programacion Lineal con 9 Variables en Orden Descendente
# 3) Programacion Lineal con 6 Variables en Orden Descendente
#
# 4) Eliminacion de Variables con 9 Variables en Orden Aleatorio
# 5) Eliminacion de Variables con 9 Variables en Orden Descendente
# 6) Eliminacion de Variables con 6 Variables en Orden Descendente
###

# Fijar una semilla para obtener los mismos resultados en cada ejecución
  set.seed(2007)

# Cargar la librería Anál. de Componentes Principales Supervisados(v. 1.03)
  library(superpc)

superpc.cv(superpc)     Cross-validation for supervised principal
                        components
superpc.decorrelate(superpc)
                        Decorrelate features with respect to competing
                        predictors
superpc.fit.to.outcome(superpc)
                        Fit predictive model using outcome of
                        supervised principal components
cor.func(superpc)       Internal superpc functions
superpc.listfeatures(superpc)
                        Return a list of the important predictors
superpc.lrtest.curv(superpc)
                        Compute values of likelihood ratio test from
                        supervised principal components fit
superpc.plot.lrtest(superpc)
                        Plot likelhiood ratio test statistics
superpc.plotcv(superpc)
                        Plot output from superpc.cv
superpc.plotred.lrtest(superpc)
                        Plot likelihood ratio test statistics from
                        supervised principal components predictor
superpc.predict(superpc)
                        Form principal components predictor from a
                        trained superpc object
superpc.predict.red(superpc)
                        Feature selection for supervised principal
                        components
superpc.predict.red.cv(superpc)
                        Cross-validation of feature selection for
                        supervised principal components
```

```
superpc.predictionplot(superpc)
                        Plot outcome predictions from superpc
superpc.rainbowplot(superpc)
                        Make rainbow plot of superpc and compeiting
                        predictors
superpc.train(superpc)
                        Prediction by supervised principal components

# www-stat.stanford.edu/~tibs
# www-stat.stanford.edu/~tibs/ftp/spca.pdf

# Lectura de datos del archivo en formato “separado por comas”
  datos<-read.csv("C:\\MU284.csv",header=T)

# Separación de la variable de interés (RMT85) y var. auxiliares
  xij<-datos[,-c(1,4)]

# Tamaño del conjunto de datos de la población (Matriz A de Nxp)
# Donde N es el tamaño de la población y p el número de variables auxiliares.
  N<-dim(datos)[1]
  p<-dim(datos)[2]

# Construcción de la matriz de prueba. Es del mismo tamaño que el conjunto de datos
# de la población.

  xij_TEST <- matrix(0,N,p)

# Selección de la variable de interés.
  yy<- datos_MU284[,4] #RMT85 Revenues from the 1985 mpal. Taxation
  tot_MU284<- sum(datos_MU284[,4])

# Centrar los datos de la matriz X. Hacer los datos con media cero y varianza uno.
  xij_c<-scale(xij,T)

# Poner etiquetas a las variables auxiliares en el objeto que necesita la función
# “superpc”
  featurenamesEXA <- names(xij[1,])

# Tamaño de muestra igual al 10% de la población
# Ver “The Elements of Statistical Learning” Hastie y col (2001)
  sTEST<-sample(1:N,round(N/10,0))
  xij_c.testEXA <- xij_c[sTEST,]
  yyTEST<-yy[sTEST]

# Crear un objeto de Datos de Entrenamiento que necesita la función "superpc"
  dataEXA<-list(x=t(xij_c),y=yy,featurenames= featurenamesEXA)

# Etiquetas del Conjunto de Prueba (Son las mismas que el de Entrenamiento)
  featurenamesTEST<-featurenamesEXA

# Crear un objeto de Datos de Prueba que necesita la función "superpc"
  dataTEST<-list(x=t(xij_c.testEXA),y=yyTEST,featurenames= featurenamesTEST)

# Cálculo de la Regresión en el Conjunto de Entrenamiento. Incluye Cálculo de Scores
  train.objEXA<- superpc.train(dataEXA, type="regression")

# Cálculo y Ordenamiento de los Scores en valor absoluto
  scores.abs<-sort(abs(train.objEXA$feature.scores),
     decreasing=TRUE)

  scores.order<-order(abs(train.objEXA$feature.scores),decreasing=TRUE)


# Ordenar las columnas de la matriz de datos de acuerdo al orden de los scores
# obtenidos por la herramienta
  xij.order<-xij[,scores.order]
> scores.order
[1] 6 2 1 7 5 3 4 9 8

  scores.spc<-scores.abs>floor(theta)
  xij.spc<-xij.order[,scores.spc]

##
#                           Muestreo Balanceado
```

```
#                              Método del Cubo
##

############
# Simulaciones de muestreo con el Método del Cubo
#  usando en la fase de aterrizaje la opción de
#            Programacion Lineal
############

# Librerías necesarias para el Método del Cubo
  library(MASS)
  library(lpSolve)
  library(sampling)

# Construcción de la matriz de variables auxiliares

# Matriz con variables auxiliares ordenadas aleatoriamente (desorden, sin orden)
  xij.random_table<-as.data.frame(xij.random,optional=TRUE)

# Matriz con variables auxiliares ordenadas descendentemente
  xij.order_table<-as.data.frame(xij.order,optional=TRUE)

# Cálculo de las probabilidades de inclusión para un diseño
# con probabilidad proporcional al tamaño (PPT)
# para las cincuenta observaciones consideradas en muestra
# 50/284 = 17.6 % de muestra aprox.
  pik=inclusionprobabilities(xij.random_table$P75,50)

# Definición de la matriz de variables de balanceo
# conteniendo el esquema de orden
  X.order<-as.matrix(xij.order_table)
  X.random<-as.matrix(xij.random_table)

# Número de Simulaciones efectuadas (previamente se hicieron prubas con 1000 y 2000)
  sim=3000

tot_PL9var_random <- tot_PL9var_order <- tot_PL6var_order <-
   rep(0,times=sim)
for(i in 1:sim)
{
   cat("Simulation number ",i,"\n")
# Cálculo de residuales
# Método aplicado: Eliminación de variables en la fase de aterrizaje

   s9_order<-samplecube(X.order,pik,1,FALSE,2)
   tot_PL9var_order[i]=HTestimator(yy_MU284[s9_order==1],pik[s9_order==1])
   s9_random<-samplecube(X.random,pik,1,FALSE,2)
   tot_PL9var_random[i]=HTestimator(yy_MU284[s9_random==1],pik[s9_random==1])
   s6_order<-samplecube(X.order[,1:6],pik,1,FALSE,2)
   tot_PL6var_order[i]=HTestimator(yy_MU284[s6_order==1],pik[s6_order==1])
}

# Función que obtiene la muestra balanceada usando el Método del Cubo
#
# samplecube(X,pik,order=1,comment=TRUE,method=1)
#
# Argumentos
# X = Matriz de variables auxiliares sobre la cual la muestra debe ser balanceada
# pik = Vector of probabilidades de inclusion.
# Orden=1, Los datos son ordenados aleatoriamente,
# Orden=2, No hay cambio en el orden de los datos,
# Orden=3, Los datos son ordenados en orden decreciente.
# comment, Un comentario es escrito durante la ejecucion si comment es TRUE.
# method=1, Para aplicar programación lineal en la fase de aterrizaje,
# method=2, Para aplicar eliminación de variables en la fase de aterrizaje.

> X.order[1,]
 ME84   P75   P85 REV84   S82  CS82  SS82    CL   REG
 2135    27    33  2836    49    13    24     1     1
>
> X.random[1,]
REV84  CS82   P85  SS82   REG    CL   S82  ME84   P75
 2836    13    33    24     1     1    49  2135    27
>
```

```
# Respaldo de la información obtenida con las simulaciones
PL_3000<-data.frame(cbind(tot_PL9var_random,tot_PL9var_order,tot_PL6var_order))
 write.csv(PL_3000, file = "C:\\PL_3000_sim.csv")
 PLvsEV_3000<-read.csv("C:\\PLvsEV_3000sim.csv",header=T)
EV_3000<-read.csv("C:\\EV_3000sim.csv",header=T)

#############
# Respaldo de información
#############
save(tot_PL9var_random, tot_PL9var_order, tot_PL6var_order, file = "C:\\
PLvsEV_3000.csv")
PL_3000<-data.frame(cbind(tot_PL9var_random,tot_PL9var_order,tot_PL6var_order))

write.csv(PL_3000, file = "C:\\PL_3000_sim.csv")
PLvsEV_3000<-read.csv("C:\\PLvsEV_3000sim.csv",header=T)
EV_3000<-read.csv("C:\\ EV_3000sim.csv",header=T)
#############



#############
##  Supression of variables method
#############

# Construcción de la matriz de variables auxiliares
    #xij.table<-as.data.frame(xij.spc,optional=TRUE)

# Cálculo de las probabilidades de inclusión
    #pik=inclusionprobabilities(xij.table$P75,50)

# Definición de la matriz de variables de balanceo
    #X<-as.matrix(xij.table)

set.seed(2007)
  sim=3000

tot_EV9var_random <- tot_EV9var_order <- tot_EV6var_order <- rep(0,times=sim)
for(i in 1:sim)
 {
  s9_order<-samplecube(X.order,pik,1,FALSE,2)
  tot_EV9var_order[i]=HTestimator(yy_MU284[s9_order==1],pik[s9_order==1])
  s9_random<-samplecube(X.random,pik,1,FALSE,2)
  tot_EV9var_random[i]=HTestimator(yy_MU284[s9_random==1],pik[s9_random==1])
  s6_order<-samplecube(X.order[,1:6],pik,1,FALSE,2)
  tot_EV6var_order[i]=HTestimator(yy_MU284[s6_order==1],pik[s6_order==1])
 }
###
# Respaldo de información
###
  EV_3000<-data.frame(cbind(tot_EV9var_random,tot_EV9var_order,tot_EV6var_order))
  write.csv(EV_3000, file = "C:\\EV_3000_sim.csv")

  tot_RMT85<-sum(datos_MU284[,4])

# Cálculo de los errores relativos
# Un conjunto de errores relativos para cada estrategia
  res_PL9ran<-(tot_PL9var_random-tot_RMT85)/tot_RMT85
  res_PL9ord<-(tot_PL9var_order-tot_RMT85)/tot_RMT85
  res_PL6ord<-(tot_PL6var_order-tot_RMT85)/tot_RMT85

  res_EV9ran<-(tot_EV9var_random-tot_RMT85)/tot_RMT85
  res_EV9ord<-(tot_EV9var_order-tot_RMT85)/tot_RMT85
  res_EV6ord<-(tot_EV6var_order-tot_RMT85)/tot_RMT85

tot_3mil<-
     cbind(tot_PL9var_random,tot_PL9var_order,tot_PL6var_order,tot_EV9var_random,tot
     _EV9var_order,tot_EV6var_order)

tot_comp<-cbind(tot_PL9var_random, tot_PL9var_order, tot_PL6var_order,
     tot_EV9var_random, tot_EV9var_order, tot_EV6var_order)

srrmse_PLvsEV_3mil_2<-""
for (k in 1:6)
```

```
{
    srrmse_PLvsEV_3mil_2[k]=sqrt(sum(tot_comp[k]^2)/sim)/tot_RMT85;
}
srmse_PLvsEV_3mil<- as.character(round(as.numeric(srrmse_PLvsEV_3mil),digits=5))

  tot_comp<-cbind(res_PL9ran,res_PL9ord,res_PL6ord,res_EV9ran,res_EV9ord,res_EV6ord)
  colnames(tot_comp) <- c("PL9ran","PL9ord","PL6ord","EV9ran","EV9ord","EV6ord")

  boxplot(data.frame(tot_comp), las=1,col="orange",main=c("Error rel. de la
     estimacion de RMT85","Prog. lineal vs Elim. de var.","3000 simulaciones"))
  abline(0,0,col="red")

error_std_PLvsEV_3mil<-sesgo_rel_PLvsEV_3mil<-ECM_rel_PLvsEV_3mil<-""
ee <- function(x){
   as.character(round(sqrt(sum((x-mean(x))^2)/sim)/tot_RMT85,5))
 }
sr <- function(x){
  as.character(round((mean(x)-tot_RMT85)/tot_RMT85,5))
}
ecm_r<- function(x){
  as.character(round(sqrt(sum((x-mean(x))^2)/sim+(mean(x)-tot_RMT85))/tot_RMT85,5))

}
for (i in 1:6)
{
error_std_PLvsEV_3mil[i] <- ee(tot_3mil[,i])
sesgo_rel_PLvsEV_3mil[i] <- sr(tot_3mil[,i])
ECM_rel_PLvsEV_3mil[i] <- ecm_r(tot_3mil[,i])
}

# # # # # # # # # # #
> ECM_rel_PLvsEV_3mil
[1] "0.0106"  "0.00865" "0.00859" "0.01015" "0.00879" "0.00907"
> error_std_PLvsEV_3mil
[1] "0.0106"  "0.00865" "0.00859" "0.01015" "0.00879" "0.00907"
> sesgo_rel_PLvsEV_3mil
[1] "0.00145" "0.00148" "0.00127" "0.00143" "0.0014"  "0.00176"
>
# El sesgo es muy pequeño y el error cuadrático medio está cargado en la varianza


 ECM_rel<-""
 for  (k in 1:6) {
   ECM_rel[k] <-
     sqrt(as.numeric(sesgo_rel_PLvsEV_3mil[k])^2+as.numeric(error_std_PLvsEV_3mil[k]
     )^2)
 }

Tabla_errores_PLvsEV_2<-
   rbind(ECM_rel_PLvsEV_3mil,sesgo_rel_PLvsEV_3mil,error_std_PLvsEV_3mil)
> Tabla_errores_PLvsEV_2
                      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
ECM_rel_PLvsEV_3mil   "0.0106"  "0.00865" "0.00859" "0.01015" "0.00879" "0.00907"
sesgo_rel_PLvsEV_3mil "0.00145" "0.00148" "0.00127" "0.00143" "0.0014"  "0.00176"
error_std_PLvsEV_3mil "0.0106"  "0.00865" "0.00859" "0.01015" "0.00879" "0.00907"
>


# # # # # # # # # # #


legend(2,0.023, "raiz del ECM relativo",cex=.8, bty="n",col="blue")
  for (i in 1:6){
    legend(i-0.9, 0.018, ECM_rel_PLvsEV_3mil[i],cex=.8, bty="n")
    legend(i-0.9, -0.013, sesgo_rel_PLvsEV_3mil[i],cex=.8, bty="n")
  }
 legend(2,-0.009, "Sesgo relativo",cex=.8, bty="n",col="blue")



write.csv(Tabla_errores_PLvsEV_2, file = "C:\\Tabla_errores_PLvsEV_2.csv")

#  F I N
```

El siguiente es el código en R usado para generar y evaluar las simulaciones de datos para el **segundo** ejercicio con los datos MU284, correspondiente a la evaluación de **Varios Escenarios de Orden** en las variables auxiliares.

```
###
#                       Simulación en los datos MU284
#                     Varios escenarios de ordenamiento
###

###
# Los escenarios de orden que se simulan son los siguientes:
#
# 1) Särndal Tienen el mismo orden que en Särndal (1992)
# 2) Worst   El peor de los ordenamientos. En orden ascendente.
# 3) Regular Las variables importantes están a la mitad del ordenamiento
# 4) Best    Un mejor ordenamiento
# 5) Topdown Ordenamiento completamente descendente
#
###

##
#                       Técnica Componentes Principales
#                                Supervisadas
##

# Fijar la semilla para obtener los mismos resultados de simulación
set.seed(2007)

# Cargar librería Anál. de Comp. Ppales. Superv. (v. 1.03)
library(superpc)

# Lectura de datos
  datos_MU284<-read.csv("C:\\MU284.csv",header=T)

# Separación de la variable LABEL (1)
# Separación de la variable de interés RMT85 (4)
  xij_MU284<-datos_MU284[,-c(1,4)]

# Tamaño de la población (Matriz A Nxp)
N<-dim(xij)[1]
p<-dim(xij)[2]

# Construcción de la matriz de prueba
xij_TEST <- matrix(0,N,p)

# Selección de la variable de interés

  yy_MU284<- datos_MU284[,4]        #RMT85 Revenues from the 1985 municipal taxation
  tot_RMT85<-sum(datos_MU284[,4])

# Centrar datos de la matriz X
  xij_c<-scale(xij,T)

# Poner Etiquetas a los campos
  featurenamesEXA <- names(xij[1,])

# Seleccionar como muestra al 10% de la población
# Ver "The Elements of Statistical Learning"
  sTEST<-sample(1:N,round(N/10,0))

# Conjunto de prueba para las var. auxiliares
  xij_c.testEXA <- xij_c[sTEST,]

# Conjunto de prueba para la var de interés
  yyTEST<-yy[sTEST]

# Crear el objeto que necesita la función superpc
dataEXA<-list(x=t(xij_c),y=yy,featurenames= featurenamesEXA)
```

```
# Etiquetas del conjunto de prueba
featurenamesTEST<-featurenamesEXA

# Crear un objeto que necesita la función "superpc"
dataTEST<-list(x=t(xij_c.testEXA),y=yyTEST,featurenames= featurenamesTEST)

# Regresión del conjunto de entrenamiento (1000 scores)
train.objEXA<- superpc.train(dataEXA, type="regression")

# Scores ordenados
scores.abs<-sort(abs(train.objEXA$feature.scores),decreasing=TRUE)

# Ordenamiento en los diferentes escenarios
scores.sarndal<-c(1,2,3,4,5,6,7,8,9)  # Tienen el mismo que en Särndal(1992)
scores.worst<-c(9,8,4,3,5,7,1,2,6)    # En orden ascendente
scores.regular<-c(9,8,4,5,1,6,2,7,3)  # Las variables importantes están a la mitad
scores.best<-c(9,8,6,2,1,7,5,3,4)     # Un mejor ordenamiento
scores.topdown<-c(6,2,1,7,5,3,4,9,8)  # Ordenamiento completamente descendente


# Reordenamiento de columnas en la matriz A
  xij.sarndal<-xij_MU284[,scores.sarndal]
  xij.worst  <-xij_MU284[,scores.worst]
  xij.regular<-xij_MU284[,scores.regular]
  xij.best   <-xij_MU284[,scores.best]
  xij.topdown<-xij_MU284[,scores.topdown]


##
#                            Muestreo Balanceado
#                              Método del Cubo
##

# Librerías necesarias para el Método del Cubo
library(MASS)
library(lpSolve)
library(sampling)

# Construcción de la matriz de variables auxiliares
xij.table_sar<-as.data.frame(xij.sarndal,optional=TRUE)
xij.table_wor<-as.data.frame(xij.worst,optional=TRUE)
xij.table_reg<-as.data.frame(xij.regular,optional=TRUE)
xij.table_bes<-as.data.frame(xij.best,optional=TRUE)
xij.table_topdown<-as.data.frame(xij.topdown,optional=TRUE)

# Cálculo de las probabilidades de inclusión
# 50/284 = 17.6% de muestra

pik_sar=inclusionprobabilities(xij.table_sar$P75,50)
pik_wor=inclusionprobabilities(xij.table_wor$P75,50)
pik_reg=inclusionprobabilities(xij.table_reg$P75,50)
pik_bes=inclusionprobabilities(xij.table_bes$P75,50)
pik_topd=inclusionprobabilities(xij.table_topdown$P75,50)

# Definición de la matriz de variables de balanceo
X_sar<-as.matrix(xij.table_sar)
X_wor<-as.matrix(xij.table_wor)
X_reg<-as.matrix(xij.table_reg)
X_bes<-as.matrix(xij.table_bes)
X_topd<-as.matrix(xij.table_topdown)

# Simulaciones
sim=3000
s_sar_3mil<-s_wor_3mil<-s_reg_3mil<-s_bes_3mil<-s_topd_3mil<-""
tot_sar_3mil<-tot_wor_3mil<-tot_reg_3mil<-tot_bes_3mil<-tot_topd_3mil<-
     rep(0,times=sim)

for(i in 1:sim)
{
   # Cálculo de residuales
   # "Eliminación de variables" en la fase de aterrizaje

       s_sar_3mil<-samplecube(X_sar,pik_sar,1,FALSE,2)
     tot_sar_3mil[i]<-HTestimator(yy_MU284[s_sar_3mil==1],pik_sar[s_sar_3mil==1])
```

```
        s_wor_3mil<-samplecube(X_wor,pik_wor,1,FALSE,2)
      tot_wor_3mil[i]<-HTestimator(yy_MU284[s_wor_3mil==1],pik_sar[s_wor_3mil==1])

        s_reg_3mil<-samplecube(X_reg,pik_reg,1,FALSE,2)
      tot_reg_3mil[i]<-HTestimator(yy_MU284[s_reg_3mil==1],pik_reg[s_reg_3mil==1])

        s_bes_3mil<-samplecube(X_bes,pik_bes,1,FALSE,2)
      tot_bes_3mil[i]<-HTestimator(yy_MU284[s_bes_3mil==1],pik_bes[s_bes_3mil==1])

        s_topd_3mil<-samplecube(X_topd,pik_topd,1,FALSE,2)
     tot_topd_3mil[i]<-HTestimator(yy_MU284[s_topd_3mil==1],pik_topd[s_topd_3mil==1])

}

Varios_Escenarios_3000<-data.frame(cbind(tot_sar_3mil, tot_wor_3mil, tot_reg_3mil,
tot_bes_3mil, tot_topd_3mil))

write.csv(Varios_Escenarios_3000, file = "C:\\Varios_Escenarios_3000sim.csv")

tot_V_E_3mil<-cbind(tot_sar_3mil, tot_wor_3mil, tot_reg_3mil, tot_bes_3mil,
tot_topd_3mil)

srrmse_3mil<-""

for (k in 1:5){
srrmse_3mil[k]=sqrt(sum((tot_V_E_3mil[k]-sum(yy_MU284))^2)/sim)/sum(yy_MU284);
}

srrmse_3mil<- as.character(round(as.numeric(srrmse_3mil),digits=5))

res_V_E_3mil<-(tot_V_E_3mil-sum(yy_MU284))/sum(yy_MU284);

colnames(res_V_E_3mil) <- c("Särndal","peor","regular","mejor","descendente")

#Gráfica de caja y bigote para los residuales
boxplot(data.frame(res_V_E_3mil), las=1, col="orange",main=c("Error rel. de la
     estimacion de RMT85(MU284)","Eliminación de Variables","3000 simulaciones"))

abline(0,0,col="red")
abline(0.02,0,col="blue",lty="dashed")
abline(-0.02,0,col="blue",lty="dashed")


# Cálculo del raíz del ECM rel., la D.E. rel y sesgo rel., para los residuales

error_std_V_E<-sesgo_rel_V_E<-ECM_rel_V_E<-""

ee_rel_V_E <- function(x){
   as.character(round(sqrt(sum((x-mean(x))^2)/sim)/tot_RMT85,5))
}
s_rel_V_E <- function(x){
  as.character(round((mean(x)-tot_RMT85)/tot_RMT85,5))
 }
raiz_ecm_r_V_E<- function(x){
  as.character(round(sqrt(sum((x-mean(x))^2)/sim+(mean(x)-tot_RMT85))/tot_RMT85,5))
}


for (i in 1:5)
{
error_std_V_E[i] <-     ee_rel_V_E(tot_V_E_3mil[,i])
sesgo_rel_V_E[i] <-      s_rel_V_E(tot_V_E_3mil[,i])
  ECM_rel_V_E[i] <- raiz_ecm_r_V_E(tot_V_E_3mil[,i])
}

# Leyendas y Etiquetas para la gráfica

legend(2.0, 0.03, "raíz del ECM relativo",cex=.8, bty="n",col="blue")

for (i in 1:5){
  legend(i-0.7, 0.018, ECM_rel_V_E[i],cex=.8, bty="n")
  legend(i-0.6, -0.008, sesgo_rel_V_E[i],cex=.8, bty="n")
 }
```

```
legend(2.0,-0.018, "Sesgo relativo",cex=.8, bty="n",col="red")

Tabla_errores_V_E<-rbind(ECM_rel_V_E,sesgo_rel_V_E,error_std_V_E)

> Tabla_errores_V_E
              [,1]      [,2]      [,3]      [,4]      [,5]
ECM_rel_V_E   "0.00838" "0.01304" "0.01298" "0.01232" "0.00862"
sesgo_rel_V_E "8e-04"   "0.00187" "0.00194" "0.00242" "0.00133"
error_std_V_E "0.00838" "0.01304" "0.01298" "0.01232" "0.00862"

write.csv(Tabla_errores_V_E, file = "C:\\ Tabla_errores_V_E.csv")

#  F I N
```

El siguiente es el código en R usado para generar y evaluar las simulaciones de muestreo para el **tercer** ejercicio con los datos ENIGH2002, correspondiente a la determinación del **Número Óptimo de Variables Auxiliares**.

```
####
#                   Simulación de datos de ENIGH2002
####

# Datos tomados del sitio web de INEGI (México)
#http://www.inegi.org.mx/est/contenidos/espanol/sistemas/enigh/bd/default.asp

##
#                      Técnica Componentes Principales
#                                Supervisadas
##

# Cargar Librería
  library(superpc)

# Fijar la semilla para obtener los mismos resultados de simulación
  set.seed(2007)

# Lectura de datos
  enigh2002<-read.csv("C:\\enigh2002.csv",header=T)


> dim(enigh2002)
[1] 17167   117

# Selección del subuniverso, en este caso el Distrito Federal
# Seleccionar un subconjunto (Entidad="9", D.F.)
  df2002<-enigh2002[substr(datos[,3],1,1)=="9",]

> dim(df2002)
[1] 1232  117


#                Quitar las siguientes variables

# Las variables fijas(básicas) desde la "1" hasta la "5"
# folio, hog, ubica_geo, estrato, conapo
#
# La variable de interés             ingmon(6) Ingreso corriente monetario
#
# Las que contienen muchos ceros
# clase_hog, ed_formal, sociedad
# cuasisoc, smg, coopera

   xij_df<-df2002[,-c(1:6,14,27,49,50,53,54,61)]


> df2002[1,c(1:6,14,27,49,50,53,54,61)]
folio hog ubica_geo estrato conapo ingmon clase_hog ed_formal sociedad cuasisoc
otros smg coopera
352 911090 1501    9012     1     NA  15830         2         8        0        0
0   0       0
dim(xij_df)
[1] 1232   104

# Tamaño de la población y tamaño muestra
  N<-dim(xij_df)[1]
  p<-dim(xij_df)[2]

# Construcción de la matriz de prueba
       xij_df_TEST <- matrix(0,N,p)
       dim(xij_df_TEST)
```

```
# Selección de la variable de interés
   yy_df<- df2002[,6]        # ENIGH2002. INGMON. Ingreso Corriente Monetario

# Centrar datos de la matriz X
  xij_df_c<-scale(xij_df,T)

# Asignar Etiquetas a los campos
  featurenamesEXA <- names(xij_df[1,])

# Seleccionar como muestra al 20% de la población

# Ver "The Elements of Statistical Learning" p.215
# En el caso de 200 observaciones usar validación cruzada de cinco dobleces
# five fold cross-validation
# para obtener conjuntos de entrenamiento
# de 160 observaciones
  sTEST<-sample(1:N,round(N/5,0))
  sTEST

  xij_df_c.testEXA <- xij_df_c[sTEST,]
  xij_df_c.testEXA
  yy_df_TEST<-yy_df[sTEST]
  yy_df_TEST


# Crear objeto que necesita superpc
  dataEXA<-list(x=t(xij_df_c),y=yy_df,featurenames= featurenamesEXA)


  dim(t(xij_df_c))
# Obs u individuos --> N=284
# Características  --> p=11
>   dim(t(xij_df_c))
[1]  104 1232

# Etiquetas del conjunto de prueba
featurenamesTEST<-featurenamesEXA
featurenamesTEST

# Crear un objeto que necesita la función "superpc"
dataTEST<-list(x=t(xij_df_c.testEXA),y=yy_df_TEST,featurenames= featurenamesTEST)
  dim(t(xij_df_c.testEXA))
>   dim(t(xij_df_c.testEXA))
[1] 104 246

# Regresión del conjunto de entrenamiento (1000 scores)
train.objEXA<- superpc.train(dataEXA, type="regression")


# Scores
scores.abs<-sort(abs(train.objEXA$feature.scores),decreasing=TRUE)
  scores.abs

  >   scores.abs
  [1] 27.8249228 23.4280144 20.5699805 19.8220322 19.7501624 17.4397435
  [7] 16.2281232 15.6900661 15.1701701 13.4263941 13.1542112 12.9896066
 [13] 12.8094043 12.3804411 12.2963991 12.0975934 11.6107160 11.4205184
 [19] 11.3208973 11.3104573 10.2250167 10.1215561 10.0339390  9.8416215
 [25]  9.8284307  9.7510152  9.4965255  9.1474450  9.1113210  8.9039802
 [31]  8.8088812  8.4801343  8.3628589  8.0688548  8.0281062  7.6880088
 [37]  7.6822107  7.5495761  6.5190931  6.3882224  6.3836460  6.0999921
 [43]  5.5760795  5.3963684  4.9835640  4.9648489  4.9060349  4.6819996
 [49]  4.6375423  4.5141227  4.5097026  4.3877546  4.2716715  4.2329874
 [55]  4.2132775  4.2032914  4.1124303  3.8930370  3.8610374  3.6769200
 [61]  3.5716137  3.5485483  3.4712468  3.4334615  3.1146761  3.0396673
 [67]  2.9796421  2.9070938  2.8924565  2.5330650  2.4100379  2.3700418
 [73]  2.3602939  2.2933066  2.1835404  2.1728283  2.1241083  2.1183542
 [79]  2.1174876  2.1162625  2.0944362  2.0136887  1.8324543  1.6675885
 [85]  1.6328821  1.6130323  1.5504262  1.5203098  1.4989572  1.4484053
 [91]  1.2274591  0.9412636  0.8034831  0.7823755  0.6689530  0.5626493
 [97]  0.5433453  0.3880910  0.3074604  0.1923432  0.1749608  0.1710187
[103]  0.1107983  0.1009125

# Ordenamiento de scores
```

```
scores.order<-order(abs(train.objEXA$feature.scores),decreasing=TRUE)
## Muestra del 20% en la entidad Distrito Federal
  scores.order

>   scores.order
[1]   4  14   5   6  17  34  79  39  43  31  38  37  91  88  78  90  32  73
[19] 72  36  89  93  99  40  48  23   2  55  84  59  80  70  87  33  64  16
[37] 18  75  81   1  15  58  41  69  86  57  66  54  76  77  92  52  71  30
[55] 35  67  28 102  49  29  63  44  56 104  24  68  26  42  61  65  97  11
[73]100 101  96  45  22   9   7  12  46  95  74  82  62  25   8  13  19  60
[91] 83  51  20  50  94  21  47   3  10  98 103  85  53  27


scores_enigh2002.order <- c(
     4, 14,  5,  6, 17, 34, 79, 39, 43, 31, 38, 37, 91, 88,78, 90, 32, 73,
    72, 36, 89, 93, 99, 40, 48, 23,  2, 55, 84, 59, 80, 70,87, 33, 64, 16,
    18, 75, 81,  1, 15, 58, 41, 69, 86, 57, 66, 54, 76, 77,92, 52, 71, 30,
    35, 67, 28,102, 49, 29, 63, 44, 56,104, 24, 68, 26, 42,61, 65, 97, 11,
   100,101, 96, 45, 22,  9,  7, 12, 46, 95, 74, 82, 62, 25, 8, 13, 19, 60,
    83, 51, 20, 50, 94, 21, 47,  3, 10, 98,103, 85, 53, 27
     )

# Reordenamiento de variables
  xij_enigh2002_df.order<-xij_df[,scores_enigh2002.order]

# Validación cruzada

# Divide (usando quantiles) el intervalo de coeficientes de regresión en 30 umbrales
# dentro del rango de coeficientes de regresión

cv30.objEXA<-superpc.cv(train.objEXA, dataEXA, n.fold=10,n.threshold=30)

# Gráfica del estadístico de prueba
# Para ENIGH2002 104 variables
par(mfrow=c(1,1))
superpc.plotcv(cv30.objEXA)

# Para ENIGH2002 104 variables yy=ingmon
# Cálculo del umbral óptimo (theta)
  umbral30<-cbind(cv30.objEXA$scor[1,],cv30.objEXA$thresholds)

# Theta óptimo
theta30<-umbral30[which(umbral30[,1]==max(umbral30[,1],na.rm=T)),2]
theta30[1]
> theta30[1]
[1] 13.94050

train.objEXA$feature.scores[train.objEXA$feature.scores>theta30[1]]
> train.objEXA$feature.scores[train.objEXA$feature.scores>theta30[1]]
[1] 27.82492 20.56998 19.82203 23.42801 19.75016 17.43974 15.69007 15.17017
[9] 16.22812



# Variables seleccionadas


#   xij_enigh2002_df_3 <- xij_enigh2002_df.order[,1:3]

#   xij_enigh2002_df_5 <- xij_enigh2002_df.order[,1:5]

#   xij_enigh2002_df_7 <- xij_enigh2002_df.order[,1:7]

#   xij_enigh2002_df_8 <- xij_enigh2002_df.order[,1:8]

#   xij_enigh2002_df_9 <- xij_enigh2002_df.order[,1:9]

#   xij_enigh2002_df_10 <- xij_enigh2002_df.order[,1:10]

#   xij_enigh2002_df_11 <- xij_enigh2002_df.order[,1:11]

  1 ingcor
  2 ingtot
  3 gasmon
```

```
  4 gascor
  5 gastot
  6 limpieza
  7 cuidados
  8 trabajo
  9 sueldos
 10 alimentos
 11 personal


##
#
#           Muestreo Balanceado
#
#             Metodo del CUBO
#
##

#############
##  Metodo de Eliminacion de Variables
#############

library(MASS)
library(lpSolve)
library(sampling)
#


# datos ENIGH2002 para las simulaciones
   xij_df_3.table<-as.data.frame(xij_enigh2002_df_3,optional=TRUE)
   xij_df_5.table<-as.data.frame(xij_enigh2002_df_5,optional=TRUE)
   xij_df_7.table<-as.data.frame(xij_enigh2002_df_7,optional=TRUE)
   xij_df_8.table<-as.data.frame(xij_enigh2002_df_8,optional=TRUE)
   xij_df_9.table<-as.data.frame(xij_enigh2002_df_9,optional=TRUE)
   xij_df_10.table<-as.data.frame(xij_enigh2002_df_10,optional=TRUE)
   xij_df_11.table<-as.data.frame(xij_enigh2002_df_11,optional=TRUE)


# Definicion de la matriz de variables de balanceo
  X_df.3<- as.matrix(xij_df_3.table)
  X_df.5<- as.matrix(xij_df_5.table)
  X_df.7<- as.matrix(xij_df_7.table)
  X_df.8<- as.matrix(xij_df_8.table)
  X_df.9<- as.matrix(xij_df_9.table)
  X_df.10<-as.matrix(xij_df_10.table)
  X_df.11<-as.matrix(xij_df_11.table)

# Calcular las probabilidades de inclusión respecto a la var gastot
pik=inclusionprobabilities(xij_df_11.table$gastot,50)


> dim(X)          # Se considera al Distrito Federal como el 10% de muestra
[1] 1232  104

  X24.1<-as.matrix(xij241.table)
  X24.2<-as.matrix(xij242.table)

   X6.1<-as.matrix(xij61.table)
   X6.2<-as.matrix(xij62.table)


# Simulaciones

sim=4000

  tot_df_3<-tot_df_5<-tot_df_7<-tot_df_8<-tot_df_9<-tot_df_10<-tot_df_11<-
  rep(0,times=sim)


for(i in 1:sim)

{

        s7_df<-samplecube(X_df.7,pik,1,FALSE,2)
```

```
        tot_df_7[i]=HTestimator(yy_df[s7_df==1],pik[s7_df==1])

        s8_df<-samplecube(X_df.8,pik,1,FALSE,2)
        tot_df_8[i]=HTestimator(yy_df[s8_df==1],pik[s8_df==1])

        s9_df<-samplecube(X_df.9,pik,1,FALSE,2)
        tot_df_9[i]=HTestimator(yy_df[s9_df==1],pik[s9_df==1])

        s10_df<-samplecube(X_df.10,pik,1,FALSE,2)
        tot_df_10[i]=HTestimator(yy_df[s10_df==1],pik[s10_df==1])

        s11_df<-samplecube(X_df.11,pik,1,FALSE,2)
        tot_df_11[i]=HTestimator(yy_df[s11_df==1],pik[s11_df==1])

  }


ENIGH_EV_peak_4000<-data.frame(cbind(tot_df_3,    tot_df_5,    tot_df_7,    tot_df_8,
tot_df_9, tot_df_10, tot_df_11))

write.csv(ENIGH_EV_peak_4000,        file        =        "C:\\Drive        D\\Jesus
Suarez\\Maestría\\Cuatrimestre         VI\\TESIS\\Resultados         de         las
simulaciones\\ENIGH_EV_peak_4000sim.csv")

ENIGH_EV_peak_4000<-read.csv("C:\\Drive    D\\Jesus    Suarez\\Maestría\\Cuatrimestre
VI\\TESIS\\Resultados de las simulaciones\\ENIGH_EV_peak_4000sim.csv",header=T)

tot_peak_4mil=cbind(tot_df_7, tot_df_8, tot_df_9, tot_df_10, tot_df_11)

tot_peak_4mil=cbind(ENIGH_EV_peak_4000[,4], ENIGH_EV_peak_4000[,5],
ENIGH_EV_peak_4000[,6], ENIGH_EV_peak_4000[,7], ENIGH_EV_peak_4000[,8])

#Raiz del ECM relativo

r_ECM_rel_4mil<-""
for (k in 1:5){
r_ECM_rel_4mil[k]=sqrt(sum((tot_peak_4mil[k]-sum(yy_df))^2)/sim)/sum(yy_df);
}
r_ECM_rel_4mil<- as.character(round(as.numeric(r_ECM_rel_4mil),digits=5))

res_4mil<-(tot_peak_4mil-sum(yy_df))/sum(yy_df);

colnames(res_4mil) <- c("EV_7var","EV_8var","EV_9var","EV_10var","EV_11var")

boxplot(data.frame(res_4mil), las=1,col="orange",main=c("Error rel. de la estimacion
de   INGMON(ENIGH2002)","Metodo   del   Cubo:   Eliminacion   de   variables","   4000
simulaciones"))

abline(0,0,col="red")
abline(0.06,0,col="blue",lty="dashed")
abline(-0.04,0,col="blue",lty="dashed")

error_std_peak<-sesgo_rel_peak<-r_ECM_rel_peak<-""

  ee <- function(x){
    as.character(round(sqrt(sum((x-mean(x))^2)/sim)/sum(yy_df),5))
  }

sr <- function(x){
   as.character(round((mean(x)-sum(yy_df))/sum(yy_df),5))

raiz_ecm_r_peak<- function(x){
  as.character(round(sqrt(sum((x-mean(x))^2)/sim+(mean(x)-sum(yy_df))/sum(yy_df),5))
}

> ENIGH_EV_peak_4000[1:3,]
  X tot_df_3 tot_df_5 tot_df_7 tot_df_8 tot_df_9 tot_df_10 tot_df_11
1 1        0        0 31668629 30870239 31123794  32563657  31899568
2 2        0        0 32464962 32189025 30931347  31689008  31067910
3 3        0        0 31634295 32530899 32324733  31754096  31423572

for (i in 1:5)
{
   error_std_peak[i] <- ee(ENIGH_EV_peak_4000[,i+3])
```

```
   sesgo_rel_peak[i] <- sr(ENIGH_EV_peak_4000[,i+3])
   r_ECM_rel_peak[i] <- raiz_ecm_r_peak(ENIGH_EV_peak_4000[,i+3])
}

Tabla_errores_LR_peak<-rbind(r_ECM_rel_4mil,sesgo_rel_peak,error_std_peak)

> Tabla_errores_LR_peak
               [,1]      [,2]      [,3]      [,4]      [,5]
r_ECM_rel_peak "0.01863" "0.01882" "0.0184"  "0.01849" "0.01876"
sesgo_rel_peak "0.00491" "0.00556" "0.00552" "0.00545" "0.00613"
error_std_peak "0.01863" "0.01882" "0.0184"  "0.01849" "0.01876"

write.csv(Tabla_errores_LR_peak, file = "C:\\Tabla_errores_LR_peak.csv")

legend(2, 0.04, "raiz del ECM relativo",cex=.8, bty="n",col="blue")

for (i in 1:5){
   legend(i-0.7, 0.03, r_ECM_rel_peak[i],cex=.8, bty="n")
   legend(i-0.7, -0.013, sesgo_rel_peak[i],cex=.8, bty="n")
 }

legend(2,-0.008, "sesgo relativo",cex=.8, bty="n",col="red")



#  F I N
```

El siguiente es el código en R usado para generar y evaluar las simulaciones de datos para el **cuarto** ejercicio con los datos ENIGH2002, correspondiente a **Varias Regiones** de México: Chiapas, Distrio Federal, Jalisco, Nuevo León y Puebla.

```
####
#                   Simulación de datos de ENIGH2002
#                            Varias Regiones
####

# Datos tomados del sitio web de INEGI (México)
#http://www.inegi.gob.mx/est/contenidos/espanol/sistemas/enigh/bd/default.asp

##
#                      Técnica Componentes Principales
#                               Supervisadas
##
Library(superpc)

# Fijar la semilla para obtener los mismos resultados de simulación

set.seed(2007)

# rm(list=ls(all=TRUE))   # Cleaning workarea

# Lectura de datos

  enigh2002<-read.csv("C:\\ enigh2002.csv",header=T)

> dim(enigh2002)
[1] 17167   117

chs2002<-enigh2002[enigh2002[,3]>=7000&enigh2002[,3]<8000,]
> dim(chs2002)
[1] 640 117

df2002<-enigh2002[enigh2002[,3]>=9000&enigh2002[,3]<10000,]
> dim(df2002)
[1] 1232  117

jal2002<-enigh2002[enigh2002[,3]>=14000&enigh2002[,3]<15000,]
>  dim(jal2002)
[1] 541 117

nln2002<-enigh2002[enigh2002[,3]>=19000&enigh2002[,3]<20000,]
> dim(nln2002)
[1] 420 117

pue2002<-enigh2002[enigh2002[,3]>=21000&enigh2002[,3]<22000,]
> dim(pue2002)
[1] 580 117

# De la tabla de datos, quitar las siguientes variables
# folio, hog, ubica_geo, estrato, conapo
# clase_hog, ed_formal, trabajo
# sociedad, cuasisoc, smg, coopera

   xij_chs<-chs2002[,-c(1:6,14,27,49,50,53,54,61)]
   xij_df <- df2002[,-c(1:6,14,27,49,50,53,54,61)]
   xij_jal<-jal2002[,-c(1:6,14,27,49,50,53,54,61)]
   xij_nln<-nln2002[,-c(1:6,14,27,49,50,53,54,61)]
   xij_pue<-pue2002[,-c(1:6,14,27,49,50,53,54,61)]
> df2002[1,c(1:6,14,27,49,50,53,54,61)]

     folio  hog ubica_geo estrato conapo ingmon clase_hog ed_formal sociedad
cuasisoc otros smg coopera
352 911090 1501      9012       1     NA  15830         2         8        0
0     0   0       0
> dim(xij_chs)
[1] 640 104
 dim(xij_df)
[1] 1232   104
> dim(xij_jal)
[1] 541 104
```

```
> dim(xij_nln)
[1] 420 104
> dim(xij_pue)
[1] 580 104


Saltar todo lo de componentes principales supervisadas
Ya que se va a usar el ordenamiento obtenido para DF
Al ejecutar comp. ppales. superv. para las cinco entidades en un solo conjunto
el programa arroja el siguiente error.
#####################################
#fold= 1
#1Error: no se puede ubicar un vector de tamaño  1.8 Gb
#Además: Warning messages:
#1: In matrix(0, p, p) :
#  Reached total allocation of 1014Mb: see help(memory.size)
#####################################

Del código de la simulación anterior, brincar hasta el_ordenamiento scores.order
(correr yy_chs)


# Tamaño de la población y tamaño muestra

  N<-dim(xij_df)[1]
  p<-dim(xij_df)[2]


# Construcción de la matriz de prueba

        xij_df_TEST <- matrix(0,N,p)
        dim(xij_df_TEST)

# Selección de la variable de interés

   yy_df <-  df2002[,6]
   yy_jal<- jal2002[,6]
   yy_nln<- nln2002[,6]
   yy_pue<- pue2002[,6]


>   dim(t(xij_df_c))
[1]  104 1232

        # Etiquetas del conjunto de prueba
#featurenamesTEST<-featurenamesEXA[sTEST]

>   dim(t(xij_df_c.testEXA))
[1] 104 246

# Para evitar ejecutar nuevamente, se capturan los scores obtenidos en la simulación
anterior
   scores.abs <- c(
      27.8249228, 23.4280144, 20.5699805, 19.8220322, 19.7501624, 17.4397435,
      16.2281232, 15.6900661, 15.1701701, 13.4263941, 13.1542112, 12.9896066,
      12.8094043, 12.3804411, 12.2963991, 12.0975934, 11.6107160, 11.4205184,
      11.3208973, 11.3104573, 10.2250167, 10.1215561, 10.0339390,  9.8416215,
       9.8284307,  9.7510152,  9.4965255,  9.1474450,  9.1113210,  8.9039802,
       8.8088812,  8.4801343,  8.3628589,  8.0688548,  8.0281062,  7.6880088,
       7.6822107,  7.5495761,  6.5190931,  6.3882224,  6.3836460,  6.0999921,
       5.5760795,  5.3963684,  4.9835640,  4.9648489,  4.9060349,  4.6819996,
       4.6375423,  4.5141227,  4.5097026,  4.3877546,  4.2716715,  4.2329874,
       4.2132775,  4.2032914,  4.1124303,  3.8930370,  3.8610374,  3.6769200,
       3.5716137,  3.5485483,  3.4712468,  3.4334615,  3.1146761,  3.0396673,
       2.9796421,  2.9070938,  2.8924565,  2.5330650,  2.4100379,  2.3700418,
       2.3602939,  2.2933066,  2.1835404,  2.1728283,  2.1241083,  2.1183542,
       2.1174876,  2.1162625,  2.0944362,  2.0136887,  1.8324543,  1.6675885,
       1.6328821,  1.6130323,  1.5504262,  1.5203098,  1.4989572,  1.4484053,
       1.2274591,  0.9412636,  0.8034831,  0.7823755,  0.6689530,  0.5626493,
       0.5433453,  0.3880910,  0.3074604,  0.1923432,  0.1749608,  0.1710187,
       0.1107983,  0.1009125
       )
# Verificación
  scores.abs
```

```
  >   scores.abs
  [1] 27.8249228 23.4280144 20.5699805 19.8220322 19.7501624 17.4397435
  [7] 16.2281232 15.6900661 15.1701701 13.4263941 13.1542112 12.9896066
 [13] 12.8094043 12.3804411 12.2963991 12.0975934 11.6107160 11.4205184
 [19] 11.3208973 11.3104573 10.2250167 10.1215561 10.0339390  9.8416215
 [25]  9.8284307  9.7510152  9.4965255  9.1474450  9.1113210  8.9039802
 [31]  8.8088812  8.4801343  8.3628589  8.0688548  8.0281062  7.6880088
 [37]  7.6822107  7.5495761  6.5190931  6.3882224  6.3836460  6.0999921
 [43]  5.5760795  5.3963684  4.9835640  4.9648489  4.9060349  4.6819996
 [49]  4.6375423  4.5141227  4.5097026  4.3877546  4.2716715  4.2329874
 [55]  4.2132775  4.2032914  4.1124303  3.8930370  3.8610374  3.6769200
 [61]  3.5716137  3.5485483  3.4712468  3.4334615  3.1146761  3.0396673
 [67]  2.9796421  2.9070938  2.8924565  2.5330650  2.4100379  2.3700418
 [73]  2.3602939  2.2933066  2.1835404  2.1728283  2.1241083  2.1183542
 [79]  2.1174876  2.1162625  2.0944362  2.0136887  1.8324543  1.6675885
 [85]  1.6328821  1.6130323  1.5504262  1.5203098  1.4989572  1.4484053
 [91]  1.2274591  0.9412636  0.8034831  0.7823755  0.6689530  0.5626493
 [97]  0.5433453  0.3880910  0.3074604  0.1923432  0.1749608  0.1710187
[103]  0.1107983  0.1009125


#  scores.order<-order(abs(train.objEXA$feature.scores),decreasing=TRUE)
  scores.order


## Muestra del 10% en el DF      sTEST<-sample(1:N,round(N/10,0))
>   scores.order
  [1]   4  14   5   6  17  34  79  39  43  31  38  37  91  88  78  90  32  73
 [19]  72  36  89  93  99  40  48  23   2  55  84  59  80  70  87  33  64  16
 [37]  18  75  81   1  15  58  41  69  86  57  66  54  76  77  92  52  71  30
 [55]  35  67  28 102  49  29  63  44  56 104  24  68  26  42  61  65  97  11
 [73] 100 101  96  45  22   9   7  12  46  95  74  82  62  25   8  13  19  60
 [91]  83  51  20  50  94  21  47   3  10  98 103  85  53  27

## Muestra del 20% en el DF      sTEST<-sample(1:N,round(N/5,0))
>   scores.order
  [1]   4  14   5   6  17  34  79  39  43  31  38  37  91  88  78  90  32  73
 [19]  72  36  89  93  99  40  48  23   2  55  84  59  80  70  87  33  64  16
 [37]  18  75  81   1  15  58  41  69  86  57  66  54  76  77  92  52  71  30
 [55]  35  67  28 102  49  29  63  44  56 104  24  68  26  42  61  65  97  11
 [73] 100 101  96  45  22   9   7  12  46  95  74  82  62  25   8  13  19  60
 [91]  83  51  20  50  94  21  47   3  10  98 103  85  53  27

# Para evitar ejecutar nuevamente, se captura manualmente
   scores.order <-c(
       4,  14,  5,   6,  17,  34,  79,  39,  43,  31,  38,  37,  91,  88,  78,  90,  32,  73,
      72,  36,  89,  93,  99,  40,  48,  23,   2,  55,  84,  59,  80,  70,  87,  33,  64,  16,
      18,  75,  81,   1,  15,  58,  41,  69,  86,  57,  66,  54,  76,  77,  92,  52,  71,  30,
      35,  67,  28, 102,  49,  29,  63,  44,  56, 104,  24,  68,  26,  42,  61,  65,  97,  11,
     100, 101,  96,  45,  22,   9,   7,  12,  46,  95,  74,  82,  62,  25,   8,  13,  19,  60,
      83,  51,  20,  50,  94,  21,  47,   3,  10,  98, 103,  85,  53,  27
       )


#  xij_chs.order<-xij_chs[,scores.order]

  xij_chs.order<-xij_chs[,scores.order]
  xij_df.order <- xij_df[,scores.order]
  xij_jal.order<-xij_jal[,scores.order]
  xij_nln.order<-xij_nln[,scores.order]
  xij_pue.order<-xij_pue[,scores.order]

# Brincar validacion cruzada
# Y pasar a las variables seleccionadas (umbrales)


       # Theta óptimo
#      theta30<-umbral30[which(umbral30[,1]==max(umbral30[,1],na.rm=T)),2]
       theta30[1]

       > theta30[1]
[1] 13.94050

#      train.objEXA$feature.scores[train.objEXA$feature.scores>theta30[1]]
```

```
  > train.objEXA$feature.scores[train.objEXA$feature.scores>theta30[1]]
[1] 27.82492 20.56998 19.82203 23.42801 19.75016 17.43974 15.69007 15.17017
[9] 16.22812

# Variables seleccionadas

#  umbrales_sel_df<-umbral30[which(umbral30[,2]<=theta30[1])]
#  umbrales_sel_df

 >   umbrales_sel_df
 [1]  631.2859  643.1844  646.9021  680.6617  700.7425  722.1821  726.3637
 [8]  727.0753  729.9288  745.9520  745.6433  758.8271  755.6447  763.5888
[15]  728.6007  752.6958  712.8236  732.6832  816.8076  892.8887  936.8109
[22] 1014.5390 1117.0657 1144.4116

# Para evitar ejecutar nuevamente, se captura manualmente
   umbrales_sel_df <- c(
      631.2859,  643.1844,  646.9021,  680.6617,  700.7425,  722.1821,  726.3637,
      727.0753,  729.9288,  745.9520,  745.6433,  758.8271,  755.6447,  763.5888,
      728.6007,  752.6958,  712.8236,  732.6832,  816.8076,  892.8887,  936.8109,
     1014.5390, 1117.0657, 1144.4116
     )

# Brincar hasta el filtrado de las 9 var. seleccionadas


   xij_chs_9 <- xij_chs.order[,1:9]
   xij_df_9  <-  xij_df.order[,1:9]
   xij_jal_9 <- xij_jal.order[,1:9]
   xij_nln_9 <- xij_nln.order[,1:9]
   xij_pue_9 <- xij_pue.order[,1:9]


  1 ingcor
  2 ingtot
  3 gasmon
  4 gascor
  5 gastot
  6 limpieza
  7 cuidados
  8 trabajo
  9 sueldos
 10 alimentos
 11 personal
 12 educacion
 13 cuidado
 14 comunica
 15 energia
 16 esparci
 17 vestido_c
 18 ves_3ymas
 19 vestido
 20 transporte
 21 educa
 22 transfe
 23 pago_tar
 24 negocio


#############
#
#           Balanced Sampling
#
#             CUBE Method
#
#############
# Selection of a balanced sample using the MU284 population,
# simulation and comparison of the variance with
# unequal probability sampling of fixed sample size.
#############

#############
##  Linear Programming
```

```
#############

#Ningún proceso

#############
##  Supression of variables method
#############

  library(MASS)           # Internal MASS functions
  library(lpSolve)        # Linear Programming
  library(sampling)


  xij_chs_9.table<-as.data.frame(xij_chs_9,optional=TRUE)
  xij_df_9.table<-as.data.frame(xij_df_9,optional=TRUE)
  xij_jal_9.table<-as.data.frame(xij_jal_9,optional=TRUE)
  xij_nln_9.table<-as.data.frame(xij_nln_9,optional=TRUE)
  xij_pue_9.table<-as.data.frame(xij_pue_9,optional=TRUE)

# inclusion probabilities compute


# Calcular las probabilidades de inclusión respecto a la var gastot
# pik=inclusionprobabilities(xij.table$P75,50)

NOTA.
Las pik son distintas para cada entidad

#inclusionprobabilities(a,n)
# pik=inclusionprobabilities(xij.order$gastot,50)

pik_chs=inclusionprobabilities(xij_chs.order$gastot,50)
 pik_df=inclusionprobabilities( xij_df.order$gastot,50)
pik_jal=inclusionprobabilities(xij_jal.order$gastot,50)
pik_nln=inclusionprobabilities(xij_nln.order$gastot,50)
pik_pue=inclusionprobabilities(xij_pue.order$gastot,50)


# Definition of the matrix of balancing variables


 X_chs.9<-as.matrix(xij_chs_9.table)
 X_df.9 <-as.matrix( xij_df_9.table)
 X_jal.9<-as.matrix(xij_jal_9.table)
 X_nln.9<-as.matrix(xij_nln_9.table)
 X_pue.9<-as.matrix(xij_pue_9.table)


   dim(X)
> dim(X)          # Distrito Federal 10% de muestra
[1] 1232  104

  sim=5000
tot_chs_9var_4mil<-tot_df_9var_4mil<-tot_jal_9var_4mil<-tot_nln_9var_4mil<-
tot_pue_9var_4mil<-rep(0,times=sim)


  for(i in 1:sim)
  {
    cat("Simulation number ",i,"\n")

# Computation of the Horvitz-Thompson estimator for a balanced sample

    s9_chs<-samplecube(X_chs.9,pik_chs,1,FALSE,2)
  tot_chs_9var_4mil[i]<-HTestimator(yy_chs[s9_chs==1],pik_chs[s9_chs==1])

    s9_df<-samplecube(X_df.9,pik_df,1,FALSE,2)
  tot_df_9var_4mil[i]<-HTestimator(yy_df[s9_df==1],pik_df[s9_df==1])

    s9_jal<-samplecube(X_jal.9,pik_jal,1,FALSE,2)
  tot_jal_9var_4mil[i]<-HTestimator(yy_jal[s9_jal==1],pik_jal[s9_jal==1])

    s9_nln<-samplecube(X_nln.9,pik_nln,1,FALSE,2)
  tot_nln_9var_4mil[i]<-HTestimator(yy_nln[s9_nln==1],pik_nln[s9_nln==1])
```

```
    s9_pue<-samplecube(X_pue.9,pik_pue,1,FALSE,2)
  tot_pue_9var_4mil[i]<-HTestimator(yy_pue[s9_pue==1],pik_pue[s9_pue==1])


  }

tot_chs_9var_5mil<-tot_chs_9var_4mil
tot_df_9var_5mil<-tot_df_9var_4mil
tot_jal_9var_5mil<-tot_jal_9var_4mil
tot_nln_9var_5mil<-tot_nln_9var_4mil
tot_pue_9var_5mil<-tot_pue_9var_4mil

ENIGH_EV_regiones_5000<-data.frame(cbind(tot_chs_9var_5mil, tot_df_9var_5mil,
tot_jal_9var_5mil, tot_nln_9var_5mil, tot_pue_9var_5mil))

write.csv(ENIGH_EV_regiones_5000, file = "C:\\ENIGH_EV_regiones_5000sim.csv")
t_chs<-sum(yy_chs)
 t_df<-sum(yy_df)
t_jal<-sum(yy_jal)
t_nln<-sum(yy_nln)
t_pue<-sum(yy_pue)


 par(mfrow=c(1,5))

 hist(res_chs_9-t_chs,col="orange")
 hist(res_df_9- t_df, col="orange")
 hist(res_jal_9-t_jal,col="orange")
 hist(res_nln_9-t_nln,col="orange")
 hist(res_pue_9-t_pue,col="orange")


tot_regiones_5mil=cbind(tot_chs_9var_5mil, tot_df_9var_5mil, tot_jal_9var_5mil,
tot_nln_9var_5mil, tot_pue_9var_5mil)



     res_5mil_chs<-(tot_regiones_5mil[,1]-t_chs)/t_chs;
     res_5mil_df<-(tot_regiones_5mil[,2]-t_df)/t_df;
     res_5mil_jal<-(tot_regiones_5mil[,3]-t_jal)/t_jal;
     res_5mil_nln<-(tot_regiones_5mil[,4]-t_nln)/t_nln;
     res_5mil_pue<-(tot_regiones_5mil[,5]-t_pue)/t_pue;


r_ECM_rel_regiones_5mil<-
as.character(round(as.numeric(r_ECM_rel_regiones_5mil),digits=4))


res_5mil<-cbind(res_5mil_chs, res_5mil_df, res_5mil_jal,
                          res_5mil_nln, res_5mil_pue)
   colnames(res_5mil) <- c("CHS_9var","DF_9var","JAL_9var","NLN_9var","PUE_9var")


boxplot(data.frame(res_5mil), las=1,col="orange",main=c("Error rel. de la estimación
de INGMON (ENIGH2002)","Eliminación de variables","5000 simulaciones"))

  abline(0,0,col="red")
  abline(0.05,0,col="blue",lty="dashed")
  abline(-0.05,0,col="blue",lty="dashed")

tot_reg<-c(sum(yy_chs),sum(yy_df),sum(yy_jal),sum(yy_nln),sum(yy_pue))

error_std_regiones<-sesgo_rel_regiones<-r_ECM_rel_regiones<-""

  ee <- function(x,y){
        as.character(round(sqrt(sum((x-mean(x))^2)/sim)/y,5))
        }
  sr <- function(x,y){
        as.character(round((mean(x)-y)/y,5))
        }
  raiz_ecm_r<- function(x,y){
        as.character(round(sqrt(sum((x-mean(x))^2)/sim+(mean(x)-y))/y,5))
        }
```

```
for (i in 1:5)
{
  error_std_regiones[i] <- ee(tot_regiones_5mil[,i],tot_reg[i])
  sesgo_rel_regiones[i] <- sr(tot_regiones_5mil[,i],tot_reg[i])
  r_ECM_rel_regiones[i] <- raiz_ecm_r(tot_regiones_5mil[,i],tot_reg[i])
}

Tabla_errores_regiones<-
rbind(r_ECM_rel_regiones,sesgo_rel_regiones,error_std_regiones)

write.csv(Tabla_errores_regiones, file = "C:\\Tabla_errores_regiones.csv")

> Tabla_errores_regiones
                    [,1]      [,2]      [,3]      [,4]      [,5]
r_ECM_rel_regiones "0.02668" "0.019"   "0.01328" "0.01076" "0.01189"
sesgo_rel_regiones "0.00631" "0.00565" "0.00399" "0.00294" "0.00375"
error_std_regiones "0.02668" "0.019"   "0.01328" "0.01076" "0.01189"

legend(2.5, 0.07, "raiz del ECM relativo",cex=.8, bty="n",col="blue")

for (i in 1:5){
    legend(i-0.6, 0.03, Tabla_errores_regiones[1,i],cex=.8, bty="n")
    legend(i-0.6, -0.01, Tabla_errores_regiones[2,i],cex=.8, bty="n")
 }

legend(2.5,-0.05, "Sesgo Relativo",cex=.8, bty="n",col="red")


# Grafica de los scores

plot(scores.abs[scores.abs>theta30],type="b",main="ENIGH2002. Scores por
variable",xlab="Variable",ylab="Scores")


#  F I N
```

Apéndice C

Descripción de variables de los datos ENIGH2002

La base de datos de la ENIGH2002 está conformada por seis tablas que contienen toda la información captada en el levantamiento. A continuación se detalla el nombre de estas tablas, el número de registros y su contenido:

Tabla C1. Lista de archivos que contienen los datos y metadatos de la encuesta ENIGH2002.

Nombre	Registros	Contenido
HOGARES.DBF	17 167	Características de los hogares, de las viviendas que habitan y el factor de expansión.
POBLACION.DBF	72 602	Características socio demográficas y ocupacionales de los miembros de los hogares.
INGRESOS.DBF	56 980	Ingresos y percepciones de capital de cada uno de los miembros de los hogares.
GASTOS.DBF	1 029 761	Gastos realizados por el hogar.
EROGACIONES.DBF	7 651	Erogaciones de capital por hogar.
NOMONETARIO.DBF	138 015	Gastos o ingresos realizados por hogar y algunos por persona.

Para obtener cualquier tipo de información se requiere expandirla, esto es multiplicar el valor de la variable en estudio por el factor de expansión, por lo que generalmente es necesario "pegar" el factor a cualquiera de las otras tablas.

Tabla C2. Descripción de variables de los datos de hogares de ENIGH2002.

Núm.	Nombre	Descripción
1	FOLIO	Identificador del hogar
2	HOG	Factor de expansión
3	ESTRATO	Estrato (Tamaño de localidad)
4	UBICA_GEO	Ubicación geográfica (entidad, municipio)
5	TENENCIA	Tenencia
6	CLASE_HOG	Clase de hogar
7	EDAD	Edad del jefe del hogar
8	ED_FORMAL	Nivel de instrucción del jefe del hogar
9	TAM_HOGAR	Población total
10	HOMBRES	Población total hombre
11	MUJERES	Población total mujeres
12	TOT_RESI	Miembros del hogar
13	TOT_HOM	Miembros del hogar hombres
14	TOT_MUJ	Miembros del hogar mujeres
15	MAYORES	Miembros del hogar mayores
16	MENORES	Miembros del hogar menores
17	P0A11	Miembros del hogar de 0 a 11 años
18	P12_64	Miembros del hogar de 12 a 64 años
19	P65MAS	Miembros del hogar de 65 años y más
20	N_OCUP	Número de ocupados
21	PERING	Perceptores de ingreso corriente monetario
22	PEROCU	Perceptores ocupados y con ingreso corriente monetario
23	INGTOT	Ingreso total
24	INGCOR	Ingreso corriente total
25	INGMON	Ingreso corriente monetario
26	TRABAJO	Remuneraciones al trabajo asalariado
27	SUELDOS	Sueldos y horas extras
28	OTRA_REM	Otras remuneraciones
29	NEGOCIO	Renta empresarial
30	NO_AGROP	Negocios no agropecuarios
31	AGROPE	Negocios agropecuarios
32	COOPERA	Cooperativa de producción
33	SOCIE	Ingresos de sociedades
34	RENTAS	Renta de la propiedad
35	TRANSFER	Transferencias
36	JUBILA	Jubilaciones y pensiones
37	BECA_DON	Becas y donativos
38	DONATIVO	Regalos o donativos en dinero provenientes de otros hogares
39	OTROS	Otros ingresos corrientes
40	GASNOM	Ingreso corriente no monetario
41	AUTO	Autoconsumo
42	PAGO	Pago en especie
43	REGA	Regalos
44	ESTI	Estimación del alquiler de la vivienda
45	PERTOT	Percepciones financieras y de capital totales
46	PERMON	Percepciones financieras y de capital monetarias

47	RETIRO	Retiro de inversiones, tandas, cajas de ahorro, etcétera.
48	PRESTAMO	Préstamos recibidos de personas no miembros del hogar o instituciones, excluye préstamos hipotecarios.
49	OTRAS_PR	Otras percepciones
50	ERONOM	Percepciones financieras y de capital no monetarias.
51	GASTOT	Gasto total
52	GASCOR	Gasto corriente total
53	GASMON	Gasto corriente monetario
54	ALIMENTOS	Alimentos y bebidas consumidas dentro y fuera del hogar y tabaco
55	ALI_DENT	Alimentos y bebidas consumidas dentro del hogar.
56	CEREALES	Cereales
57	CARNES	Carnes
58	PESCADO	Pescados y mariscos
59	LECHE	Leche y sus derivados
60	HUEVO	Huevo
61	ACEITES	Aceites
62	TUBÉRCULO	Tubérculo
63	VERDURAS	Verduras
64	FRUTAS	Frutas
65	AZUCAR	Azúcar
66	CAFÉ	Café
67	ESPECIAS	Especias
68	OTRO_ALI	Otros alimentos
69	BEBIDAS	Bebidas
70	FUERA_HOG	Alimentos y bebidas consumidas fuera del hogar
71	TABACO	Tabaco
72	VESTIDO	Vestido y calzado
73	VESTIDO_C	Vestido
74	VES_3YMAS	Vestido para personas de 4 años y más
75	VES_3MEN	Vestido para personas de 0 a 3 años
76	CALZADO	Calzado y su reparación
77	VIVIENDA	Vivienda, servicios de conservación, energía eléctrica y combustibles
78	ALQUILER	Alquileres brutos, impuestos, predial y cuotas por servicio de conservación
79	AGUA	Agua
80	ENERGIA	Energía eléctrica y combustibles
81	LIMPIEZA	Artículos y servicios para la limpieza y cuidados de la casa, enseres domésticos, muebles, cristalería, utensilios domésticos y blancos.
82	CUIDADOS	Artículos y servicios para la limpieza y cuidados de la casa
83	ENSERES	Enseres domésticos y muebles, cristalería, utensilios domésticos y blancos.
84	SALUD	Cuidados médicos y conservación de la salud
85	ATEN_PRI	Atención primaria o ambulatoria
86	HOSPITAL	Atención hospitalaria, servicios médicos y medicamentos durante el embarazo y parto, aparatos ortopédicos, terapéuticos y seguros médicos.
87	MEDICA	Medicamentos sin receta
88	TRANSPORTE	Transporte, adquisición, mantenimiento y accesorios para vehículos y comunicaciones
89	TRANS_AD	Transporte, adquisición, mantenimiento y accesorios para vehículos
90	PUBLICO	Transporte público
91	FORANEO	Transporte foráneo terrestre, aéreo y otros servicios especiales de transporte
92	VEHICULO	Adquisición de vehículos de uso particular, mantenimiento y accesorios para vehículos

93	COMUNICA	Comunicaciones
94	EDUCACION	Servicios y artículos de educación y esparcimiento, paquetes turísticos y para fiestas, hospedaje y alojamiento
95	EDUCA	Artículos y servicios de educación
96	ESPARCI	Artículos y servicios de esparcimiento, paquetes para: fiestas, turísticos, hospedaje y alojamiento
97	PERSONAL	Artículos y servicios para el cuidado personal, accesorios y efectos personales, otros gastos diversos y transferencias
98	CUIDADO	Artículos y servicios para el cuidado personal
99	ACCESORIO	Accesorios y efectos personales
100	TRANSFE	Otros gastos diversos y transferencias
101	EROTOT	Erogaciones financieras y de capital totales
102	EROMON	Erogaciones financieras y de capital monetarias
103	CUOTA	Cuota pagada por la vivienda propia
104	MATERIAL	Materiales para reparación, mantenimiento y/o ampliación de la vivienda
105	SERVICIO	Servicios de reparación, mantenimiento y/o ampliación de la vivienda
106	DEPOSITO	Depósito de cuenta de ahorros, tandas, cajas de ahorro, etc.
107	TERCEROS	Préstamos a terceros
108	PAGO_TAR	Pago por tarjeta de crédito al banco o casa comercial
109	DEUDAS	Pago de deudas a empresa donde trabaja y/o a otras personas o instituciones
110	MONEDAS	Compra de monedas, metales preciosos, joyas y obras de arte
111	CASAS	Compra de casas, condominios, locales y terrenos
112	BALANCE	Balance negativo en negocios propiedad del hogar
113	OTRA_ERO	Otras erogaciones
114	SMG	Salarios mínimos generales

REFERENCIAS

ALTER, O., BROWN, P.O., AND BOSTEIN, D. (2000) Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling. *Academy of Science*, **97**, 10101-10106.

ABDI, H. (2007). Partial least squares regression. *Encyclopedy of Measurement and Statistics*. California: SAGE .

BAIR et al. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*. **101**, 119–137.

BAIR, E. AND TIBSHIRANI, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*. **2**(4): 511-522.

BREIMAN, L. (2001). Statistical modeling: The two cultures. *Statistical Science*. **16**, 199-231.

DALLAL, G. (2001). The little handbook of statistical practice. Multiple Linear Regression. Which predictors are more important? *http://www.tufts.edu/~gdallal/LHSP.HTM*.

DEVILLE, J.-C. AND TILLÉ, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*. **91**, 893–912.

DEVILLE, J.-C. AND SARNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association,* **87**, 376-382.

HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

INEGI. (2002). *Manual para la Elaboración de Conteos, Listados y Actualización Cartográfica del Marco Nacional de Viviendas 2001*. Aguascalientes, México: INEGI.

NEYMAN, JERZY. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, **97**, 558-625

NGUYEN, D., AND ROCKE, D. (2002). Partial least squares proportional hazard regression for application to DNA microarrays. *Bioinformatics*, **18**, 1625-1632.

NGUYEN, D., AND ROCKE, D. (2004). On partial least squares dimension reduction for microarray-based classification: a simulation study. *Computational Statistics & Data Analysis*, **46**, 407-425

PEÑA, D. (2002). *Análisis de datos multivariantes.* Madrid: McGraw-Hill.

RICE, JOHN A. (1995). *Mathematical statistics and data analysis*. 2nd Edition. International Thomson Publishing.

ROYAL, R. Y HERSON, J. (1973). Robust estimation in finite populations I. *J. Am. Statist. Assoc.* **68**, 880–9.

SCOTT, A.J., K.R.W. BREWER, AND E.W.H. HO. (1978). Finite population sampling and robust estimation. *Journal of the American Statistical Association*, **73**, 359-361.

SÄRNDAL, C.-E., SWENSSON, B. AND WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.

TILLÉ,Y. (2006). *Sampling Algorithms*. New York: Springer.

VALLIANT, R., DORFMAN, A. H. AND ROYAL, R. M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons.

WINSTON, W. L. (2005). *Investigación de Operaciones. Aplicaciones y Algoritmos*. 4a. Edición. International Thomson Editores.

ZHU, M. AND HASTIE, T. (2003). Feature extraction for non-parametric discriminant analysis. *Journal of Computational and Graphical Statistics,* **12**(1), 101-120.

"lpSolve"

Michel Berkelaar and others (2006). lpSolve: Interface to Lp_solve v.5.5 to solve linear/integer programs. R package version 5.5.7.

"MASS"

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

"R"

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

"sampling"

Yves Tillé and Alina Matei (2007). sampling: Survey Sampling. R package version 0.9.

"superpc"

Eric Bair and R. Tibshirani (). superpc: Supervised principal components. R package version 1.03. http://www-stat.stanford.edu/~tibs/superpc

ÍNDICE

U

V