CIMAT

Centro de Investigación en Matemáticas, A.C.

Estructuras con variables latentes como alternativa para modelar la autoselección

T E S I S

Que para obtener el grado de

Maestro en Ciencias con Especialidad en Probabilidad y Estadística

P r e s e n t a

Oscar Santiago Sánchez

Director de Tesis:

Dr. Rogelio Ramos Quiroga

Guanajuato, Gto. Julio de 2012

ÍNDICE TEMÁTICO

INTRODUCCIÓN

Uno de los principales problemas que aborda la Estadística es el establecimiento de causalidades. Este es un tema delicado en diversas áreas, desde las ciencias sociales (como la economía), hasta las ciencias ambientales, urbanas y de la salud (como la psicología). Es bien sabido que antes de establecer una relación causal entre dos variables, debe identificarse una correlación entre las mismas. La Estadística permite, de forma relativamente sencilla, identificar correlaciones que no sean obra de la casualidad. Ahora, una vez identificadas relaciones *estadísticamente significativas*, explorar una relación causal no es tarea sencilla.

Esta tesis surge con la necesidad primordial de estudiar el problema de autoselección en econometría, el cual se aborda en breve. En el estudio de dicho problema, y su solución, surgen diversos conceptos, así como otros problemas relacionados con la causalidad, que dan pie a la exploración de distintas metodologías que tienen como objetivo establecer relaciones causales. En esta investigación brindamos una introducción al conocimiento de herramientas tradicionales utilizadas para esta tarea.

Procederemos revisando conceptos teóricos de modelación y de estimación. Por la extensa variedad de temas estudiados, sólo brindaremos una perspectiva general de los mismos. Con lo desarrollado en esta tesis, el problema de autoselección podrá ser entendido, modelado y resuelto con técnicas estadísticas probadas. Sin embargo, el potencial de las técnicas estudiadas nos obligará, en ocasiones, a no sólo hablar en términos del problema de selección, sino a mencionar otros problemas que pueden ser resueltos con dichas técnicas.

Los conceptos que surgirán en esta investigación son, principalmente, variables latentes, modelos de elección, modelación con sistemas de ecuaciones estructurales, métodos de simulación de Montecarlo (muestreador de Gibbs) y el algoritmo *Expectation Maximization*.

La estructura de este trabajo se describe enseguida. En el segundo capítulo se introduce el problema de autoselección y se brinda un marco teórico para el estudio del mismo. Enseguida se brindan estrategias de modelación basadas en sistemas de ecuaciones estructurales. El cuarto capítulo versa sobre el algoritmo *Monte Carlo*

Expectation Maximization, enfocándose en éste como estrategia de estimación para los modelos del tercer capítulo. Finalmente, se exponen las conclusiones, así como algunas consideraciones finales relevantes.

Los aspectos computacionales tratados en esta tesina se desarrollan con ayuda del software R. Los códigos pertinentes están disponibles mediante petición expresa al autor de esta tesis, a la dirección osantiago@cimat.mx.

II. PROBLEMA DE SELECCIÓN: CONCEPTOS Y TEORÍA

En este capítulo introduciremos conceptualmente el problema de autoselección. Comenzamos en una primera sección motivando los conceptos subyacentes en la modelación del problema. En segunda instancia, esquematizamos estadísticamente el problema y comentamos métodos de estimación usuales para modelos sencillos que lo abordan; en particular mencionaremos los métodos de mínimos cuadrados en dos etapas, que son clásicos en el tratamiento del problema de autoselección. Finalmente, introducimos el concepto de variables latentes a partir de modelos de utilidad, lo que dará pauta para estudiar al modelo probit; en esta última sección hablaremos de problemas de identificabilidad y estimación por máxima verosimilitud, tópicos que volverán a tratarse en capítulos y modelos posteriores.

El objetivo primordial en este capítulo es que el lector se familiarice con los conceptos y la problemática del modelo de autoselección, además de identificar problemas estadísticos que deben ser resueltos mediante modelos que pueden resultar no convencionales. Además, se introducen y encausan las estrategias de modelación y estimación que se detallarán en capítulos posteriores.

A. EJEMPLO INTRODUCTORIO

Como mencionamos en la introducción de este capítulo, antes de establecer en forma explícita el problema de autoselección, podemos motivarlo mediante el siguiente ejemplo.

Consideremos una muestra (representativa de la población) de n profesores. Supongamos que a los profesores se les aplica un examen con el fin de incrementar/disminuir su salario, en función del resultado que obtengan en la prueba. Consideramos que tomar o no el examen es decisión voluntaria de los profesores. Suponga que se desea medir la influencia de cierta característica x_1 en el resultado del examen (que denotamos con la variable aleatoria y); además interesa saber cuál es el resultado que, en promedio, obtienen los profesores que presentan un nivel de $x_1 = 0$ (esto es de interés si, por ejemplo, en promedio $x_1 = 0$).

Se propone el siguiente modelo:

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \qquad (1.1)$$

$$i = 1,2, \dots, n.$$

En el modelo anterior, x_{1i} representa el nivel de la variable x_1 del i-ésimo profesor, mientras que ε_i son características no observadas del mismo individuo, cuya distribución asumimos normal centrada e independiente de x_{1i}. Además asumimos que ε_i es independiente de ε_j. En el contexto de este ejemplo, considere que x_1 es una medida de los años de experiencia. Puede pensarse también que el término ε captura el nivel de inteligencia de los profesores.

Se puede notar que en la muestra existirán valores de y_i que no observamos, puesto que no todos los profesores tomarán el examen. En algunos contextos se considera la idea, quizá *natural*, de tomar sólo los datos de los profesores que tomaron la prueba y estimar los parámetros de interés en el modelo basándose sólo en dicha información. Una alternativa más es asignar el valor intermedio de la escala del examen como variable y_i para los profesores que no tomaron la prueba. Típicamente, los problemas que son tratados con la metodología utilizada en esta tesis involucran gran cantidad de datos. En este sentido, podría pensarse que se cuenta con un número muy grande de profesores en la muestra e, incluso, el número de profesores que presentaron el examen podría seguir siendo muy grande. Lo anterior lleva a pensar (de forma errónea) que métodos usuales, como Mínimos Cuadrados Ordinarios (MCO), serán consistentes y, por lo tanto, producirán estimadores sensatos.

Considere ahora la decisión de presentar o no el examen. Suponga que cada profesor tiene una propensión al éxito en el examen[1] (llamaremos z a una medida de dicha propensión), y ésta depende linealmente de sus años de experiencia y de su salario actual (llamaremos x_2 a la variable que se refiere al salario actual). Si suponemos que valores altos de x_2 corresponden a niveles de salario bajos, es de esperarse que los profesores con valores altos de x_2 tengan mayor propensión de éxito en el examen,

[1] No sólo nos referimos a que al profesor le vaya a ir bien en el examen, sino a que, si su nivel de salario es bajo y no puede caer mucho más, sus opciones al tomar el examen son incrementar su salario (éxito) o quedarse igual. Además, profesores con un nivel salarial muy alto (digamos inmejorable) no tendrán incentivos para tomar el examen.

puesto que no podrían perder mucho (su salario no podría bajar demasiado) y en cambio sí podrían obtener muchos beneficios si obtienen un buen resultado en la prueba. Para la medida de propensión de éxito en el examen se tiene el siguiente modelo:

$$z_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \eta_i \tag{1.2}$$

$$i = 1,2, \dots, n.$$

En la expresión anterior η_i representa características no observadas por el investigador, pero que determinan el valor de z_i. Diremos que cada profesor decide tomar el examen si su medida de propensión al éxito es al menos tan grande como un nivel a. Asumiremos que η_i sigue una distribución normal centrada, independiente de x_{1i} , x_{2i} y η_j. Si $cov(\varepsilon_i, \eta_i) \neq 0$ entonces la estimación por mínimos cuadrados del modelo en (1.1) producirá estimadores sesgados e inconsistentes.

La idea es que para tener una estimación consistente de los parámetros inmersos en el modelo propuesto en (1.1) hay que tener en consideración que los profesores toman la decisión de hacer o no el examen en función de lo determinado en (1.2). En particular, la decisión del *i-ésimo* profesor depende de la variable (no observada) η_i, la cual puede capturar, por ejemplo, habilidades cognitivas, por lo que es plausible suponer que $cov(\varepsilon_i, \eta_i) > 0$. Intuitivamente, tendremos que los profesores más *listos*/*capaces* son los que, en general, tomarán el examen.

Para ilustrar este ejemplo, procedemos a simular una estructura en la que los parámetros son: $\beta_0 = \beta_1 = \alpha_0 = \alpha_1 = \alpha_2 = a = 1$; ε_i y η_i son normales estándar y $cov(\varepsilon_i, \eta_i) = .8$. El tamaño total de la población es $n = 500$. Procedemos ahora a estimar los parámetros que inicialmente eran de interés ($\beta_0 \, y \, \beta_1$) bajo los siguientes escenarios:

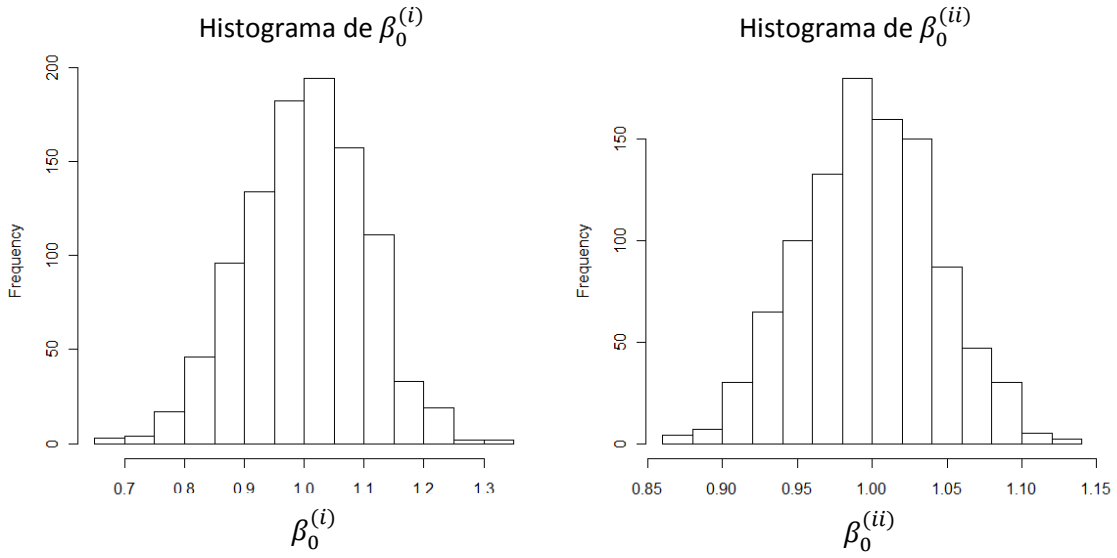
i) Estimación de máxima verosimilitud de un modelo que incorpore la información en (1.1) y (1.2). Este es el enfoque adecuado, lo describiremos a detalle más adelante.
ii) Estimación de máxima verosimilitud (equivalente a MCO) del modelo en (1.1) considerando que todos los profesores realizaron el examen. En la simulación propuesta tuvimos que crear el resultado obtenido de todos los profesores, por lo que podemos usar esta información. En un contexto real esto es impensable,

puesto que no podemos saber cuál habría sido el resultado en el examen de aquellos profesores que optaron por no presentarlo.

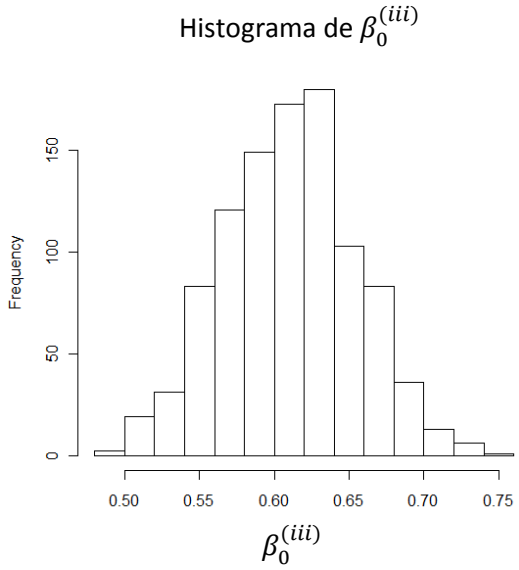
iii) Estimación de máxima verosimilitud (equivalente a MCO) del modelo en (1.1) colocando un cero (que para la simulación realizada, resulta un valor cercano al valor intermedio en el rango de resultados[2]) como resultado de los profesores que optaron por no tomar el examen. Claramente ésta es una mala alternativa, este ejemplo puede servir para ver qué tan malo es este enfoque.

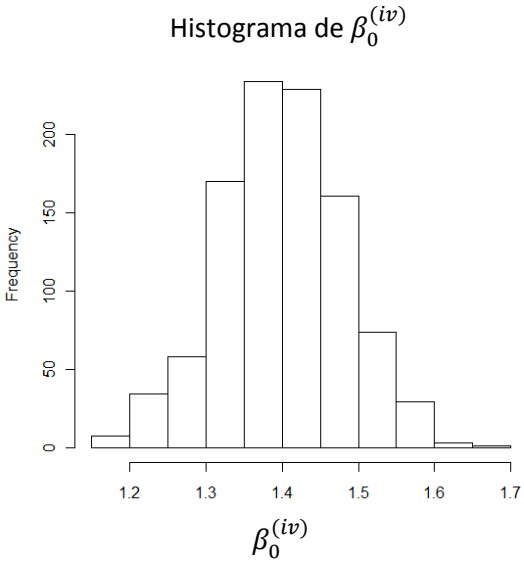
iv) Estimación de máxima verosimilitud (equivalente a MCO) del modelo en (1.1), considerando sólo la submuestra de profesores que presentaron el examen. Este enfoque es menos irrazonable, pero sigue siendo una mala alternativa, que sigue siendo empleada en la práctica.

Procedimos simulando 1000 veces el escenario antes descrito y calculando según cada enfoque los estimadores de β_0 y β_1. A continuación mostramos los histogramas correspondientes. Las gráficas corresponden a los estimadores obtenidos con los enfoque *i)*, *ii)*, *iii)* y *iv)* respectivamente. Comenzamos con las obtenidas para el parámetro β_0:



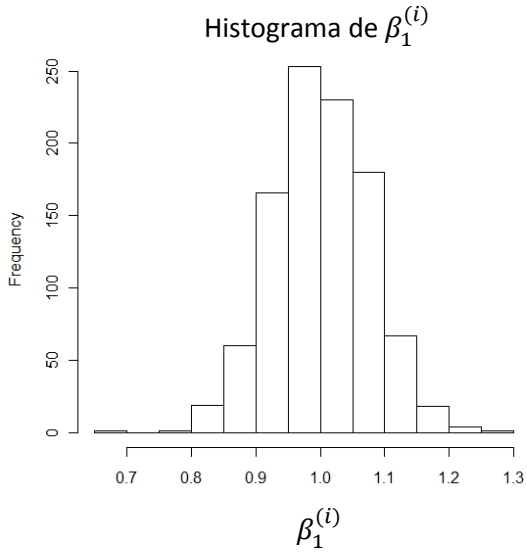
[2] Nos referimos a los resultados latentes; los que obtuvieron quienes realizaron la prueba y los que habrían obtenido quienes eligieron no hacerlo.

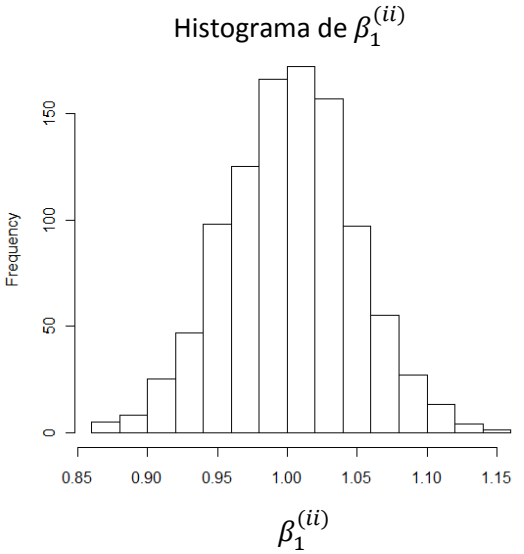


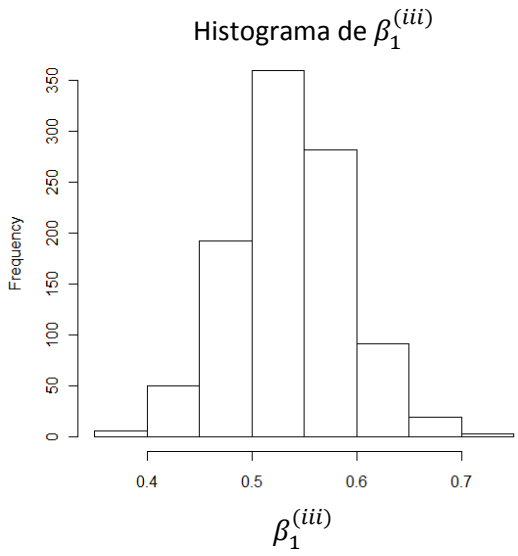


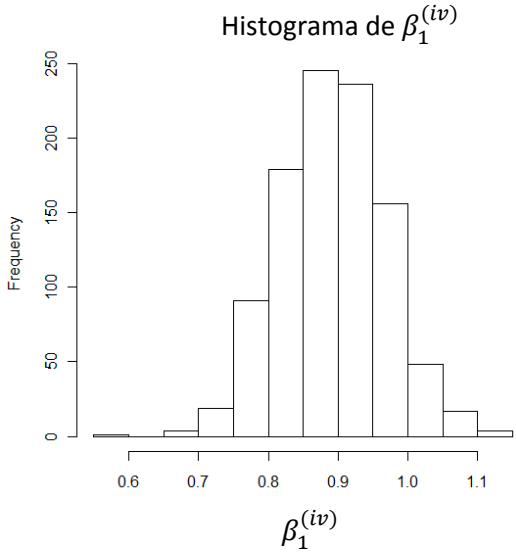
Los resultados que anticipamos se muestran claramente: los mejores estimadores vienen de la configuración *ii)*, que en la práctica no puede obtenerse. Lo más importante es comparar los estimadores de las configuraciones *i)* y *iv)*, los primeros son insesgados y consistentes, mientras que los segundos son sesgados e inconsistentes. En el contexto del ejemplo, las estimaciones del enfoque *iv)* nos llevan a sobreestimar el desempeño promedio que tendrían los profesores al presentar el examen (lo cual resulta lógico puesto que vimos que tienden a presentarlos los más inteligentes).

Enseguida vemos las estimaciones obtenidas para β_1:









Encontramos resultados similares a los de la estimación de β_0. En el contexto del ejemplo estudiado, subestimaríamos el efecto de los años de experiencia en el desempeño en el examen de los profesores.

Si este análisis fuera hecho con el fin de evaluar y/o mejorar parte del sistema educativo en México, se puede ver que las políticas públicas originadas a raíz de la (errónea) inferencia estadística podrían resultar ineficientes. Es en este tipo de escenarios donde aparece el problema de autoselección, que es adecuadamente ilustrado mediante el análisis de la decisión-resultado de la evaluación de los profesores.

B. Caracterización Estadística del Problema

En términos estadísticos, el problema de selección es generado (en el contexto del ejemplo descrito) porque, en el modelo escrito en (1.1):

$$E[y_i|z_i > 1] = \beta_0 + \beta_1 x_{1i} + E[\varepsilon_i|\eta_i > 1 - \alpha_0 - \alpha_1 x_{1i} - \alpha_2 x_{2i}]. \quad (1.3)$$

Y como η no es independiente de ε, tenemos que $E[\varepsilon_i|\eta_i > 1] \neq 0$. De modo que el modelo especificado en (1.1) no produce estimadores consistentes, este problema puede traducirse en términos estadísticos, como el problema ocasionado por error de especificación vía variables omitidas (en el contexto de regresión). Una de las primeras

alternativas para conseguir estimadores consistentes de β_0 y β_1 es considerar el siguiente modelo:

$$y = E[y_i|z_i > 1] + \tau_i = \beta_0 + \beta_1 x_{1i} + E[\varepsilon_i|\eta_i > 1 - \alpha_0 - \alpha_1 x_{1i} - \alpha_2 x_{2i}] + \tau_i =$$

$$= \beta_0 + \beta_1 x_{1i} + E\left[\varepsilon_i\middle|\eta_i > -h_i^{'}\theta\right] + \tau_i. \quad (1.4)$$

En la especificación anterior hemos considerado $h_i^{'} = (1, x_{1i}, x_{2i})$ y $\theta^{'} = (\alpha_0 - 1, \alpha_1, \alpha_2)$. El término τ_i es un término de error aleatorio, normalmente distribuido, con media 0, independiente a través de los individuos e independiente del resto de los términos que aparecen en la expresión. Usando teoría básica sobre la distribución normal bivariada, se puede ver que:

$$E\left[\varepsilon_i\middle|\eta_i > -h_i^T\theta\right] = \frac{cov(\varepsilon_i,\eta_i)}{\sigma_\eta}\lambda(H_i) = \rho\sigma_\varepsilon\lambda(H_i). \quad (1.5)$$

En la expresión anterior tenemos que ρ es el coeficiente de correlación entre ε y η. La razón $\lambda(.) = \phi(.)/\Phi(.)$ es conocida en la literatura como Razón Inversa de Mill y finalmente $H_i = -h_i^T\theta/\sigma_\eta$. Las letras ϕ y Φ representan la función de densidad y de distribución acumulada (respectivamente) de la distribución normal estándar.

Antes de continuar, vale la pena mencionar que, mediante la estimación de un modelo probit,[3] puede estimarse de manera consistente θ/σ_η, y por lo tanto pueden conseguirse estimadores consistentes de $\lambda(H_i)$. Sustituyendo por los valores estimados de H_i, y haciendo $\beta = \rho\sigma_\varepsilon$, podemos escribir, a partir de (1.4)**:**

$$y = \beta_0 + \beta_1 x_{1i} + \beta\lambda\left(\hat{H}_i\right) + \tau_i. \quad (1.6)$$

La estimación por mínimos cuadrados del modelo anterior produce estimadores consistentes. Los estimadores obtenidos son conocidos como los estimadores de Mínimos Cuadrados en 2 Etapas. Las propiedades de estos estimadores han sido estudiadas desde mucho tiempo atrás y pueden encontrarse en Heckman (1979). En 1.6

[3] Revisaremos este tema en breve.

puede notarse con mayor claridad por qué el problema de sesgo de selección puede introducirse como un problema ocasionado por error de especificación vía variables omitidas.

Los modelos de selección son empleados generalmente en el estudio del impacto de *tratamientos* (en el sentido usual en el que estos se conceptualizan en Diseño de Experimentos). En el mejor de los casos, contamos con unidades experimentales y aleatoriamente les asignamos un tratamiento para posteriormente determinar cuál es la diferencia en los resultados que es explicada por la aplicación del tratamiento respectivo. El trato en econometría, sustancialmente diferente, se debe en muchos casos al problema de autoselección. Si nos referimos a programas de asistencia social, el problema de autoselección puede entenderse así: a los individuos se les ofrece tomar un tratamiento y ellos deciden si lo toman o no, posteriormente ellos toman nuevas decisiones y estas pueden o no estar influenciadas por haber recibido o no haber recibido el tratamiento. Muchas preguntas surgen en este contexto: ¿de qué depende que un individuo desee o no tomar el tratamiento? ¿Las características del individuo afectan igual su desempeño independientemente de haber tomado el tratamiento? ¿Cómo estarían los individuos bajo tratamiento sin él? ¿Cómo estarían los individuos sin tratamiento si lo hubieran recibido? Todas estas preguntas, y más, pueden formularse en términos estadísticos y ser respondidas bajo este enfoque.

Otra situación[4] donde aparecen los modelos de selección es la determinación de *salarios/utilidad* de reserva. Si pensamos en el mercado laboral de las mujeres y queremos evaluar el impacto de la educación en el nivel salarial, supongamos que disponemos actualmente de datos concernientes a la educación y salario de mujeres en México. Un *problema* claro es que observaremos muchos “ceros” respecto a salario (muchas mujeres no trabajan). Podemos pensar, por ejemplo, que las mujeres con mayor escolaridad tienden a tener mayor utilidad al incorporarse al mercado laboral; sin embargo, algunas mujeres con poca escolaridad también se habrán incorporado al mercado laboral (posiblemente debido a características que no observamos). Esta situación provoca que para estimar de forma adecuada el impacto de educación en salario requiera de las técnicas tratadas en este trabajo.

[4] De hecho, el problema de salario de mujeres en los 70’s es el problema seminal en la metodología de modelos de selección. Un aspecto importante en la solución de este problema es determinar de qué depende que las mujeres decidan o no trabajar (esto se relaciona a su expectativa salarial y su salario de reserva).

C. Variables Latentes, Modelo Probit y Autoselección

Enseguida introduciremos el concepto de variables latentes, enfocándonos en los modelos de elección binaria, sin embargo, el concepto podrá trascender a otros contextos y será utilizado recurrentemente en este trabajo. Una variable latente es una variable que es observada parcialmente (sólo observamos algún indicador sobre dicha variable, o no la observamos para todos los individuos tratados). En el contexto binario las variables latentes pueden aparecer de dos formas. En primera instancia puede representar un índice/propensión para la ocurrencia de un evento. En segunda instancia (contexto de elección) puede representar la diferencia, en términos de utilidad, entre tomar una opción u otra.

Enseguida trabajaremos con los conceptos mencionados hasta el momento para introducir el modelo probit (binario). Supongamos que un individuo tiene la opción de elegir el producto *A* o el producto *B*. Supongamos que la función de utilidad de elegir cada uno de los productos es aleatoria (normalmente distribuida) y lineal en ciertas características:[5]

$$U(opciónA) = \boldsymbol{W}'\boldsymbol{\gamma_A} + \boldsymbol{X_A}'\boldsymbol{\beta_A} + \varepsilon_A \tag{1.7}$$

$$U(opciónB) = \boldsymbol{W}'\boldsymbol{\gamma_B} + \boldsymbol{X_B}'\boldsymbol{\beta_B} + \varepsilon_B \tag{1.8}$$

En la notación utilizada, $\boldsymbol{W}$ es un vector con características (generales) del individuo que realiza la elección. $\boldsymbol{X_A}$ y $\boldsymbol{X_B}$ son características relevantes para su utilidad según escoja la opción A o B, respectivamente. $\boldsymbol{\gamma_A}$, $\boldsymbol{\gamma_B}$, $\boldsymbol{\beta_A}$ y $\boldsymbol{\beta_B}$ representan los vectores de parámetros relevantes para la determinación de la utilidad en cada caso. ε_A y ε_B representan características desconocidas para nosotros, pero conocidas por el tomador de la decisión (podemos identificarlas como los componentes aleatorios, normalmente distribuidos y centrados en 0). Así, podemos crear la siguiente variable que representa la utilidad neta de elegir *A* sobre *B*:

[5] Este supuesto no es descabellado en muchísimos escenarios. En particular, si suponemos una función de utilidad Cobb-Douglas, con error multiplicativo que tiene distribución log-normal (algo que es sensato en muchos contextos) y trabajamos con modelos log-log entonces la modelación se justifica perfectamente. Las funciones de utilidad aleatorias que cumplen esta propiedad se denominan ARUM (en inglés hacen referencia a *additive random utility model*).

$$y^* = U_A - U_B = (\boldsymbol{W}'\boldsymbol{\gamma_A} + \boldsymbol{X_A}'\boldsymbol{\beta_A} + \varepsilon_A) - (\boldsymbol{W}'\boldsymbol{\gamma_B} + \boldsymbol{X_B}'\boldsymbol{\beta_B} + \varepsilon_B) = \boldsymbol{R}'\boldsymbol{\theta} + \omega \quad (1.9)$$

Aquí asumimos que $\boldsymbol{R}$, $\boldsymbol{\theta}$ y ω son la matriz de características, el vector de parámetros y la componente aleatoria adecuadas, respectivamente. Respecto al componente aleatorio ω, por los supuestos iniciales, podremos suponer normalidad (con media 0). Lo único que hemos hecho es escribir a la variable anterior en una forma típica de modelo lineal. El problema es que no observamos a y^*. En vez de esto, sólo observamos cuál fue la elección del agente. Si el agente elige la opción *A* entonces tenemos una variable que tomará el valor de 1, y 0 en caso de que el agente elija *B*:

$$y = \begin{cases} 1 & si \quad y^* \geq 0 \\ 0 & si \quad y^* < 0 \end{cases} \quad (1.10)$$

Conviene hacer ciertas generalizaciones sobre el modelo. Supongamos que tenemos una respuesta binaria, y se puede suponer que la variable latente (subyacente) es bien aproximada por un modelo lineal; es decir, si y^* representa la variable latente, podemos suponer que:

$$y^* = \boldsymbol{R}'\boldsymbol{\theta} + \omega. \quad (1.11)$$

Entonces, podemos presumir que la variable respuesta observada cambia (digamos de 0 a 1) cuando y^* pasa de ser negativa a positiva, es decir, el *umbral* que establecemos como válido es 0 (en otras palabras, suponemos (1.10) válido). Para ver esto, supongamos que el verdadero umbral es el nivel a (como en el ejemplo inicial). Supongamos también que en el vector $\boldsymbol{R}$ existe un término asociado a un intercepto (este supuesto es crucial para el modelo probit). Entonces tenemos que:

$$y = \begin{cases} 1\ si\ \boldsymbol{R}'\boldsymbol{\theta} + \omega \geq a \\ 0\ si\ \boldsymbol{R}'\boldsymbol{\theta} + \omega < a \end{cases} = \begin{cases} 1\ si\ -a + \boldsymbol{R}'\boldsymbol{\theta} + \omega \geq 0 \\ 0\ si\ -a + \boldsymbol{R}'\boldsymbol{\theta} + \omega < 0 \end{cases} = \begin{cases} 1\ si\ \boldsymbol{r}'\boldsymbol{\theta} + \omega \geq 0 \\ 0\ si\ \boldsymbol{r}'\boldsymbol{\theta} + \omega < 0 \end{cases} \quad (1.12)$$

Vea que la forma escrita en (1.11) sigue siendo válida. Típicamente no conocemos en qué unidades se mide la variable latente (en economía podríamos decir que son

"unidades de utilidad"), por eso es absurdo en la mayoría de las aplicaciones intentar investigar sobre el umbral, y como hemos visto, no se pierde generalidad si se impone cualquier valor. Por simplicidad en el tratamiento analítico, la convención es utilizar el 0.

A continuación, podemos ver que existe un problema de identificación[6] en la estimación del modelo descrito en (1.10) y (1.11). Para ver esto, establezcamos $\tilde{\theta} = a\theta$ y $\tilde{\omega} = a\omega$, entonces $\boldsymbol{R'\theta} + \omega > 0$ si y solamente si $\boldsymbol{R'\tilde{\theta}} + \tilde{\omega} > 0$. De este modo, es necesario fijar la varianza del componente aleatorio para realizar la estimación de manera única.[7] Desde otra perspectiva, supongamos que la "verdadera" utilidad neta se escribe como en (1.11), donde el componente aleatorio tiene desviación estándar $b \neq 0$. Como en realidad nosotros no sabemos sobre las unidades de medida de la "utilidad neta" podemos trabajar con el modelo:

$$\tilde{y}^* = \frac{y^*}{b} = \frac{\boldsymbol{R'\theta}}{b} + \frac{\omega}{b} = \boldsymbol{R'\tilde{\theta}} + \tilde{\omega} \tag{1.13}$$

$$\tilde{y} = \begin{cases} 1 & si \quad \tilde{y}^* \geq 0 \\ 0 & si \quad \tilde{y}^* < 0 \end{cases}. \tag{1.14}$$

El modelo anterior es conocido como modelo probit, y en él ya se asume sin pérdida de generalidad que el umbral es 0 y el parámetro de escala es 1.

La estimación del modelo puede ser realizada bajo la metodología de Modelos Lineales Generalizados. Note que la probabilidad de observar un *éxito* en determinado individuo es:

$$p_i = \Phi(R_i'\theta) \tag{1.15}$$

Por otra parte, podemos ver que la log-verosimilitud del modelo es:

[6] Podemos definir formalmente el problema de identificación de la siguiente forma. Supongamos que trabajamos con la variable aleatoria x con soporte en W y cuya función de distribución de probabilidad es F, donde F está determinada por el parámetroθ cuyo espacio parametral es Θ. Decimos que θ es identificable si para cualquier valor $\theta' \in \Theta$, no se cumple que $F(x|\theta) = F(x|\theta') \; \forall \; x \in W$.

[7] Antes habíamos mencionado que la estimación típica en estos contextos es la de θ/σ_ω. Se dice que el modelo está identificado salvo un parámetro de escala.

$$\log L(\theta) = \sum_{i=1}^{n} [y_i \log \Phi(R_i'\theta) + (1 - y_i)\log(1 - \Phi(R_i'\theta))]. \quad (1.16)$$

Las condiciones de primer orden generadas a partir de la función anterior son:

$$\sum_{i=1}^{n} w_i \big(y_i - \Phi(R_i'\theta)\big) R_i = 0. \quad (1.17)$$

Donde el peso

$$w_i = \frac{\phi(R_i'\theta)}{\Phi(R_i'\theta)(1-\Phi(R_i'\theta))}. \quad (1.18)$$

varía a lo largo de las observaciones. El método usual de optimización (Newton-Rhapson) es utilizado para hallar los estimadores.

Los efectos marginales del modelo probit son:

$$\frac{\partial p_i}{\partial R_{ij}} = \phi(R_i'\theta)\theta_j = \phi(\Phi^{-1}(p_i))\theta_j. \quad (1.19)$$

La estimación de estos modelos puede hacerse de manera habitual en R utilizando la función glm, estableciendo la liga probit para la familia binomial. Una generalización del modelo anterior es el modelo conocido como probit ordinal. Éste tiene como estructura de los datos observados (considerando (1.11)):

$$y_{i1} = \begin{cases} b_1 & si \quad -\infty < y^* \leq a_1 \\ b_2 & si \quad a_1 < y^* \leq a_2 \\ & \vdots \\ b_k & si \quad a_{k-1} < y^* \leq \infty \end{cases} \quad (1.20)$$

Para profundizar en los tópicos relacionados con la identificabilidad, estimación máximo-verosímil y aplicaciones de este modelo, se recomienda consultar Greene (2012) y Cameron y Trivedi (2005). Sin embargo, en el siguiente capítulo se propondrá un modelo donde se utilizará la estructura antes mencionada.

Enseguida, enunciamos el modelo básico de selección (una primera generalización de lo estudiado en el ejemplo introductorio). La notación la hemos tomado de Toomen y Henningsen (2008). Tenemos el siguiente sistema estructural (latente, no observado):

$$y_i^{S*} = \beta^{S\prime} x_i^S + \varepsilon_i^S \tag{1.21}$$

$$y_i^{O*} = \beta^{O\prime} x_i^O + \varepsilon_i^O \tag{1.22}$$

En la especificación anterior, y_i^{S*} representa la variable latente asociada a la decisión de aceptar o no el tratamiento, mientras que y_i^{O*} representa el outcome potencial (que observaremos en caso de que el i-ésimo individuo acepte el tratamiento). Tenemos que x_i^S y x_i^O (posiblemente distintas) son variables explicativas en cada proceso. Tenemos que los datos son observados de acuerdo a:

$$y_i^S = \begin{cases} 0 & si\ \ y_i^{S*} < 0 \\ 1 & si\ \ y_i^{S*} \geq 0 \end{cases} \tag{1.23}$$

$$y_i^O = \begin{cases} 0 & si\ \ y_i^S = 0 \\ y_i^{O*} & si\ \ y_i^S = 1 \end{cases} \tag{1.24}$$

Tenemos que la distribución de las componentes aleatorias es normal bivariada. Asumimos que el parámetros de escala de ε^S es 1 para garantizar la identificabilidad del modelo probit correspondiente (tal y como se explicó anteriormente). De este modo:

$$\begin{pmatrix} \varepsilon^S \\ \varepsilon^O \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix} \right) \tag{1.25}$$

Ya antes hemos mencionado que conforme crece (en magnitud) el parámetro ρ el problema de selección es más importante (de hecho sólo se evita cuando dicho parámetro es 0). Si $\rho = 0$, podemos pensar que la asignación/participación en tratamientos es aleatoria (en el sentido de que no depende del posible resultado en el tratamiento).

Utilizando teoría sobre la distribución normal bivariada, es fácil notar que la log-verosimilitud del modelo antes descrito es:

$$l = \sum_{\{i:y_i^S=0\}} \log\Phi\left(-\beta^{S\prime} x_i^S\right) + + \sum_{\{i:y_i^S=1\}} \left[\log\Phi\left(\frac{\beta^{S\prime} x_i^S + \frac{\rho}{\sigma}\left(y_i^O - \beta^{O\prime} x_i^O\right)}{\sqrt{1-\rho^2}} \right) - \frac{1}{2} log 2\pi - \log\sigma - \frac{1}{2}\frac{(y_i^O - \beta^{O\prime} x_i^O)^2}{\sigma^2} \right]. \quad (1.26)$$

Para la estimación del modelo, Toomen y Henningsen (2008) implementan una paquetería (sampleSelection) para R, que trabaja con la log-verosimilitud antes escrita, optimizándola con el algoritmo Newton-Rhapson.[8] En dicho artículo se versa también sobre el modelo de regresión cambiante, que será tratado más adelante.

Es bien sabido que los estimadores de máxima verosimilitud tienen propiedades muy deseables, y en particular, en este contexto, son más eficientes que los estimadores de MCO (incluso los estimadores en 2 etapas que se han propuesto). Si bien el anterior punto no se pone en duda, debe mencionarse que los estimadores en 2 etapas siguen siendo utilizados hoy en día, quizá por tradición, ya que durante mucho tiempo los trabajos enmarcados en el problema de autoselección utilizaban dichos estimadores.[9]

Hasta el momento hemos establecido cuál es el problema generado por la selección no aleatoria[10] de la muestra con la que se cuenta. Ante este escenario establecer relaciones de causalidad no resulta una tarea sencilla, entre otras razones, porque los potenciales errores de especificación conllevan a estimaciones inconsistentes que propician un modelo muy lejano a la realidad.

Es posible establecer algunas extensiones del modelo de selección que hasta ahora hemos tratado. Enseguida las enunciaremos y en los siguientes capítulos propondremos alternativas para la modelación y estimación en dichos escenarios.

i) En principio, en el modelo que establecimos sólo observamos la respuesta (outcome) en caso de que el individuo haya decidido recibir el tratamiento. Un

[8] También se habla sobre métodos de estimación en 2 etapas y se mencionan algunas opciones alternativas para trabajar con la verosimilitud.
[9] Si bien desde antes (80's) se había propuesto el tratamiento máximo verosímil, los costos computacionales de aquella época eran 15 y 700 dólares para la estimación.
[10] Incluida la autoselección de los individuos.

esquema más general se presenta cuando tenemos escenarios donde observamos respuestas *con tratamiento* y *sin tratamiento.* Cuando éste es el caso, es común el uso de lo que en la literatura se identifica como *modelo de regresión cambiante*.

ii) Podemos pensar en contextos donde un individuo decide tomar o no un tratamiento, y dicha decisión condiciona alguna(s) variable(s) latente(s) que a su vez generan otras respuestas. Podemos notar que este escenario engloba incluso el propuesto en *i*).[11]

En el siguiente capítulo introduciremos la modelación mediante sistemas de ecuaciones estructurales y veremos cómo las generalizaciones propuestas en *i)* y *ii)* pueden ser tratadas bajo dicho enfoque. En particular, el algoritmo MCEM, que trataremos en el cuarto capítulo de esta tesis, surge como una alternativa sensata para los modelos desprendidos del escenario *ii)*.

[11] Es decir, pensamos en respuestas latentes *per se*, que se observan no sólo condicionalmente en tomar o no el tratamiento, sino condicionadas, por ejemplo, en rebasar cierto umbral.

III. MODELACIÓN MEDIANTE SISTEMAS ESTRUCTURALES

En este capítulo hablaremos sobre la modelación ante distintos escenarios donde pueda presentarse el problema de autoselección. Además, se discutirán técnicas que permiten modelar situaciones donde las variables latentes son relevantes.

En la primera sección de este capítulo se introducirán los modelos de ecuaciones estructurales clásicos, hablaremos de los *diagramas de trayectoria* y los alcances de esta modelación para el tratamiento de variables latentes. Veremos que la versatilidad de la metodología permite tratar ciertos casos particulares. La segunda sección de este capítulo tratará sobre un ejemplo de *análisis de factores*, puesto que es una técnica que permite el tratamiento de variables latentes y se presta para la utilización del algoritmo que se estudiará en el capítulo IV, además de circunscribirse a los modelos estudiados en la primera sección. En la tercera sección se discutirá el problema que en la literatura es conocido bajo el nombre de *regresiones aparentemente no relacionadas*. Finalmente, en la cuarta y última sección, generalizaremos el esquema de dicho problema para modelar situaciones más complejas compatibles con escenarios de variable latente.

A. ECUACIONES ESTRUCTURALES Y VARIABLES LATENTES

Para describir la modelación mediante ecuaciones estructurales con variables latentes utilizaremos la notación de Bollen (1989). Es importante distinguir entre variables exógenas y endógenas. Haremos esto desde una perspectiva económica, pero el concepto quedará claro aun en el marco estadístico. Las variables endógenas son las que se intentan explicar dentro del modelo. Las variables exógenas, en cambio, son determinadas fuera del modelo. Por ejemplo, si decimos que el consumo de cierto bien depende del clima, la modelación correspondiente tendrá como variable endógena (la explicada por el modelo) al consumo y como variable exógena (determinada *fuera* del modelo) al clima.

En algunas ocasiones las estructuras son complejas y una variable puede ser exógena y endógena a la vez.[12] Por ejemplo, la inversión en seguridad de cierta zona

[12] Si este es el caso, diremos que la variable es endógena.

residencial depende, entre otros factores, del nivel de inseguridad que exista (si hay muchos asaltos, creeremos que los vecinos tendrán incentivos para incrementar la vigilancia). Para modelar el esquema anterior, una opción natural es asumir que la variable asociada a la inversión en seguridad es endógena y es explicada por la variable (que diremos exógena) nivel de inseguridad. El problema en la especificación anterior es que el nivel de inseguridad puede explicarse también en función de la inversión en seguridad (por ejemplo, mayor presencia policial ocasiona menos asaltos), de modo que podríamos decir ahora que ambas variables son endógenas y exógenas a la vez. Si se quiere estudiar el impacto de una variable endógena en cierta variable, suponiendo un modelo lineal, los estimadores que se obtienen con la metodología de MCO resultan inconsistentes.[13] Este problema es conocido en la literatura de econometría como el problema de endogeneidad. Existen diversas pruebas de hipótesis para determinar la endogeneidad de una variable (la más popular es la prueba tradicional de Hausman), y la gama de técnicas que abordan el problema es también amplia (incluyen, por ejemplo, el uso de variables instrumentales). Estos conceptos y técnicas pueden estudiarse con detalle en Greene (2012).

Entrando en materia, los modelos de ecuaciones estructurales tienen dos componentes esenciales. El primero, llamado componente de medición,[14] relaciona variables latentes con variables observadas. El segundo componente, llamado componente de variables latentes,[15] explica la estructura que genera a cada una de las variables latentes inmersas en el sistema estructural. Recomendamos al lector poner especial atención en el ejemplo que se expondrá una vez explicados los componentes.

Las ecuaciones estructurales del *componente de variables latentes* están dadas por:

$$\eta = B\eta + \Gamma\xi + \zeta \tag{2.1}$$

En la especificación anterior, η es un vector de dimensión $m \times 1$ que contiene a las variables latentes endógenas; ξ es un vector de dimensión $n \times 1$ y representa las

[13] El problema, puesto en términos del modelo de regresión lineal, es que el componente de error covaría con la variable explicativa (regresor).
[14] En inglés se encontrará que dicho componente es llamado *measurement model*.
[15] En inglés conocido como *latent variable model*.

variables latentes exógenas; B es una matriz $m \times m$ con los coeficientes que determinan la relación entre las variables endógenas;[16] Γ es una matriz de dimensión $m \times n$ que contiene la información de los efectos de las variables exógenas en las endógenas. Finalmente, decimos que ζ es el componente de *error aleatorio* en el modelo, éste es tal que $E[\zeta] = 0$ y $cov[\zeta, \xi] = 0$.[17]

Enseguida presentamos el segundo componente del modelo, que es conocido como *componente de medición*. Éste está dado por:

$$y = \Lambda_y \eta + \epsilon \tag{2.2}$$

$$x = \Lambda_x \xi + \delta \tag{2.3}$$

En las expresiones anteriores tenemos que y y x son vectores de dimensión $p \times 1$ y $q \times 1$, respectivamente, que representan variables observadas. Λ_y y Λ_x son matrices de dimensión $p \times m$ y $q \times n$ que determinan la relación entre y y η y entre x y ξ, respectivamente. Tenemos que ϵ y δ son vectores de dimensión $p \times 1$ y $q \times 1$ que representan errores de medición de y y x respectivamente. Se presume que los errores de medición tienen centro en el origen y son independientes entre si e independientes de ξ y η. Para simplificar cálculos, se asume que $\eta,\ y,\ \xi$ y x están escritas en términos de desviaciones respecto de su media. La identificación de la escala en las variables latentes no es posible,[18] por lo que debe imponerse una escala para dichas variables, generalmente, ésta se fija como la escala de alguna de las observables en y.

Vale la pena contemplar algunos escenarios particulares de la modelación anterior. Por ejemplo, supongamos que las variables latentes son observadas en su totalidad, lo que implica que no hay error de medición y que en realidad el *componente de medición* mide con exactitud cuáles fueron las realizaciones de las variables latentes, en términos del modelo, se tiene que: $\Lambda_y = I_m,\ \Lambda_x = I_n,\ var(\delta) = 0_{q \times q},\ var(\epsilon) = 0_{p \times p}$. Si éste es el caso, entonces trabajamos con la siguiente forma particular de (2.1):

[16] La matriz tendrá ceros en la diagonal, para no trivializar el modelo y decir que una variable se causa a si misma. Se asume también que la matriz $(I - B)$ es no singular.
[17] Note que bajo la definición de endogeneidad, el hecho de que $cov[\zeta, \xi] = 0$, es implícito a partir de que nombramos a ξ como variable exógena.
[18] La razón es la misma que origina el problema de identificabilidad en el contexto de los modelos probit.

$$y = By + \Gamma x + \zeta \tag{2.4}$$

Esta especificación es adecuada para los modelos de ecuaciones estructurales que se brindan en textos clásicos de Econometría,[19] y en particular puede ser útil en el contexto de *regresiones aparentemente no relacionadas*, que se especificará más adelante. La estructura determinada en (2.1-2.3) es también útil en otros contextos, especialmente en *análisis de factores*;[20] en este trabajo estudiaremos un caso especial de *análisis de factores* que permite la implementación del algoritmo MCEM (estudiado en el capítulo IV) para estimar los parámetros relevantes.

En la estructura propuesta tenemos que las matrices que contienen a los parámetros relevantes son: B, Γ, Λ_y, Λ_x, Φ, Ψ, Θ_ϵ y Θ_δ. Las matrices, Φ, Ψ, Θ_ϵ y Θ_δ son las matrices de covarianzas de ξ, ζ, ϵ y δ, respectivamente. Es necesario imponer *a priori* una estructura para las matrices de parámetros involucradas.[21] La idea es fijar algunos valores de los parámetros en las matrices correspondientes.[22] La teoría y el conocimiento sobre el fenómeno que se esté estudiando es fundamental en esta tarea. El tipo de preguntas que nos ayudan a esclarecer la forma de los parámetros del modelo son, entre otras: ¿cómo se relacionan las variables endógenas?, ¿qué estructura podemos suponer para los errores en ζ?, ¿qué variables latentes endógenas están determinadas por qué variables exógenas?, etcétera.

Para ilustrar la técnica expuesta hasta el momento, conviene considerar un ejemplo.[23] Se desea investigar sobre la posible relación entre la condición democrática de países desarrollados entre 1960 y 1965, y su relación con los índices de industrialización en 1960. En Ciencia Política existen tesis importantes en el sentido de que la industrialización promueve regímenes democráticos, además los comportamientos democráticos se fortalecen (es decir, niveles de democracia altos favorecen comportamientos democráticos en el futuro). Trataremos al nivel de democracia en 1960 (η_1) y al nivel de democracia en 1965 (η_2) como las variables latentes endógenas. Por otra

[19] Sobre este punto, puede consultarse el capítulo 10 de Greene (2012) y la sección 2.4 y 6.9 de Cameron y Trivedi (2005).
[20] Para más detalles puede consultarse el capítulo 7 de Bollen (1989).
[21] No hemos abundado al respecto, pero si dejamos todos los parámetros libres es muy probable que aparezcan problemas de identificación.
[22] Típicamente se colocan 0's o 1's donde la teoría sugiera que es sensato hacerlo. En economía además, existen ocasiones en que, con base en la teoría económica, se puede imponer la forma de ciertos parámetros en B, Γ, Λ_y y Λ_x (por ejemplo que $B_{1,2} = B_{2,1}$ o que $\Gamma_{2,1} - \Gamma_{2,3} = \Gamma_{2,2}$, etcétera).
[23] Tomado, igual que la explicación general del modelo, de Bollen (1989).

parte, trataremos al nivel de industrialización en 1960 (ξ_1) como variable latente exógena, y que tienen un efecto (contemporáneo) sobre los índices de democratización en 1960 y que también afecta los índices de democratización de 1965. Por supuesto, no tiene sentido suponer que la condición democrática de 1965 causó la de 1960. Tenemos entonces que, en este ejemplo, (2.1) es de la forma:

$$\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} = \begin{bmatrix}0 & 0\\ \beta_{21} & 0\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} + \begin{bmatrix}\gamma_{11}\\ \gamma_{21}\end{bmatrix}[\xi_1] + \begin{bmatrix}\zeta_1\\ \zeta_2\end{bmatrix} \tag{2.5}$$

En este ejemplo, asumimos que las componentes aleatorias ζ_1 y ζ_2 son no correlacionadas, por lo que Ψ es diagonal. Tenemos que Φ sólo contiene un elemento (la varianza de ξ_1) que es ϕ_{11}.

Respecto al *componente de medición*, correspondiente al vector η, representado en (2.2), se toman mediciones de 4 variables en ambos años. Éstas son: libertad de prensa (y_1, y_5), libertad del grupo político de la oposición (y_2, y_6), transparencia en el sistema electoral (y_3, y_7) y calidad del sistema legislativo (y_4, y_8). Se asume que las variables sólo tienen efecto contemporáneo (la variable latente ocurrida en cierto año sólo determina las variables observadas correspondientes a ese mismo año). La escala de las variables η_1 y η_2 son escogidas como las escalas de y_1 y y_2, respectivamente. Se establece que el impacto de la variable latente sobre la variable observada es el mismo a través del tiempo (por ejemplo, la transparencia del sistema electoral en 1960 determina el nivel democrático de 1960 en la misma medida que la calidad del sistema electoral en 1965 determina el nivel democrático en 1965). Con estas consideraciones, tenemos que (2.2) es:

$$\begin{bmatrix}y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\\ y_7\\ y_8\end{bmatrix} = \begin{bmatrix}1 & 0\\ \lambda_2 & 0\\ \lambda_3 & 0\\ \lambda_4 & 0\\ 0 & 1\\ 0 & \lambda_2\\ 0 & \lambda_3\\ 0 & \lambda_4\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} + \begin{bmatrix}\epsilon_1\\ \epsilon_2\\ \epsilon_3\\ \epsilon_4\\ \epsilon_5\\ \epsilon_6\\ \epsilon_7\\ \epsilon_8\end{bmatrix} \tag{2.6}$$

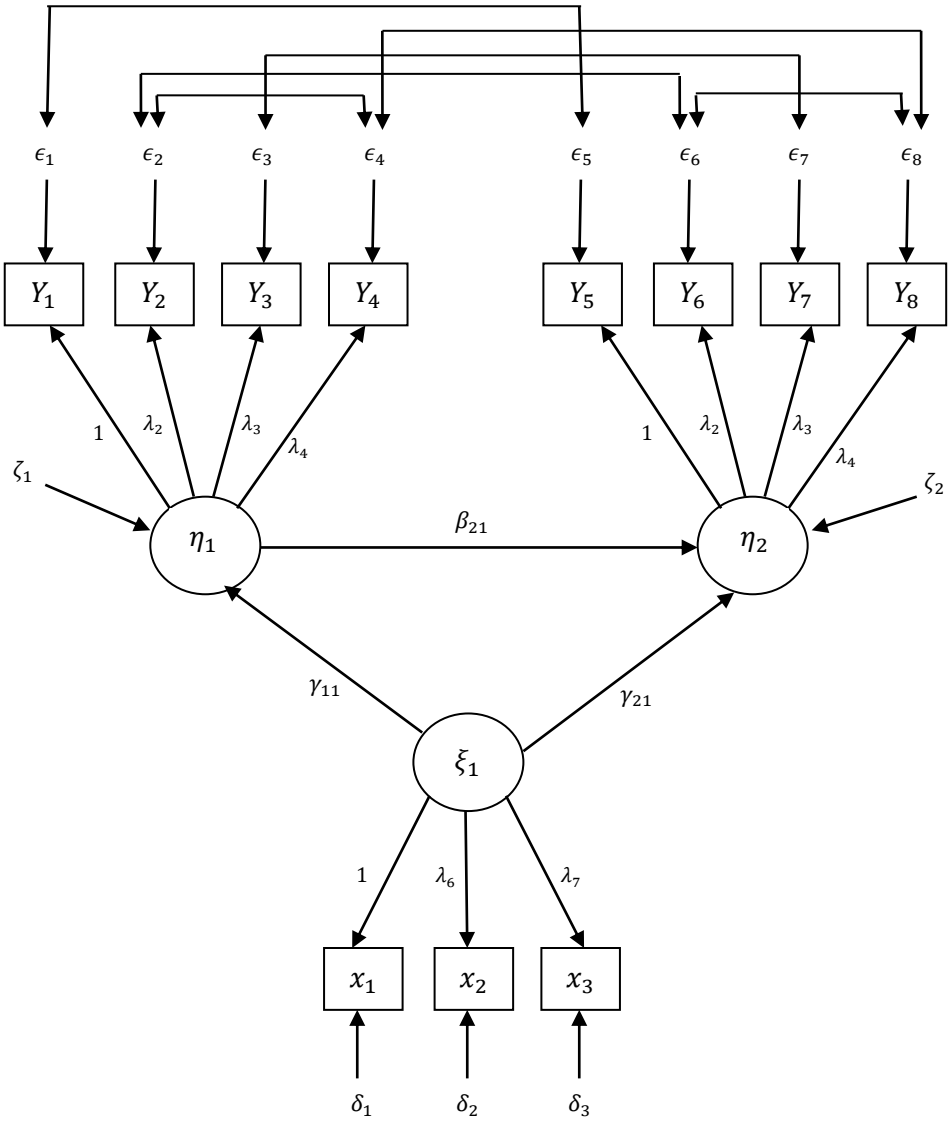
En la modelación se tiene que los errores de medición de los observables son independientes entre distintos años. Además suponemos que son independientes entre si, salvo los asociados a la libertad del grupo político de la oposición y la calidad del sistema legislativo. De este modo, asumimos que los elementos de Θ_ϵ que no son cero, son todos los de la diagonal y los elementos (4,2) y (8,6), correspondientes a la covarianza de los errores de medición de la libertad del grupo político de la oposición y la calidad del sistema legislativo en 1960 y 1965, respectivamente.

Para el *componente de medición*, correspondiente a la variable ξ_1 (que en este caso es el único componente del vector ξ), tenemos 3 variables observables. Éstas son el producto nacional bruto *per cápita* (x_1), el consumo de energía *per capita* (x_2) y la proporción de fuerza laboral ubicada en actividades industriales (x_3). La escala de ξ_1 es fijada como la escala de x_1. De este modo, tenemos que (2.3) es:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ \lambda_6 \\ \lambda_7 \end{bmatrix} [\xi_1] + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} \tag{2.7}$$

En esta ocasión no hay razones para suponer que los errores de medición sean dependientes entre si, por lo que se tiene que Θ_δ es diagonal.

Para la determinación de las estructuras que se analizan mediante sistemas de ecuaciones estructurales, se ha convertido en una tradición dibujar el *diagrama de trayectoria*, que resume de forma gráfica todas las relaciones modeladas en el sistema estructural. La realización de dicho diagrama puede ayudar a imponer algunos valores sobre las matrices de parámetros. Enseguida presentamos el *diagrama de trayectoria* correspondiente al ejemplo que hemos estudiado sobre democracia e industrialización.



A continuación, consideramos que todos los parámetros por estimar están contenidos en el vector θ. Utilizaremos la notación Σ_{qr} para referirnos a la matriz de covarianzas entre cualquier vector q y cualquier vector r. La metodología típica en la modelación mediante ecuaciones estructurales, se basa en estimar los parámetros en θ de modo que la matriz de covarianzas *teórica* de los observables (que denotaremos como Σ) se aproxime a la matriz de covarianzas muestral generada a partir de los datos observados (típicamente se denota como S). Dicho lo anterior, analicemos cómo es la matriz de covarianzas en función del modelo propuesto en (2.1-2.3).

Primero veamos cómo es la matriz de covarianzas del vector y en términos de θ:

$$\Sigma_{yy}(\theta) = E(yy') = E\left[\left(\Lambda_y\eta + \epsilon\right)\left(\eta'\Lambda_y' + \epsilon'\right)\right] = \Lambda_y E(\eta\eta')\Lambda_y' + \Theta_\epsilon \quad (2.8)$$

Ahora notemos que, a partir de la relación escrita en (2.1) podemos utilizar el álgebra de matrices para aislar del lado izquierdo de la ecuación al vector η. De este modo, tenemos que $\eta = (I - B)^{-1}(\Gamma\xi + \zeta)$.[24] Aprovechando la forma escrita de η, es fácil calcular $E(\eta\eta')$. Haciendo esto y remplazando dicha esperanza en (2.8) tenemos:

$$\Sigma_{yy}(\theta) = \Lambda_y(\mathrm{I} - \mathrm{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)[(\mathrm{I} - \mathrm{B})^{-1}]'\Lambda_y' + \Theta_\epsilon \quad (2.9)$$

Enseguida, calculamos la matriz de covarianzas (en términos de θ) entre las variables x y y. Tenemos:

$$\Sigma_{yx}(\theta) = E(yx') = E\left[\left(\Lambda_y\eta + \epsilon\right)(\xi'\Lambda_x' + \delta')\right] = \Lambda_y E(\eta\xi')\Lambda_x' \quad (2.10)$$

De nueva cuenta, utilizamos la forma reducida de η para calcular de forma sencilla el término $E(\eta\xi')$ y sustituyendo el valor de dicha esperanza en (2.10) tenemos:

$$\Sigma_{yx}(\theta) = \Lambda_y(\mathrm{I} - \mathrm{B})^{-1}\Gamma\Phi\Lambda_x' \quad (2.11)$$

Finalmente, para completar los elementos de Σ necesitamos calcular la matriz de covarianzas de x, en términos de θ. Veamos:

$$\Sigma_{xx}(\theta) = E(xx') = E[(\Lambda_x\xi + \delta)(\xi'\Lambda_x' + \delta')] = \Lambda_x E(\xi\xi')\Lambda_x' + \Theta_\delta = \Lambda_x\Phi\Lambda_x' + \Theta_\delta \quad (2.12)$$

Por último, utilizando los resultados anteriores, podemos hacer explícita la matriz de covarianzas de los datos observados en función de θ:

[24] En Econometría e incluso en otra literatura asociada a modelos estructurales, esta parametrización del modelo es conocida como la forma reducida de η.

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix} =$$

$$= \begin{bmatrix} \Lambda_y(\mathrm{I}-\mathrm{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)[(\mathrm{I}-\mathrm{B})^{-1}]'\Lambda_y' + \Theta_\epsilon & \Lambda_y(\mathrm{I}-\mathrm{B})^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x\Phi\Gamma'[(\mathrm{I}-\mathrm{B})^{-1}]'\Lambda_y' & \Lambda_x\Phi\Lambda_x' + \Theta_\delta \end{bmatrix} \tag{2.13}$$

Para que el modelo estructural sea estimable, debe verificarse la identificabilidad del mismo. Bollen (1989) habla sobre esta cuestión, proponiendo condiciones necesarias y suficientes para la identificabilidad del modelo propuesto. Como hemos mencionado, la idea en la estimación es establecer los parámetros de θ que aproximen la matriz Σ a la matriz S. Si se presupone que los vectores aleatorios inmiscuidos en la modelación siguen una distribución normal multivariada (y por tanto S sigue la distribución de Wishart), es posible[25] construir estimadores de máxima verosimilitud. Otras opciones de estimación se tienen a partir del método de Mínimos Cuadrados. Es posible estimar en R los parámetros de los sistemas de ecuaciones estructurales. Para esta tarea está disponible la paquetería sem. Si el lector está interesado en estimar estos modelos con dicha paquetería, es altamente recomendable leer el artículo de Fox (2006).

B. Análisis de Factores

Como se ha mencionado anteriormente, no es el interés primario en esta tesis estudiar en profundidad los sistemas de ecuaciones estructurales. Sin embargo, por la relevancia y éxito en el tratamiento de variables latentes consideramos oportuno hacer una introducción al tema desarrollado en la sección anterior. Enseguida, nos concentraremos en un problema enmarcado en la metodología de *análisis de factores* que implica trabajar con sistemas estructurales que incorporan variables latentes. La estimación del modelo propuesto será realizada con el algoritmo MCEM, que se estudiará en el siguiente capítulo. Para este ejemplo, consideramos las ideas propuestas por Yu, Lam y Lo (2005).

En su artículo, Yu, Lam y Lo (2005) se enfocan en analizar datos provenientes de una encuesta aplicada a 1,500 individuos para determinar cuáles son los factores más

[25] Mediante teoría de máxima verosimilitud para la distribución normal multivariada. Para más detalles puede consultarse la obra de Bollen (1989).

relevantes en la búsqueda de trabajo. A los encuestados se les propició una lista de 13 criterios a considerar en la búsqueda de trabajo, y se les solicitó ordenarlos según su criterio, del más al menos importante.[26] La idea del tratamiento es que las 13 variables se relacionan a algunas variables latentes.[27] Los autores concluyen que hay 3 variables latentes que son las que resultan más importantes al buscar un empleo. Éstas son: la proyección futura que ofrece el trabajo en función del talento y habilidad del trabajador, la responsabilidad/carga de trabajo que se enfrenta y finalmente un contraste entre la escala de la empresa y el salario ofrecido.[28] Enseguida nos avocaremos a explicar la metodología utilizada para este problema.

Comenzamos explicando el modelo a considerar. Suponga que tenemos una muestra de n individuos a los que se les solicita ordenar, según sus preferencias, k objetos.[29] Asumiremos que el ordenamiento del *i-ésimo* individuo depende de *utilidades latentes*, $x_{i1}, x_{i2}, \dots, x_{ik}$ que cumplen:[30]

$$x_{ij} = z_i^T a_j + b_j + \varepsilon_{ij}$$

$$i = 1, \dots, n, \qquad j = 1, \dots, k (> d) \tag{2.14}$$

Tenemos que z_i es un vector aleatorio que contiene d variables aleatorias normales estándar que son independientes; la idea es conceptualizar que el vector z_i contiene a los factores latentes.[31] En la especificación anterior, $b = (b_1, \dots, b_k)^T$ representa el vector de medias de las utilidades latentes (reflejando la importancia relativa de cada factor) mientras que $a_j = (a_{1j}, \dots, a_{dj})^T$ se forma con los coeficientes asociados a cada factor. Finalmente, ε_{ij} es un término de error que sigue una distribución normal con media 0 y varianza σ_j^2, los errores son independientes de los factores latentes.

[26] En realidad sólo se les pidió elegir los 3 aspectos que consideraban más importantes y se les pidió ordenarlos de forma descendiente en función de la importancia que le dan a cada uno.
[27] La cantidad de éstas es desconocida, y se determina, por ejemplo, mediante el criterio de Akaike.
[28] La interpretación de estos *factores importantes* no siempre es posible, y requiere de gran capacidad de modelación e interpretación por parte del estadístico. La interpretación de las variables latentes es parecida a la interpretación de los componentes principales obtenidos mediante la metodología del mismo nombre.
[29] Los "objetos" pueden sustituirse por criterios, productos, situaciones, etcétera.
[30] Se puede ver que el modelo propuesto puede anidarse dentro de un modelo general de *análisis de factores*.
[31] En el ejemplo introductorio, sobre criterios en la búsqueda de empleo, recuerde que se identificaron 3 factores: la proyección futura que ofrece el trabajo en función del talento y habilidad del trabajador, la responsabilidad/carga de trabajo que se enfrenta y finalmente un contraste entre la escala de la empresa y el salario ofrecido.

Ahora considere el vector $r_i = (r_{i1}, \dots, r_{ik})^T$, que contiene los índices de ordenamiento que hace el *i-ésimo* individuo. El elemento r_{ij} refleja el orden de importancia que da el individuo i al objeto j. Supondremos que el ordenamiento está hecho de mayor a menor.[32] Por ejemplo, si tenemos un catálogo con $j = 3$ objetos, y registramos que $r_i = (2,3,1)^T$, tenemos entonces que el vector de utilidades latentes (no observadas) que generó el ordenamiento es $x_i = (x_{i1}, x_{i2}, x_{i3})^T$ donde $x_{i2} < x_{i1} < x_{i3}$.

Consideraremos ahora algunos aspectos de notación. Tenemos que la matriz $A_{d\times k} = (a_1 \dots a_k)$, que contiene los coeficientes que asocian los factores latentes a cada objeto bajo consideración. Consideremos que la matriz de covarianzas de los errores es $\Psi_{k\times k}$, que resulta ser diagonal con $diag(\Psi) = (\sigma_1^2, \dots, \sigma_k^2)$. Tenemos entonces que los parámetros a estimar en este modelo están dados en el vector $\theta = \{A, b, \Psi\}$. Llamaremos $\mathrm{X}_{n\times k}$ y $\mathrm{Z}_{n\times d}$ a las matrices que concentran a las utilidades latentes y a los factores latentes. Cada renglón representa a un individuo. Finalmente consideraremos a $R_{n\times k} = (r_1, \dots, r_n)^T$ como la matriz con los ordenamientos.

El modelo propuesto es bastante sencillo, pero útil en diversos contextos. Además servirá para ilustrar el algoritmo MCEM que detallaremos en el siguiente capítulo,[33] y que puede ser útil en algunas generalizaciones del modelo propuesto.

Más adelante versaremos sobre los conceptos de *datos faltantes* versus *datos observados*. Por ahora, note que $\{X, Z\}$ son matrices con datos faltantes, mientras que R es una matriz con datos observados. Por el momento, pedimos al lector que considere que todas las variables inmiscuidas en (2.14) son observadas. Si este fuera el caso, entonces la verosimilitud sería:[34]

$$l(\theta|X, Z, R) = -\frac{n}{2}\sum_j \log\left(\sigma_j^2\right) - \frac{1}{2}\sum_i \sum_j \frac{(x_{ij} - z_i^T a_j - b_j)^2}{\sigma_j^2} \qquad (2.15)$$

[32] Vea, por ejemplo, que si $r_{ij} = 1$, quiere decir que el individuo i considera al objeto j como el más importante/preferido/útil.

[33] En el artículo original, Yu, Lam y Lo (2005) mencionan que el modelo se puede generalizar a situaciones en las que a los individuos se les solicita ordenar los q objetos más importantes de los k disponibles. Entonces si el individuo i determina que los q objetos más importantes son $\mathrm{j}_1, \dots, \mathrm{j}_\mathrm{q}$, con el orden de importancia $1, \dots, \mathrm{q}$ respectivamente. Esto implicaría una generalización que considere ahora que el vector de utilidades latentes que generó el ordenamiento es $\mathrm{x}_{\mathrm{i}1}, \dots, \mathrm{x}_{\mathrm{ik}}$ donde lo único que sabemos sobre dicho vector es que $\mathrm{x}_{\mathrm{ij}_1} > \mathrm{x}_{\mathrm{ij}_2} > \cdots > \mathrm{x}_{\mathrm{ij}_\mathrm{q}} > \mathrm{x}_{\mathrm{ij}_{\mathrm{q}+1}}, \dots, \mathrm{x}_{\mathrm{ij}_\mathrm{k}}$.

[34] Note que si el problema se intentara resolver con máxima verosimilitud considerando sólo la información observada sería muy difícil establecer la distribución de cada vector r_i.

En el siguiente capítulo introduciremos un algoritmo que permite dar estimadores consistentes de los parámetros contenidos en θ.

C. Regresiones Aparentemente no Relacionadas

A continuación introduciremos el modelo de regresiones aparentemente no relacionadas (SUR). Este modelo es muy útil en y popular en economía, por el tipo de escenarios que permite modelar. Básicamente, el modelo se concentra en sistemas de ecuaciones donde los términos de error están correlacionados (entre ecuaciones) pero no aparece el problema de endogeneidad.[35] En este modelo tampoco tratamos con variables latentes, sino que asumimos que todas las variables son observadas y además son en sí mismas las variables que quieren explicarse.[36]

La idea de explicar este modelo, es que además de ser útil en algunos escenarios, permite hacer las generalizaciones necesarias (incluir como regresores variables endógenas e incluir el problema censura/variables latentes) para atacar el problema descrito en el primer capítulo. Además, en el contexto sencillo de SUR, se propondrá un ejemplo sencillo bajo un esquema de inferencia Bayesiana, que permitirá ilustrar el muestreador de Gibbs, mismo que estudiaremos en el siguiente capítulo.

Motivemos la metodología asociada a SUR bajo un enfoque econométrico en el contexto de la estimación de demandas. Suponga que tenemos N individuos y analizamos su demanda por M bienes. Si nos fijamos en el i-ésimo individuo, piense que la variable dependiente (demanda) por el bien m es el gasto total que hace para la compra de ese bien, denotamos esta variable como y_{im}. Suponga que para explicar dicho comportamiento, se cuenta con ciertas características, exógenas, contenidas en el vector x_{im}. Tenemos entonces que el modelo (lineal) a considerar es:

$$y_{im} = x'_{im}\beta_m + \varepsilon_{im},$$

$$m = 1,2,\dots,M, \qquad i = 1,2,\dots,N. \tag{2.16}$$

[35] Puede decirse que los sistemas que aparecen están ya en la *forma reducida* del sistema estructural.
[36] Puede verse que la modelación mediante sistemas estructurales, en el caso particular de la especificación resumida en 2.4, resulta apta para modelar este escenario.

Tenemos que $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{im})^T$ es un vector que agrupa ruidos aleatorios normalmente distribuidos, y que tiene media 0 y matriz de covarianzas Σ. El vector β_m contiene los parámetros asociados a las variables en x_{im}. Se utiliza el término *aparentemente no relacionadas* porque en ocasiones es plausible suponer que los términos de error aleatorio intraindividuos (es decir, los términos aleatorios referentes a las ecuaciones) están correlacionados aunque generalmente se asumen independientes entre individuos.[37] Vea además que las demandas estudiadas pueden depender de distintas covariables.

El problema estadístico que aparece, si no se incorpora la estructura de covarianzas que gobierna el proceso descrito en (2.16) es la ineficiencia de los estimadores. A diferencia de lo ocurrido cuando tratamos los modelos de autoselección, al estimar (2.16) por MCO, considerando cada ecuación por separado, sí obtenemos estimadores consistentes. Pese a esto, la varianza de dichos estimadores no es la menor posible.

Consideremos como ejemplo, un escenario (partiendo de (2.16)) en el que $n = 50\ y\ m = 2$. Consideraremos, como regresores, un intercepto y una sola variable explicativa para cada ecuación (las simularemos de la distribución normal estándar). Asumiremos que los componentes aleatorios tienen varianza 1 y covarianza $-.5$. Escribiendo esto, tenemos:

$$y_{i1} = x'_{i1}\beta_1 + \varepsilon_{i1},$$

$$y_{i2} = x'_{i1}\beta_2 + \varepsilon_{i2},$$

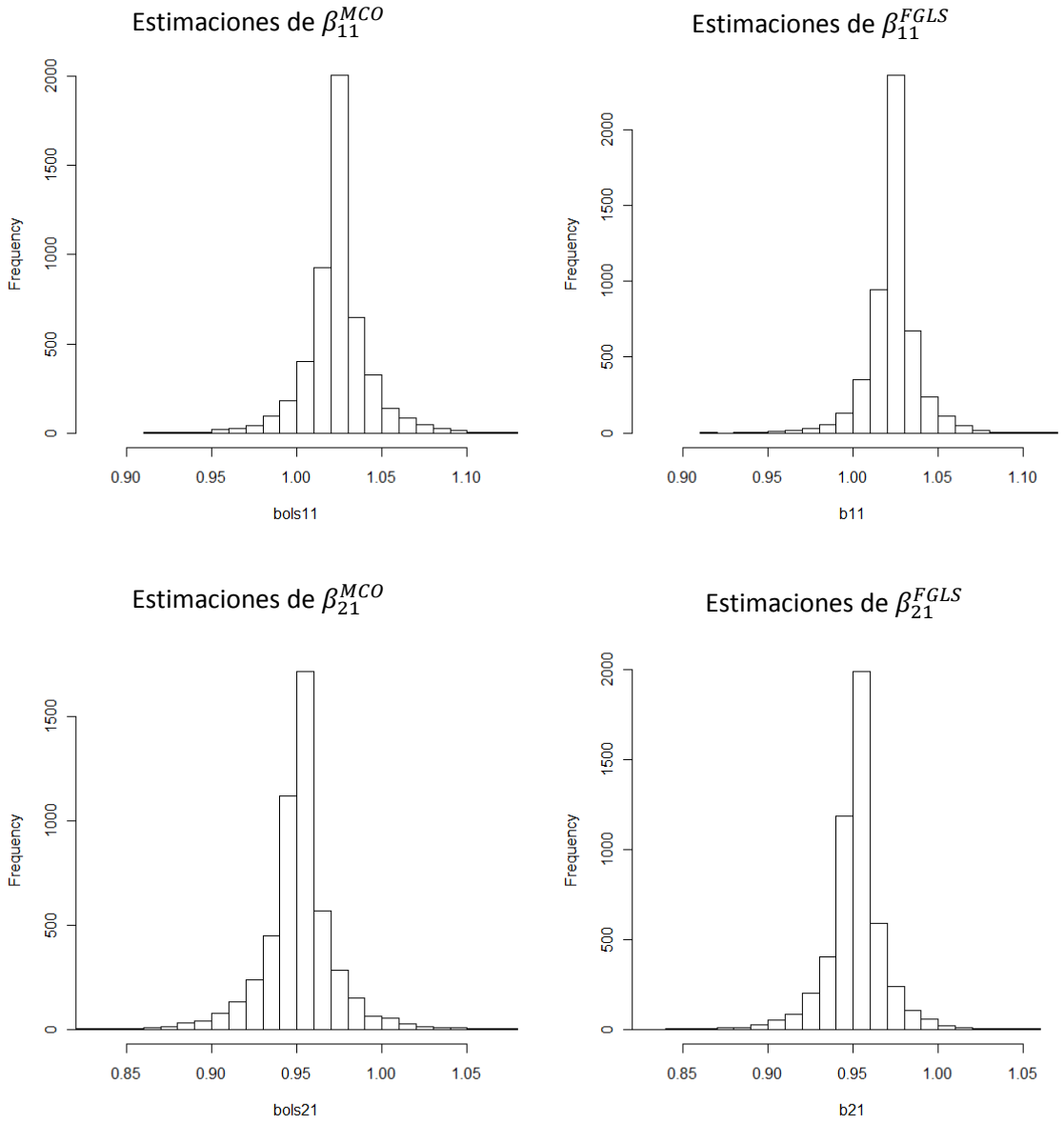
$$i = 1,2, \dots ,50 \qquad (2.17)$$

Donde $\beta_1 = \beta_2 = (1,1)'$, además $var(\varepsilon_{i1}) = var(\varepsilon_{i2}) = 1$ y $cov(\varepsilon_{i1}, \varepsilon_{i2}) = -.5$. El escenario en (2.17) fue simulado 5000 veces y en cada vez se estimaron los parámetros

[37] Un contexto donde puede entenderse que los errores de medición estén correlacionados es el siguiente. Imagine que las variables dependientes son *Gasto en escuelas* y *Gasto en hospitales*, y los individuos que se consideran son los estados de la República. Como generalmente cada estado tiene su sistema de medición, es posible suponer que los errores de medición, dentro de cada estado, estén correlacionados (por ejemplo si su sistema de contabilidad hace que se midan simultáneamente las variables dependientes). De cualquier modo es plausible suponer que, entre estados, los errores de medición son independientes.

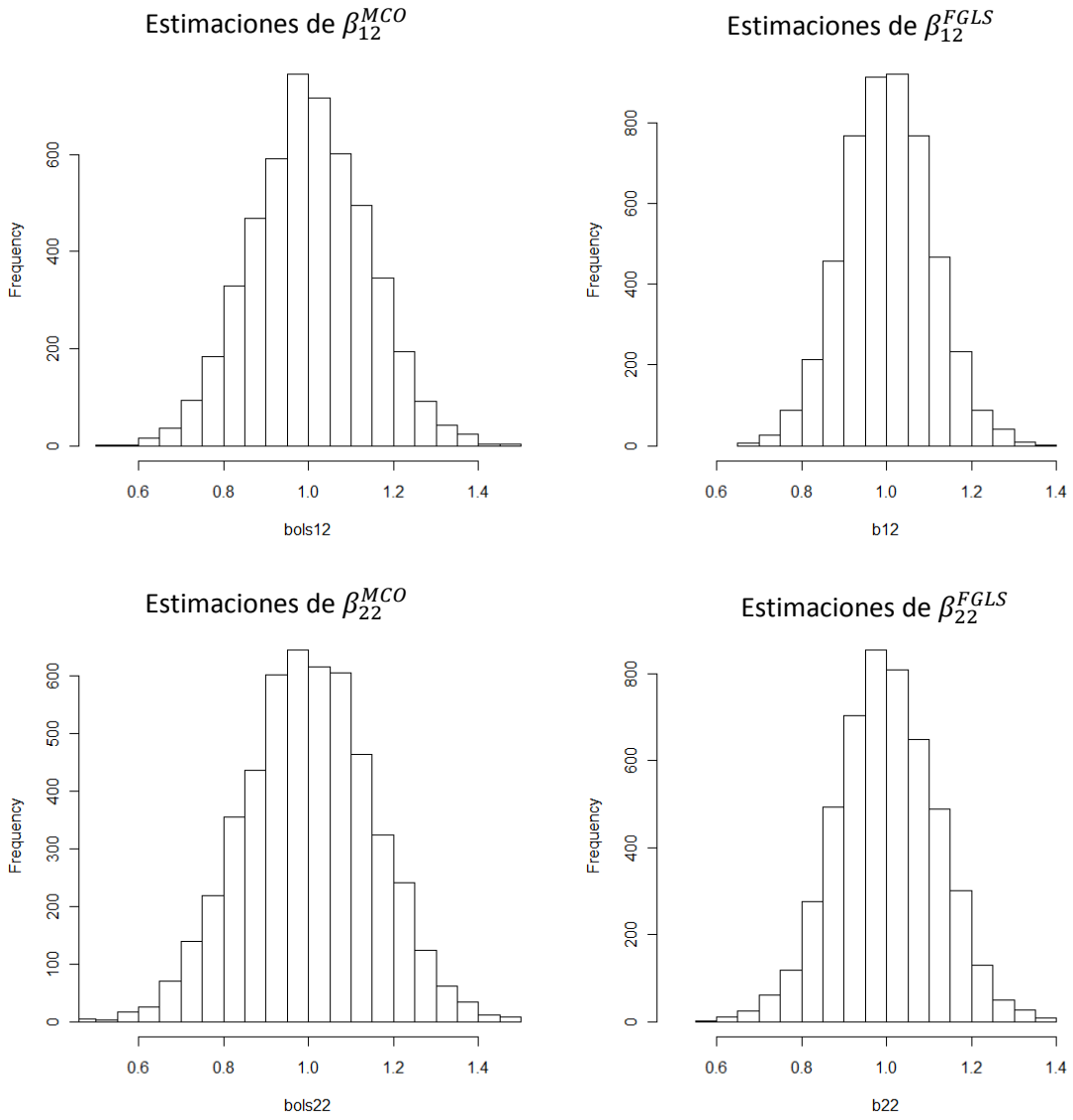
en $\beta_1 \, y \, \beta_2$. Las estimaciones fueron hechas mediante MCO y mediante Mínimos Cuadrados Generalizados Factibles (FGLS).[38] Mostramos los histogramas de los estimadores obtenidos, para evidenciar la mayor eficiencia del método FGLS.

Mostramos gráficamente el comportamiento de los estimadores obtenidos mediante MCO (lado izquierdo) y el método FGLS (lado izquierdo). Puede apreciarse que la varianza de los estimadores obtenidos mediante el segundo método es más pequeña.[39]



[38] Más adelante exploraremos, de manera muy superficial, esta metodología.
[39] La reducción de varianza en las estimaciones no es tanta, pero en problemas de mayor cardinalidad puede incrementarse. En este ejercicio, la varianza de las estimaciones se redujo, en promedio, en un 40% (la desviación estándar se redujo aproximadamente en 23%) para los 4 estimadores obtenidos.

Enseguida mostramos los histogramas obtenidos para la estimación de las pendientes.



Enseguida estructuramos formalmente el modelo SUR. Es conveniente agrupar a los N individuos según su estructura en cada ecuación. Haciendo esto obtenemos:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & X_m \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_M \end{bmatrix} = X\beta + \varepsilon \qquad (2.18)$$

En la expresión anterior, y_j, X_j, β_j y ε_j contienen a las respuestas, los regresores, los coeficientes y los errores de correspondientes a la $j-ésima$ ecuación. Recuerde que para el $n-ésimo$ individuo, el vector de errores es normalmente distribuido, con matriz de covarianzas:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2M} \\ & \vdots & & \\ \sigma_{M1} & \sigma_{M2} & \dots & \sigma_{MM} \end{bmatrix} \tag{2.19}$$

En (2.18), tenemos el modelo de regresión habitual, pero el supuesto de homocedasticidad no se cumple (la correspondiente matriz de covarianzas no es siquiera diagonal[40]). De hecho, con base en (2.19), podemos notar que la matriz de covarianzas de ε es de la siguiente forma:

$$\Omega = \Sigma \otimes \mathrm{I} \tag{2.20}$$

En la expresión anterior el operador $\otimes$ representa al producto Kronecker. Utilizando las propiedades de dicho operador, podemos obtener la inversa de la matriz de covarianzas:

$$\Omega^{-1} = \Sigma^{-1} \otimes \mathrm{I} \tag{2.21}$$

Teniendo esto en consideración, es bien sabido que el estimador de Máxima Verosimilitud (y de Mínimos Cuadrados Generalizados) para β es:

$$\hat{\beta} = [X^{-1}\Omega^{-1}X]^{-1}X^{-1}\Omega^{-1}y = [X^{'}(\Sigma^{-1} \otimes \mathrm{I})X]^{-1}X^{'}(\Sigma^{-1} \otimes \mathrm{I})\mathrm{y} \tag{2.22}$$

Claro está que para poder utilizar los estimadores de (2.22), la matriz de covarianzas debería ser conocida. Esto no ocurre. Antes de continuar, debemos mencionar que el estimador de mínimos cuadrados (bajo los supuestos habituales) es

[40] Cuando esto ocurre, decimos que existe autocorrelación.

consistente. Recuerde que dicho estimador es $\hat{\beta} = (X'X)^{-1}X'Y$. Con base en este estimador, y utilizando los residuales de la regresión correspondiente, podemos obtener estimadores consistentes de las covarianzas en (2.19). La expresión para obtener las covarianzas mencionadas es:

$$\hat{\sigma}_{ij} = S_{ij} = e_i^T e_j / N \tag{2.23}$$

Con estas estimaciones, podemos construir la matriz de covarianzas en (2.19), de modo que el estimador consistente de tal matriz es $\hat{\Sigma} = S$, donde:

$$\mathrm{S} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1M} \\ S_{21} & S_{22} & \dots & S_{2M} \\ & \vdots & & \\ S_{M1} & S_{M2} & \dots & S_{MM} \end{bmatrix} \tag{2.24}$$

Con este estimador, podemos construir de forma inmediata un estimador consistente de Ω y con ello obtener un estimador "factible" de mínimos cuadrados generalizados. Éste es el que denotamos como estimador FGLS:

$$\hat{\beta} = [X^{-1}\Omega^{-1}X]^{-1}X^{-1}\Omega^{-1}y = [X'(\Sigma^{-1} \otimes \mathrm{I})X]^{-1}X'(\Sigma^{-1} \otimes I)y \tag{2.22}$$

Los estimadores en (2.22) resultan consistentes y más eficientes que los obtenidos mediante MCO. Una vez obtenidos los estimadores, pueden re-calcularse los residuos y proponerse nuevos estimadores para la matriz de covarianza siguiendo los pasos expuestos en (2.23-2.24) y β puede reestimarse, produciendo estimaciones aun más eficientes. El proceso puede iterarse hasta convergencia y es éste el que es llamado *iterated FGLS*. Los estimadores obtenidos (tras la convergencia) son equivalentes a los estimadores de Máxima Verosimilitud.[41]

El método *iterated FGLS* es el que fue utilizado para producir los histogramas mostrados en páginas anteriores. Esta metodología puede implementarse en R mediante la función systemfit, especificando "SUR" como parámetro para *method*.

[41] Oberhofer y Kmenta (1974) *apud* Greene (2012).

Enseguida analizaremos un enfoque Bayesiano para la estimación del modelo SUR, particularmente, consideramos la expresión brindada en (2.18). Recuerde que el principal objetivo de este ejemplo es mostrar, en el capítulo siguiente, la utilidad e implementación del muestreador de Gibbs. Considerando entonces el modelo enmarcado en (2.18) (cuyos supuestos distribucionales pueden consultarse en párrafos anteriores) se proponen las siguientes distribuciones *a priori*:

$$\beta \sim N(\beta_0, B_0^{-1}) \tag{2.25}$$

$$\Sigma^{-1} \sim Wishart(v_0, D_0) \tag{2.26}$$

Puede notarse que las distribuciones *a priori* propuestas son independientes. Esto facilitará los cálculos que deben hacerse en este ejemplo, sin embargo, la modelación propuesta es utilizada y conveniente para algunos ejemplos.[42] La forma de la densidad posterior de los parámetros puede verse a partir de la siguiente expresión:

$$f(\beta, \Sigma, \mathrm{y}|x) = N(\beta_0, B_0) \times Wishart(v_0, D_0) \times \prod_i N(x_i^T \beta, \Sigma) \tag{2.2V}$$

En la expresión anterior, las densidades se evalúan en β, Σ y y_i, respectivamente. Puede observarse que no se obtendrá el kernel de una distribución conocida. Esta situación (típica en el contexto bayesiano) puede abordarse mediante métodos computacionalmente intensivos. Los algoritmos de Metropolis-Hastings constituyen una herramienta útil para simular de la densidad posterior deseada. En el siguiente capítulo hablaremos de ellos,[43] haciendo énfasis en el muestreador de Gibbs, que utilizaremos para poder dar solución a este problema.

[42] Cameron y Trivedi (2005).
[43] De forma muy general.

D. Estructuras para Modelar la Autoselección

En esta sección brindaremos una estrategia de modelación estructural que permita abordar el problema de autoselección. Conviene hacer ciertas generalizaciones al modelo SUR que permitirán dar modelación y hacer inferencia en el contexto del problema de selección. En particular, las generalizaciones que haremos son:

i) Permitir que las variables endógenas sean latentes.
ii) Permitir que las variables endógenas aparezcan como regresores.

Como veremos, estas simples generalizaciones permitirán modelar contextos de selección (sin ser esta su única utilidad). Comencemos teniendo en mente el modelo más básico de selección, descrito en el primer capítulo.

La estructura del modelo elemental para el problema de selección se encuentra en (1.21-1.24). Recuerde que en este modelo se tenía una ecuación de selección y una ecuación de respuesta, donde sólo se observaban respuestas de los individuos que tomaban el tratamiento.[44]

Una primera generalización natural, es el caso en el que observamos la variable respuesta para todos los individuos, independientemente de que hayan tomado o no el tratamiento. Si éste es el caso, estaremos interesados en responder preguntas como ¿cuál es la ganancia esperada al tomar el tratamiento? ¿Afectan de distinta forma las covariables a la variable respuesta en función de tomar o no el tratamiento? Para ilustrar un ejemplo en el que el contexto descrito está presente, podemos pensar en el análisis de los beneficios de una empresa en función de si ésta opera en el sector formal o informal.

Describiremos la primera generalización en función del ejemplo de los beneficios de la empresa considerando el sector formal e informal. Supongamos que, en primera instancia, las empresas deciden si operarán en uno u otro sector. Presumimos que las actividades económicas inician en el sector informal, por lo que la decisión sería *pasarse al sector formal*. Operar en el sector formal tiene un costo mayor (burocracia, impuestos, etcétera) pero suponemos que a la par de estos costos hay algunos beneficios que podrían incentivar la decisión en los empresarios (acceso a tecnología, financiamiento

[44] Los individuos que toman el tratamiento son aquellos para los que observamos un 1 en la etapa de selección. Observe que el modelo antes descrito tiene muchas similitudes con el modelo SUR, excepto porque la variable endógena en la primera ecuación es latente (corresponde a un modelo probit) y porque la variable endógena en la segunda ecuación dependerá de la primera.

gubernamental, etcétera). Entonces, el primer componente del modelo es la ecuación de selección. Ésta se basa en la variable latente que describe la utilidad que tiene el empresario por cambiarse del sector informal al formal.[45] Esta parte del modelo es:

$$y_{1i}^{*} = X_{1i}\beta + \varepsilon_{1i} \tag{2.27}$$

$$y_{1i} = \begin{cases} 0 \;\; si \;\; y_{1i}^{*} < 0 \\ 1 \;\; si \;\; y_{1i}^{*} \geq 0 \end{cases} \tag{2.28}$$

En la ecuación (2.27) representamos la utilidad latente obtenida por operar en el sector formal. Los regresores[46] en X_{1i} pueden ser factores que disminuyan los costos de tomar la decisión (experiencia en trámites burocráticos, contactos en el sector público), además de las características que determinan la productividad de la empresa (nivel de capital, tecnología, tipo de estructura de la empresa, etcétera). Las características capturadas por la variable de error aleatorio ε_{1i} (además del error de medición) pueden referirse, por ejemplo, a las habilidades no observadas del empresario. Si la utilidad latente es mayor que 0 entonces los individuos optarán por operar en el sector formal. Esta información es la que observamos, de acuerdo a (2.28).

Para la ecuación de respuesta, a diferencia de la estructura determinada por la (1.22) y (1.24), ahora siempre observamos la variable respuesta, sólo que el proceso que genera dicha respuesta cambia en función de si se toma o no el tratamiento. Así, en vez de registrar un número arbitrario (típicamente 0) para los individuos que no toman el tratamiento, registramos la respuesta que reportan.[47] Así, tenemos que la variable que genera las respuestas (en este caso los beneficios de las empresas) puede escribirse como:

$$y_{2i}^{*} = X_{2i}\beta_2 + y_{1i}X_{2i}\alpha + \varepsilon_{2i} \tag{2.29}$$

[45] Esta ecuación puede representar las diferencias en utilidades entre uno y otro sector, incorporando el costo de hacer el cambio. Asumimos que las características (observables y no observables) determinan dicha utilidad de manera lineal.

[46] Presumimos que los regresores son todos exógenos en el sentido que se definió previamente en este trabajo.

[47] En el contexto de mercado laboral femenino, por ejemplo, no tenemos opción y hay que registrar un 0 o un *No Trabaja* para las mujeres que deciden no entrar al mercado laboral.

En la expresión anterior, X_{2i} son las características que determinan la capacidad de la empresa para obtener beneficios (nivel de capital, experiencia en el sector, nivel de tecnología, etcétera). Colocamos el * en la variable dependiente sin que esta vez se indique que la variable es *latente*, ya que la observamos plenamente, esto más bien se hace para conservar consistencia en la notación. Se puede notar que la ecuación permite que el impacto de los regresores sea distinto dependiendo del sector en el que se opere. Haciendo $\gamma = \alpha + \beta_2$ podemos resumir la estructura que genera las respuestas de la siguiente forma:

$$\begin{array}{ll} y_{1i} = 0 & y_{1i} = 1 \\ y_{2i}^{*} = X_{2i}\beta + \varepsilon_{2i} & y_{2i}^{*} = X_{2i}\gamma + \varepsilon_{2i} \end{array} \qquad (2.30)$$

Las hipótesis de interés pueden ser, en principio, corroborar que el intercepto y las pendientes contenidas en $\beta\ y\ \gamma$ son iguales (esto equivale a la hipótesis de que $\alpha = 0$). También es pertinente en ocasiones contrastar no todos los parámetros en los vectores sino concentrarnos solamente en algunos. En el contexto del ejemplo utilizado en esta sección, podríamos ver si, *ceteris paribus*, una empresa obtiene distintas ganancias si labora en el sector formal o informal, o si los retornos a la inversión en tecnología son iguales en ambos sectores, etcétera. El modelo aquí propuesto puede compararse con el modelo de regresión cambiante, que podemos encontrar en la obra de McFadden (1984), en el que se propone que el componente de error aleatorio cambia según si se toma o no el tratamiento.[48] Dicho modelo es utilizado de forma vasta en la literatura econométrica, y su estimación puede realizarse con la función selection de la paquetería sampleSelection en R.[49] Para el modelo que hemos propuesto en (2.27-2.30) podemos obtener los estimadores de máxima verosimilitud sin mayor problema. La log-verosimilitud a maximizar[50] es:

$$l(\beta, \gamma | y, X) = \sum_{y_{1i}=1} l\big(f(\varepsilon_{2i}|y_{1i} = 0)\big) F_{\varepsilon_{1i}}(-X_i\beta) + \sum_{y_{1i}=0} l\big(f(\varepsilon_{2i}|y_{1i} = 1)\big)(1 - F_{\varepsilon_{1i}}(-X_i\beta)) \qquad (2.31)$$

[48] Si por ejemplo, observamos que las desviaciones típicas son muy distintas, una vez que controlamos por covariables, en ambos sectores, podría ser más adecuado el modelo de regresión cambiante.
[49] La estimación se hace mediante máxima verosimilitud. Para más detalles puede consultarse Toomey y Henningsen (2008).
[50] Note la gran similitud con la expresión escrita en (1.26).

A continuación realizaremos una última generalización del modelo de selección tratado en este trabajo. Hasta ahora hemos estudiado escenarios donde, en la etapa de selección, los individuos sólo deciden entre tomar o no algún tratamiento. Además, la última generalización que teníamos permitía siempre ver la respuesta (en el contexto del ejemplo sobre sector formal o informal, siempre veríamos los beneficios de la empresa). Por otra parte, desde el punto de vista estadístico, teníamos que la (única) respuesta (en el contexto de ejemplo sobre sector formal e informal nos referimos a los beneficios) sólo dependía distribucionalmente de la variable observada (0 ó 1) y no de la variable latente que genera la ecuación de selección.

Primero, consideraremos que en la etapa de selección, los individuos se seleccionan entre niveles de tratamiento. Supondremos que esta elección puede realizarse de acuerdo a un modelo probit ordinal, descrito en el capítulo II (este escenario es plausible, por ejemplo, cuando los individuos seleccionan niveles de aseguramiento). Tenemos entonces que la etapa de selección es modelada de acuerdo a:

$$y_{1i}^{*} = X_{1i}\beta + \varepsilon_{1i} \tag{2.32}$$

$$y_{i1} = \begin{cases} b_1 & si \quad -\infty < y_{1i}^{*} \leq a_1 \\ b_2 & si \quad a_1 < y_{1i}^{*} \leq a_2 \\ & \vdots \\ b_k & si \quad a_{k-1} < y_{1i}^{*} \leq \infty \end{cases} \tag{2.33}$$

Enseguida no limitaremos el modelo a tener una sola ecuación de respuesta. Posiblemente existan escenarios donde estemos interesados en evaluar distintas respuestas que están ligadas a una ecuación de selección. Por ejemplo, en el contexto de programas sociales, interesan diversas características en función de la participación en los programas: cuidado en salud, educación de los hijos, empoderamiento de las mujeres, etcétera. Además, tampoco restringiremos a que las variables respuesta sean observadas; podremos asumir estructuras latentes. Por ejemplo, si estudiamos la relación entre migración y trabajo, el componente de selección tendrá una ecuación para modelar la decisión de migrar o no migrar, mientras que en ocasiones no podremos observar el salario, para aquellos sujetos que no hayan decidido ingresar al mercado laboral. Otro ejemplo respecto al último punto es que las ecuaciones de respuesta pueden nuevamente

modelarse mediante un modelo probit ordinal (por ejemplo, si una de las ecuaciones de respuesta es el número de veces que un individuo asiste a cierto tipo de capacitación). Finalmente consideraremos que el impacto del componente de selección en las respuestas no es mediante la variable observada (tratamiento o no) sino mediante la variable latente del componente de selección.[51]

Incorporando todas las extensiones que hemos descrito en el párrafo anterior, considerando $R-1$ respuestas, tenemos que la modelación adecuada para las mismas es:[52]

$$y_{ri}^{*} = \gamma_r y_{1i}^{*} + x_{ri}\beta_r + \varepsilon_{ri} \tag{2.34}$$

$$y_{ri} = \begin{cases} y_{ri}^{*} \; si \; y_{ri}^{*} > 0 \\ 0 \quad si \; y_{ri}^{*} \leq 0 \end{cases} \tag{2.35}$$

$$r = 2,3,\dots R$$

Para que la modelación en (2.33-2.35) tenga sentido, deben verificarse los supuestos de identificación habituales. Generalmente, estos pueden hacerse sin pérdida de generalidad, o con base en la teoría micro o macroeconómica que den pie al modelo.[53] La estimación de los modelos descritos en (2.33-2.35) no es un tema menor y la discusión al respecto sigue abierta. En esta tesis exploramos el algoritmo MCEM como una alternativa para este fin.

[51] En ocasiones este último aspecto puede llevar a problemas de identificación en el modelo por lo que antes de pasar a la estimación es necesario reflexionar al respecto.

[52] Asumiremos aquí, para ilustrar un contexto típico de selección, que la estructura latente de las ecuaciones de respuesta corresponde a utilidades latentes (sólo observamos respuesta si ésta es mayor que 0), además la utilidad latente del componente de selección sólo afectará mediante un cambio en el intercepto correspondiente a cada ecuación. Reiteramos, la versatilidad de los modelos que se pueden proponer es amplia, adoptamos la forma antes descrita únicamente con fines ilustrativos y teniendo en cuenta que se desea tener herramientas para atender el problema de selección.

[53] Por ejemplo, si en el componente de selección tenemos una respuesta binaria (modelo probit) podemos asumir, sin pérdida de generalidad, que el parámetro de escala de ε_{1i} es 1, y que el umbral que determina la participación en el tratamiento es 0. Si en cambio, el componente de selección es de acuerdo al modelo probit ordinal, entonces el componente aleatorio de la ecuación también debe tener escala fija (habitualmente 1) y algún parámetro umbral en (2.33) debe ser fijo (típicamente $a_1 = 0$). Para más detalles veáse Greene (2012).

Para ilustrar la implementación y el desempeño del MCEM, utilizaremos (además del ejemplo de *análisis de factores* que mencionamos antes) un modelo útil para contextos de autoselección, éste puede encontrarse planteado en Smith (2005).

Suponga que tenemos una muestra de n individuos. De los individuos, observamos 2 respuestas acorde a una utilidad latente censurada en 0.[54] Asumimos que además los individuos seleccionan recibir (o no) cierto tratamiento. Las respuestas observadas tendrán una estructura que será, entre tratados y no tratados, esencialmente la misma, salvo por un cambio en el intercepto (homogéneo dentro de cada grupo, según hayan o no recibido el tratamiento). Asumiremos que las variables latentes inmiscuidas tienen una distribución normal cuya media depende linealmente de parámetros.

Las variables latentes que intervienen pueden modelarse como:

$$\begin{aligned} y_{1i}^{*} &= x_{1i}\beta_1 + \varepsilon_{1i} \\ y_{2i}^{*} &= \gamma_2 y_{1i} + x_{2i}\beta_2 + \varepsilon_{2i} \\ y_{3i}^{*} &= \gamma_3 y_{1i} + x_{3i}\beta_3 + \varepsilon_{3i} \end{aligned} \tag{2.36}$$

Estas utilidades latentes generan las siguientes variables observadas:

$$y_{1i} = \begin{cases} 1 \; si \; y_{1i}^{*} > 0 \\ 0 \; si \; y_{1i}^{*} \leq 0 \end{cases} \tag{2.37}$$

$$y_{2i} = \begin{cases} y_{2i}^{*} \; si \; y_{2i}^{*} > 0 \\ 0 \quad si \; y_{2i}^{*} \leq 0 \end{cases} \tag{2.38}$$

$$y_{3i} = \begin{cases} y_{3i}^{*} \; si \; y_{3i}^{*} > 0 \\ 0 \quad si \; y_{3i}^{*} \leq 0 \end{cases} \tag{2.39}$$

En las expresiones anteriores, x_{1i}, x_{2i} y x_{3i} son vectores de covariables relevantes para la distribución de las utilidades latentes. El supuesto habitual de independencia entre

[54] Por ejemplo, las respuestas pueden ser inversión en ciertos rubros: los individuos sólo invertirán (cantidades positivas) en la medida en que la utilidad por inversión en cada rubro sea grande.

individuos es conservado. Es decir, si hacemos $y_i^* = (y_{1i}^*, y_{2i}^*, y_{3i}^*)^T$, entonces y_i^* es independiente de y_j^*. Considere $\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i})^T$. Tenemos que ε_i sigue una distribución normal multivariada con media $\mu = (0,0,0)^T$ y matriz de covarianzas Σ, donde:

$$\Sigma = \begin{pmatrix} 1 & \sigma_{\varepsilon_1\varepsilon_2} & \sigma_{\varepsilon_1\varepsilon_3} \\ \sigma_{\varepsilon_1\varepsilon_2} & \sigma_{\varepsilon_2}^2 & \sigma_{\varepsilon_2\varepsilon_3} \\ \sigma_{\varepsilon_1\varepsilon_3} & \sigma_{\varepsilon_2\varepsilon_3} & \sigma_{\varepsilon_3}^2 \end{pmatrix} = \begin{pmatrix} 1 & \rho_{\varepsilon_1\varepsilon_2}\sigma_{\varepsilon_2} & \rho_{\varepsilon_1\varepsilon_3}\sigma_{\varepsilon_3} \\ \rho_{\varepsilon_1\varepsilon_2}\sigma_{\varepsilon_2} & \sigma_{\varepsilon_2}^2 & \rho_{\varepsilon_2\varepsilon_3}\sigma_{\varepsilon_2}\sigma_{\varepsilon_3} \\ \rho_{\varepsilon_1\varepsilon_3}\sigma_{\varepsilon_3} & \rho_{\varepsilon_2\varepsilon_3}\sigma_{\varepsilon_2}\sigma_{\varepsilon_3} & \sigma_{\varepsilon_3}^2 \end{pmatrix} \tag{2.40}$$

En principio, hemos determinado que el 1 en la primera entrada de la matriz de covarianzas se establece por las razones habituales en el modelo probit. En el siguiente capítulo especificáremos valores para esta matriz de covarianzas, además de la forma particular que tomarán los valores de las covariables, así como los valores para cada uno de los $\beta_j's$; todo esto con el objetivo de ilustrar el algoritmo MCEM como herramienta para obtener estimaciones para los parámetros en el modelo.

IV. ALGORITMO MONTE CARLO EXPECTATION MAXIMIZATION

La optimización numérica resulta fundamental para poder dar solución a problemas que otrora eran dejados a un lado. Los modelos de alta dimensionalidad (en parámetros), así como funciones de verosimilitud que no tienen "forma cerrada", aparecen cada vez más en el quehacer estadístico. La idea central en este capítulo es ilustrar algunos métodos computacionalmente intensivos como alternativas para la estimación, particularmente concentrándonos en los modelos que incorporan variables latentes y los modelos de selección.

La primera sección del capítulo versa sobre el muestreador de Gibbs. En la segunda sección se escribe sobre el algoritmo EM y su implementación cuando el paso E es mediante simulación de Montecarlo (lo que genera el algoritmo MCEM). La tercera sección trata la implementación del MCEM al ejemplo de análisis de factores estudiado en la sección anterior. El objetivo ulterior en este capítulo es realizar la estimación del modelo de autoselección propuesto en (3.36-3.40), esto se realiza en la cuarta y última sección.

A. MUESTREADOR DE GIBBS

La inferencia estadística que incorpora el enfoque bayesiano tiene como principal objetivo establecer una distribución posterior para los parámetros (establecida con base en la distribución a priori y la evidencia proporcionada por los datos). En muchas situaciones, como exploraremos en breve, no se tiene una forma conocida para dicha distribución. La opción inmediata (en vista de que el kernel de la distribución es siempre conocido) es simular de la misma, para obtener información sobre los valores plausibles de los parámetros a la luz de los datos y la distribución posterior. Los métodos de Monte Carlo, en particular el algoritmo Metropolis-Hastings, son herramientas útiles para esta tarea. El algoritmo Metropolis-Hastings se basa en la construcción de una cadena de Markov cuya distribución asintótica es la distribución objetivo (posterior). Enseguida, introduciremos el muestreador de Gibbs, otro método de Montecarlo común en la práctica.[55]

[55] Puede verse al muestreador de Gibbs como una configuración particular del algoritmo Metropolis Hastings.

La idea central es que tenemos $X = (X_1, \dots, X_p)$, un vector aleatorio del que queremos simular. Supongamos que para dicho vector se tiene una distribución conjunta de la que no existen métodos *ad hoc* para simular.[56] Las entradas $X_i's$ del vector anterior las asumimos como variables aleatorias unidimensionales o multidimensionales, que tienen como principal característica que podemos simular fácilmente de ellas. Es decir, tenemos las distribuciones $f_1, \dots, f_p$ tales que:

$$X_i | x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_p \sim f_i(x_i | x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_p)$$

$$i = 1,2, \dots, p \tag{4.1}$$

Las distribuciones condicionales exhibidas anteriormente reciben el nombre de condicionales totales. La idea central del muestreador es construir una cadena de Markov $X^{(t)}$ que tenga como distribución asintótica la distribución objetivo f. Enseguida mostramos el algoritmo que genera la cadena requerida.

ALGORITMO MUESTREADOR DE GIBBS

Se tienen valores iniciales $X^{(0)}$ que cumplen $f(X^{(0)}) > 0$. En la iteración $t = 1,2, \dots$ dado $x^{(t)} = (x_1^{(t)}, \dots, x_p^{(t)})$, generar

1. $X_1^{(t+1)} \sim f_1(x_1 | x_2^{(t)}, \dots, x_p^{(t)})$;
2. $X_2^{(t+1)} \sim f_2(x_2 | x_1^{(t+1)}, x_3^{(t),} \dots, x_p^{(t)})$;

$\vdots$

p. $X_p^{(t+1)} \sim f_p(x_p | x_1^{(t+1)}, \dots, x_{p-1}^{(t+1)})$.

La teoría sobre la convergencia (a la distribución objetivo) del algoritmo puede consultarse en el texto de Robert y Casella (2004). Otro aspecto importante es que, aun asumiendo que la cadena se encuentra en su distribución estacionaria, los vectores aleatorios que encontramos en cada iteración estarán correlacionados. De este modo, y como se hace de manera usual con los métodos de Montecarlo, consideramos una etapa

[56] Esta *condición* es más bien una condición práctica; si los métodos ad hoc están disponibles, presumimos que serán preferibles a la simulación mediante el muestreador de Gibbs.

inicial (etapa de quemado) para permitir que la cadena llegue a la distribución estacionaria, y después consideramos un espacio entre las variables generadas para poder obtener una muestra cuyas observaciones podamos presumir como independientes e idénticamente distribuidas.

Presumiendo una distribución unimodal, un criterio de convergencia heurístico es identificar que los valores simulados se concentren en una zona común durante un periodo considerable de iteraciones. Los ejemplos que trataremos necesitan un periodo de quemado bastante corto.[57] Para considerar la independencia de las observaciones generadas por el muestreador, utilizaremos el gráfico de autocorrelaciones que se tiene en cada muestra. No es la idea en este trabajo ahondar en estos temas, pero recomendamos al lector ser cuidadoso en el tratamiento de los mismos, con el objetivo de no obtener simulaciones que pudieran resultar inadecuadas.

Para ilustrar los conceptos enunciados hasta el momento, consideremos el siguiente ejemplo. Suponga que se desea simular de $(X, Y) \sim N_2(0, \Sigma)$, donde:

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{4.2}$$

Las condicionales totales en este ejemplo son:

$$X|Y \sim N(\rho Y, (1-\rho^2)) \tag{4.3}$$

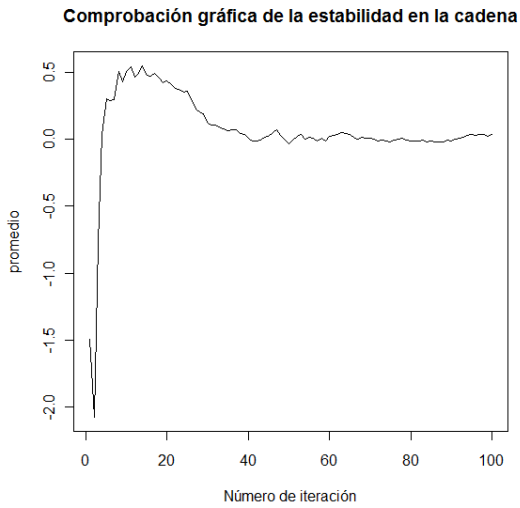
$$Y|X \sim N(\rho X, (1-\rho^2)) \tag{4.4}$$

En primera instancia notamos de manera trivial que si $\rho = 0$ el muestreador de Gibbs converge desde la primera iteración, y las observaciones generadas en cada iteración son independientes. De hecho, podemos ver cómo es la estructura de correlación de las variables, generadas, note por ejemplo que:

$$X_{t+k}|X_k \sim N(\rho^{2k}, 1-\rho^{4k}) \tag{4.5}$$

[57] Cameron y Trivedi (2005) y Yu *et al*. (2005).

De dicho resultado se observa que sin importar el valor inicial, asintóticamente la distribución de la cadena converge a la distribución marginal objetivo. En ejemplos más complejos determinar la correlación entre las variables suele ser mucho más difícil. Enseguida ilustramos los métodos gráficos que suelen usarse en dicha circunstancia. Consideremos $\rho = .2$. En la siguiente gráfica observamos el comportamiento de los promedios obtenidos a partir de la muestra simulada mediante la cadena:



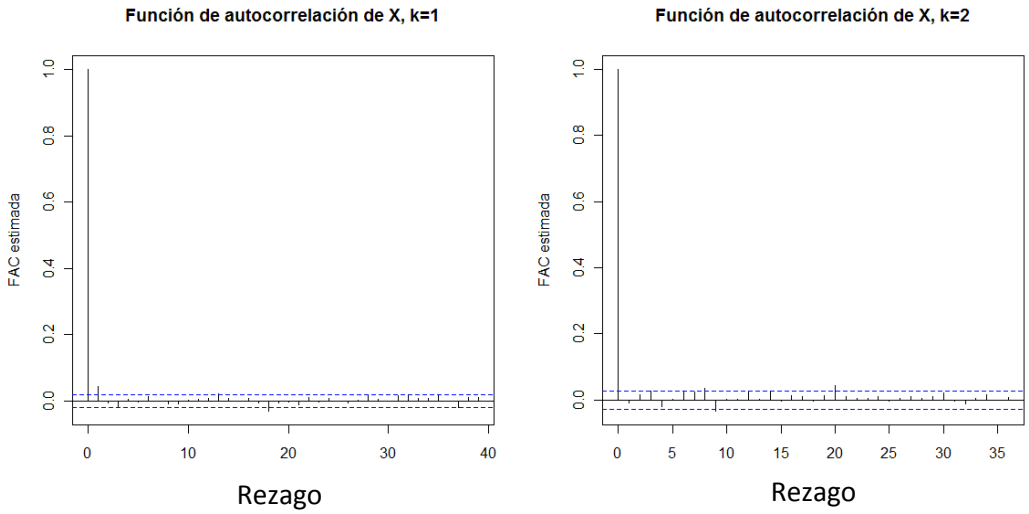
Puede percibirse que alrededor de las 50 iteraciones la cadena se estabiliza. Por supuesto este experimento debe ser repetido hasta asegurarse de que se tiene mayor certeza del número de iteraciones necesarias para converger a la distribución estacionaria.[58] Así, podemos determinar un periodo de quemado de 50 iteraciones.[59] Ahora supongamos que una vez en la distribución estacionaria, decidimos tomar simulaciones consecutivas, digamos hasta obtener una colección de 1000 observaciones. Una forma de verificar que la correlación entre las mismas es lo suficientemente pequeña como para tratar la muestra obtenida como independiente es estudiar la función de autocorrelación.

Enseguida mostramos las funciones de autocorrelación considerando que muestreamos de manera equidistante cada 1 y 2 iteraciones respectivamente. La idea es

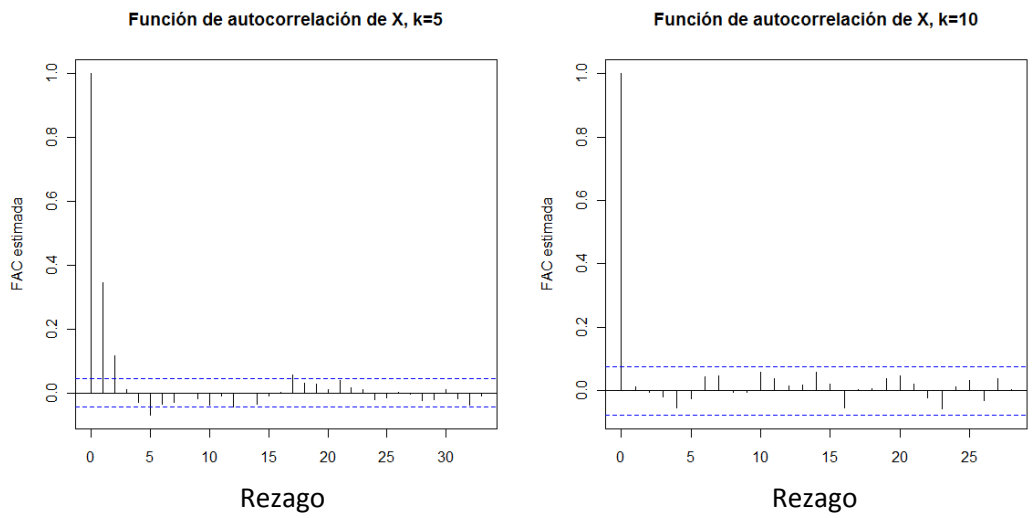
[58] Aquí tratamos con un ejemplo conocido, y sabemos que la moda de X está en 0. En ejemplos más complejos (por ejemplo de muy alta dimensionalidad con variables multimodales) no podemos asegurar que hemos llegado a la distribución estacionaria, pese a observar que la cadena permanece en una zona constante un período prolongado de tiempo.

[59] Este es un número poco conservador, y debido al costo casi nulo que tiene cada iteración podríamos imponer un número mucho más alto. De cualquier modo este ejemplo es sólo para ilustrar los conceptos.

establecer k, el número de simulaciones que dejaremos pasar para conseguir una muestra que sea convincente en cuanto a los supuestos de independencia que deseamos.



Para convencerse de la importancia de la correlación, estudiaremos un caso extremo en el que $\rho = .9$ (ya mencionamos el caso en el que las entradas son independientes y un caso intermedio en el que $\rho = .2$). El periodo de quemado obtenido con base en resultados gráficos es bastante similar al ejemplo estudiado recientemente. Como el lector sospechará, el tema más importante a estudiar es la correlación en las simulaciones (que con base en el resultado analítico en (4.5) esperamos sea alta). Veremos las gráficas de autocorrelación para X muestreando cada 5 y cada 10 simulaciones, respectivamente.



La gráfica anterior ilustra cómo aun muestreando cada 5 observaciones (un número usual en la práctica) pueden existir problemas de correlación. Cuando se emplea el muestreador de Gibbs es siempre importante verificar dicha cuestión, usualmente puede corregirse el problema muestreando con una mayor distancia en iteraciones, sin que esto incremente excesivamente el tiempo de cómputo (en el lado derecho de la gráfica puede verificarse que la autocorrelación es casi eliminada al muestrear cada 10 iteraciones).

Enseguida ilustraremos la utilidad e implementación del muestreador de Gibbs en un ejemplo menos sintético. Recordemos el modelo SUR desarrollado en el capítulo anterior. Consideraremos sólo 2 respuestas para cada individuo, es decir, se tiene el siguiente modelo:

$$y_{ji} = x'_{ji}\beta_j + \varepsilon_j$$

$$j = 1,2; \qquad i = 1, \dots, N \tag{4.6}$$

$$\varepsilon_i \sim N_2(0, \Sigma) \tag{4.7}$$

Tal y como se explicó en el capítulo anterior, este modelo puede ser útil, por ejemplo, para estudiar la demanda por 2 bienes. Recordando el proceso generador de los datos, y las distribuciones a priori tenemos:

$$y_i|x_i, \beta, \Sigma \sim N(x_i'\beta, \Sigma) \tag{4.8}$$

$$\beta \sim N(\beta_0, B_0^{-1}) \tag{4.9}$$

$$\Sigma^{-1} \sim Wishart(v_0, D_0) \tag{4.9}$$

Del kernel de la función de distribución posterior expresada en 2.2V tenemos que las densidades condicionales (condicionales totales de $\beta\ y\ \Sigma$) son:

$$\beta|\Sigma,\ y, X \sim N[C_0 A_0, C_0] \tag{4.10}$$

$$A_0 = \left(B_0\beta_0 + \sum_{i=1}^{N} x_i'\ \Sigma^{-1} y_i\right) \tag{4.11}$$

$$C_0 = \left(B_0 + \sum_{i=1}^{N} x_i'\ \Sigma^{-1} x_i\right)^{-1} \tag{4.12}$$

$$\Sigma^{-1}|\beta,\ y, X \sim Wishart[v_0 + N, H_0] \tag{4.13}$$

$$H_0 = \left(D_0^{-1} + \sum_{i=1}^{N} u_i u_i'\right)^{-1} \tag{4.14}$$

$$u_i = y_i - x_i'\beta \tag{4.15}$$

Para las distribuciones *a priori* establecemos los siguientes valores:

$$\beta_0 = 0\ ;\ B_0 = \tau I;\quad D_0 = I;\quad v_0 = 5 \tag{4.16}$$

En la especificación anterior τ es un parámetro que refleja nuestra certeza sobre la distribución de β. Para este ejercicio hemos escogido utilizar $\tau = 1$. Los parámetros elegidos para la simulación son todos 1 (ambos interceptos, pendientes, varianzas) salvo el coeficiente de correlación que fue escogido como -.5. Hemos escogido $N = 500$ y hemos simulado los regresores x_{1i} y x_{2i} de la distribución normal estándar. Con los componentes vistos hasta el momento, y en vista de que las densidades posteriores condicionales de β y Σ han resultado ser distribuciones conocidas, podemos implementar el muestreador de Gibbs y obtener así la distribución posterior de los parámetros.

En este ejemplo particular, los pasos que seguimos para la simulación quedan de la siguiente forma:

Como valores iniciales tenemos $\beta^{(0)} = 0$ y $\Sigma^{(0)} = I$.[60] En la iteración t construimos:

$$A_0^{(t)} = \left(B_0\beta_0 + \sum_{i=1}^{N} x_i' \, \Sigma^{-1^{(t-1)}} y_i\right) \tag{4.17}$$

$$C_0^{(t)} = \left(B_0 + \sum_{i=1}^{N} x_i' \, \Sigma^{-1^{(t-1)}} x_i\right)^{-1} \tag{4.18}$$

Con estos valores simulamos un vector X de la distribución normal multivariada con media $A_0^{(t)}C_0^{(t)}$ y matriz de covarianzas $C_0^{(t)}$.

Fijamos:

$$\beta^{(t)} = X. \tag{4.19}$$

Construimos:

$$u_i^{(t)} = y_i - x_i'\beta^{(t)}, \; i = 1,2,\dots,N \tag{4.20}$$

para formar:

$$H_0^{(t)} = \left(D_0^{-1} + \sum_{i=1}^{N} u_i^{(t)} \left(u_i^{(t)}\right)'\right)^{-1} \tag{4.21}$$

Finalmente, simulamos Y de la distribución Wishart con $v_0 + N$ grados de libertad y matriz de escala $H_0^{(t)}$.

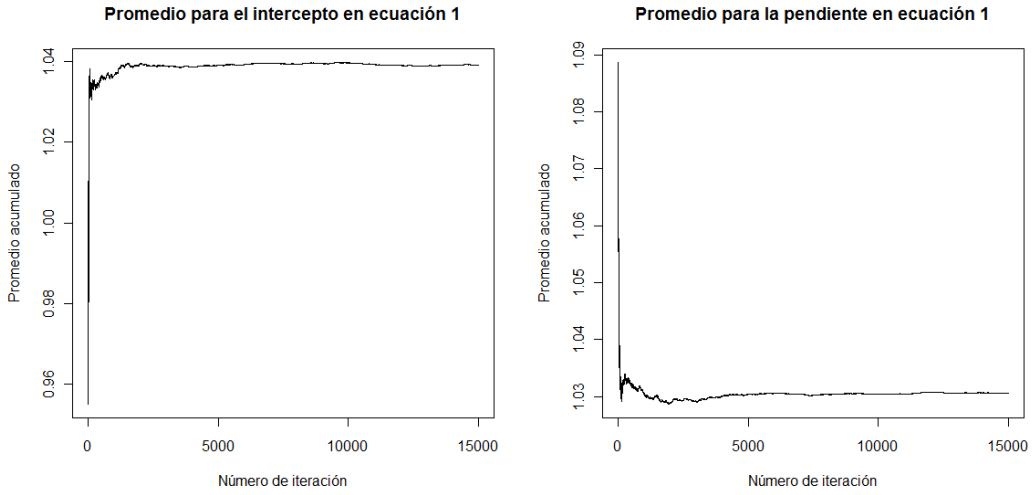
Fijamos:

[60] El lector podrá notar que sólo serán relevantes los valores asignados a la matriz $\Sigma^{(0)}$, puesto que en el algoritmo propuesto sólo este valor es utilizado en la primera iteración.

$$\Sigma^{(t)} = Y^{-1} \tag{4.22}$$

El algoritmo propuesto fue implementado, y con base en criterios gráficos determinamos un periodo de quemado de 5000 iteraciones. No detectamos correlación importante de las simulaciones, pero para mayor protección muestreamos cada 2 iteraciones a partir del periodo de quemado hasta conseguir 5000 simulaciones de los parámetros de interés.[61]

Para ilustrar lo descrito en el párrafo anterior, mostramos el comportamiento de los promedios del intercepto de la primera ecuación (β_{01}) y de la pendiente en la misma ecuación (β_{11}) con el fin de mostrar que cuando hemos hecho 5000 iteraciones se estabilizan.[62] De ahí que este valor fuera escogido como el número de iteraciones en el periodo de quemado.

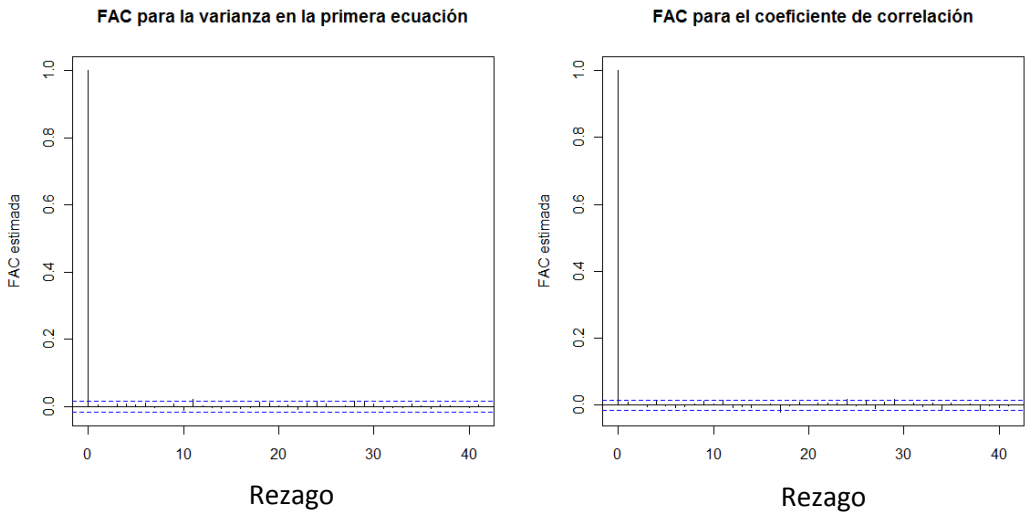


Enseguida evaluamos la correlación de las muestras generadas; en esta ocasión elegimos a la varianza del componente estocástico de la primera ecuación (σ_1^2) y al coeficiente de correlación (ρ). Observamos que no existen correlaciones importantes, por

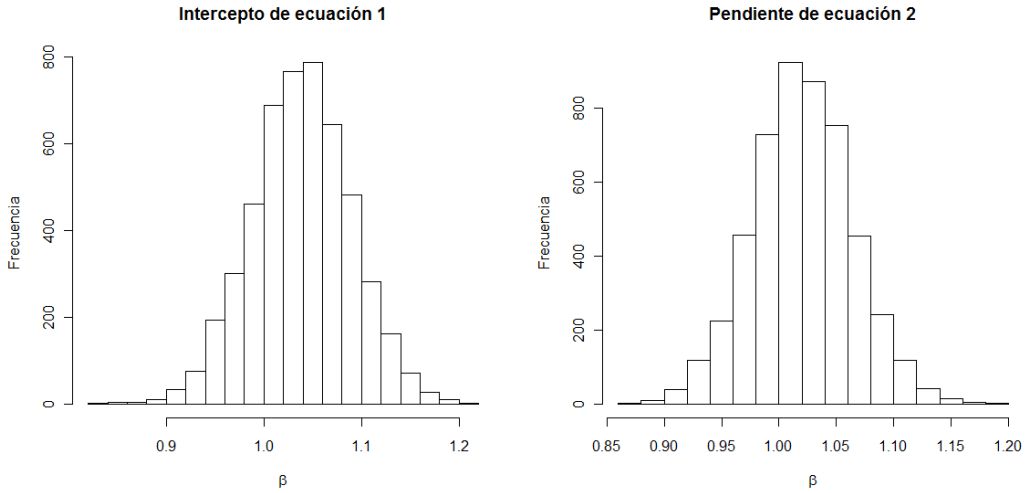
[61] En la práctica, podríamos obtener muestras mucho más grandes (de tamaño 100,000 o 500,000 por ejemplo) con el fin de tener mayor información sobre la distribución posterior. En este ejemplo con fines ilustrativos hemos escogido un número bastante menor (5,000).

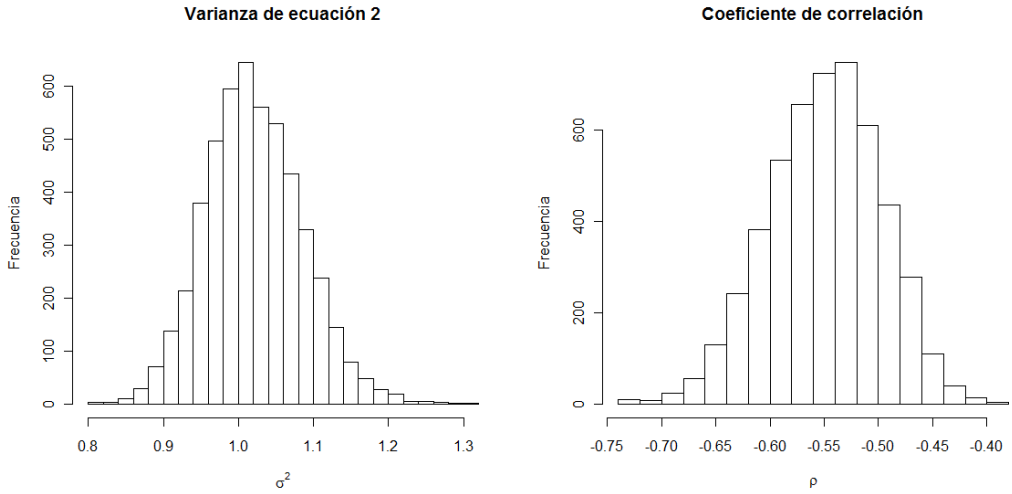
[62] Este ejercicio se realiza con todos los parámetros, pero por cuestiones de espacio ilustramos el concepto sólo con los dos parámetros referidos.

lo que podríamos muestrear todas las observaciones simuladas a partir de la iteración 5001, sin embargo, en una actitud conservadora (como suele hacerse en la práctica), decidimos muestrear cada 2 iteraciones.



Finalmente y como verificación de que el muestreador de Gibbs y el modelo propuesto son sensatos, graficamos la distribución (vía histogramas) de algunos de los parámetros en el modelo. Éstos serán el intercepto de la primera ecuación (β_{01}), la pendiente de la segunda ecuación (β_{12}), la varianza del componente aleatorio en la segunda ecuación (σ_2^2) y el coeficiente de correlación (ρ). Podremos observar que el resultado de la inferencia realizada es adecuado.





Los métodos estadísticos computacionalmente intensivos no sólo surgen de manera natural bajo el enfoque bayesiano. Por ejemplo, cuando se hace inferencia utilizando únicamente el criterio de Máxima Verosimilitud es posible encontrar que la función de verosimilitud resulta difícil de optimizar. Enseguida hablaremos del algoritmo EM y su implementación en estos casos.

B. Algoritmo EM y Algoritmo MCEM

Especialmente en modelos de variables latentes y modelos con censura o mezcla de distribuciones, surgen funciones de verosimilitud que están en términos de integrales, típicamente de alta dimensionalidad, cuya optimización puede resultar un problema más enmarcado en los métodos numéricos que en la estadística misma.[63]

Para resolver el problema planteado en líneas anteriores, y particularmente en el contexto de datos faltantes, desde 1977 Dempster *et al.* introdujeron como alternativa el algoritmo Expectation Maximization. La idea básica con este algoritmo es conceptualizar que si tuviéramos más información disponible, la función de verosimilitud con la que trabajaríamos sería más fácil de optimizar; como esto no ocurre, optimizamos el valor

[63] En este sentido, es posible que los métodos numéricos sean vistos como una caja negra y su uso pueda no ser deseable.

esperado de la función de verosimilitud tratable. Enseguida formalizamos la estructura y noción del algoritmo.[64]

Supongamos que tenemos una muestra $X_1, \dots, X_n$ de datos. La verosimilitud es siempre proporcional a la densidad conjunta de la muestra, que denotaremos por $g(x|\theta)$. Si esta función es fácil de maximizar, entonces los estimadores de máxima verosimilitud podrán hallarse sin mayor dificultad. Suponga que existe una densidad f tal que g cumple que:

$$g(x|\theta) = \int_Z f(x, z|\theta) dz. \tag{4.24}$$

Queremos hallar $\hat{\theta} = argmax\, L(\theta|x) = argmax\, g(x|\theta)$. Observe que en vez de trabajar únicamente con los datos obtenidos en la muestra (X) tendremos ahora un vector de datos aumentados, que en la literatura conocemos como datos completos, que denotamos como (X, Z) cuya distribución tiene densidad $f(x, z|\theta)$. A partir de esta relación y con la marginalización correspondiente, obtenemos la densidad condicional de los datos faltantes dados los datos observados: $k(z|\theta, x) = f(x, z|\theta)/g(x|\theta)$.

Con base en la relación establecida anteriormente, y tomando logaritmos y esperanzas respecto de $k(z|\theta_0, x)$ es evidente que:

$$logL(\theta|x) = \mathbb{E}_{\theta_0}[logL^C(\theta|x, Z)] - \mathbb{E}_{\theta_0}[logk(Z|\theta, x)] \tag{4.25}$$

Como buscamos maximizar la expresión anterior nos concentraremos en el primer elemento del lado derecho. En cada iteración debemos hallar la esperanza de (esto constituye el paso E):

$$Q(\theta|\theta_0, x) = \mathbb{E}_{\theta_0}[logL^C(\theta|x, Z)] \tag{4.26}$$

Ya que obtenemos la esperanza anterior, es momento de maximizarla (esto constituye el paso M). Si en la iteración t tenemos que el valor de que maximiza $Q(\theta|\theta_0, x)$

[64] La mayor parte de la notación e ideas utilizadas en esta sección fueron tomadas de Robert y Casella (2010).

es $\hat{\theta}$ entonces en la iteración $t+1$ reemplazamos θ_0 por $\hat{\theta}$. Repetimos el procedimiento hasta convergencia. La idea es obtener una sucesión de estimadores $\{\hat{\theta}_{(j)}\}_j$ que sea consistente. Observe entonces que $\hat{\theta}_{(j)}$ se obtiene al maximizar $Q(\theta|\hat{\theta}_{(j-1)}, x)$, es decir:

$$Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j-1)}, x) = max_{\theta}\, Q(\theta|\hat{\theta}_{(j-1)}, x) \tag{4.27}$$

En el siguiente recuadro presentamos un resumen del algoritmo EM.

ALGORITMO EM

Elija un valor inicial de $\hat{\theta}_{(0)}$

Repita

1. Compute (*Paso E*)

$$Q(\theta|\hat{\theta}_{(m)}, x) = \mathbb{E}_{\hat{\theta}_{(m)}}[logL^C(\theta|x, Z)]$$

Donde la esperanza es con respecto a $k(z|\hat{\theta}_{(m)}, x)$.

2. Maximice $Q(\theta|\hat{\theta}_{(m)}, x)$ en θ y tome (*Paso M*)

$$\hat{\theta}_{(m+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\hat{\theta}_{(m)}, x)$$

Actualice $m = m + 1$

Esto se realiza hasta convergencia, es decir, hasta que $\hat{\theta}_{(m+1)} = \hat{\theta}_{(m)}$.

El algoritmo se presume convergente[65] en vista de que $L(\hat{\theta}_{(j+1)}|x) \geq L(\hat{\theta}_{(j)}|x)$ y porque la igualdad se alcanza si y sólo si $Q(\hat{\theta}_{(j+1)}|\hat{\theta}_{(j)}, x) = Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j)}, x)$. Note que la presunta convergencia está garantizada si la función a optimizar es unimodal. En otro caso, lo único que nos garantiza el resultado que mencionamos anteriormente, es que el algoritmo nos arrojará una sucesión que converge a algún punto crítico de la función

[65] Obviamente nos referimos a la *convergencia* al máximo global de la función objetivo.

objetivo, que no necesariamente es el máximo global. En vista de lo anterior, una recomendación general cuando se implementa el EM es que, a menos que se esté muy seguro sobre la ubicación del máximo (o se sepa que la función objetivo es unimodal) se pruebe el algoritmo con distintos valores de inicio.[66]

Para ilustrar el algoritmo y familiarizarse con las ideas que involucra consideremos el siguiente ejemplo. En confiabilidad, típicamente se trabaja con tiempos de vida. Los experimentos comunes en el área consisten en evaluar el tiempo de vida de componentes bajo ciertas condiciones. El objetivo es caracterizar la distribución del tiempo de vida de dichos componentes. Los costos de experimentación provocan que no se pueda "concluir" el experimento, en el sentido de que éste se considera por terminado aun cuando algunos componentes no hayan fallado. Asumiremos que el verdadero tiempo de vida, observado o no, tiene una distribución exponencial.

En el ejemplo anterior podemos conceptualizar como variable latente al tiempo de vida (real) de los componentes; mientras tanto, consideraremos los tiempos de vida registrados como los datos observados. En sincronía con la notación utilizada en capítulos anteriores podemos escribir el siguiente modelo:

$$y_i{}^* \sim exp(\lambda) \tag{4.28}$$

$$y_i = \begin{cases} y_i{}^* si\ y_i{}^* < D \\ D\ si\ y_i{}^* \geq D \end{cases} \tag{4.29}$$

Podemos considerar en este ejemplo que y_i^*'s son los datos faltantes (variables latentes) y y_i son los datos observados. En este ejemplo, cabe resaltar, podría maximizarse la verosimilitud exacta sin necesidad de utilizar el algoritmo EM, pero lo utilizamos para introducir e ilustrar el algoritmo. La log-verosimilitud de los datos completos es:

$$l^c(\theta|y^*) = \sum_{y_i<D} [log(\lambda) - \lambda y_i^*] + \sum_{y_i \geq D} [log(\lambda) - \lambda y_i^*] \tag{4.30}$$

[66] Para mayores detalles sobre las propiedades de convergencia, una referencia básica es la de Wu (1983).

Tomando la esperanza, condicional en los datos observados y el parámetro λ determinado en la iteración anterior, tenemos que, en la iteración t:

$$E[l^c(\theta|y^*)|y_i,\lambda_{t-1}] = \sum_{y_i<D} E\left[log(\lambda) - \lambda y_i^*|y_i\right] + \sum_{y_i\geq D} E\left[log(\lambda) - \lambda y_i^*|y_i\right] =$$

$$\sum_{y_i<D}\left[log(\lambda) - \lambda y_i\right] + \sum_{y_i\geq D}\left[log(\lambda) - \lambda E(y_i^*|y_i)\right] \quad (4.31)$$

En este caso, debemos notar que la esperanza que falta calcular en la ecuación anterior es:

$$E[y_i^*|y_i] = E[y_i^*|y_i^* \geq D] = \frac{\lambda D+1}{\lambda} \quad (4.32)$$

Con esto completamos el paso E y tenemos, en la iteración t:

$$E[l^c(\theta|y^*)|y_i,\lambda_{t-1}] = \sum_{y_i<D}\left[log(\lambda) - \lambda y_i\right] + \sum_{y_i\geq D}\left[log(\lambda) - \lambda\cdot\left(\frac{\lambda_{t-1}D+1}{\lambda_{t-1}}\right)\right] = n\,log(\lambda) -$$

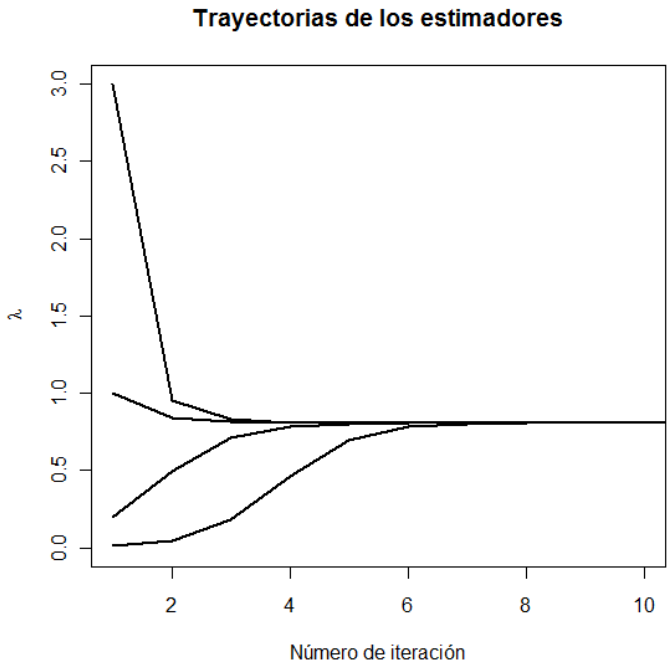
$$\lambda\left[K + m\left(\frac{\lambda_{t-1}D+1}{\lambda_{t-1}}\right)\right] \quad (4.33)$$

En la ecuación anterior tenemos que m es el número de observaciones censuradas y $K = \sum_{y_i<D} y_i$. Enseguida tendremos que completar el paso M, en el que estamos interesados en maximizar la ecuación anterior. Por la condición de primer orden es fácil ver que el óptimo de la función anterior se encuentra en (durante la iteración t):

$$\lambda_t = n\left[K + m\left(\frac{\lambda_{t-1}D+1}{\lambda_{t-1}}\right)\right]^{-1} \quad (4.34)$$

Para ilustrar el algoritmo, simulamos 100 datos considerando $\lambda = 1$ y $D = 2$. La simulación realizada arroja 79 observaciones por debajo del umbral y 21 datos censurados. Procedemos en cada iteración, actualizando el estimador λ con base en (4.34). Presumimos la convergencia cuando el valor del parámetro estimado se estabiliza.

En este ejemplo la convergencia se da dentro de las 8 primeras iteraciones, sin importar los valores iniciales. Para finalizar con el ejemplo colocamos una gráfica que muestra la evolución de los estimadores partiendo de distintos puntos iniciales.



Pese a lo útil del algoritmo EM cuando están presentes escenarios con variables latentes, en muchas ocasiones la esperanza del paso E no es sencilla de calcular. Cuando éste es el caso, se pueden utilizar Métodos de Montecarlo para simular la log-verosimilitud de los datos completos y posteriormente aproximar la esperanza de la misma como el simple promedio de las log-verosimilitudes simuladas. Este método es conocido en la literatura como el algoritmo Monte Carlo Expectation Maximization (MCEM). En muchas ocasiones las variables aleatorias a simular se prestan para aplicar el muestreador de Gibbs mencionado al inicio de este capítulo.

C. Implementación al ejemplo de Análisis de Factores

Enseguida, estudiaremos cómo el algoritmo MCEM puede adaptarse a algunos ejemplos que hemos tratado en esta tesis. En primera instancia, exploraremos su potencial para el modelo de *análisis de factores* descrito en el capítulo anterior. Nos enfocaremos en detallar cuáles son los pasos a seguir y cómo puede implementarse el algoritmo para dar

solución al problema de estimación correspondiente. Por último, recapitularemos la generalización del modelo de selección planteada en la parte final del capítulo anterior. Analizaremos detenidamente el modelo e ilustraremos cómo el algoritmo MCEM puede dar solución al problema de estimación correspondiente.[67] Finalizaremos simulando una estructura de selección y realizaremos la estimación correspondiente utilizando el algoritmo MCEM.

Consideremos el modelo de *análisis de factores* descrito en la ecuación 2.14 y justificado en el segundo capítulo de este trabajo. Tenemos que la matriz con los ordenamientos, R, constituye la información disponible, mientras que la matriz con las utilidades latentes (X) y los factores latentes (Z) constituyen los datos faltantes.[68] La log-verosimilitud de los datos completos es:

$$l^c(\theta|X,Z) = -\frac{n}{2}\sum_j log\left(\sigma_j^2\right) - \frac{1}{2}\sum_i\sum_j \frac{(x_{ij}-z_i^T a_j - b_j)^2}{\sigma_j^2} \qquad (4.35)$$

Es inmediato notar que las estadísticas suficientes que determinan la log-verosimilitud antes mencionada son $\{X^TX,\ Z^TZ, Z^TX, X^T1, Z^T1\}$. Recuerde que las variables z_i y x_i siguen conjuntamente una distribución normal multivariada. En consecuencia, las distribuciones condicionales de dichas variables son también normales multivariadas.

En la iteración t del algoritmo MCEM, para obtener la esperanza de la log-verosimilitud, condicional en los datos observados (ordenamientos) debemos simular del vector $(z_i, x_i|\theta_{t-1}, R)$ y posteriormente calcular las estadísticas suficientes inmersas para tener entonces una simulación de la log-verosimilitud de los datos completos condicional en los datos observados. Repetimos esto un número M de veces y fijamos la esperanza requerida en el paso E como el promedio de las log-verosimilitudes simuladas.

Para la simulación del vector $(z_i, x_i|\theta, r_i)$podremos utilizar de manera natural el muestreador de Gibbs, ya que las condicionales totales $(z_i|x_i, \theta, r_i)$ y $(x_i|z_i, \theta, r_i)$ siguen una distribución normal multivariada con ciertas restricciones. En el proceso de simulación

[67] Como se mencionó antes, este es el objetivo central en este capítulo.
[68] Observe cómo los datos faltantes determinan unívocamente los datos observados (de hecho X determina total y unívocamente a *R*), pero no viceversa. Esto ocurre típicamente en modelos de variable latente.

comenzaremos por simular de $z_i|x_i, r_i, \theta$ aprovechando la ventaja de que la distribución condicional de $z_i|x_i, r_i, \theta$ no depende de los ordenamientos:[69]

$$f(z_i|x_i, r_i, \theta) = f(z_i|x_i, \theta) \tag{4.36}$$

Entonces, utilizando teoría básica de la distribución normal multivariada (recuerde que z_i, x_i son conjuntamente normales) podemos ver que la distribución condicional de z_i dados los datos y dado el vector x_i es:

$$z_i|x_i, \theta \sim N_d \{A(A^T A + \Psi)^{-1}(x_i - b), I - A(A^T A + \Psi)^{-1}A^T\} \tag{4.37}$$

De este modo, tenemos que en la iteración m del muestreador de Gibbs, estando en la iteración t del algoritmo MCEM debemos simular un vector y tal que:

$$y \sim N_d \{A_{t-1}\left(A_{t-1}{}^T A_{t-1} + \Psi_{t-1}\right)^{-1}\left(x_i^{(m-1)} - b_1\right), I - A_{t-1}\left(A_{t-1}{}^T A_{t-1} + \Psi_1\right)^{-1} A_{t-1}{}^T\} \tag{4.38}$$

Hecho lo anterior, establecemos:

$$z_i^{(m)} = y \tag{4.39}$$

Dentro de la iteración t del algoritmo MCEM, la simulación anterior constituye el paso 1, en cada una de las M simulaciones que deben realizarse durante el paso E (en el que se utiliza el muestreador de Gibbs).

Es conveniente tener el vector de ordenamientos $\langle j_1, \ldots, j_k \rangle_i$ mencionado en el capítulo anterior. Definimos $x_{i,j_0} = \infty$ y $x_{i,j_{k+1}} = -\infty$. Así, en el paso 2 del muestreador de Gibbs procedemos a simular de $x_i|z_i, r_i, \theta$. Una alternativa es simular una variable aleatoria w tal que:

[69] Más que *no depender de los ordenamientos* nos referimos a que x_i determina unívocamente los ordenamientos; es decir, la densidad de (x_i, r_i) es la densidad de x_i.

$$w \sim N_k({A_{t-1}}^T z_i^{(m)} + b_{t-1}) \tag{4.40}$$

Hecho lo anterior, verificamos que $w_{j_s-1} > w_{j_s} > w_{j_{s+1}}$ para $s = 1,2,\dots,k$. Si esto ocurre, entonces fijamos:

$$x_i^{(m)} = w \tag{4.41}$$

Si esto no ocurriera, repetimos la simulación de w hasta encontrar una simulación que cumpla las restricciones mencionadas. El lector podrá sospechar que el algoritmo propuesto puede resultar bastante ineficiente. Revisaremos otra alternativa. Note primero que:

$$x_{i,j_s} | \, x_{i,j_1}, \dots, x_{i,j_{s-1}}, x_{i,j_{s+1}}, \dots, x_{i,j_k}, r_i, z_i, \theta \sim N(z_i^T a_{j_s} + b_{j_s}, \sigma_{j_s}^2) \tag{4.42}$$

Sujeta a que $x_{i,j_{s-1}} > x_{i,j_s} > x_{i,j_{s+1}}$. Es decir, x_{i,j_s} tiene distribución normal truncada. Entonces podemos simular, empezando desde $s = 1$, variables w_{i,j_s} de la distribución $N\left(\left(z_i^{(m)}\right)^T \left(a_{j_s}\right)_{t-1} + \left(b_{j_s}\right)_{t-1}, \left(\sigma_{j_s}^2\right)_{t-1}\right)$ truncada de forma que $w_{i,j_{s-1}} > w_{i,j_s} > w_{i,j_{s+1}}$, y entonces tenemos que las entradas de $x_i^{(m)}$ se fijarán considerando $x_{i,j_s}^{(m)} = w_{i,j_s}$.

Una vez hechas las M iteraciones (con los 2 pasos descritos anteriormente) podremos obtener una muestra del vector $(Z, X|\theta_{t-1}, R)$,[70] y por ende una muestra de los estadísticos suficientes necesarios para determinar la log-verosimilitud. Obteniendo un promedio de las log-verosimilitudes simuladas tendremos una aproximación de la esperanza $E[l^c(\theta|X,Z)|\theta_{t-1}, R]$. Hecho lo anterior, podemos proceder al paso M de la iteración t en el algoritmo MCEM. Como es de esperarse, el problema de maximización que enfrentamos se ha convertido en un problema de regresión lineal multivariada, por lo que de inmediato podemos encontrar los óptimos de la función de verosimilitud.

[70] Para esto habrá que determinar el periodo de quemado y explorar la autocorrelación de las variables en pos de obtener una muestra casi independiente que pueda ser utilizada para aproximar la esperanza requerida.

Para el paso M el problema se convierte en el caso clásico de regresión multivariada, por lo que las estimaciones de los parámetros en la iteración t son:[71]

$$\begin{pmatrix} \hat{A}_t \\ \hat{b}_t^T \end{pmatrix} = ((Z\ 1)^T(Z\ 1))^{-1}(Z\ 1)^T X = \begin{pmatrix} Z^TZ & Z^T1 \\ 1^TZ & 1^T1 \end{pmatrix}^{-1} \begin{pmatrix} Z^TX \\ 1^TX \end{pmatrix}, \tag{4.43}$$

$$\hat{\Psi}_t = \frac{1}{n} diag\{(X - Z\hat{A} - 1\hat{b}^T)^T(X - Z\hat{A} - 1\hat{b}^T)\} = \frac{1}{n} diag\big(X^TX - 2\hat{A}^TZ^TX - 2\hat{b}1^TX + \hat{A}^TZ^TZ\hat{A} + 2\hat{b}1^TZ\hat{A} + n\hat{b}\hat{b}^T\big).[72] \tag{4.44}$$

Una vez actualizados los parámetros estimados en la iteración t repetimos todo el proceso descrito hasta obtener convergencia. En este caso, debido a la aleatoriedad extra que se obtiene como consecuencia de la simulación en el paso E, es posible observar que, aunque el algoritmo haya llegado a un punto donde se ha alcanzado una buena aproximación de un punto crítico, en las siguientes iteraciones se presenten oscilaciones cercanas a ese valor, que no son provocadas por falta de convergencia, sino por la simulación hecha en el paso E.

Un aspecto interesante del algoritmo MCEM es que, si las simulaciones en el paso E son hechas con pocas iteraciones, es posible que la aleatoriedad provocada propicie que el algoritmo no se estanque en máximos locales. Sin embargo, desearíamos que conforme el algoritmo aproxima al máximo global, el algoritmo se estabilice y la aleatoriedad del paso E tenga menos consecuencias. Una idea en este sentido es que las M iteraciones hechas en el paso E crezcan conforme el algoritmo MCEM avanza en las iteraciones t's. Esta y otras ideas, así como referencias sobre la convergencia del algoritmo pueden consultarse en Smith (2005) y Yu *et al.*(2005).

[71] Hemos omitido, por cuestión de espacio, la parte explícita que indica que los estadísticos suficientes inmersos en los cálculos son esperanzas condicionales.
[72] De nuevo, por cuestiones de espacio, omitimos en el lado derecho de la ecuación el subíndice t para las matrices A y B, que indica que las varianzas son calculadas con base en los residuos actualizados

D. IMPLEMENTACIÓN DEL MCEM AL MODELO DE AUTOSELECCIÓN

Para finalizar el capítulo, estudiaremos el proceso de estimación mediante el algoritmo MCEM aplicado al contexto de autoselección; en particular, nos referiremos al ejemplo detallado en (2.36-2.40).[73] El objetivo principal de este capítulo es dar solución a este problema. En esta parte explicaremos nuevamente el modelo, lo relacionaremos con algunos contextos econométricos, simularemos una estructura a partir del modelo y finalmente ilustraremos el algoritmo MCEM como herramienta útil para la estimación.

Comenzaremos recordando que existen variables latentes, que en general no son observadas. En este ejemplo tenemos que dichas variables latentes, de las que se genera el modelo, son:

$$y_{1i}^{*} = x_{1i}\beta_1 + \varepsilon_{1i} \tag{4.45}$$

$$y_{2i}^{*} = \gamma_2 y_{1i} + x_{2i}\beta_2 + \varepsilon_{2i} \tag{4.46}$$

$$y_{3i}^{*} = \gamma_3 y_{1i} + x_{3i}\beta_3 + \varepsilon_{3i} \tag{4.47}$$

En general, x_{ji} contiene a todas las covariables que la teoría del problema estudiado sugiere que deben incluirse. Cuando llegue el momento de simular la estructura, sólo consideraremos que en x_{ji} se incluye un intercepto y una covariable. Note además que participar en el tratamiento sólo origina un desplazamiento del intercepto (en comparación con los individuos no tratados).

Recordemos que en vez de observar realizaciones de las variables latentes, la información de la que disponemos es generada con base en las siguientes ecuaciones:

$$y_{1i} = \begin{cases} 1 \; si \;\; y_{1i}^{*} > 0 \\ 0 \; si \;\; y_{1i}^{*} \leq 0 \end{cases} \tag{4.48}$$

$$y_{2i} = \begin{cases} y_{2i}^{*} \;\; si \;\; y_{2i}^{*} > 0 \\ 0 \quad\;\; si \;\; y_{2i}^{*} \leq 0 \end{cases} \tag{4.49}$$

[73] En gran medida, este ejemplo es tomado de Smith (2005).

$$y_{3i} = \begin{cases} y_{3i}^* & si \;\; y_{3i}^* > 0 \\ 0 & si \;\; y_{3i}^* \leq 0 \end{cases} \tag{4.50}$$

Vea que establecemos que la ecuación de selección genera una respuesta binaria (participar o no en el tratamiento), además las 2 respuestas tienen una estructura censurada en 0. En el modelo especificado hasta ahora, tenemos que el vector $\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i})^T$ es un vector que contiene las características no observadas de los individuos; dicho vector es independiente de $x_i = (x_{1i}, x_{2i}, x_{3i})^T$. Se tiene que $\varepsilon_i \sim N_3(0, \Sigma)$, donde:

$$\Sigma = \begin{pmatrix} 1 & \sigma_{\varepsilon_1\varepsilon_2} & \sigma_{\varepsilon_1\varepsilon_3} \\ \sigma_{\varepsilon_1\varepsilon_2} & \sigma_{\varepsilon_2}^2 & \sigma_{\varepsilon_2\varepsilon_3} \\ \sigma_{\varepsilon_1\varepsilon_3} & \sigma_{\varepsilon_2\varepsilon_3} & \sigma_{\varepsilon_3}^2 \end{pmatrix} = \begin{pmatrix} 1 & \rho_{\varepsilon_1\varepsilon_2}\sigma_{\varepsilon_2} & \rho_{\varepsilon_1\varepsilon_3}\sigma_{\varepsilon_3} \\ \rho_{\varepsilon_1\varepsilon_2}\sigma_{\varepsilon_2} & \sigma_{\varepsilon_2}^2 & \rho_{\varepsilon_2\varepsilon_3}\sigma_{\varepsilon_2}\sigma_{\varepsilon_3} \\ \rho_{\varepsilon_1\varepsilon_3}\sigma_{\varepsilon_3} & \rho_{\varepsilon_2\varepsilon_3}\sigma_{\varepsilon_2}\sigma_{\varepsilon_3} & \sigma_{\varepsilon_3}^2 \end{pmatrix} \tag{4.51}$$

Esta estructura de covarianzas ya ha sido explicada anteriormente. En particular debemos recordar que el 1 de la primera entrada se establece por las razones de identificabilidad en el modelo probit de la ecuación de selección. Permitimos que las respuestas covaríen entre sí (esto es lo que se planteaba en el modelo SUR, explicado anteriormente) y además covaríen con la variable latente de la ecuación de selección (ya habíamos mencionado desde el segundo capítulo que características no observadas que condicionan la participación en cierto tratamiento pueden condicionar las respuestas estudiadas).

Con base en lo anterior, notamos que para cada individuo i existen dos escenarios posibles. En primera instancia, si el individuo decidió tomar el tratamiento, entonces la información que obtenemos de él se resume en:

$$y_{1i} = 1$$

$$y_{1i}^* = x_{1i}\beta_1 + \varepsilon_{1i} \tag{4.52}$$

$$y_{2i}^* = \gamma_2 + x_{2i}\beta_2 + \varepsilon_{2i} \tag{4.53}$$

$$y_{3i}^* = \gamma_3 + x_{3i}\beta_3 + \varepsilon_{3i} \tag{4.54}$$

Por otra parte, si el individuo decidió no tomar el tratamiento, entonces tenemos:

$$y_{1i} = 0$$

$$y_{1i}^* = x_{1i}\beta_1 + \varepsilon_{1i} \tag{4.55}$$

$$y_{2i}^* = x_{2i}\beta_2 + \varepsilon_{2i} \tag{4.56}$$

$$y_{3i}^* = x_{3i}\beta_3 + \varepsilon_{3i} \tag{4.57}$$

Como se ha mencionado, esta sencilla modelación abarca una cantidad importante de contextos en los que la autoselección es un problema importante. Pensemos, por ejemplo, que se desea evaluar el impacto de un programa gratuito (y de participación voluntaria) orientado a mejorar la salud física y mental de las personas en cierta población (por ejemplo pacientes de un hospital). La ecuación de selección por supuesto modela la decisión de ingresar o no al programa. Las variables y_{2i} y y_{3i} pueden ser, por ejemplo, horas dedicadas al ejercicio físico y horas dedicadas a la lectura, respectivamente.[74]

Para comprender exhaustivamente cómo se implementará el algoritmo MCEM, comenzaremos por escribir la verosimilitud de los datos completos:

$$L^c(\theta, \Sigma|x) = \prod_i f(y_{1i}^*, y_{2i}^*, y_{3i}^*) = \prod_i \left[\frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} exp\left(-\frac{\varepsilon_i \Sigma^{-1} \varepsilon_i}{2}\right)\right] \tag{4.58}$$

$$\theta = (\beta_1, \gamma_2, \beta_2, \gamma_3, \beta_3) \tag{4.59}$$

$$\varepsilon_i = \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \end{pmatrix} = \begin{pmatrix} y_{1i}^* - X_{1i}\beta_1 \\ y_{2i}^* - \gamma_2 y_{1i} - X_{2i}\beta_2 \\ y_{3i}^* - \gamma_3 y_{1i} + X_{3i}\beta_3 \end{pmatrix} \tag{4.60}$$

Una observación importante es que, en realidad, los datos observados y los no observados son y_i y y_i^*, respectivamente. Sin embargo, podemos resumir que la densidad

[74] En la estructura que simularemos tendremos distintas covariables para la ecuación de selección y las respuestas, además presumimos que el impacto de las covariables en las respuestas no depende de la decisión de tomar o no el tratamiento. La modelación podría incluir estos componentes y el algoritmo MCEM también resulta una buena herramienta en dichos escenarios. Escogimos un escenario más sencillo y simple porque permitirá ilustrar mejor los conceptos presentes en la modelación y estimación, y pese a ser un ejemplo sintético, ya hemos explicado cómo puede ajustarse a escenarios donde se desea modelar la autoselección.

de los *datos completos* es en realidad la densidad de y_i^*, ya que dicho vector determina totalmente a y_i. Tomando el logaritmo de la expresión en (4.58) tenemos que la log-verosimilitud de los datos completos es:

$$l^c(\theta, \Sigma|x) = -\frac{3N}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_i tr(\Sigma^{-1}\varepsilon_i\varepsilon_i') \tag{4.61}$$

En el paso E debemos hallar la esperanza de $l^c(\theta, \Sigma|x)$ condicional en los datos observados y asumiendo que los parámetros que determinan la distribución de los componentes aleatorios son los encontrados como óptimos en la más reciente iteración. Así, en la iteración $m+1$, durante el paso E, debemos hallar:

$$E\left[l^c(\theta, \Sigma|x)|\theta^{(m)}, \Sigma^{(m)}, y_i\right] = -\frac{3N}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}tr\left(\Sigma^{-1}\sum_i \mathrm{E}\left[\varepsilon_i\varepsilon_i'|\theta^{(m)}, \Sigma^{(m)}, y_i\right]\right) \tag{4.62}$$

De este modo, podemos identificar que el paso E se completa cuando hallamos, para cada individuo i, las cantidades:

$$Q_i\left(\theta\middle|\theta^{(m)}, \Sigma^{(m)}, y_i\right) = \mathrm{E}\left[\varepsilon_i\varepsilon_i'|\theta^{(m)}, \Sigma^{(m)}, y_i\right] =$$

$$\mathrm{E}\left[\begin{pmatrix} y_{1i}^* - X_{1i}\beta_1 \\ y_{2i}^* - \gamma_2 y_{1i} - X_{2j}\beta_2 \\ y_{3i}^* - \gamma_3 y_{1i} + X_{3j}\beta_3 \end{pmatrix} \begin{pmatrix} y_{1i}^* - X_{1i}\beta_1 \\ y_{2i}^* - \gamma_2 y_{1i} - X_{2j}\beta_2 \\ y_{3i}^* - \gamma_3 y_{1i} + X_{3j}\beta_3 \end{pmatrix}^T |\theta^{(m)}, \Sigma^{(m)}, y_i\right] =$$

$$\sigma_i^{2(m)} + \begin{pmatrix} \mu_{y_{1i}^*}^{(m)} - X_{1i}\beta_1 \\ \mu_{y_{2i}^*}^{(m)} - \gamma_2 y_{1i} - X_{2i}\beta_2 \\ \mu_{y_{3i}^*}^{(m)} - \gamma_3 y_{1i} + X_{3i}\beta_3 \end{pmatrix} \begin{pmatrix} \mu_{y_{1i}^*}^{(m)} - X_{1i}\beta_1 \\ \mu_{y_{2i}^*}^{(m)} - \gamma_2 y_{1i} - X_{2i}\beta_2 \\ \mu_{y_{3i}^*}^{(m)} - \gamma_3 y_{1i} + X_{3i}\beta_3 \end{pmatrix}^T \tag{4.63}$$

Para la expresión anterior, definimos:

$$\sigma_i^{2(m)} = Cov\left(y_{1i}^*, y_{2i}^*, y_{3i}^* | \theta^{(m)}, \Sigma^{(m)}, y_i\right) = \begin{pmatrix} \sigma_{y_{1i}^*}^{2(m)} & \sigma_{y_{1i}^* y_{2i}^*}^{(m)} & \sigma_{y_{1i}^* y_{3i}^*}^{(m)} \\ \sigma_{y_{1i}^* y_{2i}^*}^{(m)} & \sigma_{y_{2i}^*}^{2(m)} & \sigma_{y_{2i}^* y_{3i}^*}^{(m)} \\ \sigma_{y_{1i}^* y_{3i}^*}^{(m)} & \sigma_{y_{2i}^* y_{3i}^*}^{(m)} & \sigma_{y_{3i}^*}^{2(m)} \end{pmatrix} \tag{4.64}$$

$$\begin{pmatrix} \mu_{y_{1i}^*}^{(m)} \\ \mu_{y_{2i}^*}^{(m)} \\ \mu_{y_{3i}^*}^{(m)} \end{pmatrix} = \begin{pmatrix} \mathrm{E}\left[y_{1i}^* | \theta^{(m)}, \Sigma^{(m)}, y_i\right] \\ \mathrm{E}\left[y_{2i}^* | \theta^{(m)}, \Sigma^{(m)}, y_i\right] \\ \mathrm{E}\left[y_{3i}^* | \theta^{(m)}, \Sigma^{(m)}, y_i\right] \end{pmatrix} \tag{4.65}$$

Por la estructura censurada y la dependencia entre las entradas del vector y_i puede resultar muy difícil calcular los momentos representados en (4.64) y (4.65).[75] Por esta situación, proponemos el muestreador de Gibbs como una alternativa para hallar las cantidades $Q_i's$. Enseguida veremos que las condicionales totales provienen de la densidad normal univariada (censurada según la información en y_i).

Con base en (4.46) y (4.51), conviene recordar que respecto de la distribución (no condicional) de las variables latentes en la iteración $m+1$ se tiene que $y_i^* | \theta^{(m)} \sim N\left(\mu_i^{(m)}, \Sigma^{(\mathrm{m})}\right)$, donde:

$$\mu_i^{(m)} = \begin{pmatrix} X_{1i}\beta_1^{(m)} \\ \gamma_2^{(m)} y_{1i} + X_{2i}\beta_2^{(m)} \\ \gamma_3^{(m)} y_{1i} + X_{3i}\beta_3^{(m)} \end{pmatrix} \tag{4.66}$$

$$\Sigma^{(m)} = \begin{pmatrix} 1 & \sigma_{\varepsilon_1\varepsilon_2}^{(m)} & \sigma_{\varepsilon_1\varepsilon_3}^{(m)} \\ \sigma_{\varepsilon_1\varepsilon_2}^{(m)} & \sigma_{\varepsilon_2}^{(m)} & \sigma_{\varepsilon_2\varepsilon_3}^{(m)} \\ \sigma_{\varepsilon_1\varepsilon_3}^{(m)} & \sigma_{\varepsilon_2\varepsilon_3}^{(m)} & \sigma_{\varepsilon_3}^{(m)} \end{pmatrix} \tag{4.67}$$

Para implementar el muestreador de Gibbs como opción para simular de $y_i^* | y_i, \theta^{(m)}, \Sigma^{(m)}$ debemos calcular las condicionales totales. Primero consideremos las densidades condicionales totales a partir del vector $y_i^* | \theta^{(m)}, \Sigma^{(m)}$. En vista de que el vector completo sigue una distribución normal multivariada, es bien sabido que la distribución

[75] Observe que en la notación, hemos obviado el hecho de que $\mu_{y_{ji}^*}^{(m)}$ y $\sigma_{y_{ri}^* y_{si}^*}^{(m)}$ representan momentos de las variables latentes, condicionales en las variables observadas.

que la j-ésima entrada del vector y_i^* condicional en en las otras entradas de y_i^* (otras 2 entradas en este ejemplo) seguirá una distribución normal univariada, de modo que $y_{ij|i(-j)}^* \sim N(\mu_{ji|i(-j)}^{(m)}, \sigma_{j|-j}^{2(m)})$ donde:[76]

$$\mu_{ji|i(-j)}^{(m)} = \mathrm{E}\left(y_{ji}^* | y_{i|-j}^*, \theta^{(m)}, \Sigma^{(m)}\right) =$$

$$= X_{ji}\beta_j^{(m)} + cov\left(y_{ji}^* | y_{i|-j}^*, \Sigma^{(m)}\right)\left[cov\left(y_{i|-j}^* \middle| \Sigma^{(m)}\right)\right]^{-1}(y_{i|-j}^* - \gamma_{-j}^{(m)} - X_{i|-j}\beta_{-j}^{(m)}) \tag{4.68}$$

$$\sigma_{j|-j}^{2(m)} = \mathrm{var}\left(y_{ji}^* | y_{i|-j}^*, \theta^{(m)}, \Sigma^{(m)}\right) =$$

$$= var\left(y_{ji}^* \middle| \Sigma^{(m)}\right) - cov\left(y_{ji}^* \middle| y_{i|-j}^*, \Sigma^{(m)}\right)\left[cov\left(y_{i|-j}^* \middle| \Sigma^{(m)}\right)\right]^{-1} cov(y_{ji}^* | y_{i|-j}^*, \Sigma^{(m)})' \tag{4.69}$$

Para simular de $y_i^* | \theta^{(m)}, \Sigma^{(m)}$ mediante el muestreador de Gibbs, en la iteración k del algoritmo, procedemos de la siguiente forma:

1. Simulamos z_1 de la densidad normal univariada que caracteriza $y_1^* | y_2^{*(k-1)}, y_3^{*(k-1)}$ y establecemos $y_1^{*(k)} = z_1$.
2. Simulamos z_2 de la densidad normal univariada que caracteriza $y_2^* | y_1^{*(k)}, y_3^{*(k-1)}$ y establecemos $y_2^{*(k)} = z_2$.
3. Simulamos z_3 de la densidad normal univariada que caracteriza $y_3^* | y_1^{*(k)}, y_2^{*(k)}$ y establecemos $y_3^{*(k)} = z_3$.

No estamos interesados en simular de la distribución incondicional de $y_i^* | \theta^{(m)}, \Sigma^{(m)}$; de hecho si éste fuera el caso, podríamos utilizar métodos *ad hoc* para simular de la distribución normal multivariada. En cambio, deseamos simular de la distribución condicional $y_i^* | y_i, \theta^{(m)}, \Sigma^{(m)}$. Note que las variables observadas pueden

[76] Hemos utilizado notación clásica en la literatura referente al muestreador de Gibbs. Si a es un vector aleatorio de k entradas, entonces al escribir $a_{i|-(i)}$ nos referimos a la variable aleatoria en la entrada i condicional en el resto de las $(k-1)$ variables aleatorias en el vector a.

indicar: i) la forma en la que estará censurada la distribución de la variable latente ó ii) el valor realizado de la variable latente.

Los pasos 1, 2 y 3 para el muestreador de Gibbs cambian de la siguiente forma (respecto a los propuestos para simular de la distribución no condicional de $y_i^*|\theta^{(m)}$):

1. Consideraremos (a) si $y_{1i} = 0$ y (b) si $y_{1i} = 1$.
    a. Simulamos z_1 de la densidad normal univariada que caracteriza $y_1^*|y_2^{*(k-1)}, y_3^{*(k-1)}$ y, si z_1 <0, establecemos $y_1^{*(k)} = z_1$. Si z_1 >0 simulamos de nuevo, hasta obtener un valor de z_1 adecuado.
    b. Simulamos z_1 de la densidad normal univariada que caracteriza $y_1^*|y_2^{*(k-1)}, y_3^{*(k-1)}$ y, si z_1 >0, establecemos $y_1^{*(k)} = z_1$. Si z_1 <0 simulamos de nuevo, hasta obtener un valor de z_1 adecuado.
2. Consideraremos (a) si $y_{2i} = 0$ y (b) si $y_{2i} > 0$.
    a. Simulamos z_2 de la densidad normal univariada que caracteriza $y_2^*|y_1^{*(k)}, y_3^{*(k-1)}$ y, si $z_2 < 0$, establecemos $y_2^{*(k)} = z_2$. Si z_2 >0 simulamos de nuevo, hasta obtener un valor de z_2 adecuado.
    b. Establecemos $y_2^{*(k)} = y_{2i}$.
3. Consideraremos (a) si $y_{3i} = 0$ y (b) si $y_{3i} > 0$.
    a. Simulamos z_3 de la densidad normal univariada que caracteriza $y_3^*|y_1^{*(k)}, y_2^{*(k)}$ y, si $z_3 < 0$, establecemos $y_3^{*(k)} = z_3$. Si z_3 >0 simulamos de nuevo, hasta obtener un valor de z_3 adecuado.
    b. Establecemos $y_3^{*(k)} = y_{3i}$.

Observe que en los casos en los que la simulación es requerida, se propone utilizar el método de rechazo para obtener observaciones de las densidades normales truncadas correspondientes. Este método puede resultar muy ineficiente, y de hecho nosotros utilizamos en esta tesis la paquetería truncnorm, como en ejemplos anteriores. Si escogimos escribir el algoritmo utilizando el método de rechazo es porque consideramos que revela con mayor claridad las ideas subyacentes del muestreo Gibbs para este ejemplo.

Para concluir el paso E, debemos reemplazar los momentos en (4.63) por los momentos muestrales obtenidos a partir de la simulación de $y_i^*|y_i, \theta^{(m)}, \Sigma^{(m)}$. Al hacer esto, el problema de estimación del paso M se simplifica al grado de que la función a optimizar es casi idéntica a la log-verosimilitud asociada al problema de regresión normal multivariada.[77] La función objetivo a maximizar queda:

$$l = -\frac{3N}{2} ln(2\pi) - \frac{N}{2} ln|\Sigma| - \frac{1}{2} tr\left(\Sigma^{-1} \sum_i \left[\sigma_i^{2(m)} + \left(\mu_i^{*(m)} - X\theta\right)\left(\mu_i^{*(m)} - X\theta\right)^T\right]\right) \quad (4.70)$$

En la expresión anterior, $\mu_i^{*(m)}$ y $\sigma_i^{2(m)}$ son las cantidades aproximadas en el paso E. Observamos que, respecto de θ, maximizar la función l equivale a minimizar:[78]

$$A(\theta) = tr\left(\Sigma^{-1} \sum_i \left[\left(\mu_i^{*(m)} - X\theta\right)\left(\mu_i^{*(m)} - X\theta\right)^T\right]\right) = \sum_i \left[\left(\mu_i^{*(m)} - X\theta\right)^T \Sigma^{-1} \left(\mu_i^{*(m)} - X\theta\right)\right] \geq$$
$$\geq 0. \quad (4.71)$$

Así, se evidencia que los estimadores de θ pueden obtenerse (teniendo los datos completos) como estimadores de Mínimos Cuadrados Generalizados. El lector podrá recordar los estimadores FGLS, en este caso para hacer los estimadores "factibles", supondremos que $\Sigma^{-1} = (\Sigma^{-1})^{(m)}$. En el paso $m+1$, el estimador para θ queda:

$$\theta^{(m+1)} = \left[X^T \left(\Sigma^{-1^{(m)}} \otimes I\right) X\right]^{-1} X^T \left(\Sigma^{-1^{(m)}} \otimes I\right) \mu_y^{*(m)} \quad (4.72)$$

En una segunda etapa, incorporamos esta información a la función log-verosimilitud, y tenemos que el problema consiste ahora en optimizar:

$$B(\Sigma) = -\frac{N}{2} \ln|\Sigma| - \frac{N}{2} tr\left(\Sigma^{-1} \frac{H}{N}\right) \quad (4.74)$$

$$H = \sum_i \left[\sigma_i^{2(m)} + \left(\mu_i^{*(m)} - X\theta^{(m+1)}\right)\left(\mu_i^{*(m)} - X\theta^{(m+1)}\right)^T\right]. \quad (4.75)$$

[77] El lector podrá recordar que algo similar ocurría con el ejemplo de *análisis de factores*.
[78] Éste es el único término que depende de θ.

Este problema, nuevamente similar al de máxima verosimilitud en el contexto de modelos SUR. Tenemos que al minimizar respecto de Σ, se obtiene:

$$\Sigma^{(m+1)} = \frac{H}{N}. \tag{4.76}$$

Con esto hemos completado el paso M. Repetimos el algoritmo hasta convergencia. En este caso, como se explicó anteriormente, conviene evaluar la convergencia de la verosimilitud, además de observar la convergencia de los estimadores *per se*, siendo en ambos casos conscientes de la aleatoriedad que puede generarse por la simulación del paso E.

En el ejemplo que estamos desarrollando, escogimos que todos los coeficientes fueran 1 (exceptuando el intercepto en la ecuación de selección, fijado igual a .5). De este modo, las ecuaciones que caracterizan al modelo son:

$$y_{1i}^{*} = .5 + x_{1i} + \varepsilon_{1i}$$

$$y_{2i}^{*} = y_{1i} + 1 + x_{2i} + \varepsilon_{2i}$$

$$y_{3i}^{*} = y_{1i} + 1 + x_{3i} + \varepsilon_{3i}$$

Vea que en la notación inicial, donde x_{ji} podía representar vectores, tendríamos que los parámetros del modelo son $\beta_1 = (\beta_{01}, \beta_{11})^{T} = (.5,1)$, $\beta_2 = (\beta_{02}, \beta_{12})^{T} = (1,1)$, $\beta_3 = (\beta_{03}, \beta_{13})^{T} = (1,1)$ y finalmente, $\gamma_2 = \gamma_3 = 1$. Para las variables x_{ji} hemos simulado de distribuciones normales estándar independientes. Recuerde que los datos que podremos observar se originan con base en 4.48-4.50. Finalmente, hemos considerado la siguiente estructura de covarianzas:

$$\Sigma = \begin{pmatrix} 1 & .5 & 0 \\ .5 & 1 & .3 \\ 0 & .3 & 1 \end{pmatrix}.$$

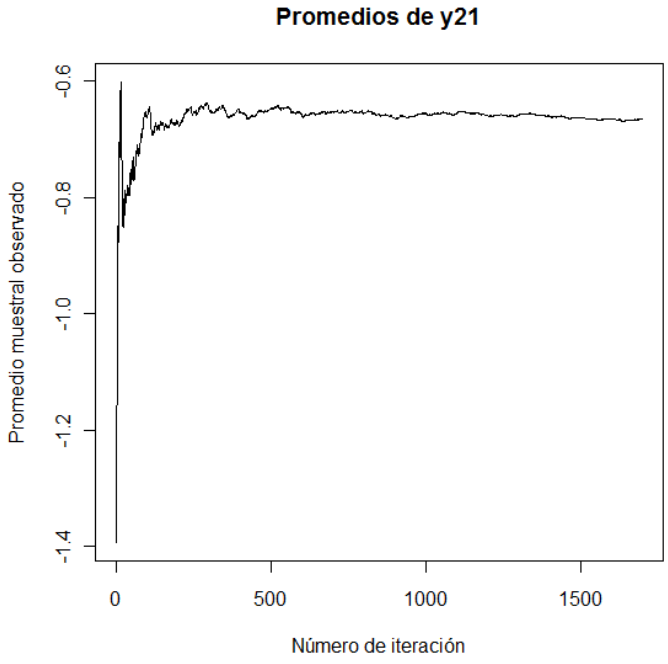
Para el algoritmo MCEM usamos, respecto de la maximización, una matriz identidad como valor inicial de Σ. Todos los 8 parámetros en θ se inician en 0. Respecto al muestreador de Gibbs, consideramos un periodo de quemado de 700 iteraciones y a partir de ahí muestreamos cada 10 vectores simulados. No observamos problemas de correlación. Para ilustrar estos 2 aspectos, estudiaremos la simulación correspondiente al primer individuo de la estructura simulada. Resulta ser que para éste, tenemos:

$$y_1 = \begin{cases} y_{11} \\ y_{21} \\ y_{31} \end{cases} = \begin{cases} 0 \\ 0 \\ 2.493 \end{cases}$$

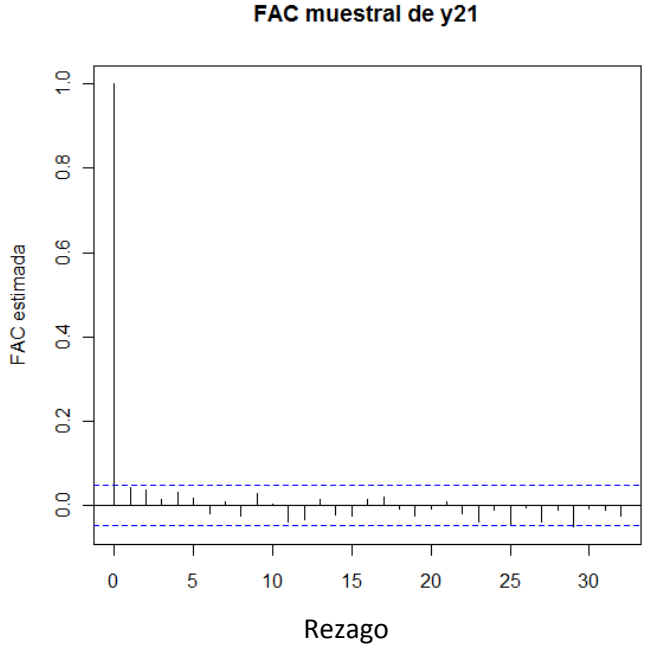
Los valores iniciales de y_1^* se fijan como los valores de y_1. Simularemos entonces la distribución de y_1^* condicional en los datos observados anteriormente, tomamos la matriz identidad como valor inicial de Σ mientras que todos los coeficientes en θ son 0.[79] Ahora, para ver la convergencia veremos cómo se estabiliza el valor de y_{21}^*, utilizando herramientas gráficas que hemos estudiado antes.

Primero mostramos el comportamiento de los promedios de la variable en cuestión. Esperamos determinar cuántas iteraciones deben transcurrir para cerciorarnos de que la cadena de Markov simulada ha convergido a su distribución asintótica.

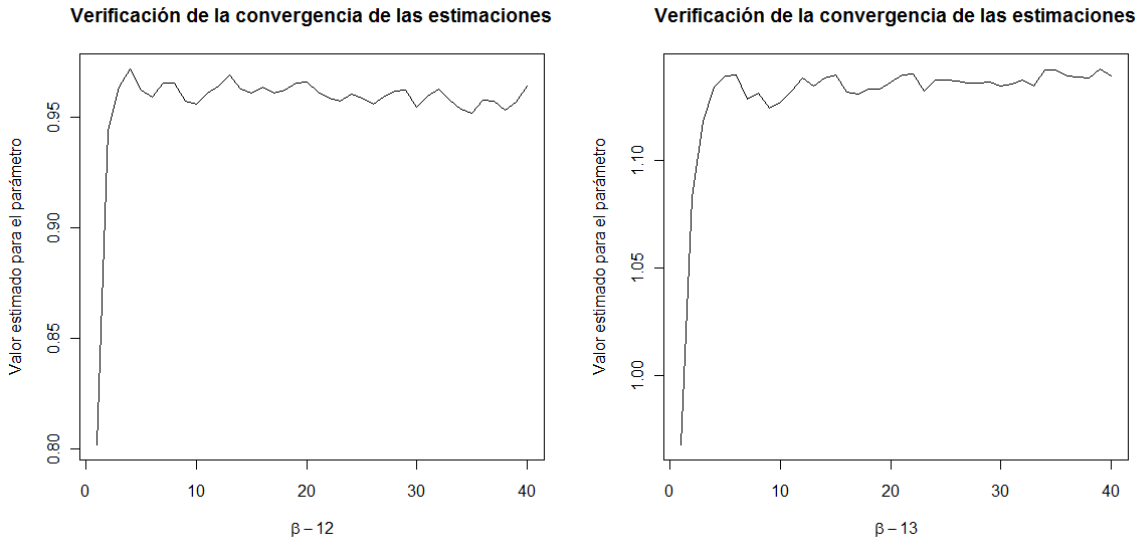
[79] En realidad estos valores no son determinantes para la velocidad de convergencia, por ello en las demás iteraciones (respecto al paso M) aunque los valores de θ sean no nulos, no hay cambios significativos en cuanto a la velocidad de convergencia. La sensibilidad de dicha velocidad es respecto a la estructura que se imponga en los datos observados; esto es lógico en tanto que si observamos ambas respuestas sólo habrá que simular de la ecuación de selección.



Enseguida, observaremos la función de autocorrelación estimada. Hemos visto antes que si, como en este caso, requerimos de una muestra iid (en esta ocasión para estimar momentos) entonces debemos cerciorarnos de que las observaciones muestreadas no covaríen de forma importante.



Para finalizar, colocamos la evolución del algoritmo a través de verificar el comportamiento de los estimadores, con el fin de estudiar la convergencia. Con el fin de ilustrar este punto, colocamos el comportamiento de $\hat{\beta}_{12}$ y $\hat{\beta}_{13}$, que son los efectos estimados de las covariables sobre las respuestas (recuerde que el valor real de los parámetros es igual a 1). Puede observarse que los estimadores llegan a una zona (que presuntamente contiene al máximo) y se estabilizan parcialmente alrededor de ella.



El algoritmo es exitoso en cuanto a que se consiguen hallar los parámetros con buen grado de aproximación, pero resulta ineficiente (las iteraciones del algoritmo MCEM son lentas), con especial demora en el paso E del algoritmo. De cualquier manera, algunas comparaciones mencionadas en la literatura[80] indican que esta opción de estimación resulta una buena alternativa ante el problema de integración numérica alto dimensional que implicaría atacar la verosimilitud directamente, y resulta mucho más ineficiente (en términos de tiempo de cómputo).

Con esto, hemos resuelto el problema de modelación y estimación para contextos donde la autoselección desea ser tratada. Cabe resaltar que el código implementado hasta el momento, puede ser mejorado en distintas direcciones y el tiempo de cómputo puede reducirse considerablemente, esto se planea hacer en el futuro. Sin embargo, la idea de este capítulo era ilustrar en buena medida el algoritmo MCEM, así como su

[80] Especialmente en Smith (2005)

implementación en el ejemplo concreto de modelos de autoselección. Si el lector está interesado en emplear el algoritmo y profundizar en temas como la convergencia del mismo o la obtención de errores estándar, las referencias mencionadas en este trabajo pueden resultarle de gran utilidad.

V. CONCLUSIONES Y CONSIDERACIONES FINALES

En este trabajo de investigación hemos expuesto como aspecto seminal el problema de autoselección. Comenzamos desarrollando la conceptualización del problema, mediante ejemplos y explicación de contextos (sobre todo económicos) donde ignorar la autoselección es algo que condiciona que las conclusiones de la inferencia estadística puedan ser erradas.

Como alternativa para la modelación de dicho problema se introdujeron las variables latentes. En particular, hemos estudiado cómo establecer estructuras donde intervienen variables latentes y variables observadas. Se exploró el alcance y utilidad de este tipo de modelación y se mostró, en particular, que es útil para resolver escenarios donde la autoselección esté presente.

Este trabajo refiere distintas fuentes donde se puede consultar cómo proceder para emplear métodos usuales[81] con el fin de estimar los modelos más tradicionales que incorporan variables latentes. En particular estudiamos un escenario[82] no estándar que incorpora utilidades latentes y el problema de autoselección. Discutimos el potencial de dicho modelo en contextos de autoselección y propusimos como alternativa para su estimación el algoritmo MCEM. La parte *estocástica* en dicho algoritmo se implementó mediante el muestreador de Gibbs. Los conceptos, aplicaciones y referencias sobre el algoritmo se detallaron en el capítulo IV de este trabajo.

En esta investigación, hemos cumplido básicamente con 2 objetivos principales:

i) Estudiar, exponer e ilustrar el problema de autoselección.

ii) Proponer y estudiar alternativas de modelación y estimación para atacar dicho problema.

En el transcurso del texto se estudian *grosso modo*, diversos temas relacionados con los puntos i) y ii) recientemente mencionados. Cuando es necesario, se orienta al lector sobre posibles fuentes de consulta para profundizar en dichos temas, sin perder de vista que los objetivos principales se concentran en el problema de autoselección. Este trabajo resulta adecuado para contextualizarse en el problema mencionado y conocer herramientas útiles para abordarlo, si bien no se limitan a ello.

[81] Pueden estudiarse, por ejemplo, la construcción teórica de los modelos o bien las alternativas (especialmente paqueterías del software R) para realizar la estimación.

[82] Ampliamente documentado en la parte final de los capítulos III y IV.

El trabajo tiene un carácter de introductorio, en el sentido de que no se pretende enfocar el problema desde todas las perspectivas, ni darle solución en profundidad; este trabajo más bien se introduce en el problema desde una visión inclinada a la econometría, y explora una solución que conlleva el uso de métodos computacionalmente intensivos.

Existen aspectos que en la práctica deben considerarse al modelar y hacer inferencia. Entre ellos el lector podrá notar que, hemos tratado siempre modelos lineales, si bien las generalizaciones a modelos no lineales siguen el mismo camino que en toda la teoría de Mínimos Cuadrados y/o Máxima Verosimilitud, es importante conocer las técnicas y la literatura que se enfoca en estos modelos. Otros aspectos son las estructuras de heterocedasticidad y/o autocorrelación que pueden incorporarse en la estimación de los modelos propuestos. Tampoco se profundizó en el aspecto de la obtención de errores estándar con la implementación del algoritmo MCEM, lo que sin duda resulta indispensable para poder hacer inferencia en contextos reales. Las referencias mencionadas a lo largo del texto dan cauce en buena medida a la solución o incorporación de las cuestiones antes descritas, sin embargo existen también problemas abiertos a la discusión académica y sobre los que todavía no existe consenso.

Como trabajo futuro y producto de la continuación de esta investigación, se podría pensar en el perfeccionamiento del algoritmo propuesto al final de la sección IV, además de una supervisión exhaustiva de la implementación en el software R. Consideramos que esfuerzos en ese sentido resultarían valiosos para dar acceso fácil para la estimación de los modelos enmarcados en esta tesis. En particular, el código implementado (y que estará en constante mejoría y extensión) en este trabajo para el contexto de selección está disponible solicitándolo a osantiago@cimat.mx.

BIBLIOGRAFÍA

Bollen, K., *Structural equations with latent variables*, Wiley, New York, 1989.

Cameron, A., P. Trivedi, *Microeconometrics: Methods and Applications*, Cambridge University Press, New York, 2005.

Fox, J., *Structural Equation Modeling With the sem Package in R*, Structural Equation Modeling, Vol. 13, 206.

Greene, W., *Econometric Analysis*, Prentice Hall, Seventh Edition, New York, 2012.

Heckman, J., "Sample selection bias as a epecification error", *Econometrica*, Vol. 47, No.1, 1979.

McFadden, D., "Econometric Analysis of Qualitative Response Models" en *Handbook of Econometrics*, Z. Griliches, M. Intriligator (Eds.), Vol. 2, North-Holland, Amsterdam, 1984.

Oberhofer, W., J. Kmenta, "A general procedure for obtaining maximum likelihood estimates in generalized regression models", *Econometrica*, Vol. 42, No. 3, 1974.

Robert, C., G. Casella, *Monte Carlo Statistical Methods*, Springer, New York, 2004.

Robert, C., G. Casella, *Introducing Monte Carlo Methods with R*, Springer, New York, 2010.

Smith-Ramírez, R. (2005), *On the Evaluation of Conservation Cost-Sharing Programs. An Application of a Monte Carlo EM Algorithm*, Disertación Doctoral, Universidad de Maryland.

Toomet, O., A. Henningsen, "Sample selection models in R: Package sampleSelection", *Journal of Statistical Software*, Vol. 27, No. 7, 2008.

Wu, C., "On the convergence properties of the EM algorithm", *Annals Statistics*, Vol. 11, 983.

Yu, P., K. Lam, S. Lo, "Factor analysis for ranked data with application to a job selection attitude survey", *Journal of the Royal Statistical Society (Series A)*, Vol. 168, No. 3, 2005.