CONSIDERACIONES SOBRE EL ANALISIS DE LA INFORMACION

La información existente sobre desempleo puede ser analizada utilizando diferentes técnicas estadísticas, dependiendo de los objetivos de la investigación.

El enfoque tradicional -no por ello el menos útil-, consiste en analizar la información a través de los cuadros de clasificación de dos o más entradas. Estos cuadros permiten caracterizar a los desempleados usando diferentes variables. En general, la caracterización se hace a través del arreglo (distribución) de frecuencias resultante.

Un análisis más realista consiste en reconocer que el arreglo de frecuencias resultante proviene de una muestra aleatoria y por tanto, un método de análisis recomendable consiste en usar tablas de contingencia para estudiar la relación entre las categorías de empleado y desempleado, con variables tales como educación, sexo y edad.

El análisis a través de tablas de contingencia, supone que todas las variables son categóricas (discretas) o que aquellas que son continuas, pueden ser categorizadas sin perder información. Si el supuesto anterior se satisface, entonces, todo el análisis del problema del desempleo puede efectuarse a través de tablas de contingencia.

Tradicionalmente, el análisis de tablas de contingencia consiste en estudiar, por ejemplo, la independencia o dependencia de las categorías de empleado o desempleado con variables como las mencionadas arriba. Actualmente existen modelos estadísticos que proporcionan más información

sobre la relación entre las variables que se estudian y al mismo tiempo facilitan el análisis de tablas de contingencia de tres o más dimensiones. Ver por ejemplo Fienberg (1977). Una desventaja de estos modelos es que requieren de datos agrupados para su aplicación. Esta desventaja es particularmente importante en el caso de las Ciencias Sociales, donde es muy frecuente encontrar información sobre variables que no se desean categorizar.

Otro problema que a menudo enfrentan los investigadores no sólo en las Ciencias Sociales sino también en otras áreas, consiste en relacionar una variable categórica con una o más variables que aquí llamaremos "explicativas", las cuales pueden ser de naturaleza continua o discreta. Este problema también puede verse como un problema de clasificación aunque no satisface las condiciones para ser analizado a través de tablas de contingencia.

En el caso de un estudio sobre desempleo parece razonable pensar en un modelo estadístico que relacione las categorías de empleado o desempleado con un conjunto de variables explicativas, las cuales pueden ser de naturaleza continua o discreta. Este será el enfoque que se usará para estudiar el desempleo.

En general, el problema de relacionar una variable categórica con un conjunto de variables explicativas, aparece frecuentemente en áreas de ciencias como Medicina, Economía, Agricultura y Sociología entre otras. Algunos ejemplos de estas aplicaciones pueden verse en Mantel (1973) y Press y Wilson (1978) para el caso de aplicaciones en Medicina, Dhrymes (1978) y

McFadden (1973) en Economía, Nerlove y Press (1973) en Agricultura y García-Hernández (1980) en Sociología. Estas aplicaciones del modelo logístico pone de manifiesto que aunque en este estudio se analizará el problema del desempleo, el modelo propuesto es relevante para el estudio de una gama muy amplia de problemas.

EL MODELO

Suponiendo que un individuo pueda clasificarse sin ambiguedad como empleado o desempleado, es posible relacionar su condición de empleado o - desempleado con un conjunto de variables explicativas. Si denotamos a "Y" como la variable aleatoria dicotómica observada, la cual toma los valores Y=1 si el individuo está empleado y Y=0 si está desempleado, y a las "p" variables explicativas como X_1, X_2, ... X_p entonces la relación matemática entre "Y" y X_1, X_2, ... X_p puede escribirse como: $Y = F(x_1, x_2, \ldots x_p)$ donde la función F tiene que ser especificada.

Para especificar F, es posible suponer una relación monotónica entre X_i (i = 1, ..., p) y el valor de "Y". Por ejemplo, se puede pensar, que a medida que X_i aumenta "Y" se aproximará a uno, y en el caso contrario "Y" se aproximará a cero. Dado que la función "F" depende de varias variables, podemos usar una suma ponderada de las variable explicativas de la forma $X'\beta = \sum_{i=1}^{p} \beta_i X_i$, donde las β's representan las ponderaciones. Al mismo tiempo, los valores de "Y" pueden interpretarse como probabilidades - de tal manera que ahora suponemos una relación monotónica entre Y y $X'\beta$. Es

to último, implica que la función "F" tendrá que especificarse de tal manera que a medida que $X'\beta$ aumente (disminuya) "Y" se aproximará a uno (cero). La relación inversa también puede presentarse, pero esto no requiere una interpretación adicional.

Varios modelos han sido propuestos para estimar la relación entre Y y $X'\beta$. El modelo de mínimos cuadrados ordinarios ha sido descartado entre otras razones porque el modelo es heteroscedástico. Además, el supuesto de normalidad de las Y_i's no es válido y las pruebas de significancia de los coeficientes estimados no se aplican. Para eliminar el problema de heteroscedasticidad, Goldberger (1964) ha sugerido el uso del método de mínimos cuadrados generalizados para estimar la relación entre "Y" y $X'\beta$. La desventaja es que este método no restringe los valores estimados de "Y" a que estén entre cero y uno, lo que puede generar valores negativos para algunas varianzas. Una exposición clara y bien fundamentada teóricamente, sobre las desventajas de usar mínimos cuadrados ordinarios o mínimos cuadrados generalizados cuando se relaciona una variable categórica con un subconjunto de variables explicativas, puede verse en Nerlove y Press (1973). De los métodos de estimación, de la relación entre "Y" y $X'\beta$, que evitan las dificultades teóricas mencionadas, los más usados son el análisis "probit" y la regresión logística.

En este estudio se hará uso únicamente de la regresión logística - debido a que este modelo tiene ventajas prácticas sobre el análisis "probit". En primer lugar, la función acumulativa de probabilidad logística aproxima bastante bien a la normal y en segundo lugar es mucho más fácil

de computarla debido a que no involucra la evaluación de integrales como ocurre con el análisis "probit". Además, el modelo de regresión logística puede ser justificado formalmente sin hacer supuestos muy fuertes sobre la distribución conjunta de las variables aleatorias a considerar.

FORMULACION MATEMATICA DEL MODELO

Sea "E" el evento empleado y "$\bar{E}$" el evento desempleado. La variable aleatoria tomará el valor Y=1 si E ocurre y Y=0 si $\bar{E}$ ocurre. Si X es un vector de variables aleatorias continuas con densidad $h(x \mid \theta)$, donde θ es una matriz de parámetros que indica la distribución, entonces por el teorema de Bayes:

$$P(E \mid X) = \frac{P(E)\, h(x \mid E, \theta)}{P(E)\, h(x \mid E, \theta) + P(\bar{E})\, h(x \mid \bar{E}, \theta)} \qquad (1)$$

donde $P(\cdot)$ denota la función discreta de probabilidad y $P(E) + P(\bar{E}) = 1$.

Es fácil ver que (2) puede expresarse como:

$$P(E \mid X) = \frac{1}{1 + \left[\frac{1-P(E)}{P(E)}\right]\left[\frac{h(x \mid \bar{E}, \theta)}{h(x \mid E, \theta)}\right]} \qquad (2)$$

Si suponemos que, dado (E, θ), X se distribuye normalmente con media θ_1 y matriz de covarianzas Σ, y que dado $(\bar{E}, \theta)$, X se distribuye también normal con θ_2 y la misma matriz de covarianzas Σ, entonces:

$$\frac{h(x \mid \bar{E},\theta)}{h(x \mid E,\theta)} = e^{-x' \Sigma^{-1} (\theta_1-\theta_2) + \frac{1}{2}\left[(\theta_1+\theta_2)' \Sigma^{-1} (\theta_1-\theta_2)\right]}$$

También:

$$\frac{1-P(E)}{P(E)} = e^{\ln\left[\frac{1-P(E)}{P(E)}\right]} \qquad (3)$$

Finalmente (3) quedará expresada en la forma

$$P(E \mid X) = \left[\frac{1}{1+e^{-\alpha - x'\beta}}\right] \qquad (4)$$

donde $\beta = \Sigma^{-1} (\theta_1-\theta_2)$ y $\alpha = -\ln\left[\frac{1-P(E)}{P(E)}\right] - \frac{1}{2} (\theta_1+\theta_2)'\beta$

Del resultado obtenido en (4) es claro que $P(E \mid x)$ tiene la forma de la función de probabilidad acumulada de la logística.

La derivación de (4) se obtuvo bajo el supuesto de que condicionado a $(\bar{E},\theta)$ $\left[(\bar{E},\theta)\right]$, X se distribuye normalmente con media θ_1 (θ_2) y matriz de covarianzas Σ. Este supuesto es muy difícil de satisfacer en muchos casos prácticos, pero J.A. Anderson (1972) ha demostrado que para que $P(E \mid x)$ tenga la forma de la función de probabilidad acumulada de la logística es suficiente que $h(x \mid E,\theta)$ y $h(x \mid \bar{E},\theta)$ pertenezcan a la familia exponencial. Esto permite establecer que el resultado presentado en (4) es mucho más robusto de lo que se supuso al derivarlo.