# Un modelo para estudiar el Desempleo en el Area Metropolitana de Monterrey

Francisco García



CENTRO DE INVESTIGACIONES ECONOMICAS

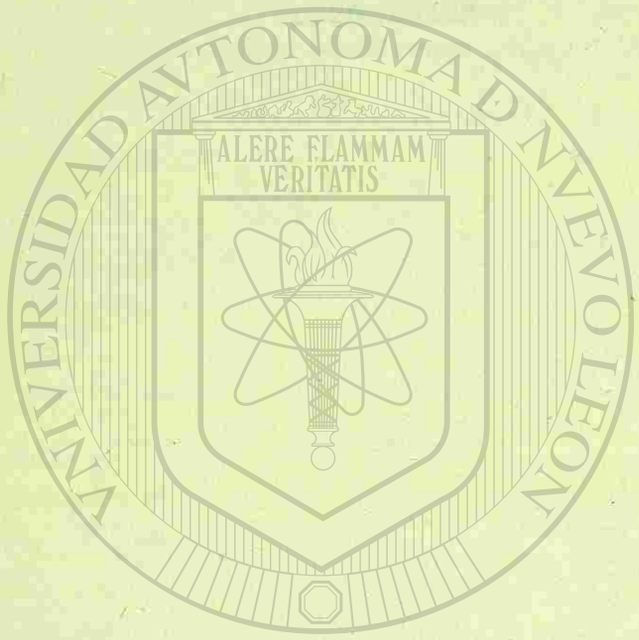
1980

# Un modelo para estudiar el Desempleo en el Area Metropolitana de Monterrey

Francisco Cerda



FACULTAD DE ECONOMIA
CENTRO DE INVESTIGACIONES ECONOMICAS




# Un modelo para estudiar el Desempleo en el Area Metropolitana de Monterrey


Francisco García



FACULTAD DE ECONOMIA
CENTRO DE INVESTIGACIONES ECONOMICAS



Primera edición, 1980.



# UN MODELO PARA ESTUDIAR EL DESEMPLEO EN EL AREA METROPOLITANA DE MONTERREY

## INTRODUCCION

Es de aceptación general que uno de los objetivos de la política económica consiste en disminuir la tasa de desempleo. El estudio cuantitativo del problema del desempleo es importante porque éste, debe preceder al diseño de los instrumentos de política económica a través de los cuales se pretenda influir en la tasa de desempleo.

En el Area Metropolitana de Monterrey existe un antecedente importante en relación a estudios cuantitativos sobre desempleo.

El Centro de Investigaciones Económicas de la Universidad Autónoma de Nuevo León (CIE), ha llevado a cabo varias encuestas por muestreo con el fin de estudiar -entre otros- el problema del desempleo en el Area Metropolitana de Monterrey. La información obtenida ha sido publicada por el mismo Centro, en diversos números de la publicación "Ocupación y Salarios".

Actualmente la Secretaría de Programación y Presupuesto lleva a cabo una encuesta trimestral, encuesta continua de mano de obra, a través de la cual se obtiene información sobre desempleo. Esta información también ha sido publicada por el CIE en los números más recientes de la publicación Ocupación y Salarios.

Lo anterior muestra que existe un volumen considerable de información, tanto de corte transversal como a través del tiempo, sobre desempleo en el Area Metropolitana de Monterrey.

## 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 $X_1, X_2, \ldots X_p$ entonces la relación matemática entre "Y" y $X_1, X_2, \ldots X_p$ 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 $X_i$ $(i = 1, \ldots, p)$ 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 $\beta$'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 $\theta$ 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 $(E, \theta)$, X se distribuye normalmente con media $\theta_1$ y matriz de covarianzas $\Sigma$, y que dado $(\bar{E}, \theta)$, X se distribuye también normal con $\theta_2$ y la misma matriz de covarianzas $\Sigma$, 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 $\theta_1$ $(\theta_2)$ y matriz de covarianzas $\Sigma$. 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.

## ESTIMACION DE LOS PARAMETROS DEL MODELO

Para estimar $P(E \mid x)$ es necesario estimar el factor de ponderaciones $\beta$, y la constante $\alpha$ del modelo. Dos métodos de estimación de $\alpha$ y $\beta$ se consideran a continuación: a) el método basado en la función lineal discriminante y b) el método de máxima verosimilitud.

### 1.- Estimación de los Parámetros $\alpha$ y $\beta$ a través de la función lineal discriminante

Si se supone que tanto $h(x \mid E,\theta)$ son funciones de densidad normales multivariadas con la misma matriz de covarianzas $\Sigma$ pero con diferente vector de medias, podemos identificar al escalar $x'\beta$, donde $\beta = \Sigma^{-1}.(\theta_1-\theta_2)$, como la función lineal discriminante. Es claro que un estimador de $\beta$ puede obtenerse usando los estimadores consistentes de $\Sigma$, $\theta_2$ y $\theta_1$. Además dado que

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

podemos obtener un estimador de $\alpha$ usando el estimador de $\beta$, los estimadores consistentes de $\theta_1$ y $\theta_2$ y como un estimador de $P(E)$ a la razón del número de casos en la muestra con la característica que se estudia (empleado) entre el tamaño de la muestra.

Los estimadores consistentes de $\Sigma$, $\theta_1$ y $\theta_2$ pueden obtenerse de la siguiente manera. Se forman dos grupos de observaciones. En el primer

grupo se incluyen las observaciones de los individuos que poseen la característica bajo estudio (empleado) y en el segundo los restantes. Suponiendo que cada grupo consiste de $n_1$ y $n_2$ observaciones respectivamente y que cada observación consiste de p variables podemos representar las observaciones por grupo y las medias por variables de la siguiente manera:

$$\begin{array}{cc}
X'_{11}, X'_{12}, \ldots, X'_{1p} & X^{\circ}_{11}, X^{\circ}_{12}, \ldots, X^{\circ}_{1p} \\
X'_{21}, X'_{22}, \ldots, X'_{2p} & X^{\circ}_{21}, X^{\circ}_{22}, \ldots, X^{\circ}_{2p} \\
\vdots \quad \vdots \quad \vdots & \vdots \quad \vdots \quad \vdots \\
X'_{n_1 1}, X'_{n_1 2}, \ldots, X'_{n_1 p} & X^{\circ}_{n_2 1}, X^{\circ}_{n_2 2}, \ldots, X^{\circ}_{n_2 p} \\
\hline
\bar{X}'_1 \quad \bar{X}'_2 \quad \ldots \quad \bar{X}'_p & \bar{X}^{\circ}_1 \quad \bar{X}^{\circ}_2 \quad \ldots \quad \bar{X}^{\circ}_p
\end{array}$$

Dentro de cada grupo se pueden obtener las observaciones en forma de desviaciones con respecto a la media para cada variable y formar matrices de la forma:

$$x' = \begin{bmatrix} x'_{11} & x'_{12} & \cdots & x'_{1p} \\ x'_{21} & x'_{22} & \cdots & x'_{2p} \\ \vdots & \vdots & & \vdots \\ x'_{n_1 1} & x'_{n_1 2} & \cdots & x'_{n_1 p} \end{bmatrix}$$

$$y^{\circ} = \begin{bmatrix} x^{\circ}_{11} & x^{\circ}_{12} & \cdots & x^{\circ}_{1p} \\ x^{\circ}_{21} & x^{\circ}_{22} & \cdots & x^{\circ}_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x^{\circ}_{n_2 1} & x^{\circ}_{n_2 2} & \cdots & x^{\circ}_{n_2 p} \end{bmatrix}$$

donde:

$$x'_{11} = (X'_{11} - \bar{X}_1) \cdots x'_{n_1 p} = (X'_{n_1 p} - \bar{X}'_p)$$

y

$$x^{\circ}_{11} = (X^{\circ}_{11} - \bar{X}^{\circ}_1) \cdots x^{\circ}_{n_2 p} = (X^{\circ}_{n_2 p} - \bar{X}^{\circ}_p)$$

Usando las matrices de las observaciones en forma de desviaciones podemos obtener las matrices de productos cruzados para cada grupo, esto es:

$$S_1^2 = \frac{1}{n_1 - 1} (x^1)' (x^1)$$

$$S_2^2 = \frac{1}{n_2 - 1} (x^{\circ})'(x^{\circ})$$

Finalmente el estimador consistente de $\Sigma$ (matriz de covarianzas) está dado por:

$$\hat{\Sigma} = \frac{1}{n_1 + n_2 - 2} \left[ (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 \right]$$

Los estimadores consistentes de $\theta_1$ y $\theta_2$ están dados por:

$$\theta'_1 = (\bar{X}'_1, \bar{X}'_2, \ldots, \bar{X}'_p)$$

$$\theta'_2 = (\bar{X}^{\circ}_1, \bar{X}^{\circ}_2, \ldots, \bar{X}^{\circ}_p)$$

Definiendo a los vectores:

$$D' = (\bar{X}'_1 - \bar{X}^{\circ}_1, \bar{X}'_2 - \bar{X}^{\circ}_2, \ldots, \bar{X}'_p - \bar{X}^{\circ}_p) \quad y$$

$$v' = (\bar{X}'_1 + \bar{X}^{\circ}_1, \bar{X}'_2 + \bar{X}^{\circ}_2, \ldots, \bar{X}'_p + \bar{X}^{\circ}_p)$$

Los estimadores de $\beta$ y $\alpha$ se expresan:

$$\hat{\beta} = \Sigma^{-1} D \quad y \quad \hat{\alpha} = -\frac{1}{2}\hat{\beta}' v - \ln\left(\frac{n_2}{n_1}\right)$$

Para hacer pruebas de significancia acerca de la diferencia entre el valor estimado de los parámetros con otro valor preestablecido, se puede usar la matriz de covarianzas asintóticas de los estimadores la cual está dada por:

$$V(\hat{\beta}) = \left[\frac{1}{n_o} + \frac{1}{n_1}\right](\Sigma)^{-1}$$

2.- Estimación de los parámetros $\alpha$ y $\beta$ a través del método de Máxima Verosimilitud.

La estimación de los parámetros de la función de distribución de

la logística a través del método de máxima verosimilitud, ha sido formalmente establecida en varios artículos. (Ver ejemplo a Nerlove y Press, 1973, y las referencias contenidas ahí).

De manera general, el método de estimación de esta sección puede establecerse como sigue:

Sea $y = (y_1, y_2, \ldots, y_n)$ y $x'_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$

donde $x_{i1} = 1$ para toda i.

La función de verosimilitud para la i-ésima observación está dada por:

$$L(y_i \mid x_i) = \left[P(y_i = 1 \mid x_i)\right]^{y_i} \left[P(y_i = 0 \mid x_i)\right]^{1-y_i}$$

La función de verosimilitud para una muestra aleatoria de n observaciones está dada por:

$$L(y \mid x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} \left[P(y_i = 1 \mid x_i)\right]^{y_i} \left[P(y_i = 0 \mid x_i)\right]^{1-y_i}$$

En el supuesto caso que P siga la función de distribución de la logística, puede demostrarse a través de manipulaciones algebraicas sencillas que la función de verosimilitud de las "n" observaciones está dada por:

$$L(y \mid x_1, x_2, \ldots, x_n) = \frac{\exp\left[\beta' \sum_{i=1}^{n} x_i y_i\right]}{\prod_{i=1}^{n} \left[1 + \exp(x'_i \beta)\right]}$$

donde: $\beta' = (\alpha, \beta_1, \ldots, \beta_{p-1})$

Se puede notar que para dados $x_1, x_2, \ldots, x_n$, $t = \sum_{i=1}^{n} x_i y_i$ es suficiente para $\beta$. Esto es una ventaja teórica importante, ya que el estimador de $\beta$ es una función de t y por tanto tendrá menor error cuadrático medio que cualquier otro estimador.

El estimador de máxima verosimilitud de $\beta$ ($\hat{\beta}$), se obtiene derivando el logaritmo natural de la función de verosimilitud con respecto a $\beta$ e igualando el vector de derivadas a cero. Es fácil probar que la función de verosimilitud es cóncava y que por tanto la función tiene un máximo global cuando $\beta = \hat{\beta}$.

El estimador de $\beta$ ($\hat{\beta}$) debe satisfacer las ecuaciones:

$$\sum_{i=1}^{n} \left[ \frac{1}{1+\exp(-x'_i \hat{\beta})} \right] x_i = \sum_{i=1}^{n} x_i y_i$$

Dado que las ecuaciones dadas arriba no son lineales en $\beta$, éstas pueden resolverse por cualquier método iterativo (por ejemplo a través del método de Newton-Raphson).

Actualmente existen programas para computadora muy eficientes que facilitan mucho la solución de las ecuaciones descritas arriba.

Una característica general del método de estimación de máxima vero

similitud es que el estimador de $\beta$ ($\hat{\beta}$) se distribuye asintóticamente - normal con media $\beta$ y matriz de covarianzas $\Sigma$.

$$\text{Si denotamos a } \lambda_i = \frac{\exp(-\hat{\beta}'x_i)}{1 + \exp(-\hat{\beta}'x_i)}, \quad (i = 1, 2, \ldots, n)$$

a

$$x' = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad (\text{donde } n = n_1 + n_0)$$

y

$$A = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

entónces $\Sigma$ estará dada por

$$\Sigma = XAX'$$

Esto permite hacer pruebas estadísticas de significancia para determinar si las $\beta$'s difieren significativamente de un valor preestablecido (generalmente cero).

3.- Consideraciones prácticas sobre los métodos de estimación de los parámetros $\alpha$ y $\beta$ descritos anteriormente.

En muchos casos prácticos el supuesto que $h(x \mid \theta)$ es una distribución normal multivariada es muy fuerte. Además, al subdividir a la muestra total en dos grupos es muy poco probable que ambos grupos tengan la misma matriz de covarianzas. Estos dos supuestos limitan mucho el uso de la función lineal discriminante para estimar $P(E \mid x)$. La ventaja más importante del uso de la función lineal discriminante consiste en que es muy facil computarla. Esta ventaja se vuelve más importante cuando las facilidades de computación son limitadas. Las desventajas teóricas del uso de la función lineal discriminante para estimar $P(E \mid x)$ se pueden resumir en que cuando el supuesto de normalidad no se cumple, los estimadores de $\alpha$ y $\beta$ son sesgados. Además, estos estimadores no son función de una estadística suficiente, y por tanto, no tienen el mínimo error cuadrático medio.

El método de máxima verosimilitud tiene todas las ventajas teóricas que ya se mencionaron. Además, el supuesto de normalidad y desigualdad de la matriz de covarianzas, no es relevante en su aplicación.

El principal problema con este método es que es difícil de computarse y cuando las facilidades de computación son limitadas es casi imposible obtener los estimadores de máxima verosimilitud.

Existen estudios empíricos que comparan, a través de métodos ad-hoc, la "bondad del ajuste" que se obtiene al estimar los parámetros de la fun-

ción de distribución de la logística cuando se usa como método de estimación la función lineal discriminante o el de máxima verosimilitud.

Ejemplos de estos estudios son los de Halperin, Blackwelder y Verter (1971) y Press y Wilson (1979). En el primer estudio los autores encontraron que en general los estimadores obtenidos a través de la función lineal discriminante son sesgados y que el sesgo no disminuye al aumentar el tamaño de la muestra, esto es, los estimadores son inconsistentes. En el caso de la intercepción del modelo lineal transformado ($\alpha$), su estimador no sólo es inconsistente sino que tiende a afectar mucho las estimaciones de $P(E \mid X)$ cuando el número de casos en la muestra con el atributo E dista mucho del 50%. En el segundo estudio Press y Wilson encontraron que el modelo de regresión logística clasifica mejor si los parámetros del modelo se estiman usando el método de máxima verosimilitud en vez del método de la función lineal discriminante.

A pesar de la evidencia presentada, en este estudio se ha usado la función lineal discriminante para estimar $P(E \mid x)$ con la idea de que una vez que esté disponible un programa para obtener los estimadores de máxima verosimilitud, ambos resultados puedan ser comparados.

APLICACION DEL MODELO A LA INFORMACION SOBRE DESEMPLEO EN EL AREA METROPOLITANA DE MONTERREY.

La información que se utilizó para estimar la probabilidad de que un individuo esté empleado o desempleado en el Area Metropolitana de Monte

ción de distribución de la logística cuando se usa como método de estimación la función lineal discriminante o el de máxima verosimilitud.

Ejemplos de estos estudios son los de Halperin, Blackwelder y Verter (1971) y Press y Wilson (1979). En el primer estudio los autores encontraron que en general los estimadores obtenidos a través de la función lineal discriminante son sesgados y que el sesgo no disminuye al aumentar el tamaño de la muestra, esto es, los estimadores son inconsistentes. En el caso de la intercepción del modelo lineal transformado ($\alpha$), su estimador no sólo es inconsistente sino que tiende a afectar mucho las estimaciones de $P(E \mid X)$ cuando el número de casos en la muestra con el atributo E dista mucho del 50%. En el segundo estudio Press y Wilson encontraron que el modelo de regresión logística clasifica mejor si los parámetros del modelo se estiman usando el método de máxima verosimilitud en vez del método de la función lineal discriminante.

A pesar de la evidencia presentada, en este estudio se ha usado la función lineal discriminante para estimar $P(E \mid x)$ con la idea de que una vez que esté disponible un programa para obtener los estimadores de máxima verosimilitud, ambos resultados puedan ser comparados.

APLICACION DEL MODELO A LA INFORMACION SOBRE DESEMPLEO EN EL AREA METROPOLITANA DE MONTERREY.

La información que se utilizó para estimar la probabilidad de que un individuo esté empleado o desempleado en el Area Metropolitana de Monte

rrey es la del trimestre Enero-Marzo de 1978. Esta se obtuvo a través de la "Encuesta Continua de Mano de Obra", que lleva a cabo la Secretaría de Programación y Presupuesto.

Se consideró conveniente hacer cambios en las definiciones que usa la Secretaría de Programación y Presupuesto, antes de usar el modelo de regresión logística. La principal razón para hacer cambios, es que no es posible establecer sin ambiguedad la condición de empleado o desempleado para cada individuo. Por ejemplo, individuos que trabajan pocas horas, pueden estar buscando trabajo, o aún más complicado, es el caso de los que trabajan en negocios familiares con o sin remuneración y que pudieran estar en esa ocupación sólo mientras encuentran trabajo. La forma en que cada individuo busca trabajo, depende de su visión del mercado de trabajo y es plausible que existan relaciones informales a través de las cuales se busca trabajo. Esto implica que cuando se hace la pregunta acerca de si el individuo busca trabajo, él considera que no está buscando trabajo.

Los cambios en las definiciones están orientados principalmente a construir una clasificación de los individuos en la población económicamente activa en empleados y desempleados.

Los cambios efectuados se resumen a continuación:

1) Se eliminaron de la población económicamente activa a todos los individuos menores de 15 años y mayores de 70.

2) De los individuos clasificados como ocupados, se considera-

ron como empleados sólo los que trabajan actualmente.

3) Se consideraron como desempleados los inactivos temporales y los desocupados.

4) De los individuos clasificados como ocupados:

a) Se eliminaron los que trabajaban en un negocio familiar sin remuneración.

b) Los que trabajaron pocas horas, menos de 20, se consideraron como desocupados.

c) Los que recibieron un ingreso semanal muy bajo, menos de $100.00, también se consideraron como desocupados.

Es conveniente señalar que los cambios se hicieron, considerando las críticas de Gunnar Myrdal (1972) a los estudios sobre desempleo en países subdesarrollados, aunque se debe reconocer que tanto las horas de trabajo como el nivel de ingreso semanal que se usó para reclasificar a los individuos, es un tanto arbitraria.

RESULTADOS EMPIRICOS

La muestra consistió de 1,236 observaciones de las cuales, una vez reclasificadas, 979 correspondieron a personas ocupadas y 257 a desocupadas.

Las variables disponibles para explicar la condición de ocupado o desocupado fueron muy limitadas y sólo fue posible usar sexo, edad y educación.

Los resultados obtenidos al ajustar un modelo de regresión logística a los datos, se presentan en el cuadro de resumen. Se ajustaron siete modelos con el fin de analizar la contribución de cada una de las variables a la explicación de la condición de empleado o desempleado. La contribución de cada una de las variables explicativas se mide usando el número de observaciones correctamente clasificadas cuando se consideran modelos con o sin la variable bajo estudio.

Usando el criterio, "número de observaciones correctamente clasificadas" puede verse en el cuadro de resumen que los modelos II, III y V son equivalentes y que el modelo I es marginalmente inferior. Además puede observarse que el modelo V que contiene a sexo como única variable explicativa clasifica tan bien como los modelos II y III y mejor que el resto de los modelos. Esto quiere decir que bajo el criterio de preferencia de modelos propuesta tanto la variable edad como la escolaridad son redundantes.

También es importante notar, que bajo el criterio usado, el modelo VII es el menos preferido y dado que este modelo tiene a escolaridad como única variable explicativa la inferencia es clara en el sentido que escolaridad es la que menos constribuye a explicar la condición de empleado o desempleado.

## CONCLUSIONES

Es común pensar en una función de bienestar social que contiene entre sus argumentos a la tasa de desempleo. Pero la tasa de desempleo depende obviamente del número de personas desempleadas, lo cual es el resultado

Los resultados obtenidos al ajustar un modelo de regresión logística a los datos, se presentan en el cuadro de resumen. Se ajustaron siete modelos con el fin de analizar la contribución de cada una de las variables a la explicación de la condición de empleado o desempleado. La contribución de cada una de las variables explicativas se mide usando el número de observaciones correctamente clasificadas cuando se consideran modelos con o sin la variable bajo estudio.

Usando el criterio, "número de observaciones correctamente clasificadas" puede verse en el cuadro de resumen que los modelos II, III y V son equivalentes y que el modelo I es marginalmente inferior. Además puede observarse que el modelo V que contiene a sexo como única variable explicativa clasifica tan bien como los modelos II y III y mejor que el resto de los modelos. Esto quiere decir que bajo el criterio de preferencia de modelos propuesta tanto la variable edad como la escolaridad son redundantes.

También es importante notar, que bajo el criterio usado, el modelo VII es el menos preferido y dado que este modelo tiene a escolaridad como única variable explicativa la inferencia es clara en el sentido que escolaridad es la que menos constribuye a explicar la condición de empleado o desempleado.

## CONCLUSIONES

Es común pensar en una función de bienestar social que contiene entre sus argumentos a la tasa de desempleo. Pero la tasa de desempleo depende obviamente del número de personas desempleadas, lo cual es el resultado

de un proceso estocástico que depende a su vez de muchos factores aleatorios.

Al investigar sobre el modelo que explica mejor la condición de empleado o desempleado también se investiga sobre el modelo que estima mejor a la tasa de desempleo, aunque la importancia de esta última es sólo secundaria. Lo que es relevante es entender lo que está detrás de la tasa de desempleo cuando se consideran las características de las personas que componen la sociedad.

En una sociedad en la que se reconoce que las oportunidades de educación están desigualmente distribuidas y en donde la mano de obra calificada escasea, parece ser que la educación debería ser importante para explicar la condición de empleado o desempleado. En cambio en este estudio se encontró que la escolaridad tiene muy bajo poder explicativo de la condición bajo estudio. Una posible explicación de este resultado es que las oportunidades de empleo, para cada nivel de educación, son aproximadamente las mismas.

El resultado, quizá esperado por algunos lectores, aunque no por el autor, de que la variable más importante para explicar la condición de empleado o desempleado sea sexo, implica obviamente una práctica discriminatoria en el mercado de trabajo. De manera sencilla, el resultado del estudio establece que una persona por el mero hecho de ser mujer tiene una probabilidad bastante alta de estar desempleada (independientemente de su edad y educación).

CUADRO DE RESUMEN

VALORES ESTIMADOS DE LOS PARAMETROS DE SIETE MODELOS DE REGRESION LOGISTICA Y SUS RESPECTIVAS "T" ASINTOTICAS

| Modelos | Estimación Puntual | | | | t asintótica | | | No. de Casos clasificados | | $(c)=\frac{(a)}{(a)+(b)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\alpha}$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $t\hat{\beta}_1$ | $t\hat{\beta}_2$ | $t\hat{\beta}_3$ | (a) Correctamente | (b) Incorrectamente | |
| $X_1$ VS $X_2$, $X_3$, $X_4$ | -1.29 | .829 | .013 | .042 | 5.06** | 2.23* | 2.35* | 828 | 408 | .67 |
| $X_1$ VS $X_2$, $X_3$ | - .823 | .803 | .008 | | 4.91** | 1.53 | | 858 | 378 | .69 |
| $X_1$ VS $S_2$, $X_4$ | - .807 | .859 | | .029 | 5.26** | | 1.70 | 858 | 378 | .69 |
| $X_1$ VS $X_3$, $X_4$ | - .742 | | .015 | .036 | | 2.66** | 2.02* | 631 | 605 | .51 |
| $X_1$ VS $X_2$ | - .582 | .831 | | | 5.12** | | | 858 | 378 | .69 |
| $X_1$ VS $X_3$ | - .357 | | .011 | | | 2.09* | | 598 | 638 | .48 |
| $X_1$ VS $X_4$ | - .141 | | | .02 | | | 1.17 | 579 | 657 | .47 |

$X_1$ = Variable dependiente dicotómica: 1 = Ocupado, 0 = Desocupado.

Variables independientes: $X_2$ = Sexo, $X_3$ = Edad, $X_4$ = Escolaridad.

** Significativa al 1%.

* Significativa al 5%.

Volviendo al concepto de la función de bienestar social, es obvio que no sólo a la tasa de desempleo debe ser un argumento de la misma, sino que también se debe considerar que las probabilidades de desempleo sean aproximadamente iguales para los diferentes grupos sociales que constituyen la sociedad.

BIBLIOGRAFIA

Anderson, J.A. (1972), "Separate sample logistic discrimination". Biometrics, Vol. 59, 19-35.

Drymes, P.J. (1978), Introductory Econometrics. Springer Verlag, New York.

Fienberg, S.E. (1977), The Analysis of Cross-Classified Categorical Data. MIT Press, Cambridge Mass.

García-Hernández, F. (1980), "A Bayesian Procedure to Compare and Choose Among Logistic Regression Models". Tesis Doctoral no publicada, University of California, Riverside.

Goldberger, A.S. (1964), Econometric Theory. John Wiley, New York.

Halperin, M., C. Blackwelder y Joel I. Verter (1971), "Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Meximum Likelihood Approaches". J. Chron. Dis. 24, 125-158.

Mantel, N. (1973), "Synthetic Retrospective Studies and Related Studies". Biometrics 29, 479-486.

McFadden, D. (1973), "Conditional Logit Análisis of Qualitative Choice Behavior". En Frontiers in Econometrics. Editado por P. Zarembka. Academic Press, New York.

Myrdal, Gunnar (1972), "Evaluación Critica de Algunos Estudios sobre Desempleo y Subempleo". En Lecturas Sobre Desarrollo Agrícola, editado por Edmundo Flores. Fondo de Cultura Económica, México.

Nerlove, M. y S.J. Press (1973), Univariate and Multivariate Log-Linear and Logistic Models. R-1306-EDAINIH, Rand Corp., Santa Mónica.

Press, S.J. y Sandra Wilson (1979), "Choosing Between Logistic Regression and Discriminante Analysis". J. Amer. Statist. Assoc. 73-699-705.

APENDICE

Programa en Fortran para estimar los parámetros de un modelo de regresión logística, usando la función lineal discriminante.

```
C PROGRAMA PARA CALCULAR LA PROBABILIDAD DE ESTAR DESEMPLEADO
C USANDO LA FUNCION LINEAL DISCRIMINANTE
C PROGRAMA EN FORTRAN PARA LA CONTROL DATA
C AUTORES:AMALIA SYLVIA ARRIAGA Y FRANCISCO GARCIA H.
      PROGRAM PROB(INPUT,OUTPUT,TAPE5=INPUT,TAPE6=OUTPUT,TAPE8,TAPE2)
      DOUBLE PRECISION S,D,F
      DIMENSION X(4),X1M(3),XBM(3),BETA(3),SUM1(3),SUMO(3)
      DIMENSION DIFME(3),SUMME(3)
      DIMENSION PNAMES(4),XNAME(4),IPT(4),XDATA(4)
      DIMENSION DX1(257,3),DX0(979,3),S2X1(3,3),S2X0(3,3),TEMP1(3,257)
      DIMENSION TEMP0(3,979),SIGMA(3,3),S(3,3),VARB(3,3)
      NO=0
      N1=0
      READ(5,35)NVAR,NNAM
   35 FORMAT(2I5)
      READ(5,36)(PNAMES(I),I=1,NNAM)
   36 FORMAT(A8,A8,A8,A8)
      READ(5,37)K
   37 FORMAT(I5)
      K1=K+1
      READ(5,36)(XNAME(I),I=1,K1)
C READ POINTERS TO VARIABLES DESIRED INTO IPT AND
C NAMES OF VARIABLES INTO PNAMES
      DO 40 I=1,K1
      IPT(I)=0
      DO 38 J=1,NNAM
      IF(PNAMES(J).EQ.XNAME(I)) GO TO 39
   38 CONTINUE
      WRITE(6,9001)XNAME(I)
 9001 FORMAT(1H1,2X,12HLA VARIABLE ,A8,1X,28HNO SE ENCONTRO EN EL ARCHIV
     1O/)
      STOP
   39 IPT(I)=J
   40 CONTINUE
      WRITE(6,9002)
 9002 FORMAT(1H1,20X,45HCALCULOS CORRESPONDIENTES AL SIGUIENTE MODELO/)
      WRITE(6,9003)(I,XNAME(I),IPT(I),I=1,K1)
 9003 FORMAT(2X,65HVARIABLES A SER USADAS Y SU LOCALIZACION EN EL ARCHIV
     1O DE ENTRADA/(1X,I5,2X,A8,I6,5X,I5,2X,A8,I6,5X,I5,2X,A8,I6))
      DO 200 I=1,K
      SUMO(I)=0.0
      SUM1(I)=0.0
  200 CONTINUE
C SELECT VARIABLES TO BE INCLUDED IN THE MODEL
      DO 43 I=1,1236
      DO 330 J=1,K1
  330 X(J)=XDATA(IPT(J))
      READ(8)(XDATA(J),J=1,NVAR)
C WRITE THE SELECTED VARIABLES ON SCATCH DISK (UNIT 2).
      WRITE(2,42)(X(J),J=1,K1)
   42 FORMAT(4F4.0)
   43 CONTINUE
      REWIND 2
      REWIND 8
C DESIRED VARIABLES ARE READY TO BE READ OFF FILE 2
    1 READ(2,42)(X(J),J=1,K1)
C CHECK FOR END OF FILE
```

```
      IF(EOF(2))5,2
C CHECK IF PERSON BELONGS TO THE GROUP OF EMPLOYED PEOPLE
C CONTROL GOES TO STATEMENT 4 IF EMPLOYED,AND TO STA 3 IF UNEMPLOYED
    2 IF(X(K+1))4,3
    3 N1=N1+1
      DO 300 J=1,K
      SUM1(J)=SUM1(J)+X(J)
  300 CONTINUE
      GO TO 1
    4 N0=N0+1
      DO 400 I=1,K
      SUM0(I)=SUM0(I)+X(I)
  400 CONTINUE
      GO TO 1
    5 CONTINUE
      RN1=FLOAT(N1)
      RN0=FLOAT(N0)
C  CALCULATE THE MEANS OF THE XSS FOR THE TWO GROUPS
      DO 500 I=1,K
      X1M(I)=SUM1(I)/RN1
      X0M(I)=SUM0(I)/RN0
  500 CONTINUE
C PRINT THE MEANS OF THE XSS FOR THE TWO GROUPS
      DO 550 I=1,K
      WRITE(6,540)I,X1M(I),I,X0M(I)
  540 FORMAT(11X,16HVALOR DE LA X1M(,I1,1H),2X,F7.2,3X,16HVALOR DE LA X0
     1M(,I1,1H),2X,F7.2/)
  550 CONTINUE
C  REWIND THE TAPE TO BE READ AGAIN
      REWIND 2
      I0=0
      I1=0
C READ DATA AGAIN TO CALCULATE THE MATRICES OF DEVIATIONS
   10 READ(2,42)(X(I),I=1,K1)
C CHECK FOR END OF FILE
      IF(EOF(2))20,15
C CHECK IF PERSON BELGONS TO FIRST GROUP (UNEMPLOYED)
   15 IF(X(K+1))17,16
   16 I1=I1+1
      DO 600 J1=1,K
      DX1(I1,J1)=X(J1)-X1M(J1)
  600 CONTINUE
      GO TO 10
   17 I0=I0+1
      DO 625 J0=1,K
      DX0(I0,J0)=X(J0)-X0M(J0)
  625 CONTINUE
      GO TO 10
C TRANSPOSE THE MATRICES OF DEVIATIONS AND PLACE THEM IN MATRICES TEMP1
C AND TEMPO
   20 DO 650 I1=1,N1
      DO 650 J1=1,K
      TEMP1(J1,I1)=DX1(I1,J1)
  650 CONTINUE
      DO 675 I0=1,N0
      DO 675 J0=1,K
      TEMP0(J0,I0)=DX0(I0,J0)
```

```
  675 CONTINUE
C  NOW,CALCULATE THE MATRICES OF CROSS PRODUCTS S2X1 AND S2X0
      DO 700 J=1,K
      DO 700 I=1,K
      S2X1(J,I)=0.0
      DO 700 L=1,N1
      S2X1(J,I)=S2X1(J,I)+TEMP1(J,L)*DX1(L,I)
  700 CONTINUE
      DO 725 J=1,K
      DO 725 I=1,K
      S2X0(J,I)=0.0
      DO 725 M=1,N0
      S2X0(J,I)=S2X0(J,I)+TEMP0(J,M)*DX0(M,I)
  725 CONTINUE
C CALCULATE THE MATRIX SIGMA WHICH IS THE SUM OF THE MATRICES OF CROSS PRODUCTS
      DO 750 I=1,K
      DO 750 J=1,K
      SIGMA(I,J)=(S2X1(I,J)+S2X0(I,J))/(RN1+RN0-2.0)
  750 CONTINUE
C CALCULATE THE VECTORS DIFME AND SUMME
      DO 775 J=1,K
      DIFME(J)=X0M(J)-X1M(J)
      SUME(J)=X1M(J)+X0M(J)
  775 CONTINUE
C INVERT MATRIX SIGMA AND PLACE THE INVERSE IN SIGMA
      DO 9 I=1,K
      DO 9 J=1,K
    9 S(I,J)=SIGMA(I,J)
      DO 8 M=1,K
      D=S(M,M)
      DO 7 I=1,K
      IF(I.EQ.M)GO TO 7
      F=S(I,M)/D
      DO 11 J=1,K
      IF(J.EQ.M) GO TO 11
      S(I,J)=S(I,J)-F*S(M,J)
   11 CONTINUE
    7 CONTINUE
      DO 12 I=1,K
      S(I,M)=-S(I,M)/D
   12 S(M,I)=S(M,I)/D
    8 S(M,M)=1.0/D
      DO 6 I=1,K
      DO 6 J=1,K
    6 SIGMA(I,J)=S(I,J)
C  CALCULATE THE VECTOR OF BETAS
      DO 800 I=1,K
      BETA(I)=0.0
      DO 800 J=1,K
      BETA(I)=BETA(I)+SIGMA(I,J)*DIFME(J)
  800 CONTINUE
C  NOW,CALCULATE ALPHA
      TEMP=0.0
      DO 825 I=1,K
      TEMP=TEMP+BETA(I)*SUMME(I)
  825 CONTINUE
      FNJ=FLOAT(NJ)
```

```
      FNL=FLOAT(N1)
      ALPHA=(-0.5*TEMP)
C  CALCULATE THE MATRIX OF VARIANCE-COVARIANCE
      DO 850 I=1,K
      DO 850 J=1,K
      VARB(I,J)=SIGMA(I,J)*(1./FND+1./FNL)
  850 CONTINUE
C  REWIND TAPE TO BE READ AGAIN
      REWIND 2
      WRITE(6,860)
  860 FORMAT(1H1,10X,46HCOMPARACION DE LA VARIABLE X(K1) Y P CALCULADA/1
     1 11X,5HXK1),2X,11HP CALCULADA/)
C  INITIALIZE COUNTERS NCOR AND NINCOR
      NCOR=0
      NINCOR=0
   25 READ(2,42)(X(I),I=1,K1)
      IF(EOF(2))30,26
C  CALCULATE P
   26 TEMP=0.0
      DO 875 I=1,K
      TEMP=TEMP+BETA(I)*X(I)
  875 CONTINUE
      P=1./(1.+EXP(-ALPHA-TEMP))
      WRITE(6,880)X(K1),P
  880 FORMAT(10X,F6.3,5X,F8.5/)
C  COMPARE THE VALUE OF P WITH THE VALUE OF THE VAR X(K+1)
      IF(P.GE.0.5.AND.P.LE.0.50) GO TO 900
      IF(X(K+1).EQ.1.0) GO TO 885
      NINCOR=NINCOR+1
      GO TO 25
  885 NCOR=NCOR+1
      GO TO 25
  900 IF(X(K+1).EQ.0.0) GO TO 910
      NINCOR=NINCOR+1
      GO TO 25
  910 NCOR=NCOR+1
      GO TO 25
   30 CONTINUE
C  PRINT RESULTS
      WRITE(6,1000)
 1000 FORMAT(1H1,9X,101HESTIMACION DE LOS PARAMETROS DE UN MODELO DE REG
     1RESION LOGISTICA A TRAVES DE LA FUNCION DISCRIMINANTE//20X,10HPARA
     2METRO ,10X,14HVALOR ESTIMADO/)
      WRITE(6,1001)ALPHA
 1001 FORMAT(12X,7HALPHA  ,10X,F14.8/)
      DO 1003 I=1,K
      WRITE(6,1002)I,BETA(I)
 1002 FORMAT(11X,5HBETA(,I2,1H),10X,F14.8/)
 1003 CONTINUE
      WRITE(6,1004)
 1004 FORMAT(1H1,45X,45HMATRIZ DE VARIANZAS-COVARIANZAS ASINTOTICAS /)
      DO 1006 I=1,K
      WRITE(6,1005)(VARB(I,J),J=1,K)
 1005 FORMAT(40X,3(F16.9,10X)/)
 1006 CONTINUE
      WRITE(6,1007)NCOR,NINCOR
 1007 FORMAT(20X,42HNUMERO DE CASOS CLASIFICADOS CORRECTAMENTE,3X,I4/20X
```

```
1,44HNUMERO DE CASOS CLASIFICADOS INCORRECTAMENTE,3X,I4/)
 STOP
 END
```

Se terminó de imprimir en diciembre de 1980, en el Departamento de Impresos de la Facultad de Economía, de la Universidad Autónoma de Nuevo León. Loma Redonda No. 1515 Pte., Col. Loma Larga, Monterrey, N.L., México. Se tiraron 500 ejemplares más sobrantes para reposición.



FACULTAD DE ECONOMIA
CENTRO DE INVESTIGACIONES ECONOMICAS