

Identificación y determinación probabilística de los grupos pobres en Santafé de Bogotá¹

Oliva Sierra G., Elkin Castaño V., Sergio Yañez Canal, Luz Angela Valencia C.

Lecturas de Economía No. 41

Introducción, 95. I. Problema de la definición e identificación de la pobreza: "línea de pobreza utilizada en este trabajo", 96. II. Modelo de clasificación, 99. III. Definición de variables y análisis preliminares, 108. IV. Resultados del modelo de discriminación logística para la ciudad de Santafé de Bogotá, 116. Conclusiones, 128.

Introducción

La asignación eficiente del gasto social y el seguimiento de su aplicación en los diferentes programas destinados a combatir la pobreza, tales como salud, educación, vivienda, etc., requieren perfeccionar los mecanismos de focalización de los grupos más desfavorecidos.

Este trabajo, al igual que el modelo elaborado en la primera fase de la investigación para la ciudad de Medellín, *Identificación y determinación probabilística de los grupos pobres: modelo aplicado para Medellín, 1992*, pretende contribuir al problema de la identificación de los grupos pobres. Su objeto es construir un modelo estadístico de discriminación logística, el cual, mediante un conjunto de variables o características fácilmente verificables, permite clasificar los hogares pobres. Igualmente

-
1. El Departamento Nacional de Planeación -DNP- autorizó la publicación de este artículo, el cual hace parte de una investigación realizada para dicha Institución. De igual manera los conceptos, estimaciones y conclusiones de dicho estudio, son responsabilidad exclusiva del grupo investigador y, por lo tanto, no compromete al DNP.

te, a partir de dicho conjunto de características referidas al hogar o a sus miembros se puede estimar la probabilidad de que un hogar sea pobre y construir un índice de ordenamiento de la pobreza entre los hogares.

La fuente de información utilizada proviene de la Encuesta de Calidad de Vida y Pobreza realizada para Santafé de Bogotá durante los meses de octubre y noviembre de 1991, por el Departamento Administrativo Nacional de Estadística -DANE- y el Programa de las Naciones Unidas para el Desarrollo -PNUD-. La base de datos inicial la conforma un total de 2.728 hogares que, después de su depuración, se reduce a 2.718.² La información está basada en el formulario ampliado de la citada encuesta. Los componentes del presente trabajo son los siguientes: 1) problema de la definición e identificación de la pobreza: "línea de pobreza" utilizada; 2) formulación del modelo de clasificación, definición de variables y análisis preliminar; 3) resultados del modelo. En esta última se explica la selección del modelo y se interpretan los resultados; además se estiman los diferentes costos de clasificación y se construye la región de indecisión. Finalmente, se exponen las conclusiones.

I. Problema de la definición e identificación de la pobreza: "línea de pobreza utilizada en este trabajo"

Para los investigadores la medición exacta de la incidencia de la pobreza e identificación de los grupos pobres son temas, que aún hoy, suscitan controversia. La literatura internacional y nacional da cuenta de una serie de medidas posibles de la pobreza,³ y aunque se ha logrado un acuerdo casi unánime en identificar como pobres el agregado de hogares y de población que no satisfacen las necesidades básicas, entre los estudiosos del tema existen divergencias respecto a la definición

-
2. Ajustes y correcciones que el DANE hizo posteriormente, no han sido tenidas en cuenta.
 3. El conocimiento de la pobreza en América Latina. Revista Comercio Exterior, Vol. 42, No 4, México, Abril de 1992 (véase Amartya K. Sen, sobre conceptos y Medidas de pobreza). Oscar Fresneda, y otros, pobreza, violencia y desigualdad: retos para una nueva Colombia. PNUD, Colombia. 1991

específica de las mismas. Esto es: ¿cuáles son en realidad básicas y los montos mínimos, bien sea en términos del gasto o del ingreso, requeridos por los hogares o las personas para su satisfacción?

La identificación de los pobres en este trabajo no se hizo con base en una línea de pobreza tal como se le define convencionalmente. Para la aplicación del modelo se optó por una medición relativa de la pobreza. El no haber acogido un método similar al de la Medición Integral de Pobreza -basado en una línea de pobreza y en indicadores de Necesidades Básicas Insatisfechas (NBI)- se explica por dos razones, que de alguna manera se interrelacionan.⁴ De un lado, por posible subregistro en los ingresos de los hogares y, de otro, por el valor aparentemente elevado de la línea de pobreza "oficial" establecida por el DANE. Ambas razones son explicadas a continuación.

En primer lugar, la línea de pobreza del DANE parece tener un valor relativamente alto cuando se le compara con los indicadores convencionales de la capacidad adquisitiva de los ingresos urbanos -o sea en términos del salario mínimo y el valor per-cápita de la canasta básica de hogares de ingresos bajos-. En efecto, el valor per-cápita mensual de la línea según el DANE -\$45.051 por mes en octubre de 1991, período en que se realizó la Encuesta de Calidad de Vida y Pobreza- es equivalente a 0.87 del valor del salario mínimo mensual en la misma fecha -\$51.716 en 1991.

!Por tanto, un hogar con supuestamente 4.5 personas requería 3.9 salarios mínimos por mes -\$202.730- para no ser pobre! Y cuando dicha línea per-cápita se le compara con el valor de la canasta básica por persona de ingresos bajos -establecida por el DANE para construir el

4 Aunque algunos de los indicadores de necesidades básicas insatisfechas y de calidad de vida se utilizaron como variables clasificatorias dentro del modelo, la identificación misma de los hogares pobres con base en NBI se descartó debido a que un alto porcentaje de hogares (18.5%) no informaron sobre aspectos referidos a la calidad de la vivienda y acceso a los servicios públicos, y dichas respuestas, *sin información*, están altamente concentradas en los estratos 2 y 3.

Índice de Precios al Consumidor -IPC- y cuyo valor por persona en octubre de 1991 asciende a \$28.676, si se supone un promedio de 4.5 personas por hogar- equivale a 1.57 veces el valor de esta última. Para la actualización de la línea de pobreza, el DANE se basa en la canasta básica de alimentos establecida para el 25% de los hogares más pobres según los datos de la Encuesta de Ingresos y Gastos realizada entre 1984-1985; dicho valor actualizado -el de la canasta de alimentos- se multiplica por un coeficiente fijo de 2.62, correspondiente al inverso de la proporción del valor de los alimentos en el gasto total de los hogares de ingresos bajos.⁵ Dadas las altas tasas de inflación en Colombia en los últimos años y los consiguientes cambios en los precios relativos de los bienes, se cree necesario revisar el valor de la canasta normativa de los alimentos y su proporción en el gasto total de los hogares, tarea por fuera del alcance y objetivos de este trabajo.

Al confrontar la línea de pobreza con el ingreso per-cápita de los hogares de la base de datos disponible para Santafé de Bogotá -Encuesta de Calidad de Vida y Pobreza- el porcentaje de hogares pobres asciende al 45%, cifra que sobrepasa en más del 10% los índices que arrojan otros estudios para la ciudad (véase estudio de Miguel Urrutia, (1993) y Libardo Sarmiento, (1993). Ahora bien, ello puede deberse no sólo al elevado valor de la línea sino que, en alguna medida, puede ser atribuible a un probable subregistro de los datos de ingresos de los hogares. Esta es la otra razón que explica el por qué no se utilizó una "línea de pobreza" en sentido estricto. Aunque los ingresos de los hogares se sometieron a una rigurosa revisión, el margen de subregistro "típico", inherente a la medición de los mismos, no pudo ser precisado, ni por lo tanto corregido. Sin embargo, es necesario señalar la depuración que se hizo de los ingresos. Infortunadamente no pudieron ajustarse mediante la estimación de los arriendos imputados, cuando la tenencia de la vivienda era diferente al arriendo o sub-arriendo.⁶ El valor del ingreso total del hogar

5. Muñoz C. Manuel y otros. *La pobreza en trece ciudades colombianas* en Boletín de Estadística del DANE, No. 429, Diciembre de 1988.

6. De 1524 hogares que debían informar el arriendo imputado, 500 hogares no lo hicieron.

se tomó con base en las respuestas dadas a una pregunta que controla el ingreso promedio estimado de los perceptores dentro de la Encuesta de Calidad de Vida y Pobreza, debido a que la sumatoria de los ingresos de éstos, según su fuente de origen, en términos generales, resultó inferior a la suma de los valores de los ingresos promedios mencionados.

En síntesis, debido a las consideraciones anteriores para el análisis de la clasificación de los hogares pobres se procedió ante todo con un criterio estadístico, el cual a su vez se apoyó en el bajo ingreso per-cápita de los hogares como concepto definitorio de la pobreza. Con dicho criterio se determinó una banda de pobreza que establece los valores máximos del ingreso per-cápita que mejor clasifica el grupo de hogares "pobres" en términos del comportamiento de las características o variables seleccionadas por el modelo. Las "líneas de pobreza" aproximadas -per-cápita mensuales- así establecidas, se mueven en un rango que parece razonable entre dos valores extremos: \$35.000 y \$45.000. Con el límite más bajo -\$35.000- el índice de hogares pobres se sitúa aproximadamente en 35%, y con \$45.000 -línea del DANE- asciende al 45%.

II. Modelo de clasificación

A. Introducción

El problema de clasificar un hogar en una de las categorías "pobre" o "no pobre" puede ser estudiado en el contexto estadístico del análisis de discriminación el cual, dado un objeto -un hogar en nuestro caso-, desarrolla reglas de asignación óptimas en algún sentido, para clasificar dicho objeto en uno de dos -o en uno de varios- grupos conocidos.

La literatura del análisis de discriminación es extensa y algunas referencias básicas se encuentran en Cacoullos (1973), Lachenbruch (1975), Huberty (1975), Goldstein y Dillon (1978), Lachenbruch y Goldstein (1979), Hosmer y Lemeshow (1989).

Las siguientes secciones discuten brevemente el análisis de discriminación multinormal, sus problemas para la aplicación al caso de la

clasificación de un hogar como pobre o no pobre, y el procedimiento de discriminación logística como una metodología alternativa más apropiada.

B. Elementos del análisis de discriminación en dos grupos

Supongamos que tenemos una población P con una proporción p_1 de individuos en el grupo G_1 y una proporción $p_2 = 1-p_1$ en el grupo G_2 .

Sea $f_i(x)$ la función de probabilidad o la función de densidad de probabilidad del vector X de p características, si X procede del grupo G_i . Para nuestro trabajo identificaremos G_1 con el grupo de hogares pobres y a G_2 con el grupo de hogares no pobres.

Consideremos además las siguientes cantidades:

i) La probabilidad de clasificación errónea $P(j/i)$, que es la probabilidad de clasificar el objeto con características X en el grupo G_j cuando en realidad procede del grupo G_i , $i, j=1, 2$, con $i \neq j$. ii) El costo por clasificación errónea $C(j/i)$, costo asociado de clasificar un objeto con características X en el grupo G_j cuando en realidad procede del grupo G_i , $i, j=1, 2$, con $i \neq j$.

iii) El costo esperado total por clasificación errónea

$$C_T = C(2/1)P(2/1)p_1 + C(1/2)P(1/2)p_2$$

La regla de asignación que minimiza el costo esperado total por clasificación errónea es:

Asigne el objeto con características X al grupo G_1 si

$$\frac{f_1(x)}{f_2(x)} > \frac{p_2 C(1/2)}{p_1 C(2/1)}$$

La regla anterior es también conocida con el nombre de procedimiento bayesiano óptimo.

Cuando los costos por mala clasificación son iguales, la regla se convierte en

$$\frac{f_1(x)}{f_2(x)} > \frac{p_2}{p_1}$$

Para la implementación en la práctica de estas reglas es necesario conocer las funciones de probabilidad o de densidad de probabilidad del vector de características en cada grupo, f_1 y f_2 , los costos de clasificación errónea $C(2/1)$, $C(1/2)$, y las probabilidades *a priori* p_1 y p_2 . Los costos de clasificación errónea son determinados de acuerdo a la importancia que tenga en la investigación clasificar mal un hogar pobre o clasificar mal un hogar no pobre. Por ejemplo, si el objetivo es reconocer con más precisión cuándo las características en el vector X corresponden a un hogar pobre, entonces $C(2/1)$ debería ser mayor que $C(1/2)$.

C. Reglas de discriminación usando la distribución multinormal

En general, el análisis de discriminación se basa en el supuesto de que el vector de características X para los objetos del grupo G_i tiene una distribución normal multivariante con vector de medias μ_i y matriz de covarianzas Σ_i .

Bajo este supuesto se generan las reglas de discriminación siguientes:

i) Si las matrices de covarianzas de las dos funciones de densidad de probabilidad multinormal son iguales, es decir $\Sigma_1 = \Sigma_2 = \Sigma$, la regla de decisión es:

asigne el objeto con características X al grupo G_1 si

$$(\mu_1 - \mu_2)' \Sigma^{-1} [x - \frac{1}{2}(\mu_1 + \mu_2)] > \log \left(\frac{p_2 C(1/2)}{p_1 C(2/1)} \right)$$

Esta es conocida como la regla de la Función Lineal Discriminante -FLD-.

ii) Si las matrices de covarianzas son diferentes, la regla de decisión es de la forma:

sea

$$Q(x) = \frac{1}{2} \log \left(\frac{\det(\Sigma_1)}{\det(\Sigma_2)} \right) - \frac{1}{2} (x - \mu_1)' \Sigma_1^{-1} (x - \mu_1) + \frac{1}{2} (x - \mu_2)' \Sigma_2^{-1} (x - \mu_2)$$

Entonces asigne el objeto con características X al grupo G_1 si:

$$Q(x) > \log \left(\frac{p_2 C^{(1/2)}}{p_1 C^{(2/1)}} \right)$$

Esta es conocida como la regla de la Función Cuadrática Discriminante -FCD-.

En la práctica los parámetros μ_i y Σ_i son desconocidos y dadas las muestras para cada grupo se obtienen los estimadores máximos verosímiles de ellos y se reemplazan en el lugar de los verdaderos parámetros.

Estas reglas son óptimas cuando se cumplen los supuestos:

- i) Las distribuciones son multinormales (para la FLD y la FCD).
- ii) Las matrices de covarianzas son iguales (para la FLD).
- iii) Las observaciones empleadas para la estimación de los parámetros están clasificadas correctamente.

Ahora bien: ¿cómo se comportan ellas cuando uno o varios supuestos no se cumplen?

Algunos autores, entre ellos Krzanowsky (1977), Lachenbruch (1975), Titterington et al (1981), Moore (1973), Lachenbruch et al (1973), Crawley

(1979), Chinganda y Subrahmaniam (1979), Clarke et al (1979), Aitchison et al (1977), Ahmed y Lachenbruch (1975, 1977) y Broffitt et al (1980), han estudiado el efecto sobre el comportamiento de las reglas de discriminación FLD y FCD cuando se viola uno o varios de dichos supuestos.

En el caso estudiado aquí, debido a que casi todas las variables discriminadoras son discretas los procedimientos descritos pueden no ser adecuados.

Crawley (1979), Byth y McLachlan (1980), y O'Hara et al (1982) han concluido que el procedimiento de discriminación logística es preferible cuando las distribuciones son claramente no-normales o las matrices de covarianza son diferentes.

D. El modelo de discriminación logística

El modelo de clasificación logística supone que

$$\log \left[\frac{f_1(x)}{f_2(x)} \right] = \alpha + \beta'X$$

Dada la regla general, el procedimiento de clasificación es de la forma:

Asigne el objeto con características X al grupo G_1 si

$$\alpha + \beta'X > \log \left(\frac{p_2 C(1/2)}{p_1 C(2/1)} \right)$$

o si

$$\alpha_0 + \beta'X > \log \left(\frac{C(1/2)}{C(2/1)} \right)$$

donde

$$\alpha_0 = \alpha + \log \left(\frac{p_1}{p_2} \right)$$

Alguna literatura sobre la discriminación logística se encuentra en Lachenbruch (1975), Anderson (1982), Anderson y Blair (1982) y Seber (1984).

Una ventaja de la aproximación logística es que si X contiene p características de la población, solamente necesitamos estimar $p+1$ parámetros en contraste con los $3p+p(p-1)/2$ parámetros del caso FLD y los $4p+p(p-1)$ para el caso FCD.

Otra ventaja importante es que las familias de distribuciones que satisfacen el modelo logístico son bastante amplias, entre las que se encuentran distribuciones con mezclas de variables.

1. Diseño del muestreo para la discriminación logística

Existen tres diseños de muestreo que proporcionan datos adecuados para la estimación de :

- i) Muestreo mixto: en este caso una muestra de n_1+n_2 elementos de la población total son aleatoriamente seleccionados de manera que los n_i son aleatorios.
- ii) Muestreo separado: en este caso una muestra de tamaño fijo n_i es tomada del grupo G_i ; $i = 1,2$.
- iii) Muestreo condicional: en este caso una muestra de $n(x_j)$ elementos de la población P son seleccionados al azar entre todos los miembros de P para los cuales $x = x_j$. En este caso $n_i(x_j)$ es aleatorio.

Las diferentes clases de muestreo tienen incidencia en la estimación máximo verosímil de α_0 . Para los muestreos condicionales y mixto su estimación es proporcionada directamente. Para el muestreo separado α_0 no puede ser estimado directamente a menos que p_1 y p_2 sean conocidos o tengan estimadores independientes. Véase por ejemplo, Seber (1982).

En el caso de estudio, el muestreo adecuado corresponde al muestreo mixto, puesto que de un tamaño muestral total de n hogares se obtuvie-

ron n_1 (aleatorios) hogares para el grupo denominado "pobre" y n_2 (también aleatorios) con $n = n_1 + n_2$.

De acuerdo con esto el procedimiento de estimación máximo verosímil arrojará directamente la estimación de α_0 y por supuesto la de los otros coeficientes β .

2. Tasa de error en la discriminación

Para estudiar el comportamiento de una regla es necesario considerar la tasa de error en la clasificación. Existen varios métodos para calcularla (véase Lachenbruch (1975) o Seber (1984)). Aquí se considera el procedimiento propuesto por Lachenbruch y Mickey (1968), llamado el método de validación cruzada o el método de dejar una observación fuera (este procedimiento es llamado algunas veces, incorrectamente, el estimador Jackknife). La técnica consiste en determinar la regla de asignación dejando por fuera una observación, y luego se usa la regla para clasificar la observación omitida. Repitiendo el procedimiento para cada una de las n_1 observaciones del grupo G_1 se obtiene,

$$e_1 = \frac{a_1}{n_1}$$

que es la proporción de observaciones de G_1 mal clasificadas. Con estas proporciones se puede definir la tasa de error como

$$e = p_1 e_1 + p_2 e_2$$

que puede ser estimada por

$$e^* = \frac{a_1 + a_2}{n_1 + n_2}$$

3. La interpretación de los coeficientes en el modelo de discriminación logística

Considérese de nuevo el modelo de discriminación logística

$$\log \left[\frac{f_1(x)}{f_2(x)} \right] = \alpha + \beta'x$$

Cuando X es un vector discreto, el cociente $f_1(x)/f_2(x)$ es la razón de dos probabilidades: $f_1(x)$ la probabilidad de pertenecer a G_1 dado X y de $f_2(x)$ la probabilidad de pertenecer a G_2 dado X . Este cociente es llamado el riesgo relativo de G_1 dado X . Esta cantidad permite medir el incremento en el riesgo a ser pobre por unidad de cambio en una de las variables.

En el modelo logístico, el vector X se relaciona linealmente con el logaritmo del riesgo, denominado logit. Por tanto, se puede estudiar el tipo de efecto y la magnitud, de un cambio en una variable discriminadora sobre el logit de G_1 . Si β_j es positivo, un cambio de una unidad en la característica X_j produce un aumento de β_j en el logit de G_1 -y de e^{β_j} en el riesgo de G_1 -, cuando las demás variables permanecen constantes. Por el contrario, si β_j es negativo, el logit disminuye y también el riesgo de G_1 .

El signo del coeficiente de una variable discriminadora indica el sentido en que ella obra en la clasificación de un hogar: si es positivo, el incremento en los valores tenderá a aumentar el riesgo de ser pobre y si es negativo hará disminuir dicho riesgo.

Estos signos deberían corresponder a las expectativas *a priori* que se tienen sobre el sentido discriminativo de las variables empleadas en el modelo.

Para el caso continuo el resultado anterior se mantiene. Seber (1984).

4. Estrategias para la construcción del modelo

En esta sección se presentan las diferentes etapas consideradas en la construcción del modelo.

a. Tratamiento de las variables categóricas

Algunas de las variables independientes relevantes del modelo son de naturaleza categórica; es decir, cada una de estas variables tiene un número fijo de clases y su escala de medida es, en general, nominal. Debido a esta característica, es inapropiado modelarlas como si fueran variables de intervalo. Se recomienda crear variables de diseño para representar las categorías de cada variable (Hosmer y Lemeshow, 1989). En el caso estudiado aquí, si una variable categórica contiene b categorías, es necesario crear $b-1$ variables indicadoras (*dummies*). De esta forma se evitan problemas de dependencia lineal en la matriz de datos del modelo. La categoría base está representada por el valor del intercepto.

b. Transformaciones sobre las variables independientes continuas

Para las variables independientes continuas se debe chequear el supuesto de linealidad en el logit. Una vez la variable ha sido identificada como importante, se debe obtener la relación paramétrica correcta o la escala de ella en el modelo. Esto puede ser realizado construyendo una gráfica del logit del modelo, en el cual entra la variable en forma lineal, contra la variable. La forma del gráfico permite determinar la forma funcional más adecuada en que la variable debería entrar al modelo. Alternativamente, se puede derivar la transformación más adecuada corriendo una regresión de Box y Tidwell (1962) entre el logit del modelo y la variable correspondiente.

c. Detección de observaciones influenciales y/o atípicas

En esta etapa se consideró la existencia de observaciones que tuvieran una influencia inusual sobre el modelo. Su detección se realizó mediante las medidas de diagnóstico de puntos influenciales basadas en los residuales de Pearson y el residual de desvío (Hosmer y Lemeshow, 1989).

III. Definición de variables y análisis preliminar

El análisis descriptivo⁷ efectuado, permitió seleccionar el conjunto de variables que tienen mayor capacidad para discriminar los hogares pobres del resto no pobre.

Estas constituyeron las hipótesis iniciales de los factores de riesgo de la pobreza urbana a nivel de hogares.

A continuación se explicita el conjunto amplio de variables que fueron sometidas al análisis de discriminación.

A. Variables socio-demográficas

1. Porcentaje de menores de 12 en el hogar

El mayor peso de menores respecto al total de personas en el hogar se espera que afecte negativamente el ingreso per-cápita de hogar mediante dos vías: por una parte, restringe las posibilidades de la participación laboral de las mujeres -sobre las que recae el cuidado de los niños- y, por otra, en general representa una mayor carga de personas dependientes para quienes están en edad de trabajar y de manera especial para los miembros que efectivamente trabajan -esto es, los ocupados en el hogar-. La mayor proporción de menores, en principio, es una variable socio-demográfica; sin embargo, puede estar afectada por causas económicas. En efecto, la imposibilidad de independencia de algunos de los miembros conlleva a que entre los pobres el hogar se amplíe por encima del núcleo básico.

7. Centro de Investigaciones Económicas -CIE-. *Identificación y determinación probabilística de los grupos pobres en Colombia. Informe preliminar*. Universidad de Antioquia, mayo de 1994.

2. Edad del jefe del hogar

La mayor edad del jefe, la cual está ligada al ciclo laboral de las personas, y en términos generales, al de los hogares, debería asociarse en forma negativa con la pobreza. De un lado, los ingresos laborales aumentan con la experiencia y, de otro, a edades mayores -intermedias- del jefe se amplía la participación laboral de los demás miembros del hogar. En las edades más avanzadas del jefe -después de 60 años-, el crecimiento en los ingresos, según lo predicen los estudios sobre el capital humano, tienden a disminuir, aunque dicho decrecimiento puede ser compensado con la acumulación de activos del hogar -por ejemplo, propiedad de la vivienda y/o con la disminución del número de personas en el hogar-.

3. Edad del jefe menos el número total de años de educación del mismo

Esta variable se construyó como indicador aproximado de la experiencia laboral del jefe. Igualmente, se tuvo en cuenta la experiencia laboral, según información directa de la encuesta. *Sin embargo no parece que la experiencia por si sola sea un factor que discrimine hogares pobres.*

4. Sexo del jefe del hogar

Los hogares en cabeza de la mujer no parecen tener un mayor riesgo de pobreza. Al igual que en Medellín no es confirmada dicha hipótesis por el modelo estadístico en el caso de Santafé de Bogotá. Quizás la definición de jefe de hogar en las encuestas dirigidas a medir el nivel de vida de los hogares merece ser revisada.

5. Estado civil y sexo del jefe

Contrariamente a lo esperado y a lo que sucede en otros países, los hogares bajo la *jefatura femenina sin cónyuge* -"mujeres separadas, viudas y solteras"-, no resultaron ser más propensos a la pobreza que el resto.

Así mismo, se consideró el sexo del jefe del hogar relacionado con su estado conyugal -con ello se aislan los hogares en cabeza de mujeres solteras-; la inestabilidad de las relaciones conyugales según análisis descriptivo, afecta negativamente el ingreso per-cápita de los hogares. Pero, a pesar de que tanto en los hogares *bajo unión libre* -tanto con jefatura femenina como masculina- como en aquellos cuyo jefe es una mujer separada se presentan los menores niveles de ingreso per-cápita por hogar, dichas características de los jefes de hogar no tienen capacidad para discriminar los hogares pobres.

6. Estrato

La estratificación socio-económica de los hogares basada en la ubicación y características físicas de la vivienda se espera, en principio, guarde una relación positiva con el ingreso per-cápita del hogar. Por tanto, a medida que el hogar ascienda de estrato debería reducirse su probabilidad de ser pobre.

B. Variables educativas

La educación constituye, sin lugar a dudas, el factor determinante de los ingresos de los ocupados. Al aumentar el nivel educativo del jefe y de los demás miembros del hogar se eleva el bienestar económico y de manera sensible se mejora la calidad de vida en general. Según M. Desai (1992) el acceso a la educación primaria y secundaria debe considerarse una parte esencial del nivel de vida, "la falta de educación es un factor determinante de la probabilidad de ser pobre según los datos de los países desarrollados, y no hay razón para dudar que ello sea cierto en otros países".⁸ Las variables educativas consideradas fueron las siguientes:

8. Meghnad Desai, Bienestar y privación vitales: Propuesta para un índice de progreso social. En: *Comercio Exterior*. Vol. 42. México, abril de 1992. p. 334.

1. El nivel educativo del jefe

Es de esperarse que éste, a la vez que presenta una asociación positiva con el ingreso per-cápita de los hogares, constituya un factor decisivo del bienestar, puesto que se traduce en mejoras en nutrición, salud, control de la natalidad, asistencia escolar, índice de mortalidad infantil, etc.

Inicialmente dicha variable se ordenó en ocho categorías: sin educación, con máximo 3 años de primaria, hasta 5 años -o primaria completa-, secundaria incompleta -máximo 4 años-, secundaria completa, universitaria incompleta -hasta 3 años-, universitaria completa -5 años-, y posgrado. Sin embargo, de acuerdo con la metodología de la regresión logística se trataron como variables *dummies*.

2. Máximo nivel educativo de los ocupados en el hogar

Según análisis descriptivo el ingreso per-cápita presenta una asociación directa con el máximo nivel educativo de sus ocupados. Así mismo, la distribución de hogares según el máximo nivel educativo de los ocupados indica también la relación positiva de esta variable con el ingreso per-cápita: mientras en los dos deciles más bajos, un poco más del 80% de los hogares, los ocupados han cursado sólo hasta máximo 4 años de secundaria, en los deciles 9 y 10 -20% más rico- dicha cifra se sitúa aproximadamente entre 11% y 9% respectivamente. Lo anterior da sustento a la hipótesis que postula una mayor probabilidad de pobreza entre menor sea el máximo nivel alcanzado de educación de los ocupados en el hogar. Al igual que el nivel educativo del jefe en dicha variable se establecieron inicialmente 6 niveles, que fueron tratados en el modelo como variables *dummies*.

3. Nivel educativo de la Población en Edad de Trabajar -PET-

El mayor capital humano reunido en el hogar (esto es, invertido en las personas en edad de trabajar) es de esperarse reduzca, al igual que las anteriores variables, el riesgo de pobreza. Dicha variable se estableció tomando el promedio de años de estudio por persona -mayor de 12 años-,

y también mediante la sumatoria de años de estudio de la población en edad de trabajar.

C. Variables laborales

Con excepción de la información de las personas ocupadas, la de las demás características referentes a la actividad principal de los miembros del hogar -inactivos, desocupados, población activa, etc.- es bastante deficiente. Por ello sólo pudieron ser tenidas en cuenta las variables relacionadas con las personas ocupadas.

1. Número de ocupados por hogar

En Colombia la pobreza urbana a nivel de los hogares no sólo puede atribuirse a la falta de educación, causante, a su turno, de los bajos salarios. Además del bajo nivel educativo y de los factores demográficos que afectan negativamente el ingreso per-cápita de los hogares: el desempleo urbano golpea más fuertemente los estratos bajos -concentran la mayoría de los pobres de nuestras principales ciudades-. *El mayor desempleo de los pobres se traduce en un menor número de ocupados por hogar.* En el análisis descriptivo dicho factor, para el caso de Santafé de Bogotá, se asocia más con la pobreza extrema. Mientras en los deciles 1 y 2 se constata la más baja relación de ocupados por hogar: 1.2 y 1.4 respectivamente, en el resto de hogares se sitúa cerca de 1.8 ocupados por hogar

2. Porcentaje de ocupados sobre el total de personas en el hogar -inverso de la tasa de dependencia-

Esta es una variable de gran importancia puesto que relaciona el tamaño del hogar con el número de ocupados. Entre mayor sea el porcentaje de personas ocupadas por hogar -*por ende menor el inverso de la tasa de dependencia económica*- la probabilidad de pobreza es más reducida. En efecto, los pobres son, en general, más numerosos y tienen un número de ocupados por hogar más bajo. En Santafé de Bogotá las tasas de dependencia en los primeros 3 deciles son en su orden de 4.3, 3.2 y 2.6 -por persona ocupada-, mientras en los 3 deciles superiores oscilan entre 1.6 y 2.0.

3. Porcentaje de ocupados en el hogar en la categoría de servicio doméstico

Se esperaba que dicha proporción fuese más frecuente en los hogares pobres. Sin embargo, según el modelo no tiene capacidad para discriminar el grupo de hogares pobres. Así mismo, se establecieron otras variables basadas en las categorías ocupacionales, tales como *número de patrones, porcentaje de trabajadores por cuenta propia con máximo primaria*, no tuvieron tampoco ningún poder clasificatorio.

4. Jornada laboral de los ocupados del hogar

En un primer ensayo se estableció la variable conformada por la sumatoria de la jornada total del hogar, pero al ser descartada por el modelo *se reemplazó por la jornada laboral por ocupado -y por hogar-*, variable que quedó incluida en el modelo definitivo. Según los resultados obtenidos las mayores jornadas de los ocupados están asociadas con una mayor probabilidad de pobreza.

En el modelo aplicado para Medellín la hipótesis sobre la asociación directa entre el empleo informal y la pobreza se confirmó mediante la construcción de una variable aproximada que relaciona la jornada de ocupados informales -cuentas propias-, patrones y asalariados en empresas de menos de 10 trabajadores con la jornada total del hogar. La Encuesta de Calidad de Vida y Pobreza para Santafé de Bogotá no permite obtener dicha información según tamaño de los establecimientos cruzada con las categorías ocupacionales.

D. Indicadores de calidad de vida o de necesidades básicas insatisfechas

Según el estudio del DANE-PNUD, en 1988, el 34.6% de la población en Santafé de Bogotá era pobre, aplicando el método de la Medición Integral de Pobreza -MIP-, el cual complementa la medición según una línea de pobreza con la insatisfacción de necesidades básicas. En dicho trabajo el 12.9% de la población es pobre según el método de necesidades básicas insatisfechas, en tanto que el porcentaje de incidencia es igual a

30.2% de acuerdo al método de la línea de pobreza. Igualmente, es importante señalar que se constata una alta proporción de pobres con al menos una necesidad básica insatisfecha que, a su vez, se encuentra por debajo de la línea de pobreza -el 66% de la población con NBI-. *Lo cual indica una alta probabilidad de ser pobre por deficiencia de ingresos cuando no se satisface al menos una necesidad básica.*

No todos los indicadores de NBI, más frecuentemente utilizados en Colombia y América Latina, pudieron ser tenidos en cuenta en el modelo estadístico trabajado aquí, porque el alto porcentaje de preguntas, referidas a la calidad de la vivienda, acceso de los servicios públicos, niveles de nutrición, morbilidad, entre otras, no fueron respondidas -aparecen con NO INFORMA en la base de datos-.

También se consideran otros indicadores de calidad de vida, de utilización menos frecuente, reportados en la Encuesta de Calidad de Vida y Pobreza, tales como afiliación a la seguridad social, menores que trabajan y grado de dificultad para acceder al pago de bienes y servicios privados básicos. A continuación se detallan las variables relacionadas con NBI e indicadores de calidad de vida:

1. Hogares cuyos miembros no están afiliados al ISS, tienen seguro de salud, ni afiliación a cajas de compensación familiar

En el País, entre los pobres es más frecuente la ausencia o baja cobertura de la seguridad social.

2. Menores de 12 años que trabajan

Este es uno de los síntomas más claros de pobreza a nivel de hogares urbanos.

3. Hogares con al menos un menor entre 6 y 12 años que no asiste a la escuela

Esta variable mide el acceso de los menores de 12 años a la educación primaria. En gran medida su carencia está relacionada en forma directa con la pobreza, más que con restricciones en la oferta educativa urbana.

4. Número de personas por cuarto -diferente a cocina, baño o garaje-

Los estudios empíricos en Colombia han tomado como un indicador de la calidad de vida el "hacinamiento crítico" -3 ó más personas por cuarto-; en general, ésta es mayor cuando aumenta la pobreza. Según la Encuesta de Calidad de Vida y Pobreza el 92% de los hogares con "hacinamiento crítico" se concentran en los primeros 4 deciles.

5. Grado de dificultad monetaria del hogar para el pago de bienes y servicios básicos

Sólo se tomó la variable referida a "mucha dificultad para cubrir los pagos de arriendos y alimentación".

6. Hogares con "alta dependencia económica"

-Con 3 ó más personas por ocupado y cuyo jefe tiene un nivel de máximo 5 años de primaria-. Según los resultados del análisis descriptivo, en el 30% de hogares pobres -3 primeros deciles- se observa la mayor proporción de hogares con "alta dependencia económica" -tal como se ha definido-. Además, los hogares en donde coincide un bajo nivel educativo del jefe y una alta tasa de dependencia, se concentran en un 73% en los tres primeros deciles.

7. Hogares con condiciones deficientes de la vivienda

Sin cocina de uso exclusivo -cocina en dormitorio o patio-; con pisos en tierra y habitación no destinada a vivienda o cuando se limita a un cuarto.

8. Hogares que carecen de los servicios públicos esenciales

Acueducto y sin servicio sanitario.

E. Análisis de problemas y consistencia de la base de datos

Se detectaron entre otros los siguientes problemas e inconsistencias en la información recibida:

1. Algunas variables agregadas mostraron inconsistencias, pues al agrupar la información de las subvariables no coincidía con la de los valores de la variable agregada o total. Por ejemplo, la suma de asalariados y no asalariados difería con la del total de ocupados. Cosa similar ocurrió con los ocupados según rama de actividad económica y con la jornada mensual de trabajo considerado de acuerdo con la formalidad o informalidad del mismo. La sumatoria de los ingresos clasificados por fuentes no coincidía tampoco con los ingresos estimados por cada persona en el hogar, fundamentalmente por la no información de algunos de esos ingresos parciales.

2. Algunos hogares informaban ingresos del hogar iguales a cero, a pesar de que manifestaban tener ocupados. Estos fueron retirados de la base para efectos de modelación.

Con algunas de las variables citadas se construyeron otras, buscando medir la interacción entre las mismas o variables no presentes en la base de datos, como la inversa de la Tasa De Dependencia -TDD- explicada posteriormente y se espera refleje condiciones de riqueza, así mismo personas por cuarto -PERCU- y porcentaje de menores -PORMEN-.

IV. Resultados del modelo de discriminación para la ciudad de Santafé de Bogotá

A. Especificación del modelo

De acuerdo con lo planteado en el Modelo de Clasificación en la Sección II, se construyó una función de discriminación logística para el caso de Santafé de Bogotá.

Se hicieron ensayos con varias líneas, iniciando con un ingreso per cápita por hogar de \$35000 llevándolo hasta \$45000 con incrementos de \$1000, con el fin de encontrar el modelo con más poder discriminador y de mayor significancia por parte de las variables consideradas.

$$\log \left[\frac{f1(X)}{f2(X)} \right] = \alpha + \beta_1 \bullet \text{PORMENT} + \beta_2 \bullet \text{TDDIT} + \beta_3 \bullet \text{NEJ23} + \beta_4 \bullet \text{NEJ45} + \beta_5 \bullet \text{NEJ6} + \beta_6 \bullet \text{E4} + \beta_7 \bullet \text{E56} + \beta_8 \bullet \text{M2} + \beta_9 \bullet \text{MA3} + \beta_{10} \bullet \text{MA4} + \beta_{11} \bullet \text{MA56} + \beta_{12} \bullet \text{HNA} + \beta_{13} \bullet \text{JOROCUP} + \beta_{14} \bullet \text{PERCUT} + \beta_{15} \bullet \text{EDADJT} + \beta_{16} \bullet \text{INAESC}$$

Donde,

- α : representa una constante
- β_i : representa el coeficiente para la i -ésima variable discriminadora
- PORMENT**: porcentaje de menores con respecto al total de personas en el hogar -variable transformada elevándola a la 0.25-.
- TDDIT**: inverso de la tasa de dependencia del hogar, número de ocupados dividido por la población total del hogar. Se tomó el inverso para evitar divisiones por cero (variable elevada al cuadrado).
- NEJ23**: agrupación de los niveles educativos de secundaria completa e incompleta del jefe del hogar.
- NEJ45**: niveles educativos con universidad incompleta y universidad completa del jefe del hogar.⁹
- NEJ6**: jefe con nivel educativo hasta postgrado.
- E4**: hogares pertenecientes al estrato socioeconómico 4.
- E56**: hogares pertenecientes al estrato socioeconómico 5 y 6
- MA2**: máximo nivel de educación de los ocupados del hogar: secundaria incompleta.
- MA3**: máximo nivel de educación de los ocupados del hogar: secundaria completa.
- MA4**: máximo nivel de educación de los ocupados del hogar: universidad incompleta.

9. La unión de las categorías se realizó después de una estimación preliminar, la cual mostró que los coeficientes individuales eran aproximadamente iguales.

- MA56: máximo nivel de educación de los ocupados del hogar: universidad completa.
- HNA: hogares cuyos miembros no están afiliados al ISS, ni tienen seguro de salud, ni afiliación a cajas de compensación familiar.
- JOROCUP: suma de la jornada principal de cada uno de los ocupados del hogar dividida por el número total de ocupados.
- PERCUT: número de personas en el hogar dividida por el número de cuartos -diferentes a cocina, baño y garaje- para dormir en la vivienda -variable logaritmizada-.
- EDADJT: edad del jefe del hogar -variable transformada elevándola a 0.5-.
- INAESC: inasistencia escolar de menores. Variable dicótoma que toma el valor de uno (1) cuando todos los menores del hogar asisten a la escuela y de cero (0) cuando al menos uno no asiste.

Empíricamente se espera que variables como el estrato, el nivel educativo del jefe del hogar, el máximo nivel educativo de los ocupados, el inverso de la tasa de dependencia, la edad del jefe del hogar y, por último, la inasistencia escolar, disminuyeran el riesgo de un hogar de pertenecer a la población pobre a medida que crecen; para que esto ocurra es necesario que el signo de los coeficientes asociados a cada una de estas variables sea negativo. Para las otras variables se espera un efecto positivo sobre el "logit", indicando una incidencia directa sobre la probabilidad de pobreza de un hogar. El signo de sus coeficientes debería ser positivo. Así por ejemplo, el signo de la variable jornada por ocupado es positivo en el sentido de que mientras mayor sea la jornada laboral que deban cumplir los ocupados del hogar, mayor será su probabilidad de pobreza.

Se corrieron los modelos para cada una de las líneas definidas anteriormente y se presentan los resultados obtenidos para el modelo

con línea de 40000 pues fue el que mejor se comporta económica y estadísticamente.

B. Interpretación de los resultados del modelo con línea de 40000

Empleando el procedimiento logístico del paquete estadístico SAS, se obtuvieron los siguientes resultados:

Número total de observaciones: 2693

Número de observaciones en el grupo 1: 1072

Número de observaciones en el grupo 2: 1621

Después de realizar diagnósticos de puntos influyentes y atípicos, 25 observaciones.

1. Criterios para contrastar el ajuste del modelo

Criterion	Intercept and Covariates		Chi-Square for Covariates
	Intercept Only		
AIC	3622.582	1980.225	
SC	3628.481	2080.498	
-2 LOG L	3620.582	1946.225	1674.357 with 11 DF (p=0.0)
Score			1215.412 with 11 DF (p=0.0001)

2. Análisis de los estimadores máximo verosímiles

Analysis of Maximum Likelihood Estimates					
Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	2.6003	0.6992	13.8315	0.0002	
PORMENT	0.3163	0.1779	3.1612	0.0754	0.067491
TDDIT	-5.1791	0.3941	172.6690	0.0001	-0.831449
NEJ23	-0.7990	0.1638	23.7956	0.0001	-0.214823
NEJ45	-1.3023	0.2603	25.0315	0.0001	-0.299549
NEJ6	-1.9521	0.6546	8.8940	0.0029	-0.216734
E4	-0.8797	0.1530	33.0539	0.0001	-0.219947
E56	-1.4361	0.3145	20.8504	0.0001	-0.227739
MA2	0.4985	0.1964	6.4400	0.0112	0.107538
MA3	-0.0288	0.1880	0.0234	0.8784	-0.006416
MA4	-0.8589	0.2671	10.3425	0.0013	-0.133405
MA56	-1.0156	0.2654	14.6412	0.0001	-0.242353
HNA	0.5555	0.1212	20.9920	0.0001	0.146009
JOROCUP	0.00763	0.00299	6.5121	0.0107	0.086451
PERCUT	1.3683	0.1202	129.6103	0.0001	0.519296
EDADJT	-0.1749	0.0647	7.3058	0.0069	-0.106132
INASESC	-0.7958	0.4314	3.4025	0.0651	-0.062068

3. Asociación de las probabilidades predichas y las respuestas observadas

Association of Predicted Probabilities and Observed Responses

Concordant	=	91.3%	Somers' D	=	0.827
Discordant	=	8.6%	Gamma	=	0.828
Tied	=	0.1%	Tau-a	=	0.397
(1737712 pairs)		c	=		0.917

La siguiente tabla muestra la capacidad de discriminación del modelo de acuerdo con la línea de pobreza:

4. Tabla de clasificación

Tabla de Clasificación

		Predicted		Total
		EVENT	NO EVENT	
Observed	EVENT	837	235	1072
	NO EVENT	236	1385	1621
Total		1073	1620	2693

Sensitivity= 78.1% Specificity= 85.4% Correct= 82.5%

False Positive Rate= 22.0% False Negative Rate= 14.5%

Nota: un evento equivale a una clasificación de "Pobre"

5. Matriz de correlación de los coeficientes estimados

Estimated Correlation Matrix						
Variable	INTERCPT	PORMENT	TDDIT	NEJ23	NEJ45	NEJ6
INTERCPT	1.00000	-0.37503	-0.22744	-0.32019	-0.31475	-0.13105
PORMEN	-0.37503	1.00000	0.23126	0.01365	0.02757	0.01293
TDDI	-0.22744	0.23126	1.00000	0.15971	0.16789	0.05368
NEJ23	-0.32019	0.01365	0.15971	1.00000	0.39798	0.14117
NEJ45	-0.31475	0.02757	0.16789	0.39798	1.00000	0.23139
NEJ6	-0.13105	0.01293	0.05368	0.14117	0.23139	1.00000
E4	-0.07219	0.01543	0.07593	-0.05062	-0.11016	-0.05827
E56	0.02404	-0.02492	0.06055	-0.07686	-0.06456	-0.05755
MA2	0.06988	-0.08835	-0.17850	-0.50197	-0.14476	-0.04640
MA3	0.10102	-0.09945	-0.13834	-0.53564	-0.15792	-0.04700
MA4	0.11479	-0.07201	-0.12160	-0.23040	-0.42855	-0.09101
MA56	0.17150	-0.09742	-0.12334	-0.26056	-0.55716	-0.25230
HNA	-0.15161	0.05550	-0.02242	0.00194	-0.03117	0.02405
JOROCUP	-0.23758	-0.07470	-0.14546	0.12393	0.15228	0.06978
PERCU	-0.07464	-0.22683	-0.12620	0.09618	0.09518	0.02272
EDADJ	-0.74621	0.31592	0.18913	0.34651	0.38701	0.10886
INASESC	-0.63319	0.12096	0.02831	0.05192	0.00077	0.05216

continuación matriz de correlación

Variable	E4	E56	MA2	MA3	MA4	MA56
INTERCPT	-0.07219	0.02404	0.06988	0.10102	0.11479	0.17150
PORMEN	0.01543	-0.02492	-0.08835	-0.09945	-0.07201	-0.09742
TDDI	0.07593	0.06055	-0.17850	-0.13834	-0.12160	-0.12334
NEJ23	-0.05062	-0.07686	-0.50197	-0.53564	-0.23040	-0.26056
NEJ45	-0.11016	-0.06456	-0.14476	-0.15792	-0.42855	-0.55716
NEJ6	-0.05827	-0.05755	-0.04640	-0.04700	-0.09101	-0.25230
E4	1.00000	0.14200	0.03175	-0.02289	0.02568	-0.10983
E56	0.14200	1.00000	0.00798	0.02134	0.03494	-0.01871
MA2	0.03175	0.00798	1.00000	0.65303	0.36179	0.35776
MA3	-0.02289	0.02134	0.65303	1.00000	0.38419	0.39432
MA4	0.02568	0.03494	0.36179	0.38419	1.00000	0.44445
MA56	-0.10983	-0.01871	0.35776	0.39432	0.44445	1.00000
HNA	-0.00073	-0.00396	0.08501	0.13754	0.14363	0.13779
JOROCUP	0.00860	-0.00863	-0.23469	-0.24601	-0.20775	-0.22914
PERCU	0.14658	0.06748	-0.03946	-0.00779	-0.00255	0.01145
EDADJ	-0.01196	-0.06713	-0.08254	-0.11837	-0.15399	-0.20380
INASESC	0.04999	-0.00055	-0.05294	-0.07859	-0.04546	-0.06092

continuación matriz de correlación

Variable	HNA	JOROCUP	PERCUT	EDADJT	INASESC
INTERCPT	-0.15161	-0.23758	-0.07464	-0.74621	-0.63319
PORMENT	0.05550	-0.07470	-0.22683	0.31592	0.12096
TDDIT	-0.02242	-0.14546	-0.12620	0.18913	0.02831
NEJ23	0.00194	0.12393	0.09618	0.34651	0.05192
NEJ45	-0.03117	0.15228	0.09518	0.38701	0.00077
NEJ6	0.02405	0.06978	0.02272	0.10886	0.05216
E4	-0.00073	0.00860	0.14658	-0.01196	0.04999
E56	-0.00396	-0.00863	0.06748	-0.06713	-0.00055
MA2	0.08501	-0.23469	-0.03946	-0.08254	-0.05294
MA3	0.13754	-0.24601	-0.00779	-0.11837	-0.07859
MA4	0.14363	-0.20775	-0.00255	-0.15399	-0.04546
MA56	0.13779	-0.22914	0.01145	-0.20380	-0.06092
HNA	1.00000	0.02317	-0.02608	0.08073	0.00884
JOROCUP	0.02317	1.00000	-0.08773	0.16174	0.02114
PERCUT	-0.02608	-0.08773	1.00000	0.13750	-0.02724
EDADJT	0.08073	0.16174	0.13750	1.00000	0.01898
INASESC	0.00884	0.02114	-0.02724	0.01898	1.00000

De los resultados anteriores se puede concluir:

De 1. los estadísticos del Score y -2Log L que proporcionan contrastes para la significancia conjunta de las variables explicativas, permiten concluir que los efectos de las variables discriminadoras son significativos. Los estadísticos AIC (Akaike Information Criterion) y SC (Schwartz Criterion) permitieron elegir este modelo frente a otros, dado que tenían valores más bajos.

De 2. se confirma el poder de discriminación de cada una de las variables seleccionadas. Para cada variable de la tabla el estadístico Chi-

cuadrado de Wald permite contrastar su significancia individual. En cada caso, excepto para MA3, se obtiene un valor p cercano a cero el cual se puede interpretar como un fuerte soporte muestral a la significancia de cada variable discriminadora.

Para las variables dicotomizadas, -nivel educativo del jefe del hogar, máximo nivel de educación de los ocupados y estrato-, se observa, en sus coeficientes, un ordenamiento de acuerdo con su comportamiento esperado *a priori*. Para el nivel educativo del jefe del hogar se pueden diferenciar cuatro categorías para discriminar pobreza, así: nivel 1 efecto recogido en el intercepto, NEJ23 unión de los niveles secundaria completa y secundaria incompleta para los cuales el valor del coeficiente es el mismo, NEJ45 unión de los niveles universidad incompleta y universidad completa y, por último, la variable NEJ6, posgrado, indicando que esta categoría tiene un poder discriminador diferente.

En lo que respecta al estrato se observa que no hay diferencias en la capacidad discriminadora de la pobreza en los estratos 1, 2 y 3, ya que sus coeficientes son bastantes cercanos entre sí. Igual cosa sucede con los estratos 5 y 6. Por consiguiente es posible diferenciar sólo tres grupos que discriminan pobreza así: E1, E2 y E3 cuyo efecto se recoge en el intercepto; E4 permanece independiente y E5, E6 que conforman una categoría diferente. Esto demuestra que no hay una clara estratificación de los hogares en la ciudad de Santafé de Bogotá.

La variable máximo nivel educativo de los ocupados en su nivel 3, MA3, secundaria completa, tiene asociado un coeficiente estadísticamente igual a cero dado su bajo nivel de significancia. Los niveles cinco y seis -universidad completa y posgrado- se unen para formar una sola categoría, pues sus coeficientes toman valores muy cercanos.

Seguidamente 3. da información sobre la capacidad predictiva del modelo. Los valores de los coeficientes D, Gamma, y C, cercanos a uno, indican buena habilidad predictiva. El coeficiente Tau-a, aunque generalmente es pequeño, es positivo; lo que señala de nuevo la capacidad de predicción del modelo. Este coeficiente no es recomendado por algunos autores en caso de existir empates.

De 4. se obtiene el comportamiento del modelo para predecir cada una de las observaciones. El procedimiento empleado fue el de la validación cruzada, descrito en la Sección llamada **Tasa de error en la discriminación**. Se concluye que el procedimiento es capaz de clasificar correctamente el 82.5% de todas las observaciones. La tasa de clasificación correcta para hogares pobres fue de 78.1% -sensibilidad- y de los hogares no pobres fue de 82.5% -especificidad-. La tasa de error del modelo fue de 17.5%

Por último de 5. se puede concluir la inexistencia de problemas de colinealidad entre las variables discriminadoras. Esto significa que cada variable está participando en la clasificación de un hogar en forma relativamente independiente.

C. Empleo de costos diferenciales de clasificación

El costo de clasificar a un pobre como rico o a un rico como pobre incide de manera diferente en la clasificación de los hogares. Si el interés del ente planificador se centra en la “captura” del mayor número de pobres, el costo de clasificar mal a un hogar perteneciente a esta población debiera ser mayor que el costo de clasificar erróneamente a un hogar rico. Al considerar tal interés en la especificación del modelo, se puede ser más estricto en el primer error y discriminar, por tanto, un porcentaje mayor de hogares pertenecientes al grupo de pobres.

A continuación se presentan algunos resultados de clasificación asumiendo costos diferentes para los dos tipos de error. El modelo general, presentado anteriormente, supone costos iguales.

COSTOS (C1/C2)

CLA	20/80	30/70	40/60	45/55	55/45	60/40	70/30	80/20
S								
Sen%	92.77	88.19	83.79	80.68	73.81	69.51	62.00	51.37
Esp%	67.53	74.29	80.20	82.90	87.02	89.54	93.11	94.22

Nota: $C1/C2$ = costo asignado a la clasificación de no pobre/costo asignado a la clasificación de pobre. Los resultados para el costo 50/50 se presentan con los demás resultados del modelo.

D. Construcción de una región de indecisión

Existen hogares cuyo vector de características puede no ser claramente clasificado por el procedimiento en uno de los dos grupos. Cuando esto ocurre, la predicción de la probabilidad de pertenecer al grupo de pobres hecha por el modelo se encuentra generalmente muy próxima a 0.5. Basados en este estimador se construyó un intervalo del 90% de confianza para la verdadera probabilidad (Cox y Snell, 1989) de pertenecer al grupo de pobres para cada hogar.

Un hogar con dificultades de ser clasificado tendrá la característica de que el intervalo para su probabilidad de pertenencia al grupo de pobres tiene el límite inferior menor que 0.5 y el límite superior mayor que 0.5. Con el conjunto de hogares cuyos intervalos cumplen con esta condición, y tomando el mínimo de los límites inferiores y el máximo de todos los límites superiores, se obtuvieron los "logits" correspondientes y se construyó una región la cual se caracteriza por contener los hogares de difícil clasificación. Se debe poner atención especial a todo hogar cuyo "logit" caiga dentro de ella. El conjunto encontrado fue el intervalo (-1.94, 1.11) el cual se llamará región de indecisión.

De acuerdo con esto la regla de clasificación puede ser modificada como sigue:

Clasifique un hogar como pobre si su "logit" es mayor que 1.11, clasifique el hogar como no pobre si su "logit" es menor que -1.94 y no lo clasifique en ninguno de los dos grupos si su "logit" cae entre -1.94 y 1.11. La proporción de hogares de la muestra que cae en esta zona es de 9.17% aproximadamente. Con esto, el modelo queda con una tasa de error de 8.3%.

E. El "logit" como un indicador del tamaño de la pobreza

Al aplicar la función de discriminación logística descrita en la Sección IV. A. (Especificación del modelo), a las variables especificadas en el modelo para cada uno de los hogares, se obtiene un valor correspondiente al logaritmo de la probabilidad de pobreza denominado "logit". La magnitud de éste puede ser considerado como un indicador de la pobreza del hogar; es decir, mientras mayor sea el logit mayor es la pobreza del hogar, y mientras menor lo sea más lejos está dicho hogar de pertenecer al grupo de los hogares pobres.

Así pues, al comparar dos hogares con un logit de 7.83 y 3.24, respectivamente, se puede concluir que el primer hogar es más pobre que el segundo.

Conclusiones

1. Los resultados representan un avance sustancialmente significativo, respecto a los obtenidos para Medellín. Sin embargo, es interesante señalar la alta coincidencia, en ambos modelos, de aquellas variables seleccionadas que según distintos trabajos, tienen mayor incidencia en el fenómeno de la pobreza.

2. Dentro de las variables con capacidad para discriminar pobres, de acuerdo al modelo, algunas de ellas corresponden a factores demográficos y socio-económicos, que además de ser verificables, tienen una alta significación económica como determinantes del ingreso per-cápita del hogar, esto es: tasa de dependencia, nivel educativo del jefe, máximo nivel educativo de los ocupados en el hogar, porcentaje de menores, edad del jefe y jornada laboral promedio de los ocupados. Otras variables, en cambio, son características o indicadores de calidad de vida estrechamente asociados a la pobreza urbana, que más bien son efectos o síntomas de la misma.¹⁰ Entre estas variables las seleccionadas fueron: hogares cuyos menores de 12 años no asisten a la escuela, ausencia de afiliación

10. Aunque en el largo plazo, estos síntomas tienden a operar como factores causales de la reproducción de la pobreza.

a la seguridad social, el estrato y el hacinamiento crítico de la vivienda. En el modelo de Medellín, por limitaciones en la información, sólo el estrato y la no afiliación a la seguridad social quedaron incluidos dentro de este subgrupo de variables.

3. El ordenamiento, de mayor a menor, de las variables según su peso relativo clasificatorio fue el siguiente: la tasa de dependencia -definida como porcentaje de ocupados en el hogar-; el índice de hacinamiento -personas por cuarto-; la educación del jefe del hogar, el máximo nivel educativo de los ocupados; inasistencia escolar de al menos un menor de 12 años, ausencia de afiliación a la seguridad social -ISS o seguro médico o Cajas de Compensación- de los miembros del hogar y el estrato -estas dos últimas variables tienen un peso similar-, porcentaje de menores, y, finalmente, edad del jefe y jornada promedio de ocupados, con pesos también similares.

4. Algunas de las hipótesis, establecidas *a priori*, o señaladas en otros estudios, que se esperaba discriminaran los hogares más pobres, *no fueron confirmadas por el modelo*. En primer lugar, el fenómeno de la feminización de la pobreza no tuvo capacidad para discriminar pobreza en el caso de la ciudad de Santafé de Bogotá -resultado similar al obtenido en el modelo aplicado para Medellín-. Así mismo, ni la jefatura femenina sin cónyuge -"mujeres solas"-, ni la inestabilidad de las relaciones conyugales -esto es, bajo unión libre y separadas-, fueron confirmados como factores significativos en la clasificación de los pobres. Tampoco pudo ponerse por deficiencias en información de la asociación directa entre grado de informalidad del empleo y pobreza, la cual quedó confirmada para Medellín. Finalmente, la conclusión anterior -no confirmación por el modelo- también se aplica a los indicadores de necesidades básicas insatisfechas referidas a las condiciones de la vivienda -con excepción del hacinamiento- y al acceso a los servicios públicos. En parte ello puede deberse a la falta de información, o bien a que la pobreza por deficiencia de ingresos -bajo ingreso per-cápita- no puede identificarse únicamente por Necesidades Básicas Insatisfechas -NBI-, debido al alto cubrimiento de los servicios básicos del Estado incluso dentro de los sectores pobres.

5. Al mover la línea de pobreza en el rango de \$35.000 a \$45.000 se observa que el mejor modelo, estadísticamente hablando, se presenta al utilizar \$40.000 como "línea" -per-cápita y por mes-. Con dicha "línea", las variables seleccionadas tienen una mejor significación y su peso relativo concuerda con el pensamiento económico en lo que se refiere a las causas y consecuencias de la pobreza. Este valor de \$40.000 como ingreso per-cápita para identificar el grupo pobre, parece sobrestimar la extensión de la pobreza en Santafé de Bogotá en 1.991. Sin embargo, por los resultados obtenidos en cuanto a la claridad de las variables desde el punto de vista económico, parece adecuado para identificar los pobres. *Debido a ello, insistimos en la conveniencia de revisar la línea de pobreza del DANE.*

6. El modelo también puede ser utilizado como ordenador del grado de pobreza de los hogares, "a mayor logit mayor pobreza".

7. Finalmente, el comportamiento del "estrato socio-económico" como variable clasificatoria, en el caso de Santafé de Bogotá no parece corresponder a una clara diferenciación del ingreso per-cápita de los hogares. Según los resultados del modelo los tres primeros estratos (1, 2 y 3) tienen una influencia similar sobre probabilidad de pobreza y lo mismo sucede con los estratos 5 y 6. La estratificación de las viviendas de la ciudad no permite por tanto hacer diferenciación en el ingreso per-cápita en los 3 primeros deciles, ni en los deciles 5 y 6 superiores.

Referencias

Ahmed, S.W. *et al.* *Discriminant analysis when one or both of the initial samples is contaminated: large sample results.* EDV. in *Medizin und Biologie*, 6. p.35-42. 1975.

Ahmed, S.W. *et al.* *Discriminant analysis when scale contamination is present in the initial sample.* In J. Van Ryzin (Ed), *Classification and Clustering*, pp. 331-353. Academic Press: New York. 1977

Aitchison, J. *et al.* *A critical comparison of two methods of statistical discrimination.* *Appl. Stat* 26. p.15-25. 1977

Amartya K. Sen. *Sobre conceptos y medidas de pobreza, en Comercio Exterior*. Vol 42 N0.4. México, abril de 1992.

Anderson, J.A. and Blair, V. *Penalized maximum likelihood estimation in logistic regression and discrimination*. *Biometrika* 69. p.123-136. 1982.

Bell, David NF. *et al. Poverty in Australia: A study of the implication of educations, House hold Formation and the labour market experience of the young*. Mec. Abril de 1991. Australia.

Broffitt et al. *The efect of Huberizing and trimming on the quadratic discriminant function*. *Stat. Theor. Methods A*, 9. 13-25. 1980.

Byth y McLachlan. *Logistic Regrssion compared to normal discrimination for non-normal populations*. *Aust. J. Stat*, 22. 188-196. 1980.

Cacoullos, T. *Discriminant Analisis and Applications*. Academic Press: New York. 1973

Callan Timand and Brian, Nolan. *Concepts of Poverty and the Poverty Line*. *Journal of Economic Surveys*. Vol. 5, No. 3.

Castaña, Elkin, y otros. *Identificación y determinación probalística de los grupos pobres: un modelo aplicado para Medellín*. En *Revista de Planeación y Desarrollo*. Vol XXIII No.2 Santafé de Bogotá, septiembre de 1992.

Chinganda, E.F. and Subrahmaniam, K. *Robustness of the linear discriminant function to nonnormality: Johnson's System* *J. Stat. Plan. Infer.*, 3. p. 69-77.

Clarke, W.R., *et al. How Non-normality affects the cuadratic discriminant function*. *Commun. Stat. Theor. Methos A*, 8, p.1285-1301. 1979

Colasanto, Diane *et al. Two subjetive definitions of Poverty: Results from the Wisconsin Basic Needs Study*.

Cox, D.R. and Snell, E.J. *The Analysis of Binary Data*, 2d Edition, London: Chapman and Hall.

Crawley, D.R. *Logistic discrimination as an alternative to Fisher's linear discrimination function*. N.Z. Stat., 14(2) p. 21-25. 1979

David W. Hosmer, Jr. and Stanley Lemeshow. *Applied Logistic Regression*. N.Y. Ed. John Wiley Sons. 1989.

Fresneda, Oscar. *Dimensión y características de la pobreza en Colombia según la metodología de Medición Integrada (MIP), en Pobreza, violencia y desigualdad para la Nueva Colombia*. Oscar Fresneda y Otros. PNUD. Proyecto Regional para la superación de la pobreza. Noviembre de 1991. Colombia p.42.

Goldstein, M. and Dillon, W.R. *Discrete Discriminant Analysis*. Wiley: New York. 1978

Hagenaars Aldi and Klaar de Vos. *The Definition and Measurement of Poverty*. The Journal of Human Resources. XXIII-2.

Huberty, C.J. *Discriminant Analysis*. Rev. Educ. Res., 45. 1975

Krzanoswsky, W.J *The performance of Fisher's linear discriminant function under non-optimal conditions*. Tecnometrics, 19 p 191-200. 1977.

Lachenbruch, P.A. *Discriminant Analysis*. Hafner. New York. 1975.

Lachenbruch, P.A. and Goldstein, M. *Discriminant Analysis*. Biometrics 35 New York. 1979.

Lachenbruch, P.A. and Mickey, M.R. *Estimation of error rates in discriminant analysis*. Technometrics 10 p. 1-11. 1968.

Lachenbruch, P.A *et al. Robustness of the linear and quadratic discriminant function to certain types of non-normality*. Commun. Stat, 1 p. 39-57. 1973.

Moore, D.H. *Evaluation of five discrimination procedures for binary variables*. J. Am. Stat. Assoc., 68, p. 339-404. 1973

Muñoz C., Manuel. *La pobreza en 13 ciudades colombianas- Medida a través de líneas de indigencia y de pobreza- Proyecto DANE-PNUD-UNICEF-DPN* Boletín de Estadística, DANE. No. 429. Bogotá. Dic/ 1988.

O'Hara et all. *A comparison of discriminant function and maximum likelihood estimates of logistic coefficients for categorical-scaled data*. J. Stat Comput. Simul 14, 169-178. 1982

Rodgers, John L. and Rodgers Joan R. *Measuring the intensity of Poverty Among. Subpopulations. Application to the United States*. The Journal of Humanan Resources XXVI-2.

Sarmiento, A Libardo. *El paquete social del neoliberalismo en Revista Foro*. No. 20, mayo de 1993. Ediciones Foro Nacional por Colombia. Bogotá.

SAS. Technical Report P-200. *Logistic Procedures*. Release 6.04. USA. 1990

Seber, G.A.F. *Multivariate Observations*. New York. 1984

Titterington, D.M. et al. *Comparison of discrimination techniques applied to a complex data set of head injured patients* J. R. Stat. Soc. A, 144 p. 144-175. 1981

Urrutia, Miguel. *Distribución del Ingreso y la pobreza en Colombia: Evolución reciente* en Revista DEl Banco de la República. Bogotá, agosto de 1993.

World Bank, *Poverty, World Development indicators*, Report 1990, Oxford.