





facultad de ciencias  
sociales e historia



## **Factores determinantes de la pobreza en base a un modelo logístico**

Berta Teitelboim  
Académica de la Escuela de Sociología  
Instituto de Investigación en Ciencias Sociales, Universidad Diego Portales  
Berta.teitelboim@udp.cl

## RESUMEN

Se tenía como antecedente que la medición de pobreza en Chile se basaba exclusivamente en el ingreso monetario de los hogares, lo que significaba que en la declaración de ingresos de los hogares, había una fuerte subdeclaración y para superar este problema, se debe realizar un ajuste de ingresos en base a las Cuentas Nacionales, lo que implica que sólo se puede realizar esta medición cuando las muestras son representativas de toda la población.

El objetivo general fue obtener un modelo que permitiera identificar en los hogares las variables que son determinantes de la pobreza de éstos, basándose en la siguiente hipótesis: ser pobre o no, está determinado por un conjunto de características estructurales del hogar, vinculadas a las siguientes dimensiones: Geográficas, Demográficas, Mercado Laboral, Educación, Vivienda, Ingresos, Patrimonio. Otro objetivo era obtener un modelo de estimación de pobreza estructural, que pueda ser utilizado en investigaciones que usen muestras representativas de la población, en los distintos niveles geográficos tales como manzana, agrupación vecinal, comuna, provincia, región, etc.

Las regresiones categóricas pueden ayudar a seleccionar las mejores variables para identificar a pobres y no pobres. Se encontraron dos modelos probabilísticos, uno para la zona urbana y otro rural, los cuales fueron validados con datos externos y mostraron un alto grado de predicción.

## **PALABRAS CLAVES**

***Pobreza, Regresión Logística, Datos Categóricos, Ciencias Sociales***

## **INDICE**

### **1.- Introducción**

### **2.- Objetivos**

- 2.1 Objetivos generales
- 2.2 Objetivos específicos
- 2.3 Hipótesis de trabajo

### **3.- Marco Teórico Conceptual de Referencia**

- 3.1 Marco teórico
- 3.2 Método de análisis
- 3.3 Modelo de regresión logística
- 3.4 Modelo de regresión Logística Múltiple

### **4.- Metodología**

- 4.1 Información primaria a utilizar
- 4.2 Universo y unidades de muestreo
- 4.3 Variables del estudio
- 4.4 Plan general de análisis

### **5.- Resultados**

- 5.1 Análisis de las variables
- 5.2 Modelo de regresión logística completo
- 5.3 Modelo en el área urbana
- 5.4 Modelo en el área rural

### **6.- Conclusiones**

### **7.- Revisión Bibliográfica**

### **8.- Anexos**

- Anexo N°1 Medición de la pobreza
- Anexo N°2 Ficha técnica de la Encuesta CASEN 2003
- Anexo N°3 Definiciones
- Anexo N°4 Tablas

## 1.- INTRODUCCIÓN:

En Chile, la mayoría de los programas sociales (educación, salud, subsidios, etc.) se asignan o se operan a nivel comunal y la asignación de recursos a éstos se efectúa de acuerdo a los niveles de pobreza que muestran las comunas. Uno de los principales objetivos de la política social en Chile es la focalización de los recursos en los sectores más pobres de la población.

La asignación de estos programas es estimada a través de diversas metodologías (Indicador de Desarrollo Humano Comunal del PNUD<sup>1</sup>, SINIM<sup>2</sup>, etc.), ya que no existe información de pobreza a nivel de todas las comunas del país (sólo para algunas comunas que están autorepresentadas en la Encuesta CASEN<sup>3</sup>) que permita ordenar o priorizar a los municipios según esta variable en forma precisa.

Poder determinar para todas las comunas de Chile y su desagregación territorial (zonas, distritos censales y manzana) su situación de pobreza, permitiría identificar a los sectores más pobres de la población, y así asignar o diseñar programas sociales dirigidos a éstos grupos que históricamente han quedado relativamente más marginados de la política social por problemas de falta de información o acceso al sistema de protección social.

Este trabajo tiene gran importancia en la política social desarrollada por el Gobierno, la que tiene como objetivo focalizar en los más pobres una parte importante de los recursos del Estado, debido a la escasez de éstos, y poder diseñar adecuadamente políticas que tengan un fuerte impacto redistributivo.

El modelo estadístico desarrollado en esta investigación permite determinar los niveles de pobreza a nivel regional, provincial, comunal, por zona, distrito y finalmente por manzana.

Para clasificar la situación que presenta un hogar, se buscó perfilarlos de acuerdo a un conjunto de características referidas al Jefe de Hogar y que tengan relación con los niveles de pobreza.

Esta investigación de tesis desarrolló un modelo probabilístico, utilizando "Regresión Logística", que permite determinar, en base a la información censal o encuestas, los niveles de pobreza a nivel regional, provincial, comunal, por zona, distrito y finalmente por manzana, lo que implicará poder focalizar los programas sociales territorialmente con gran precisión. El modelo propuesto combina información social censal con georeferenciación y modelación probabilística.

En base al resultado obtenido, se puede calcular la probabilidad de ser pobre para cada una de las familias entrevistadas por el censo realizado en el año 2002 y así obtener, en base a variables estructurales y no coyunturales, una medición alternativa de pobreza e indigencia. Esta medición se realizó para un conjunto de variables que tienen estabilidad en el tiempo<sup>4</sup> y que además pueden ser verificadas, lo que no sucede con los ingresos del hogar, variable que es muy difícil de observar, mostrando los estudios que la subdeclaración llega al 47% en algunos casos<sup>5</sup>.

## 2.- OBJETIVOS E HIPOTESIS

### 2.1 Objetivos Generales:

- Obtener un modelo que permita identificar las variables que son determinantes de la pobreza de los hogares.
- Tener un modelo de estimación de pobreza estructural, que sea usado por encuestas de muestras representativas de la población.
- Elaborar un modelo que permita focalizar social y territorialmente los programas sociales.

### 2.2 Objetivos Específicos:

- Obtener estimaciones de pobreza a nivel regional, comunal, por distrito y zona censal.
- Obtener estimaciones de pobreza e indigencia a nivel de manzanas y de hogares.
- Poder en el futuro, en base a la información censal y al modelo obtenido, georeferenciar cada una de las variables e identificar geográficamente los bolsones de pobreza existentes en el país, para poder diseñar políticas y programas sociales adecuados a realidades específicas de las familias.

### 2.3 Hipótesis de Trabajo:

La pobreza corresponde a múltiples factores, los que, según estudios recientes, se ven reflejados por variables vinculadas al hogar y su jefatura (Larrañaga, 2002, Guerrero 2004, y otros autores mencionados en la bibliografía). El ser pobre o no, está determinado por un conjunto de características estructurales del hogar, vinculadas a las siguientes dimensiones: Geográficas, Demográficas, Mercado Laboral, Educación, Vivienda, Ingresos, Patrimonio. Éstas se modelarán en base a un modelo de regresión logística, el cual necesita validar muy pocos supuestos acerca de cada una de las variables y además permite la regresión no lineal utilizando variables categóricas ordinales (Hosmer y Lemeshow (2000)).

La variable dependiente será el nivel de pobreza, pudiendo tomar dos valores: pobre y no pobre. Las variables independientes consideran las siguientes dimensiones:

<b>Geográficas:</b>	Región y Zona
<b>Demográficas:</b>	Sexo del Jefe(a) de Hogar (JH), Nº de personas en el hogar, Nº de menores de 14 años en el hogar
<b>Mercado Laboral:</b>	Actividad del JH (Ocupado, Desocupado, Inactivo), Nº de ocupados en el hogar.
<b>Educación:</b>	Nivel Educativo del JH y Años de Escolaridad del JH
<b>Vivienda<sup>6</sup>:</b>	Calidad de la Vivienda, Acceso al Agua Potable, Acceso al alcantarillado, Acceso a la energía Eléctrica, Hacinamiento

**Ingresos:** N° de perceptores de ingresos en el hogar<sup>7</sup>

**Patrimonio** Tenencia de: Teléfono, Computador, Microondas, Videograbador, Lavadora, Refrigerador, Internet y Celular

### 2.3.1 Hipótesis Específicas:

- La inserción de los Jefes de Hogar en el mercado laboral tiene una fuerte incidencia en la probabilidad de que el hogar sea pobre o no.
- La edad del Jefe de Hogar tiene incidencia en la probabilidad de que el hogar sea pobre o no, a menor edad mayor probabilidad de ser pobre.
- La escolaridad del Jefe de Hogar tiene incidencia en la probabilidad de que el hogar sea pobre o no.
- A mayor número de personas en el hogar, mayor es la probabilidad de pobreza.
- El género del Jefe de Hogar tiene incidencia en la probabilidad de que el hogar sea pobre.
- El acceso a bienes y servicios básicos incide en menores niveles de pobreza.
- El ser miembro de un pueblo originario aumenta la posibilidad de ser pobre.

## 3. MARCO TEÓRICO CONCEPTUAL DE REFERENCIA:

### 3.1 Marco teórico

La pobreza tiene muchas caras pero se puede resumir simplemente como la falta de recursos, va más allá de la carencia de bienes económicos, también involucra la falta de oportunidades para el desarrollo de una vida decente para mantener y conservar la dignidad, la autoestima y el respeto de otros, trascendiendo de los bienes materiales<sup>8</sup>.

Se caracteriza por ser un fenómeno especialmente económico con dimensiones sociales, políticas y culturales (véase Moon 1991). La pobreza está asociada a la escasa participación de las personas en los distintos ámbitos de la vida del país y se expresa en el subconsumo en los hogares. Las personas que se encuentran en esta situación se ven obligadas a elegir la satisfacción de unas necesidades sacrificando otras igualmente apremiantes para ellos.

Esta concepción de la pobreza sostiene que el foco tiene que ser la situación de bienestar de los pobres como tales, sin importar los factores que lo afecten. Esta medición se realiza simplemente contando el número de pobres; así la pobreza se expresa como la relación entre el número de pobres y la población total de la comunidad, o tasa de incidencia de la pobreza (Sen 1992). Los enfoques más utilizados para medir la pobreza son el de la desigualdad (Miller S.M. 1967) y el biológico (CEPAL 1990).

El enfoque de la desigualdad consiste en medirla en base a la privación relativa de la población, en términos de su distribución de ingresos<sup>9</sup>, y supone concebir la pobreza como un problema de desigualdad. Existen una serie de medidas propuestas para utilizar este método, dentro de ellos el más utilizado es el de Atkinson (1970).

El enfoque biológico es el que se utiliza en América Latina para medir pobreza y define a las familias en situación de "pobreza como aquellas cuyos ingresos totales resultan insuficientes para cubrir las necesidades básicas relacionadas con el mantenimiento de la simple eficiencia física". Este enfoque ha sido intensamente atacado por varios autores (véase Townsend, y Rein 1974).

Según esta metodología, para delimitar la pobreza se define un conjunto de necesidades básicas (alimentación, acceso a la educación, salud y vivienda, etc), para cada una de éstas, se especifican características mínimas de sus satisfactores, normas por debajo de las cuales se presenta una situación de insatisfacción. Los umbrales definidos tienen validez temporal y restringida, motivo por el cual deben revisarse periódicamente<sup>10</sup>.

De acuerdo con este método, a un individuo se le considera pobre si su nivel de ingreso se sitúa por debajo de un nivel mínimo que le permita satisfacer sus necesidades básicas; e indigente, si éste no le permite satisfacer sus necesidades alimentarias. Estos mínimos se denominan "línea de pobreza" y "línea de indigencia", respectivamente. Así, estas líneas constituyen el límite entre quienes son pobres y quienes no lo son, y entre quienes son indigentes y quienes no lo son<sup>11</sup>. Es el método utilizado por la CEPAL.

Entre las críticas que se realizan a este enfoque, está que los requerimientos varían según el clima, la actividad que se realiza, y características demográficas tales como edad y sexo. Además los requerimientos nutricionales son difíciles de precisar, aún más los no alimentarios, por lo que generalmente su medición es arbitraria. Las canastas básicas resultantes tienen un costo excesivamente bajo (véase Stigler 1945). No obstante, aun con estos cuestionamientos, este método presenta gran utilidad y algunos de los planteamientos se pueden ir superando, como puede ser a través del cálculo de los requerimientos nutricionales considerando las características climáticas de cada país, la estructura etaria y laboral de su población, así como también se ha avanzado en construir canastas de consumo que consideran los hábitos de la población de referencia (véase CEPAL; Medición y análisis de la pobreza, 2002).

En América Latina, en los últimos años han predominado tres métodos de medición de la pobreza, los cuales consideran solamente las necesidades económicas de las personas (enfoque biológico):

- (i) El enfoque del ingreso o Línea de Pobreza (LP)
- (ii) El de las Necesidades Básicas Insatisfechas (NBI)
- (iii) El Método Integrado de medición de la pobreza

A nivel mundial existen otras mediciones que permiten su comparación, como la propuesta por el Banco Mundial. Esta línea de pobreza se basa en el consumo de bienes y servicios:

- En los países muy carentes de recursos una línea de pobreza fijada en 1 dólar por persona diarios (US\$ 30 mensuales)
- Se sugiere una línea de pobreza de 2 dólares diarios para América Latina y el Caribe.
- Para Europa oriental y las repúblicas de la ex Unión Soviética se ha utilizado una línea de pobreza de 4 dólares diarios.
- Para la comparación entre los países industrializados, se ha utilizado una línea de pobreza que corresponde a la de los Estados Unidos, de 14,40 dólares diarios por persona.
- La Unión Europea ha sugerido que para determinar la línea de pobreza de esos países se utilice la mitad de la mediana del ingreso personal disponible



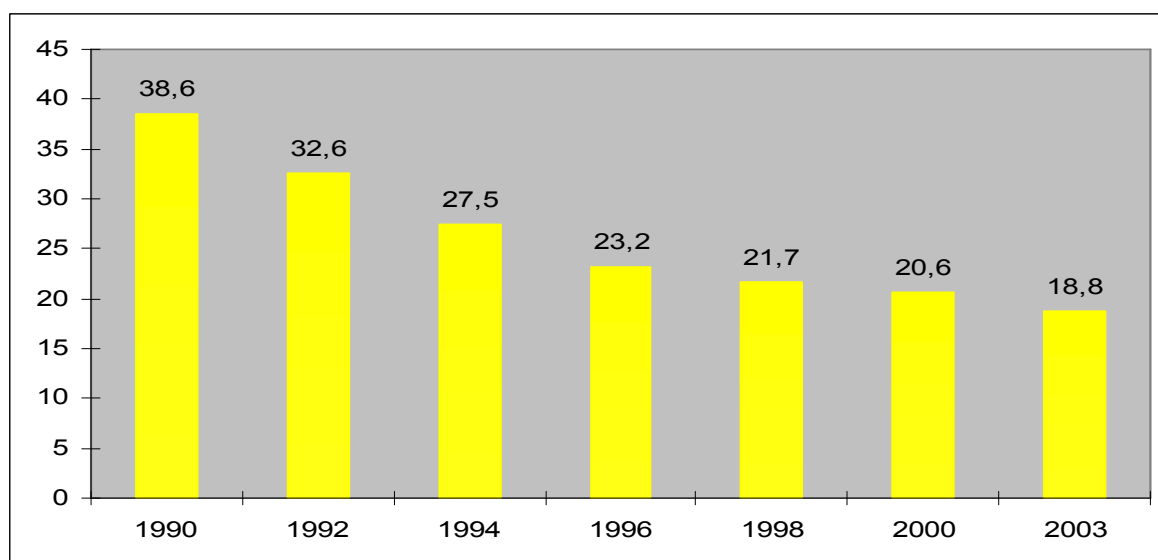
### 3.1.1 La Pobreza en Chile:

El concepto de pobreza utilizado en Chile corresponde al definido por Naciones Unidas<sup>12</sup>, que la precisa como: "Una situación que impide al individuo o a la familia satisfacer una o más necesidades básicas y participar plenamente en la vida social". El Ministerio de Planificación (MIDEPLAN)<sup>13</sup> es la institución en Chile encargada de estimar la pobreza, y lo realiza a través del "método del ingreso" o "del costo de las necesidades básicas", que como se mencionó anteriormente es el método utilizado internacionalmente y propuesto por CEPAL<sup>14</sup>, el cual tiene como desventaja que es muy coyuntural al depender solamente de los ingresos corrientes del mes pasado y no de la acumulación de bienes. Mide la pobreza en base a los ingresos monetarios percibidos por la población<sup>15</sup>. Desde el año 1987 MIDEPLAN realiza la encuesta CASEN<sup>16</sup>, cuyo principal objetivo entre otros, es medir y analizar la evolución de la pobreza de los hogares.

Se estiman dos líneas de pobreza e indigencia, una correspondiente a las zonas urbanas y la segunda a las zonas rurales<sup>17</sup>. Las diferencias entre ambas se derivan del supuesto de que en las zonas rurales hay un menor costo de los alimentos y del gasto no alimentario, según estudios previos realizados por CEPAL. En definitiva, la línea de indigencia rural representa un 77% del valor de la urbana. La "línea de pobreza" en las zonas urbanas se calcula como el doble del valor de la línea de indigencia urbana, en tanto que la de las zonas rurales aquella asciende a 1.75 veces el valor de la línea de indigencia rural. Lo cual significa que la cuantía de recursos monetarios que necesitan los hogares para superar esta situación según área geográfica es distinta, el valor de la canasta de la zona rural representa el 67,4% del valor de la canasta de la zona urbana).

Los datos provenientes de la Encuesta CASEN, muestran una importante disminución en los niveles de pobreza e indigencia en los últimos 13 años, tanto en el área rural como en la urbana, como se puede observar en los siguientes gráficos:

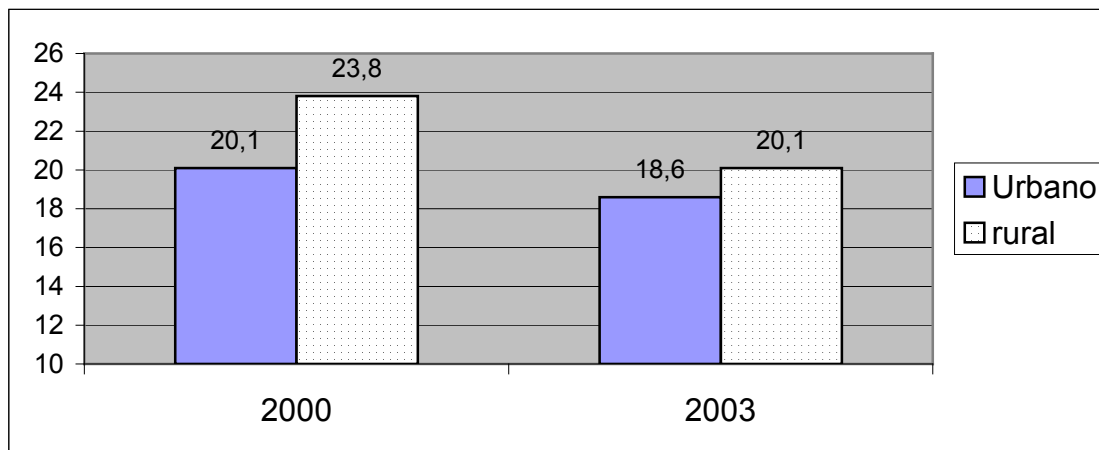
Gráfico N°1: Evolución de la Pobreza en Chile



Fuente: MIDEPLAN; Volumen 1: "Pobreza, Distribución del Ingreso e Impacto Distributivo del Gasto Social"; Santiago; agosto 2004.

A fines del año 2003 la situación de pobreza total del país baja a un 18.8%, correspondiente a 2.907.700 personas aproximadamente, de las cuales un 4,7% son indigentes (728.100 personas) y 14,1% son pobres no indigentes (2.179.600)<sup>18</sup>.

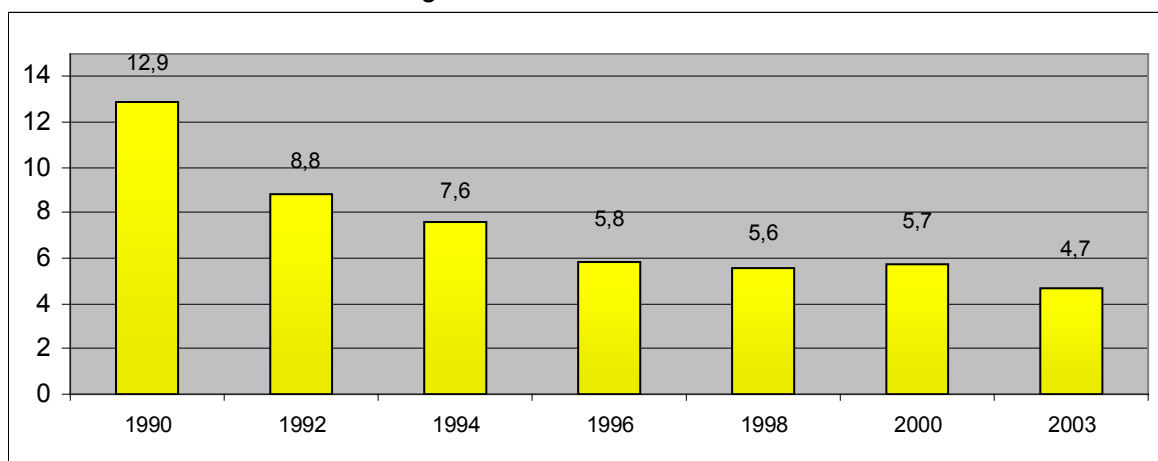
Gráfico N°2: Pobreza según área Geográfica: 2000-2003



Fuente: MIDEPLAN; Volumen 1: "Pobreza, Distribución del Ingreso e Impacto Distributivo del Gasto Social"; Santiago; agosto 2004.

En las zonas urbanas, la pobreza es menor que en las zonas rurales y en el periodo 2000-2003 disminuyó en ambas áreas. La pobreza rural registró un importante retroceso, 86 mil 200 personas salieron de esta condición, lo que corresponde a una reducción en estos tres años del 23,8 al 20,1%. La misma senda siguió la indigencia, que bajó en 1,0 punto porcentual.

Gráfico N°3: Evolución de la Indigencia en Chile



Fuente: MIDEPLAN; Volumen 1: "Pobreza, Distribución del Ingreso e Impacto Distributivo del Gasto Social"; Santiago; agosto 2004.

Como se observa en los gráficos anteriores, la evolución de la pobreza y la indigencia muestra una importante disminución a partir de la década de los '90. Desde 1996, el ritmo de esta baja en las tasas disminuye, especialmente en la situación de indigencia, o también llamada "pobreza dura", donde se mantienen los mismos niveles entre el año 1996 y el 2000. Esta tendencia presenta un considerable cambio en el año 2003, pasando de 5,7% a 4,7%. Según los primeros análisis esta baja sería producto de las políticas sociales focalizadas en los sectores más pobres, viéndose estas representadas en estos grupos por el Sistema Chile Solidario, que es un programa de atención integral a las familias más pobres que se implementó en el año 2002. Esta cifra muestra la importancia de poder producir información que permita focalizar cada vez mejor cada uno de los programas dirigidos a la superación de la pobreza.

### 3.2 Método de Análisis

Para encontrar los determinantes de la pobreza en los hogares chilenos, se utilizará un modelo que permita primero caracterizar y luego clasificar los hogares que requieren ayuda social en forma más prioritaria, para lo cual se utilizará un modelo de regresión.

En los casos en que el marco conceptual lleva a la especificación de un modelo en el cual existe una variable explicada (nivel de pobreza) por un conjunto de variables explicativas ( $X_1, X_2, X_3, \dots, X_k$ ), la técnica estadística adecuada es el análisis de regresión. En este caso como el objetivo es determinar tanto las características de la pobreza como la probabilidad de ser pobre, se usará regresión logística que permite predecir probabilidades.

La proposición teórica que lleva a emplear esta herramienta supone que la variabilidad de la variable dependiente se puede explicar a través de las variables explicativas. El modelo estadístico correspondiente agrega que, además, opera como variable independiente un factor aleatorio no observable ( $x$ ), llamado error aleatorio.

El ajuste del modelo estadístico especifica una serie de condiciones que debe satisfacer el error aleatorio y que permiten estimar, no sólo si el modelo ajusta adecuadamente, para lo cual se tienen pruebas de hipótesis e índices de bondad, sino también la significación y magnitud de los efectos de las variables explicativas.

La medición de los impactos de las variables explicativas requiere que no haya relación entre ellas o que ésta sea baja, cuando esto no ocurre se presentan dificultades para estimar la magnitud real de los efectos debido a que el modelo es incapaz de separar cuál corresponde a cada variable. Los procedimientos de estimación se complican si se detecta que no se cumplen los supuestos relativos al término de error. Por lo cual se deben realizar las pruebas estadísticas para verificar los supuestos.

El planteamiento clásico del modelo de regresión supone que todas las variables son métricas (medidas en escala de intervalo o de razón). Esta condición se convierte en una fuerte barrera para aplicar esta técnica estadística a los problemas típicos de las ciencias sociales. Sin embargo, esta limitación fue superada al introducir, *variables explicativas* dicotómicas (dummies) que dieran cuenta de la presencia o ausencia de un evento particular. Es así como la regresión se ha convertido en una herramienta de gran utilidad para las ciencias sociales. Este modelo aún mantiene la restricción a la variable dependiente, la cual implica la continuidad de ésta y que sus valores pueden variar entre menos infinito e infinito.

Estos problemas se superaron aplicando transformación logarítmicas a la variable dependiente (Agresti A., 1990)<sup>19</sup> y utilizando un modelo denominado de regresión logística que permite que la variable dependiente sea cualitativa.

En las ciencias sociales se recurre cada vez más a utilizar técnicas de análisis de variables múltiples para analizar los distintos fenómenos. En el ajuste de este tipo de modelos tienden a predominar los métodos de estimación mínimo cuadrático y de máxima verosimilitud. Si bien hay razones técnicas para preferir uno u otro dependiendo del problema de que se trate.

Finalmente, los desarrollos estadísticos recientes, fuertemente condicionados por las demandas planteadas por las ciencias sociales, tienden a abandonar la naturaleza determinística de la experiencia y a reemplazarla por una concepción que reconoce el carácter aleatorio de la misma y a reconocer el papel activo del sujeto en la construcción del objeto”<sup>20</sup>.

### 3.3 Modelo de Regresión Logística <sup>21</sup>:

Es un modelo no lineal, los datos no se ajustan a una línea recta, a las variables explicativas no se les exige una distribución determinada. Permite construir modelos donde las variables dependientes pueden ser cuantitativas o cualitativas, además éstas últimas pueden ser dicotómicas o politómicas, y dentro de éstas las variables pueden ser ordinales o nominales. Las variables explicativas pueden ser cualitativas y/o cuantitativas.

La función logística es una curva sigmoidea en forma de letra S, utilizando los logaritmos es posible linealizar el modelo.

Si la variable dependiente es dicotómica, se utiliza para predecir la probabilidad estimada de que la variable dependiente "Y" presente uno de los valores posibles (1=sí o 0=no) en función de los diferentes valores que adoptan el conjunto de variables independientes "X<sub>i</sub>"<sup>22</sup>.

Las variables predictoras "X<sub>i</sub>" pueden ser categóricas, medidas a nivel nominal u ordinal, y de intervalo o razón.

Jovell plantea que los objetivos del modelo logístico son "determinar la existencia o ausencia de relación entre una o más variables independientes y la variable dependiente; medir la magnitud de dicha relación y estimar o predecir la probabilidad de que se produzca un suceso Y=1 en función de los valores que adopten las variables independientes "X<sub>i</sub>"<sup>23</sup>.

En la mayoría de los casos las regresiones logísticas además pueden utilizarse para medir el rendimiento de los parámetros alternativos de determinación de beneficiarios de programas y transferencias sociales. Se trata de un tema importante, ya que las autoridades suelen utilizar variables representativas para evaluar el nivel de renta o de consumo de un hogar o individuo con el objeto de determinar si les corresponde o no ser incluidos en dichos programas. Las regresiones logísticas pueden ayudar a seleccionar las mejores variables representativas para identificar a pobres y no pobres, o bien más ampliamente para seleccionar beneficiarios<sup>24</sup>.

#### 3.3.1 Definición:

Sea Y una variable dependiente binaria que toma dos valores posibles (0 y 1).

Sean X<sub>1</sub>,.....,X<sub>k</sub> un conjunto de variables independientes observadas con el fin de explicar y/o predecir el valor de Y.

El objetivo es determinar:

$P[Y=1/ X_1,.....,X_k]$ , donde P indica probabilidad

Así  $P[Y=0/ X_1,.....,X_k] = 1 - P[Y=1/ X_1,.....,X_k]$ .

Se construye un modelo :

$P[Y=1/ X_1,.....,X_k] = p(X_1,.....,X_k ; \beta)$

Donde  $p(X_1,.....,X_k; \beta)$  es una función que recibe el nombre de función de enlace (función de probabilidad) cuyo valor depende de un vector de parámetros

$\beta = (\beta_1,.....,\beta_k)'$ .

### 3.3.2 Función de Verosimilitud

Con el fin de estimar  $\beta$  y analizar el comportamiento modelo, observamos una muestra aleatoria simple de tamaño  $n$  dada por  $\{(X_i', Y_i; i=1, \dots, n)\}$  donde  $X_i = (X_{i1}, \dots, X_{ik})$ , es el valor de las variables independientes e  $Y_i = \{0, 1\}$  es el valor observado de  $Y$  en el  $i$ -ésimo elemento de la muestra.

$Y_i/(X_{i1}, \dots, X_{ik})$  se distribuye Binomial  $(1, p(Y=1/ X_{i1}, \dots, X_{ik}; \beta))$ .

La función de verosimilitud es:

$$L(\beta/(x_1, y_1, \dots, x_n, Y_n)) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i} \quad (1)$$

Donde  $p_i = p(x_i', \beta) = p(x_{i1}, \dots, x_{ik}; \beta)$  con  $i=1, \dots, n$

### 3.4 Modelo de regresión logística múltiple

El modelo matemático se construye en base a probabilidades, las cuales se obtienen considerando la probabilidad de que ocurra un suceso determinado  $P(Y)$  en relación con la dependencia de que dicha probabilidad no ocurra  $1 - P(Y)$ .

La probabilidad proporciona predicciones consistentes y fáciles de los resultados en términos del "Odds" del evento  $Y=1$ .

$$\text{Odds}(Y=1) = \frac{P(y)}{1 - p(y)} \quad (2)$$

El modelo de regresión logística múltiple, relaciona la probabilidad de que ocurra un determinado suceso denotado por el vector  $X' = (X_1, \dots, X_k)$  con probabilidad condicional  $P(Y=1| X)$  en función de  $k$  variables independientes que pueden ser cuantitativas, cualitativas o ambas según sea el tipo de diseño de estudio.

El modelo logístico múltiple es:

$$L = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (3)$$

$$\text{O también: } p_i = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}} = \frac{1}{1 + e^{-z}} \quad (4)$$

Donde:

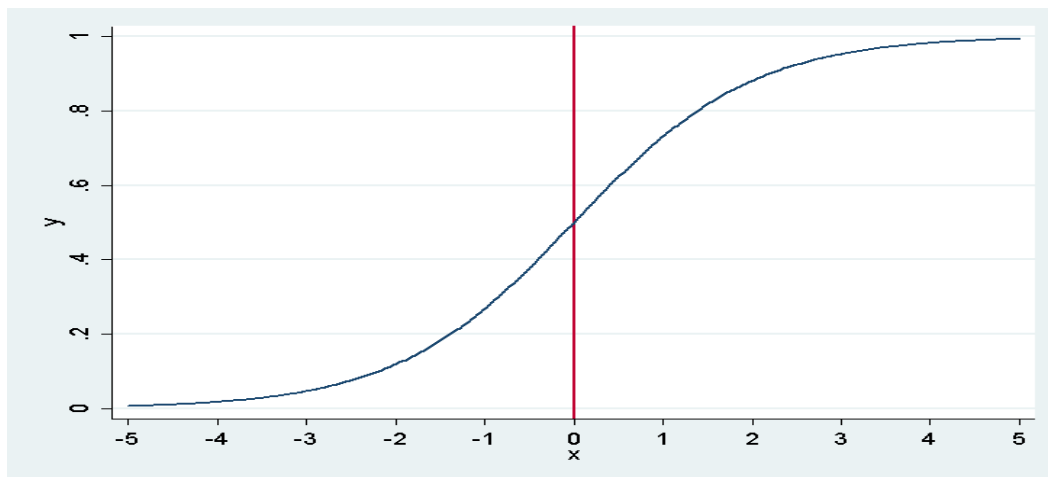
- $\ln$  es el logaritmo natural de "Odds"<sup>25</sup> también denominado "Logit" o "L"
- $\beta_0, \beta_1, \dots, \beta_k$  son constantes
- $X$  una variable explicativa que puede ser continua o discreta.

Como los coeficientes del modelo logístico no tienen restricciones éstos son fácilmente interpretables en términos de independencia o asociación entre las variables.

Gráficamente la función es simétrica

- Continua y creciente sobre el intervalo 0 y 1
- Sigue una curva sigmoidea

Gráfico n° 4 Gráfico del Modelo de Regresión Logística



### 3.4.1 Estimación de los Parámetros

Sea una muestra de  $n$  observaciones independientes definido por  $(X_{i1}, X_{i2}, \dots, X_{ik}, Y_i)$ ,  $i=1, \dots, n$ ; se elige el vector  $\beta' = (\beta_0, \beta_1, \dots, \beta_k)$  el método más usado es el de verosimilitud, con  $k+1$  ecuaciones de verosimilitud que se obtienen derivando el ln de la función de verosimilitud respecto a  $k+1$  coeficientes.

Las ecuaciones de verosimilitud son:

$$\sum_{i=1}^n (y_i - p_i) = 0$$

$$\sum_{i=1}^n X_{ij} (y_i - p_i) = 0, j = 1, \dots, k; i = 1, \dots, n$$

Mediante calculo diferencial se encuentran las soluciones a este conjunto de ecuaciones, actualmente existen software estadísticos para estimar los parámetros.

### 3.4.2 Pruebas de Significación

Una vez estimado los coeficientes del modelo, se tiene que verificar si el modelo predice de manera adecuada a la variable dependiente. Para evaluar la bondad del modelo se utiliza el logaritmo del cociente de verosimilitud y la prueba de Hosmer-Lemeshow.

#### 3.4.2.1 Modelo

Se procede de la siguiente manera

- 1.- Se prueba la hipótesis de que el modelo encontrado es el que mejor se ajusta a través de la razón de verosimilitud<sup>26</sup>.

$$D = -2 \ln \frac{\text{verosimilitud del modelo sin la variable (modelo analizado)}}{\text{verosimilitud del modelo con la variable (modelo saturado)}}$$

La diferencia entre estos dos valores de  $-2nL$  se llama **Devianza** y prueba si la o las variables  $X_i$  son significativas.

$$H_0 : \beta_0 = \beta_1 = \dots = \beta_k = 0$$

$$H_1 : \text{Algún } \beta_i \text{ distinto de } 0.$$

Estadístico de prueba  $D \sim \chi^2$  con  $n-k-1$  grados de libertad.

Región de Rechazo: si  $D > \chi^2_{\alpha, (n-k-1)}$  si se rechaza  $H_0$ , significa que al menos uno

de los coeficientes es diferente de cero y la variable correspondiente es significativa con respecto a la probabilidad de que ocurra o no el suceso en estudio<sup>27</sup>.

- 2.- Cuando el número de variables del modelo es grande y/o tiene variables independientes cuantitativas, se debe utilizar la prueba de Hosmer-Lemeshow, como consecuencia de la gran cantidad de variables, el número de pautas de variación existente entre ellas es tan elevado que puede invalidar la utilidad de los estadísticos de bondad de ajuste clásicos.

Evalúa la bondad del modelo construyendo una tabla de contingencia, divide la muestra en 10 grupos a partir de los deciles de las probabilidades estimadas (a los 10 grupos resultantes se les denomina deciles de riesgo). En cada decil de riesgo se calcula el número de casos que pertenecen a cada categoría de la variable dependiente (número observado) y el número de casos que el modelo pronostica que pertenecen a cada categoría de la variable dependiente (número esperado). A continuación se compara el número de casos observado con el número de casos esperado (esta comparación se realiza en cada una de las 20 casillas definidas por la combinación de las 2 categorías de la variable dependiente con los 10 deciles de riesgo).

Este estadístico permite contrastar la hipótesis nula de igualdad de distribuciones, es decir, la hipótesis de que la variable dependiente se distribuye de la misma manera en los 10 deciles de riesgo o, que no existen diferencias entre las frecuencias observadas y las esperadas. El estadístico de contraste tiene siempre k grados de libertad<sup>28</sup>.

Las hipótesis que se contrastan son:

$H_0$  : El modelo es adecuado

$H_1$  : El modelo ajustado no es adecuado

Decisión si el estadístico de prueba es mayor o igual que  $X^2_{\alpha, (n^{\circ} \text{ de grupos}-2)}$ , se rechaza  $H_0$  y se concluye que el modelo no es adecuado con un nivel de significación  $\alpha$ .

### 3.4.2.2 Pruebas de coeficientes

Una vez encontrado el mejor conjunto de variables explicativas que predicen la variable dependiente Y, se debe evaluar cada coeficiente para determinar cuál o cuáles ingresan al modelo, este proceso se realiza mediante el estadístico Wald.

La hipótesis que se plantea es la siguiente:

$H_0$  :  $\beta_i = 0$

$H_1$  :  $\beta_i$  no es igual a 0

Para contrastar la hipótesis señalada se usa el estadístico de Wald:

$$WALD = \frac{\beta_i^2}{S^2_{\beta_i}} \sim X^2_{\alpha, 1} \quad (5)$$

en el caso de la regresión logística multivariada corresponde a un vector donde cada celda es la división entre el coeficiente  $\beta_i$  y el error estándar de éste. Donde  $S_{\beta_i}$ , es el error estándar del coeficiente de regresión logística muestral y k es el número de variables independientes.

Se formula la hipótesis

$H_0$  : La variable independiente no influye sobre  $p_i$ .

$H_1$  : La variable independiente influye sobre  $p_i$ .

### 3.4.3 Interpretación de los Resultados

La interpretación de los resultados obtenidos se realiza a partir de los coeficientes del modelo. Para ello basta tener en cuenta que si el modelo ajustado es adecuado, entonces se dice que el modelo es significativo. Además, se debe analizar el grado de asociación estadística que existe en sus parámetros. Si:

$\beta_1$	$> 0$	el factor de riesgo será mayor que 1 y $p(X_1, X_2, \dots, X_k; \beta)$ aumentará
$\beta_1$	$< 0$	el factor de riesgo será menor que 1 y $p(X_1, X_2, \dots, X_k; \beta)$ disminuirá.
$\beta_1$	$= 0$	la variable $X_1$ no ejerce ningún efecto sobre la probabilidad de riesgo.

Si el modelo de Regresión Logística es significativo y una de variables independientes es dicotómica con valores de 0 y 1, el número  $e^{\beta_i}$  es el OR, denominado factor de riesgo o protección que implica un aumento unitario de la variable independiente.

En el caso de una variable cuantitativa,  $e^{\beta_i}$  es el número de veces que aumenta la "chance" de que ocurra el suceso, en este caso de ser pobre, por cada unidad de aumento de la variable independiente.



## 4.- METODOLOGÍA

### 4.1 Información Primaria a Utilizar

La Encuesta CASEN tiene una gran cantidad de información, que permite construir un modelo para determinar la probabilidad de ser pobre, y tiene como variable dependiente los niveles de pobreza, por lo cual se propone utilizar un modelo de regresión logística que determine esta probabilidad.

Las variables utilizadas en el modelo deben también existir en el Censo de Población, para poder determinar con las variables que contiene, niveles de pobreza en las comunas, distritos, zonas censales y manzanas. La construcción de los niveles de pobreza de estos últimos, sólo será propuesta metodológicamente, para que pueda ser utilizada por instituciones como los ministerios de Planificación (MIDEPLAN), de Salud y Educación, las secretarías regionales, las gobernaciones y los municipios.

### 4.2 Universo, Muestra y Unidades de muestreo

Se seleccionaron dos muestras —una para el área urbana y otra para el área rural—, estratificadas, proporcionales y aleatorias, en que cada región del país se consideró como un estrato. Esto significó construir 13 muestras aleatorias regionales urbanas y 13 rurales basadas en la encuesta CASEN 2003, con un número total de 2000 casos para la zona urbana y 1000 para la zona rural.

Estas muestras fueron proporcionales a la representación de cada área y región del país, ya que el tamaño muestral de la Encuesta CASEN 2003 es muy grande (68.150 hogares), lo que podría significar que la mayoría de las variables del modelo podrían ser estadísticamente significativas sólo por el tamaño de la muestra. Además, hay regiones que están sobrerrepresentadas en la muestra por razones administrativas<sup>29</sup>.

Tabla 1 Número de hogares del país

Región	Zona Urbana			Zona Rural		
	Urbana	%	Muestra	Rural	%	Muestra
I	95.822	2,7	54	5.156	0,9	9
II	110.688	3,1	62	1.678	0,3	3
III	66.428	1,9	37	5.775	1,1	10
IV	117.172	3,3	66	39.984	7,3	72
V	400.522	11,3	225	31.600	5,7	57
VI	155.168	4,4	87	66.823	12,1	121
VII	158.785	4,5	89	83.688	15,2	152
VIII	426.433	12,0	239	96.674	17,5	175
IX	159.795	4,5	90	71.399	12,9	129
X	187.781	5,3	105	95.442	17,3	173
XI	19.810	0,6	11	6.464	1,2	12
XII	43.107	1,2	24	3.578	0,7	6
R.M.	1.619.601	45,5	910	43.465	7,9	79
<b>Total</b>	<b>3.561.112</b>	<b>100,0</b>	<b>2.000</b>	<b>551.726</b>	<b>100,0</b>	<b>1.000</b>

Los resultados obtenidos con esta muestra se validaron con el total de la información proveniente de la Encuesta CASEN 2003.

### 4.3 Variables del Estudio

La variable dependiente será el nivel de pobreza, pudiendo tomar dos valores: pobre y no pobre.

Las variables independientes consideran las siguientes dimensiones y variables independientes en los modelos explicativos:

Tabla 2 Variables

Dimensión	Variable	Etiqueta	Categoría
Geográfica	z	Zona	Cualitativa
Demográficas	sexojh	Sexo del jefe de hogar	Cualitativa
	edad	Edad del jefe de Hogar	Cuantitativa
	disca	¿Tiene alguna Discapacidad:?	Cualitativa
	etnia	Pertenece a alguna etnia	Cualitativa
	numper	Número de Personas en el hogar	Cuantitativa
Mercado Laboral	ocupado	¿Es ocupado el jefe de Hogar:?	Cualitativa
	inactivo	¿Es inactivo:?	Cualitativa
Educación	esc	Escolaridad del JH	Cuantitativa
Patrimonio	lavadora	Lavadora Automática	Cualitativa
	refri	Refrigerador	Cualitativa
	fono	Teléfono fijo	Cualitativa
	video	Videograbador	Cualitativa
	micro	Microondas	Cualitativa
	computa	Computador	Cualitativa
	calefont	Cálefont	Cualitativa
	celular	Teléfono Móvil	Cualitativa
	internet	internet	Cualitativa
	tv cable	TV Cable	Cualitativa
Vivienda	agua	Agua Potable	Cualitativa
	alcanta	Alcantarillado	Cualitativa
	vivienda	calidad vivienda	Cualitativa
	nhogar	hogares en la vivienda	Cualitativa
	hacina	Hacinamiento	Cualitativa

### 4.4 Plan General de Análisis

- a) Preparación de la base de datos  
Objetivo: Verificación y recodificación de los datos originales.
- b) Análisis Exploratorio  
Objetivo: Obtención de la información acerca del comportamiento de cada variable.  
Verificación de supuestos principales.
- c) Análisis de Regresión Logística con Respuesta Binaria  
Objetivo: A partir del primer modelo predictivo completo, se seleccionaron las variables con mejor capacidad predictiva, y se analizaron los efectos de la interacción de las variables.
- d) Validación del modelo obtenido  
Objetivo: Se validó el modelo obtenido en las dos submuestras, tomando la muestra de la Encuesta CASEN completa separada por área urbana y rural.

## 5. RESULTADOS

Se construyeron 2 matrices de datos, la matriz con información de la zona urbana contiene 1999 unidades muestrales y la rural, 1.000; cada unidad muestral corresponde a un hogar<sup>30</sup>; cada unidad de respuesta está representada por un vector de dimensión igual al número de variables, éstas corresponden a 23<sup>31</sup>.

Las variables son de tipo numéricas y nominales y corresponden a la información del Jefe de hogar y al hogar que habitan. La definición de medición de pobreza es por hogar y no por individuo. Si un hogar es pobre, todos los individuos de ese hogar enfrentan la misma condición, lo que implica que la unidad de análisis es el hogar.

### 5.1 Análisis de las variables

Una vez obtenidas las muestras urbana y rural, se recodificaron los datos, para facilitar el análisis e interpretación de los coeficientes del modelo de regresión logística. Las variables discretas quedaron como "0" cuando había ausencia del evento y 1 en caso de presencia. La variable "Participación en la Fuerza de Trabajo", que tenía 3 categorías (Ocupado, Desocupado e Inactivo), se transformó en dos variables ficticias, ocupados e Inactivos).

Luego se hizo un análisis descriptivo de las variables, en donde se concluyó lo siguiente:

#### Zona Urbana

- Los jefes de hogares pobres de la zona urbana pertenecen en mayor proporción a un pueblo originario en comparación a los jefes de hogares no pobres.
- Las viviendas de los hogares pobres urbanos tienen menor acceso al agua potable, alcantarillado, y una proporción mayor de hogares hacinados y viviendas de mala calidad.

En el caso de las variables continuas, se realizaron las pruebas de normalidad y luego se hizo el análisis de diferencia de medias según situación de pobreza.

### 5.2 Modelo de Regresión Logística Completo

$$\text{Logit } P(\text{pobreza}) = X' \hat{\beta}$$

con  $X' =$  (1, NUMPER, ESC, OCUPADO, INACTIVO, SEXOJH, EDADJH, DISCA, LAVADORA, REFRI, FONON, VIDEO, MICRO, COMPUTA, CALEFONT, CELULAR, INTERNET, TVCABLE, ETNIA, AGUA, ALCANTA, VIVIENDA, NHOGAR, HACINA)

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{23})$$

#### 5.2.1 Proceso de Selección de variables

A partir del modelo completo, la selección de variables se realizó en base a la probabilidad del coeficiente de regresión estimado por cada variable con un nivel de significación  $\alpha = 0.09$ . Este ajuste se efectuó mediante el software estadístico SPSS versión 13.

Como se mencionó en la sección anterior, se construyeron dos modelos: uno para la zona urbana y otro para la zona rural. Como se explicó en el capítulo 3, la metodología de medición de pobreza indica "Canastas Alimentarias" distintas para cada una de las zonas.

La canasta de la zona rural representa el 66% del valor de la canasta de la zona urbana, por lo cual mantener como variable ficticia a "Zona" implica incluir una variable "confusora" en el modelo, ya que al construir el modelo lo más probable es que el signo del coeficiente de la variable indique que si la persona vive en la zona urbana la probabilidad de ser pobre aumenta, y esta situación sólo se explica por la metodología utilizada en la medición de pobreza y no por qué efectivamente ocurra<sup>32</sup> así.

### 5.3 Modelo Área Urbana

#### 5.3.1 Parámetros Estimados del Modelo Completo

En la tabla 3 se presentan los parámetros del modelo completo, considerando 7 iteraciones que arrojó el software SPSS.

Se utilizó el estadístico de Wald para la prueba de hipótesis de cada uno de los coeficientes.

Las hipótesis son las siguientes:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \text{ no es igual a } 0$$

Tabla 3: Variables y coeficientes del Modelo Completo

	B	S.E.	Wald	df	Sig.	Exp(B)
Paso 1(a)						
NUMPER	,373	,042	79,594	1	,000	1,452
ESC	-,039	,023	2,882	1	,090	,962
OCUPADO	-2,411	,282	73,054	1	,000	,090
INACTIVO	-1,204	,312	14,846	1	,000	,300
SEXO	,118	,173	,464	1	,496	1,125
EDAD	-,045	,007	45,322	1	,000	,956
DISCA	-,582	,359	2,624	1	,105	,559
LAVADORA	-,359	,164	4,794	1	,029	,698
REFRI	-,092	,182	,255	1	,613	,912
TELEFONO	-,739	,184	16,134	1	,000	,477
VIDEO	-,653	,213	9,367	1	,002	,520
MICROONDAS	-,394	,221	3,184	1	,074	,674
COMPUTADOR	-,192	,287	,446	1	,504	,825
CALEFONT	-,639	,165	14,971	1	,000	,528
CELULAR	-,089	,158	,317	1	,573	,915
INTERNET	-,896	,686	1,708	1	,191	,408
TVCABLE	-,640	,261	5,984	1	,014	,527
ETNIA	,555	,303	3,359	1	,067	1,741
AGUA	-,021	,364	,003	1	,954	,979
ALCANTA	-,319	,302	1,117	1	,291	,727
VIVIENDA	,461	,436	1,119	1	,290	1,586
HACINA	,466	,306	2,311	1	,128	1,593
Constant	2,349	,668	12,379	1	,000	10,478

a Variable(s) en el paso 1: NUMPER, ESC, OCUPADO, INACTIVO, SEXO, EDAD, DISCA, LAVADORA, REFRI, TELEFONO, VIDEO, MICROONDAS, COMPUTADOR, CALEFONT, CELULAR, INTERNET, TVCABLE, ETNIA, AGUA, ALCANTA, VIVIENDA, HACINA.

Como se mencionó en el capítulo 3, este estadístico se distribuye  $\chi^2_{(1)}$ . El valor crítico utilizado es de 9%, por lo cual se rechaza la hipótesis nula si el valor del estadístico de WALD es mayor que el valor crítico.

### 5.3.2 Bondad de Ajuste del modelo con 12 variables

Se eligió el mejor modelo logístico a través del método denominado “Adelante Condicional”<sup>33</sup>, el cual correspondió a un modelo con 12 variables explicativas

Las Hipótesis son:

$$H_0 : \beta_0 = \beta_1 = \dots = \beta_k = 0$$

$$H_1 : \text{Algún } \beta_i \text{ distinto de } 0.$$

Este estadístico se distribuye como  $\chi^2_{(n-k-1)}$ , y se rechaza la hipótesis nula si el valor de  $-2\log$  Likelihood es mayor que  $\chi^2_{\alpha(n-k-1)}$ , donde  $\alpha$  es el nivel de significación, siendo en este caso de 0,05. Se rechaza  $H_0$ , por lo tanto el modelo con 12 variables contribuye igual que el modelo saturado.

Tabla 4

Paso	Resumen del Modelo -2 Log likelihood
1	1350,949

Se terminó de iterar en el paso 6, ya que el parametro estimado cambia menos que 0,01.

Si se utiliza el Test de Hosmer y Lemeshow<sup>34</sup>, sugerido para un número de variables significativo, se plantea la siguiente hipótesis:

$$H_0 : \text{El modelo ajustado es significativo, se ajusta correctamente}$$

$$H_1 : \text{El modelo ajustado no es significativo, no se ajusta correctamente}$$

Tabla 5

Test de Hosmer y Lemeshow

Paso	Chi cuadrado	gl	sig.
1	8,844	8	0,356

En la tabla anterior se observa que no hay antecedentes para rechazar  $H_0$ , por lo cual se concluye que el modelo ajustado es significativo.

El mejor modelo para el área urbana, se obtuvo con 12 variables, que son las siguientes:

- Número de personas en el hogar
- Escolaridad del Jefe de Hogar
- Jefe de Hogar Ocupado
- Jefe de Hogar Inactivo
- Edad del Jefe de Hogar
- Hogar con lavadora
- Hogar con teléfono
- Hogar con video
- Hogar con microondas
- Hogar con calefont
- Hogar con TV cable
- Jefe de hogar pertenece a un pueblo originario

La mejor ecuación de regresión logística para predecir la situación de pobreza de los hogares en la zona urbana es:

$$P_i = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k X_k}} = \frac{1}{1 + e^{-z}}$$

$$\text{Logit } P(\text{pobreza}) = Z = X' \beta$$

con  $X'$  = (1, NUMPER, ESC, OCUPADO, INACTIVO, EDADJH, LAVADORA, TELEFONO, VIDEO, MICROONDAS, CALEFONT, TVCABLE, ETNIA)

y  $\beta$  = (2,705; 0,379; -0,048; -2,394; -1,201; -0,046; -0,376; -0,785; -0,727; -0,476; -0,703; -0,742; 0,509)

### 5.3.3 Interpretación de los parámetros Estimados del Modelo – ODDS Ratio

El ODDS Ratio cambia cuando la i-ésima variable explicativa regresora se incrementa en una unidad, si:

- $\beta_i > 0$  significa que el ODDS RATIO se incrementa.
- $\beta_i < 0$  significa que el ODDS RATIO decrece.
- $\beta_i = 0$  significa que el factor es igual a uno, lo cual hace que ODDS RATIO no varía.

Tabla 6

		Variables en la Ecuación						95,0% C.I. for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Paso 1	NUMPER	,379	,040	91,429	1	,000	1,460	1,351	1,578
	ESC	-,048	,022	4,740	1	,029	,953	,912	,995
	OCUPADO	-2,394	,271	78,131	1	,000	,091	,054	,155
	INACTIVO	-1,201	,297	16,296	1	,000	,301	,168	,539
	EDAD	-,046	,006	52,371	1	,000	,955	,943	,967
	LAVADORA	-,376	,159	5,612	1	,018	,687	,503	,937
	TELEFONO	-,785	,179	19,329	1	,000	,456	,322	,647
	VIDEO	-,727	,209	12,106	1	,001	,483	,321	,728
	MICROONDAS	-,476	,215	4,883	1	,027	,621	,407	,948
	CALEFONT	-,703	,158	19,783	1	,000	,495	,363	,675
	TVCABLE	-,742	,258	8,312	1	,004	,476	,287	,788
	ETNIA	,509	,295	2,978	1	,084	1,664	,933	2,965
	Constant	2,705	,496	29,765	1	,000	14,948		

a. Variable(s) : NUMPER, ESC, OCUPADO, INACTIVO, EDAD, LAVADORA, TELEFONO, VIDEO, MICROONDAS, CALEFONT, TVCABLE, ETNIA.

En la columna  $\exp(\beta)$  de la tabla 6, se observa que:

$\beta_1$  y  $\beta_{12} > 0$ , por lo tanto, por cada 1 unidad que aumente el número de personas por hogar, la posibilidad de ser pobre aumenta en 1,46 veces si se mantiene el resto de las variables constantes. De la misma manera si que el Jefe de hogar pertenece a un pueblo originario la probabilidad de que el hogar sea pobre aumenta en 1,66 veces.

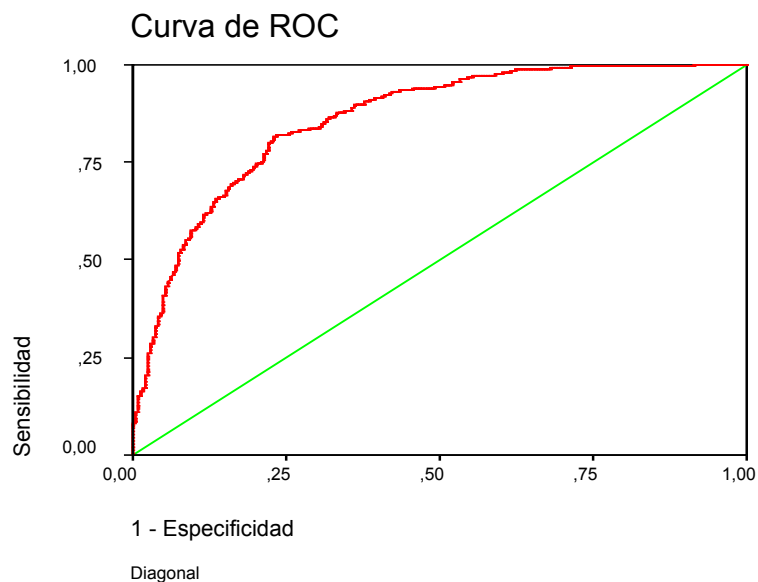
$\beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}$  y  $\beta_{11} < 0$ , por lo tanto, las variables, escolaridad, edad y actividad del Jefe de hogar, tenencia de lavadora, teléfono, video, microondas, calefón y conexión TV cable son factores que disminuyen la probabilidad de que el hogar sea pobre. Así, el hogar que posee lavadora, disminuye la probabilidad de ser pobre en 0,7 veces si mantiene constante el resto de las variables.

### 5.3.4 Curva de ROC – Sensibilidad y Especificidad

Cuando se utiliza un programa estadístico y se ejecuta una regresión logística, a cada observación se le asigna un valor de índice equivalente previsto de la regresión o también denominado valor predicho. Este valor predicho se utiliza para clasificar las unidades familiares u hogares como pobres o no pobres, los programas computacionales emplean normalmente una mitad como punto de corte (los hogares que se encuentran por encima del punto de corte se clasifican como pobres). No obstante, este punto de corte puede cambiarse.

Dado a que lo que interesa es clasificar adecuadamente a los pobres, se prefiere tener una menor proporción de coincidencias en la predicción total del modelo, pero una mejor para la categoría “pobre”, por lo cual para determinar el punto de corte se utilizó la curva de ROC basada en los cálculos de la especificidad y sensibilidad del modelo<sup>35</sup>, Para identificar el mejor punto de clasificaciones utilizó el SPSS graficando la “Curva de ROC” utilizando como variable la probabilidad predicha y la observada.

Gráfico nº 5



El eje de las abscisas, que corresponde a la proporción de falsos positivos, aparece etiquetado como 1-Especificidad y el eje de ordenadas, que corresponde a la proporción de aciertos positivos, aparece etiquetado como Sensibilidad. La línea curva escalonada representa los valores de 1-Especificidad y Sensibilidad para cada punto de corte ensayado (es decir, para cada una de las puntuaciones discriminantes incluidas en el análisis). Cada punto de esta curva se corresponde con un valor observado de la función.

Cuanto mayor es la curvatura de la línea, mayor es la capacidad de discriminación de la función. La mejor función discriminante posible sería aquella que permitiera obtener un 100% de aciertos positivos (sensibilidad = 1) Con un 0 % de falsos positivos (1-especificidad = 0). Pero en esa situación no existiría incertidumbre sobre el punto de corte óptimo ya que las distribuciones de frecuencias de cada grupo se encontrarían completamente. Por otra parte, la peor curva posible sería aquella que careciera por completo de curvatura (recorrería la bisectriz del ángulo definido por los dos ejes representados). En una situación como ésta, las distribuciones de frecuencias de los dos grupos se encontrarían completamente solapadas y un aumento de los aciertos positivos llevaría asociado un incremento idéntico de falsos positivos. El gráfico incluye una línea de referencia (la bisectriz) que corresponde a esta situación.

La situación de peor discriminación posible es la representada por la línea de referencia, diagonal, una proporción de área igual a 0,5 (sería una situación en la que se obtendrían tantos aciertos positivos como falsos positivos). Por el contrario, en la situación de mejor discriminación posible, el área existente bajo la curva contendría el 100 % del espacio; es decir, una proporción de área igual a 1. En consecuencia, el área existente bajo la curva ROC no sólo es un indicador de la curvatura de la línea, sino que, además, tiene un mínimo y un máximo conocidos.

La Tabla siguiente entrega una estimación del tamaño del área existente bajo la curva ROC. También ofrece el error típico de esa estimación (error estándar) y el nivel crítico (Sig. asintótica) resultante de contrastar la hipótesis nula de que el área existente bajo la curva vale 0,5. La tabla también incluye los límites del intervalo de confianza (calculados al 95 %) correspondientes a la estimación efectuada. Si el intervalo de confianza incluye el valor 0,5 o, lo que es lo mismo, si el nivel crítico es mayor que 0,05, no se podrá rechazar la hipótesis nula y, por tanto, no se podrá afirmar que los puntos de corte muestreados generen una curva ROC bajo la cual exista un área significativamente mayor que 0,5. Por el contrario, si el intervalo de confianza no incluye el valor 0,5, o lo que es lo mismo, si el nivel crítico es menor que 0,05, se podrá rechazar la hipótesis nula y concluir que el área existente bajo la curva ROC es significativamente mayor que 0,5.

Tabla 7 Área Bajo la Curva

Resultado del Test: Probabilidad Predicha

Area	Std. Error	Sig.(a)	Intervalo de Confianza 95%	
			Inferior	Superior
,860	,010	,000	,841	,879

A: Hipótesis Nula: Área de la Curva = 0.5

El gráfico de la curva ROC permite apreciar una curvatura muy pronunciada; y el área estimada (0,86) toma un valor próximo a 1. Además, dados los valores del nivel crítico y del intervalo de confianza, se puede concluir que esta área es significativamente mayor que 0,5, por lo cual indica una capacidad o eficacia predictiva del modelo alta.

La Tabla muestra una selección de los puntos representados (máximos y mínimos y las que se encuentran en torno al valor que corresponde al mejor punto de corte) en el gráfico, corresponde a los distintos valores que toma la función ordenados de menor a mayor, y las coordenadas correspondientes a esos valores en cada uno de los ejes del gráfico. El punto de corte va a estar dado por el cruce en el gráfico entre la especificidad y la sensibilidad.



Tabla 8 Coordenadas de la Curva ROC

Punto	Sensibilidad A	1- Especificidad B	1- Sensibilidad C	Especificidad d	a-d
0,0000	1,0000	1,0000	0,0000	0,0000	1,0000
0,0009	1,0000	0,9994	0,0000	0,0006	0,9994
0,0009	1,0000	0,9988	0,0000	0,0012	0,9988
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
0,1997	0,7832	0,2204	0,2168	0,7796	0,0036
0,2000	0,7805	0,2204	0,2195	0,7796	0,0009
0,2002	0,7778	0,2204	0,2222	0,7796	-0,0019
0,2004	0,7778	0,2198	0,2222	0,7802	-0,0025
0,2012	0,7778	0,2191	0,2222	0,7809	-0,0031
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
0,9681	0,0054	0,0000	0,9946	1,0000	-0,9946
0,9734	0,0027	0,0000	0,9973	1,0000	-0,9973
1,0000	0,0000	0,0000	1,0000	1,0000	-1,0000

El mejor punto de corte se encuentra en 0,20, el cual coincide con los niveles de pobreza urbana en Chile, que es un poco menos de 20%, como lo muestra el gráfico anterior, pero se cortará en 0,25, ya que el objetivo final de este modelo es asignar programas sociales a los hogares pobres y los estudios<sup>36</sup> muestran que hay un grupo importante de ellos que se encuentra en el borde de la línea y que transita permanentemente en estas dos situaciones y además que interesa sobremanera, de que los hogares pobres queden bien clasificados.

A continuación se observa una tabla de 2 por 2; en dicha tabla se muestran los casos observados de la situación de pobreza en los hogares. En esta tabla se enfrentan los valores estimados y observados de la situación de pobreza de los hogares calculando el porcentaje de coincidencias, hay un 80,2%.

Tabla 9 Tabla de Clasificación

Valor Observado	Valor Predicho NIVEL DE POBREZA		Porcentaje coincidencia	
	NO	SI		
NIVEL DE POBREZA	NO	1334	286	82,3
	SI	108	261	70,7
	TOTAL			80,2

Punto de Corte 0.25

### 5.3.5 Matriz de correlación e Interacciones

Se construyó la matriz de correlación entre las variables y se hicieron las pruebas correspondientes, las cuales no mostraron niveles de correlación significativos entre las variables independientes.

También se probaron en el modelo las interacciones que podían existir entre las variables Independientes. Se analizó si la variable sexo del jefe de hogar tenía un efecto de interacción con el resto de las variables utilizadas, se crearon las variables con el efecto interacción y el modelo fue testeado, no presentando en ninguno de los casos interacciones significativas.

### 5.3.6 Validación del Modelo Logístico

Para validar el modelo de regresión logística se utilizó la Encuesta CASEN 2003 completa, separada por área geográfica, encontrándose un 82% de casos coincidentes entre la variable original y la estimación del modelo.

Tabla 10 Cruce entre Niveles de Pobreza estimados y originales Zona Urbana

grupo predicho		Medición		Total	%
		POBREZA NO	original SI		
Pobres	NO	2531579	173503	2705082	93,6
	SI	476425	364769	841194	56,6
Total		3008004	538272	3546276	81,7

Fuente: Encuesta CASEN 2003.

### 5.4 Modelo Área Rural

El mismo procedimiento fue utilizado en el análisis del área rural. Las siguientes variables son significativas en la determinación de la pobreza en las zonas rurales:

- Número de personas en el hogar
- Escolaridad del Jefe de Hogar
- Jefe de Hogar Ocupado
- Jefe de Hogar Inactivo
- Edad del Jefe de Hogar
- Hogar con refrigerador
- Hogar con Alcantarillado
- Vivienda de Calidad Aceptable

Solamente con ocho variables se puede predecir si un hogar va a ser pobre o no, y se debe destacar que solamente 1 bien (el refrigerador) es determinante y que el acceso al alcantarillado no es significativo en la zona urbana y si en la rural.

Tabla 11

#### Variables en la Ecuación

		B	S.E.	Wald	df	Sig.	Exp(B)
Paso 1 <sup>a</sup>	NUMPER	,547	,065	71,739	1	,000	1,727
	ESC	-,097	,036	7,329	1	,007	,907
	OCUPADO	-2,617	,486	29,036	1	,000	,073
	INACTIVO	-1,579	,515	9,393	1	,002	,206
	EDAD	-,060	,009	43,354	1	,000	,942
	REFRIGER	-,808	,213	14,363	1	,000	,446
	ALCANTA	-,842	,244	11,858	1	,001	,431
	VIVIENDA	-,875	,413	4,484	1	,034	,417
	Constant	3,350	,803	17,412	1	,000	28,508

a. Variable(s) en paso 1: NUMPER, ESC, OCUPADO, INACTIVO, EDAD, REFRIGER, ALCANTA VIVIENDA.

Tabla 12 Tabla de Clasificación

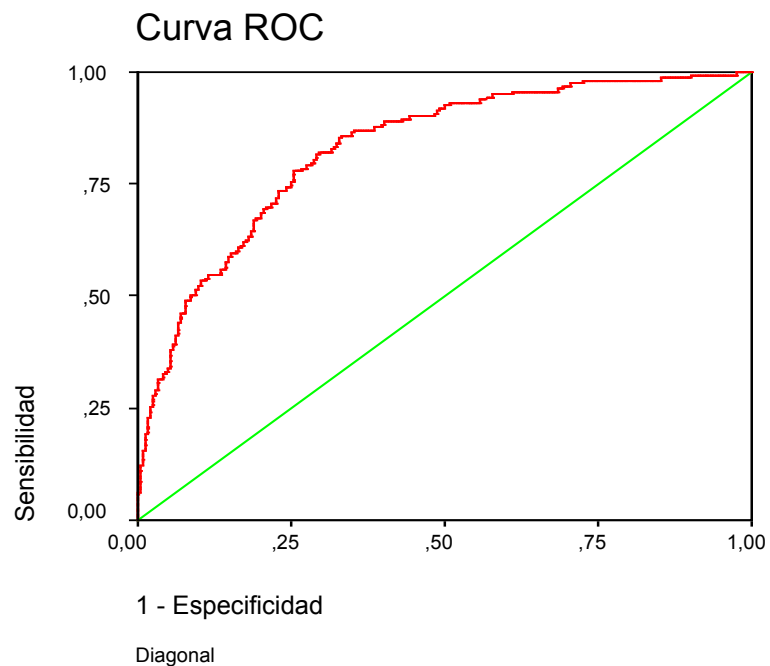
Tabla de Clasificación(a)

	Valores Observados		Valores Predichos		
			NIVEL DE POBREZA		Porcentaje de aciertos
	NO	SI	NO	SI	
Step 1	NIVEL DE POBREZA	NO	677	156	81,3
		SI	58	105	64,4
	Porcentaje				78,5

a El punto de corte es 0.21

El punto de corte seleccionado en la curva ROC es 0,21, lo que permite clasificar correctamente al 78,5% de los casos, a continuación se presenta la curva ROC y el test de significación correspondiente.

Gráfico nº 6



El gráfico de la curva ROC permite apreciar una curvatura muy pronunciada; y el área estimada 0,83 toma un valor próximo a 1. Además, dados los valores del nivel crítico y del intervalo de confianza, se puede concluir que esta área es significativamente mayor que 0,5.

Tabla 13 Área Bajo la Curva

Test Resultado: Probabilidad Predicha

Area	Std. Error(a)	Sig.(b)	95% Intervalo de Confianza	
			Mínimo	Máximo
,830	,017	,000	,797	,863

a Hipotesis nula: área = 0.5

La mejor ecuación de regresión logística para predecir la situación de pobreza de los hogares en la zona rural es:

$$p_i = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}} = \frac{1}{1 + e^{-z}}$$

Logit P(pobreza) = Z =  $\underline{X}' \beta$

con  $\underline{X}' =$  (1, NUMPER, ESC, OCUPADO, INACTIVO, EDADJH, REFRIGERADOR, ALCANTARILLADO, VIVIENDA)

y  $\beta =$  (3,350; 0,547; -0,097; -2,617; -1,579; -0,060; -0,808; -0,842; -0,875)

El cual tiene más de un 78% de validación con los datos observados, como lo muestra la tabla 12. En el modelo rural también fueron revisadas las interacciones, las que no mostraron valores significativos.

Al igual que en el análisis urbano, para validar el modelo de regresión logística se utilizó la Encuesta CASEN 2003 completa para la zona urbana, encontrándose un 81,7% casos coincidentes, entre la variable original y la estimación del modelo.

Tabla 14 Cruce entre Niveles de Pobreza estimados y originales Zona Rural

GRUPO PREDICHO		POBREZA OBSERVADA		Total	Coincidencias %
		NO	SI		
Pobres	NO	351.617	31.320	382.937	91,8
	SI	108.119	59.099	167.218	64,7
<b>Total</b>		<b>459.736</b>	<b>90.419</b>	<b>550.155</b>	<b>81,7</b>

Fuente: Encuesta CASEN 2003

## 6.- Conclusiones

Esta investigación se ha concentrado en estimar la contribución que distintos factores socioeconómicos de los hogares tienen sobre la pobreza de éstos.

La regresión Logística es una de las herramientas estadísticas con mejor capacidad para el análisis cuando la variable dependiente es categórica. Sirve para determinar los factores de riesgo y factores de prevención frente a la situación de pobreza que enfrentan los hogares. En esta investigación se obtiene un modelo que cumple con los supuestos y todos los tests estadísticos requeridos.

Con el modelo obtenido se alcanza un porcentaje de coincidencia entre la estimación de probabilidad de ser pobre y el valor observado de alrededor de un 80%, tanto en la zona rural como en la zona urbana.

En la zona urbana se obtiene un modelo que considera 12 variables significativas en la determinación de la pobreza. Aumenta la probabilidad de ser pobre: el número de personas por hogar, y la pertenencia o descendencia del jefe de hogar de algún pueblo originario, considerado en la Ley Indígena. Se destaca esta variable, la que debería ser incluida entre las variables de selección en los programas de pobreza urbana, ya que si se mantienen el resto de las variables constantes, y ésta aumenta en 1 unidad, la probabilidad de que el hogar sea pobre crece en 1,66 veces.

Las variables escolaridad, edad y actividad del Jefe de hogar, tenencia de lavadora, teléfono, video, microondas, calefón y conexión TV cable son factores que disminuyen la probabilidad de que el hogar sea pobre en la zona urbana.

Cabe destacar que cuando se realiza la prueba de medias en el capítulo 5, se demuestra que es significativa la diferencia de edad de los jefes de hogar pobres y no pobres, donde los que están a cargo de hogares pobres son más jóvenes.

En la zona rural, sólo 8 variables de las 23 utilizadas en el modelo inicial son significativas. Al igual que en la zona urbana, aumenta la probabilidad de ser pobre que exista un mayor número de personas en el hogar, y la disminuye el que los jefes de hogar tengan mayor cantidad de años de escolaridad, que estén ocupados o sean inactivos.

La diferencia con la zona urbana se observa en el tipo de bienes, ya que en la zona rural hay un solo bien que discrimina entre ser pobre o no, y es la posesión de un refrigerador. En relación a los servicios, corresponde el acceso a alcantarillado y que la calidad de la vivienda sea aceptable.

Esto último demuestra las diferentes características de la pobreza en las distintas zonas geográficas, y la necesidad de tener modelos que predigan o discriminen en forma independiente para cada una de las zonas, esto es relevante, ya que en Chile, la zona rural concentra una mayor proporción de pobres, pero tiene una importancia relativa muy baja, 14% de la población del país vive en la zona rural. Por lo tanto, los modelos generales tienden a mostrar el comportamiento urbano, lo que implica una mala focalización en los recursos y el desarrollo de las políticas sociales en los sectores rurales.

**ANEXOS****AnexoN° 1 Metodología de medición de la pobreza: El enfoque del ingreso o Línea de Pobreza (LP)**

La forma habitual de medir la pobreza utilizada en América Latina consiste en:

1. Definir una Canasta Alimentaria que satisfaga los requerimientos mínimos (calorías, proteínas, etc.) en base al consumo real de la población de cada país.
2. Se calcula el costo de esta Canasta de Alimentos y éste es considerado como la "Línea de Indigencia" que corresponde al costo mensual per cápita de una canasta básica alimentaria, cuya composición cubre las necesidades nutricionales promedio de la población y considera sus hábitos de consumo predominantes.  
Se estiman dos líneas de indigencia, una correspondiente a las zonas urbanas y la segunda a las zonas rurales. Las diferencias entre ambas se derivan del menor costo de los alimentos y del mayor requerimiento promedio de energía que, según estudios previos, caracteriza a las zonas rurales. En definitiva, la línea de indigencia rural representa el 77% del valor de la urbana.
3. Para definir la Línea de Pobreza se multiplica el costo de la Canasta Alimentaria por un factor, con el fin de obtener el costo de la Canasta necesaria para satisfacer las necesidades básicas (incluyendo las no alimentarias).
4. Este factor se elige con base en el cociente observado entre el gasto total de consumo del hogar y el gasto en alimentos, del primer estrato de hogares que satisfaga sus requerimientos nutricionales, este factor se obtiene de las Encuestas de Presupuestos Familiares (habitualmente este factor es 2).
5. Luego se calcula el porcentaje de hogares en situación de indigencia que corresponde a aquellos cuyo ingreso mensual per cápita es inferior a la Línea de Indigencia en relación al total de hogares. Una vez identificados estos hogares, es posible obtener el porcentaje de población indigente, que corresponde a las personas que son miembros de dichos hogares con relación al total de población.
6. La Línea de Pobreza en las zonas urbanas se calcula como el doble del valor de la línea de indigencia urbana, en tanto que el de las zonas rurales aquélla asciende a 1.75 vez el valor de la línea de indigencia rural.

Estas relaciones entre gastos básicos alimentarios y no alimentarios se derivan también de estudios previos realizados por CEPAL.

Los valores de la línea de pobreza son los siguientes:

Cuadro : Línea de Indigencia y Pobreza (\$nominal y real)

Zona	Tipo de Canasta	2000		2003
		\$ nominal	\$ real (Nov 2003)	\$ real (Nov 2003)
Urbana	Indigente	20.281	21.826	21.856
	Pobre no Indigente	40.562	43.653	43.712
Rural	Indigente	15.616	16.806	16.842
	Pobre no Indigente	27.328	29.410	29.473

Fuente: División Social MIDEPLAN en base a Casen 2000, 2003

## Anexo N°2 – Ficha técnica de la Encuesta CASEN 2003<sup>37</sup>

### Universo

La encuesta es representativa de la población que habita en hogares particulares del país. Es posible realizar estimaciones a nivel nacional, por zona, regional y comunal. El año 2003, es posible realizar estimaciones en 302 de un total de 341 comunas.

### Diseño Muestral

El tipo de muestreo utilizado es:

- **Estratificado:** La estratificación utilizada es de tipo geográfico. El país se dividió en estratos, entendiéndose como tal a la conjunción de división político administrativa (comuna o agrupación de comunas) y área geográfica (urbana o rural).
- **Por conglomerados:** En el área urbana están constituidos por los sectores de empadronamiento censal y en la rural como un conjunto de viviendas próximas.
- **Probabilístico:** En cada conglomerado se seleccionan sectores censales con probabilidad proporcional al tamaño de éste, medido por el número de viviendas. Las viviendas a encuestar se eligen dentro de cada sector seleccionado.

### Tamaño de la muestra

El año 2003, la muestra consideró 68.400 viviendas, se encuestaron 68.150 hogares, esto es, aproximadamente, 272.000 personas.

El error muestral a nivel nacional (hogares) es de 0,4%, a nivel urbano, 0,5% y a nivel rural, 0, 8%. El máximo error a nivel de región y zona es de 5%.

### Fecha del levantamiento de la encuesta

Entre el 8 de noviembre y el 20 de diciembre del 2003.

## Anexo N°3 Definiciones

“Línea de pobreza”: ingreso mínimo por persona para cubrir el costo de una canasta mínima individual para la satisfacción de las necesidades alimentarias y no alimentarias. Los hogares pobres son aquellos cuyos ingresos no alcanzan para satisfacer las necesidades básicas de sus miembros (\$43.712 en zona urbana, y \$29.473 en zona rural). La diferencia de líneas de corte urbano y rural se deben a la diferencia de disponibilidad de ingresos monetarios en cada una.

“Línea de indigencia”: ingreso mínimo por persona para cubrir el costo de una canasta alimentaria. Son indigentes los hogares que, aun cuando destinan todos sus ingresos a satisfacer las necesidades alimentarias de sus miembros, no logran cubrirlas adecuadamente (\$21.856 urbano y \$16.842 rural).

Hogar: aquel constituido por una persona o un grupo de personas, con o sin vínculos de parentesco, que comparten vivienda y tienen presupuesto de alimentación común.

**Anexo N°4      Tablas**

Tabla 1 Características de los hogares urbanos

Zona Urbana		EL HOGAR ES POBRE		
		NO	SI	TOTAL
El Jefe de hogar esta Ocupado:	NO	27,9	44,3	31,0
	SI	72,1	55,7	69,0
El Jefe de hogar es Inactivo:	NO	74,4	72,7	74,1
	SI	25,6	27,3	25,9
Sexo del Jefe de Hogar	MUJER	26,3	30,5	27,1
	HOMBRE	73,7	69,5	72,9
El hogar tiene Lavadora Automática	NO	38,2	68,4	43,8
	SI	61,8	31,6	56,2
Refrigerador	NO	13,1	29,5	16,1
	SI	86,9	70,5	83,9
Teléfono fijo	NO	43,8	81,4	50,8
	SI	56,2	18,6	49,2
Videograbador	NO	61,8	88,4	66,7
	SI	38,2	11,6	33,3
Microondas	NO	61,4	88,4	66,4
	SI	38,6	11,6	33,6
Computador	NO	74,0	93,8	77,6
	SI	26,0	6,2	22,4
Cálefont	NO	31,8	71,6	39,2
	SI	68,2	28,4	60,8
Teléfono Móvil	NO	55,1	66,8	57,3
	SI	44,9	33,2	42,7
Internet	NO	88,9	99,2	90,8
	SI	11,1	0,8	9,2
Tvcable	NO	75,3	94,1	78,8
	SI	24,7	5,9	21,2
Hogar pertenece a alguna etnia	NO	97,1	92,2	96,1
	SI	2,9	7,8	3,9
La vivienda tiene agua potable:	NO	3,7	11,4	5,2
	SI	96,3	88,6	94,8
La vivienda tiene Alcantarillado	NO	5,8	16,2	7,7
	SI	94,2	83,8	92,3
La vivienda es de buena calidad	NO	1,9	6,2	2,7
	SI	98,1	93,8	97,3
Más de 1 hogar en la vivienda	NO	96,4	94,3	96,0
	SI	3,6	5,7	4,0
Hay Hacinamiento en el hogar	NO	98,0	88,8	96,3
	SI	2,0	11,2	3,7



Tabla 2 Promedio de personas por hogar, escolaridad y edad promedio del Jefe de Hogar según situación de pobreza

	EL HOGAR ES POBRE		TOTAL
	NO	SI	
	Promedio	Promedio	Promedio
Número de Personas en el Hogar	3,5	4,4	3,7
Promedio de Escolaridad del Jefe de Hogar (años)	9,6	7,9	9,3
Edad del Jefe de Hogar (años)	51,4	45,7	50,4

Tabla 3 Test de significación estadística de promedio de personas por hogar, escolaridad y edad promedio del Jefe de Hogar según situación de pobreza

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Promedio de Escolaridad del Jefe de Hogar	26,692	,000	6,743	1987	,000	1,742	,258	1,235	2,248
			7,804	664,869	,000	1,742	,223	1,303	2,180
Número de Personas en el Hogar	1,273	,259	-9,289	1997	,000	-,924	,099	-1,119	-,729
			-9,318	551,161	,000	-,924	,099	-1,118	-,729
Edad del Jefe de Hogar	9,798	,002	6,377	1997	,000	5,745	,901	3,978	7,512
			6,794	592,132	,000	5,745	,846	4,084	7,406

Tabla 4

Zona Rural		EL HOGAR ES POBRE		
		NO	SI	TOTAL
El Jefe de hogar esta Ocupado:	NO	30,9	40,5	32,5
	SI	69,1	59,5	67,5
El Jefe de hogar es Inactivo:	NO	70,4	70,6	70,4
	SI	29,6	29,4	29,6
Sexo del Jefe de Hogar	MUJER	17,1	18,4	17,3
	HOMBRE	82,9	81,6	82,7
El hogar tiene Lavadora Automática	NO	72,1	87,7	74,6
	SI	27,9	12,3	25,4
Refrigerador	NO	32,5	54,0	36,0
	SI	67,5	46,0	64,0
Teléfono fijo	NO	93,3	97,5	94,0
	SI	6,7	2,5	6,0
Videograbador	NO	89,1	96,3	90,3
	SI	10,9	3,7	9,7
Microondas	NO	92,0	98,8	93,1
	SI	8,0	1,2	6,9
Computador	NO	95,7	98,8	96,2
	SI	4,3	1,2	3,8
Cálefont	NO	78,4	92,0	80,6
	SI	21,6	8,0	19,4
Teléfono Móvil	NO	63,7	77,3	65,9
	SI	36,3	22,7	34,1
Internet	NO	98,9	100,0	99,1
	SI	1,1	0,0	0,9
Tvcable	NO	97,2	100,0	97,7
	SI	2,8	0,0	2,3
Hogar pertenece a alguna etnia	NO	88,0	81,0	86,9
	SI	12,0	19,0	13,1
La vivienda tiene agua potable:	NO	61,9	75,5	64,1
	SI	38,1	24,5	35,9
La vivienda tiene Alcantarillado	NO	56,4	79,8	60,2
	SI	43,6	20,2	39,8
La vivienda es de buena calidad	NO	3,7	9,2	4,6
	SI	96,3	90,8	95,4
Más de 1 hogar en la vivienda	NO	98,7	99,4	98,8
	SI	1,3	0,6	1,2
Hay Hacinamiento en el hogar	NO	96,5	82,4	94,2
	SI	3,5	17,6	5,8

Tabla 5 Promedio de personas por hogar, escolaridad y edad promedio del Jefe de Hogar según situación de pobreza – Zona Rural

	NIVEL DE POBREZA	N	Mean	Std. Deviation	Std. Error
Numero de Personas por Hogar	NO	835	3,60	1,601	,055
	SI	163	4,74	1,661	,130
Escolaridad del Jefe de Hogar	NO	834	5,88	3,997	,138
	SI	163	5,34	3,229	,253
Edad del Jefe de Hogar	NO	835	53,89	16,309	,564
	SI	163	46,28	13,636	1,068

Tabla 6 Test de significación estadística de promedio de personas por hogar, escolaridad y edad promedio del Jefe de Hogar según situación de pobreza – zona rural

**Independent Samples Test**

	Levene's Test for Equality of Variances		t-test for Equality of Means							
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference		
								Lower	Upper	
Numero de Persona por Hogar	Equal variances assumed	,238	,626	-8,305	996	,000	-1,146	,138	-1,417	-,875
	Equal variances not assumed			-8,102	224,659	,000	-1,146	,141	-1,425	-,867
Escolaridad del Jefe Hogar	Equal variances assumed	10,880	,001	1,607	995	,108	,534	,332	-,118	1,187
	Equal variances not assumed			1,853	268,868	,065	,534	,288	-,033	1,102
Edad del Jefe de Hc	Equal variances assumed	12,767	,000	5,593	996	,000	7,617	1,362	4,945	10,290
	Equal variances not assumed			6,306	261,154	,000	7,617	1,208	5,239	9,996

Tabla 7 Variables en la ecuación del modelo completo – zona urbana

**Variables in the Equation**

Step		B	S.E.	Wald	df	Sig.	Exp(B)
1	NUMPER	,406	,112	13,236	1	,000	1,501
	ESC	-,125	,046	7,325	1	,007	,883
	OCUPADO	-2,368	,271	76,121	1	,000	,094
	INACTIVO	-1,165	,303	14,772	1	,000	,312
	EDAD	-,039	,010	14,902	1	,000	,962
	LAVADORA	-,365	,159	5,244	1	,022	,695
	TELEFONO	-,785	,179	19,291	1	,000	,456
	VIDEO	-,729	,210	12,087	1	,001	,482
	MICROONDAS	-,462	,216	4,589	1	,032	,630
	CALEFONT	-,707	,159	19,819	1	,000	,493
	TVCABLE	-,719	,259	7,708	1	,005	,487
	ETNIA	1,136	,890	1,629	1	,202	3,114
	snumper	-,023	,077	,091	1	,762	,977
	sedad	-,006	,006	1,008	1	,315	,994
sesc	,057	,030	3,679	1	,055	1,059	
setnia	-,475	,597	,633	1	,426	,622	
Constant	2,748	,499	30,343	1	,000	15,614	

a. Variable(s) entered on step 1: NUMPER, ESC, OCUPADO, INACTIVO, EDAD, LAVADORA, TELEFONO, VIDEO, MICROONDAS, CALEFONT, TVCABLE, ETNIA, snumper, sedad, sesc, setnia.

Tabla 8 Matriz de correlaciones de las variables en la zona urbana

**Correlation Matrix**

	Constant	NUMPER	ESC	OCUPADO	INACTIVO	EDAD	LAVADORA	TELEFONO	VIDEO	MICROONDAS	CALEFONT	TVCABLE	ETNIA
Step 1	Constant	1,000	-,218	-,616	-,544	-,342	-,737	,020	,096	,043	,032	-,009	-,074
	NUMPER	-,218	1,000	,017	-,099	,036	-,092	-,109	-,136	-,057	-,020	-,017	-,025
	ESC	-,616	,017	1,000	,026	,074	,456	-,125	-,075	-,150	-,071	-,154	-,026
	OCUPADO	-,544	-,099	,026	1,000	,753	,135	,002	,032	-,025	,060	,011	-,008
	INACTIVO	-,342	,036	,074	,753	1,000	-,217	,016	,002	-,026	,045	,009	-,025
	EDAD	-,737	-,092	,456	,135	-,217	1,000	-,026	-,127	-,021	-,004	-,094	,006
	LAVADORA	,020	-,109	-,125	,002	,016	-,026	1,000	-,085	-,104	-,140	-,189	-,024
	TELEFONO	,096	-,136	-,075	,032	,002	-,127	-,085	1,000	-,052	-,185	-,190	-,047
	VIDEO	,043	-,057	-,150	-,025	-,026	,021	-,104	-,052	1,000	-,184	-,024	-,085
	MICROONDAS	,032	-,020	-,071	,060	,045	-,004	-,140	-,185	-,184	1,000	-,083	-,060
	CALEFONT	-,009	,022	-,154	,011	,002	-,094	-,189	-,190	-,024	-,083	1,000	-,058
	TVCABLE	-,009	-,017	-,026	,008	,009	,006	-,024	-,047	-,085	-,060	-,058	1,000
	ETNIA	-,074	-,025	,049	-,032	-,025	,070	-,012	,039	,034	-,031	,020	-,054

## 7.- Revisión bibliográfica

Agresti, A.; Categorical data analysis. Wiley Interscience. Canadá 2002.

Altimir, Oscar; La dimensión de la pobreza en América Latina, Naciones Unidas, Cuadernos de la CEPAL, N°27, Santiago, 1979.

Atkinson, A.B.. On the Measurement of Inequality, en Journal of Economic Theory, núm. 2, 1970

BRIONES, G. (1995) Métodos y Técnicas de Investigación para las Ciencias Sociales, Editorial Trillas. México.

Cea Dàncona; M. Ángeles; Análisis Multivariable. Teoría y práctica en la Investigación Social; Editorial Síntesis. Madrid, 2002.

CELADE; Metodología de localización de bolsones de pobreza intracomunales y espacios de inversión. (mimeo, primer informe), Santiago, noviembre 1991.

CEPAL; Determinación de las necesidades de energía y proteínas para la población de nueve países latinoamericanos; (mimeo) Santiago, octubre 1990.

CEPAL; Una estimación de la magnitud de la pobreza en Chile; Santiago, octubre 1990.

CEPAL; Medición y análisis de la pobreza: Notas Técnicas, 26 de abril del 2002 (mimeo).

CEPAL; Uso de Regresiones Categóricas para comprobar el Rendimiento de los Parámetros de Determinación de Beneficiarios.

Chossudovsky; Global Poverty in the Late 20th Century” Journal of International Affairs, Vol 52, n°1, 1998.

Contreras, D.; Cooper, R.; Herman J. Y Neilson C.; Dinámica de la Pobreza y Movilidad Social: Chile 1996-2001; Depto. De Economía; U. De Chile (mimeo) agosto 2004.

Córtés Fernando; Regresión Logística en la investigación social: potencialidades y limitaciones; CES, COLMEX, página web.

Cumsille; Francisco; Análisis de Datos Categóricos; Apuntes de clase, Magíster Bioestadística, Escuela de Salud Pública; Facultad de Medicina; Universidad de Chile.

Feres, J.C. y Mancero X.; El método de las necesidades básicas insatisfechas y sus aplicaciones en América Latina; Estudios Estadísticos y prospectivos, CEPAL, mimeo, Santiago 2001.

Flores Manrique; L.; Análisis Estadístico de los Factores de Riesgo que Influyen en la Enfermedad Angina de Pecho. Tesis Digitales, Universidad Nacional de San Marcos, Lima, Perú, 2000.

French-Davis, Ricardo, Distribución y Pobreza en Chile; ICHEH, 1996 (mimeo).

Guerrero, Miguel, Metodología para la caracterización socioeconómica, págs 11-28 de “Como ha cambiado la vida de los chilenos”; INE, Chile 2004.

Gujarati, Damodar; "Econometría"; INE, Chile 2004. Ed. Mc Graw Hill, 3ª edición, Colombia, 1997.

Hair J.F.; Anderson R.E.; Tatham R.L. y Black W.C. Análisis Multivariante; Ed. Prentice Hall; 1999.

Hosmer D. Y Lemeshow S.; Applied Logistic Regression; Wiley Series in Probability and Statistics; 2ª edición, Canadá, 2000.

Hosmer D. Y Lemeshow S.; Applied Logistic Regression; Wiley Interscience Publication; Solutions Manual to Accompany, Canadá, 2001.

Jovel; Albert J.; "Análisis de Regresión Logística" Centro de Investigaciones Sociológicas; Madrid, 1995.

INE; Censo de Población y Vivienda. Chile 2002; Santiago; 2003.

INE; Metodología de Clasificación Socioeconómica de los Hogares Chilenos; Santiago; agosto 2003.

Kaztman, Rubén; La heterogeneidad de la pobreza. El caso de Montevideo. Revista de la CEPAL N°37. Santiago, 1990.

Larrañaga, Osvaldo y equipo; Mejoramiento del Factor Discriminatorio de la Ficha CAS; Informe Final; Santiago, diciembre 2002.

Levy J.P. y Varela Jesús; Análisis Multivariante para Ciencias Sociales; Ed. Prentice Hall; 2003.

López O.; P. Evaluación de la Precisión del Diagnóstico de Edad Gestacional Fetal basado en los Datos de la Mujer; Tesis de Grado Magíster en Bioestadística; Escuela de salud Pública, Facultad de Medicina; U. De Chile, Santiago, 2002.

LOHR S. L. (2000): Muestreo: diseño y análisis. Thomson 1ª. Edición.

MIDEPLAN; Volumen 1: "Pobreza, Distribución del Ingreso e Impacto Distributivo del Gasto Social"; Santiago; agosto 2004.

MIDEPLAN; Distribución e Impacto Distributivo del Gasto Social en los Hogares, 1996. Santiago, mayo 1998. Mimeo.

MIDEPLAN - PNUD, Desarrollo Humano en las Comunas de Chile; Temas de Desarrollo Humano Sustentable, N°3, Santiago, 2000.

Miller S.M.; Rein M., Roby M. y Cross B., Poverty, Inequality and Conflict en Annals of the American Academy of Political Science, 1967.

Moon, Bruce, The Political Economy of Basic Human Needs; Cornell University Press, 1991.

Ortúzar, Juan de Dios; Modelos Econométricos de Elección Discreta. Ediciones Universidad Católica de Chile; 2000.

Pagano, M. Y Gauvreau, K. Principles of Biostatistics; Duxbury, Thomson Learning; USA, 2000.

Pérez, César; Análisis Estadístico Multivariado con SPSS; Ed. Prentice Hall, Madrid, España, 2004.

Pérez, César; Técnicas Estadísticas con SPSS; Ed. Prentice Hall, Madrid, España, 2003.

PNUD, Desarrollo sin Pobreza. II Conferencia Regional sobre la Pobreza en América Latina y el Caribe, Quito, Ecuador, 1990.

Raczynski, Dagmar, Políticas Sociales y Programas de Combate a la Pobreza en Chile: balance y desafíos, Colección Estudios CIEPLAN, 39, Santiago, Chile, 1994.

Rajaraman, Indira; Constructing the Poverty Line: Rural Punjab, 1960-1961, Discussion Paper, núm. 43, Programa de Investigación en Desarrollo Económico, Universidad de Princeton.

Rein, M. ; Problems in the Definition and Measurement of Poverty, en Peter Townsend, The Concept of Poverty, Heineman, Londres, 1971, p. 46.

Sen, Amartya y Nussbaum M.; La calidad de vida. ; Fondo de Cultura Económica; México, 1993.

Sen, Amartya, Sobre conceptos y medidas de Pobreza. Revista de Comercio Exterior, vl 42, número 4; México, abril 1992.

Sen, Amartya, Sobre conceptos y medidas de Pobreza. Revista de Comercio Exterior, vl 42, número 4; México, abril 1992.

Silva Aycaguer; L.C., , Excursión a la Regresión Logística en Ciencias de la Salud..Ediciones Diaz de Santos; Madrid 1995.

Skoufias, Emmanuel; Evaluación de la selección de hogares beneficiarios en PROGRESA. Programa de Educación, Salud y Alimentación. PROGRESA; México 2004; mimeo.

Stigler, G. J. ; The Cost of Subsistence, en Journal of Farm Economics, núm. 27,1945;

Townsend, Peter; Poverty as Relative Deprivation: Resources and Styles of Living, 3 (ed.), en Dorothy Wedderburn; Poverty, Inequality and Class Structure, Cambridge University Press, Cambridge, 1974

Sen, Amartya, Sobre conceptos y medidas de Pobreza. Revista de Comercio Exterior, vl 42, número 4 México, abril 1992.

Vivanco, M., Análisis Estadístico Multivariado, Teoría y práctica. Ed. Universitaria, Santiago, Chile, 1999.

## Notas al Pié de Página

<sup>1</sup> MIDEPLAN - PNUD, Desarrollo Humano en las Comunas de Chile; Temas de Desarrollo Humano Sustentable, N°3, Santiago, 2000.

<sup>2</sup> Sistema Nacional de Indicadores Municipales, Subsecretaría de Desarrollo Regional, Ministerio del Interior, Chile.

<sup>3</sup> El número de comunas de las cuales se puede realizar inferencias estadísticas, varía en cada versión de la Encuesta, el diseño de la muestra es regional y por zona geográfica.

<sup>4</sup> La metodología de medición de pobreza utilizada actualmente en Chile es exclusivamente en base a los ingresos per cápita del hogar.

<sup>5</sup> MIDEPLAN; Distribución e Impacto Distributivo del Gasto Social en los Hogares, 1996. Santiago, mayo 1998. Mimeo.

<sup>6</sup> Los Indicadores utilizado en vivienda corresponden a las definiciones realizadas por el Ministerio de Planificación, MIDEPLAN. en el documento Situación habitacional en Chile, 1990-1998; Santiago, julio 1999.

<sup>7</sup> Se define como preceptores de ingresos en el hogar a las personas que tienen ocupación, perciben rentas, jubilaciones o subsidios monetarios.

<sup>8</sup> Altimir, Oscar; La dimensión de la pobreza en América Latina, Naciones Unidas, Cuadernos de la CEPAL, N°27, Santiago, 1979.

<sup>9</sup> En este enfoque, se consideramos la naturaleza y la magnitud de las diferencias entre el 20 o el 10 por ciento más pobre de la población y el resto de ella. El interés se centra en cerrar las brechas entre los que están abajo y los que están mejor en cada dimensión de la estratificación social (Sen). La discusión que se ha planteado frente a este enfoque es que la desigualdad es fundamentalmente un problema distinto de la pobreza. La desigualdad y la pobreza están relacionadas pero ninguno de los conceptos subsume al otro..

<sup>10</sup> En Chile, la última revisión se realizó en el año 1987.

<sup>11</sup> La línea de pobreza representa el ingreso mínimo necesario por persona para cubrir el costo de una canasta mínima individual para la satisfacción de las necesidades básicas, alimentarias y no alimentarias. La línea de indigencia representa el ingreso mínimo necesario por persona para cubrir el costo de una canasta alimentaria.

<sup>12</sup> PNUD, Desarrollo sin Pobreza. II Conferencia Regional sobre la Pobreza en América Latina y el Caribe, Quito, Ecuador, 1990.

<sup>13</sup> Institución en Chile encargada de medir los niveles de pobreza en el país.

<sup>13</sup> CEPAL; Medición y análisis de la pobreza: Notas Técnicas, 26 de abril del 2002 (mimeo).

<sup>15</sup> Se clasifican los hogares en 3 categorías : Indigentes, pobres no indigentes y no pobres

<sup>16</sup> La Encuesta CASEN se aplica a hogares de todos los estratos socioeconómicos seleccionados por muestreo estratificado (región y zona), aleatorio dentro de los estratos y finalmente por conglomerados, obteniéndose así una muestra representativa de todo el país. Además se obtiene representatividad regional, por zona y en algunas comunas. La Encuesta se ha realizado en noviembre de los años 1987, 1990, 1992, 1994, 1996, 1998 y 2000; la próxima se realizará en noviembre del 2003. (ver Anexo N°1).

<sup>17</sup> Ver Anexo N°1

<sup>18</sup> MIDEPLAN; Volumen 1: "Pobreza, Distribución del Ingreso e Impacto Distributivo del Gasto Social"; Santiago; agosto 2004.

<sup>19</sup> Córtes Fernando; Regresión Logística en la investigación social: potencialidades y limitaciones; CES, COLMEX, página web.

<sup>20</sup> Córtes Fernando; Regresión Logística en la investigación social: potencialidades y limitaciones; CES, COLMEX, página web.

<sup>21</sup> La formulación matemática de este modelo está basada en el libro de Hosmer D. Y Lemeshow S.; Applied Logistic Regression; Wiley Series in Probability and Statistics; 2ª edición, Canadá, 2000.

<sup>22</sup> Jovel; Albert J.; "Análisis de Regresión Logística" Centro de Investigaciones Sociológicas; Madrid, 1995.

<sup>23</sup> Vivanco, M., Análisis Estadístico Multivariado, Teoría y práctica. Ed. Universitaria, Santiago, Chile, 1999.

<sup>24</sup> Boltvinik Julio.

<sup>25</sup> Corresponde a la razón de las probabilidades  $P(Y)$  y  $1 - p(y)$ .

<sup>26</sup> Hosmer D. Y Lemeshow S.; Applied Logistic Regression; Wiley Series in Probability and Statistics; 2ª edición, Canadá, 2000.

---

<sup>27</sup> Jovel; Albert J.; "Análisis de Regresión Logística" Centro de Investigaciones Sociológicas; Madrid, 1995.

<sup>28</sup> Hosmer D. Y Lemeshow S.; Applied Logistic Regression; Wiley Series in Probability and Statistics; 2ª edición, Canadá, 2000.

<sup>29</sup> El objetivo de la muestra es realizar inferencias estadísticas a nivel regional y en 302 comunas del país, por lo cual no es proporcional en cada región.

<sup>30</sup> Hogar: aquel constituido por una persona o un grupo de personas, con o sin vínculos de parentesco, que comparten vivienda y tienen presupuesto de alimentación común.

<sup>31</sup> Se excluye la variable zona.

<sup>32</sup> Fue analizado computacionalmente y efectivamente Zona actúa como variable confusora.

<sup>33</sup> Adelante Condicional. Método de selección por pasos *hacia adelante* que, partiendo del modelo nulo, va incorporando aquellas variables cuyo *estadístico de puntuación*, siendo significativo, posee la probabilidad asociada más pequeña. Tras incorporar al modelo una nueva variable, todas las variables incluidas hasta ese momento son revisadas para determinar si existe alguna que debe ser excluida (es decir, para determinar si, como consecuencia de la nueva incorporación, el *estadístico de puntuación* de alguna variable ha dejado de ser significativo). El proceso se detiene cuando entre las variables no incluidas en el modelo no queda ninguna cuyo estadístico de puntuación sea significativo.

<sup>34</sup> Esta prueba es utilizada cuando el modelo contiene un gran número de variables independientes. Evalúa la bondad del modelo construyendo una tabla de contingencia, divide la muestra en 10 grupos a partir de los deciles de las probabilidades estimadas (a los 10 grupos resultantes se les denomina deciles de riesgo). En cada decil de riesgo se calcula el número de casos que pertenecen a cada categoría de la variable dependiente (número observado) y el número de casos que el modelo pronostica que pertenecen a cada categoría de la variable dependiente (número esperado). Se compara el número de casos observado con el número de casos esperado. Este estadístico permite contrastar la hipótesis nula de igualdad de distribuciones, es decir, la hipótesis de que la variable dependiente se distribuye de la misma manera en los 10 deciles de riesgo o, que no existen diferencias entre las frecuencias observadas y las esperadas.

<sup>35</sup> Jovel; Albert J.; "Análisis de Regresión Logística" Centro de Investigaciones Sociológicas; Madrid, 1995 y Cea Dàncona; M. Ángeles; Análisis Multivariable. Teoría y Práctica en la Investigación Social; Editorial Síntesis. Madrid, 2002.

<sup>36</sup> Contreras, D.; Cooper, R.; Herman J. Y Neilson C.; Dinámica de la Pobreza y Movilidad Social: Chile 1996-2001; Depto. De Economía; U. De Chile (mimeo) agosto 2004.

<sup>37</sup> MIDEPLAN; Volumen 1: "Pobreza, Distribución del Ingreso e Impacto Distributivo del Gasto Social"; Santiago; agosto 2004.