



PROELCE



PROGRAMA DE ACTIVIDADES CONJUNTAS ELAS-CELADE

0022 | 0048700
ha recibido: 2/11/16
ARCHIVO de DOCUMENTOS
Original NO SALE de la oficina
I 424

// A PROPOSITO DE LA MEDICION CUANTITATIVA DEL
EFECTO DE LAS VARIABLES CUALITATIVAS
EN LA INVESTIGACION SOCIO-DEMOGRAFICA. //

Fernando Cortés .
Laura Gougain A. (autor)

1110

THE UNIVERSITY OF CHICAGO
DIVISION OF THE PHYSICAL SCIENCES
DEPARTMENT OF CHEMISTRY

Chicago, Ill.

PROELCE

Programa de Actividades Conjuntas
ELAS-CELADE

Santiago de Chile, Junio de 1976.

FOR THE
SECRETARY OF THE
UNITED STATES
DEPARTMENT OF
THE ARMY

P R E F A C I O

En esta publicación hemos reunido tres trabajos, que se caracterizan porque sus desarrollos han sido relativamente independientes. El que forma el primer capítulo, corresponde a una inquietud nacida en el seno de la cátedra de Estadística Social del Programa Docente regular de FLACSO, en tanto que el segundo y tercero, han surgido como respuestas a algunos problemas que se han constituido en típicos de la investigación sociodemográfica emprendida en PROELCE.

Esta génesis diferencial, nos permite sostener que no sería posible pretender encontrar una lógica que articule de manera absolutamente coherente los tres capítulos. Sin embargo, si consideramos que el objeto de referencia es común, la medida de juntarlos en una sola publicación pareciera no ser totalmente arbitraria.

En efecto, el hilo conductor central está constituido por una serie de situaciones de investigación, que han desembocado en ecuaciones de regresión en que los factores explicativos pueden ser de naturaleza cualitativa. Nos preocupamos por distinguir los modelos en que todas las variables independientes son no métricas, de aquellos que combinan variables cuantitativas y cualitativas. A los segundos los hemos calificado como modelos mixtos, mientras que a los primeros como cualitativos.

Una segunda línea de conexión - que puede ser catalogada como subsidiaria -, es la que nos provee el nivel de agregación de las observaciones: los datos pueden referirse a agregados, o a las unidades individuales que los constituyen.

El cruce entre el nivel de medición de las variables, con el nivel de agregación, genera cuatro tipos alternativos de modelos: mixtos agregados, mixtos no agregados, cualitativos agregados, y cualitativos no agregados. Estos dos últimos han sido tratados en el primer y tercer capítulos, mientras que en el segundo y tercero nos hemos preocupado por los modelos mixtos no agregados. Ni la inquietud docente, ni la dinámica de la investigación nos ha conducido a centrar nuestra atención sobre modelos del tipo mixto agregado, por ello no se encontrará ninguna referencia a esta clase de modelos.

Por último, queremos señalar que la naturaleza específica del tercer capítulo consiste en un programa de computación que - al interior del paquete SPSS -, permite transformar las distintas categorías asociadas a las variables en regresores mudos.

I N D I C E

	Pág.
I. REGRESION CON FACTORES EXPLICATIVOS CUALITATIVOS	
1. Introducción	1
2. El ajuste de una función por pasos	4
3. El ajuste de un plano discreto	15
4. Una alternativa al análisis porcentual	23
5. Consideraciones técnicas adicionales	35
6. Conclusiones	38
II. PROBLEMAS DE ESTIMACION EN MODELOS CON REGRESORES MUDOS	
1. Introducción	43
2. Estimación de una ecuación de regresión con datos individuales a partir de datos agregados..	45
3. Estimación de un modelo agregativo y cualitativo	50
4. Modelos mixtos	55
5. Algunos problemas adicionales	65
6. Conclusiones	77
III. UN PROGRAMA PARA CALCULAR REGRESION MULTIPLES CON VARIABLES MUDAS, CON EL SISTEMA SPSS	
1. Objetivos	81
2. Acerca de las variables mudas	81
3. Algunos ejemplos	83
4. Acerca del sistema SPSS	88
5. Una forma sencilla de crear variables mudas	95

SECRET

100

1. The following information was obtained from a confidential source who has provided reliable information in the past.

2. It is the policy of the Department to maintain the confidentiality of all information received from confidential sources.

3. The information received from this source is being provided to you for your information only and should not be disseminated to other personnel.

4. The source has provided information regarding the activities of certain individuals who are active in the area of [redacted].

5. The information received from this source is being provided to you for your information only and should not be disseminated to other personnel.

6. The source has provided information regarding the activities of certain individuals who are active in the area of [redacted].

7. The information received from this source is being provided to you for your information only and should not be disseminated to other personnel.

8. The source has provided information regarding the activities of certain individuals who are active in the area of [redacted].

9. The information received from this source is being provided to you for your information only and should not be disseminated to other personnel.

10. The source has provided information regarding the activities of certain individuals who are active in the area of [redacted].

11. The information received from this source is being provided to you for your information only and should not be disseminated to other personnel.

12. The source has provided information regarding the activities of certain individuals who are active in the area of [redacted].

13. The information received from this source is being provided to you for your information only and should not be disseminated to other personnel.

14. The source has provided information regarding the activities of certain individuals who are active in the area of [redacted].

15. The information received from this source is being provided to you for your information only and should not be disseminated to other personnel.

16. The source has provided information regarding the activities of certain individuals who are active in the area of [redacted].

17. The information received from this source is being provided to you for your information only and should not be disseminated to other personnel.

SECRET

I. REGRESION CON FACTORES EXPLICATIVOS CUALITATIVOS.

Fernando Cortés.

1. Introducción.

Es frecuente, que en el proceso de construcción de hipótesis teóricas, el cientista social movilice nociones conceptuales cuyos orígenes se encuentran en los más diversos y variados campos del conocimiento. Como consecuencia de ello, algunos sectores del pensamiento teórico admiten una formalización más o menos inmediata, mientras que en otros, es necesario plantearse una labor de traducción. En efecto, en este caso, es necesario buscar en el interior del lenguaje matemático, el concepto que capture de manera adecuada la idea teórica. Por cierto, debemos consignar que hay situaciones, en que por variadas razones, se hace imposible tan solo emprender la labor de traducción. Es claro que, dada la naturaleza de nuestro interés, sólo focalizaremos nuestra atención sobre aquellos problemas que son susceptibles de ser formalizados. En resumidas cuentas, el proceso de formalización matemática puede ser visto como el conjunto de recursos que nos permiten pasar del lenguaje cotidiano, a uno de carácter matemático.

De otra parte, la relación entre las Matemáticas y la Estadística adopta formas variadas y múltiples, sin embargo, en este trabajo, sólo nos interesa la ligazón que nos permite transitar desde funciones hacia ecuaciones de regresión.

Ahora bien, pareciera que en presencia de problemas formalizables, los cientistas sociales son capaces de recorrer con

relativa fluidez el camino que permite expresar las concatenaciones conceptuales en articulaciones entre símbolos matemáticos, aunque estas últimas sólo sean del nivel de la geometría euclidiana. Sin embargo, el paso de la Matemática a la Estadística pareciera encerrar dificultades mayores. Las razones deben ser múltiples, sin embargo, queremos consignar aquellas, que a nuestro juicio, son las más claramente observables.

La primera, se encuentra en la reticencia de traducir por escrito, las hipótesis teóricas. Vale decir, hay obstáculo para animarse a poner sobre el papel los símbolos abstractos que pudieran representar el pensamiento teórico. Es paradójico que esta resistencia se minimice, cuando se recurre al lenguaje geométrico en circunstancias que uno de los principios básicos de la geometría analítica establece una relación biunívoca entre la representación gráfica y las ecuaciones.

La segunda razón es de un carácter más técnico. A través de los libros de texto de estadística^{1/}, se ha difundido la noción que postula que una de las limitaciones más importantes en la aplicación del modelo de regresión, se encuentra en el nivel de medición de las variables: sólo se podría estimar ecuaciones de regresión si todas y cada una de sus variables han sido expresadas al nivel métrico.

^{1/} Un ejemplo conspicuo es el libro de Sidney Siegel: Nonparametric Statistics for the Behavioral Sciences. Mc Graw Hill. 1956.

La primera causa, impediría el tránsito desde el dominio de las Matemáticas al de la regresión, por cuanto este último requiere que la aseveración teórica adopte la forma de una ecuación. La segunda, nos permitiría sostener que aun cuando dicha deducción se haga explícita, no sería posible obtener estimaciones de sus parámetros, debido a que la regresión supondría que las variables deben ser por lo menos intervalares.

La Economía, en general, y la Econometría, en particular, han abordado las complejidades derivadas del nivel de medición de las variables. A través de la definición de variables mudas - también denominadas ficticias, donde ambos nombres corresponden al vocablo inglés dummy -, han incorporado al modelo de regresión variables tales como sexo, categorías ocupacionales, ausencia o presencia de conflictos internacionales, etc.^{1/}

Uno de nuestros propósitos consiste en examinar las potencialidades de este instrumento, en relación a las situaciones de investigación que caracterizan a la Sociología Demográfica. Más específicamente, sabemos que la Sociología en general, y por consecuencia la socio-demografía, utilizan con profusión el instrumento denominado análisis de variables múltiples. En este trabajo, nos proponemos indagar respecto a la posibilidad de sustituirlo, por el modelo de regresión que incorpore variables ficticias. Este intento de reemplazo se justifica por el mayor potencial analítico de este último.

1/ Ver, por ejemplo, Arthur Goldberger: Econometric Theory. John Wiley, 1964. Págs.226-227.

Sin embargo, enfatizaremos por sobre los problemas de estimación, el enlace entre el pensamiento teórico y su expresión matemática. En particular, supondremos la existencia de hipótesis sustantiva que orienten la descomposición de la variable explicada en función de los factores explicativos, y a partir de este conocimiento nos preguntamos por la ecuación matemática que mejor lo refleja.

El hecho que hayamos decidido privilegiar la formalización, no implica que desestimemos el tratamiento de algunas de las complejidades que caracterizan a los problemas de estimación. En las páginas que siguen se entrelazan la formalización y la estimación, pero se pone el acento sobre la primera.

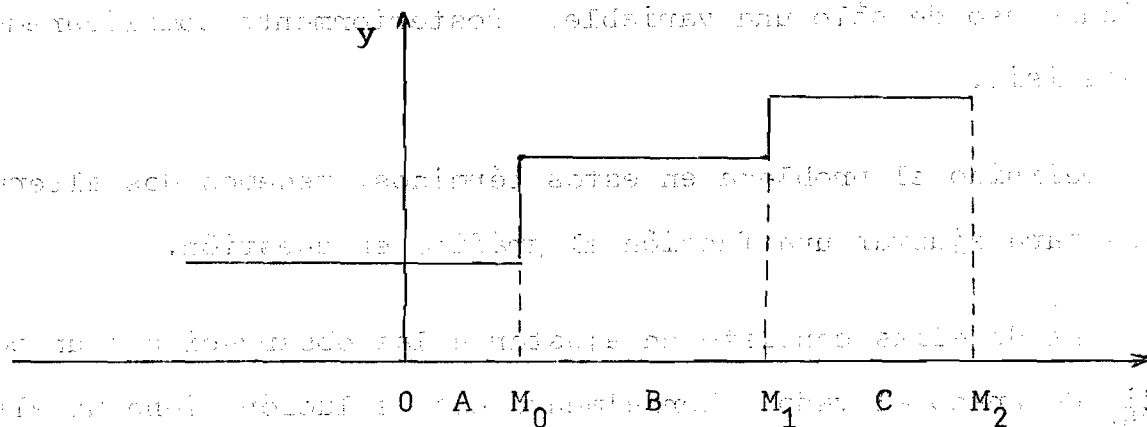
2. El ajuste de una función por pasos.

Con el objeto de darle contenido al problema que abordaremos, supongamos que un analista social, después de una serie de operaciones teóricas, llega a establecer que existe una relación discontinua entre el grado de urbanización y el volumen de producción industrial. La vinculación entre estas variables descansa en la noción de que las calidades asociadas al espacio físico condicionarían la actividad industrial. En este sentido, pareciera ser evidente el papel que juegan las disponibilidades de servicios fundamentales, como por ejemplo, energía eléctrica, agua, etc.; así como las facilidades de comunicación física con el ambiente exterior, en todo lo que tiene que ver con el movimiento de los

insumos necesarios y los productos terminados.

Esta aseveración de carácter general se especifica en el momento en que el investigador sostiene que, si bien esta relación existe, ella no es continua. Vale decir, entre determinados niveles de urbanización la producción industrial es relativamente constante, pero a partir de cierto punto, o de cierto grado de urbanización en adelante, la producción industrial sufre un aumento manteniéndose constante hasta que alcanza otro punto, donde nuevamente la producción sufre una modificación importante. Este proceso se repite un gran número de veces.

La expresión correspondiente a este esquema de pensamiento es:



donde el eje de abscisas mide el grado de urbanización (X) y, el de la ordenada, el volumen de la producción (Y)^{1/}, expresada en términos per-cápita.

^{1/} Con el único propósito de facilitar la redacción, usaremos como sinónimos las expresiones: volumen de producción industrial e industrialización.

La función representada en el gráfico se denomina función por pasos y sirve para expresar matemáticamente una situación perfectamente general.

En todos aquellos casos donde definimos clases estadísticas, la idea básica es la de homogeneidad de la clase. Se espera que el estrato así definido presente a lo menos una característica común. En otras palabras, se puede hablar de una clase estadística en la medida que la variabilidad intraclase sea mínima a la vez que la interclase sea significativa. Aún cuando generalmente se requiere de más de una característica o variable para definir una clase estadística, supondremos en un comienzo, que ello es posible haciendo uso de sólo una variable. Posteriormente complicaremos el análisis.

Definido el problema en estos términos, tenemos dos alternativas para ajustar una función al gráfico en cuestión.

Una de ellas consiste en ajustar a las observaciones un polinomio de grado elevado. Normalmente esta solución tiene un alto costo en términos de grados de libertad, lo que no la hace aconsejable. La otra alternativa es la de usar variables mudas (dummy). La cual, como veremos más adelante, si bien también tiene un costo en términos de grados de libertad, en la mayoría de los casos, es bastante menor que el que hay que pagar cuando se ajusta un polinomio.

Abordemos el problema de ajuste postulando un modelo lineal múltiple, para el caso en que tenemos tres clases o categorías definidas de la manera siguiente:

Clase A, está formada por todos aquellos X, tales que:

$$X_1 = \begin{cases} 1 & \text{si } X \leq M_0 \\ 0 & \text{para todo otro } X \end{cases} \quad (1)$$

Clase B, formada por todas las observaciones en que:

$$X_2 = \begin{cases} 1 & \text{si } M_0 < X \leq M_1 \\ 0 & \text{para todo otro } X \end{cases}$$

Clase C, se define para todo X, tal que

$$X_3 = \begin{cases} 1 & \text{si } M_1 < X \leq M_2 \\ 0 & \text{para todo otro } X \end{cases}$$

Se trata de ajustar la función:

$$(1) \quad Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

en que el término e, recibe el nombre de término de error estocástico y se agrega al valor esperado, para dar cuenta de la variabilidad intra-clase. Por lo tanto, puede ser utilizado como un indicador del éxito o fracaso al definir las clases^{1/}.

Si se hacen los supuestos clásicos respecto al término de error, vale decir:

^{1/} En particular, el término de error e, se usa como uno de los elementos del coeficiente de determinación R².

1. $E(e) = 0$

2. $Var(e) = \sigma^2$, donde σ^2 es una constante.

3. $Cov(e) = 0$; es decir, ausencia de correlación entre los residuos, se puede realizar un ajuste mínimo-cuadrático ordinario.

Pero, antes de ello hagamos uso del supuesto $E(e)=0$, para escribir el modelo (1) de la forma:

$$(2) \quad E \{Y/X\} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Si suponemos que $X \in M_0$, vale decir, X es tal que pertenece a la clase A, entonces X_1 asumirá el valor 1 y, X_2 y X_3 , tomarán valores cero. Aplicando la ecuación (2) tenemos:

$$E \{Y/A\} = \beta_1(1) + \beta_2(0) + \beta_3(0) = \beta_1$$

donde $E(Y/A)$, es el valor esperado de Y dado que la observación pertenece a la clase A. Por lo tanto, β_1 es el efecto esperado de la pertenencia a la clase A, sobre la variable explicada. Además, es constante, característica esencial que deseabamos capturar.

Si la observación pertenece a la clase B, entonces:

$$E\{Y/B\} = \beta_1(0) + \beta_2(1) + \beta_3(0) = \beta_2$$

donde β_2 , es el efecto esperado de la pertenencia a la clase B sobre la variable industrialización.

Del mismo modo se tiene: $E \{Y/C\} = \beta_3$

donde β_3 , es el efecto de pertenencia a la clase C.

En el ejemplo que estamos usando como ilustración, esto significa que β_1 es el efecto del primer nivel de urbanización sobre la industrialización; β_2 , es el impacto del segundo nivel; β_3 es el del tercer nivel. Por analogía se puede extender el procedimiento sobre cualquier número de clases. La única limitación radicarán en las disponibilidades de información numérica.

Las ecuaciones normales que se deben calcular para realizar el ajuste son:

$$\Sigma Y X_1 = \hat{\beta}_1 \Sigma X_1^2 + \hat{\beta}_2 \Sigma X_1 X_2 + \hat{\beta}_3 \Sigma X_1 X_3$$

$$\Sigma Y X_2 = \hat{\beta}_1 \Sigma X_1 X_2 + \hat{\beta}_2 \Sigma X_2^2 + \hat{\beta}_3 \Sigma X_2 X_3$$

$$\Sigma Y X_3 = \hat{\beta}_1 \Sigma X_1 X_3 + \hat{\beta}_2 \Sigma X_2 X_3 + \hat{\beta}_3 \Sigma X_3^2$$

Para todos los X pertenecientes a la clase A, $X_1 = 1$, en tanto que $X_2 = X_3 = 0$. Al imponer estas condiciones sobre las ecuaciones, la primera se transforma en:

$$\Sigma Y X_1 = \hat{\beta}_1 \Sigma X_1^2$$

mientras que la segunda y tercera se hacen iguales a cero. Las sumatorias de ella se extienden sobre todos los X, que pertenecen a la clase estadística A. Si suponemos que el tamaño de esta clase es N_A tendremos:

$$\hat{\beta}_1 = \frac{\sum Y}{\sum X_1^2} = \frac{\sum Y}{N_A} = \bar{Y}_A$$

debido a que $X_1 = 1$, por lo tanto, $X_1^2 = 1$, lo que implica que:

$$\sum X_1^2 = \sum X_1 = N_A$$

En el caso que se trabaje con las observaciones correspondientes a la clase B, la segunda ecuación normal toma la forma:

$$\sum Y = \hat{\beta}_2 \sum X_2$$

y,

$$\hat{\beta}_2 = \frac{\sum Y}{N_B} = \bar{Y}_B$$

mientras que la primera y tercera ecuaciones se hacen idénticamente iguales a cero.

Por medio de un procedimiento análogo se llega a:

$$\hat{\beta}_3 = \bar{Y}_C$$

En conclusión: los estimadores mínimos-cuadráticos de los efectos de pertenencia a clase, son las medias aritméticas de la variable explicada, distinguiendo por clase estadística.

Como la ecuación (1) no incluye un término libre y, generalmente es de buena costumbre considerarlo, supongamos que escribamos el modelo de la forma:

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

sen que X_0 es siempre igual a 1. Por tanto, esta ecuación es igual a la siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

En este caso la inclusión del término libre nos crea un problema bastante serio, por cuanto, siempre se cumplirá que:

$$X_0 = X_1 + X_2 + X_3 = 1$$

lo que genera un problema denominado técnicamente "multicolinealidad" que en esencia consiste en que las ecuaciones normales no son linealmente independientes y, por consiguiente, tenemos más incógnitas que ecuaciones, por lo que no es posible determinar los estimadores mínimo-cuadráticos.

Luego, al ajustar una función por pasos no se debe incluir un término libre.

Un lector inquieto puede preguntarse acerca de la dependencia de los resultados encontrados, respecto a la forma de definición de las variables. Es claro que, aún cuando se acepte las variables dicotómicas como una forma de simbolizar presencia o ausencia de atributos, hay varias formas de definirlas. Con el objetivo de estudiar la sensibilidad de los resultados respecto a la definición de las variables mudas consideremos que:

$$X_1 = 1 \text{ para todo } X$$
$$X_2 = \begin{cases} 1 & \text{para } M_0 < X \leq M_1 \\ 0 & \text{para todo otro } X \end{cases}$$
$$X_3 = \begin{cases} 1 & \text{para } M_1 < X \leq M_2 \\ 0 & \text{para todo otro } X \end{cases}$$

En este caso, aun sigue siendo válida la ecuación (1) y sus correspondientes ecuaciones normales. Sin embargo, hay que recordar que el regresor X_1 es igual a 1, para todas las clases. Este hecho nos introduce una ligera modificación sobre las ecuaciones normales.

Al reemplazar estos nuevos valores de las variables mudas, en las sumatorias constituyentes de las ecuaciones normales obtenemos:

$$\hat{\beta}_1 N + \hat{\beta}_2 N_B + \hat{\beta}_3 N_C = \Sigma Y_i$$

$$\hat{\beta}_1 + \hat{\beta}_2 N_B = \bar{Y}_B$$

$$\hat{\beta}_1 + \hat{\beta}_3 = \bar{Y}_C$$

donde N representa el total de las observaciones, mientras que N_B y N_C simbolizan los tamaños de las clases B y C respectivamente.

Resolviendo el sistema de ecuaciones se llega a:

$$\hat{\beta}_1 = \bar{Y}_A$$

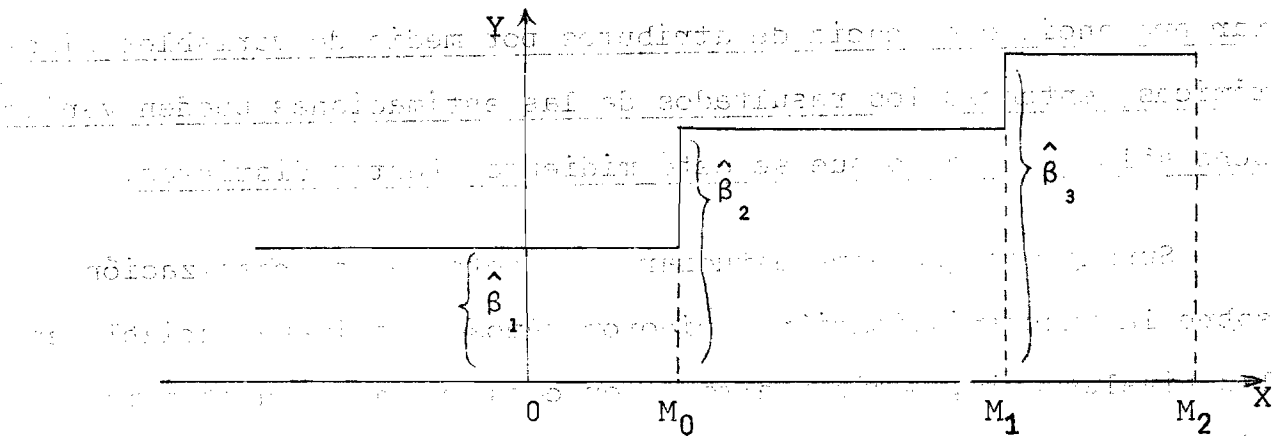
$$\hat{\beta}_2 = \bar{Y}_B - \bar{Y}_A$$

$$\hat{\beta}_3 = \bar{Y}_C - \bar{Y}_A$$

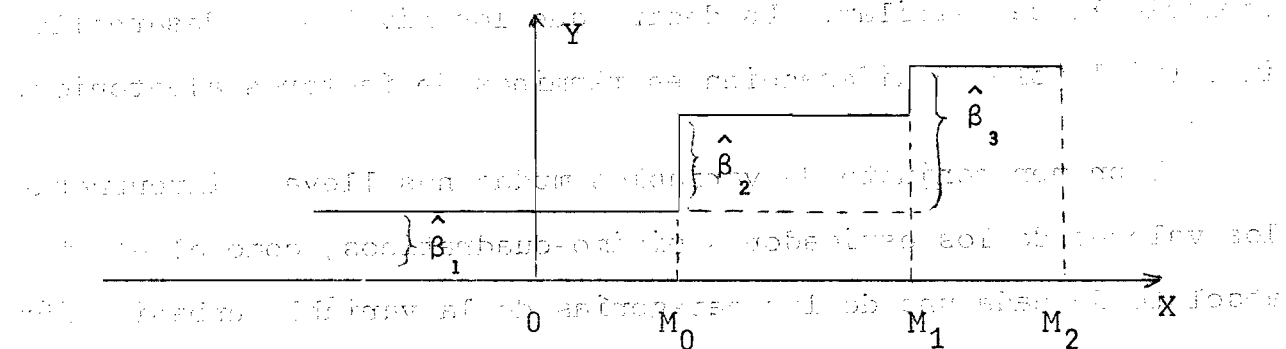
Pareciera que los resultados dependen de la manera como se definan las variables. Si esto fuese cierto el procedimiento no tendría valor. Sin embargo, la diferencia es solo aparente por cuanto en ambos casos se está midiendo "cosas" distintas. Como

consecuencia los resultados deben ser diferentes. El primer conjunto de variables mudas mide el impacto absoluto de cada clase estadística sobre la variable explicada, mientras que el segundo mide el impacto relativo, tomando como base de comparación la primera clase.

En términos gráficos: con el primer conjunto de variables se está midiendo las alturas:



en tanto que, con el segundo se está midiendo el efecto diferencial, tal como se señala en el siguiente gráfico:



Si la interpretación gráfica es correcta, al sumar $\beta_1 + \beta_2$, para el segundo conjunto de variables mudas, debe obtenerse el

mismo resultado que el $\hat{\beta}_2$, para el primer conjunto. Lo mismo debe ocurrir al sumar $\hat{\beta}_1$ a $\hat{\beta}_3$, para el segundo conjunto. Es decir, la suma tiene que ser igual al $\hat{\beta}_3$ correspondiente a las primeras definiciones.

$$\hat{\beta}_2 = (\bar{Y}_B - \bar{Y}_A) + \bar{Y}_A = \bar{Y}_B$$

$$\hat{\beta}_3 = (\bar{Y}_C - \bar{Y}_A) + \bar{Y}_A = \bar{Y}_C$$

Por consiguiente, podemos concluir que si aceptamos representar presencia o ausencia de atributos por medio de variables dicotómicas, entonces los resultados de las estimaciones pueden variar pero sólo en el caso que se esté midiendo efectos distintos.

Supongamos que para estudiar el efecto de la urbanización sobre la industrialización, tricotomizamos la primera variable en los niveles bajo, medio y alto. Por consiguiente, nuestra medición se hará sobre unidades geográficas que se clasificarán por urbanización, en una de las tres categorías. Esperamos que el nivel de industrialización correspondiente a cada categoría de la variable X, sea similar. Es decir, que los niveles de desarrollo industrial sólo se diferencien en términos de factores aleatorios.

El primer conjunto de variables mudas nos lleva a interpretar los valores de los estimadores mínimo-cuadráticos, como el efecto absoluto de cada una de las categorías de la variable urbanización. El segundo conjunto, muestra el impacto diferencial (o relativo) de los niveles medio y alto, en comparación al del nivel bajo. La elección de la base de comparación es arbitraria. Podríamos haber

elegido la categoría media o la alta, pero, lo que no es arbitrario es la definición de las variables para cada caso.

Es un ejercicio conveniente, para lograr un conocimiento más profundo, encontrar las variables mudas para cada una de estas situaciones.

3. El ajuste de un plano discreto.

Una vez que nuestro investigador ha aplicado la técnica expuesta en la sección 2, no se encuentra satisfecho con los resultados obtenidos. La variabilidad intraclase es demasiado grande y, por lo tanto, el coeficiente de determinación es bajo.

Planteado el problema, ha revisado su esquema teórico y después de un análisis detenido decide incluir una variable clasificatoria adicional.

En términos estadísticos esto significa que no se ha tenido éxito en definir la clase. La variabilidad interna es demasiado grande. La constitución del estrato depende de más de una variable.

La decisión del teórico es la de incluir la variable región. Su argumento básico es que el uso de los servicios necesarios para la producción, es una función de las disponibilidades de recursos de cada región. Para simplificar, la exposición, supongamos que este país se caracteriza porque la zona Norte es esencialmente

minera, la zona central es un complejo de actividades minero, agrícolas, de servicios, e industriales. La zona Sur es esencialmente agrícola.

En resumen, se pretende explicar o estimar el impacto de la urbanización condicionada por la región en la cual se desarrolla, sobre el nivel de industrialización. En este momento nuestro investigador necesita elaborar un poco más su esquema teórico.

Veremos que la técnica en sí plantea algunas interrogantes que reenvían sobre la teoría realizando preguntas muy específicas.

Esta es la línea argumental central que organiza esta sección del trabajo.

La forma típica de presentación de la información, es una tabla como la siguiente:

		URBANIZACION		
		Baja (B)	Media (M)	Alta (A)
R	Norte	Y_{11k_1}	Y_{12k_2}	Y_{13k_3}
E				
G				
I	Centro	Y_{21k_4}	Y_{22k_5}	Y_{23k_6}
O				
N				
E				
S	Sur	Y_{31k_7}	Y_{32k_8}	Y_{33k_9}

donde $k_i = 1, 2, \dots, n_i$
para todo $i = 1, 2, \dots, 9$

La variable Y simboliza el nivel de industrialización. El subíndice K, es variable por cuanto el número de observaciones por casillas es distinto y, es igual a n_1 , para la primera; n_2 para la segunda y así sucesivamente hasta llegar a n_9 para la última casilla. Además se cumple que:

$$\sum_{f=1}^g n_f = n$$

$$P=1$$

Para poder continuar el análisis es necesario preguntar al teórico respecto a la forma de concebir el impacto de las variables. Sus impactos son agregativos?; son interactivos? Si son interactivos, dónde o a qué niveles se produce la interacción?, etc. Es claro que las respuestas deben buscarse al nivel teórico, haciendo uso de todos los conocimientos disponibles. En todo caso, lo que nos interesa en este trabajo es exponer algunas alternativas de análisis y por consiguiente, supondremos que el técnico abre un abanico de posibilidades, tomando en cuenta sólo los aspectos de su incumbencia.

Caso 1: Un modelo agregativo simple.

Supongamos que se argumenta que el impacto de la urbanización sobre la industrialización es independiente de la región de que se trate y que lo mismo es válido para la región.

Podríamos replantear la tabla, considerando los efectos de las variables explicativas sobre la explicada:

		URBANIZACION		
		B	M	A
R E G I O N E S	N	β_1	$\beta_1 + \beta_4$	$\beta_1 + \beta_5$
	C	$\beta_1 + \beta_2$	$\beta_1 + \beta_2 + \beta_4$	$\beta_1 + \beta_2 + \beta_5$
	S	$\beta_1 + \beta_3$	$\beta_1 + \beta_3 + \beta_4$	$\beta_1 + \beta_3 + \beta_5$

β_1 , mide el nivel de industrialización esperado en las localidades con bajos niveles de urbanización y que, además, se encuentran situadas en la zona norte. Esta categoría es la que se usa como base de comparación.

Si restamos a la segunda línea la primera, nos encontramos con que para los tres pares de casillas, dicha diferencia es igual a β_2 . Por lo tanto, este parámetro mide el impacto relativo asociado a la categoría centro. Manteniendo constante el nivel de urbanización, esperamos un efecto igual a β_2 , por el hecho de que la localidad se encuentre en la zona central en lugar de la zona norte. Además, suponemos que este impacto es independiente del nivel de urbanización. El parámetro β_2 es el mismo para cualquier nivel de urbanización.

Una interpretación similar se puede realizar para β_3 . β_3 mide el efecto relativo de la zona sur sobre el nivel de industrialización, en que se toma como base de comparación la zona norte. β_4 , se puede obtener restando a la segunda columna la primera. Por lo tanto, mide el efecto relativo de la clase estadística urbanización media sobre el nivel de industrialización. Sin embargo, en este caso usamos como base de comparación el nivel bajo de urbanización. De otra parte, como β_4 , es el mismo para cualquier región, entonces el nivel de urbanización es independiente de la región.

El parámetro β_5 , se interpreta de manera similar que β_4 . Es decir, mide el impacto relativo del grado alto de urbanización,

sobre la industrialización. La base de comparación es el nivel bajo de la misma variable.

En esta forma de descomponer la variable explicada o a explicar, hemos realizado dos supuestos básicos: 1) la variable industrialización se puede descomponer en una serie de factores agregativos; 2) el impacto de la variable explicativa urbanización, es independiente del nivel de la variable regionalización. La aseveración inversa también es válida.

Estas afirmaciones son supuestos para el técnico. Sin embargo, para el teórico constituyen hipótesis a demostrar. Normalmente, estas hipótesis son o deben ser obtenidas a base del desarrollo del esquema teórico.

Una vez establecido el modelo que descompone la variable industrialización, en una serie de efectos simbolizados por los distintos valores de β , debemos proceder a la estimación de ellos.

Consideremos el siguiente modelo de regresión:

$$(3) \quad E(Y/X) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Si el primer elemento del par ordenado (i,j) simboliza las líneas y el segundo las columnas, entonces cuando una observación cae en la casilla $(1,1)$ las variables asumen los siguientes valores:

$$(4) \quad X_1 = 1; X_2 = 0; X_3 = 0; X_4 = 0; X_5 = 0$$

Al reemplazar estos valores en la ecuación (3) tenemos:

$$E(Y/X) = \beta_1(1) + \beta_2(0) + \beta_3(0) + \beta_4(0) + \beta_5(0)$$

Este resultado es idéntico con el de la casilla (1,1), de la tabla en que se descompone la variable urbanización en "efectos".

Por lo tanto, para cada unidad geográfica que cumpla con las características de nivel bajo de urbanización y, que geográficamente se ubique en la zona norte, las variables tomarán los valores expresados por (4):

Del mismo modo podemos desarrollar el conjunto típico de valores que tomarán las distintas variables para cada una de las casillas:

Casillas	X_1	X_2	X_3	X_4	X_5	Y
(1-1)	1	0	0	0	0	Y_{11}
(2-1)	1	1	0	0	0	Y_{12}
(3-1)	1	0	1	0	0	Y_{31}
(1-2)	1	0	0	1	0	Y_{12}
(2-2)	1	1	0	1	0	Y_{22}
(3-2)	1	0	1	1	0	Y_{32}
(1-3)	1	0	0	0	1	Y_{13}
(2-3)	1	1	0	0	1	Y_{23}
(3-3)	1	0	1	0	1	Y_{33}

Para realizar el ajuste mínimo-cuadrático disponemos de un conjunto n de valores de variables en que habrá n_1 iguales a la primera línea de la tabla, n_2 iguales a la segunda línea y así hasta n_g iguales a la de la última línea de la tabla.

Al tener el listado correspondiente a las variables explicativas y explicada, podemos aplicar los procedimientos tradicionales de estimación mínimo-cuadrática, obteniéndose de ese modo las estimaciones de los parámetros.

Caso 2: La interacción en modelos agregativos:

Consideraremos el caso en que no existe independencia entre las variables explicativas, urbanización y regionalización.

Supongamos que el teórico establece que para todas aquellas unidades geográficas que se encuentran en la zona central y, que poseen un nivel medio de urbanización, la variable industrialización debe descomponerse considerando no sólo los impactos individuales de cada variable, sino que también un efecto inducido por la conjunción de ambas categorías.

En este caso la tabla de descomposición de la variable explicada (industrialización) asume la siguiente forma:

		URBANIZACION		
		B	M	A
REGIONES	N	β_1	$\beta_1 + \beta_4$	$\beta_1 + \beta_5$
	C	$\beta_1 + \beta_2$	$\beta_1 + \beta_2 + \beta_4 + \beta_6$	$\beta_1 + \beta_2 + \beta_5$
	S	$\beta_1 + \beta_3$	$\beta_1 + \beta_3 + \beta_4$	$\beta_1 + \beta_3 + \beta_5$

La única diferencia con la tabla correspondiente para el caso 1, se encuentra en que hemos agregado un coeficiente β_6 , que

mide el impacto de la interacción entre las categorías, zona central y urbanización media, sobre la industrialización.

Como el modelo debe contemplar la inclusión de un nuevo parámetro debemos incluir una variable X_6 :

$$(4) \quad E(Y/X) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

El listado de observaciones típicas será:

Casillas	X_1	X_2	X_3	X_4	X_5	X_6	Y
(1-1)	1	0	0	0	0	0	Y_{11}
(2-1)	1	1	0	0	0	0	Y_{21}
(3-1)	1	0	1	0	0	0	Y_{31}
(1-2)	1	0	0	1	0	0	Y_{12}
(2-2)	1	1	0	1	0	1	Y_{22}
(3-2)	1	0	1	1	0	0	Y_{32}
(1-3)	1	0	0	0	1	0	Y_{13}
(2-3)	1	1	0	0	1	0	Y_{23}
(3-3)	1	0	1	0	1	0	Y_{33}

Cuando se trabaja el modelo de regresión con variables métricas, se acostumbra a representar la interacción por medio del producto de dos o más variables. Es conveniente notar que la variable X_6 , podría haber sido definida como $X_2 X_4$.

Siguiendo un procedimiento similar al expuesto se podría continuar considerando interacciones entre las distintas categorías. El límite estará dado por un modelo en que se considere interacciones entre todos los niveles de las variables. Si deseamos

trabajar con este orden de generalidad no hay razones para considerar el conjunto de observaciones como si fuesen una sola unidad. En este caso lo más conveniente es trabajar cada casilla de manera separada, utilizando la técnica expuesta en la sección 1.

Aun cuando las formas de descomposición de la variable explicada (industrialización) son múltiples y variadas, creemos haber establecido principios relativamente generales para construir modelos de análisis. Sin embargo, mostraremos algunos desarrollos alternativos en la siguiente sección.

4. Una Alternativa al Análisis Porcentual.

En la sección 3, hemos considerado el caso en que la variable explicada es métrica. En esta estudiaremos la forma de plantear el modelo cuando la variable Y es dicotómica y se transforma en una variable cuantitativa, al definirla en términos de proporciones o porcentajes. Además de plantear algunos modelos alternativos de análisis, los cuales deben ser entendidos como complementarios a los desarrollos en la sección anterior, consideraremos los problemas de estimación que son característicos en este tipo de modelos.

4.1. Un breve resumen de una teoría.

Para intentar explicar la migración rural-urbana se dispone de algunas ideas teóricas las cuales se pueden sintetizar en ideas relativas al grado de modernismo del campesino, consideraciones de carácter estructural y, por último, factores particulares de expulsión de mano de obra campesina.

La conclusión última del primer enfoque es que la tasa de migración es mayor a medida que mayor es el nivel de modernización del campesino. El enfoque estructural establece que la proporción de migrantes potenciales depende de las relaciones de producción y de las relaciones sociales que se establecen al interior de la explotación agrícola. En concreto, si ha existido un proceso de reforma agraria que ha producido cambios estructurales y observamos las tasas o proporciones de migración al interior del área reformada éstas serán sustancialmente menores que las de las explotaciones agrícolas tradicionales. La investigación del profesor Arguello^{1/} establece que el grado de modernización de los campesinos afecta a la migración en aquellos contextos sociales que no han sufrido cambios estructurales. Mientras que las variaciones estructurales juegan un rol fundamental en la explicación de la migración, cuando se ha llevado a cabo reales procesos de Reforma Agraria. En este caso, la modernización juega sólo un papel de carácter interactivo.

^{1/} Arguello, Omar: "Estructura agraria, participación y migraciones internas". Migración y Desarrollo, N°3, CLACSO, Buenos Aires, 1974.

Además de estas ideas, existen otras explicaciones en que se considera el papel explicativo de variables aisladas. Por ejemplo, se plantea que el nivel de ingreso o que la educación, etc., etc., juegan un "papel". Para nuestros objetivos consideraremos que la migración potencial puede ser explicada por las características estructurales de la explotación agrícola, por el grado de modernización del campesino y por el nivel de ingreso.

Para la explicación estructural distinguiremos dos categorías; a saber, explotaciones pertenecientes al área no reformada y reformada. La variable modernización será dicotomizada en moderno y no moderno y consideraremos tres niveles de ingreso: alto, medio y bajo. La variable a explicar es la proporción de migrantes potenciales.

4.2. Algunos modelos.

Si bien podríamos considerar como adecuados para el estudio de la migración potencial, algunos de los modelos establecidos en la sección 3 con el propósito de acopiar otras alternativas, consideremos el siguiente esquema de descomposición:

	Area No-Reformada			Area Reformada		
	Ingreso Bajo	Ingreso Medio	Ingreso Alto	Ingreso Bajo	Ingreso Medio	Ingreso Alto
Modernos	β_1	$\beta_1 + \beta_2$	$\beta_1 + \beta_3$	$\beta_1 + \beta_{21}$	$\beta_1 + \beta_2 + \beta_{21}$	$\beta_1 + \beta_3 + \beta_{21}$
No-Modernos	$\beta_1 + \beta_{12}$	$\beta_1 + \beta_2 + \beta_{12}$	$\beta_1 + \beta_3 + \beta_{12}$	$\beta_1 + \beta_{22}$	$\beta_1 + \beta_2 + \beta_{22}$	$\beta_1 + \beta_3 + \beta_{22}$

El parámetro β_1 mide el efecto absoluto de las categorías Area no reformada, nivel de ingreso bajo y Modernismo del campesino; sobre la migración potencial. El valor de él es usado como base de comparación.

β_2 y β_3 miden el impacto relativo de los niveles de ingreso medio y alto respectivamente sobre la migración potencial, y se caracterizan por ser independientes del tipo de explotación productiva agrícola.

β_{12} nos indica el efecto esperado sobre la migración debido a la "caída" del nivel de modernismo. Además, se supone que este impacto es distinto en función del contexto productivo.

Al interior del área reformada, se siguen manteniendo los efectos agregativos del nivel de ingresos. Sin embargo, el impacto de la categoría "área reformada" sobre la migración potencial es diferencial, dependiendo del nivel de modernismo del campesino. En efecto, la categoría área reformada, produce un impacto diferente si interactúa con la categoría modernos, que si lo hace con la de no moderno. En el primer caso dicho impacto es β_{21} y, en el segundo β_{22} .

Siguiendo un procedimiento similar al empleado en la sección 3 podemos establecer el modelo:

$$(5) \quad E\{Y/X\} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_4 + \beta_{21} X_5 + \beta_{22} X_6$$

en que el listado exhaustivo de las variables es:

Casillas	X_1	X_2	X_3	X_4	X_5	X_6	Y
(1,1)	1	0	0	0	0	0	Y_{11}
(2,1)	1	0	0	1	0	0	Y_{21}
(1,2)	1	1	0	0	0	0	Y_{12}
(2,2)	1	1	0	1	0	0	Y_{22}
(1,3)	1	0	1	0	0	0	Y_{13}
(2,3)	1	0	1	1	0	0	Y_{23}
(1,4)	1	0	0	0	1	0	Y_{14}
(2,4)	1	0	0	0	0	1	Y_{24}
(1,5)	1	1	0	0	1	0	Y_{15}
(2,5)	1	1	0	0	0	1	Y_{25}
(1,6)	1	0	1	0	1	0	Y_{16}
(2,6)	1	0	1	0	0	1	Y_{26}

Como en este caso estamos trabajando con las proporciones por casilla, en realidad tendremos tantas observaciones como casillas haya.

En toda la sección 3 trabajamos con un conjunto de n observaciones. Al considerar las proporciones asociadas a cada casilla, pasamos de una situación con n observaciones a una con un número de observaciones igual al total de casillas. En nuestro ejemplo tenemos 12.

De otra parte, el ajuste mínimo-cuadrático presenta algunas complicaciones que estudiaremos en 4.3.

Supongamos que el investigador descompone la proporción de migrantes potenciales según el siguiente esquema:

		Area No-Re formada		
		Ingreso Bajo	Ingreso Medio	Ingreso Alto
Modernos		$\beta_1 + \alpha_1 \gamma_1$	$\beta_1 + \beta_2 + \alpha_1 \gamma_1$	$\beta_1 + \beta_3 + \alpha_1 \gamma_1$
	No-Modernos	$\beta_1 + \alpha_1$	$\beta_1 + \beta_2 + \alpha_1$	$\beta_1 + \beta_3 + \alpha_1$

		Area Reformada		
		Ingreso Bajo	Ingreso Medio	Ingreso Alto
Modernos		$\beta_1 + \alpha_2 \gamma_1$	$\beta_1 + \beta_2 + \alpha_2 \gamma_1$	$\beta_1 + \beta_3 + \alpha_2 \gamma_1$
	No-Modernos	$\beta_1 + \alpha_2$	$\beta_1 + \beta_2 + \alpha_2$	$\beta_1 + \beta_3 + \alpha_2$

en que las interpretaciones para β_1 , β_2 y β_3 , son las mismas que en el caso anterior. La diferencia radica en los términos del tipo $\alpha\gamma$.

Según este modelo de descomposición de la migración potencial, el parámetro γ_1 , puede ser interpretado como el efecto del nivel de modernización del campesinado. Sin embargo, su impacto final depende del contexto productivo en que opera, ya que hemos supuesto que en el Area no Reformada actúa sobre un nivel α_1 , en tanto que en el Area Reformada, sobre un nivel α_2 , el cual se

caracteriza por ser distinto a α_1 . La diferencia entre casillas análogas, definidas en los dos tipos de contextos productivos, nos muestra que para el mismo nivel de modernización e ingresos, se cumple que $(\alpha_1 - \alpha_2) \gamma_1$, es decir, el impacto del tipo de organización productiva sobre la migración potencial está condicionado por el hecho de que el campesino presente características que lleven a catalogarlo de moderno.

El modelo se puede expresar como:

$$(6) \quad E \{Y/X\} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \alpha_1 \gamma_1 X_4 + \alpha_1 X_5 + \alpha_2 \gamma_1 X_6 + \alpha_2 X_7$$

y el listado exhaustivo de las variables será:

Casillas	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y
(1,1)	1	0	0	1	0	0	0	Y_{11}
(2,1)	1	0	0	0	1	0	0	Y_{21}
(1,2)	1	1	0	1	0	0	0	Y_{12}
(2,2)	1	1	0	0	1	0	0	Y_{22}
(1,3)	1	0	1	1	0	0	0	Y_{13}
(2,3)	1	0	1	0	1	0	0	Y_{23}
(1,4)	1	0	0	0	0	1	0	Y_{14}
(2,4)	1	0	0	0	0	0	1	Y_{24}
(1,5)	1	1	0	0	0	1	0	Y_{15}
(2,5)	1	1	0	0	0	0	1	Y_{25}
(1,6)	1	0	1	0	0	1	0	Y_{16}
(2,6)	1	0	1	0	0	0	1	Y_{26}

4.3. Problemas de Estimación.

Como nuestro interés se centra en la descomposición de una variable expresada en términos de proporciones, en función de ciertos factores determinantes, debemos tomar consciencia que

dicha variable no puede tomar valores que excedan la unidad ni que sean menores que cero. Para garantizar el cumplimiento de estos límites de variación para la variable explicativa, podemos restringirnos a utilizar funciones que cumplan con dichos requerimientos. Esta alternativa tiene un costo expresado en cantidad de funciones disponibles^{1/}.

Otra forma de lograr el mismo propósito es por medio de la transformación de la variable explicada de manera tal que desaparezcan las restricciones de recorrido. Hay varias alternativas disponibles, como la transformación probit^{2/}, la transformación tangente y la transformación logit^{3/}. En esencia estos tres tipos de transformaciones son equivalentes, por ello podemos centrarnos en la logit, sin pérdida de generalidad.

Sea P_{ij} , la probabilidad correspondiente a la casilla ubicada en la i -ésima línea y j -ésima columna. El logit de P_{ij} se define como:

$$L_{ij} = \log. \frac{P_{ij}}{1-P_{ij}}$$

donde la razón $P_{ij}/1-P_{ij}$, mide las chances a favor de la variable en cuestión. Por ejemplo si $P_{ij} = 0,8$ entonces $P_{ij}/1-P_{ij} = 4/1$. La chance a favor del suceso simbolizado por P , es de 4 a 1.

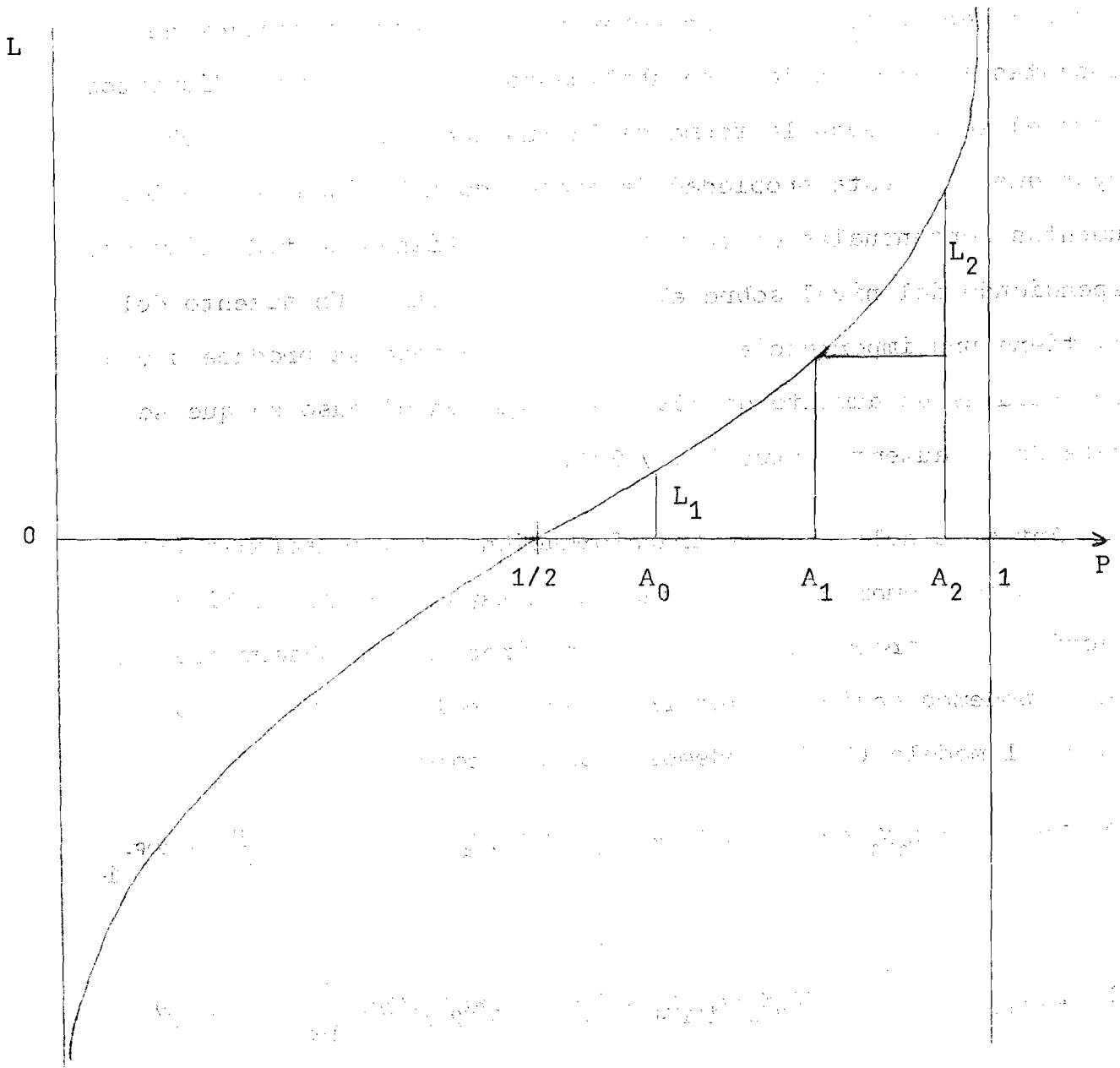
^{1/} Goldberger Arthur, "Econometric Theory". John Wiley, New York, 1964. Págs.222-224.

^{2/} Goldberger Arthur, *op.cit.*, págs.248-251.

^{3/} Theil Henry, "Statistical Decomposition Analysis". North Holland, Amsterdam, 1972, págs.166-173.

Si P se refiere a la proporción de migrantes potenciales dicho valor nos dice que por cada 5 campesinos hay 4 dispuestos a migrar, y sólo uno está en la situación contraria.

La función logit responde a la siguiente representación gráfica:



A partir de este gráfico debemos destacar dos hechos notables: al realizar la transformación han desaparecido las restricciones del recorrido de la variable. El logit varía entre más y menos infinito. De otra parte, los aumentos absolutos de la misma magnitud, en general, tienen efectos distintos sobre el logit. En el gráfico la distancia entre $1/2$ y A_0 , es igual a la distancia entre A_1 y A_2 . Sin embargo, el efecto de variaciones absolutas iguales en las probabilidades, tienen efectos distintos sobre el logit: dada la forma de la función, siempre L_2 será mayor que L_1 . Esta propiedad da contenido a la idea de que los aumentos porcentuales de un mismo tamaño, tienen sentido distinto dependiendo del nivel sobre el cual se aplican. Un aumento del 3%, tiene una importancia relativa menor cuando se produce a consecuencia de un aumento de 50% a 53%, que en el caso en que se trata de un aumento entre 90% y 93%.

Antes de aplicar esta transformación a las ecuaciones (5) y (6), debemos tomar en cuenta que en la mayoría de las aplicaciones prácticas las proporciones P , no son directamente observables, lo que si podemos medir son las frecuencias relativas h , por lo tanto, el modelo (5) lo podemos escribir como:

$$(7) \quad \log. \frac{h}{1-h} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_4 + \beta_{21} X_5 + \beta_{22} X_6 + (\log. \frac{h}{1-h} - \log. \frac{P}{1-P})$$

y en el (6):

$$(8) \quad \log. \frac{h}{1-h} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \alpha_1 \gamma_1 X_4 + \alpha_1 X_5 + \alpha_2 \gamma_1 X_6 + \alpha_2 X_7 + (\log. \frac{h}{1-h} - \log. \frac{P}{1-P})$$

Para aplicar los procedimientos tradicionales de estimación, debe cumplirse que la esperanza del término de error ($\log \frac{h}{1-h} - \log \frac{P}{1-P}$), debe ser igual a cero, su varianza debe ser constante y los errores no deben estar autocorrelacionados.

Siguiendo el argumento de Theil^{1/}, en que se supone que las observaciones de cada casilla han sido obtenidas independientemente, y que el número de observaciones por casilla sea lo suficientemente grande, se puede demostrar que la esperanza del término de error será cero. La ausencia de correlación entre los residuos está garantizada por haber considerado muestras independientemente. Sin embargo, hay dificultades con el supuesto de varianza constante (homocedasticidad).

Como se supone que en cada casilla se ha obtenido una serie de muestras aleatorias de distintos tamaños, en que se cumple el supuesto de independencia estadística y la característica de dicotomía, entonces podemos explicar el número de éxitos mediante el modelo probabilístico binomial. Se sabe que en este modelo la varianza de la proporción es igual a:

$$\text{Var}(h) = \frac{P(1-P)}{n}$$

Tanto los P como los tamaños de muestras por casilla (n), son distintos, entonces las varianzas de las proporciones para cada casilla serán distintas. Con este resultado se puede demostrar

^{1/} Theil, op.cit., págs.176-177.

que si por casillas disponemos de un número grande de observaciones, la varianza del término de error de los modelos (7) y (8) es aproximadamente igual a: $1/nh(1-h)$ tanto n como h se refieren a una sola casilla.

La teoría econométrica ha establecido que cuando se tiene este tipo de problemas (denominado de heterocedasticidad), el método más eficiente de ajuste es el mínimo cuadrático ponderado. La idea subyacente en este procedimiento de ajuste, es la de otorgar menos importancia a las observaciones con mayor varianza y más importancia a las con varianzas menores. La aplicación de este principio lleva al resultado de que las varianzas serán todas iguales.

Si en nuestros modelos multiplicamos las ecuaciones por: $\sqrt{nh(1-h)}$, donde hay tantas de estas expresiones como casillas, se llega a modelos homocedásticos con varianza unitaria.

Dado el carácter de este artículo sólo hemos bosquejado el problema debido a que es conocimiento básico de aquellas personas especializadas en este tipo de temas. Los no iniciados deben recurrir a un buen programa de computación. Nuestro interés se centra más bien en el problema de construcción de modelos que en los problemas de estimación.

5. Consideraciones técnicas adicionales.

El realizar el análisis de una tabla de contingencia a través del método de regresión presenta un conjunto apreciable de ventajas. Estas ventajas se encuentran fundamentalmente al nivel de una serie de conceptos calculables que enriquecen el análisis. En este párrafo pondremos interés especial en el coeficiente de determinación y en la prueba χ^2 .

Es sabido que en análisis de asociación existen una serie de coeficientes que intentan capturar la fuerza de la relación entre dos o más variables. El problema se presenta en el momento en que todos ellos entregan valores distintos y, a veces la diferencia es sustancial. Las personas que han trabajado sobre el tema parecen haber llegado al acuerdo de que los valores numéricos son distintos porque los coeficientes miden "cosas" diferentes^{1/}.

Al aplicar el método de regresión obtenemos como subproducto de la estimación de parámetros, el coeficiente de determinación el cual conceptualmente mide la proporción de la varianza total explicada por el modelo.

Sin embargo, es necesario que consideremos que el coeficiente de determinación es una función de la forma de la relación y del tipo de variables que se usan para realizar el ajuste. En concreto, lo que nos interesa es conocer el grado de ajuste entre los valores predichos por el modelo y los valores observados.

Para ello, recurrimos a la definición de R^2 , como:

1/ Ver, L. Goodman y W. Kruskal: Measures of Association for Cross Classification. Journal of the American Statistical Association. Vol.49, N°268, Diciembre, 1954. Págs.732-764.

$$R^2 = 1 - \frac{\sum e^2}{\sum (Y_i - \bar{Y})^2}$$

en que e se define como la discrepancia entre los valores observados y pronosticados.

Si simbolizamos la frecuencia estimada por \hat{h} , y operamos algebraicamente sobre la ecuación (7):

$$\frac{\hat{h}}{1-\hat{h}} = \exp.\{\hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_{12} X_4 + \hat{\beta}_{21} X_5 + \hat{\beta}_{22} X_6\}$$

despejando \hat{h} se llega a:

$$(9) \quad \hat{h} = (1 + \exp.\{\hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_{12} X_4 + \hat{\beta}_{21} X_5 + \hat{\beta}_{22} X_6\})^{-1}$$

Como $e = h - \hat{h}$, se dispone de todos los elementos para calcular R^2 .

Operando de la misma forma sobre la ecuación (8) se llega a:

$$(10) \quad \hat{h} = (1 + \exp.\{\hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\alpha}_1 \gamma_1 X_4 + \hat{\alpha}_1 X_5 + \hat{\alpha}_2 \gamma_1 X_6 + \hat{\alpha}_1 X_7\})^{-1}$$

Los modelos que hemos considerado no toman en cuenta el número de observaciones por casillas. Hemos reemplazado el total de observaciones por 12. El término de error estimado e (discrepancia entre valores observados y estimados), nos da una indicación respecto al grado de ajuste de nuestro modelo, pero no toma en cuenta el número de observaciones por casillas.

Se puede derivar un test χ^2 que cumpla con esta condición.

Uno de los teoremas elementales de la estadística matemática establece que si el tamaño de muestra es lo suficientemente grande, entonces la distribución binomial tiende a la normal.

De otra parte χ^2 , se define como una suma de variables aleatorias normales independientes elevadas al cuadrado, donde los grados de libertad son iguales al número de sumandos^{1/}.

Como hemos supuesto que al interior de cada casilla de nuestra tabla, se cumplen las condiciones necesarias para poder aplicar la distribución binomial, entonces para cada casilla, la expresión:

$$\frac{n(h-P)^2}{P(1-P)}$$

seguirá una distribución χ^2 , con un grado de libertad. Ahora bien, como a la vez hemos supuesto que las muestras se han obtenido independientemente entonces, las distintas χ^2 serán estadísticamente independientes y como en nuestro ejemplo tenemos doce casillas, entonces tendremos que la suma:

$$\chi_{1.1}^2 + \chi_{1.2}^2 + \chi_{1.3}^2 + \chi_{1.4}^2 + \chi_{1.5}^2 + \chi_{1.6}^2 + \chi_{2.1}^2 + \chi_{2.2}^2 + \chi_{2.3}^2 + \chi_{2.4}^2 + \chi_{2.5}^2 + \chi_{2.6}^2 = \chi^2$$

o bien,

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^6 \chi_{ij}^2$$

^{1/} Estos teoremas se encuentran en cualquier libro de estadística matemática. Una buena sistematización se expone en: John E. Freund: "Mathematical Statistics". Prentice-Hall, Ind. Englewood, N.J. 1962, Cap.8.

sigue una distribución χ^2 con doce grados de libertad.

Podríamos someter a prueba la consistencia entre las frecuencias relativas observadas \hat{h} y, las proporciones teóricas P , pero como no las conocemos y sólo disponemos de estimaciones (obtenidas por las ecuaciones del tipo (9) y (10)), podemos reemplazar los P por los h . Sin embargo, hay que corregir el cálculo de grados de libertad, por cuanto hay un teorema que nos dice que debemos restar un grado de libertad por cada parámetro que hayamos estimado^{1/}. En el caso del modelo (7) hemos estimado 6 parámetros, por lo tanto, tenemos 6 grados de libertad. En el caso del modelo (8) hemos estimado 7 parámetros, lo que origina 5 grados de libertad. En todos los demás aspectos esta es una prueba χ^2 tradicional, donde la región crítica se encuentra ubicada en la cola derecha de la distribución.

6. Conclusiones.

Las secciones 2 y 3, nos han permitido mostrar una equivalencia analítica entre el análisis de varianza y el análisis de regresión por variables mudas. Este resultado es ampliamente conocido en la literatura técnica^{2/}.

1/ Hoel Paul: "Introduction to Mathematical Statistics". John Wiley, New York, 1962, pág.250.

2/ Este resultado está en cualquier libro de Econometría. Una de las exposiciones más simples se encuentra en: Wonnacott and Wonnacott: "Econometrics". John Wiley, New York, 1970, págs.77-79.

En la sección 4 hemos abordado el mismo problema, con la única diferencia que hemos impuesto restricciones al recorrido de la variable explicada. En particular, nos hemos preocupado de la variable proporción. Para el tratamiento de las denominadas tablas de contingencia o tablas de entradas múltiples, se ha desarrollado una serie de instrumentos de análisis que cubren desde el análisis de asociación hasta el análisis de variables múltiples a la Lazarsfeld.

La limitación más importante de dichas técnicas radica en la imposibilidad de medir el efecto que tiene cada categoría sobre la variable explicada. Como hemos visto a lo largo de este trabajo, la ventaja fundamental del análisis de regresión con variables mudas radica en la descomposición de la variable explicada en un conjunto de impactos o efectos, asociados a las categorías de las variables explicativas. Además, estos efectos pueden ser independientes o interactivos.

Al realizar un análisis de regresión sobre una tabla de contingencia, podemos calcular no sólo los impactos de cada una de las categorías o de cada uno de los regresores, sino también intervalos de confianza, intervalos de predicción, etc.

Además, obtenemos una medida para la fuerza de la relación; el coeficiente de determinación R^2 . Este coeficiente se interpreta como el porcentaje de la varianza total que es explicada por el

modelo. En general, se abren las puertas para hacer uso de todas las construcciones elaboradas en torno al poderoso modelo de regresión.

La diferencia básica entre el modelo de regresión tradicional y el que usa variables mudas, radica en que en la mayoría de los casos el primero, permite medir efectos de variables, en tanto que el segundo sólo permite medir efectos de regresores. En todos los desarrollos que hemos presentado se ha establecido reglas de correspondencia entre regresores y categorías. A pesar de ello, podemos investigar la significación del efecto de una variable.

Supongamos que, al aplicar el modelo (5), nos preguntamos acerca de la significación de la variable ingreso. Para responder podemos establecer la hipótesis nula:

$$H_0 = \beta_2 + \beta_3 = 0$$

en contra de la alternativa:

$$H_A = \beta_2 + \beta_3 < 0 \quad \text{ó} \quad H_A = \beta_2 + \beta_3 > 0 \quad \text{ó} \quad H_A = \beta_2 + \beta_3 \neq 0$$

La respuesta depende del resultado de la prueba.

Sabemos que una de las limitaciones fundamentales en el análisis de tablas de contingencias viene dada por el número de observaciones. El número de casillas aumenta en progresión geométrica al número de variables cruzadas y al número de categorías

por variable. Esta restricción es común a todas las técnicas de análisis de contingencia y por lo tanto, es perfectamente aplicable al modelo de regresión con variables mudas. Sin embargo, el método de regresión impone una restricción adicional. En todo modelo de regresión uno de los conceptos que juega un papel central es el de grados de libertad, que se calcula como el número de observaciones menos el número de parámetros a estimar. En el análisis de la sección 4, hemos reemplazado las n observaciones originales por un número igual al de casillas. Por lo tanto, hemos sufrido una pérdida importante de grados de libertad. Esta disminución repercutirá a través de todo el modelo de regresión.

Por último, queremos destacar que aún cuando no hemos estudiado todas las posibilidades de descomposición de la variable explicada hemos entregado una serie de principios respecto a su construcción. En realidad el procedimiento que hemos utilizado contiene los siguientes pasos:

- 1) Descomponga la variable explicada en una suma de efectos.
Se supone que para ello se hace uso de la teoría.
- 2) Asocie a cada efecto o parámetro un regresor X .
- 3) Escriba el modelo. Es decir, plantee una ecuación que consista de parámetros y variables X .
- 4) Imponga valores numéricos a las distintas X , de manera que se puede realizar una asociación entre el conjunto de valores de X y cada casilla de la tabla de contingencia.

5) Una vez definido el modelo y conocidos los valores de las variables proceda al ajuste. Si la variable a explicar es una proporción, recuerde que debe usar el criterio de mínimo cuadrático ponderado y, la conveniencia de la transformación de la variable.

6) Finalmente, proceda al análisis de resultados. Tenga en cuenta que sus resultados no son independientes de su teoría.

II. PROBLEMAS DE ESTIMACION EN MODELOS CON REGRESORES MUDOS.

por Fernando Cortés.

1. Introducción.

En el capítulo anterior, hemos puesto especial énfasis sobre las potencialidades que se derivan de la formalización. En especial nos hemos dedicado a considerar expresiones matemáticas que permitan incorporar factores explicativos de naturaleza cualitativa. En este trabajo, nuestro interés central consiste en estudiar las dificultades que surgen de la estimación de modelos cualitativos. Aún más, sólo tomaremos en cuenta aquellas dificultades que sobrepasen los límites establecidos por el método mínimo cuadrático.

Con el propósito de especificar con algún detalle nuestro objetivo central consideremos el modelo cuya variable a explicar es el número de hijos por familia y en que las explicativas están compuestas por variables cualitativas y cuantitativas. Entre las primeras contamos con categorías ocupacionales y ciudades y, entre las segundas, con edad actual de la mujer, edad al casarse, socialización, etc., etc.

Si las variables definidas se refieren a observaciones individuales se nos plantean algunas alternativas de análisis.

De una parte, podemos cruzar las variables cualitativas de manera de definir un conjunto de subpoblaciones o submuestras y realizar un análisis de regresión entre las variables cuantitativas al interior de cada una de ellas.

La otra posibilidad consiste en la incorporación al interior del modelo de las variables cualitativas. Para ello se recurre a la definición de variables mudas. Una vez que establecemos estas variables los procedimientos de estimación no se modifican por cuanto, estamos en condiciones de calcular todas las medidas estadísticas que nos interesen.

La diferencia esencial que existe entre ambos tipos de modelos es que el segundo nos permite medir directamente el impacto que tienen las variables cualitativas sobre la variable explicada.

Desde el punto de vista gráfico, el primer tipo de modelos consiste en el ajuste independiente de un conjunto de hiperplanos de regresión. Uno para cada casilla definida por el cruce de las variables cualitativas. En cambio, el segundo es más restrictivo en la medida que considera un ajuste relacionado de los mismos. La asociación está dada por el hecho de que los distintos hiperplanos poseen las mismas inclinaciones y se diferencian, por sus niveles, son estos últimos los que recogen los impactos de las variables mudas.

Por lo tanto, el modelo absolutamente cuantitativo (en la medida que sólo incorpora dicho tipo de variables) se puede caracterizar por el hecho de que los coeficientes de regresiones variarán de subpoblación a subpoblación. El modelo cuantitativo-cualitativo impone la restricción de que los impactos de las variables métricas son los mismos para todas y cada una de las subpoblaciones y, que el efecto de las variables cualitativas sólo modifican el término libre.

Además, consideraremos un modelo absolutamente cualitativo (todos sus regresores son cualitativos y la variable explicada es cuantitativa). Este tipo de situación es tradicionalmente abordada por medio del análisis de varianza. No hemos seguido esta óptica debido a que las condiciones de construcción del instrumento de análisis son difícilmente cumplidas por el tipo de información de que disponemos. De otra parte, la relación que existe entre el modelo cualitativo y el análisis de varianza, ha sido ampliamente estudiada por la Econometría^{1/}.

2. Estimación de una Ecuación de Regresión con Datos

Individuales a partir de Datos Agregados.

Trataremos este problema como punto de partida, aún cuando no diga relación directa con los modelos construidos. Sin embargo, nos permite relevar y dar una solución fácil al tipo de problemas que debemos enfrentar posteriormente.

Supongamos que nos interesa ajustar un modelo lineal en que todas y cada una de las variables se han construido a partir de observaciones individuales.

En términos generales, podemos representar la relación entre el conjunto de observaciones y regresores por medio de:

^{1/} Este tema se encuentra tratado en la mayoría de los textos de Econometría. A nuestro entender los mejores desarrollos están en: Arthur Goldberger, "Econometric Theory", John Wiley, 1964. New York, págs.227-231 y en Wonnacott y Wonnacott, "Econometrics", John Wiley, 1970. New York, págs.77-80.

$$(1) \quad Y = XB + U$$

donde X es una matriz de orden (nxk) y de rango k (n es el número de observaciones y K es el número de parámetros a estimar), donde k es menor o igual que n. Y y U son vectores columnas de orden $(nx1)$ y B es un vector de orden $(kx1)$.

Además se agregan los supuestos tradicionales relativos al término de error, vale decir:

$$E(U) = 0$$

$$E(UU') = \sigma^2 I$$

Ahora bien, supongamos que en lugar de disponer de la información original sólo contamos con datos agregados y que conocemos las medias de X, Y y U simbolizadas por $(\bar{X}, \bar{Y}$ y $\bar{U})$. Si hemos formado m grupos $(k < m < n)$ podemos establecer el modelo:

$$(2) \quad \bar{Y} = \bar{X} B + \bar{U}$$

Como disponemos de la información para realizar el ajuste del modelo (2), estaremos en condiciones de estimar el vector de parámetros. Sin embargo, la conexión no es tan directa como parece en una primera aproximación. Para que los estimadores mínimo cuadráticos del segundo modelo cumplan con las propiedades Gauss-Markov, es necesario que estudiemos el procedimiento por medio del cual a partir de (1) se llega a (2).

Para ello definamos una matriz de agrupamiento P de orden (mxn) cuya característica es que cada línea se refiere a un grupo

y que, además, está compuesta por las frecuencias relativas al interior de cada grupo y por ceros. Las primeras aparecen cuando las observaciones se encuentran incluidas en el grupo y, los ceros cuando la observación no corresponde al grupo a que se refiere la línea. Tomemos como ejemplo la tercera línea de P, es decir, el tercer grupo. Si este agregado está formado por tres observaciones: la cuarta, séptima y n-ésima, entonces el tercer, séptimo y n-ésimo elemento de la tercera línea de P serán iguales a un tercio y todos los restantes serán iguales a cero.

Dada la matriz P estamos en condiciones de establecer las siguientes relaciones entre las variables individuales y agrupadas:

$$\bar{Y} = PY$$

$$\bar{X} = PX$$

$$U = PU$$

Dadas estas ecuaciones podemos reexaminar los supuestos relativos al vector estocástico de error:

$$E(\bar{U}) = PE(U) = 0$$

$$E(\bar{U}\bar{U}') = E(PUU'P') = \sigma^2 PP'$$

Resulta claro que no se introducen modificaciones en torno al primer supuesto, sin embargo, la matriz de varianzas y covarianzas de los errores se encuentra afectada. Dada la naturaleza de la matriz P el producto matricial PP' da origen a una matriz diagonal de orden $m \times m$. Como, en general, estos términos serán distintos entre sí, tenemos que la transformación lineal operada sobre las variables individuales ha generado un problema de heterocedasticidad,

el cual afecta a la propiedad de varianza mínima. La solución se encuentra en la utilización de los mínimos cuadráticos generalizados o procedimiento de Aitken^{1/}.

Al aplicar el método de estimación mínimo cuadrático generalizado las expresiones correspondientes al vector de estimadores (b) y a la matriz de varianzas y covarianzas {var(b)} son:

$$(3) \quad b = \{ \bar{X}' (PP')^{-1} \bar{X} \}^{-1} \bar{X}' (PP')^{-1} \bar{Y}$$

$$(4) \quad \text{Var}(b) = \sigma^2 \{ \bar{X}' (PP')^{-1} \bar{X} \}^{-1}$$

En resumen, si deseamos estimar la relación que existe entre un conjunto de variables individuales y sólo contamos con información grupal debemos tomar en cuenta que el agrupamiento destruye la propiedad de homocedasticidad que se ha supuesto al nivel individual. Debido a ello el procedimiento de estimación más adecuado es el mínimo cuadrático generalizado.

Antes de abandonar esta sección debemos realizar algunas acotaciones adicionales.

En la definición de la matriz de regresores X no se impone ninguna restricción sobre el tipo de variables, por lo tanto, puede estar compuesta tanto por variables métricas como por variables mudas, de modo que las conclusiones obtenidas son de carácter general.

^{1/} A.C. Aitken: "On Least-Squares and linear combinations of observations". Pro. Royal. Soc., Vol. 55, págs.42-48, 1934.

Al aplicar el procedimiento de estimación recomendado, se genera un problema de predicción cuya solución se debe a A. Goldberger^{1/}. Este autor demuestra que el mejor predictor lineal insesgado (mejor en el sentido varianza mínima), en aquellos casos en que se ha usado el procedimiento de Aitken es:

$$(5) \hat{Y} = X_0 b + W' (P P')^{-1} U$$

donde X_0 es un vector línea que contienen los regresores de predicción. W , es un vector columna definido por:

$$W = E(U_0 U) = \begin{bmatrix} E(U_0 U_1) \\ E(U_0 U_2) \\ \vdots \\ E(U_0 U_n) \end{bmatrix}$$

en que U_0 es el residuo estocástico correspondiente al vector de predicción X_0 . El vector columna U es el que contiene los errores correspondientes a la aplicación del método de estimación generalizado.

^{1/} Referencia tomada de J. Johnston: "Econometric Methods". John Wiley. New York, 1972. Págs. 212-213.

3. Estimación de un Modelo Agregativo y Cualitativo.

Nos interesa estimar efectos de variables cualitativas sobre variables cuantitativas agregadas. Supongamos que nuestro problema se centra en la estimación de un modelo de regresión en que para explicar el número medio de hijos utilizamos como variables explicativas categorías ocupacionales y ciudades importantes de algunos países latinoamericanos.

Ahora bien, por medio del cruce de las variables cualitativas podemos construir la tabla:

CIUDADES	OBREROS NO ESPECIALIZ.	OBREROS ESPECIALIZADOS	EMPLEADOS	DIRECTIVOS
GUAYAQUIL	α	$\alpha + \beta_1$	$\alpha + \beta_1 + \beta_2$	$\alpha + \beta_1 + \beta_2 + \beta_3$
CARACAS	$\alpha + \gamma_1$	$\alpha + \beta_1 + \gamma_1$	$\alpha + \beta_1 + \beta_2 + \gamma_1$	$\alpha + \beta_1 + \beta_2 + \beta_3 + \gamma_1$
RIO DE JANEIRO	$\alpha + \gamma_1 + \gamma_2$	$\alpha + \beta_1 + \gamma_1 + \gamma_2$	$\alpha + \beta_1 + \beta_2 + \gamma_1 + \gamma_2$	$\alpha + \beta_1 + \beta_2 + \beta_3 + \gamma_1 + \gamma_2$
BUENOS AIRES	$\alpha + \gamma_1 + \gamma_2 + \gamma_3$	$\alpha + \beta_1 + \gamma_1 + \gamma_2 + \gamma_3$	$\alpha + \beta_1 + \beta_2 + \gamma_1 + \gamma_2 + \gamma_3$	$\alpha + \beta_1 + \beta_2 + \beta_3 + \gamma_1 + \gamma_2 + \gamma_3$

en cuyo interior tenemos la descomposición teórica del número medio de hijos por familia.

Todos los coeficientes de esta tabla exceptuando α , miden impactos relativos en que, por construcción del modelo, las categorías tienen efectos independientes. Es decir, no se consideran interacciones. La no consideración de dependencia entre las categorías de clasificación no debe interpretarse como una limitación

inherente al uso de variables mudas. Sólo significa que no las hemos considerado en el modelo.

α , mide el número medio esperado de hijos de las familias cuya cabeza es obrero no especializado, y residente de Guayaquil.

β_1 , muestra el impacto que tiene sobre la variable explicada el hecho de que el jefe de familia sea obrero especializado en lugar de no especializado.

β_2 , mide el efecto de ser empleado en lugar de obrero especializado.

β_3 , indica el impacto que tiene sobre el número medio de hijos el hecho de que el jefe de familia sea directivo en lugar de empleado.

Los coeficientes γ se refieren a los efectos relativos de ciudades. De este modo, γ_1 , γ_2 , γ_3 , muestran los impactos de residir en Caracas en lugar de Guayaquil; en Río de Janeiro en vez de Caracas y en Buenos Aires en lugar de hacerlo en Río de Janeiro, respectivamente.

Una vez planteada la descomposición teórica del número medio de hijos debemos proceder a la asignación de un conjunto de variables mudas o regresores que nos permita movernos a través de las distintas casillas de la tabla.

Los regresores definidos son los siguientes:

Casillas	X ₀	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	Y (Nº Medio de Hijos)
(1,1)	1	0	0	0	0	0	0	\bar{Y}_{11}
(1,2)	1	1	0	0	0	0	0	\bar{Y}_{12}
(1,3)	1	1	1	0	0	0	0	\bar{Y}_{13}
(1,4)	1	1	1	1	0	0	0	\bar{Y}_{14}
(2,1)	1	0	0	0	1	0	0	\bar{Y}_{21}
(2,2)	1	1	0	0	1	0	0	\bar{Y}_{22}
(2,3)	1	1	1	0	1	0	0	\bar{Y}_{23}
(2,4)	1	1	1	1	1	0	0	\bar{Y}_{24}
(3,1)	1	0	0	0	1	1	0	\bar{Y}_{31}
(3,2)	1	1	0	0	1	1	0	\bar{Y}_{32}
(3,3)	1	1	1	0	1	1	0	\bar{Y}_{33}
(3,4)	1	1	1	1	1	1	0	\bar{Y}_{34}
(4,1)	1	0	0	0	1	1	1	\bar{Y}_{41}
(4,2)	1	1	0	0	1	1	1	\bar{Y}_{42}
(4,3)	1	1	1	0	1	1	1	\bar{Y}_{43}
(4,4)	1	1	1	1	1	1	1	\bar{Y}_{44}

El mejor procedimiento de estimación del modelo:

$$(7) \quad \bar{Y} = XB + \bar{U}$$

es el mínimo cuadrático generalizado, por cuanto debemos tomar en cuenta que las varianzas de las medias muestrales para cada una de las casillas serán, en general, distintas. Por consiguiente no podemos mantener el supuesto de homocedasticidad^{1/}.

^{1/} En el modelo de regresión se cumple que la varianza del término de error es igual a la varianza de la variable explicada; sea,

$$E(Y/X_1, X_2, \dots, X_n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

(Continúa nota en pág. siguiente)

Las varianzas estimadas de la variable a explicar están dadas por la fórmula:

$$(8) \quad \text{Var}(\bar{Y}_{ij}) = \frac{S_{ij}^2}{n_{ij}} \left(1 - \frac{n_{ij}}{N_{ij}}\right)$$

en que S_{ij}^2 es varianza muestral de las observaciones incluidas en la (i,j) ésima casilla, n_{ij} es el tamaño de la subpoblación.

Estas varianzas estimadas se pueden disponer en una matriz diagonal:

$$V = \begin{bmatrix} \text{Var } \bar{Y}_{11} & 0 & 0 & \dots & 0 \\ 0 & \text{Var } \bar{Y}_{22} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \text{Var } \bar{Y}_{nn} \end{bmatrix}$$

1/ (Continuación) El valor esperado de Y, dado el conjunto de regresores X. Además, tenemos:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + U$$

Restando la primera de la segunda ecuación:

$$Y - E(Y/X_1, X_2, \dots, X_n) = U$$

$$E \{Y - E(Y/X_1, X_2, \dots, X_n)\}^2 = E(U)^2$$

El término de la izquierda es la varianza de la variable explicada y el de la derecha es la varianza del término de error ya que:

$$E(U) = 0$$

Por lo tanto: $\text{Var}(Y) = \text{Var}(U)$

En este caso la fórmula del vector de estimadores y de la matriz de varianzas y covarianzas son:

$$(9) \quad b = (X' V^{-1} X)^{-1} (X' V^{-1} Y) \quad (9)$$

$$(10) \quad \text{Var}(b) = (X' V^{-1} X)^{-1}$$

Los elementos de la diagonal principal de la matriz de varianzas y covarianzas son los cuadrados de los errores estándares de los estimadores. De este modo, la raíz cuadrada del primer término de la diagonal de V entrega el error estándar de $\hat{\alpha}$, el segundo el de $\hat{\beta}_1$ y el séptimo el de $\hat{\gamma}_3$.

Los elementos de la matriz de varianzas y covarianzas se determinan sobre la base de las fórmulas correspondientes a la varianza de la variable explicada. Así, si la variable que pretendemos explicar son proporciones en lugar de medias aritméticas debemos utilizar:

$$(11) \quad \text{Var}(p) = \frac{pq}{n}$$

De otra parte, debemos agregar que en aquellos casos en que la variable explicada tiene limitaciones en su recorrido es recomendable someterla a transformaciones. Entre las transformaciones más utilizadas se encuentra la logit, probit y tangente.

En resumen, el procedimiento que hemos utilizado en esta sección se puede describir en los siguientes términos:

- 1) Disponemos de un cuadro que contiene 16 observaciones. Siguiendo una de las normas básicas del trabajo estadístico,

planteamos un modelo de análisis que permita describir de manera resumida los tipos de conexiones entre las variables, originándose de este modo, un conjunto sintético de medidas. Estas son esencialmente los coeficientes de regresiones, los cuales nos permiten discernir conceptualmente los efectos de las variables cualitativas.

2) Al establecer un modelo en los términos planteados, se generan algunos problemas de estimación, debido a que se rompe el supuesto de homocedasticidad. Por lo tanto, debemos recurrir al procedimiento mínimo cuadrático generalizado que equivale a una transformación de las variables.

3) Si bien el modelo que hemos desarrollado no impone, desde el punto de vista teórico ninguna restricción en el recorrido de la variable explicada, no existen mayores dificultades al considerar variables con recorrido limitado. Para ello es recomendable someter previamente dicha variable a alguna transformación.

4. Modelos Mixtos.

El modelo que orienta la exposición de esta sección ha sido calificado como mixto debido a que las variables explicativas son tanto cualitativas como cuantitativas.

La variable a explicar es el número de hijos por familia, las variables explicativas cualitativas son la categoría ocupacional del jefe de familia y algunas de las principales ciudades latino-

americanas. Las cuantitativas son: edad actual de la mujer, edad de la mujer al casarse, socialización urbana de la pareja, feminismo, religiosidad, tipo de matrimonio y participación ocupacional de la mujer.

El análisis se desarrolla al nivel de observaciones individuales las cuales se disponen en una tabla de doble entrada construida a base de las variables cualitativas. Al interior de cada una de las casillas tenemos un conjunto de observaciones individuales. Este conjunto está definido por la variable número de hijos por familia y por las variables explicativas métricas. El número de observaciones por casillas (n_{ij}) es igual al tamaño de cada una de las submuestras.

El modelo teórico de descomposición y la asignación del conjunto de variables mudas se presenta en las páginas siguientes. Las variables cuantitativas se han denotado por Z. De esta forma, Z_1 define la edad actual de la mujer, Z_2 la edad de la mujer al casarse y así sucesivamente hasta llegar a Z_7 ; participación ocupacional de la mujer.

El modelo escrito de manera explícita es:

$$(12) \quad Y = \alpha X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \gamma_2 X_8 + \gamma_3 X_9 + \gamma_4 X_{10} + \gamma_5 X_{11} + \gamma_6 X_{12} + \gamma_7 X_{13} + \gamma_8 X_{14} + \gamma_9 X_{15} + \gamma_{10} X_{16} + \delta_1 Z_1 + \delta_2 Z_1^2 + \delta_3 Z_2 + \delta_4 Z_3 + \delta_5 Z_4 + \delta_6 Z_5 + \delta_7 Z_6 + \delta_8 Z_7 + U$$

En que X_1, X_2, \dots, X_9 denotan las variables mudas que se construyen con el propósito de incorporar en la explicación el

papel que teóricamente jugarán las ciudades y los grupos laborales.

De otra parte Z simboliza variables métricas.

α , mide el número esperado de hijos para los obreros no especializados, residentes en Guayaquil. Los coeficientes gamas se refieren al efecto de "ciudades" tomando como base de comparación Guayaquil. De este modo γ_2 mide el impacto que tiene sobre el número de hijos el hecho de que el jefe de familia resida en Quito en lugar de Guayaquil y así sucesivamente hasta llegar a γ_{10} que muestra el efecto que tiene sobre la variable explicada el hecho de residir en Buenos Aires en comparación a Guayaquil.

Los coeficientes betas miden el impacto de la categoría ocupacional tomando como base de comparación fija la categoría de obreros no especializados. Por ejemplo, el coeficiente β_5 mide el efecto relativo que tiene sobre el número de hijos el hecho de que el jefe de familia sea empleado en vez de obrero no especializado.

Los coeficientes delta representan el impacto lineal que tienen las variables métricas sobre el número de hijos. Esta interpretación es válida para todos los deltas a excepción de δ_1 y δ_2 , por cuanto el efecto de la variable edad de la mujer (Z_1) es no lineal. El efecto de Z_1 sobre Y , manteniendo constantes todos los otros regresores es:

$$\frac{\delta Y}{\delta Z_1} = \delta_1 + 2\delta_2 Z_1$$

es decir, su efecto depende del nivel de Z_1 y de los dos parámetros deltas, en que δ_1 muestra la parte constante de la variación y δ_2 , la mitad de la tasa de cambio del mismo (aceleración). La variación de la tasa de cambio es constante e igual a $2\delta_2$, como se puede apreciar por medio de:

$$\frac{\delta^2 Y}{\delta Z_1^2} = 2\delta_2$$

Una vez descrito el modelo debemos preocuparnos por su estimación. Con este objeto escribamos el modelo (12) de manera compacta haciendo uso de matrices:

$$(13) \quad Y = W B + U$$

W es una matriz de orden $n \times k$ y de rango completo, en cuyas columnas podemos distinguir los regresores del tipo X y los del tipo Z . n es el número de observaciones y k es el número de parámetros.

Y y U son vectores columnas de orden $n \times 1$, y B es el vector columna que contiene los parámetros y su orden es $k \times 1$.

Si a este modelo agregamos los supuestos tradicionales respecto al término de error:

$$E(U) = 0$$
$$E(UU') = \sigma^2 I$$

completamos su especificación teórica.

A pesar de que en este momento estamos en condiciones de proceder al ajuste, puede resultar de cierto interés plantear el mismo

LISTADO DE VARIABLES MUDAS

CASILLAS	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	X ₁₁	X ₁₂	X ₁₃	X ₁₄	X ₁₅	X ₁₆
(1,1)	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(1,2)	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
(1,3)	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
(1,4)	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
(1,5)	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
(1,6)	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
(1,7)	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
(1,8)	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
(1,9)	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
(1,10)	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
(2,1)	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(2,2)	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
(2,3)	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
(2,4)	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
(2,5)	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
(2,6)	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
(2,7)	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
(2,8)	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
(2,9)	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
(2,10)	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
(3,1)	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
(3,2)	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
(3,3)	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
(3,4)	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
(3,5)	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
(3,6)	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
(3,7)	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
(3,8)	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
(3,9)	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
(3,10)	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
(4,1)	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
(4,2)	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0

(Continúa en pág. siguiente)

SYSTEM RELIABILITY AND OPTIMALITY

α_1^1	α_1^2	α_1^3	α_1^4	α_1^5	α_1^6	α_1^7	α_1^8	α_1^9	α_1^{10}	α_1^{11}	α_1^{12}	α_1^{13}	α_1^{14}	α_1^{15}	α_1^{16}	RELIABILITY
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,1)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,2)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,3)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,4)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,5)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,6)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,7)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,8)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,9)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,10)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,11)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,12)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,13)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,14)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,15)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,16)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,17)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,18)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,19)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,20)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,21)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,22)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,23)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,24)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,25)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,26)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,27)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,28)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,29)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,30)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,31)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,32)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,33)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,34)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,35)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,36)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,37)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,38)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,39)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,40)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,41)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,42)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,43)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,44)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,45)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,46)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,47)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,48)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,49)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(1,50)

TABLA DE DESCOMPOSICION TEORICA

CIUDADES GRUPOS OCUPACIONALES	GUAYAQUIL	QUITO	C. GUA TEMALA	C. MEXICO	SAN JOSE	CARACAS	BOGOTA	RIO DE JANEIRO	C. PA JAMA	BUENOS AIRES
OBRERO NO ESPECIALIZADO	α	$\alpha + \gamma_2$	$\alpha + \gamma_3$	$\alpha + \gamma_4$	$\alpha + \gamma_5$	$\alpha + \gamma_6$	$\alpha + \gamma_7$	$\alpha + \gamma_8$	$\alpha + \gamma_9$	$\alpha + \gamma_{10}$
ARTESANOS	$\alpha + \beta_2$	$\alpha + \gamma + \beta_{2 \ 2}$	$\alpha + \gamma + \beta_{3 \ 2}$	$\alpha + \gamma + \beta_{4 \ 2}$	$\alpha + \gamma + \beta_{5 \ 2}$	$\alpha + \gamma + \beta_{6 \ 2}$	$\alpha + \gamma + \beta_{7 \ 2}$	$\alpha + \gamma + \beta_{8 \ 2}$	$\alpha + \gamma + \beta_{9 \ 2}$	$\alpha + \gamma + \beta_{10 \ 2}$
OBREROS ESPECIALIZADOS	$\alpha + \beta_3$	$\alpha + \gamma + \beta_{2 \ 3}$	$\alpha + \gamma + \beta_{3 \ 3}$	$\alpha + \gamma + \beta_{4 \ 3}$	$\alpha + \gamma + \beta_{5 \ 3}$	$\alpha + \gamma + \beta_{6 \ 3}$	$\alpha + \gamma + \beta_{7 \ 3}$	$\alpha + \gamma + \beta_{8 \ 3}$	$\alpha + \gamma + \beta_{9 \ 3}$	$\alpha + \gamma + \beta_{10 \ 3}$
SERVICIO INDEPENDIENTE	$\alpha + \beta_4$	$\alpha + \gamma + \beta_{2 \ 4}$	$\alpha + \gamma + \beta_{3 \ 4}$	$\alpha + \gamma + \beta_{4 \ 4}$	$\alpha + \gamma + \beta_{5 \ 4}$	$\alpha + \gamma + \beta_{6 \ 4}$	$\alpha + \gamma + \beta_{7 \ 4}$	$\alpha + \gamma + \beta_{8 \ 4}$	$\alpha + \gamma + \beta_{9 \ 4}$	$\alpha + \gamma + \beta_{10 \ 4}$
EMPLEADOS	$\alpha + \beta_5$	$\alpha + \gamma + \beta_{2 \ 5}$	$\alpha + \gamma + \beta_{3 \ 5}$	$\alpha + \gamma + \beta_{4 \ 5}$	$\alpha + \gamma + \beta_{5 \ 5}$	$\alpha + \gamma + \beta_{6 \ 5}$	$\alpha + \gamma + \beta_{7 \ 5}$	$\alpha + \gamma + \beta_{8 \ 5}$	$\alpha + \gamma + \beta_{9 \ 5}$	$\alpha + \gamma + \beta_{10 \ 5}$
EMPLEADORES	$\alpha + \beta_6$	$\alpha + \gamma + \beta_{2 \ 6}$	$\alpha + \gamma + \beta_{3 \ 6}$	$\alpha + \gamma + \beta_{4 \ 6}$	$\alpha + \gamma + \beta_{5 \ 6}$	$\alpha + \gamma + \beta_{6 \ 6}$	$\alpha + \gamma + \beta_{7 \ 6}$	$\alpha + \gamma + \beta_{8 \ 6}$	$\alpha + \gamma + \beta_{9 \ 6}$	$\alpha + \gamma + \beta_{10 \ 6}$
DIRECTIVOS	$\alpha + \beta_7$	$\alpha + \gamma + \beta_{2 \ 7}$	$\alpha + \gamma + \beta_{3 \ 7}$	$\alpha + \gamma + \beta_{4 \ 7}$	$\alpha + \gamma + \beta_{5 \ 7}$	$\alpha + \gamma + \beta_{6 \ 7}$	$\alpha + \gamma + \beta_{7 \ 7}$	$\alpha + \gamma + \beta_{8 \ 7}$	$\alpha + \gamma + \beta_{9 \ 7}$	$\alpha + \gamma + \beta_{10 \ 7}$

modelo desde otro ángulo.

Si a la matriz W la particionamos distinguiendo las variables mudas y las cuantitativas y realizamos la misma operación sobre el vector de parámetros en que distinguimos, de una parte, los coeficientes betas y gamas, y de otra, los deltas, tendremos:

$$(14) \quad Y = Z \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + U$$

Realizando la operación del producto matricial indicada, concluimos que el sistema de ecuaciones representado por (13) se transforma en:

$$(13) \quad Y = Z \beta_1 + X \beta_2 + U$$

Sabemos que la aplicación del criterio mínimo cuadrático a la ecuación (13) entrega por resultados:

$$(16) \quad b = (W'W)^{-1} W' Y$$

Realizando operaciones convenientes sobre (16):

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} Z'Z & Z'X \\ X'Z & X'X \end{bmatrix}^{-1} \begin{bmatrix} Z'Y \\ X'Y \end{bmatrix}$$

Una vez que se invierte la matriz particionada y después de un tedioso pero no difícil proceso de manipulación algebraica obtenemos fórmulas que nos permiten estimar por separado los coeficientes de regresión asociados a las variables mudas y a los regresores métricos:

UNIVERSIDAD DE LOS ANDES

$$(17) \quad \begin{cases} 17.1 & b_1 = (Z'Z)^{-1} Z'NY \\ 17.2 & b_2 = (X'MX)^{-1} X'MY \end{cases}$$

En que M y N se definen como:

$$M = I - Z(Z'Z)^{-1} Z'$$
$$N = I - X(X'MX)^{-1} X'M$$

Hemos explicitado uno de los modelos mixtos que nos permiten realizar análisis de variables demográficas consideradas como variables a explicar. Sin embargo, los principios que hemos establecido son de carácter absolutamente general.

Es evidente que en el modelo expuesto en esta sección pueden aparecer algunos problemas econométricos que dificultan el proceso de estimación, por ello en la próxima sección haremos una breve discusión de aquellos que a nuestro juicio tienen una mayor probabilidad de surgir en investigaciones del tipo que hemos expuesto.

5. Algunos Problemas Adicionales.

En esta sección recopilamos una serie de problemas susceptibles de presentarse en el tratamiento de los modelos cualitativos agregados y mixtos.

Además de plantear la denominada "trampa de las variables mudas", dedicamos espacio a las formas específicas que pueden

asumir la multicolinealidad y la heterocedasticidad. Agregamos a ello, las implicaciones que se derivan de la presencia de dichos problemas de manera que el investigador tenga medios que le permitan reconocerlos y detectarlos, así como también, mostramos algunas alternativas para su solución.

Agregamos algunas pruebas de hipótesis que pueden ser utilizadas en el caso de modelos cualitativos agregados. Si bien, la teoría que los sustenta no es privativa de ellos, hay razones de tipo práctico que nos impulsan a presentarlas en ese contexto.

Por último, nos preocupamos por la forma correcta de interpretar los parámetros que afectan a las variables mudas y, por la manera adecuada de proceder al cálculo del coeficiente de determinación.

5.1. La trampa de las variables mudas.

En los modelos que incluyen variables mudas no hemos considerado explícitamente un término libre. Por ejemplo, en la ecuación (12) se puede apreciar que todos y cada uno de los parámetros se encuentran afectados por una variable X . Si hubiésemos incluido explícitamente el coeficiente de posición se generaría un problema de multicolinealidad perfecta. Agregar la constante implica adicionar un regresor que siempre asume el valor unitario. Como el regresor X_1 siempre toma el valor uno, tendríamos como consecuencia que en la matriz de regresores aparecerían dos columnas idénticas y, por lo tanto, su inversa no existiría.

Si bien este problema se presenta con mucha claridad al intentar plantear el modelo teórico y es susceptible de ser eliminado con facilidad, desde el punto de vista computacional pueden surgir algunas complicaciones que nos obligan a estar alertas.

En algunos programas de computación no existe la opción de incluir o no un término libre. Por lo tanto, si disponemos de un programa que incorpora automáticamente una columna de valores unitarios en la matriz de observaciones X , y el usuario alimenta a la máquina con todos los valores de los regresores - incluido X_1 , que se caracteriza por asumir sólo el valor unitario -, entonces habrá dificultades en la estimación. En efecto, el programa es de naturaleza tal, que a los K regresores se le agrega automáticamente una columna de 1, lo que conduce a que, por una parte, dispongamos de las estimaciones correspondientes a $K+1$ parámetros - en circunstancias que sólo se han definido teóricamente K de ellos -, y, por otra, se materialice un problema de rango en la matriz de observaciones. Esta dificultad genera lo que, en la literatura econométrica se conoce con el nombre de colinealidad, o de multicolinealidad.

Ahora bien, si el modelo a ser ajustado sólo incorpora regresores mudos, el problema se detecta fácilmente a posteriori. El computador no entregará resultados y nos alertará respecto a la singularidad de la matriz de observaciones.

Sin embargo, si el modelo es tal que combina variables mudas con variables métricas, suele suceder de que a pesar de haberse

cometido el error señalado podemos obtener resultados relativos a todas las medidas que nos interesan: estimaciones de coeficientes de regresiones, R^2 , errores estándares de los parámetros, etc., etc. Esto se explica debido a que en el cálculo de la matriz inversa, se recurre a una serie de aproximaciones de manera tal que se obtienen resultados a pesar de la singularidad. Pero, si se pide el cálculo del producto $(X'X)^{-1}(X'X)$ acontece que el resultado dista bastante de la matriz identidad.

Si el investigador no es cuidadoso en la selección de insumos que se entregan al computador, puede no darse cuenta de que algo anda mal. El peligro es aún mayor en aquellos casos en que se toman los resultados y se comienza el análisis sin detenerse previamente en la consideración de los problemas econométricos fundamentales.

A modo de resumen, podemos afirmar que si pretendemos estimar ecuaciones del tipo de la relación (12) y el programa de computación produce automáticamente el coeficiente de posición, el listado de variables debe excluir los regresores del tipo de X_1 .

En todo caso, siempre es conveniente investigar respecto a multicolinealidad, considerando los frecuentes errores de entrada de datos y la colinealidad real que puede existir entre las variables explicativas métricas.

5.2. Algunas consideraciones sobre multicolinealidad.

En el tipo de modelos que estamos tratando, este problema puede surgir desde varias fuentes.

En los modelos cualitativos se puede generar a raíz de una mala asignación de valores a las variables mudas^{1/}. En los modelos mixtos se puede presentar a partir de problemas computacionales, o bien debido a la relación lineal estrecha que se pueda establecer entre algunos de los regresores métricos.

En todo caso, lo que importa en una primera aproximación es reconocer la existencia del problema. Cuando hay multicolinealidad perfecta la matriz que se debe invertir para obtener los estimadores de los parámetros y la matriz de varianzas y covarianzas es de rango incompleto. Esto significa que la inversa no existe y, por lo tanto, no es posible obtener las estimaciones.

Este caso se relativiza bastante en la medida en que la multicolinealidad no es perfecta. Cuando ello ocurre es posible obtener los valores de los estimadores de los parámetros, sus varianzas, calcular el coeficiente de determinación, etc., etc. Pero, debido a la relación lineal estrecha las varianzas de los estimadores resultan ser exageradamente grandes. Generalmente, los resultados son hasta cierto punto contradictorios ya que de una parte, tenemos coeficientes de regresiones estimados no significativos; y de otra, suele aparecer un coeficiente de determinación grande.

^{1/} Cortés Fernando, ver artículo anterior, en esta misma publicación.

La solución econométrica a este problema radica en acopiar información externa al modelo que permita establecer una relación de proporcionalidad entre los parámetros que están asociados a las variables colineales. Si dicha relación es conocida por medio de simples sustituciones algebraicas se puede evitar las dificultades.

Además, consignamos que la solución de uso cotidiano en la investigación empírica radica en la eliminación de una de las variables colineales. Esta forma de abordaje del problema no es recomendable en la medida que se presentan dificultades en la asignación de contenido a los parámetros^{1/}.

1/ Para evaluar las consecuencias de una eliminación de una variable explicativa, en caso de relación lineal estrecha entre los regresores, consideremos el modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

Ahora bien, partamos del supuesto que la teoría que origina esta ecuación nos señala que ambas son variables "relevantes" y que estamos interesados en evaluar el efecto de X_1 sobre Y , manteniendo constante X_2 (β_1) y el impacto de X_2 controlando X_1 (β_2). En el ajuste del modelo constatamos que existe una relación lineal estrecha entre los regresores:

$$X_1 = a + bX_2 + e$$

la cual nos impide encontrar resultados estadísticamente significativos. Supongamos que decidimos eliminar X_1 del modelo original. Debido a que la eliminación no significa control tendremos problemas en asignar contenido sustantivo a los valores estimados de los parámetros:

Reemplazando la segunda ecuación en la primera:

$$Y = (\beta_0 + \beta_1 a) + (\beta_2 + \beta_1 b)X_2 + (\beta_1 e + U)$$

Luego el coeficiente de X_2 no mide el impacto de X_2 manteniendo constante X_1 sino que es una combinación lineal de ese efecto y el impacto de X_1 (β_1) a través de la relación entre X_2 y X_1 ($\beta_1 b$). De forma similar se puede interpretar la estimación del término libre.

Nótese que si $b=0$, el coeficiente estimado tiene el mismo sentido que en la relación original.

5.3. Heterocedasticidad.

En secciones anteriores hemos visto que este problema se genera esencialmente en la agregación.

Además, hemos argumentado que, en aquellas situaciones en que no se puede sustentar el supuesto de homocedasticidad, el método de estimación más eficiente, es el mínimo cuadrático generalizado. Si aplicamos el método mínimo cuadrático ordinario las varianzas de los estimadores resultarán ser mayores. Por lo tanto, el método de estimación de uso corriente será más impreciso que el mínimo cuadrático generalizado.

El procedimiento de Aitken se basa en el principio de dar menos ponderación a aquellas observaciones que tienen mayor dispersión que a las que tienen menor. Se "cree" menos a aquellas observaciones que generan mayores errores observados que a las que generan menores errores.

Otra fuente que origina heterocedasticidad, es la posible relación que en ocasiones se puede observar entre los residuos y las variables explicativas. En aquellos casos en que existe este tipo de relación no sólo se generan problemas de sesgo en los estimadores, sino también de consistencia^{1/}.

La solución a este problema se encuentra en el uso del método mínimo cuadrático generalizado, aún cuando surgen dificultades en

^{1/} Wonnacott y Wonnacott, Op.cit., págs.149-155.

la determinación de los elementos de la matriz de varianzas.

Si por ejemplo, hemos establecido una relación proporcional entre un regresor y el término de error:

$$\sigma_i = kx_i$$

podemos reemplazar los elementos de la matriz diagonal V:

$$V = k^2 \begin{bmatrix} x_1^2 & 0 & \dots & 0 \\ 0 & x_2^2 & \dots & 0 \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ 0 & 0 & \dots & x_n^2 \end{bmatrix}$$

y posteriormente proceder a aplicar el método general.

En todo caso antes de intentar una solución a un problema, es necesario tener elementos para detectarlo. En esta línea hay dos tipos de pruebas de hipótesis que prestan utilidad.

Se puede recurrir a la prueba de Goldfeld y Quandt^{1/}, la cual sólo presenta utilidad práctica para aquellos casos en que tenemos un reducido número de datos (los programas de regresión disponibles presentan restricciones). En el modelo cualitativo, hemos controlado problemas de heterocedasticidad. En el modelo mixto,

^{1/} Goldfeld, S.M. y Quandt, R.E.: "Some Tests for Homocedasticity". Journal of the American Statistical Association 60. Págs.539-547

en que trabajamos con alrededor de 13.000 observaciones individuales y con un programa de regresión que no entrega residuos, se nos plantean una serie de dificultades adicionales para aplicar este método.

Por ello resulta más conveniente recurrir a la prueba de igualdad de varias varianzas. Esta prueba se realiza con la F de Snedecor y presenta la dificultad de que sus resultados dependen de los grupos que se forman para calcular las varianzas. Sin embargo, en nuestros modelos mixtos no tenemos tal dificultad por cuanto los agrupamientos realizados tienen plena validez teórica.

En resumen, el problema de heterocedasticidad se encuentra controlado en el caso de los modelos cualitativos agregados. En los modelos mixtos hay que investigar respecto a la dependencia de la varianza de los errores respecto a las variables métricas. Para ello lo más conveniente es aplicar la prueba F. Una vez detectado el problema, debe investigarse la naturaleza de la relación. Una vez que la conocemos definimos la matriz V y, por último, aplicamos el método mínimo cuadrático generalizado.

5.4. Algunas Pruebas de Hipótesis.

En los modelos cualitativos agregativos, disponemos de dos procedimientos para someter a prueba el modelo considerado como un todo.

De una parte, tenemos la d6cima F tradicional que nos permite concluir con alg6n grado de probabilidad si el conjunto de coeficientes de regresi6n es estadisticamente distinto de cero. De otra, disponemos de una prueba ji cuadrado especifica, para este tipo de modelos. En efecto, la prueba F es independiente del n6mero de observaciones por casillas, sin embargo, es intuitivamente obvio que el grado de bondad de ajuste del modelo debe ser funci6n de la base de las proporciones.

Si aceptamos que se cumplen las condiciones para aplicar el teorema central del l6mite sobre los residuos estoc6sticos, entonces la variable media muestral se distribuir6 tambi6n de la misma forma, por cuanto no es m6s que una transformaci6n lineal de los residuos.

Por lo tanto, si normalizamos la variable media aritm6tica y la elevamos al cuadrado estaremos en condiciones de construir por cada casilla una ji-cuadrado con un grado de libertad. Si adem6s suponemos que cada una de las submuestras se han generado a partir de procesos de selecci6n independientes, podemos construir la suma de un conjunto de variables ji-cuadrada estadisticamente independientes con tantos grados de libertad como sumandos haya:

$$X^2 = \sum_i \sum_j \frac{(\bar{X}_{ij} - \mu_{ij})^2}{\sigma_{ij}^2 / n_{ij}}$$

Ahora bien, como no conocemos los valores de μ_{ij} , procedemos a su estimaci6n sobre la base de la ecuaci6n de regresi6n cuidada

de restar tantos grados de libertad como parámetros tenga el modelo^{1/}. De esta forma llegamos a:

$$\chi^2 = \sum_i \sum_j \frac{(\bar{x}_{ij} - \mu_{ij})^2}{\sigma_{ij}^2 / n_{ij}}$$

donde μ_{ij} es el valor de la media estimada a base del modelo de regresión.

Para el caso del modelo de la sección 3, el número de grados de libertad será igual a 9.

La prueba F tradicional nos permite aseverar estadísticamente si el modelo existe o no, la prueba ji-cuadrada nos permite evaluar la validez de la especificación del modelo.

Especial cuidado debe ponerse al realizar las pruebas de hipótesis respecto a los parámetros de las variables mudas. Para ello es necesario en primer lugar encontrar el sentido que tiene cada uno de los parámetros. En el caso del modelo (7) los parámetros muestran el impacto de las categorías de las variables explicativas sobre el número medio de hijos. Esos efectos son relativos y de base variable por cuanto la comparación se hace respecto a la categoría inmediatamente contigua. En el modelo representado por la ecuación (12) la base de comparación es fija. Se mide el impacto de las categorías respecto a la ciudad de Guayaquil y en relación a la categoría ocupacional obreros no especializados.

^{1/} Hoel Paul, "Introduction to Mathematical Statistics". John Wiley, New York, 1962. Págs.249-250.

Supongamos que en este último modelo estamos interesados en someter a prueba los efectos de categorías contiguas. Específicamente si nos preocupamos por saber si el pasaje de la categoría empleado a empleador tienen algún efecto estadísticamente significativo sobre el número de hijos, debemos saber en primer lugar cómo se mide este efecto. Observando la tabla de descomposición teórica del número de hijos concluimos que este efecto es medido por la diferencia entre β_6 y β_5 . Por lo tanto, debemos llevar a cabo una prueba de diferencia de coeficientes en que el error estándar viene dado por:

$$\text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) - 2 \text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$$

y donde la covarianza entre los estimadores se obtiene de la matriz de varianzas y covarianzas.

En esta sección hemos mostrado una prueba ji cuadrada que nos permite evaluar la bondad de la especificación del modelo. Esta prueba de hipótesis nos permite ir más allá de la prueba F tradicional en la medida que esta última sólo nos dice si el modelo existe o no. Además, nos hemos preocupado acerca del sentido de las pruebas relativas a los parámetros de las variables mudas. Para realizar dósimas con sentido debemos conocer previamente cómo se mide el impacto que nos interesa someter a prueba.

5.5. El coeficiente de determinación.

Si consideramos que la selección de formas funcionales alternativas envuelven un compromiso entre varios criterios, incluyendo los sustantivos y los de simpleza de la función, es posible argumentar que a menor número de parámetros mayor simpleza. Es convencional recoger esta idea y comparar coeficientes de determinación corregidos mediante la fórmula:

$$R_c^2 = R^2 - \frac{k}{n-k-1} (1-R^2)$$

Este coeficiente penaliza a aquellas funciones que tienen un gran número de parámetros (k).

Por lo tanto, si debemos comparar el grado de bondad de ajuste de modelos alternativos y con distinto número de regresores se hace necesario entregar tanto el coeficiente de correlación corriente como el corregido.

6. Conclusiones.

El enfoque general que establece los límites de este trabajo sólo consiste en la aplicación del principio que debiera orientar toda investigación cuyo sustrato empírico descansa en el modelo de regresión:

- a) Investigar hasta qué punto se cumple el conjunto de supuestos que caracterizan la aplicación de la estimación mínimo cuadrática ordinaria.

- b) Desarrollar las formas alternativas de detección de problemas de estimación, y
- c) Aplicar, en consecuencia, los mejores métodos de estimación disponibles.

La parte (a) de este principio nos ha llevado a plantear los problemas de multicolinealidad y heterocedasticidad.

La multicolinealidad se introdujo a raíz de la "trampa de las variables mudas" y se dedicó algún espacio a su posible presencia en los modelos que sirven de referencia a este trabajo.

El tratamiento de heterocedasticidad se incorpora a raíz de la agregación de observaciones, así como a la posible relación funcional entre los residuos y los regresores.

La parte (b) nos ha enviado sobre una breve discusión respecto a la forma más adecuada de someter a prueba la presencia de esos problemas. En cuanto a las fuentes de heterocedasticidad hemos planteado dos alternativas. Una que podemos calificar como esencialmente teórica que consiste en reconocer que si trabajamos con datos agregados es difícil sostener el supuesto de homocedasticidad, bajo la condición de haberlo aceptado para datos individuales. La otra, puede ser denominada como fuente empírica por cuanto descansa en la existencia de relación funcional entre la varianza del error y a lo menos uno de los regresores. Cuando este es el caso, hemos concluido que es posible aplicar la prueba tradicional de igualdad de varias varianzas.

La presencia de relación lineal entre los regresores, se detecta a través del examen de los errores estándares estimados, o bien, por medio de las correspondientes pruebas de hipótesis de los parámetros. El habitual examen de la matriz de intercorrelaciones, que normalmente se realiza con el propósito de detectar multicolinealidad, puede llevar a resultados engañosos: el hecho de que las correlaciones simples se distribuyan en el entorno del valor cero, no necesariamente implica ausencia de colinealidad. En efecto, a veces se desliza al interior del modelo, a consecuencia de combinaciones lineales estrechas entre un conjunto de regresores.

Debido a la presencia de heterocedasticidad ha sido necesario incorporar el método de estimación mínimo cuadrático generalizado. En cada caso nos hemos preocupado por especificar la forma que asume la matriz de ponderaciones, que ha sido, simbolizada indistintamente por P o por V.

El símbolo P ha sido utilizado en el modelo que pretende estimar la relación entre variables individuales, pero la información disponible es la media de los agregados. Los elementos de la matriz diagonal P están definidos por el peso relativo de cada observación en el grupo a que pertenece. Cuando los elementos de la matriz de ponderaciones son varianzas, hemos utilizado la letra V.

Por último hemos tratado una serie de tópicos de menor trascendencia. Hemos considerado una prueba ji-cuadrado que nos

permite evaluar la bondad de la especificación del modelo, así como nos preocupamos por establecer la necesidad de encontrar el contenido sustantivo de los coeficientes de regresiones asociados a las variables mudas. Esto último es de importancia fundamental por cuanto nos permite construir pruebas de hipótesis con sentido sobre los parámetros.

III. UN PROGRAMA PARA CALCULAR REGRESION MULTIPLE CON VARIABLES MUDAS, CON EL SISTEMA SPSS.

Laura Gougain A.

1. Objetivos:

En este trabajo se presenta un programa en el sistema SPSS, para el análisis de regresión múltiple con variables mudas para una técnica de regresión que incorpora variables cualitativas. Para facilitar la comprensión y los alcances prácticos de los procedimientos involucrados, se ha estimado necesario describir brevemente tanto el significado de las variables mudas como los elementos centrales de este sistema computacional. Debe entenderse claramente que el propósito de este trabajo no es el desarrollo de aspectos teóricos de las variables mudas en sí mismas y que las delimitaciones conceptuales que se han hecho sólo corresponden a la finalidad mencionada.

2. Acerca de las variables mudas.

Existe la idea bastante generalizada que el procedimiento estadístico denominado Regresión Múltiple necesita una serie de supuestos para su utilización, siendo una de sus limitaciones más insalvables, el nivel de medición de las variables utilizadas, y como con frecuencia los análisis requieren del manejo de variables cualitativas quedaría fuera de toda posibilidad la aplicación de las técnicas de regresión y por tanto sus

derivaciones más útiles (por ejemplo el cálculo de R^2). La utilización de variables mudas es un recurso desarrollado por los econométricos, el cual permite elaborar modelos de regresión que incluyen variables cualitativas. Como se sabe, habitualmente las hipótesis que involucran este tipo de variables son puestas a prueba utilizando tests de contingencia o coeficientes de asociación, que como herramientas analíticas no ofrecen las ventajas del análisis de regresión.

Variable muda es una variable cualitativa que consta solamente de dos valores, cero y uno, indicativos de ausencia o presencia de una determinada cualidad.

Pueden indicarse aquí como delimitaciones conceptuales algunos alcances contenidos en trabajos de econometría. J. Johnston^{1/} señala su utilidad para representar variables cualitativas (atributos) anotando como ejemplo el sexo, estado civil, situación ocupacional y nivel socio-económico, al mismo tiempo que hace presente su utilidad en el tratamiento de variables cuantitativas cuando es necesario y tiene sentido realizar agrupamientos con sus valores.

En el trabajo de Wonnacott y Wonnacott^{2/} se describe la utilidad de las variables mudas en el análisis de series temporales.

^{1/} Johnston: Métodos de la Econometría. Editorial Vicens-Vives, 1970. (Págs. 221).

^{2/} Wonnacott y Wonnacott: Econometrics. Wiley International Edition, 1970. Págs. 68-80.

tomando diversos ejemplos ligados a la función de consumo, demanda, etc.

3. Algunos ejemplos.

Al trabajar con variables mudas es posible construir innumerables modelos, dependiendo del tipo de análisis que se proponga realizar. Se presentarán algunos casos frecuentes:

3.1. Supongamos que se desea relacionar una variable cualitativa tricotómica (considerada como independiente) con una variable dependiente cuantitativa, midiendo además los impactos de cada una de las categorías de la primera sobre la última. En este caso se generan 3 variables mudas (X_1 , X_2 , X_3), una por cada categoría de la variable cualitativa.

Formalmente el problema planteado queda formulado del siguiente modo:

$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

Al ajustar esta función debe tenerse presente que:

β_1 = efecto de la primera categoría sobre la variable dependiente.

β_2 = efecto de la segunda categoría sobre la variable dependiente.

β_3 = efecto de la tercera categoría sobre la variable dependiente.

e = término de error.

Puede apreciarse la omisión del regresor β_0 , ya que la inclusión de este término libre, conlleva problemas de multicolinealidad^{1/}.

3.2. Otro caso interesante puede ocurrir al relacionar dos variables independientes (cualitativas tricotómicas) con una variable dependiente (cuantitativa) para determinar el impacto que sobre la última tiene el paso de una a otra categoría de cada una de las dos primeras. Aquí se generan 5 variables mudas.

Para mayor claridad se detalla en el cuadro siguiente la disposición de los coeficientes correspondientes a estas variables.

^{1/} Para mayor profundización en el tema ver primer artículo de esta misma publicación.

VI ₂ \ VI ₁	A	B	C
a	β_1	$\beta_1 + \beta_2$	$\beta_1 + \beta_3$
b	$\beta_1 + \beta_4$	$\beta_1 + \beta_2 + \beta_4$	$\beta_1 + \beta_3 + \beta_4$
c	$\beta_1 + \beta_5$	$\beta_1 + \beta_2 + \beta_5$	$\beta_1 + \beta_3 + \beta_5$

La función que debe ajustarse es:

$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + e$$

donde β_1 = nivel de la variable independiente esperado en la situación que combina la categoría A de VI₁ (variable independiente número 1) y la categoría a de VI₂ (segunda variable independiente). Se usa como base de comparación, razón por la cual aparece en todas las subcasillas siguientes.

β_2 = impacto que sobre la variable dependiente tiene el cambio de la categoría A hacia la B en VI₁, manteniendo constantes VI₂.

β_3 = efecto del traslado de la categoría A hacia la C en VI₁, permaneciendo constante VI₂.

β_4 = impacto del cambio de la categoría a hacia la b en VI₂ manteniendo constante VI₁.

β_5 = efecto del traslado de la categoría a hacia la c en VI_2 , permaneciendo constante VI_1 .

e = término estocástico.

Las variables mudas adoptan valores 1 en las casillas en donde se encuentre el coeficiente beta correspondiente y 0 cuando éste no aparezca.

Listado de variables mudas

a_{ij}	X_1	X_2	X_3	X_4	X_5
a_{11}	1	0	0	0	0
a_{12}	1	1	0	0	0
a_{13}	1	0	1	0	0
a_{21}	1	0	0	1	0
a_{22}	1	1	0	1	0
a_{23}	1	0	1	1	0
a_{31}	1	0	0	0	1
a_{32}	1	1	0	0	1
a_{33}	1	0	1	0	1

Así en la casilla a_{22} : ($\beta_1 + \beta_2 + \beta_4$), se observa la combinación $X_1=1; X_2=1; X_3=0; X_4=1; X_5=0$. De modo que para esta casilla aplicando los correspondientes elementos del dominio:

$$y_{22} = \beta_1(1) + \beta_2(1) + \beta_3(0) + \beta_4(1) + \beta_5(0)$$

Para la casilla a_{33} :

$$y_{33} = \beta_1(1) + \beta_2(0) + \beta_3(1) + \beta_4(0) + \beta_5(1)$$

y así respecto a cada una de las a_{ij} .

Si existe interacción entre las variables independientes se agregan nuevas variables mudas.

Con este modelo se miden los impactos del cambio de cada una de las categorías respecto de la primera.

3.3. Si se deseara medir (para las mismas variables definidas en el caso 2) los impactos que producen cambios entre categorías sucesivas el modelo varía ligeramente, modificándose algunos valores de las variables mudas, ya que los coeficientes beta aparecen en mayor número de casillas que en el modelo anterior. La tabla de los coeficientes sería la siguiente:

$VI_2 \backslash VI_1$	A	B	C
a	β_1	$\beta_1 + \beta_2$	$\beta_1 + \beta_2 + \beta_3$
b	$\beta_1 + \beta_4$	$\beta_1 + \beta_2 + \beta_4$	$\beta_1 + \beta_2 + \beta_3 + \beta_4$
c	$\beta_1 + \beta_4 + \beta_5$	$\beta_1 + \beta_2 + \beta_4 + \beta_5$	$\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5$

La función a ajustarse no varía:

$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + e$$

en cambio varía el significado de algunos coeficientes:

β_1 , β_2 y β_4 = igual que en el modelo anterior.

β_3 = efecto del traslado de la categoría B a la categoría C en VI_1 , permaneciendo constante VI_2 .

β_5 = efecto del traslado de la categoría b a la categoría c en VI_2 , permaneciendo constante VI_1 .

La diferencia más notable entre el modelo de regresión con variables mudas y el de regresión normalmente utilizado es que mientras éste último permite medir efectos de variables, el primero mide efectos de categorías de las variables.

Tal diferencia constituye la ventaja fundamental del análisis de regresión con variables mudas, ya que al incorporar variables explicativas otorgando valores a las categorías que les corresponden es posible evaluar los impactos o efectos asociados a ellas. En esta forma es posible distinguir el peso de los distintos factores considerados como explicativos.

4. Acerca del Sistema SPSS.

El Statistical Package for the Social Sciences (SPSS), es un sistema de programas generalizados destinados especialmente al análisis de datos que provienen de investigaciones sociales y socio-demográficas. Las características de este sistema son:

- Gran capacidad para el manejo de datos y su procesamiento estadístico.
- Posibilidad de transformar y recodificar variables. De esta

manera no es necesario alterar los datos originales para definir nuevas variables a partir de otras ya existentes (propiedad que sirve especialmente, en nuestro caso, para la generación de variables mudas).

- Está estructurado en base a subprogramas de cálculo, que ejecuta los distintos procedimientos estadísticos.

Tal como ocurre en la utilización de cualquier programa, en el SPSS distinguiremos dos conjuntos de tarjetas:

- Las tarjetas de control del sistema operativo;
- Las tarjetas del sistema (SPSS).

Supondremos en adelante que los datos se encuentran almacenados en cinta magnética.

4.1. Las tarjetas de control del sistema operativo.

Son la serie de instrucciones en lenguaje especial (Job Control Language) que están dirigidas a los programas que controlan la operación interna del computador. Estas tarjetas se perforan siempre a partir de la primera columna y pueden ocupar hasta la columna 71 inclusive. Siempre comienzan con un slash (/). Estas tarjetas iniciarán siempre el programa.

```
//Clave de identificación del usuario
//CLASS=A
//AQG EXEC PGM=SPSS,REGION=240K,PARM=100K,TIME=120,COND=EVEN
//STEPLIB DD UNIT=SYSDA,VOL=SER=,SORT01,DSN=SPSSH,LOAD=V501,DISP=OLD
//FT01F001 DD UNIT=SYSDA,SPACE=(TRK,(350,125))
//FT02F001 DD UNIT=SYSDA,SPACE=(CYL,(25,5))
```

```
//FT03F001 DD UNIT=2400,LABEL=(1,NL),DISP=OLD,VOL=SER=_____1/
```

```
//DCB=BLKSIZE=4000
```

```
//FT06F001 DD SYSOUT=A
```

```
//FT05F001 DD *//
```

Si se va a crear un nuevo archivo, deben agregarse dos tarjetas, antes de la tarjeta //FT05F001 DD

```
//FT04F001 DD UNIT=2400,LABEL=(1,NL),DISP=(NEW,PASS),VOL=SER=_____2/
```

```
//DCB=BLKSIZE=4000
```

4.2. Las tarjetas del sistema:

Están compuestas de 2 campos:

- Campo de control: ocupa las columnas 1 a 15 inclusive, contiene la palabra clave de identificación (ej.: RUN NAME, GET FILE, COMPUTE STATISTICS);
- Campo de especificación: ocupa las columnas 16 a 80 inclusive, su contenido varía de acuerdo a la tarjeta.

Para la creación y procesamiento de variables mudas la disposición de las tarjetas es la siguiente:

-
- 1/ Se especifica el nombre de la cinta que tiene los datos que se utilizarán para el cálculo (esta cinta debe ir sin anillo, para que no se corra el riesgo de borrar los datos del archivo original).
 - 2/ Se especifica el nombre de la cinta que va a contener el nuevo archivo (esta cinta debe ir con anillo, para que pueda ser grabada con los nuevos datos).

```

Col 1                               Col 16
// Clave de identificación del usuario
.
.
.
//FT05F001 DD *
RUN NAME          REGRESION MULTIPLE CON VARIABLES MUDAS
GET FILE          ...
COMPUTE          ...
.
.
PROCESS SBFILES  ...
REGRESION        VARIABLES= ...
REGRESION= ... WITH...
STATISTICS      ALL
FINISH
/*

```

Tarjetas de control del sistema operativo.

Tarjetas para la creación de las variables mudas

Si se está creando un nuevo archivo se agregarán las siguientes tarjetas.

```

FILE NAME ...
(SAVE FILE ...)

```

Después de GET FILE y antes de FINISH

1. Tarjeta RUN NAME.

Desde la columna 16 a 80 se perfora el rótulo del trabajo que se va a realizar. En este caso se trata de regresión múltiple con variables mudas. Si se desea podrán colocarse algunas otras indicaciones, siempre que queden dentro del espacio ya señalado.

2. Tarjeta GET FILE.

Desde la columna 16 se perfora el nombre del FILE (archivo) en el cual se encuentran los datos que se van a procesar.

3. Tarjeta PROCESS SBFILES.

Generalmente el archivo tiene la forma de subarchivos.

(Por ejemplo las encuestas de PECFAL tienen un subarchivo para cada país, lo que hace más manejable la información separada de cada uno de ellos). Si esto sucede, debe obligatoriamente señalarse cómo se procesarán estos subarchivos. Los casos más comunes son:

- a) PROCESS SBFILES ALL. Indica que se ignora la estructura de subarchivos y que se procesa la información en bloque.
- b) PROCESS SBFILES EACH. Indica que se procesa cada subarchivo separadamente.
- c) PROCESS SBFILES SUB 1 o PROCESS SBFILES SUB 3 o PROCESS SBFILES SUB 5. Indica que se procesa solamente un subarchivo, el 1, el 3 o el 5, según el caso.
- d) PROCESS SBFILES (SUB 1. SUB 2)(SUB 3. SUB 4)(SUB 5)SUB 6)
o PROCESS SBFILES (SUB 1)(SUB 2. SUB 3. SUB 4)(SUB 5)SUB 6)
o PROCESS SBFILES (SUB 1. SUB 2. SUB 3)(SUB 4. SUB 5)(SUB 6)

Estos u otros arreglos indican que se procesarán juntos los subarchivos indicados entre paréntesis y posteriormente se irán procesando los demás conjuntos señalados.

4. Tarjeta REGRESSION.

El campo de especificación indica en primer lugar las variables que se incluyen en la regresión y posteriormente indica cuál es la variable dependiente y cuáles son las variables independientes. Si se desea regresión múltiple paso a paso (step wise)^{1/} se colocará al final de las variables independientes un número impar entre paréntesis. Ejemplo:

Col 1	Col 16
REGRESSION	VARIABLES = X ₁ , X ₂ , X ₃ , X ₄ , X ₅ , X ₆ , X ₇ , X ₈
	REGRESSION = X ₈ with
	X ₁ , X ₂ , X ₃ , X ₄ , X ₅ , X ₆ , X ₇ (1)

Si no se desea regresión step wise se colocará al final de las variables independientes un número par entre paréntesis.

5. Tarjeta STATISTICS.

En el campo de especificación se coloca la palabra ALL, con lo cual se logran todos los estadísticos que el computador calcula para obtener la regresión múltiple (para citar algunos: medias, varianzas, matriz de correlaciones, coeficientes alfa, coeficientes beta, correlaciones parciales, correlaciones múltiples, porcentajes de explicación de cada variable y acumulado, error standard de beta).

^{1/} Regresión step wise es la que ingresa en la regresión a las variables independientes de acuerdo al porcentaje de explicación de la variable dependiente que cada una de ellas entrega.

6. Tarjeta FINISH

Indica al sistema que se termina el proceso.

7. Tarjeta slash asterisco(/*).

Indica que no hay más tarjetas en el programa.

Si se está creando un nuevo archivo, todas las tarjetas de transformación de variables que no vayan con asterisco^{1/} hacen que las nuevas variables sean agregadas al final de éste. Las tarjetas que se deben agregar en este caso significan:

8. FILE NAME.

Da un nombre al nuevo archivo (FILE) que se creará.

9. SAVE FILE.

Indica que un nuevo archivo será creado.

Si se desea eliminar algunas variables debe colocarse una de estas dos tarjetas: DELETE VARS o KEEP VARS inmediatamente antes de SAVE FILE.

Como ya hemos señalado siempre que exista una tarjeta SAVE FILE en las tarjetas del sistema, debe haber una tarjeta FT04F001 en las tarjetas de control del sistema operativo. Si esto no sucede no se podrá crear el nuevo archivo.

1/ Las tarjetas COMPUTE IF y RECODE (*RECODE,*COMPUTE,*IF) con asterisco indican que la recodificación o transformación de las variables es temporal, sin asterisco indica que es definitiva, de modo tal que si se está creando un archivo se incluirán todas las variables que provengan de COMPUTE IF o RECODE sin asterisco.

El formato general de las nuevas tarjetas es:

Col 1	Col 16
DELETE VARS.	Lista de variables
KEEP VARS	Lista de variables

Se usará DELETE cuando en el archivo sean menos las variables que se deseen eliminar que las que se guardan y KEEP cuando sean menos las variables que se desee guardar, que las que se eliminan.

5. Una forma sencilla de crear variables mudas.

Como las variables mudas tienen sólo valores cero y uno, pueden inicializarse dando el valor cero a todas las variables que se crearán, fijando a continuación las condiciones bajo las cuales adoptarán valores uno.

	Col 1	Col 16
Con la tarjeta TMISS y el siguiente formato:	TMISS	0

se inicializan en cero todas las variables a generar.

Con la tarjeta IF se dan las condiciones bajo las cuales estas variables adoptarán el valor uno.

Tarjeta IF.

Permite efectuar transformaciones de variables sobre condiciones lógicas, su formato general es:

Col 1	Col 16
-------	--------

IF (expresión lógica) variable a calcular = expresión aritmética

El cálculo se efectuará en caso de que se cumpla la condición lógica.

La expresión lógica está constituida por uno o más conjuntos de relaciones: los operadores de relación pueden ser:

EQ igual GT mayor que

GE mayor o igual LT menor que

LE menor o igual NE no igual

Para combinar relaciones en una expresión lógica más compleja existen dos operadores lógicos:

AND - el resultado es verdadero si y sólo si ambas relaciones lo son,

OR - el resultado es falso si y sólo si ambas relaciones lo son.

Puede también usarse el operador NOT. Su efecto es invertir el valor de la expresión lógica que precede.

Si la transformación es temporal, o sea se desea para sólo un procedimiento y se necesita ignorarla en el que sigue hay que colocar un asterisco antes de la palabra clave IF.

Para explicar cómo se crean las variables mudas se trabajará con algunos ejemplos, los cuales han sido tratados sustantivamente en la investigación de Adolfo Aldunate "Reproducción de la población en 10 ciudades de América Latina" (PROELCE).

1. Se desea conocer los impactos que sobre la variable N° de hijos nacidos vivos tienen algunas categorías, de las variables "grupo ocupacional del jefe de familia" según grado de calificación

de la fuerza de trabajo y ciudades latinoamericanas según nivel de desarrollo. En este caso es necesario construir un modelo agregativo y cualitativo (ordinal) en el cual los coeficientes, exceptuando β_1 , miden impactos relativos de las categorías. Estas últimas tienen efectos independientes.

ARCHIVO: PECFAL URBANO. Consta de 10 subarchivos con información de cada una de las 10 ciudades estudiadas.

Variable dependiente^{1/}: N° de hijos nacidos vivos (VI64)

Variabes independientes: 1) Grupo ocupacional (BTT01)
2) Ciudad (CIUD)

Se consideran sólo algunas categorías de las variables independientes, las cuales son presentadas en el siguiente cuadro:

Grupos Ciudades	Obreros no Especializ.	Obreros Especializ.	Empleados	Directivos
Guayaquil	β_1	$\beta_1 + \beta_2$	$\beta_1 + \beta_2 + \beta_3$	$\beta_1 + \beta_2 + \beta_3 + \beta_4$
Caracas	$\beta_1 + \beta_5$	$\beta_1 + \beta_2 + \beta_5$	$\beta_1 + \beta_2 + \beta_3 + \beta_5$	$\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5$
Río de Janeiro	$\beta_1 + \beta_5 + \beta_6$	$\beta_1 + \beta_2 + \beta_5 + \beta_6$	$\beta_1 + \beta_2 + \beta_3 + \beta_5 + \beta_6$	$\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6$
Buenos Aires	$\beta_1 + \beta_5 + \beta_6 + \beta_7$	$\beta_1 + \beta_2 + \beta_5 + \beta_6 + \beta_7$	$\beta_1 + \beta_2 + \beta_3 + \beta_5 + \beta_6 + \beta_7$	$\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7$

Puesto que a cada coeficiente beta corresponde una variable muda, tenemos en total 7 variables mudas. La función a ajustar

^{1/} Las claves que designan a las variables corresponden a los nombres de éstas en el archivo PUCC2, que contiene la información sobre casadas y convivientes de PECFAL URBANO (CELADE).

es la siguiente:

$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + e$$

dónde:

β_1 = N° de hijos esperado en una familia cuyo jefe es obrero especializado residente en Guayaquil.

β_2 = impacto que sobre la variable dependiente tiene el hecho de que el jefe de familia sea obrero especializado en lugar de no especializado.

β_3 = efecto de ser empleado en lugar obrero especializado.

β_4 = efecto de ser directivo en lugar de empleado.

β_5 = impacto que sobre la variable explicada tiene el hecho de residir en Caracas en lugar de Guayaquil.

β_6 = impacto de residir en Río de Janeiro en lugar de Caracas.

β_7 = impacto de residir en Buenos Aires en lugar de Río de Janeiro.

e = término estocástico.

Se ha mencionado anteriormente que las variables mudas adoptan valores cero en las casillas en que no se encuentra su coeficiente beta asociado y valores uno en las casillas en que aparece.

Por lo tanto, β_1 adopta siempre el valor uno. β_2 adopta el valor uno cuando la variable grupo ocupacional se refiere a los obreros especializados, empleados y directivos y toma el valor cero cuando no se cumple esa condición. β_3 adopta el valor uno en las categorías Empleados y Directivos, cero si no se da esta condición. β_4 adopta el valor uno en la categoría Directivo, cero en las

restantes. β_5 adopta el valor uno cuando la variable ciudad a la que pertenecen los datos se refiere a Caracas, Río de Janeiro y Buenos Aires y toma el valor cero cuando no se cumple esta condición. β_6 adopta el valor uno en la categoría Buenos Aires y cero en las restantes.

Las tarjetas del sistema para la solución de este problema serán:

```
RUN NAME      REGRESION MULTIPLE CON VARIABLES MUDAS.MODELO CUALITATIVO
GET FILE      PUCC2 (PECFAL URBANO,CASADAS Y CONVIVIENTES,VERSION Nº2
COMPUTE       X1=1
TMISS        0
IF            (BTT01 EQ 6 OR BTT01 EQ 3 OR BTT01 EQ 2) X2=1
IF            (BTT01 EQ 3 OR BTT01 EQ 2) X3=1
IF            (BTT01 EQ 2) X4=1
IF            (CIUD EQ 7 OR CIUD EQ 2 OR CIUD EQ 1) X5=1
IF            (CIUD EQ 2 OR CIUD EQ 1) X6=1
IF            (CIUD EQ 1) X7=1
PROCESS SBFIL (GUAYAQUI,CARACAS,JANEIRO,BAIRES)
REGRESSION    VARIABLES = X2, X3, X4, X5, X6, X7,VI64
REGRESSION=VI64 with X2,X3,X4,X5,X6,X7 (1)
STATISTICS   ALL
FINISH
/*
```

NOTA: BTT01 tiene las siguientes categorías: 1. Patrón. 2. Directivo. 3. Empleado. 4. Servidor independiente. 5. Artesanos. 6. Obrero especializado. 7. Obrero no especializado.

CIUD tiene las siguientes categorías: 0. Guatemala. 1. Buenos Aires. 2. Río de Janeiro. 3. Bogotá. 4. San José.

5. México. 6. Panamá. 7. Caracas. 8. Quito. 9. Guayaquil.

No se incluye X_1 en la regresión, ya que este programa crea una columna de unos para generar el término libre, lo que equivale a X_1 .

Este ejemplo fue desarrollado con un computador de mesa, disponible en PROELCE: Hewlett Packard 9810 A, el cual tiene una capacidad de memoria limitada y que como máximo puede resolver en este momento inversión de matrices con $n \leq 50$.

Para la resolución del ejemplo mencionado debieron utilizarse programas de multiplicación de matrices, inversión de matrices, determinación de "y" estimados, cálculos de residuos y cálculo de R^2 .

La ecuación de regresión fue la siguiente:

$$y = 4.59X_1 - 0.70X_2 - 0.39X_3 - 0.04X_4 - 0.42X_5 - 0.82X_6 - 0.80X_7$$

De los resultados se puede destacar que β_6 , β_7 , y β_2 son los efectos de mayor importancia, o sea, pasar de Caracas a Río de Janeiro, pasar de Río de Janeiro a Buenos Aires y pasar de obrero no especializado a especializado respectivamente. Una vez realizadas las pruebas de significación, el

único coeficiente no significativo resultó ser β_4 (efecto de pasar de la categoría Empleado a la categoría Directivo).

Nuevamente interesa conocer los impactos que sobre la variable número de hijos nacidos vivos tienen las categorías de las variables grupo ocupacional y ciudad a la que corresponden los datos (se utilizan ahora todas las categorías de estas variables) además interesa detectar los impactos que otras variables independientes (cuantitativas) tienen sobre número de hijos nacidos vivos, estas son: edad de la entrevistada, edad cuadrática, edad de la entrevistada al casarse, socialización urbana, feminismo, religiosidad, tipo de unión del matrimonio y participación laboral femenina^{1/}.

ARCHIVO: el mismo utilizado en el modelo anterior.

Variable dependiente: Número de hijos nacidos vivos (V 164).

VARIABLES INDEPENDIENTES: Grupo ocupacional marido (BTT01)
Ciudad a la que corresponden los datos (CIUD)
Edad de la entrevistada (B174)
Edad cuadrática (X72)
Edad de la entrevistada al casarse (EDLEAC)
Socialización urbana (BSU01)
Feminismo (BEF06)
Religiosidad (BRE01)
Tipo de unión del matrimonio (B140)
Participación laboral femenina (V033)

El cuadro con los coeficientes beta del modelo se presenta a continuación.

^{1/} Exceptuando la edad, se trata de un conjunto de índices elaborados en: "Reproducción de la Población en 10 Ciudades de América Latina: aproximación a un análisis grupal", Adolfo Aldunate, PROELCE.

GRUPO	CIUDAD									
	GUAYAQUIL	QUITO	GUATEMALA	MEXICO	SAN JOSE	CARACAS	BOGOTA	RIO DE JANEIRO	PANAMA	BUENOS AIRES
Obrero No Especializado	β_1	$\beta_1 + \beta_2$	$\beta_1 + \beta_3$	$\beta_1 + \beta_4$	$\beta_1 + \beta_5$	$\beta_1 + \beta_6$	$\beta_1 + \beta_7$	$\beta_1 + \beta_8$	$\beta_1 + \beta_9$	$\beta_1 + \beta_{10}$
Artesano	$\beta_1 + \beta_{11}$	$\beta_1 + \beta_2 + \beta_{11}$	$\beta_1 + \beta_3 + \beta_{11}$	$\beta_1 + \beta_4 + \beta_{11}$	$\beta_1 + \beta_5 + \beta_{11}$	$\beta_1 + \beta_6 + \beta_{11}$	$\beta_1 + \beta_7 + \beta_{11}$	$\beta_1 + \beta_8 + \beta_{11}$	$\beta_1 + \beta_9 + \beta_{11}$	$\beta_1 + \beta_{10} + \beta_{11}$
Obrero Especializado	$\beta_1 + \beta_{12}$	$\beta_1 + \beta_2 + \beta_{12}$	$\beta_1 + \beta_3 + \beta_{12}$	$\beta_1 + \beta_4 + \beta_{12}$	$\beta_1 + \beta_5 + \beta_{12}$	$\beta_1 + \beta_6 + \beta_{12}$	$\beta_1 + \beta_7 + \beta_{12}$	$\beta_1 + \beta_8 + \beta_{12}$	$\beta_1 + \beta_9 + \beta_{12}$	$\beta_1 + \beta_{10} + \beta_{12}$
Servicio Independiente	$\beta_1 + \beta_{13}$	$\beta_1 + \beta_2 + \beta_{13}$	$\beta_1 + \beta_3 + \beta_{13}$	$\beta_1 + \beta_4 + \beta_{13}$	$\beta_1 + \beta_5 + \beta_{13}$	$\beta_1 + \beta_6 + \beta_{13}$	$\beta_1 + \beta_7 + \beta_{13}$	$\beta_1 + \beta_8 + \beta_{13}$	$\beta_1 + \beta_9 + \beta_{13}$	$\beta_1 + \beta_{10} + \beta_{13}$
Empleado	$\beta_1 + \beta_{14}$	$\beta_1 + \beta_2 + \beta_{14}$	$\beta_1 + \beta_3 + \beta_{14}$	$\beta_1 + \beta_4 + \beta_{14}$	$\beta_1 + \beta_5 + \beta_{14}$	$\beta_1 + \beta_6 + \beta_{14}$	$\beta_1 + \beta_7 + \beta_{14}$	$\beta_1 + \beta_8 + \beta_{14}$	$\beta_1 + \beta_9 + \beta_{14}$	$\beta_1 + \beta_{10} + \beta_{14}$
Patrón	$\beta_1 + \beta_{15}$	$\beta_1 + \beta_2 + \beta_{15}$	$\beta_1 + \beta_3 + \beta_{15}$	$\beta_1 + \beta_4 + \beta_{15}$	$\beta_1 + \beta_5 + \beta_{15}$	$\beta_1 + \beta_6 + \beta_{15}$	$\beta_1 + \beta_7 + \beta_{15}$	$\beta_1 + \beta_8 + \beta_{15}$	$\beta_1 + \beta_9 + \beta_{15}$	$\beta_1 + \beta_{10} + \beta_{15}$
Directivo	$\beta_1 + \beta_{16}$	$\beta_1 + \beta_2 + \beta_{16}$	$\beta_1 + \beta_3 + \beta_{16}$	$\beta_1 + \beta_4 + \beta_{16}$	$\beta_1 + \beta_5 + \beta_{16}$	$\beta_1 + \beta_6 + \beta_{16}$	$\beta_1 + \beta_7 + \beta_{16}$	$\beta_1 + \beta_8 + \beta_{16}$	$\beta_1 + \beta_9 + \beta_{16}$	$\beta_1 + \beta_{10} + \beta_{16}$

El modelo con que se trabaja mide los efectos de las categorías respecto a la primera de ellas que sirve de base de comparación fija (modelo cualitativo nominal). Se generan 16 variables mudas; como hay otras variables independientes (cuantitativas), el modelo puede denominarse mixto. La función a ajustar será la siguiente:

$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \\ + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{17} B_{174} + \beta_{18} X_{72} + \beta_{19} \text{EDLEAC} + \\ + \beta_{20} \text{BSU01} + \beta_{21} \text{BEF06} + \beta_{22} \text{BRE01} + \beta_{23} B_{140} + \beta_{24} \text{V033}$$

donde: β_1 = N° esperado de hijos para los obreros no especializados de Guayaquil.

β_2 = jefe de familia reside en Quito en lugar de Guayaquil.

β_3 = impacto de la residencia en Guatemala en lugar de Guayaquil.

β_4 = impacto por residir en México en lugar de Guayaquil.

β_5 = impacto por residir en San José en lugar de Guayaquil.

β_6 = impacto por residir en Caracas en lugar de Guayaquil.

β_7 = impacto por residir en Bogotá en lugar de Guayaquil.

β_8 = impacto por residir en Río de Janeiro en lugar de Guayaquil.

β_9 = impacto por residir en Panamá en lugar de Guayaquil.

β_{10} = impacto por residir en Buenos Aires en lugar de Guayaquil.

β_{11} = efecto relativo que tiene sobre el N° de hijos el hecho de que el jefe de familia sea artesano en lugar de no especializado.

β_{12} = efecto debido a ser obrero especializado en lugar de obrero no especializado.

β_{13} = efecto debido a ser servidor independiente en lugar de obrero no especializado.

β_{14} = efecto debido a ser empleado en lugar de obrero no especializado.

β_{15} = efecto debido a ser patrón en lugar de obrero no especializado.

β_{16} = efecto debido a ser director en lugar de obrero no especializado.

β_{17} = efecto sobre la variable dependiente debido a la edad de la entrevistada.

β_{18} = efecto sobre la variable dependiente debido a la edad considerada al cuadrado.

β_{19} = efecto producido sobre la variable dependiente debido a la edad de la entrevistada al momento de casarse.

β_{20} = efecto sobre la variable dependiente debido a la socialización urbana.

β_{21} = efecto sobre la variable dependiente debido al feminismo de la entrevistada.

β_{22} = efecto sobre la variable dependiente debido a la religiosidad.

β_{23} = efecto sobre la variable dependiente debido al tipo de unión del matrimonio.

β_{24} = efecto sobre la variable dependiente debido a la participación laboral de la mujer.

Los valores de las variables mudas serán los siguientes:

β_1 siempre tendrá valores uno.

β_2 tendrá valores uno cuando CIUD sea Quito y cero si no se da esta condición.

β_3 tendrá valores uno cuando CIUD sea Guatemala, cero si no se da esta condición.

β_4 tendrá valores uno cuando CIUD sea México, cero si no se da esta condición.

β_5 tendrá valores uno cuando CIUD sea San José, cero si no se da esta condición.

β_6 tendrá valores uno cuando CIUD sea Caracas, cero si no se da esta condición.

- β_7 = valores uno cuando CIUD sea Bogotá, cero si no se da esta condición.
- β_8 = valores uno cuando CIUD sea Río de Janeiro, cero si no se da esta condición.
- β_9 = valores uno cuando CIUD sea Panamá, cero si no se da esta condición.
- β_{10} = valores uno cuando CIUD sea Buenos Aires, cero si no se da esta condición.
- β_{11} = valores uno cuando BTT01 sea Artesanos, cero si no se da esta condición.
- β_{12} = valores uno cuando BTT01 sea Obreros especializados, cero si no se da esta condición.
- β_{13} = valores uno cuando BTT01 Servidores independientes, cero si no se da esta condición.
- β_{14} = valores uno cuando BTT01 sea Empleados, cero si no se da esta condición.
- β_{15} = valores uno cuando BTT01 sean Patrones, cero si no se da esta condición.
- β_{16} = valores uno cuando BTT01 sean Directivos, cero si no se da esta condición.

Las tarjetas del sistema para la solución de este problema son:

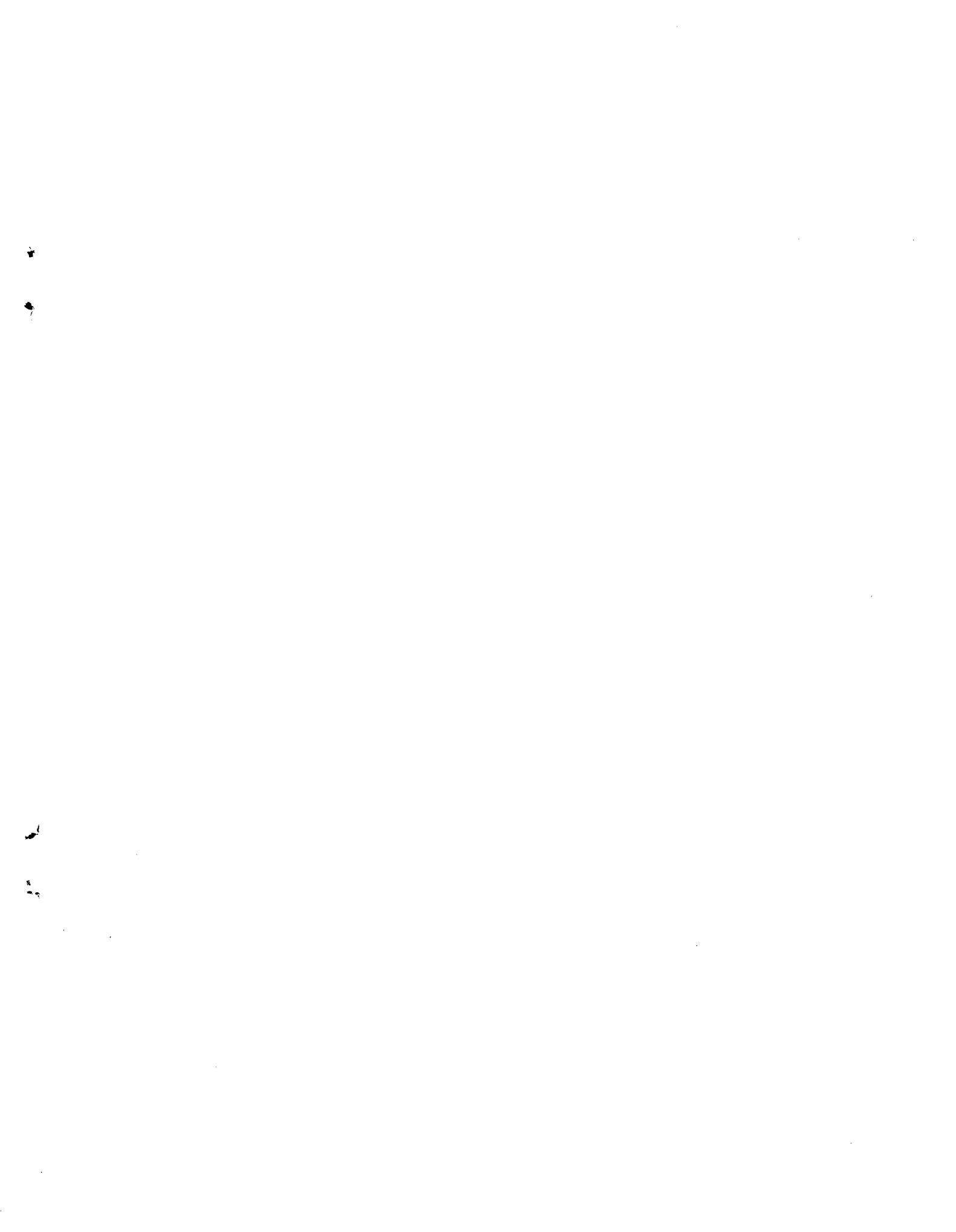
```
RUN NAME      REGRESION MULTIPLE CON VARIABLES MUDAS. MODELO MIXTO
GET FILE      PUCC2
COMPUTE       X1=1
TMISS        0
IF            (CIUD EQ 8) X2 = 1
IF            (CIUD EQ 0) X3 = 1
IF            (CIUD EQ 5) X4 = 1
IF            (CIUD EQ 4) X5 = 1
IF            (CIUD EQ 7) X6 = 1
IF            (CIUD EQ 3) X7 = 1
IF            (CIUD EQ 2) X8 = 1
IF            (CIUD EQ 6) X9 = 1
```

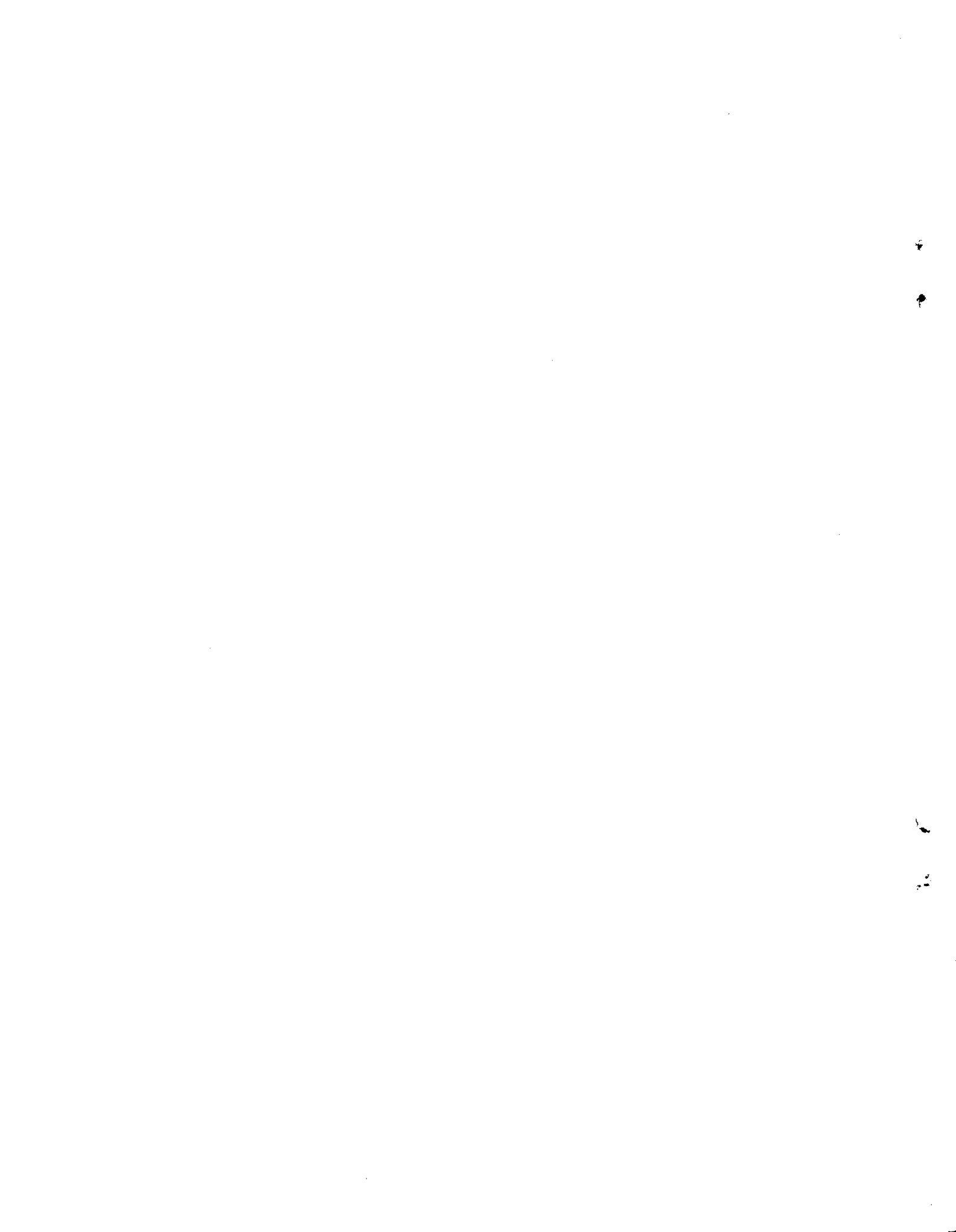
```
IF (CIUD EQ 1) X10 = 1
IF (BTT01 EQ 5) X11 = 1
IF (BTT01 EQ 6) X12 = 1
IF (BTT01 EQ 4) X13 = 1
IF (BTT01 EQ 3) X14 = 1
IF (BTT01 EQ 1) X15 = 1
IF (BTT01 EQ 2) X16 = 1
REGRESSION VARIABLES =
X2,X3,X4,X5,X6,X7,X8,X9, X10,X11,X12,X13,X14,X15,
X16,B174,X72,EDLEAC,BSU01,BEF06,BRE01,B140,V033,
V164
REGRESSION: V 164 with
X2,X3,X4,X5,X6,X7,X8,X9, X10,X11,X12,X13,X14,X15,
X16,B174,X72,EDLEAC,BSU01,BEF06,BRE01,B140,V033(1)
STATISTICS ALL
FINISH
/*
```

NOTA: Las categorías de las variables usadas para la creación de las variables mudas (grupo ocupacional y ciudad fueron dadas a conocer en la nota al programa anterior).

Como se señaló anteriormente no puede tomarse en cuenta X_1 , pues su inclusión distorsionaría el modelo, generándose un problema de multicolinealidad perfecta; puesto que el término libre creado automáticamente por el programa de regresión y la variable muda X_1 serían idénticas.

Ya que el propósito de este trabajo es la descripción de un programa para utilizar variables mudas en estudios de regresión no se entrará acá en desarrollos sustantivos de análisis de resultados.







**Centro Latinoamericano
de Demografía
(CELADE)
J. M. Infante 9
Casilla 91 - Teléfono 257806
Santiago de Chile**

**Escuela Latinoamericana
de Sociología de la Facultad
Latinoamericana de Ciencias
Sociales, FLACSO
(ELAS)
J. M. Infante 51
Casilla 3213 - Teléfono 251043
Santiago de Chile**