UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

ANALISIS DE REGRESION APLICADO AL PROBLEMA DE EDUCACION DE LA POBLACION HISPANICA EN CALIFORNIA.

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I O

PRESENTA

PATRICIA COVARRUBIAS AGUIRRE

MEXICO, D. F. 1983

INDICE

INTRODUCCION

En México, como en cualquier otro país en vías de desarrollo, existe una gran cantidad de problemas tanto económicos como sociales, que son atendidos en mayor o menor grado de acuerdo a las políticas y prioridades de cada país en particular.

Durante esta última década, la mayoría de los países Latino-Americanos han venido sufriendo problemas económicos; en el caso de México, estos problemas han alcanzado un punto crítico y requieren de drásticas medidas para su resolución. Uno de los problemas ocasionados por esta inestabilidad económica es el hecho de que muchas personas, al ver que no se encuentran en condiciones o que no tienen los elementos necesarios para lograr las condiciones de vida que desean, decidan salir del país en busca de trabajos mejor remunerados, que les permitan mantenerse y mantener a sus familias en condiciones decorosas.

Actualmente existe una gran emigración hacia los

Estados Unidos, los trabajadores Mexicanos que no cuentan con los medios necesarios para sobrevivir en su país, tratan desesperadamente de entrar en los Estados Unidos, pagando en ocasiones grandes sumas de dinero a personas que se dedican especialmente a introducir trabajadores ilegales en la Unión Americana, muchos de ellos logran entrar y establecerse, sin embargo la vida allá no es tan fácil como esperaban y tienen que enfrentarse con muchos problemas. Además de encontrar empleos, vivienda y medios para sostenerse, existe otro problema muy importante, al que yo considero no se le ha dado suficiente atención, éste es el problema de la educación.

Las necesidades de educación por parte de la población hispánica en California aumentan día a día y no se encuentran satisfechas debido a que requieren de la implementación de programas especiales con costos muy elevados.

Durante el desarrollo de este trabajo, trataré de analizar estadisticamente este problema y encontrar -- una medida de satisfacción de las necesidades de educa--- ción de la población hispánica en California.

Localización

California se encuentra localizada en la Costa Oeste de los Estados Unidos de América; colinda al Norte con el estado de Oregon, al Noreste con Nevada, al Sureste con el estado de Arizona, Al Oeste con el Océano Pacífico y al Sur tiene Frontera Internacional con México, siendo vecino del estado de Baja California Norte.

Antecedentes Históricos

California ha sido un estado de gran interés para los historiadores por el hecho de ser una tierra de grandes contrastes culturales. La historia del hombre en California es relativamente reciente; excepto por alguna evidencia de la presencia del hombre cerca de Calico en el Desierto de Mohave hace unos cincuenta mil años, la antigüedad generalmente aceptada del hombre en California es de treinta mil años, aún cuando unicamente se tiene evidencia directa de diez mil años.

Cuando los europeos llegaron, California tenía probablemente una densidad de población mayor que las otras áreas del Norte de México. El área de la historia

empezó con la exploración de Baja California bajo la dirección de Hernán Cortés en los 1530's; a pesar de que la península fué poco hospitalaria para los europeos, que primero la exploraron y después trataron de establecerse en ella, Juan Rodríguez Cabrillo dirigió una exploración hacia el Norte y el 28 de Septiembre de 1542 descubrió la Bahía de San Diego y se embarcó aún más hacia el Norte, llegando probablemente hasta Oregon.

Los primeros Anglo-americanos que llegaron a California fueron comerciantes de Boston y fueron seguidos rapidamente por otros comerciantes y cazadores alrededor de 1820. Antes de 1840 había menos de cien Anglo-americanos residiendo en la California Mexicana, pero para el final de 1846 este número se había incrementado hasta setecientos; estos residentes empezaron a explotar el potencial agrícola de la región ocasionando gran descontento entre los oficiales mexicanos.

La penetración de los Anglo-americanos facilitó la adquisición de California, que en 1846 se proclamaba como una República independiente. Antes de que un nuevo gobierno pudiera ser organizado ocurrió la guerra entre Estados Unidos y México, y California fué rapidamente ocu

pada por las fuerzas militares Norteamericanas, convirtiéndose finalmente en un estado de los Estados Unidos de América. Posteriormente se encontró oro en el estado y la --- "Fiebre de Oro" trajo nuevos inmigrantes de México, Chile, Perú, la costa Este de los Estados Unidos, Francia, Inglaterra y de lugares tan lejanos como Hawaii, Australia y -- China.

En la última dêcada del Siglo XIX, Japoneses, Filipinos e Indostanos se aumentaron al flujo de inmigrantes y al principio del Siglo XX empezó una constante inmigra-- ción de Mexicanos, legales e ilegales. Entre 1900 y 1920, California había recibido miles de trabajadores mexicanos, muchos de ellos fueron repatriados y sus trabajos en la -- agricultura fueron tomados por refugiados de Arkansas y -- Oklahoma.

A pesar de los salarios tan bajos pagados a los trabajadores en los campos, la afluencia de Mexicanos en - California se ha incrementado en grandes proporciones, produciendo desequilibrios tanto en la economía de México como en la de los Estados Unidos.

Consideraciones Generales

Dentro de los Estados Unidos de América, de acuerdo al Censo de 1970 (United States Bureau of the Census) -- había una población de orígen hispánico * de 9,294,509, -- sin embargo la inmensa mayoría de estas personas se encuentran en los estados del Sur; primero, porque estos estados pertenecieron alguna vez a México y segundo, por su situación geográfica y sus políticas respecto a la admisión de extranjeros y refugiados; sin embargo es importante hacer notar que aproximadamente el 33 por ciento de toda la población hispánica en los Estados Unidos se encuentra en -- California, aproximadamente 3,100,500 hispanos.

Otro punto importante es el hecho de que en California no solo hay grandes concentraciones de hispánicos, sino también de Japoneses, Vietnamitas y otras nacionalidades y grupos étnicos, esto es consecuencia de que California ha recibido refugiados de guerra en proporciones -- mucho más altas que los otros estados.

Dada tan heterogénea población en el estado, es natural que tanto el Gobierno Federal como el Departamen--

* Ethnic Groups and Public Education in California, Research Report Number Three.

to de Educación estén tratando de modificar día a día los sistemas de educación, con el fin de satisfacer las necesidades especiales de estos individuos, que tienen diferentes culturas, razas y formas de vida; sin embargo, las restricciones de presupuesto, entre otras cosas, no han permitido que estos esfuerzos hayan fructificado.

Si tomamos en consideración el hecho de que el trabajador hispánico en los Estados Unidos normalmente está empleado en actividades agrícolas, es natural que cerca de las zonas de cultivo la concentración de hispánicos sea mucho más alta que en las regiones dedicadas a otras cosas; además, dada la discriminación de que son objeto, tratan de establecer colonias o barrios hispánicos, de tal forma que en muchos casos, aún después de muchos años de residir en los Estados Unidos, no tienen dominio del idioma ni se han adaptado al tipo de vida.

Durante los últimos años, los Estados Unidos han intensificado el control en la frontera con México para evitar el paso de trabajadores ilegales; sin embargo muchos de ellos logran entrar al país, legalizar su situación, obtener empleos y establecerse con sus familias;

una vez hecho esto, deben inscribir a sus hijos menores en la escuela, pues la ley de California obliga a todas las personas menores de dieciseis años, o hasta terminar la preparatoria, a asistir a la escuela; sin embargo, de acuerdo a los datos reportados por el Departamento de Educación, el 55 por ciento de los estudiantes hispánicos inscritos en las escuelas públicas no terminan la preparatoria, esta cifra es, sin duda, indicativa de que los esfuerzos realizados para dar educación a estos estudiantes no han tenido éxito. Las causas principales son la situación económica del trabajador hispánico, que prefiere que sus hijos trabajen para ayudar a mejorar la situación familiar en vez de asistir a la escuela; y el hecho de que el estudiante hispánico en general no domina el idioma y por ende su capacidad de aprender y asimilar conocimientos en una escuela en donde las clases se imparten en inglés es muy reducida, lo cual lo coloca en una situación desventajosa con respecto a los estudiantes norteamericanos y le impide aspirar a empleos mejor remunerados una vez que necesita trabajar.

Entre los intentos de solucionar este proble-

ma, el Departamento de Educación ha tratado de establecer un sistema de Educación Bicultural y Bilingüe, mediante el cual se dá capacitación al personal docente para impartir sus clases tanto en inglés como en español, o en el idioma necesario de acuerdo a la población escolar. Una de las principales razones por la cual este sistema no ha dado buenos resultados es la falta de incentivos para los profesores, pues al tener que preparar sus clases en dos idiomas diferentes, la cantidad de trabajo aumenta al doble y los incrementos en sueldo no compensan esto; por esta razón, aún cuando hay muchos profesores con la capacidad para participar en estos programas, simplemente deciden no hacerlo; otra razón del fracaso es la discriminación, generalmente los profesores capacitados para participar en este tipo de programas son de orígen hispánico o asiático, que aún bajo el sistema tradicional de enseñanza reciben sueldos menores que los de los profesores anglos, y no están dispuestos a trabajar más para recibir apenas lo que un profesor anglo recibe sin hacer ningún esfuerzo extra.

Otra de las causas por las que este programa no ha arrojado los frutos esperados es el hecho de que

dentro del Sistema Educativo, el sector administrativo - ha crecido en una proporción relativamente alta con respecto al sector académico, y el ya bastante reducido presupuesto dedicado a la educación va, en su mayor parte, a pagar sueldos del personal administrativo dejando muy poco para los profesores; por esta razón, el número de - nuevos profesores ha sido bastante menor al esperado durante los últimos años.

Con este trabajo no pretendo en modo alguno solucionar uno de los muchos problemas que nuestros compatriotas, en su afán de superación, tienen que afrontar - al tratar de sobrevivir en un país extraño, sino hacer - un somero análisis del problema y mediante él, lograr obtener una medida de "Satisfacción de las necesidades de educación de la población extranjera en California".

Para hacer esto, haré uso de las herramientas estadísticas y mediante el análisis de regresión, trataré de encontrar una relación entre el número de estudiantes que no son norteamericanos, al menos en su orígen, y el número de profesores en California.

ANALISIS PRELIMINAR

California está dividida politicamente en 58 condados, en los cuales se encuentran 1052 distritos - escolares distribuídos de acuerdo a la población absoluta de cada condado y a las políticas del Departamento de Educación.

En Octubre de 1977 se les pidió a los Distritos escolares que entregaran a la Barra de Educación - un reporte con el número de estudiantes y profesores, clasificados por grupos étnicos. La Barra de Educación en 1966 empezó a llevar a cabo encuestas de este tipo para tomar las medidas necesarias para disminuir la segregación de los estudiantes pertenecientes a los llamados grupos minoritarios, así como para iniciar el desarrollo de programas Bilingues y Biculturales.

Los resultados de la encuesta de 1977 mostraron que 36.5% de los 4.3 millones de estudiantes en escuelas públicas de California eran miembros de grupos raciales identificados y grupos étnicos minoritarios, comparado con 25% en 1967; de 1,562,310 estudiantes de

los grupos minoritarios, 27.55% eran negros; 9.55% asiáticos o de las Islas del Pacífico, aproximadamente 3.3% Filipinos y 2.48% Indios Americanos, mientras que el grupo Hispánico alcanzaba el 57%.

A continuación se muestra una tabla en la que pueden observarse los siguientes hechos importantes: Todos los grupos minoritarios han aumentado, desde 15.6% los negros hasta 194% los indios americanos; el mayor incremento en términos absolutos fué el de los hispánicos, de 616,229 en 1967 a 892,113 en 1977, aproximadamente un 45% de incremento; es también importante notar que el único grupo que ha disminuído es el de los blancos, de 3,308,828 en 1967 a 2,722,995 en 1977, constituyendo un decremento de casi 18%.

NOTAS:

1) No se realizó encuesta racial en las escuelas públicas de California en 1975,

2) En 1969 y 1973 el grupo Filipino fué incluído dentro del grupo Asiático, aún cuando algunos pudieron haberse incluído con otros grupos.

DISTRIBUCION DE ALUMNOS EN ESCUELAS PUBLICAS DE CALIFORNIA

POR GRUPOS ETNICOS (KINDER- 6º AÑO DE BACHILLERATO)

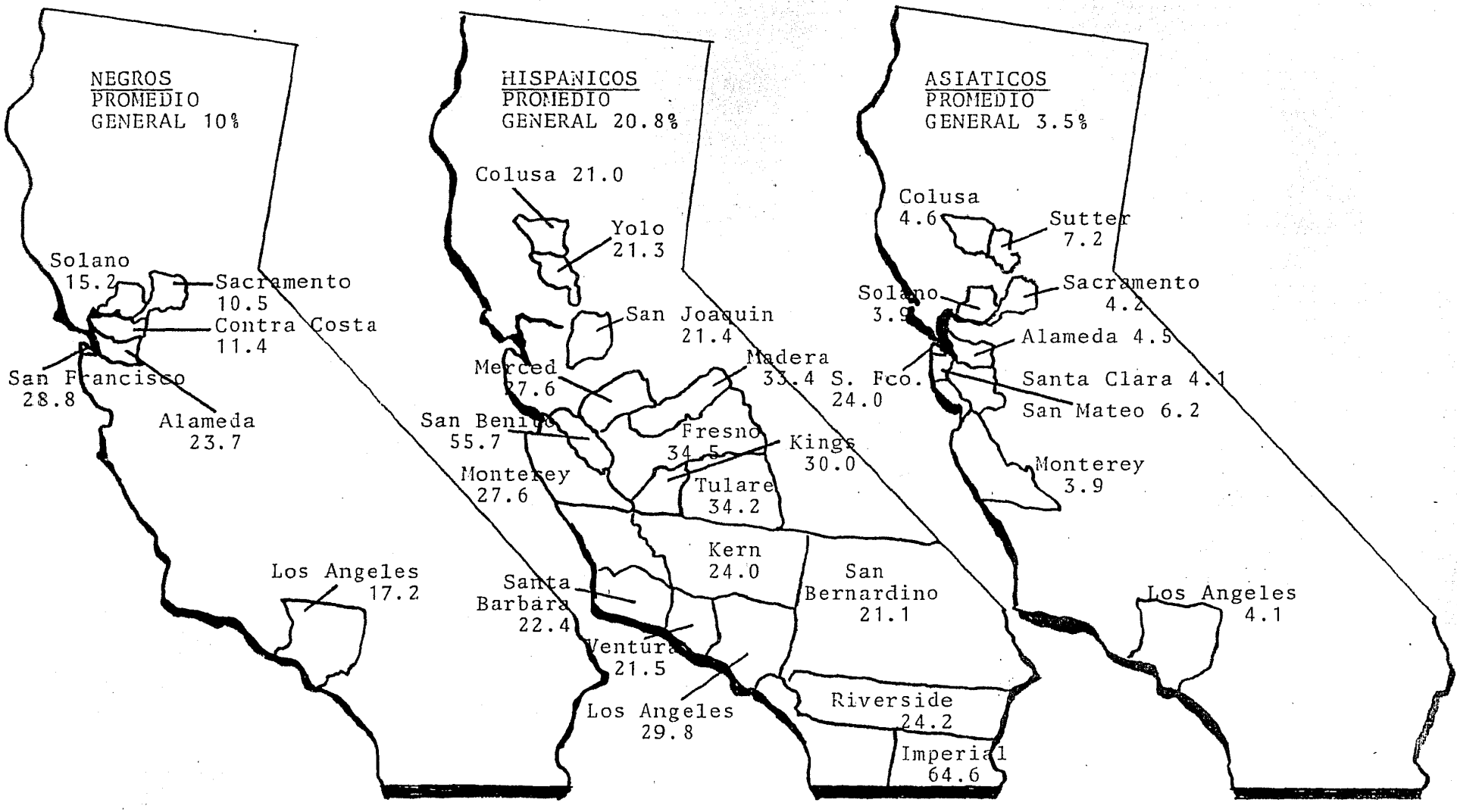
	INDIOS AMERICANOS	ASIATICOS	FILIPINOS	NEGROS	HISPANICOS	TOTAL GR. MINORIT.	BLANCOS	TOTAL
Oct. 1977	38,799	149,132	51,899	430,367	892,113	1,562,310	2,722,995	4,285,305
Porcentaje	0.9	3.5	1.2	10.0	20.8	36.5	63.5	
Oct. 1973	22,316	133,430	- -	432,418	765,419	1,353,583	3,088,758	4,442,341
Porcentaje	0.5	3.0		9.7	17.2	30.5	69.5	
Oct. 1971	19,319	97,978	49,704	422,945	725,227	1,315,173	3,230,106	4,545,279
Porcentaje	0.4	2.2	1.1	9.3	16.0	28.9	71.1	
Oct. 1969	15,663	96,845	- -	404,272	684,432	1,201,212	3,358,397	4,559,609
Porcentaje	0.3	2.1		8.9	15.0	26.3	73.7	
Oct. 1967	13,195	91,455	30,141	372,150	616,226	1,123,167	3,308,828	4,431,995
Porcentaje	0.3	2.1	0.7	8.4	13.9	25.3	74.7	

La distribución geográfica de los condados en los que el número de estudiantes negros, asiáticos e -- hispánicos excedió al promedio general del estado se -- muestra a continuación. La proporción de negros estuvo arriba del promedio en 6 condados: Alameda, Contra Cos- ta, Los Angeles, Sacramento, San Francisco y Solano; de los cuales unicamente el Condado de Los Angeles estuvo arriba del promedio en estudiantes hispánicos. Además - de Los Angeles, otros condados arrojaron datos de pobla ción hispánica arriba del promedio general: Colusa, -- Fresno, Imperial, Kern, Kings, Madera, Merced, Monterey, Riverside, San Benito, San Bernardino, San Joaquín, San ta Barbara, Tulare, Ventura y Yolo.

Los estudiantes asiáticos e isleños del Pacífico estuvieron arriba del promedio en Alameda, Los Angeles, Colusa, Monterey, Sacramento, San Francisco, San Mateo, Santa Clara y Sutter; y las escuelas en los 25 - condados que excedieron al promedio general tenían al - 74% del total de estudiantes en escuelas públicas en to do el estado de California.

De acuerdo a estos datos, parece ser que la -

CONDADOS CON INSCRIPCIONES ARRIBA DEL PROMEDIO DEL ESTADO.

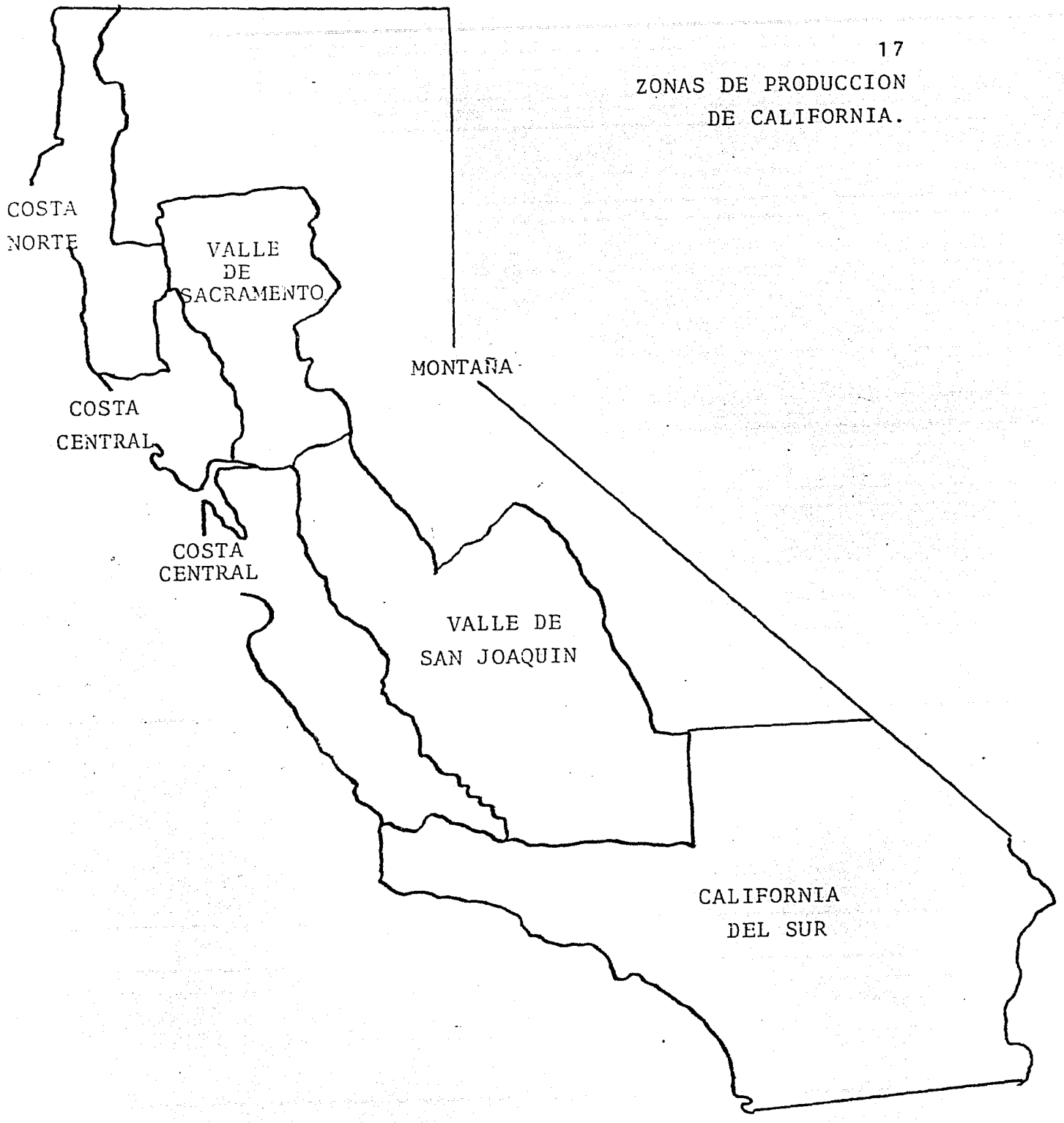


segregación de estudiantes pertenecientes a los grupos minoritarios continúa siendo una práctica común, a pesar de los esfuerzos que se han hecho por eliminarla.

En cuanto a la distribución de estudiantes hispánicos, existen dos factores que parecen ser determinantes: primero, la situación geográfica, los condados del sur en general tienen una gran población de estudiantes hispánicos, y segundo, la localización de las zonas agrícolas.

Con respecto a producción, California está dividida en seis regiones: Costa Norte, Costa Central, California del Sur, Valle de Sacramento, Valle de San Joaquín y Montaña. Dos de estas regiones son muy importantes en la producción agrícola, tanto en el estado como a nivel nacional: California del Sur y el Valle de San Joaquín; en la Costa Central y el Valle de Sacramento la agricultura tiene regular importancia y en la Costa Norte y la Montaña, la agricultura no constituye una rama de importancia.

ZONAS DE PRODUCCION
DE CALIFORNIA.



El Departamento de Educación de California ha clasificado a los estudiantes hispánicos en tres - grupos, de acuerdo al grado de dominio del idioma inglés:

a) EP.- Estudiantes con habilidad de hablar y entender inglés, y que por lo tanto no deben tener problemas en su desarrollo - académico, debidos al hecho de que el inglés no es su lengua materna.

b) LEP.- Estudiantes cuya capacidad de entender y hablar inglés es limitada y requieren por lo tanto de cursos especiales para así poder lograr un buen desarrollo académico.

c) NEP.- Estudiantes que no entienden ni hablan inglés, este grupo ha constituído - hasta el momento la mayor fuente de deserción estudiantil en lo que al grupo - hispánico se refiere y requiere, por este motivo, mucha atención.

En la siguiente tabla se muestran las propor

ciones de estudiantes hispánicos pertenecientes a los grupos LEP Y NEP como porcentaje del total de estudiantes hispánicos por condado:

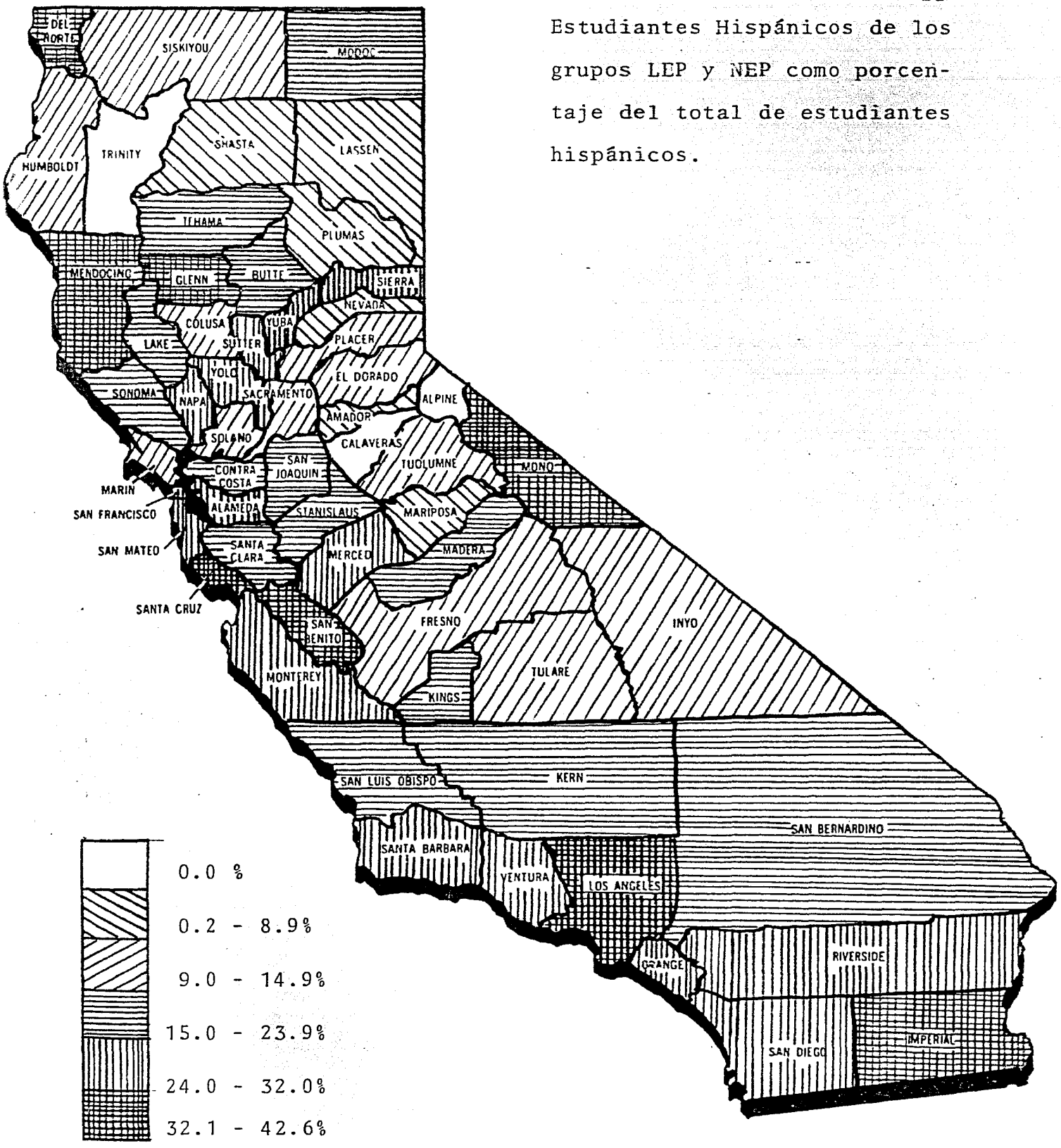
No.	CONDADO	LEP y NEP / TOT. HISP.
1	ALAMEDA	24.8
2	ALPINE	0.0
3	AMADOR	1.7
4	BUTTE	18.8
5	CALAVERAS	0.0
6	COLUSA	14.9
7	CONTRA COSTA	19.1
8	DEL NORTE	35.4
9	EL DORADO	12.8
10	FRESNO	14.0
11	GLENN	32.6
12	HUMBOLDT	9.9
13	IMPERIAL	34.0
14	INYO	10.1
15	KERN	18.6
16	KINGS	19.8
17	LAKE	21.9

No.	CONDADO	LEP y NEP TOT. HISP.
18	LASSEN	1.9
19	LOS ANGELES	33.1
20	MADERA	17.4
21	MARIN	12.6
22	MARIPOSA	6.6
23	MENDOCINO	35.7
24	MERCED	26.2
25	MODOC	22.6
26	MONO	32.6
27	MONTEREY	26.3
28	NAPA	30.5
29	NEVADA	1.4
30	ORANGE	28.3
31	PLACER	9.3
32	PLUMAS	3.8
33	RIVERSIDE	25.9
34	SACRAMENTO	13.4
35	SAN BENITO	37.2
36	SAN BERNARDINO	20.6
37	SAN DIEGO	28.3
38	SAN FRANCISCO	31.0

No.	CONDADO	LEP y NEP / TOT. HISP.
39	SAN JOAQUIN	18.2
40	SAN LUIS OBISPO	20.6
41	SAN MATEO	29.5
42	SANTA BARBARA	26.0
43	SANTA CLARA	17.9
44	SANTA CRUZ	42.6
45	SHASTA	0.2
46	SIERRA	32.0
47	SISKIYOU	14.3
48	SOLANO	14.7
49	SONOMA	18.0
50	STANISLAUS	21.2
51	SUTTER	26.8
52	TEHAMA	17.7
53	TRINITY	0.0
54	TULARE	14.5
55	TUOLOMNE	10.0
56	VENTURA	24.9
57	YOLO	29.3
58	YUBA	25.4

Fuente; 1980 R-30 LC. 1979 R-30 D/C
California Department of Education
Office of Bilingual-Bicultural Education
Data BICAL
Cortesía del Dr. Norman C. Gold

Estudiantes Hispánicos de los grupos LEP y NEP como porcentaje del total de estudiantes hispánicos.



Estimación de los parámetros del modelo mediante el método de MINIMOS CUADRADOS

El método que emplearé para estimar los parámetros del modelo es el llamado de Mínimos Cuadrados; en este método, los parámetros se determinan de tal manera que minimizan la cantidad llamada "Suma de los cuadrados de los residuales".

Los parámetros del modelo (β's) son desconocidos pero constantes y los estimadores ($\hat{\beta}$'s) son variables aleatorias que dependen de la muestra seleccionada en cada caso.

Supónganse dos cantidades X y Y, y la relación entre X y Y está dada por una función desconocida f , de tal forma que:

$$Y = f(X)$$

Sería extremadamente caro y tardado el realizar un censo sobre toda la población de interés y observar el valor de la variable Y para todos los valores posibles de X y así tener completamente defini-

da la función f (salvo errores de medición y otros factores). La Estadística nos permite, mediante la selección de una muestra estudiar la variable f y obtener una aproximación de ella sin necesidad de considerar todos y cada uno de los elementos de la población.

Suponiendo que tenemos una muestra de n elementos, observamos los valores x_i de X y los valores y_i de Y, entonces la relación entre X y Y puede ser escrita como sigue:

$$Y_i = f(x_i) + \xi_i$$

donde ξ_i es un error aleatorio que representa la variabilidad en el proceso de medición y debida a otros factores, que generalmente se asume que son despreciables.

Supongamos ahora que la forma de la función f puede ser aproximada mediante una línea recta, por lo tanto f(X) puede ser escrita como $\beta_o + \beta_1 x_i$ para algunos β_o y β_1 ,

$$f(x_i) = \beta_o + \beta_1 x_i + \delta_i$$

donde δ_i es un error fijo que refleja la falta de precisión de la línea recta en modelar la función f. Para que un modelo de regresión lineal sea útil, se requiere que la parte fija $\delta_{i's}$ sean pequeños comparados con la parte aleatoria $\xi_{i's}$.

El error se define entonces como $\varepsilon_i = \delta_i + \xi_i$ para obtener así el modelo de regresión lineal simple:

$$Y_i = \beta_o + \beta_1 x_i + \varepsilon_i \qquad i= 1,2,\ldots,n$$

donde los términos $\varepsilon_{i's}$ consisten de un componente fijo y un componente aleatorio.

Es importante hacer notar que durante el desarrollo anterior, se ha considerado que las $X_{i's}$ son medidas sin error; pues en general, el hecho de incluir errores en las mediciones de la variable X complica el análisis y siempre que sea posible es útil asumir que los errores en X son relativamente pequeños.

La estimación del modelo arriba mencionado mediante una línea recta es:

$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 x_i$$

El error "observable" e_i o residual está entonces definido como la diferencia del valor observado de Y menos el valor estimado de Y mediante la línea recta, es decir:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_o + \hat{\beta}_1 x_i)$$

y la suma de los cuadrados de los residuales $\sum_{i=1}^{n} e_i^2$ es la cantidad a minimizar mediante el método de mínimos cuadrados; para el caso de regresión lineal simple, los parámetros β_o y β_1 son estimados mediante:

$$\hat{\beta}_o = \frac{1}{n} \left(\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i \right) = \bar{y} - \hat{\beta}_1 \bar{x}$$

y

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{SXY}{SXX}$$

estos estimadores minimizan la suma $\sum_{i=1}^{n} e_i^2$.

Formalmente, el modelo de regresión lineal simple puede ser escrito como:

$$Y_i = \beta_o + \beta_1 x_i + \varepsilon_i$$

donde

Y_i es el valor de la variable dependiente Y en la observación i,

β_o y β_1 son parámetros,

X_i constante conocida, valor de la variable X en la observación i,

ε_i error aleatorio tal que $E(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = \sigma^2$; ε_i y ε_j no están correlacionados, i. e., $Cov(\varepsilon_i, \varepsilon_j) = 0$ para toda $i \neq j$.

Bajo estas condiciones $\hat{\beta}_o$ y $\hat{\beta}_1$ son estimadores insesgados y de varianza mínima entre la clase de estimadores lineales.

En el caso de regresión lineal múltiple, cuando se incluyen en el modelo dos o más variables independientes, los estimadores de mínimos cuadrados son obteni------

dos de la misma manera y en este caso minimizan la cantidad $SCE = \sum_{i=1}^{n} e_i^2$.

En este caso el modelo de regresión puede ser escrito como sigue:

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

donde

$\beta_o, \beta_1, \ldots, \beta_p$ son parámetros,

$X_1, X_2, \ldots, X_p$ son constantes desconocidas

$\varepsilon \sim NID(0, \sigma^2)$.

Usando notación matricial, tenemos que:

$$\underline{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\underline{\beta} = \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \underline{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_o \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

La matriz de datos $\underline{X}$ está definida como:

$$\underline{\bar{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & & x_{np} \end{bmatrix}$$

Sea X_i^τ el renglón i de la matriz $\underline{\bar{X}}$

$$\Rightarrow \quad \hat{Y}_i = X_i^\tau \hat{\underline{\beta}}$$

de aquí, el residual se define nuevamente como la diferencia del valor observado menos el valor estimado de Y_i, es decir

$$e_i = Y_i - \hat{Y}_i = Y_i - X_i^\tau \hat{\underline{\beta}}$$

de donde, finalmente podemos escribir el modelo de regresión lineal múltiple de la siguiente manera:

$$\hat{\underline{Y}} = \underline{\bar{X}} \hat{\underline{\beta}}$$

y el vector columna de residuales:

$$\underline{e} = \underline{Y} - \hat{\underline{Y}} = \underline{Y} - \underline{\bar{X}} \hat{\underline{\beta}}$$

por lo tanto la suma de los cuadrados de los residuales,

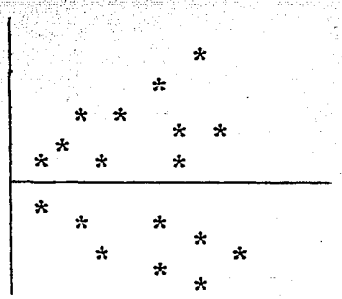
$$SCE = \underline{e}^\tau \underline{e} = (\underline{Y} - \hat{\underline{Y}})^\tau (\underline{Y} - \hat{\underline{Y}}) = (\underline{Y} - \underline{\bar{X}} \hat{\underline{\beta}})^\tau (\underline{Y} - \underline{\bar{X}} \hat{\underline{\beta}})$$

El estimador de mínimos cuadrados de está dado en este caso por:

$$\hat{\underline{\beta}} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y}$$

suponiendo que la matriz $(\underline{X}^T \underline{X})^{-1}$ existe.

En este caso particular, se hizo primero una regresión directa, pero la gráfica de los residuales mostró un patrón del tipo:



Por esta razón, y debido a que el modelo de regresión multiple requiere del hecho de que los errores tienen varianza constante σ^2 y que son independientes e identicamente - distribuídos, un análisis directo no resulta apropiado.

La gráfica indica que la suposición de que la - varianza de los errores es constante para toda i, puede - estar siendo violada, por lo tanto, en vez de una regre-- sión directa, utilizaré el método de ponderación y co---

mo variable ponderadora el número total de estudiantes por condado, esta variable es representativa del tamaño del -- condado y se espera que ayude a estabilizar las varianzas.

Bajo las condiciones mencionadas anteriormente - en el modelo de regresión múltiple, los estimadores de mínimos cuadrados son también los estimadores de máxima verosimilitud, son los estimadores insesgados de varianza mínima, consistentes y suficientes.

Sin embargo, raramente se tiene un conocimiento específico del comportamiento de las varianzas, pero en el caso de que las varianzas sean conocidas, o conocidas hasta cierta constante multiplicativa, existe una metodología para incorporar esta nueva información dentro del análisis. Supongamos que conocemos el valor de una matriz simétrica positiva definida Σ, tal que la matriz de covarianzas del vector $\underline{\varepsilon}$ está dada por $Cov(\underline{\varepsilon}) = \sigma^2\Sigma$, con $\sigma^2 > 0$ pero no necesariamente conocida.

Es razonable esperar que bajo estas circunstancias, el estimador $\hat{\underline{\beta}}$ obtenido mediante el método ordinario de mínimos cuadrados, aún cuando insesgado, ya no --

tendrá varianza mínima, pues ignora información muy im-- portante.

En el caso del método generalizado de mínimos cuadrados se tiene:

$$SCE = \underline{\varepsilon}^{T}\Sigma^{-1}\underline{\varepsilon}$$

el uso de este método reconoce que algunos de los resi-- duales son más importantes que otros; en particular, --- los residuales correspondientes a errores con mayor va-- rianza tienen menos importancia que aquellos correspon-- dientes a errores con menor varianza. En este caso el -- estimador $\hat{\underline{\beta}}$ está dado por:

$$\hat{\underline{\beta}} = (\underline{X}^{T}\Sigma^{-1}\underline{X})^{-1}\underline{X}^{T}\Sigma^{-1}\underline{Y}$$

Para el caso del método de ponderación, los e- lementos del vector $\underline{\varepsilon}$ no están correlacionados, pero las varianzas no necesitan ser todas iguales; la matriz Σ - está dada por:

$$\Sigma = \begin{bmatrix} W_1 & 0 & \cdots & 0 \\ 0 & W_2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & W_n \end{bmatrix}$$

donde σ^2 es generalmente un parámetro desconocido, pero las W_i's > 0 son "pesos" conocidos o proporciones relativas. Los casos con mayor W_i son relativamente más variables que los casos con pequeños W_i.

Como Σ es diagonal, también lo es Σ^{-1} ---

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{W_1} & & & 0 \\ & \frac{1}{W_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{W_n} \end{bmatrix}$$

de aquí

$$C = \begin{bmatrix} \frac{1}{\sqrt{W_1}} & & & 0 \\ & \frac{1}{\sqrt{W_2}} & & \\ & & \ddots & \\ 0 & & & \frac{1}{\sqrt{W_n}} \end{bmatrix}$$

y $C^TC = \Sigma^{-1}$; por lo tanto la matriz $\underline{\overline{W}} = C\underline{\overline{X}}$ y el vector $\underline{Z} = C\underline{Y}$ pueden ser obtenidos facilmente:

$$\underline{\overline{W}} = \begin{bmatrix} \frac{1}{\sqrt{W_1}} & \frac{x_{11}}{\sqrt{W_1}} & \frac{x_{12}}{\sqrt{W_1}} & \cdots & \frac{x_{1p}}{\sqrt{W_1}} \\ \frac{1}{\sqrt{W_2}} & \frac{x_{21}}{\sqrt{W_2}} & \frac{x_{22}}{\sqrt{W_2}} & \cdots & \frac{x_{2p}}{\sqrt{W_2}} \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{1}{\sqrt{W_n}} & \frac{x_{n1}}{\sqrt{W_n}} & \frac{x_{n2}}{\sqrt{W_n}} & & \frac{x_{np}}{\sqrt{W_n}} \end{bmatrix}$$

$$\underline{Z} = \begin{bmatrix} \frac{y_1}{\sqrt{W_1}} \\ \frac{y_2}{\sqrt{W_2}} \\ \vdots \\ \frac{y_n}{\sqrt{W_n}} \end{bmatrix}$$

Una vez hecho esto, el problema puede ser resuelto vía mínimos cuadrados ordinarios, utilizando la matriz $\underline{W}$ en lugar de la matriz $\underline{X}$ y el vector $\underline{Z}$ en vez del vector $\underline{Y}$.

La mayoría de los paquetes estadísticos implementados en las computadoras pueden resolver este tipo de problemas aún cuando no cuenten con la opción WEIGHT.

Lo que realmente se está haciendo mediante este procedimiento es un tipo de transformación de las variables en la que originalmente se supone que la varianza de los errores ε_i , $\text{Var}(\varepsilon_i) = \sigma^2 W_i$, al transformar el modelo de

$$Y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

al modelo

$$\frac{Y_i}{\sqrt{W_i}} = \frac{\beta_o}{\sqrt{W_i}} + \frac{\beta_1 x_{1i}}{\sqrt{W_i}} + \frac{\beta_2 x_{2i}}{\sqrt{W_i}} + \ldots + \frac{\beta_p x_{pi}}{\sqrt{W_i}} + \frac{\varepsilon_i}{\sqrt{W_i}}$$

Como $\text{Var}(\varepsilon_i) = \sigma^2 W_i$, de aquí se obtiene que

$$\text{Var}\left\{\frac{\varepsilon_i}{\sqrt{W_1}}\right\} = \frac{1}{W_i} \text{Var}(\varepsilon_i) = \sigma^2$$

y ya es entonces posible aplicar las mismas fórmulas computacionales que se utilizan en el método ordinario de mínimos cuadrados.

ANALISIS DE REGRESION

Mediante un análisis de regresión múltiple es posible encontrar la relación entre el número de profesores en California y el número de estudiantes, para -- así lograr obtener una medida de satisfacción de las necesidades de educación en el estado y determinar cuales de los grupos étnicos son estadisticamente significantes en esta relación.

Las variables a utilizar serán definidas posteriormente, las unidades de observación son los cin--- cuenta y ocho condados que componen California y se pretende que el modelo de regresión obtenido sea aplicable a las sub-unidades o unidades más pequeñas que son los distritos; sin embargo, de antemano, como resultado del análisis preliminar, sabemos que la distribución de los estudiantes (clasificados por grupos étnicos), no es - uniforme a lo largo de todo el estado y quizás sea conveniente considerar este hecho dentro de un futuro análisis, ya sea mediante una variable indicadora de la localización geográfica del condado o construyendo dos modelos diferentes para: regiones Norte y Sur de Califor-

nia. Con esto quedaría considerado también el factor agricultura, debido a su alta correlación con la separación del estado en regiones Norte y Sur.

Definición de las variables

Y Variable dependiente o respuesta, número de profesores calificados para impartir clases dentro del sistema de educación pública en California.

Variables independientes:

AMIN Número de estudiantes indios, nativos de los Estados Unidos de América.

ASFIL Número de estudiantes asiáticos, filipinos y de las Islas del Pacífico.

NEGRO Número de estudiantes negros, no se consideran dentro de este grupo los negros de orígen hispánico.

HISP Número de estudiantes hispânicos, este grupo incluye estudiantes Mexìcanos, Porto-Riqueños, Centro-Americanos, Cubanos, Latino-Americanos y personas de orígen español.

STUD Número total de estudiantes por condado, sin importar raza ni nacionalidad.

Dentro de las variables bajo estudio, se han considerado también las siguientes:

ANGLO Número de estudiantes Blancos, de descendencia Europea, considerados como los precursores de la raza Norte-Americana.

LOC Variable indicadora de la localización geográfica del condado bajo observación.

Los datos que se utilizarán para llevar a cabo el análisis se encuentran en la siguiente tabla:

No.	CONDADO	INDIOS AMERICANOS	ANGLOS	ASIATICOS/ FILIPINOS	NEGROS	HISPANICOS	TOTAL	PROFESORES
1	Alameda	1,789	114,832	13,309	47,011	21,870	198,811	9,749
2	Alpine	35	97	0	0	0	132	11
3	Amador	86	3,042	28	7	93	3,256	158
4	Butte	484	20,187	160	382	1,089	22,302	1,103
5	Calaveras	95	3,394	26	16	101	3,632	199
6	Colusa	15	1,894	124	35	551	2,619	155
7	Contra Costa	672	99,638	4,987	14,603	8,931	128,831	6,354
8	Del Norte	558	2,873	12	11	87	3,541	185
9	El Dorado	225	13,757	151	46	392	14,571	662
10	Fresno	897	59,384	2,451	6,723	36,569	106,024	5,290
11	Glenn	78	3,984	40	11	521	4,634	268
12	Humboldt	1,790	16,873	194	94	494	19,445	1,082
13	Imperial	354	6,865	468	602	15,064	23,353	1,139
14	Inyo	382	2,945	13	6	160	3,506	199
15	Kern	909	54,851	1,487	5,398	19,747	82,392	4,207
16	Kings	126	9,870	429	995	4,909	16,329	794

No.	CONDADO	INDIOS AMERICANOS	ANGLOS	ASIATICOS/ FILIPINOS	NEGROS	HISPANICOS	TOTAL	PROFESORES
17	Lake	153	4,757	37	39	197	5,183	260
18	Lassen	180	3,654	18	59	195	4,106	223
19	Los Angeles	6,952	618,184	65,912	224,574	388,540	1,304,162	58,984
20	Madera	267	7,093	105	514	4,013	11,992	625
21	Marin	78	34,676	1,054	957	896	37,661	1,901
22	Mariposa	70	1,514	12	4	61	1,661	77
23	Mendocino	769	11,293	101	92	605	12,860	697
24	Merced	95	18,469	531	1,852	7,979	28,926	1,397
25	Modoc	83	1,751	5	4	107	1,950	128
26	Mono	73	1,192	1	6	35	1,307	96
27	Monterey	411	29,075	4,183	3,341	14,093	51,103	2,506
28	Napa	140	14,623	433	110	1,318	16,624	825
29	Nevada	56	6,959	42	7	116	7,180	323
30	Orange	1,501	300,747	12,821	5,048	50,472	370,589	17,459
31	Placer	320	21,486	227	93	1,598	23,724	1,132
32	Plumas	219	2,838	16	38	91	3,202	189
33	Riverside	1,252	78,673	1,357	8,210	28,481	117,973	5,422

No.	CONDADO	INDIOS AMERICANOS	ANGLOS	ASIATICOS/ FILIPINOS	NEGROS	HISPANICOS	TOTAL	PROFESORES
34	Sacramento	2,574	104,748	7,249	15,094	14,785	144,450	7,140
35	San Benito	27	2,089	88	18	2,796	5,018	269
36	San Bernardino	1,358	116,932	2,355	10,609	35,109	166,363	7,527
37	San Diego	1,449	225,345	16,614	22,485	52,733	318,626	14,875
38	San Francisco	364	14,134	21,941	18,431	9,167	64,037	3,649
39	San Joaquin	395	42,971	4,006	4,788	14,190	66,350	3,239
40	San Luis Obispo	278	20,414	397	320	2,728	24,137	1,135
41	San Mateo	468	66,588	9,588	8,876	11,623	97,143	5,136
42	Santa Barbara	350	36,618	1,580	1,923	11,625	52,096	2,750
43	Santa Clara	1,925	183,234	14,385	10,986	51,575	262,105	12,576
44	Santa Cruz	196	22,551	799	280	6,247	30,073	1,414
45	Shasta	1,086	20,693	107	161	347	22,394	1,017
46	Sierra	16	624	1	0	31	672	65
47	Siskiyou	475	6,463	48	128	306	7,420	430
48	Solano	790	28,666	3,351	6,482	3,479	42,768	2,045
49	Sonoma	1,051	45,950	887	803	3,534	52,225	2,564

No.	CONDADO	INDIOS AMERICANOS	ANGLOS	ASIATICOS/ FILIPINOS	NEGROS	HISPANICOS	TOTAL	PROFESORES
50	Stanislaus	729	42,170	719	728	8,571	52,917	2,749
51	Sutter	60	8,173	769	87	1,190	10,279	530
52	Tehama	131	6,586	54	8	372	7,151	365
53	Trinity	148	2,045	11	2	33	2,239	135
54	Tulare	1,067	31,621	865	902	17,912	52,367	2,649
55	Tuolumne	362	5,012	46	16	214	5,650	298
56	Ventura	1,042	79,557	2,948	2,473	23,433	109,453	4,890
57	Yolo	319	13,567	619	291	3,958	18,754	973
58	Yuba	738	8,082	336	485	1,047	10,688	591
	T O T A L	38,512	2,706,603	200,497	427,264	886,380	4,258,956	202,810

En un intento de detectar si es que existe discriminación de los estudiantes hispánicos en cuanto a la asignación de profesores, se muestran a continuación una tabla con la proporción de estudiantes hispánicos por condado y la razón estudiantes/profesores, así como una gráfica de estas dos variables.

En esta gráfica no parece haber una tendencia marcada en cuanto a discriminación; sin embargo, tampoco parece observarse que el número de profesores aumente de acuerdo al incremento en la proporción de estudiantes hispánicos.

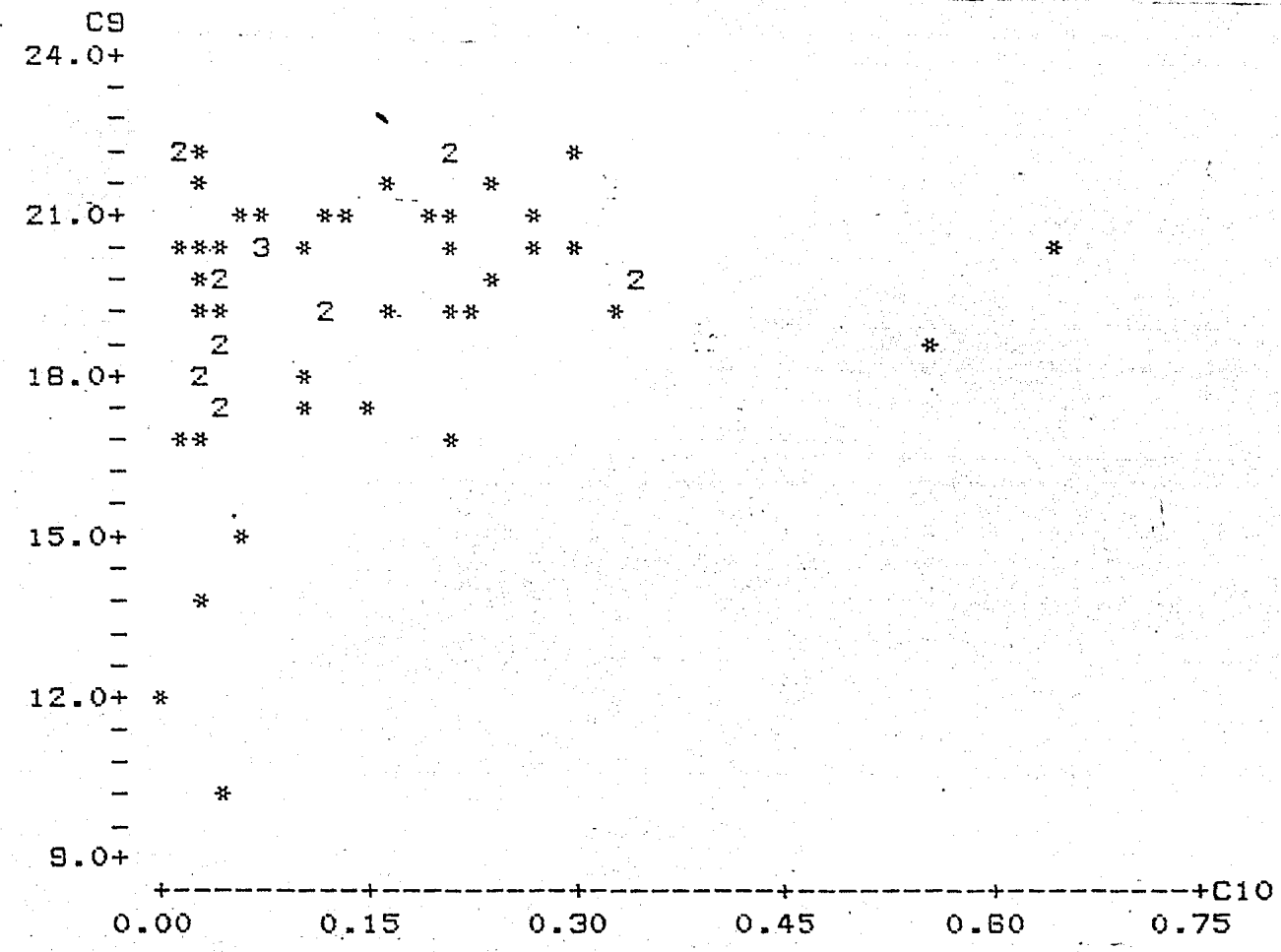
El Departamento de Educación de California trata de asignar aproximadamente 1 profesor por cada 19 estudiantes, sin embargo, en algunos condados no hay suficientes profesores para lograr esto y en otros la discriminación y la segregación siguen practicándose.

No.	CONDADO	LOCALIZACION	NUMERO DE DISTRITOS	RAZON EST./PROF.	PROPORCION HISPANICOS
1	Alameda	Sur	19	20.39	0.1100
2	Alpine	Norte	1	12.00	0.0000
3	Amador	Norte	3	20.61	0.0286
4	Butte	Norte	15	20.22	0.0488
5	Calaveras	Norte	4	18.25	0.0278
6	Colusa	Norte	4	16.90	0.2104
7	Contra Costa	Sur	18	20.28	0.0693
8	Del Norte	Norte	1	19.14	0.0246
9	El Dorado	Norte	15	22.01	0.0269
10	Fresno	Sur	53	20.04	0.3449
11	Glenn	Norte	10	17.29	0.1124
12	Humboldt	Norte	35	17.97	0.0254
13	Imperial	Sur	16	20.50	0.6451
14	Inyo	Sur	7	17.62	0.0456
15	Kern	Sur	49	19.58	0.2397
16	Kings	Sur	14	20.57	0.3006
17	Lake	Norte	7	19.93	0.0380
18	Lassen	Norte	11	18.41	0.0475
19	Los Angeles	Sur	82	22.11	0.2979
20	Madera	Sur	12	19.19	0.3346

No.	CONDADO	LOCALIZACION	NUMERO DE DISTRITOS	RAZON EST./PROF.	PROPORCION HISPANICOS
21	Marin	Norte	20	19.81	0.0238
22	Mariposa	Sur	1	21.57	0.0367
23	Mendocino	Norte	10	18.45	0.0470
24	Merced	Sur	21	20.71	0.2758
25	Modoc	Norte	3	15.23	0.0549
26	Mono	Sur	2	13.61	0.0268
27	Monterey	Sur	25	20.39	0.2758
28	Napa	Norte	5	20.15	0.0793
29	Nevada	Norte	11	22.23	0.0162
30	Orange	Sur	29	21.23	0.1362
31	Placer	Norte	19	20.96	0.0674
32	Plumas	Norte	1	16.94	0.0284
33	Riverside	Sur	24	21.76	0.2414
34	Sacramento	Norte	16	20.23	0.1024
35	San Benito	Sur	11	18.65	0.5572
36	San Bernardino	Sur	35	22.10	0.2110
37	San Diego	Sur	43	21.42	0.1655
38	San Francisco	Sur	1	17.55	0.1432
39	San Joaquin	Sur	18	20.48	0.2139

No.	CONDADO	LOCALIZACION	NUMERO DE DISTRITOS	RAZON EST./PROF.	PROPORCION HISPANICOS
40	San Luis Obispo	Sur	14	21.27	0.1130
41	San Mateo	Sur	23	18.91	0.1196
42	Santa Barbara	Sur	24	18.94	0.2231
43	Santa Clara	Sur	33	20.84	0.1968
44	Santa Cruz	Sur	11	21.27	0.2077
45	Shasta	Norte	28	22.02	0.0155
46	Sierra	Norte	1	10.34	0.0461
47	Siskiyou	Norte	30	17.26	0.0412
48	Solano	Norte	6	20.91	0.0813
49	Sonoma	Norte	49	20.37	0.0677
50	Stanislaus	Sur	29	19.25	0.1620
51	Sutter	Norte	12	19.39	0.1158
52	Tehama	Norte	18	19.59	0.0520
53	Trinity	Norte	12	16.59	0.0147
54	Tulare	Sur	49	19.77	0.3420
55	Tuolumne	Sur	12	18.96	0.0379
56	Ventura	Sur	20	22.38	0.2141
57	Yolo	Norte	5	19.27	0.2110
58	Yuba	Norte	5	18.08	0.0980

NUMERO DE ALUMNOS
POR PROFESOR.



PROPORCION DE
ESTUDIANTES
HISPANICOS.

Para comenzar el análisis, la matriz de correlación simple entre las variables, nos dá valiosa información sobre cuales de las variables independientes pueden sér útiles dentro del modelo.

La matriz de correlación de las variables originales es:

	AMIN	ANGLO	ASFIL	NEGRO	HISP	STUD	TEACH
AMIN	1.000						
ANGLO	0.899	1.000					
ASFIL	0.862	0.903	1.000				
NEGRO	0.883	0.861	0.939	1.000			
HISP	0.895	0.913	0.927	0.969	1.000		
STUD	0.922	0.978	0.949	0.945	0.976	1.000	
TEACH	0.923	0.980	0.951	0.942	0.971	0.999	1.000

y la matriz de correlación de las variables transformadas, después del método de ponderación, es:

	AMIN	ANGLO	ASFIL	NEGRO	HISP	STUD	TEACH
AMIN	1.0000						
ANGLO	0.6429	1.0000					
ASFIL	0.3725	0.6582	1.0000				
NEGRO	0.4159	0.6868	0.8126	1.0000			
HISP	0.4746	0.8157	0.6656	0.8159	1.0000		
STUD	0.6151	0.9730	0.7516	0.8173	0.9114	1.0000	
TEACH	0.6272	0.9740	0.7652	0.8113	0.9007	0.9984	1.000

Como se puede observar en la matriz de correlación, parece ser que la variable HISP es la que tiene el mayor poder de estimación sobre el número de profesores, pues la variable ANGLO no será considerada dentro del modelo, dado que el propósito de éste es estimar el número de profesores en función de la población estudiantil extranjera.

Selección de variables estadisticamente significantes

Existen varios métodos y criterios para tratar de decidir entre diferentes modelos, cual de ellos es el más apropiado; sin embargo, estos procedimientos llevan, en la mayoría de los casos, a diferentes modelos, y por lo tanto de ninguna manera se puede afirmar que cierto modelo es definitivamente el mejor.

A continuación se muestran los resultados obtenidos mediante el paquete estadístico BMDP(1R):

MODELO	$\hat{\beta}_0$	AMIN $\hat{\beta}_1$	ASFIL $\hat{\beta}_2$	NEGRO $\hat{\beta}_3$	HISP $\hat{\beta}_4$	g. l.	SCE	R^2
1	99.60462				0.18386	56	6,279,883	81.12%
2	- 63.83314	2.30913				56	20,181,765	39.34%
3	131.95611		0.56817			56	13,787,946	58.56%
4	149.76226			0.30375		56	11,371,068	65.82%
5	106.78126			0.08561	0.14578	55	5,698,232	82.87%
6	6.70957	0.94905			0.15889	55	4,567,110	86.27%
7	96.44361		0.22095		0.14343	55	4,638,788	86.06%
8	16.06099	0.92086		0.07760	0.12511	54	4,090,805	87.70%
9	95.76978		0.22758	-0.00691	0.14529	54	4,636,480	86.06%
10	11.08678	0.87551	0.20301		0.12200	54	3,191,999	90.41%
11	10.34897	0.87562	0.21015	-0.00745	0.12568	53	3,189,316	90.41%
0	192.76135					57	33,269,858	

Nota: Estos resultados fueron obtenidos utilizando el paquete estadístico BMDP, haciendo uso de la opción WEIGHT, y utilizando para este efecto la variable STUD, que es el número total de estudiantes por condado.

Es claro que al incrementar el número de variables, se logra aumentar el valor del coeficiente de determinación R^2, que es una medida descriptiva que mide la proporción de variabilidad en la variable dependiente Y (TEACH), que puede ser explicada por la inclusión del conjunto de variables independientes en el modelo:

$$R_i^2 = \frac{SCE_o - SCE_i}{SCE_o}$$

donde SCE_i es la suma de los cuadrados de los residuales en el modelo i. Sin embargo, los grados de libertad para la estimación de la varianza σ^2, disminuyen cada vez que una nueva variable es incorporada dentro del modelo y llega un momento en que el aumento logrado en el coeficiente de determinación R^2 y la disminución en la cantidad SCE no son suficientemente grandes como para justificar la inclusión de una nueva variable en el modelo, en este caso se dice que la variable no es estadisticamente significante.

Una forma de saber si una variable es estadisticamente significante cuando ya hay otras variables en el modelo, es mediante la prueba de F; por ejemplo si queremos probar el modelo 0 contra el modelo 1, tenemos:

$H_o : Y_i = 192.76135 + \varepsilon_i$

contra

$H_1 : Y_i = 99.60462 + 0.18386\ HISP + \varepsilon_i$

donde, para la hipótesis nula tenemos:

$gl_o^* = 57$ y $SCE_o = 33,269,858$

y para la hipótesis alternativa:

$gl_1 = 56$ y $SCE_1 = 6,279,883$

La estadística de prueba F, queda entonces definida como:

$$F_{gl_o - gl_1, gl_o} = \frac{(SCE_o - SCE_1)/ (gl_o - gl_1)}{SCE_1 / gl_1}$$

y la regla de decisión está dada por:
Se rechaza la hipótesis nula si el valor de la estadística de prueba F, es mayor que el valor $F^*_{gl_o - gl_1,\ gl_1}(\alpha)$.

Sea $\alpha = 0.01$, entonces

$$F_{1,56} = \frac{(33,269,858 - 6,279,883)/ (57-56)}{6,279,883/ 56} = 240.68$$

como 240.68 es mayor que $7.126 = F^*_{1,56}(.01)$, se rechaza la hipótesis nula con un nivel de significancia del .01 , por lo tanto la variable HISP es estadisticamente significante y debe ser incluída.

* gl_i = grados de libertad del error en el modelo i.

Es muy importante hacer notar que la variable HISP, por sí sola, logra explicar aproximadamente el 81% de la variabilidad en el número de profesores. Este hecho confirma la importancia que tiene la población hispánica, ya que ninguna otra de las variables independientes considerada logra alcanzar un coeficiente de determinación tan alto como éste.

De acuerdo a los coeficientes en la matriz de correlación, la siguiente variable que puede sernos útil es la variable NEGRO, sin embargo, no puede saberse aún si dicha variable será importante cuando la variable HISP está ya considerada dentro del modelo, debido a la alta correlación que existe entre ambas variables, (Corr(HISP, NEGRO)=0.8159).

En este caso tenemos:

H_1: $Y_i = 99.60462 + 0.18386\ HISP + \varepsilon_i$

H_5: $Y_i = 106.78126 + 0.08561\ NEGRO + 0.14578\ HISP + \varepsilon_i$

$gl_1 = 56$, $SCE_1 = 6,279,883$

$gl_5 = 55$, $SCE_5 = 5,698,232$

y la estadística de prueba:

$$F_{1,55} = \frac{(6,279,883 - 5,698,232)/ 1}{5,698,232/55} = 5.6141$$

como 5.6141 es menor que $F^{*}_{1,55}(.01) = 7.1375$, no podemos rechazar la hipótesis nula y por lo tanto se concluye que la variable NEGRO no es significante cuando ya la variable HISP está incluída dentro del modelo.

Siguiendo el mismo procedimiento con la variable ASFIL:

$$H_1 : Y_i = 99.60462 + 0.18386 \text{ HISP} + \varepsilon_i$$
$$H_7 : Y_i = 96.44361 + 0.22095 \text{ ASFIL} + 0.14343 \text{ HISP} + \varepsilon_i$$

$gl_1 = 56$, $SCE_1 = 6,279,883$

$gl_7 = 55$, $SCE_7 = 4,638,788$

$$F_{1,55} = 19.45 > 7.1375 = F^{*}_{1,55}\ (.01)$$

por lo tanto se rechaza la hipótesis nula y se agrega la variable ASFIL, por ser estadisticamente significante.

Considerando ahora la variable AMIN, modelo 7 contra modelo 10, se obtiene una $F_{1,54} = 24.47$, que es mayor -

al valor $F^{*}_{1,54}$ (.01) = 7.149; por lo tanto la variable - AMIN es también considerada.

De acuerdo a este procedimiento el modelo 10 parece ser al más apropiado; sin embargo, como se mencionó anteriormente existen varios métodos para selección de - variables, y este método en particular tiene la desventaja de que unicamente puede comparar modelos en los cuales el conjunto de variables incluídas en la hipótesis - nula es un subcónjunto de las variables incluídas en la hipótesis alternativa.

Además de este método, existe otro en el cual todas las variables son incorporadas en el modelo desde el principio y se van eliminando aquellas que no son significantes. Este método tiene la desventaja de que las variables que quedan finalmente en el modelo pueden ser colineares.

Con el fin de evitar estos dos problemas, el método que utilizaré para seleccionar el modelo definitivo es aquel conocido como "STEPWISE", que consiste en una - combinación de los dos métodos descritos anteriormente.

En este método se incorpora primeramente la variable más significante, (mayor R^2), y cada vez que una nueva variable es incluída en el modelo, se tiene la opción de sacar de él a aquella variable menos significante estadisticamente.

Las reglas fundamentales del método STEPWISE son las siguientes:

SW1.- Si hay al menos dos variables en el modelo y una o varias de ellas tienen un valor de F menor que F-EX*, la variable con la F más pequeña es excluída del modelo.

SW2.- Si hay dos o más variables en el modelo, la que tiene el valor de F más pequeño es excluída si su exclusión resulta en un valor de R^2 mayor al obtenido previamente para un modelo con el mismo número de variables.

SW3.- Si hay dos o más variables en el modelo, una de ellas es intercambiada por otra que

* F-EX .- Valor escogido previamente para excluír variables.

está fuera del modelo si este intercambio se traduce en un incremento en el coefi-- ciente de determinación R^2.

SW4.- Una variable es incluída en el modelo si tiene la mayor F; suponiendo, por supuesto, que este valor es mayor que la cantidad F-IN* y que el criterio de tolerancia se satisface.

Existen variantes en cuanto al orden de ejecución de estas reglas y en algunos casos no todas son -- aplicadas; el paquete BMDP permite 4 diferentes opciones: F (SW1, seguida por SW4); FSWAP (SW1, SW3 y después SW4); R (SW2, seguida por SW4) y RSWAP (SW2, SW3 y SW4).

Elección de las cantidades F-IN, F-EX y la tolerancia.

La tolerancia debe ser suficientemente baja, de tal forma que el algoritmo se detenga unicamente si real

* F-IN.- Valor escogido previamente para incluír variables en el modelo.

mente existe una dependencia lineal entre los datos ó si los estimadores resultantes estuvieran fuertemente influenciados por errores de redondeo. Para este caso se usará una tolerancia de 0.01, que en la mayoría de los casos resulta apropiada y es también la utilizada por el programa BMDP2R.

La elección de las cantidades F-IN y F-EX depende del criterio del analista, aún cuando en la mayoría de los paquetes estadísticos se usan F-IN=2 y F-EX=4; estos son los valores usados por BMDP2R; sin embargo algunos autores (Kennedy y Bancroft, 1971) sugieren usar el percentil 25 de la distribución F, obteniendo así valores de F-IN entre 2 y 4, y el percentil 10 de la distribución F para F-EX.

En este caso se utilizan F-IN=4 y F-EX=2, pues, como ya se mencionó, son los valores disponibles mediante BMDP.

Se considera primeramente el modelo sin variables independientes, para el cual tenemos: $SCE_{ex} = 33,269,858$ y $gl_{ex} = 57$.

El primer paso en este procedimiento es incluír la variable con mayor significancia; es decir, aquella que logra la mayor reducción en la cantidad SCE, que en este paso es la misma que tiene el mayor coeficiente de determinación R^2.

Los modelos a considerar en este paso son 1,2, 3 y 4 y las cantidades de interés son las siguientes:

MODELO	SCE_{in}	g.l.	VARIABLE
1	6,279,883	56	HISP
2	20,181,765	56	AMIN
3	13,787,946	56	ASFIL
4	11,371,068	56	NEGRO

Como puede observarse la inclusión de la variable HISP es la que logra la mayor reducción en la suma de los cuadrados de los errores, por lo tanto, como primer paso tenemos:

1er. paso: ¿ Incluir la variable HISP?

$$F = \frac{(33,269,858 - 6,279,883)/1}{6,279,883/56} = 240.68$$

Como F= 240.68 > 2 = F-IN, la variable HISP

es incluída dentro del modelo.

Consideramos para el segundo paso el modelo que incluye la variable HISP, es decir el modelo 1 y como modelos alternativos a aquellos que tienen otra variable -- además de HISP:

MODELO	SCE_{in}	g.l.	VARIABLES
5	5,698,232	55	HISP, NEGRO
6	4,567,110	55	HISP, AMIN
7	4,638,788	55	HISP, ASIL

2o. paso: ¿Incluír la variable AMIN?

$$F = \frac{(6,279,883 - 4,567,110)/1}{4,567,110/55} = 20.62 > 4 = F\text{-}IN$$

De aquí, se incluye la variable AMIN en el modelo.

3er. paso: ¿Excluír la variable HISP?

En caso de que la variable HISP fuera excluída del modelo, se obtendría otro modelo en el que la única variable sería AMIN, por lo tanto, en este caso tenemos: SCE_{ex} = 20,181,765 y $g.l._{ex}$ = 56; SCE_{in}=4,567,110 y

$g.l._{in} = 55.$

$$F = \frac{(20,181,756 - 4,567,110)/1}{4,567,110/\ 55} = 188.04 > 2 = F\text{-}EX$$

por lo tanto, la variable HISP es significante y debe ser conservada dentro del modelo.

4o. paso: ¿Incluír ASFIL?

Nótese que la suma de los cuadrados de los errores en el modelo 10, que incluye las variables AMIN, ASFIL e HISP, es 3,191,999 y los grados de libertad en este caso son 54, por lo tanto

$$F = \frac{(4,567,110 - 3,191,999)/\ 1}{3,191,999\ /\ 54} = 23.26 > 4 = F\text{-}IN$$

la variable ASFIL es también incluída dentro del modelo.

5o. paso: ¿ Excluír HISP?

En caso de que la variable HISP fuera excluída, tendríamos un modelo con las variables AMIN y ASFIL unicamente, sin embargo, este modelo no ha sido considerado dentro de los 12 modelos obtenidos, por lo que se supondrá --

que la variable HISP es significante aún cuando ya las va-- riables AMIN y ASFIL estén consideradas dentro del modelo, debido a que los coeficientes de determinación alcanzados por AMIN (0.3934) y ASFIL (0.5856) separadamente son muy -- pequeños comparados con el alcanzado por la variable HISP - por sí sola (0.8112).

6o. paso ¿Excluír AMIN?

En este caso tendríamos el modelo 7, cuyas únicas variables son ASFIL e HISP; SCE_7 = 4,638,788 y $g.l._7$ =55, - por lo tanto:

$$F = \frac{(4,638,788 - 3,199,999) / 1}{3,199,999 / 54} = 24.475 > 2 = F\text{-}EX$$

la variable AMIN es conservada dentro del modelo.

7o. paso: ¿Incluír NEGRO?

Si la variable NEGRO fuera incluída, obtendríamos el modelo 11, donde SCE_{11} = 3,189,316 entonces:

$$F = \frac{(3{,}191{,}999 - 3{,}189{,}316) / 1}{3{,}189{,}316 / 53} = 0.044 < 4 = F\text{-IN}$$

Por lo tanto, la variable NEGRO no es incluída dentro del modelo, por no ser estadisticamente significante después de que ya se han considerado las demás variables.

Como hemos visto, este método nos lleva también al modelo 10, que será, por el momento, considerado como el más apropiado.

Otro criterio utilizado en la selección de variables es la estadística C_p de Mallows, que se define como sigue:

$$C_p = \frac{SCE_p}{\hat{\sigma}^2} + 2p - n$$

donde p es el número de parámetros en el modelo y $\hat{\sigma}^2$ es la estimación de la varianza (SCE/ g.l. error) en el modelo que incluye todas las variables; n es el tamaño de muestra. Esta cantidad es, de hecho, una estimación de la cantidad J_p usando datos observados para estimar σ^2 y $ecm(\hat{Y}_i)$.

$$E(J_p) = \frac{1}{\sigma^2} ecm(\hat{Y}_i) = \frac{1}{\sigma^2} \Sigma(Var(\hat{Y}_i) + (E(\hat{Y}_i) - E(Y_i))^2)$$

donde $ecm(\hat{Y}_i)$ = Error cuadrático medio de $\hat{Y}_i = E(\hat{Y}_i - Y_i)^2$

y J_p está definida como sigue:

$$J_p = \frac{1}{\sigma^2} \left(\underbrace{\sum_{1=1}^{n} (\nu_i - \eta_i)^2}_{\text{Componente debido al sesgo}} + \underbrace{\sum_{i=1}^{n} \sigma^2(\hat{Y}_i)}_{\text{Componente debido al error aleatorio}} \right)$$

donde

$\nu_i = E(Y_i)$ de acuerdo con la relación de regresión real.

$\eta_i = E(Y_i)$ de acuerdo con el modelo ajustado

$\sigma^2(\hat{Y}_i)$ = varianza del valor estimado Y_i

σ^2 = Varianza real del error.

Existe también una relación entre C_p y la estadística F_p definida como sigue:

$$F_p = \frac{(SCE_p - SCE_{k'}) / (k'-p)}{SCE_{k'} / (n-k')}$$

donde p es el número de parámetros en el modelo mayor y k' el número de parámetros en el modelo menor.

La relación entre C_p y F_p está dada como sigue:

$$F_p = 1 + \frac{C_p - p}{k' - p}$$

y Mallows sugiere que un buen modelo tiene una estadística $C_p - p$ muy pequeña o negativa. A continuación se encuentra un cuadro con las estadísticas C_p y $C_p - p$ para los doce modelos obtenidos anteriormente:

MODELO	PARAMETROS (p)	Cp	$C_p - p$
1	2	50.36	48.36
2	2	281.38	279.38
3	2	175.13	173.13
4	2	134.96	132.96
5	3	42.69	39.69

MODELO	NUMERO DE PARAMETROS (p)	C_p	C_p-p	
6	3	23.90	20.90	
7	3	25.09	22.09	
8	4	17.98	13.98	
9	4	27.05	23.05	
10	4	3.04	-0.96	**
11	5	5.00	0.00	
0	1	496.88	495.88	

En vista de que, de acuerdo a los tres criterios anteriores, el modelo 10 parece ser más apropiado que los demás, el resto del análisis se llevará a cabo considerando unicamente dicho modelo.

Una vez que el modelo a utilizar ha sido determinado, es necesario realizar un análisis del comportamiento de los residuales para verificar si, en efecto, los supuestos necesarios para que el modelo sea válido se satisfacen (varianza constante, errores independientes e identicamente distribuídos como normal con media 0 y varianza σ^2, en el caso de mínimos cuadrados ordina-

rios y para el caso del método de ponderación los errores son independientes y no están correlacionados, sin embargo la suposición de varianza constante no es necesaria, pues ya dentro de la estimación de los coeficientes del modelo ha sido considerado el hecho de que las varianzas se incrementan de acuerdo con el tamaño de cada condado).

Para realizar este análisis, las gráficas de los residuales contra el valor estimado, así como contra cada una de las variables independientes son herramientas muy importantes para tratar de determinar si los supuestos se satisfacen; ya que de no ser así, las inferencias realizadas a partir del modelo pueden ser erróneas.

La prueba de F utilizada anteriormente es robusta ante no normalidad, pero puede perder validez fácilmente si las suposiciones acerca de la varianza no se satisfacen.

A continuación se muestran los resultados obtenidos mediante BMDP1R para este modelo:

COVARIANCE MATRIX

		ID	AMIN	ANGLO	ASFIL	NEGRO	HISP	STUD	TEACH	DIST
		1	2	3	4	5	6	7	8	10
ID	1	329.4								
AMIN	2	941.3	.4306E+05							
ANGLO	3	.3786E+05	.1389E+07	.1083E+09						
ASFIL	4	1866.	.7954E+05	.7050E+07	.1059E+07					
NEGRO	5	1715.	.1761E+06	.1459E+08	.1706E+07	.4164E+07				
HISP	6	7167.	.3686E+06	.3178E+08	.2563E+07	.6231E+07	.1401E+08			
STUD	7	.4954E+05	.2056E+07	.1632E+09	.1246E+08	.2687E+08	.5495E+08	.2595E+09		
TEACH	8	2591.	.9944E+05	.7746E+07	.6016E+06	.1265E+07	.2575E+07	.1229E+08	.5837E+06	
DIST	10	48.27	863.2	.3756E+05	1508.	2852.	10000.	.5279E+05	2587.	40.53

CORRELATION MATRIX

		ID	AMIN	ANGLO	ASEIL	NEGRO	HISP	STUD	TEACH	DIST
		1	2	3	4	5	6	7	8	10
ID	1	1.0000								
AMIN	2	0.2499	1.0000							
ANGLO	3	0.2004	0.6429	1.0000						
ASEIL	4	0.0999	0.3725	0.6582	1.0000					
NEGRO	5	0.0463	0.4159	0.6868	0.8126	1.0000				
HISP	6	0.1055	0.4746	0.8157	0.6656	0.8159	1.0000			
STUD	7	0.1695	0.6151	0.9730	0.7516	0.8173	0.9114	1.0000		
TEACH	8	0.1868	0.6272	0.9740	0.7652	0.8113	0.9007	0.9984	1.0000	
DIST	10	0.4177	0.6534	0.5668	0.2302	0.2196	0.4197	0.5147	0.5319	1.0000

VARIABLE NO.	NAME	MINIMUM LIMIT	MAXIMUM LIMIT	MISSING CODE	CATEGORY CODE	CATEGORY NAME	INTERVAL RANGE GREATER THAN	INTERVAL RANGE LESS THAN OR = TO
9	LOC							
					1.00000	SOUTH		
					2.00000	NORTH		

BMDP1R - MULTIPLE LINEAR REGRESSION

6TH JULY 1983 AT 13:26
REGRESSION TITLE IS

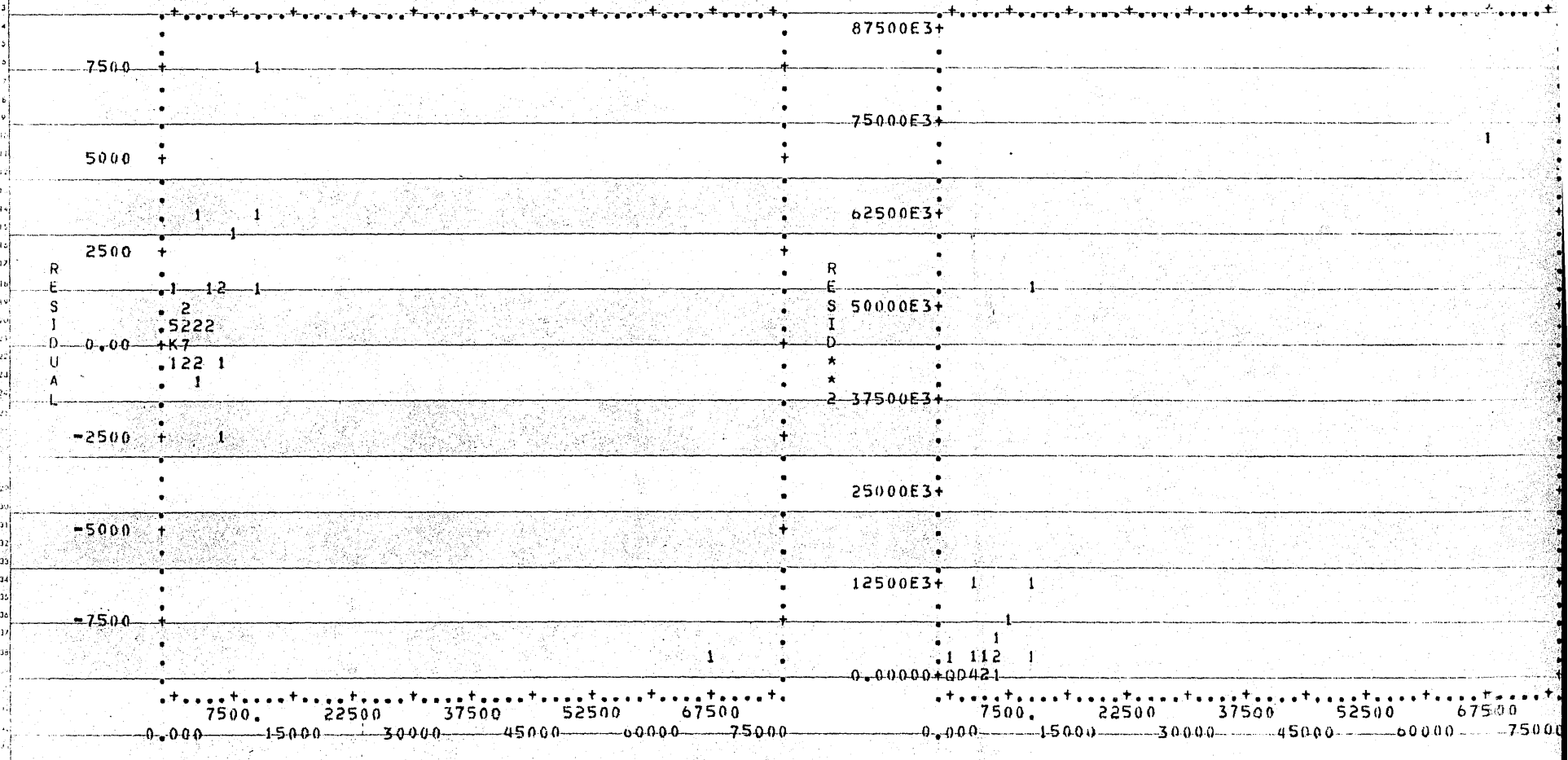
DEPENDENT VARIABLE. . . . . 8 TEACH
TOLERANCE 0.0100
ALL DATA CONSIDERED AS A SINGLE GROUP

MULTIPLE R 0.9508 STD. ERROR OF EST. 243.1277
MULTIPLE R-SQUARE 0.9041

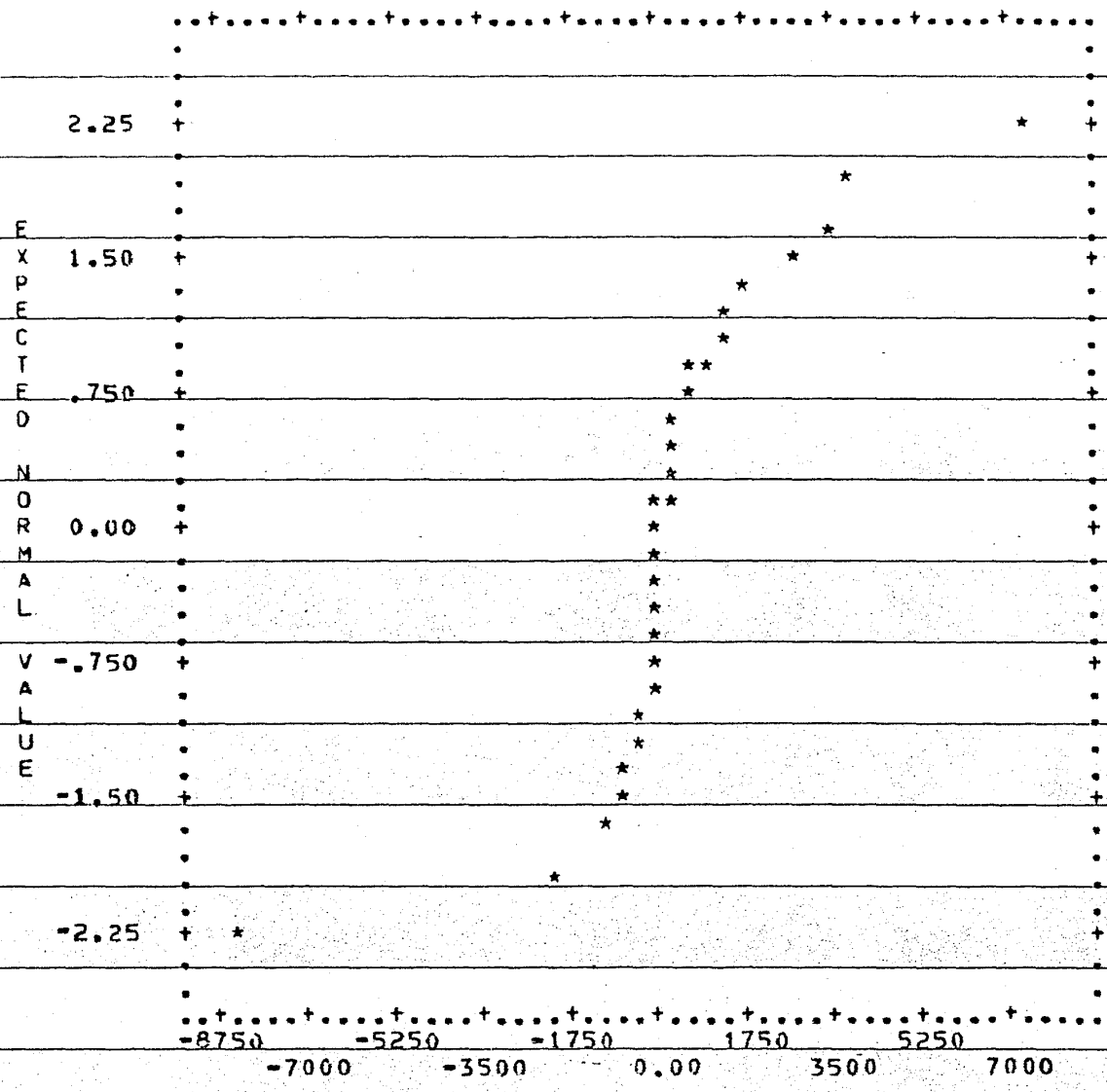
ANALYSIS OF VARIANCE

	SUM OF SQUARES	DF	MEAN SQUARE	F RATIO	P(TAIL)
REGRESSION	30077860.2239	3	10025953.4080	169.612	0.0000
RESIDUAL	3191999.4982	54	59111.1018		

VARIABLE		COEFFICIENT	STD. ERROR	STD. REG COEFF	T	P(2 TAIL)	TOLERANCE
INTERCEPT		11.08678					
AMIN	2	0.87551	0.17697	0.238	4.947	0.0000	0.76897
ASFIL	4	0.20301	0.04209	0.273	4.823	0.0000	0.55289
HISP	6	0.12367	0.01220	0.606	10.136	0.0000	0.49734


7500
5000
2500
0.00
-2500
-5000
-7500
RESIDUAL
7500.
22500
37500
52500
67500
0.000
15000
30000
45000
60000
75000
PREDICTD
87500E3+
75000E3+
62500E3+
50000E3+
37500E3+
25000E3+
12500E3+
0.00000+00
RESID
PREDICTD

NORMAL PROBABILITY PLOT OF RESIDUALS



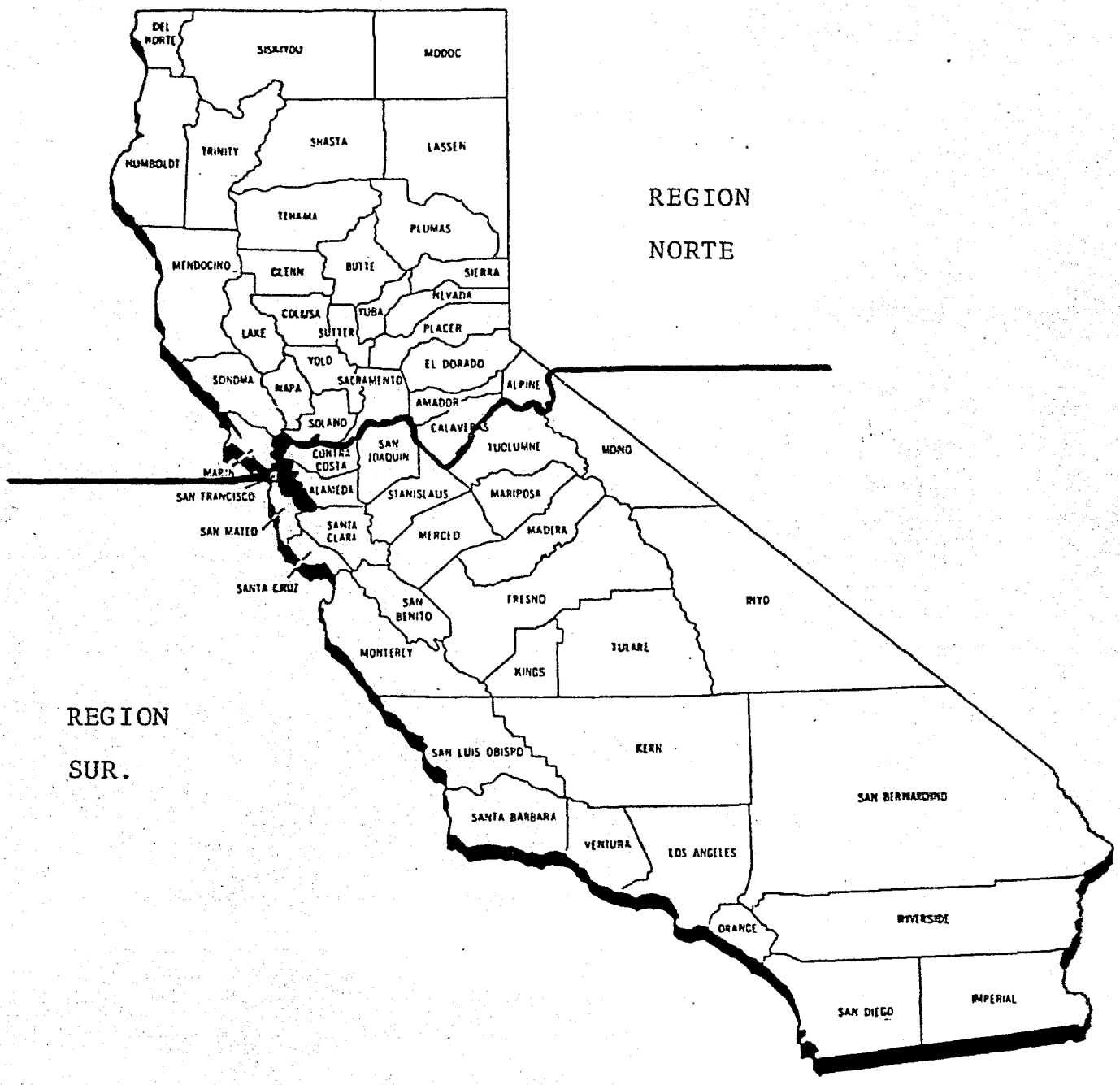
NUMBER OF INTEGER WORDS OF STORAGE USED IN PRECEDING PROBLEM 1424
CPU TIME USED IN PRECEDING PROBLEM 2.200 SECONDS
CUMULATIVE CPU TIME USED 61.217 SECONDS

En las gráficas anteriores se puede observar - que es muy posible que las varianzas no sean constantes, sin embargo, esta inconveniencia ya ha sido considerada al realizar un análisis ponderado en vez de un análisis directo, con lo cual se logra que las observaciones cu-- yos residuales provienen de errores con menor varianza - pesen más dentro del modelo, que aquellas en las que la varianza de los errores es mayor.

Al hacer la identificación de los residuales,. se observó que, en general, los mayores de ellos pertenecen a las observaciones en los condados del Sur, mientras que los residuales menores corresponden a los condados del Norte; este hecho es una indicación de que el factor LOCALIZACION puede ser una variable independiente importante que ha sido omitida en el modelo..En un - intento por mejorar dicho modelo se considerará este -- factor que, como podemos notar por simple inspección de los datos y por los resultados del análisis preliminar, tiene gran importancia; pues existe una gran diferencia, tanto en el número de profesores como en el de estudiantes, dependiendo de si se trata de un condado en el Sur ó en el Norte.

Para considerar este factor que, como ya se indicó anteriormente, está altamente correlacionado con las zonas agrícolas de California, se partirá la población en dos sub-poblaciones, de acuerdo a la localización geográfica. En cada una de las dos subpoblaciones se considerarán 29 condados y debido a que la división del estado en las 6 zonas de producción no permite, - en general, hablar de un condado como agrícola o no -- agrícola; pues hay condados que tienen una parte en la región agrícola mientras que la otra parte está localizada en la Montaña o en zonas que no son agrícolas por excelencia; la partición de la población se hará me--- diante una línea arbitraria, separando 29 condados al Norte y 29 al Sur. Esta partición se muestra en el siguiente mapa. y posteriormente se encuentran los resultados obtenidos después de correr los mismos modelos separadamente para las regiones Norte y Sur.

PARTICION DEL ESTADO EN REGIONES NORTE Y SUR.



MODELO	$\hat{\beta}_o$	AMIN $\hat{\beta}_1$	ASFIL $\hat{\beta}_2$	NEGRO $\hat{\beta}_3$	HISP $\hat{\beta}_4$	g. l.	SCE	R^2
1	56.23537				0.48364	27	597,558	74.66%
2	18.98180	1.05782				27	1,197,524	49.22%
3	83.05733		0.96183			27	816,879	65.36%
4	102.64907			0.45788		27	1,166,220	50.55%
5	60.61820			0.09022	0.42464	26	577,484	75.51%
6	16.58150	0.55475			0.38330	26	354,075	84.99%
7	60.73813		0.30312		0.36132	26	557,088	76.38%
8	20.59772	0.54709		0.07141	0.33798	25	341,543	85.52%
9	59.22830		0.47391	-0.08330	0.34689	25	552,821	76.56%
10	21.12600	0.55283	0.29670		0.26392	25	315,303	86.63%
11	17.35278	0.56708	0.60780	-0.15182	0.23510	24	301,291	87.23%
0	119.70679					28	2,358,469	

Resultados obtenidos para los 29 condados de la Región Norte.

MODELO	$\hat{\beta}_o$	AMIN $\hat{\beta}_1$	ASFIL $\hat{\beta}_2$	NEGRO $\hat{\beta}_3$	HISP $\hat{\beta}_4$	g. l.	SCE	R^2
1	135.41180				0.17811	27	12,970,729	85.15%
2	-494.16659	5.49555				27	27,257,848	68.80%
3	327.66815		0.53562			27	36,535,738	58.18%
4	378.51849			0.29124		27	27,225,311	68.84%
5	169.19406			0.07804	0.14291	26	11,558,177	86.77%
6	-121.54550	1.85160			0.13601	26	10,304,047	88.21%
7	130.49517		0.19686		0.14284	26	9,022,290	89.67%
8	-79.66573	1.77281		0.07149	0.10556	25	9,123,568	89.56%
9	128.35274		0.20125	-0.00470	0.14417	25	9,019,143	89.68%
10	-101.42389	1.6734	0.18452		0.10700	25	6,859,710	92.15%
11	-104.57246	1.6747	0.19060	-0.00651	0.10882	24	6,853,670	92.16%
0	581.32513					28	87,367,120	

Resultados obtenidos para los 29 condados de la Región Sur.

Después de aplicar el procedimiento STEPWISE para seleccionar las variables estadisticamente significantes en cada una de las dos regiones, se obtuvieron los siguientes modelos:

Para la región Norte (Modelo 6):

$$Y_i = 16.5815 + 0.55475\ AMIN + 0.3833\ HISP + \varepsilon_i$$

y para la región Sur (Modelo 10):

$$Y_i = -101.42389 + 1.6734\ AMIN + 0.18452\ ASFIL + 0.26392\ HISP + \varepsilon_i$$

donde, por principio, se puede notar que la variable - ASFIL (Estudiantes asiáticos y Filipinos) es estadisticamente significante en el Sur, pero no en el Norte.

Las variables AMIN e HISP son significantes tanto en el Norte como en el Sur, mientras que la variable NEGRO parece no tener influencia sobre el número de profesores.

Las estadísticas C_p y C_p-p se muestran a continuación:

REGION NORTE				REGION SUR			
MODELO	NUMERO DE PAR. (p)	C_p	C_p-p	MODELO	NUMERO DE PAR. (p)	C_p	C_p-p
1	2	22.60	20.60	1	2	20.42	18.42
2	2	70.39	68.39	2	2	70.45	68.45
3	2	40.07	38.07	3	2	102.94	100.94
4	2	67.90	65.90	4	2	70.34	68.34
5	3	23.00	20.00	5	3	17.47	14.47
6	3	5.20	2.20	6	3	13.08	10.18
7	3	21.38	18.38	7	3	8.59	5.59
8	4	6.21	2.21	8	4	10.95	6.95
9	4	23.04	19.04	9	4	10.58	6.58
10	4	4.12*	0.12	10	4	3.02*	-0.98
11	5	5.00	0.00	11	5	5.00	0.00
0	1	748.74	747.74	0	1	278.94	277.94

Como puede observarse en la tabla anterior, en ambos casos la estadística C_p es mínima para el modelo -- 10, sin embargo, bajo el procedimiento STEPWISE, se ob--- tiene que la variable ASFIL no es significante en la re-- gión Norte, por lo tanto para esta región se considerará el modelo 6 como el más apropiado. Además, como ya se mencionó anteriormente, la determinación del modelo depende del criterio y procedimientos aplicados y no se puede decir que un modelo es definitivamente mejor que otro, pues en muchas ocasiones los procedimientos conducen a diferentes modelos; pero, considerando el hecho de que es más fácil trabajar con un modelo que tiene menos variables, se utilizará el modelo 6 para la región Norte.

Debido a que el interés de este trabajo está en la importancia de la población estudiantil hispánica, para el resto del análisis se considerará unicamente la parte Sur de California, pues como ya se observó los modelos obtenidos difieren sustancialmente para el Norte y para el Sur, aún cuando se ha visto que la variable HISP es -- también significante en el Norte.

El modelo a considerar será entonces el siguiente:

$$Y_i = -101.42389 + 1.6734\ AMIN + 0.18452\ ASFIL + 0.26392\ HISP + \varepsilon_i$$

Interpretación de los estimadores de los parámetros.

En muchos problemas de regresión, el modelo es solo una ficción, que se sugiere para poder aplicar al problema, las técnicas de análisis de datos. Como resultado de esto, el estimador $\hat{\beta}$ puede no estimar una cantidad real, más aún si los datos fueron tomados sobre las mismas variables pero sobre diferentes rangos. Las estimaciones obtenidas tienen sentido sólo cuando la forma del modelo es una aproximación más o menos exacta de los datos. En general, el valor obtenido de $\hat{\beta}$ depende tanto del proceso que se usa para establecer el modelo como de la manera en que se obtuvieron los datos. Dependiendo de estos factores, la interpretación de los coeficientes de regresión puede complicarse.

Los estimadores pueden interpretarse de acuerdo a su magnitud y a su signo:

MAGNITUD.- Usualmente un estimador se interpreta

como una tasa de cambio, por ejemplo, en este caso el coeficiente 0.26392 de la variable HISP, indica que el incremento de una unidad en el número de estudiantes hispánicos aumentará el número de profesores en 0.26392, suponiendo que las demás variables en el modelo permanecen constantes; para esta interpretación, la suposición de que, de hecho una variable puede ser cambiada en una unidad sin afectar a las otras, es fundamental.

SIGNO.- El signo indica la dirección de la relación entre la variable independiente y la dependiente; en este caso, como era de esperarse, las variables independientes incluídas en el modelo están correlacionadas positivamente con la variable dependiente, pues es natural que mientras más grande sea el número de estudiantes, sin importar raza ni nacionalidad, más grande sea el número de profesores necesarios para enseñar a estos estudiantes. Sin embargo, esto no ocurre en todos los casos, si observamos los modelos 4, 9 y 11 para la región Sur, podemos notar que en el modelo 4 el signo del coeficiente de la variable NEGRO es positivo, mientras que en los modelos 9 y 11 es negativo; esto no implica que al aumentar el número de estudiantes negros el de profesores disminui-

rá, sino que la correlación entre las variables NEGRO y - TEACH ignorando las otras variables es positiva, mientras que la correlación parcial entre NEGRO y TEACH ajustada - por las otras variables, es negativa.

A continuación se muestran los resultados, así como las gráficas obtenidos mediante el paquete BMDP(1R) para el modelo de la región Sur:

COVARIANCE MATRIX

		ID	AMIN	ANGLO	ASFIL	NEGRO	HISP	STUD	TEACH	DIST
		1	2	3	4	5	6	7	8	10
ID	1	126.								
AMIN	2	448.	.7108E+05							
ANGLO	3	.3521E+05	.5049E+07	.5452E+09						
ASFIL	4	2635.	.3681E+06	.3850E+08	.6328E+07					
NEGRO	5	-1213.	.8896E+06	.8315E+08	.1017E+08	.2532E+08				
HISP	6	676.	.1904E+07	.1817E+09	.1501E+08	.3778E+08	.8375E+08			
STUD	7	.4376E+05	.8283E+07	.8536E+09	.7038E+08	.1573E+09	.3201E+09	.1410E+10		
TEACH	8	2156.	.3906E+06	.4015E+08	.3389E+07	.7375E+07	.1492E+08	.6623E+08	.3120E+07	
DIST	10	31.13	1696.	.1276E+06	6514.	.1437E+05	.5095E+05	.2012E+06	9691.	78.35

CORRELATION MATRIX

		ID	AMIN	ANGLO	ASFIL	NEGRO	HISP	STUD	TEACH	DIST
		1	2	3	4	5	6	7	8	10
ID	1	1.0000								
AMIN	2	0.1494	1.0000							
ANGLO	3	0.1339	0.8111	1.0000						
ASFIL	4	0.0930	0.5488	0.6555	1.0000					
NEGRO	5	-0.0214	0.6631	0.7076	0.8038	1.0000				
HISP	6	0.0648	0.7805	0.8501	0.6519	0.8203	1.0000			
STUD	7	0.1035	0.8274	0.9736	0.7452	0.8326	0.9316	1.0000		
TEACH	8	0.1084	0.8295	0.9735	0.7628	0.8297	0.9228	0.9986	1.0000	
DIST	10	0.3123	0.7188	0.6175	0.2925	0.3227	0.6289	0.6053	0.6198	1.0000

BMDP1R - MULTIPLE LINEAR REGRESSION

6TH JULY 1983 AT 13:26

REGRESSION FOR GROUP 1 SOUTH
REGRESSION TITLE IS

DEPENDENT VARIABLE. . . . . 8 TEACH
TOLERANCE 0.0100

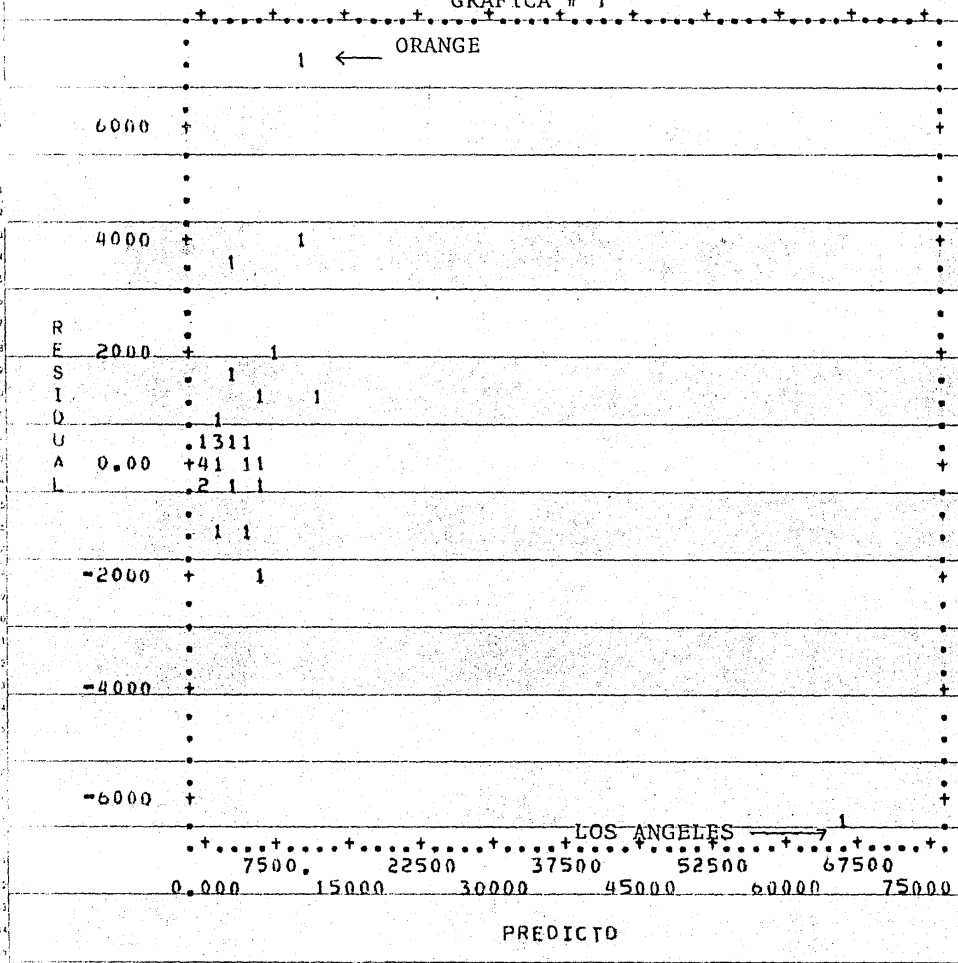
MULTIPLE R 0.9599 STD. ERROR OF EST. 523.8210
MULTIPLE R-SQUARE 0.9215

ANALYSIS OF VARIANCE

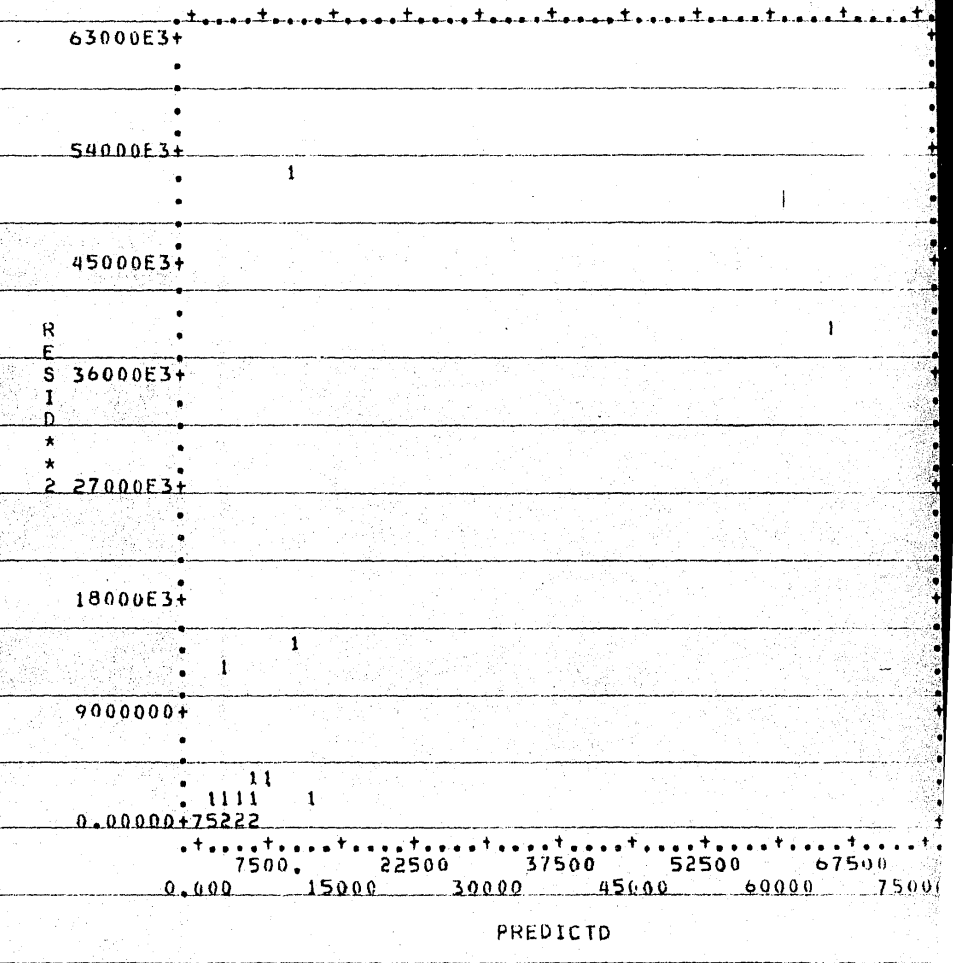
	SUM OF SQUARES	DF	MEAN SQUARE	F RATIO	P(TAIL)
REGRESSION	80507411.5416	3	26835803.8472	97.802	0.0000
RESIDUAL	6859710.0248	25	274388.4010		

VARIABLE		COEFFICIENT	STD. ERROR	STD. REG COEFF	T	P(2 TAIL)	TOLERANCE
INTERCEPT		-101.42389					
AMIN	2	1.67340	0.59607	0.253	2.807	0.0095	0.38802
ASEIL	4	0.18452	0.05208	0.263	3.543	0.0016	0.57094
HISP	6	0.10700	0.01914	0.554	5.590	0.0000	0.31930

GRAFICA # 1



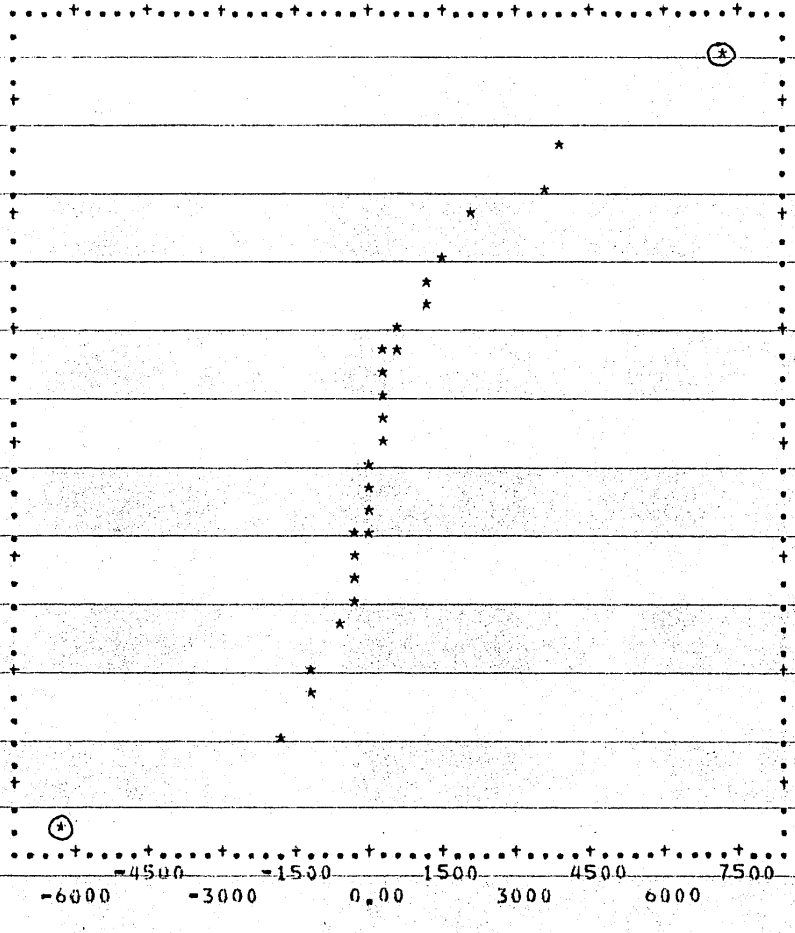
GRAFICA # 2



AGE 97 BMDP1R

GRAFICA # 3

NORMAL PROBABILITY PLOT OF RESIDUALS



En esta gráfica pueden notarse 2 puntos extremos aberrantes (Outliers), correspondientes a Los Angeles y Orange.

NUMBER OF INTEGER WORDS OF STORAGE USED IN PRECEDING PROBLEM 1424

CPU TIME USED IN PRECEDING PROBLEM 1.467 SECONDS

Prueba F de Linearidad

Para probar si, en efecto, existe una relación entre la variable dependiente Y (TEACH) y el conjunto de variables independientes AMIN, ASFIL e HISP, cuyos coeficientes en el modelo son $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\beta}_4$, es decir, probar:

$$H_o : \beta_1 = \beta_2 = \beta_4 = 0$$

contra

$$H_1 : \beta_i \neq 0 \text{ , para alguna i} \qquad i=1,2,4.$$

se usa la estadística de prueba F^* = CMR/ CME, que en este caso es igual a 97.802 y que comparada con el valor en tablas $F_{3,25}$, aún con un nivel de significancia de 0.001, $F_{3,25}\ (0.001) \doteq 7.55$, resulta mayor. Por lo tanto se rechaza la hipótesis nula y se dice que hay una relación, sin embargo, la existencia de esta relación, por sí sola, no asegura que puedan lograrse buenas estimaciones a partir del modelo.

A continuación se muestra una tabla de los residuales para cada una de las observaciones de la región Sur:

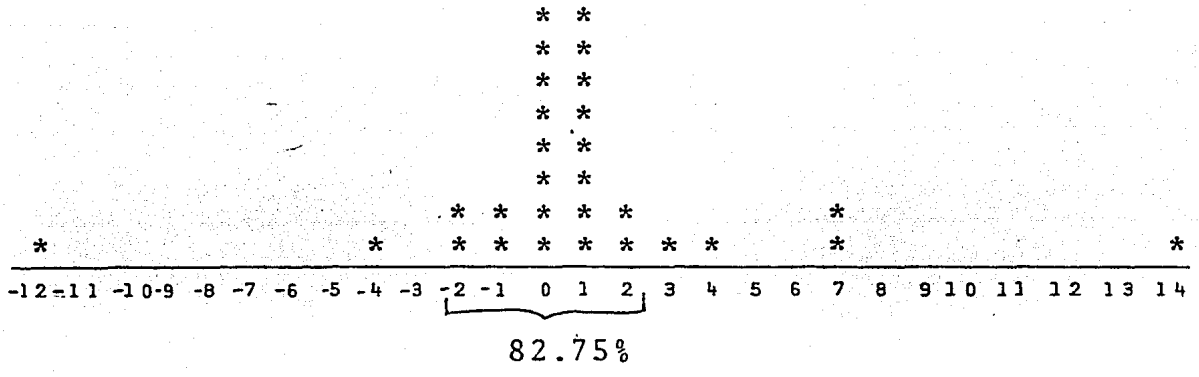
* CMR = Cuadrado medio de la regresión.
  CME = Cuadrado medio del error

CONDADOS DEL SUR, Tabla de Residuales

No.	CONDADO	Y (TEACH)	$\hat{Y}$	RESIDUAL	RESIDUAL ESTANDARIZADO
1	ALAMEDA	9,749	7,688	2,061	3.93
7	CONTRA COSTA	6,354	2,899	3,455	6.60
10	FRESNO	5,290	5,765	- 475	-0.91
13	IMPERIAL	1,139	2,189	-1,050	-2.00
14	INYO	199	557	- 358	-0.68
15	KERN	4,207	3,807	400	0.76
16	KINGS	794	714	80	0.15
19	LOS ANGELES	58,984	65,268	-6,284	-12.00
20	MADERA	625	794	- 169	0.32
22	MARIPOSA	77	24	53	0.10
24	MERCED	1,397	1,009	388	0.74
26	MONO	96	25	71	0.14
27	MONTEREY	2,506	2,866	- 360	0.69
30	ORANGE	17,459	10,177	7,282	13.90
33	RIVERSIDE	5,422	5,292	130	0.25
35	SAN BENITO	269	259	10	0.02
36	SAN BERNARDINO	7,527	6,362	1,165	2.22
37	SAN DIEGO	14,875	11,031	3,844	7.34

No.	CONDADO	Y (TEACH)	$\hat{Y}$	RESIDUAL	RESIDUAL ESTANDARIZADO
38	SAN FRANCISCO	3,649	5,537	-1,888	-3.60
39	SAN JOAQUIN	3,239	2,817	422	0.81
40	SAN LUIS OBISPO	1,135	729	406	0.78
41	SAN MATEO	5,136	3,695	1,441	2.75
42	SANTA BARBARA	2,750	2,020	730	1.39
43	SANTA CLARA	12,576	11,293	1,283	2.45
44	SANTA CRUZ	1,414	1,042	372	0.71
50	STANISLAUS	2,749	2,168	581	1.11
54	TULARE	2,649	3,760	-1,111	-2.12
55	TUOLOMNE	298	536	- 238	-0.45
56	VENTURA	4,890	4,694	196	0.37

Histrograma de los residuales estandarizados



Al observar, tanto la gráfica # 3 así como el - histograma de los residuales, podemos notar que es muy posible que los errores vengan de una distribución con co-- las más pesadas que las de una normal, ya que entre los - valores -2 y 2 se encuentran aproximadamente el 82% de -- ellos y no el 95%, que es el porcentaje aproximado esperado cuando los errores son normales.

Puede observarse también que hay 2 puntos extremos, los correspondientes a los condados de Los Angeles y Orange.

Después de revisar nuevamente los datos correspondientes a estos dos condados, se decidió eliminar al - condado de Los Angeles, debido a que por su gran tamaño y población, puede considerarse como un caso excepcional, - tanto en el número total de estudiantes como en el de profesores y las observaciones de las variables para este -- condado se encuentran en un rango muy alejado del resto de los condados.

Es muy importante mencionar que la omisión de - este condado no significa que no sea importante, sino al

contrario, Los Angeles es el único condado en todo California en el que el promedio de Negros, Asiáticos, Filipinos e Hispánicos está arriba del promedio general del Estado; precisamente por esta razón constituye un caso muy particular, que requiere atención individual en cuanto a la educación de los grupos minoritarios y no sería apropiado considerarlo dentro del modelo; además, dado que el rango de observación de las variables está tan alejado del de los demás condados, el tratar de estimar el valor de la variable dependiente TEACH para este condado sería casi una extrapolación del modelo.

Una vez omitido este condado, se obtuvo el modelo cuyos resultados se muestran a continuación:

BMDP1R - MULTIPLE LINEAR REGRESSION

6TH JULY 1983 AT 13:31

REGRESSION FOR GROUP 1 SOUTH

REGRESSION INTERCEPT	NON-ZERO
GROUPING VARIABLE	LOC
WEIGHT VARIABLE	W
PRINT COVARIANCE MATRIX	YES
PRINT CORRELATION MATRIX	YES
PRINT CORRELATION OF REGRESSION COEFFICIENTS	NO
PRINT RESIDUALS	NO
PRINT NORMAL PROBABILITY PLOT	YES
PRINT DETRENDED NORMAL PROBABILITY PLOT	YES

NUMBER OF CASES READ	57
CASES WITH GROUPING VALUES NOT USED	29
REMAINING NUMBER OF CASES	28

VARIABLE		WEIGHTED MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION	MINIMUM	MAXIMUM
1	ID	26.92906	11.26836	0.41844	1.00000	56.00000
2	AMIN	193.63888	238.18750	1.23006	27.00000	1925.00000
3	ANGLO	7491.26526	20688.56307	2.76169	1192.00000	300747.00000
4	ASFIL	455.56964	2231.87282	4.92046	1.00000	21941.00000
5	NEGRO	627.99072	3080.92775	4.90601	4.00000	47011.00000
6	HISP	2385.91940	6055.77005	2.53845	35.00000	52733.00000
7	STUD	11152.10409	29708.85333	2.66397	1307.00000	370589.00000
8	TEACH	503.48902	1429.89749	2.53758	77.00000	17459.00000
10	DIST	7.12247	8.75794	1.22962	1.00000	53.00000

COVARIANCE MATRIX

		ID	AMIN	ANGLO	ASFIL	NEGRO	HISP	STUD	TEACH	DIST
		1	2	3	4	5	6	7	8	10
ID	1	127.0								
AMIN	2	400.3	.5673E+05							
ANGLO	3	.3680E+05	.3751E+07	.4280E+09						
ASFIL	4	2804.	.2286E+06	.2591E+08	.4981E+07					
NEGRO	5	-653.3	.4119E+06	.3999E+08	.5550E+07	.9492E+07				
HISP	6	7656.	.1081E+07	.1073E+09	.7028E+07	.1046E+08	.3667E+08			
STUD	7	.4707E+05	.5529E+07	.6050E+09	.4369E+08	.6591E+08	.1025E+09	.8826E+09		
TEACH	8	2306.	.2062E+06	.2092E+08	.2184E+07	.3245E+07	.7798E+07	.4242E+08	.2045E+07	
DIST	10	31.37	1539.	.1134E+06	4972.	9088.	.4187E+05	.1708E+06	8321.	76.70

CORRELATION MATRIX

		ID	AMIN	ANGLO	ASEIL	NEGRO	HISP	STUD	TEACH	DIST
		1	2	3	4	5	6	7	8	10
ID	1	1.0000								
AMIN	2	0.1737	1.0000							
ANGLO	3	0.1579	0.7611	1.0000						
ASEIL	4	0.1115	0.4300	0.5610	1.0000					
NEGRO	5	-0.0186	0.5613	0.6274	0.8072	1.0000				
HISP	6	0.1122	0.7495	0.6564	0.5200	0.5608	1.0000			
STUD	7	0.1406	0.7813	0.9842	0.6590	0.7200	0.9034	1.0000		
TEACH	8	0.1431	0.7817	0.9777	0.6843	0.7367	0.9006	0.9965	1.0000	
DIST	10	0.3178	0.7376	0.6256	0.2544	0.3368	0.7895	0.6565	0.6645	1.0000

BMDP1R - MULTIPLE LINEAR REGRESSION

6TH JULY 1983 AT 13:33

REGRESSION FOR GROUP 1 SOUTH
REGRESSION TITLE IS

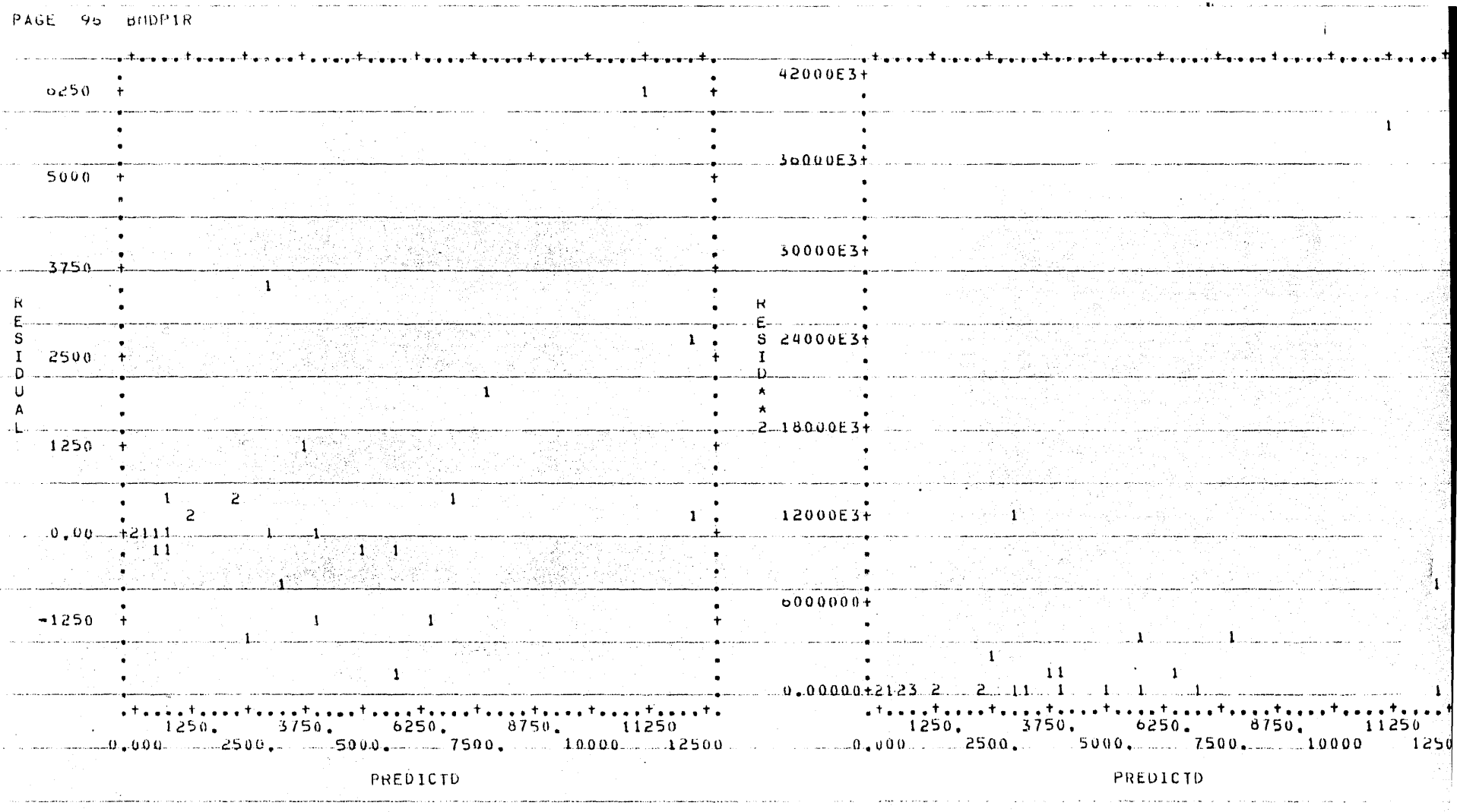
DEPENDENT VARIABLE. . . . . 8 TEACH
TOLERANCE. 0.0100

MULTIPLE R 0.9464 STD. ERROR OF EST. 489.8925
MULTIPLE R-SQUARE 0.8957

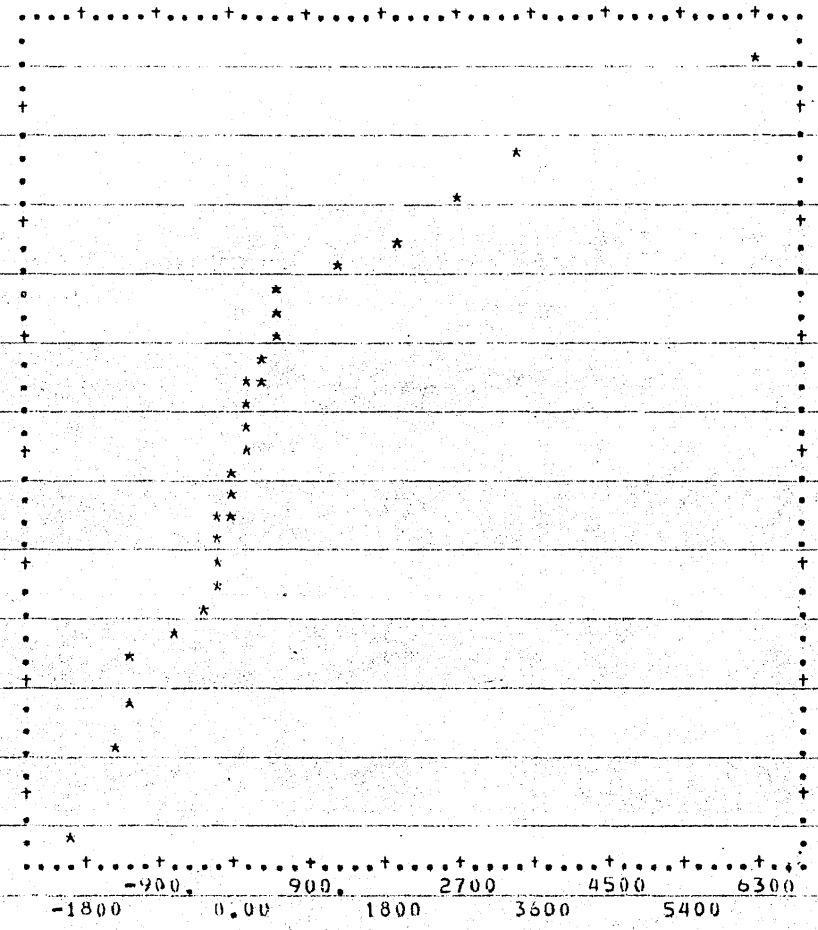
ANALYSIS OF VARIANCE

	SUM OF SQUARES	DF	MEAN SQUARE	F RATIO	P(TAIL)
REGRESSION	49444512.1968	3	16481504.0656	68.674	0.0000
RESIDUAL	5759872.1751	24	239994.6740		

VARIABLE		COEFFICIENT	STD. ERROR	STD. REG COEFF	T	P(2 TAIL)	TOLERANCE
INTERCEPT		-104.11475					
AMIN	2	1.30526	0.59946	0.217	2.177	0.0395	0.43600
ASFIL	4	0.18201	0.04958	0.284	3.671	0.0012	0.72592
HISP	6	0.13929	0.02492	0.590	5.589	0.0000	0.39027


RESIDUAL
6250
5000
3750
2500
1250
0.00
-1250
1250.
2500.
3750.
5000.
6250.
7500.
8750.
10000
11250
12500
0.000
PREDICTD
RESID**2
42000E3
36000E3
30000E3
24000E3
18000E3
12000E3
6000000
0.00000
1250.
2500.
3750.
5000.
6250.
7500.
8750.
10000
11250
12500
0.000
PREDICTD

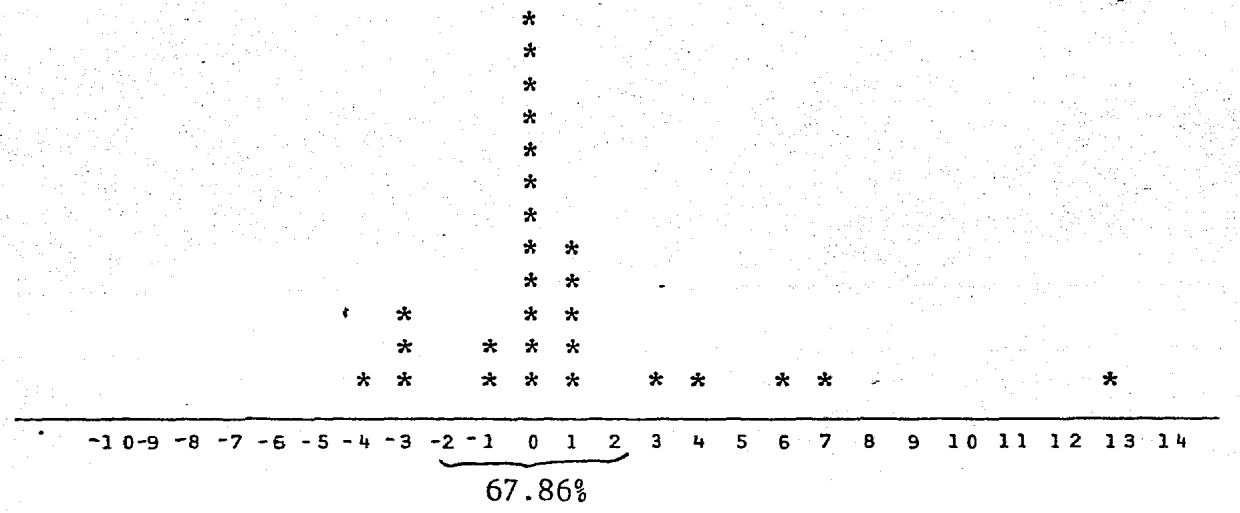
NORMAL PROBABILITY PLOT OF RESIDUALS



NUMBER OF INTEGER WORDS OF STORAGE USED IN PRECEDING PROBLEM 1424
CPU TIME USED IN PRECEDING PROBLEM 1.47% SECONDS

No.	CONDADO	Y (TEACH)	$\hat{Y}$	RESIDUAL	RESIDUAL ESTANDARIZADO
40	SAN LUIS OBISPO	1,135	711	424	0.87
41	SAN MATEO	5,136	3,871	1,265	2.58
42	SANTA BARBARA	2,750	2,260	490	1.00
43	SANTA CLARA	12,576	12,211	365	0.75
44	SANTA CRUZ	1,414	1,167	247	0.50
50	STANISLAUS	2,749	2,172	577	1.18
54	TULARE	2,649	3,941	-1,292	-2.64
55	TUOLOMNE	298	407	- 109	-0.22
56	VENTURA	4,890	5,057	- 167	-0.34

Histograma de los residuales estandarizados



CONDADOS DEL SUR, excluyendo a Los Angeles

Tabla de Residuales

No.	CONDADO	Y (TEACH)	$\hat{Y}$	RESIDUAL	RESIDUAL ESTANDARIZADO
1	ALAMEDA	9,749	7,700	2.049	4.18
7	CONTRA COSTA	6,354	2,925	3,429	7.00
10	FRESNO	5,290	6,607	-1,317	-2.69
13	IMPERIAL	1,139	2,541	-1,402	-2.86
14	INYO	199	419	- 220	-0.45
15	KERN	4,207	4,104	103	0.21
16	KINGS	794	822	- 28	-0.06
20	MADERA	625	822	- 197	-0.40
22	MARIPOSA	77	- 2	79	0.16
24	MERCED	1,397	1,228	169	0.34
26	MONO	96	- 4	100	0.20
27	MONTEREY	2,506	3,157	- 651	-1.33
30	ORANGE	17,459	11,219	6,240	12.74
33	RIVERSIDE	5,422	5,744	- 322	-0.66
35	SAN BENITO	269	337	- 68	-0.14
36	SAN BERNARDINO	7,527	6,987	540	1.10
37	SAN DIEGO	14,875	12,156	2,719	5.55
38	SAN FRANCISCO	3,649	5,641	-1,192	-4.07
39	SAN JOAQUIN	3,239	3,117	122	0.25

Nuevamente, después de omitir al condado de Los Angeles, las gráficas y el histograma muestran que los errores pueden venir de una distribución con colas más pesadas que las de una normal; pero como ya antes se mencionó, las pruebas de F son robustas ante no normalidad, y el tamaño de muestra (n=28) es suficientemente grande como para que esta desviación de normalidad no constituya un problema serio.

Pruebas de Hipótesis e Intervalos de Confianza

Para probar si los parámetros del modelo son diferentes de cero, se usa la siguiente prueba:

$$t^* = \frac{\hat{\beta}_k}{s(\hat{\beta}_k)} \quad , \qquad k=0,1,2,\ldots,p$$

donde

$s(\hat{\beta}_k)$ es la desviación estándar estimada para el estimador $\hat{\beta}_k$ y la regla de decisión es como sigue:

Rechazar la hipótesis nula si $|t^*| > t(1-\alpha/2, n-p)$ donde p es el número de parámetros en el modelo, n el tamaño de muestra y α el nivel de significancia de la prueba.

Para el caso del parámetro β_2, tenemos:

$H_o : \beta_2 = 0$

$H_a : \beta_2 \neq 0$

con n= 28, p= 4 y sea α= 0.01

$$t^* = \frac{1.30526}{0.59946} = 2.177 < 2.797 = t(1-.01/2,24)$$

$\therefore$ la hipótesis nula no se rechaza.

Es curioso observar que, aún cuando mediante el procedimiento STEPWISE se obtuvo que la variable AMIN era significante, en la prueba de hipótesis no se puede rechazar que β_2 sea igual a cero, esto se debe al nivel de significancia considerado aquí y a la elección de las cantidades F-IN y F-EX que se utilizaron durante la selección de las variables; además la desviación estandar estimada para $\hat{\beta}_2$ es relativamente grande en comparación con el estimador $\hat{\beta}_2$. En cualquier caso, la consideración de esta variable dependerá del criterio del lector, para lo cual se menciona el nivel de significancia descriptivo, que es el nivel de significancia más pequeño para el cual la hipótesis nula puede ser rechazada, en este caso:

$$0.01 \leq \text{Nivel de significancia descriptivo} \leq 0.025$$

Para el caso del parámetro β_4:

$H_o : \beta_4 = 0$

$H_a : \beta_4 \neq 0$

La estadística de prueba t*= 0.18201/ 0.04958 = 3.67 , y el nivel de significancia descriptivo en este caso está entre 0.005 y 0.0005, por lo que a un nivel de significancia de 0.005 la hipótesis nula puede ser rechazada.

Para el caso de la variable HISP, cuyo coeficiente es β_6 la estadística t* es:

$t^* = 0.13929 / 0.02492 = 5.589$

y el nivel de significancia descriptivo es menor o igual que 0.0005.

Mediante estas pruebas, se puede inferir que las variables utilizadas en el modelo tienen influencia sobre la variable TEACH, y nuevamente se tiene evidencia de que los estudiantes hispánicos tienen una gran impor-

tancia en la determinación del número de profesores.

Mediante el procedimiento anterior, al probar si los parámetros son iguales o diferentes de cero, unicamente se trata de ver si en realidad existe una relación entre cada una de las variables independientes y la variable dependiente, es decir un sí o no a la relación, pero no se tiene una aproximación confiable de cuales son los parámetros ; para esto es conveniente obtener intervalos de confianza para dichos parámetros.

Para la determinación de los intervalos de confianza, se utilizará un nivel se significancia $\alpha = 0.01$; estos intervalos están dados por:

$$\hat{\beta}_k \pm t(1- \alpha/2, n-p)\, s(\hat{\beta}_k)$$

de aquí, los intervalos de confianza del 99% son:

para β_2 : $1.30526 \pm 2.797\ (0.59946) = (-0.3714, 2.9819)$

para β_4 : $0.18201 \pm 2.797\ (0.04958) = (0.0433, 0.3207)$

para β_6 : $0.13929 \pm 2.797\ (0.02492) = (0.0696, 0.2090)$

Nuevamente, el único de los intervalos de confianza que contiene al cero como un posible valor, es el intervalo para β_2, coeficiente de la variable AMIN.

Un problema que se presenta al calcular este tipo de intervalos, es el hecho de que el nivel de significancia $\alpha = 0.01$ es válido para cada uno de ellos individualmente, pero no para los tres conjuntamente. Cuando se consideran los tres intervalos conjuntos, el nivel de significancia es mayor y la confianza disminuye.

Para evitar esto se pueden utilizar las llamadas regiones de confianza conjuntas, una de ellas está dada por:

$$\frac{(\hat{\underline{\beta}} - \underline{\beta})^T X^T X (\hat{\underline{\beta}} - \underline{\beta})}{p\ CME} = F(1-\alpha\ ;\ p,\ n-p)$$

sin embargo, esta región es generalmente difícil de obtener y de interpretar, en vez de ella, utilizaré los intervalos conjuntos de Bonferroni, que están dados por:

$$\hat{\beta}_k \pm B\ s(\hat{\beta}_k)$$

donde $B = t(1-\alpha/2m;\ n-p)$, m es el número de intervalos que desean obtenerse, es decir, el número de parámetros para los cuales se quieren obtener los intervalos conjuntos.

Sea $\alpha = 0.01$

Tenemos $n=28$, $p=4$, $m=3$.

Los intervalos cuyo nivel de significancia es 0.01 , conjuntamente, están dados por:

$$\hat{\beta}_k \pm 3.499\ s(\hat{\beta}_k)$$

donde el valor $3.499 \doteq t_{(0.998,\ 24)}$ ha sido obtenido mediante interpolación de los valores:

$t_{(0.995,\ 24)} = 2.797$

$t_{(0.9995,\ 24)} = 3.745,$

de aquí, los intervalos conjuntos del 99% de confianza para β_2, β_4 y β_6 son:

para β_2 : $1.30526 \pm 3.499\ (0.59946) = (-0.792,\ 3.403)$

para β_4 : $0.18201 \pm 3.499\ (0.04958) = (0.009\ ,\ 0.355)$

para β_6 : $0.13929 \pm 3.499\ (0.02492) = (0.052,\ 0.226)$

Como puede observarse, estos tres intervalos son más anchos que los obtenidos anteriormente, pero el nivel de significancia conjunto para los tres es 0.01, y aquí otra vez observamos que el único intervalo que contiene al cero como un posible valor es el correspondiente a β_2, lo que puede indicar que la variable AMIN no es tan significante dentro del modelo, mientras que las variables ASFIL e HISP si lo son.

COMENTARIOS AL ANALISIS

El análisis anterior se desarrolló como un intento de medir la importancia, al menos en cantidad, de la población hispánica en las necesidades de educación en California; de ninguna manera puede decirse que el método y las técnicas utilizados sean los únicos aplicables; de hecho, después de haber realizado el análisis, pienso que quizás una transformación logarítmica de los datos ó el trabajar con proporciones hubieran evitado algunos problemas respecto a las varianzas; sin embargo, este análisis puede considerarse como un primer paso en el estudio de un problema tan complejo como el de la educación en un lugar como California.

Además, después de observar el comportamiento de los residuales, cabe la posibilidad de intentar un modelo de regresión polinomial, en vez de lineal; pero, como todos sabemos, en Estadística siempre hay varios criterios aplicables al mismo problema y siempre hay también, manera de mejorar los anteriores y dar un paso adelante.

CONCLUSIONES

Como resultado del análisis anterior, hemos visto que en efecto, la población estudiantil hispánica tiene gran importancia en California y que, por lo tanto, es de primordial importancia el atender sus necesidades académicas.

Se ha llegado a decir que, de continuar la migración como se ha presentado hasta ahora, California se convertirá en el estado "tercer-mundista" de los Estados Unidos, esto parece ser una broma, pero podría suceder en realidad.

A pesar de los esfuerzos que el estado ha hecho para satisfacer las necesidades de estos individuos, no se han logrado buenos resultados, quizás porque los enfoques y las medidas aplicadas al problema no son aceptados por la ideología de las personas hispánicas.

En las condiciones en las que se encuentra nuestro país, donde la alimentación y los problemas -

internos deben ser el principal foco de atención, sería muy difícil tratar de implementar programas conjuntos - que ayudaran a resolver el problema de la población mexicana en California y sería por otro lado, fomentar la emigración hacia los Estados Unidos; sin embargo, a pesar de que esta emigración representa, en ciertos aspectos, un problema para los Estados Unidos, en otros les ha sido de utilidad, pues ayuda a mantener estables los precios de los productos agrícolas provenientes de California, que es uno de los primeros estados en esta rama.

Quizá también para México ha sido una manera de conservar un cierto equilibrio económico y disminuír la tasa de desempleo, sin embargo, esta no parece ser - la política más sana ni para estas personas, ni para -- los dos paises.

Yo pienso que en la medida en que México re-- suelva sus problemas internos y sea capaz de producir - suficientes fuentes de trabajo, el problema de la emi-- gración disminuirá, mientras que una alternativa de solución al problema de las gentes que ya actualmente se

encuentran fuera del país, sería la implementación de programas conjuntamente con los Estados Unidos, mediante los cuales se les diera educación y enseñanza; así como convenios para el procuramiento de un trato justo y el acceso a empleos bien remunerados y con las prestaciones y garantías a que todo ser humano tiene derecho.

BIBLIOGRAFIA

1.- CALIFORNIA DEPARTMENT OF EDUCATION,
Office of Bilingual-Bicultural Education
Data BICAL, 1980 R-30 LC. 1979 R-30 D/C.

2.- CALIFORNIA STATE DEPARTMENT OF EDUCATION,
Characteristics of Professional Staff in California
Public Schools 1981-82
Wilson Riles, Superintendent of Public Instruction,
Sacramento, 1982.

3.- CALIFORNIA STATE DEPARTMENT OF EDUCATION,
Racial and Ethnic Distribution of Staff ans Students
in California Public Schools 1981-82
Sacramento, 1982.

4.- CALIFORNIA SCHOOL FINANCE REFORM PROJECT AND CALIFORNIA ASSOCIATION FOR BILINGUAL EDUCATION,
Ethnic Groups and Public Education in California, Research Report Number Three
San Diego State University, College of Education.

5.- CENTRO DE ESTUDIOS EDUCATIVOS, A. C.
Revista Latinoamericana de Estudios Educativos
Vol. IX, Segundo trimestre de 1979, Número 2
México, D. F.

6.- HOEL PAUL G., PORT SIDNEY C AND STONE CHARLES J.
Introduction to Probability Theory
University of California, Los Angeles
Houghton Mifflin Company, 1971.

7.- HOGG ROBERT V. AND CRAIG ALLEN T.
Introduction to Mathematical Statistics
The University of Iowa
Macmillan Publishing Company Inc., 1978.

8.- JOHNSON RICHARD A. AND WICHERN DEAN W.
Applied Multivariate Statistical Analysis
University of Wisconsin-Madison
Prentice-Hall Inc., 1982.

9.- LARSEN RICHARD J. AND MARX MORRIS L.
An Introduction to Mathematical Statistics and its Applications.
Vanderbilt University and University of Oklahoma
Prentice-Hall Inc., 1978.

10.-MEYER PAUL L.
Probabilidad y Aplicaciones Estadísticas
Washington State University
Fondo Educativo Interamericano, S. A.

11.- NETER JOHN AND WASSERMAN WILLIAM
Applied Linear Statistical Models
University of Georgia and Syracuse University
Richard D. Irwin Inc., 1974.

12.- OFFICE OF BILINGUAL-BICULTURAL EDUCATION,
Basic Principles for the Education of Language-Minority Students. An Overview
Sacramento, Ca.

13.- REZABEK DALE J.
Horizon, An Overview of Vocational Education and Employment Training Services for Limited English Proficient Persons in California
California Advisory Council on Vocational Education.

14.- VALENCIA GUSTAVO R., MENDOZA MANUEL R. Y ARANDA FRANCISCO
Introducción a la Inferencia Estadística
Departamento de Matemáticas, Facultad de Ciencias
Comunicaciones Internas No. 42, 1978.

15.- WEISBERG SANFORD
Applied Linear Regression
University of Minnesota
John Wiley and Sons, Inc., 1980.