Universidad Nacional Autónoma de México

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES ACATLAN.

Analisis del Costo de Operación para una compañia de Aviación

T E S I S

QUE PARA OBTENER EL TITULO DE:

LICENCIADO EN ACTUARIA

P R E S E N T A:

J. GERARDO MILLAN CAMPOS

Noviembre de 1986

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

INDICE

INTRODUCCION

La Aviación Mexicana tuvo su auge durante el período de 1950-1960. Finalmente en nuestros días, su crecimiento se ve interrumpido, debido a las continuas devaluaciones que está sufriendo nuestro país.

En particular, sobre una de las compañías más importantes del país, la compañía Aeroméxico, a partir de 1959 le ocurren cambios significativos que facilitaron crecimiento. Se pueden citar, entre otros, la nacionalización de la compañía Aeroméxico, S.A., la adhesión de pilotos a "ASPA", así como el cambio de imagen, incluyendo su logotipo.

En nuestros días Aeronaves de México, S.A.- Aeroméxico cuenta en Latinoamérica con la mayor flota de aviones Douglas Tipo DC-9, incluyendo 4 del modelo 80 de la última generación, participando así con el 56% de las 52 aeronaves comerciales de este tipo de operación actualmente. Además, ocupa el 3er. lugar en Latinoamérica en lo que a unidades de flota aérea se refiere, participando con el 12% (38) de las 306 unidades turboreactoras operando comercialmente para las líneas Aéreas regulares de servicio internacional. Sin embargo, de acuerdo a datos fidedignos, para 1984 la Compañía Aeroméxico se encontraba con problemas financieros bastante agudos.

Por lo anterior es de interés realizar un estudio acerca del comportamiento del costo de operación, utilizando variables que expliquen verazmente el fenómeno.

Para llevar a cabo el análisis del comportamiento de la variable respuesta (costo de operación) se aplica el problema uno de los métodos estadísticos más usados: "Análisis de Regresión".

En el análisis de regresión se tiene dos grupos de variables: variables explicativas y variable respuesta. Como su nombre lo indica, las primeras se utilizan para tratar de "explicar" el comportamiento de la segunda. Esta explicación se efectúa mediante un modelo (usualmente lineal) y nuestros objetivos son de descripción y predicción.

La bondad del modelo se juzga de acuerdo, a la precisión con que se logran los objetivos.

Recientemente, se ha dado un auge en la investigación en esta disciplina. Principalmente técnicas de diagnóstico post-ajuste. Tales técnicas buscan proveer de información para mejorar el modelo, o para aprender más acerca de la relación que intenta abstraer. Hoy en día se han desarrollado nuevos métodos cuya importancia es fundamental si se desea llevar a cabo un análisis completo de los datos.

Con el fin de analizar e interpretar los resultados que proporciona el método, se estudiaron sus diferentes técnicas, tanto pasadas como actuales, llegando a un modelo que consideramos óptimo, siendo éste el objetivo principal de este trabajo.

Con el propósito de no caer en meros mecanismos sin conocer sus fundamentos, todas las técnicas utilizadas en este trabajo están acompañadas de su correspondiente demostración, la cual se realiza de la manera más sencilla posible.

La tesis se desarrolla inicialmente considerando el contexto general de la compañía Aeroméxico, S.A., llevando a cabo el planteamiento de un modelo inicial, observando y analizando todas sus ventajas y desventajas.

El trabajo continúa con el análisis más profundo de diferentes modelos, haciendo hincapíe en las ganancias o pérdidas que se adquieren al transformar al modelo original. Al final, se hace una recapitulación y análisis completo de los diferentes modelos que se propusieron en los distintos capítulos.

Es de interés hacer notar que se llevó a cabo un análisis comparativo, tanto al eliminar variables, como al transformar los modelos, con el fin de proponer, de

acuerdo a todas sus características, cual es uno de los mejores.

Finalmente, se proporciona un anexo que contiene las demostraciones, que en los diferentes capítulos no se desarrollan; además se explica en otro anexo, la importancia de aplicar el paquete estadístico "Simulación y Análisis de Problemas de Regresión". (SAREG) en la solución del problema.

CAPITULO I

ANTECEDENTES DE LA COMPAÑIA AERONAVES DE MEXICO, REFERENTE AL COSTO DE OPERACION.

La Compañía Aeroméxico ha sido un pilar en el transporte aéreo mexicano, por lo cual se ha considerado como punto de referencia para el desarrollo de este trabajo.

El desarrollo de esta importante empresa se puede resumir en siete puntos fundamentales, que son:

i) Los primeros vuelos;
ii) La consolidación;
iii) La expansión;
iv) La internacionalización;
v) Patrimonio nacional;
vi) La era del jet;
vii) La nueva era.

i. Los Primeros Vuelos

La Compañía Aeronaves de México, S.A. (Aeroméxico), con cuatro aeronaves, inicia sus actividades en forma oficial el día 14 de septiembre de 1934, con el vuelo México-Acapulco-México, utilizando un avión Stinson de 5 plazas, con motor Lyncoming de 215 Hp., después de haberse realizado varios vuelos de prueba.

El fundador de esta empresa aérea fue el Sr. Antonio Díaz Lombardo, quedando posteriormente, como presidente de la misma. Pero no fue únicamente por el deseo de un hombre que se formó, ni tampoco fue de la noche a la mañana. La participación activa de los pilotos: Leonardo Enriquez, Julio Zinser y de los hermanos González, fue fundamental para llevar a cabo tan ambicioso proyecto.

Es de interés mencionar que en aquella época, los vuelos comerciales se realizaban en condiciones muy precarias, ya que, en muchas ocasiones, los pilotos no contaban con los recursos monetarios para adquirir el combustible necesario para llevar a cabo operaciones aéreas; otra de las incompatibilidades que se tenían en el sistema era que se utilizaba una pista inadecuada para realizar aterrizajes, conocida como: "Area Despejada", la cual se encontraba junto a la Playa de Hornos.

Con la implementación de este transporte, en la ruta México Acapulco, el puerto de Acapulco pronto llegaría a ser uno de los puertos turísticos más cotizados.

En el año de 1940, son adquiridos por Pan American, el 40% de las acciones de "Aeronaves"; se compra la Aerolínea Transportes Aéreos del Pacífico y se comienzan con los servicios entre La Paz y Mazatlán, donde se ha-

cía conexión con Mexicana de Aviación.

En 1942 se incorpora a Aeronaves de México la Compañía Aeronaves de Michoacán, que volaba de Acapulco a Uruapan, con varias escalas (Tecpan, Petatlán, Zihuatanejo, La Unión, Melchor Ocampo, Playa Azul, Arteaga y Apatzingán).

ii) La **Consolidación**

En 1943 se compra la Compañía Taxi Aéreo de Oaxaca, se adquieren tres aviones C-39, los cuales posteriormente son denominados DC-1/2, y que operaban en la ruta México-Acapulco.

Un año después se compra la Compañía Líneas Aéreas Jesús Sarabia, prestando servicios en la ruta México-Puebla-Córdoba-Veracruz-Coatzacoalcos. Esta ruta no era redituable, por lo cual pronto fue cancelada.

A sólo una década de su fundación, Aeronaves de México, S.A. contaba con una considerable flota, una amplia red de rutas y una eficiente organización, que hicieron posible su consolidación.

En 1946, Aeronaves pone en servicio aviones tipo DC-3, los cuales llegarían a volar en todas las rutas existentes.

Con las recientes adquisiciones, la empresa contaba con los elementos necesarios para brindar un magnífico servicio a la comunidad en general. Sus operaciones se realizaban con mucha regularidad en todas las rutas, proporcionando altos índices de ocupación y obteniéndose resultados económicos positivos.

En 1948, Aeronaves pone en servicio el avión C-47, dedicado exclusivamente al servicio de carga. Esta aeronave proporcionaba una rápida comunicación con el Estado de Oaxaca, y, posteriormente se asigna al resto del sistema.

En 1949, se compra un DC-4, con capacidad para 54 pasajeros, el cual realiza tres vuelos México-Acapulco diariamente.

Es en el año 1950, cuando se establece la ruta México-Tepic-Mazatlán, que se transformaría en México-Guadalajara-Mazatlán, que enlazó varias ciudades como Culiacán, Los Mochis, Cd. Obregón, Nogales y Guaymas, entre otras.

iii) **La Expansión**

Para el año de 1952, se incorpora a Aeronaves, la Compañía Líneas Aéreas Mexicanas, S.A. (LAMSA), la cual cubría rutas del Noroeste.

En el año de 1954, se integra al sistema Aeronaves, la Compañía Aerovías Reforma. Con esta incorporación, Aeronaves vuela a Tijuana, vía Gudalajara, Culiacán y Guaymas, de México a Manzanillo, con escalas en Morelia, Coacolman y Colima, y de Guadalajara a Manzanillo.

En ese mismo año se adquieren cuatro aviones Convair 340 con cabina presurizada y con capacidad para 44 pasajeros. Estas naves cubrieron tanto la ruta de Acapulco como la de Tijuana siendo las operaciones de "oro" de la Compañía.

iv) **La Internacionalización.**

Con el crecimiento de la Compañía Aeronaves de México, S.A., se fueron presentando una serie de repercuciones, entre otras, la cancelación de algunas rutas debido a que unas eran incosteables y otras porque ya se contaba con medios de transportación terrestre. La Compañía al enfrentarse a estos problemas decide transpasar las fronteras realizando vuelos internacionales.

Para el año de 1957, Aeronaves de México, S.A., (A.M.) adquiere dos aviones marca Lockheed L-49, con capacidad para 58 pasajeros, los cuales prestarían servicio en las rutas de México-Acapulco y México-Tijuana.

El 40% de las acciones, que pertenecían a Pan American, fueron adquiridas por inversionistas mexicanos, lo

cual propició que el capital social fuera totalmente de ciudadanos mexicanos.

A fines de 1957 se lleva a cabo la reestructuración de las rutas internacionales y se le otorga a A.M. la ruta México-Nueva York sin escalas, que es inaugurada el 8 de diciembre, con un cuatrimotor Bristol Britania 302, con capacidad para 92 pasajeros.

En 1958 se compran 3 naves marca Constellation L-749, con capacidad para 68 pasajeros, los cuales sustituyeron a los L-049.

En el siguiente año, se produce un paro nacional de pilotos, por lo que el gobierno de la República interviene a A.M., nombrando administrador al Ing. Jorge Pérez y Bouras.

v) **Patrimonio Nacional**

Un suceso importante que cambiaría la organización de A.M. ocurre el 28 de julio de 1959, cuando por órdenes del entonces Presidente Adolfo López Mateos, se adquieren las acciones que estaban en poder de los particulares, pasando así a ser patrimonio nacional. El Sr. Pérez y Bouras es nombrado Director General, siendo antes Administrador de la misma. Así mismo, A.M. es la primera arolínea que firma un contrato colectivo con la Aso

ciación Sindical de Pilotos Aviadores (ASPA).

Las cifras arrojadas al final del año 1959, son las siguientes:

2,045 empleados;
22 aviones;
416,419 pasajeros transportados;
23 ciudades servidas en la República y en el extranjero;
13,365 kilómetros recorridos.

En A.M. se contaba con un equipo bastante diferente, por lo que se decidió sustituir los Constellatión, Convair y DC-4 por aviones Douglas DC-6 cuatrimotores de cabina presurizada, equipados con radar y capacidad para 79 pasajeros.

Al pasar A.M. a ser propiedad pública se modifica el emblema de un águila en vuelo por el de un "caballero águila", el cual representa nuestras raíces históricas, y como consecuencia, las naves fueron bautizadas como "guerreros prehispánicos".

vi) La Era del Jet

Para 1960, A.M. pone en servicio un cuatrimotor Douglas DC-8 en la ruta México-Nueva York, con capacidad

para 138 pasajeros y con velocidad de crucero superior a los 800 kilómetros por hora.

En el año de 1960, la Compañía Aerolíneas Mexicanas, S.A., se ve afectada por un paro de pilotos, por lo que el gobierno decide pasar a A.M. la responsabilidad de dicha empresa.

En 1961, se abren dos nuevas rutas internacionales: una a Tucson y la otra a San Antonio.

En 1962, se le incorporan también otras aerolíneas, las cuales son: Trans Mar de Cortés y Guest Aerovías México. Con estas anexiones, A.M. establece una ruta entre La Paz-Loreto-Santa Rosalía-Tijuana y La Paz-Tijuana-Los Angeles. Posteriormente, se hace cargo de varias rutas internacionales.

En 1964, es inaugurada la ruta internacional México-Toronto-Montreal, con el equipo Douglas DC-8. También se inauguraron las rutas siguientes: México-Detroit, Acapulco-México-Nueva York y Acapulco-La Paz-Los Angeles.

A los treinta años de su fundación, A.M. arroja las siguientes cifras:*

45,000 km. de rutas;
40 ciudades con servicio aéreo;
2,980 empleados;

---

* Folleto "50 Aniversario" Aeroméxico 1934-1984.

673,329 pasajeros transportados.

En 1965, se establece otra ruta: México-Torreón-Monterrey-Chihuahua-Hermosillo-Tijuana.

En 1966 se extiende la ruta Miami-Madrid hasta Roma, y en 1967 entran en servicio los primeros DC-9 para rutas cortas. Con su incorporación al sistema se propicia la baja de los aviones de pistón DC-3 y DC-6.

La flota de A.M. queda integrada únicamente por aviones jet, a partir del día 4 de marzo de 1968.

Para 1969 las Compañías Aeromaya, Vega y del Istmo se anexan a A.M. con las cuales se forma el sistema de Aeronaves Alimentadoras.

Una nueva etapa de esta empresa se inicia en el año de 1971, período en el cual se adopta el nombre comercial de Aeroméxico, debido a la simplicidad y fácil pronunciación del vocablo.

Para 1974, se compran aviones tipo jumbo DC-10-30 McDonnell Douglas, que estaban equipados con poderosos reactores y con capacidad de 301 pasajeros. Estas naves pueden realizar vuelos transatlánticos.

Para 1977, el Lic. Pedro Vázquez Colmenares es designado Director General de la empresa, quien pone en

marcha un plan de rentabilidad a base de incrementar la productividad, alentar la responsabilidad del personal y reestructurar rutas, frecuencias y servicios, lo que propició un auge financiero para la aerolínea.

El Lic. Enrique M. Loaeza Tovar es designado para 1980, Director General de Aeroméxico, y en su administración se incorporan nuevos aviones DC-10-15 y DC-9 Super 80.

vii) **La Nueva Era**

A fines de 1982, es designado como Director General, el Sr. Sigfrido Paz Paredes, quien adopta medidas a fin de reducir costos y aprovechar los efectos positivos de la devaluación, que han hecho a nuestro país sumamente atractivo en el extranjero.

Una de las principales acciones que se realizaron en este período, fue la de ubicar a la flota y adaptar itinerarios y frecuencias a las necesidades cambiantes de la demanda, sobre todo, en la zona fronteriza.

Durante el año de 1983, se establecieron nuevos servicios nacionales e internacionales.

Al cumplirse en 1984 el cincuentenario de su fundación, Aeroméxico cuenta con los recursos huma

nos y técnicos para hacer frente a ciertas necesidades del país y se suma a las demás aerolíneas para hacer de México un país mejor.

Es así como Aeroméxico es una de las aerolíneas - más fuertes de la nación.

Por todo lo anterior, es conveniente estudiar el comportamiento de la variable costo de operación, que llamaremos respuesta, para efectos de este trabajo.

Una forma adecuada es suponer que la citada respuesta puede ser aproximada por una relación funcional en la cual se consideren todas aquellas variables que pudiesen afectar a la respuesta. Idealmente, los valores de la - respuesta (Y) podrían obtenerse a partir de la relación:

$$Y = f(X_1, X_2, \ldots, X_n)$$

donde f es una función arbitraria y X_q , $q = 1, \ldots, n$ denotan a las variables que se supone afectan a la respuesta.

Al establecer tal expresión, son evidentes dos - posibles dificultades:

1) La forma analítica de δ puede ser desconocida o, aún siendo conocida, muy complicada;

2) El número n de variables requeridas, puede ser tan grande que sea prácticamente imposible manejar δ , - aún en el caso de conocerla.

Como posibles alternativas, pueden considerarse - las siguientes:

a) Aproximar δ mediante una función sencilla, h, posiblemente un polinomio.

b) Cancelar todas aquellas variables cuya influencia sobre la respuesta se considere de poca importancia.

La segunda alternativa tendrá como resultado que las variables no unidas causen fluctuaciones en la respuesta, las cuales pueden considerarse, para efectos de análisis, como aleatorias.

Entonces quedamos con la relación:

$$Y = h(X_1, X_2, \ldots X_k) + \varepsilon$$

donde ε denota una variable cuyos valores estarán determinados por todas aquellas variables eliminadas en (b), y por errores que pudiesen cometerse en la medición de las variables explicativas incluidas.

Si se supone que la función es contínua, al menos en el rango de valores de las variables $X_1, X_2 \ldots X_n$ que interesa, una conveniente forma para h es la de un polinomio

$$y = \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p + \varepsilon$$

donde Z_j, $j=1,\ldots,p$ es alguna función real conocida de las variables $X_1, X_2 \ldots X_k, \beta_j, j=1,\ldots\ldots,p$ son parámetros desconocidos y ε es una variable aleatoria.

En el análisis de regresión se usan este tipo de modelos imponiendo ciertas condiciones, ya sea sobre los datos, o sobre la estructura de la componente aleatoria.

Para llevar a cabo el análisis del costo de operación de la Cía. Aeroméxico, S.A., se consideran nueve variables que tratan de explicar el comportamiento de la variable respuesta (costo de operación). La selección de ta

les variables fue realizada con la ayuda de un experto en el área, y una variable aleatoria (ε) que engloba todos los errores de medición y todas aquellas variables que se consideraron de poca importancia.

CAPITULO II

APLICACION DEL MODELO MULTIPLE

A: ASPECTOS GENERALES

i. Modelo Múltiple

En un modelo de regresión múltiple actuan P variables explicativas y una variable respuesta, donde las variables independientes o explicativas son X_1, X_2, X_3,... X_p y la variable respuesta o dependiente es Y.

El modelo puede ser especificado en una ecuación lineal y escrito de la siguiente manera.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots \beta_p X_p + \varepsilon \quad (2.1)$$

Donde los β_j, $j = 0, 1, 2, \ldots p$ son parámetros desconocidos y los ε's son los errores estadísticos.

Si $p=3$ entonces la ecuación (2.1) representaria una superficie tridimensional en un espacio cuadri-dimensional (X_1, X_2, X_3, Y).

La notación matricial es una herramienta matemática que simplifica enormemente todos los desarrollos algebráicos, requeridos para la estimación de los parámetros desconocidos.

Sea el modelo considerado para las n-observaciones

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots \beta_p x_{ip} + \varepsilon_i \quad i=1,\ldots n \quad (2.2)$$

las variables del modelo (2.2) se pueden expresar como sigue:

Caso No.	Y	1	X_1	X_2	X_3	. . .	X_p
1	y_1	1	x_{11}	x_{12}	x_{13}	. . .	x_{1p}
2	y_2	1	x_{21}	x_{22}	x_{23}	. . .	x_{2p}
.	.	.	.	.	.	. . .	.
.	.	.	.	.	.	. . .	.
.	.	.	.	.	.	. . .	.
n	y_n	1	x_{n1}	x_{n2}	x_{n3}		x_{np}

Y y ε en su forma matricial son:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix}_{nx1} \quad ; \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}_{nx1}$$

También podemos definir al vector de parámetros β y a la matriz de datos u observaciones

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix}_{(p+1)x1} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}_{nx(p+1)}$$

Al haber aplicado este tipo de notación, el modelo queda reducido y se muestra a continuación en términos de matrices

$$Y = X\beta + \varepsilon \qquad [2.3]$$

Este modelo supone que los errores tienen esperanza 0 y varianza σ^2, es decir $E(\varepsilon) = 0$ y $V(\varepsilon) = \sigma^2 I_n$, donde I_n es la matriz identidad de nxn, 0 es un vector de ceros y $V(\varepsilon)$ es la matriz de varianzas y covarianzas. Si además se supone que los errores están distribuidos normalmente, podríamos escribir que

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

El método de mínimos cuadrados, trata de minimizar la suma de errores al cuadrado, con el propósito de encontrar los estimadores del vector β donde $\hat{e}$ es el vector de residuales y el "gorrito $\hat{}$ "

significa aproximación puntual al parámetro.

Para poder especificar $\hat{\varepsilon}$ es necesario encontrar el vector $\hat{\beta}$ para lo cual tratamos de minimizar al error.

Tomando (2.3)

$$\varepsilon = y - X\beta$$

entonces

$$\varepsilon^1\varepsilon = (y - X\beta)^1 (y - X\beta)$$

$$= y^1y - y^1X\beta - (X\beta)^1 y + (X\beta)^1 (X\beta)$$

$$\varepsilon^1\varepsilon = y^1y - y^1X\beta - \beta^1X^1y + \beta^1X^1X\beta$$

$$\frac{\partial \varepsilon^1\varepsilon}{\partial \beta} = -2X^1y + 2X^1X\beta$$

Buscando el mínimo igualamos a cero $\frac{\partial \varepsilon^1\varepsilon}{\partial \beta} = 0$

$$\rightarrow -2X^1y + 2X^1X\hat{\beta} = 0$$

$$\rightarrow X^1X\hat{\beta} = X^1y \quad \text{que son las ecuaciones normales}$$

$$\rightarrow \hat{\beta} = (X^1X)^{-1} X^1y \leftrightarrow (X^1X)^{-1} \text{ existe, si } X^1X$$

es cuadrada y cuando su determinante es no cero o no singular. El $\hat{\beta}$ es el estimador por mínimos cuadrados del vector β y el valor ajustado de Y es

$$\hat{y} = X\hat{\beta} \qquad (2.4)$$

ii). **Propiedades de los estimadores.**

Posteriormente es importante observar las propiedades de los estimadores*.

Bajo la hipótesis de que $E(\varepsilon)=0$ y $Var(\varepsilon)=\sigma^2 I_n$ entonces $\hat{\beta}$ es un estimador insesgado, es decir $E(\hat{\beta})=\beta$. Además se tiene que la varianza de $\hat{\beta}$ es:

$$V(\hat{\beta}) = \sigma^2 (X^1X)^{-1} \quad * \qquad (2.5)$$

También es factible estimar σ^2* de la siguiente manera

$$\hat{\sigma}^2 = \frac{SCR}{n-p^1} \qquad (2.6)$$

donde $SCR = \hat{\varepsilon}^1\hat{\varepsilon} = (Y-\hat{Y})^1 (Y-\hat{Y})$ y $p' = (p+1)$

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

$$= Y^1Y - YX\hat{\beta} - \hat{\beta}^1X^1Y + \hat{\beta}^1X^1X\hat{\beta}$$

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

$$= 0$$

$$= Y^1Y - Y^1X\hat{\beta}$$

Si además asumimos que los ε se distribuyen normalmente, entonces $\frac{\hat{\sigma}^2}{\sigma^2} \sim X^2_{(n-p^1)}$* (2.7)

---

* Demostración en el Apendice I

Con la expresión (2.5) es fácil encontrar la varianza estimada de $\hat{\beta}$ que es:

$$\hat{Var}(\hat{\beta}) = \hat{\sigma}^2 (X^1X)^{-1} \qquad (2.8)$$

Por otro lado el error estándar de $\hat{\beta}_j$ $j=1,\ldots p$ es la raiz cuadrada del j-ésimo elemento de la diagonal de la matriz $\hat{\sigma}^2(X^1X)^{-1}$

La covarianza es de suma importancia para encontrar la dependencia entre variables o estimadores y la forma de calcularla es la siguiente:

$$Cov(\hat{\beta}_j, \hat{\beta}_l) = \sigma^2 (X^1X)^{-1} \qquad (2.10)$$

Es importante hacer notar que si $i=j$ entonces estamos en el caso de la $\hat{Var}(\hat{\beta}_i)$, para encontrar la covarianza es necesario calcular la matriz inversa $(X^1X)^{-1}$ y buscar la intersección del renglón i contra la columna j y ese valor multiplicarlo por σ^2.

Valores Ajustados y Predicciones

Es factible encontrar el error estándar del valor ajustado de cualquier renglón de la matriz , o sea $\hat{Y} = x' \hat{\beta}$ y la notación adecuada es: esa $(\hat{Y}|x)=\hat{\sigma}(x^1(X^1X)^{-1}x)^{1/2}$

(2.11)

---

* Demostración Apendice I

El error estándar de predicción es de suma importancia, ya que mide que tan mal predecimos, o sea una predicción en x es dada por $\tilde{y} = x^1\hat{\beta} + \hat{\varepsilon}$ donde $\hat{\varepsilon}$ es un error aleatorio que contribuye al error estándar de predicción (esp) y es:

$$esp(\tilde{y}/x) = \hat{\sigma}\{1 + x^1(X^1X)^{-1}x\}^{1/2} \quad * \qquad (2.12)$$

iii) El análisis de varianza es una de las técnicas más relevantes que se utiliza para descomponer la variabilidad y para comparar modelos que incluyan diferentes conjuntos de variables.

Si tenemos el modelo $Y = X\beta + \varepsilon$ y lo queremos comprar contra el modelo $Y = 1\beta_0 + \varepsilon$, donde 1 es un vector de unos, $\hat{\beta}_0 = \bar{Y}$, su estimador y $SYY = Y'Y$ su suma de cuadrados de residuales. Para el modelo completo $\hat{\beta}$ es el estimador y su respectiva suma de cuadrados de residuales es SCR. Claramente $SCR < SYY$, la diferencia $SYY-SCR = SCreg$ corresponde a la suma de cuadrados en Y explicada por el modelo completo que no es explicada por el modelo reducido.

Todos los resultados son resumidos generalmente en una tabla de análisis de varianza (ANOVA), y se mues-

---

* Demostracción en Apendice No. 1

tra a continuación: (tabla 2.1)

FUENTE	Grados de Libertad(GL)	Suma de Cuadrados(SC)	Cuadrados Medios(CM)	F
Regresión	p	$SCreg$	$SCreg/p$	$\frac{SCreg/p}{SCR/n-p'}$
Residual	$n-p'$	SCR	$SCR/_{n-p'}$	
Total	$n-1$	SYY		

Si suponèmos que los $\varepsilon \sim N(0,\sigma^2 In)$ entonces la hipotesis a probar para la prueba F es

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \ldots = \beta_p = 0$$

$$H_\Delta: \text{Algún } \beta_j \neq 0 \quad j = 1,2 \ldots p$$

El coeficiente de determinación da la proporción de variabilidad en Y explicada por la regresión en X es:

$$R^2 = \frac{SYY - SC_R}{SYY} = \frac{SCreg}{SYY} \qquad (2.13)$$

Se puede demostrar* que el coeficiente de determinación es igual al coeficiente de correlación múltiple al cuadrado, es decir, es el cuadrado de la correlación entre Y y $\hat{y}$

Es necesario analizar los problemas que acarrea el considerar un modelo incorrecto por lo cual se requiere examinar tanto la carencia de ajuste como el error

---

* Demostración en el apendice I.

puro.

La suma de cuadrados de residuales representa a tanto la variación debida al error como la variación debida al modelo.

Notación

$Y_{K1}, Y_{K2}, \cdots Y_{KL_K}$ son L_K observaciones repetidas en X_K

$$n = \sum_{j=1}^{K} \sum_{\ell=1}^{Lj} 1 = \sum_{j=1}^{K} Lj \quad \text{No. de observaciones}$$

El error puro (EP) puede ser expresado de la siguiente manera:

$$EP = \sum_{j=\ell}^{K} \sum_{\ell=1}^{Lj} (Y_{j\ell} - \bar{Y}j)^2$$

y los grados de libertad son:

$$GLEP = \sum_{j=1}^{K} (Lj - 1) = n-K$$

El modelo es correcto si no existe un sesgo en la estimación de σ^2, por lo que es factible llamar al resumen de arriba tabla de varianza, pasando a calcular el estadístico carencia de ajuste, para el cual se requiere de la siguiente expresión:

$$\sum_{j=1}^{K} \sum_{\ell=1}^{Lj} (Y_{j\ell} - Yj)^2 = \sum_{j=1}^{K} \sum_{\ell=1}^{Lj} (Y_{j\ell} - \bar{Y}j)^2 + \sum_{j=1}^{K} Lj\,(\bar{Y}j - \bar{\bar{Y}}j)^2$$

bajo el supuesto de que existan repeticiones

donde el término del lado izquierdo es la suma de cuadrados de residuo; el primer término del lado derecho es la suma de cuadrados de errores puros y al segundo término del lado derecho le llamaremos suma de cuadrados de carencia de ajuste se calcula directamente puesto que son la diferencia entre los grados de libertad de SCR y los grados de libertad de EP. $(n-p'-(n-K)=K-p')$.

Los resultados obtenidos pueden ser resumidos de la siguiente manera: (tabla 2.2)

FUENTE	G.L.	S.C.	C.M.	F.
Residual	$n-p'$	SCR		
Carencia de A	$K-p'$	CA	$CA/_{K-p'}$	$\dfrac{CA/_{K-p'}}{EP/_{n-K}}$
Error Puro	$n-K$	EP	$EP/_{n-K}$	

Es importante hacer notar que la prueba asociada a las expresiones de arriba es H_0: No hay carencia de ajuste VS H_1: Hay carencia de ajuste, el estadístico de prueba es F y se compara contra la distribución $F_{(K-p',n-K)}^{(1-\alpha)}$ a fin de rechazar o no H_0.

iv). **Analisis de Residuales**

Los residuales nos dan información sobre el com-

portamiento del Modelo.

Sea. $Y = X\beta + \varepsilon$

bajo

$E(\varepsilon) = 0 \;\; ; \;\; Var(\varepsilon) = \sigma^2 \, In \quad Cov(\varepsilon_i, \varepsilon_j) = 0$

$\forall \, i,j \;\; i \neq j$

Es decir los error son homocedásticos y no correlacionados.

Sea

$$\hat{\varepsilon} = Y - \hat{Y}$$

Sustituyendo la expresión 2.4 obtenemos que

$$\hat{\varepsilon} = Y - X\hat{\beta}$$

y si además sustituimos $\hat{\beta} = (X^1X)^{-1} X^1\underline{Y}$ se tiene que $\hat{\varepsilon} = Y - X(X^1X)^{-1} X^1\underline{Y} = (In - X(X^1X)^{-1} X^1)\underline{Y}$ sea $H = X(X^1X)^{-1} X^1$ entonces (Apendice I)

$\hat{\varepsilon} = (In-H)\underline{Y}$ ..... (2.14)

Los elementos de H son dados por la siguiente ecuación.

$$h_{ij} = x_i^1 \, (X^1X)^{-1} \, x_j$$

y para los elementos de la diagonal

$$h_{ii} = x_i^1 \, (X^1X)^{-1} \, x_i$$

donde x_i^1 y x_j^1 son, respectivamente el i-ésimo y el j-ésimo renglón de la matriz de datos X.

Pasamos a demostrar que $E(\hat{\varepsilon}) = 0$ y que $V(\hat{\varepsilon})=\sigma^2(I-H)$

$E(\hat{\varepsilon}) = E\{(In-H)Y\} = (In-H)E(\underline{Y}) = (In-H)X_\beta$

$E(\hat{\varepsilon}) = 0$

$Var(\hat{\varepsilon}) = Var\{(In-H)\underline{Y}\} = (In-H)\ Var(\underline{Y})$. Por ser $(In-H)$ Idempotente y simétrica: (demostración en apendice I).

$$\rightarrow \quad Var(\hat{\varepsilon}) = (In-H)\sigma^2 \qquad (2.16)$$

Vemos directamente que la varianza del i-ésimo residual es:

$$Var(\hat{\varepsilon}_i) = \sigma^2(1-h_{ii}) \leftrightarrow 0 \leq h_{ii} \leq 1. \qquad (2.17)$$

La $Var(\hat{\varepsilon}_i)$ es pequeña si h_{ii} es grande, la covarianza entre el i-ésimo y j-ésimo residual es:

$$Cov(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = -\sigma^2 h_{ij} \quad i \neq j \qquad (2.18)$$

Además la correlación entre el i-ésimo y j-ésimo residual es:

$$Corr(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = -\frac{h_{ij}}{(1-h_{ii})^{1/2}\ (1-h_{jj})^{1/2}} \qquad (2.19)$$

Por lo cual la correlación no depende de la variable respuesta.

Es importante hacer notar que el rango de H es igual al rango de X igual a P^1, verificandose que es un modelo de rango completo. Por otro lado tenemos

que la suma de los elementos de la diagonal, llamada traza es:

$$\sum_{i=1}^{n} h_{ii} = \text{rango}(X) = p'$$

y que $\sum_j h_{ij}^2 = h_{ii}$ el cual se verifica por la definición de idempotencia. En regresión multiple, h_{ii} mide la distancia del punto x_i al centro de los datos ($\bar{X}$ vector).

La única desventaja de los h_{ii} es que dependen exclusivamente de los valores de las variables explicativas y no involucran a la variable respuesta correspondiente.

Para mejorar la definición de residuales tenemos lo que se llama residuales estudentizados y se define como:

$$\hat{e}_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\ (1-h_{ii})^{1/2}} \qquad i = 1 \ldots n \qquad (2.20)$$

a estos también se les llama residuales estandarizados.

La expresión (2.20) tiene las siguientes características:

$$\sum_{i=1}^{n} \hat{e}_i \neq 0 \quad ; \quad E(\hat{e}_i) = 0$$

Los $\hat{e}_i$ estan ligeramente correlacionados con los $\hat{y}_i$ y la ventaja es que $Var(\hat{e}_i) = 1 \ \forall i$ si el modelo es correcto.

La distribución de $\hat{e}_i$ es una transformación monótona de una distribución *t-student*. Esto último será de gran utilidad para probar observaciones discrepantes. Una aproximación posible es tratar $\hat{e}_i$ como si fueran variables normales estándar (0,1).

v) **La Función Influencia**

Un problema importante es como evaluar el impacto del valor observado para el *i-ésimo* caso sobre las estimaciones hechas.

Esto puede enfocarse de la siguiente manera: 1º retirar la *i-ésima* observación de los datos, quedando *n-1* casos para el análisis.

Segundo, examinar la diferencia entre los estimadores de los parámetros, incluyendo el *i-ésimo* caso y sin incluirlo.

Al parecer el procedimiento es sumamente tedioso, pero solo se requieren los valores de $\hat{e}_i$, h_{ii}, $\hat{\sigma}^2$, n y p'.

La Función Empírica de Influencia.

Una versión empírica de la función de influencia muy util para nuestos fines se obtiene al tomar la diferencia

$$\hat{\beta} - i - \hat{\beta}$$

donde $\hat{\beta}-i$ es el estimador de β excluyendo la i-ésima observación ($i=1,\ldots n$). Una observación (o caso) será llamado influyente si al retirarla resulta un cambio substancial en la estimación del parámetro.

Cook (1) propone que la influencia del i-ésimo caso se mida mediante la siguiente expresión

$$\mathcal{D}_i = \frac{(\hat{\beta}-i - \hat{\beta})^1 \ (X^1X) \ (\hat{\beta}-i - \hat{\beta})}{(p+1)\hat{\sigma}^2} \qquad (2.21)$$

Entonces los casos con un valor grande $\mathcal{D}_i$ tendrán gran influencia sobre la estimación de β y retirarlos puede propiciar importantes alteraciones en las conclusiones.

$\mathcal{D}_i$ se compara con el valor $F(\alpha, p+1, n-p-1)$ para una cierta α, juzgando así su influencia.

Es importante hacer notar que $\mathcal{D}_i$ no se distribuye como F, ésta comparación permite tan sólo una escala familiar.

La expresión (2.21) puede reescribirse usando cantidades conocidas; la mejor forma computacional es:

---

1 Revista Techonometrics.

$$Di = \frac{1}{p+1} \hat{e}_i^2 \left[\frac{h_{ii}}{1-h_{ii}}\right]$$

El segundo factor es el cociente de la varianza del i-ésimo valor ajustado entre la varianza del correspondiente residual. Ocurre que Di va a tomar valores grandes, si el primero o segundo factor es grande. Estos factores miden dos características separadas de cada observación o caso.

1. Valor grande de $\hat{e}_i$, lo que refleja la carencia de ajuste del modelo en el i-ésimo caso.
2. Gran distancia del promedio de los otros vectores de variables explicativas para x_i evidenciado por la h_{ii}.

En sí, la distancia Di proporciona una metodología específica para examinar el cambio en la estimación de β cuando un caso individual es retirado, en relación a ciertos elipsoides de confianza.

La interpretación es como sigue: Si una Di es exactamente igual a $F(\alpha, p+1, n-p-1)$ esto significa que eliminando el caso i propiciaría mover el estimador de β al límite de una región de $(1-\alpha)$x 100% de confianza basada en los datos completos.

La teoría del análisis de regresión se lleva actualmente a la práctica. Sus técnicas son aplicadas a muchos campos de estudio, incluyendo ciencias sociales ciencias biológicas y físicas, en negocios y tecnología.

Las técnicas expuestas en este capítulo serán aplicadas a continuación a un problema práctico, con el fin de analizar el costo de operación de la compañía Aeroméxico, S.A.

Planteamiento del Modelo Inicial

Con el objeto de investigar el costo de operación de la compañía Aeronaves de México se recopiló la siguiente información.

y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
207.51	77.25	13	63	745	0	0	0	0	8.02
217.38	77.50	13	72	745	1	0	0	0	8.12
270.71	77.83	7	80	886	0	0	0	1	11.35
272.37	78.17	12	50	822	0	0	0	0	11.45
280.36	78.83	12	71	886	1	0	0	1	11.45
284.88	79.20	12	63	886	0	0	0	1	11.78
288.48	79.63	9	48	821	0	0	1	0	12.41
289.66	79.95	15	76	530	1	0	1	0	21.42
317.21	80.23	14	59	457	0	0	0	0	12.81
345.39	80.54	13	51	514	0	1	1	0	12.73
350.63	80.58	12	64	560	0	0	0	0	13.30
394.36	80.92	13	65	850	0	0	0	1	13.16
402.59	81.00	13	47	790	0	1	0	0	13.60
412.18	81.21	15	62	530	0	0	0	0	14.20
423.32	81.34	11	67	778	0	0	0	0	15.3
443.22	81.50	10	85	1065	0	0	0	0	15.34
452.05	81.60	10	73	1065	0	0	0	0	15.42
457.12	81.67	15	55	822	1	0	0	0	17.5
460.05	81.72	14	46	687	0	1	0	0	16.14
473.64	81.79	19	44	538	0	0	0	0	18.19
490.88	81.97	16	59	1050	0	0	0	0	18.28
495.58	82.03	17	52	1050	0	0	0	0	19.7
567.79	82.32	11	70	913	0	0	1	1	19.15
608.80	82.40	19	58	821	1	0	0	0	20.18
621.45	82.47	16	59	786	0	0	0	0	20.30
642.23	82.32	11	78	1065	1	1	0	0	21.12

652.32	82.43	11	67	1065	0	1	0	0	21.42
665.99	82.63	22	57	828	1	1	0	0	21.20
690.19	83.00	12	71	792	1	1	0	1	21.30
697.14	83.25	20	57	1130	1	0	0	0	21.40
712.27	83.53	18	66	845	1	1	0	0	21.70
881.24	83.70	15	67	1090	1	0	0	0	21.90

Se tienen 32 observaciones, nueve variables independientes y una variable respuesta. El Modelo inicial es:

Modelo 1.1

$$Y = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8+\beta_9X_9+\varepsilon$$

bajo los supuestos $E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2 I_n$

$$Cov(\varepsilon_i,\varepsilon_j) = 0 \quad \forall i, \; j \quad i \neq j$$

Donde:

Y = Costo de Operación. (Por 10^4pesos)

X_1= Fecha de evaluación (años)

X_2= Seguro y renta de aviones (Por 10^2x 2 en dolares)

X_3= Sobrecargos y Pilotos (sueldos en diezmiles de pesos)

X_4= Mantenimiento (materiales) (en dólares)

X_5= Apertura de Nuevas Rutas (dicotómica si=1,no=0)

X_6= Gas subsidiado (dicotómica si=1, no=0)

X_7= Compras de nuevas naves (dicotómica si=1, no=0)

X_8= Renta de al menos un avión (dicotómica si=1,no=0)

X_9= Salarios variables (por 10^3 pesos)

Al suponer esperanza cero y varianza sigma cuadrada se puede estimar el vector β, es decir si los errores son homocedásticos y no correlacionados. Al aplicar la técnica de mínimos cuadrados se obtuvieron los siguientes resultados.

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}^3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \\ \hat{\beta}_6 \\ \hat{\beta}_7 \\ \hat{\beta}_8 \\ \hat{\beta}_9 \end{bmatrix} = \begin{bmatrix} -2555.7278 \\ 31.7413 \\ -1.0047 \\ 0.4848 \\ 0.0574 \\ 39.8310 \\ 15.2211 \\ -25.1039 \\ 0.2591 \\ 22.8311 \end{bmatrix}$$

$$\rightarrow \hat{Y} = X\hat{\beta}$$

Es importante realizar los diagramas de dispersión, puesto que nos ayudan a darnos una idea de las tendencias y de alguna manera observar la correlación que se da entre las variables. Por lo cual pasamos a graficar al vector $\hat{Y}$ contra las columnas X_1, X_2, X_3, X_4 y X_9, las variables X_5, X_6, X_7 y X_8 no son graficadas puesto que son categóricas o binarias y no tendrá mucho senti-

do por el momento analizar sus diagramas de dispersión. (gráficas 1, 2, 3, 4 y 5).

Posteriormente pasamos a calcular las varianzas estimadas de los betas "gorros".

$$\hat{Var}\ (\hat{\beta}_0) = 141004.7986$$
$$\hat{Var}\ (\hat{\beta}_1) = 262.2963$$
$$\hat{Var}\ (\hat{\beta}_2) = 29.8305$$
$$\hat{Var}\ (\hat{\beta}_3) = 1.6579$$
$$\hat{Var}\ (\hat{\beta}_4) = 0.0044$$
$$\hat{Var}\ (\hat{\beta}_5) = 698.3947$$
$$\hat{Var}\ (\hat{\beta}_6) = 608.2725$$
$$\hat{Var}\ (\hat{\beta}_7) = 872.8899$$
$$\hat{Var}\ (\hat{\beta}_8) = 651.1522$$
$$\hat{Var}\ (\hat{\beta}_9) = 56.0575$$

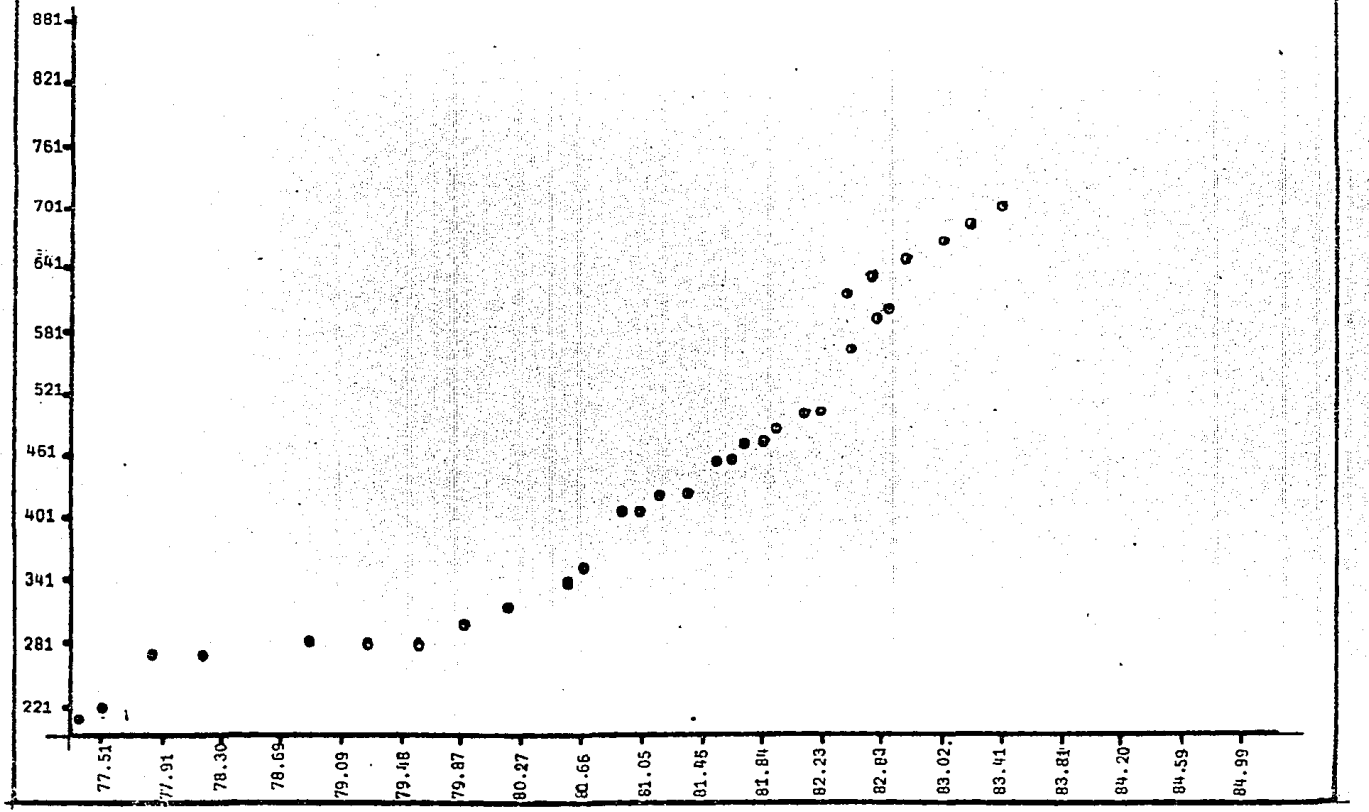
Estos cálculos son de suma importancia para comparar modelos y observar ganancias ó pérdidas al realizar transformaciones. Otro estadístico que nos ayuda a realizar comparaciones entre modelos es $\hat{\sigma}^2$, la cual es calculada como se muestra en el punto (2.6), es decir:

$$\hat{\sigma}^2 = \frac{y'y - y'X\beta}{n - p'} = 2528.173$$

Pasamos a calcular la tabla de análisis de varianza para comparar modelos que contengan un conjunto

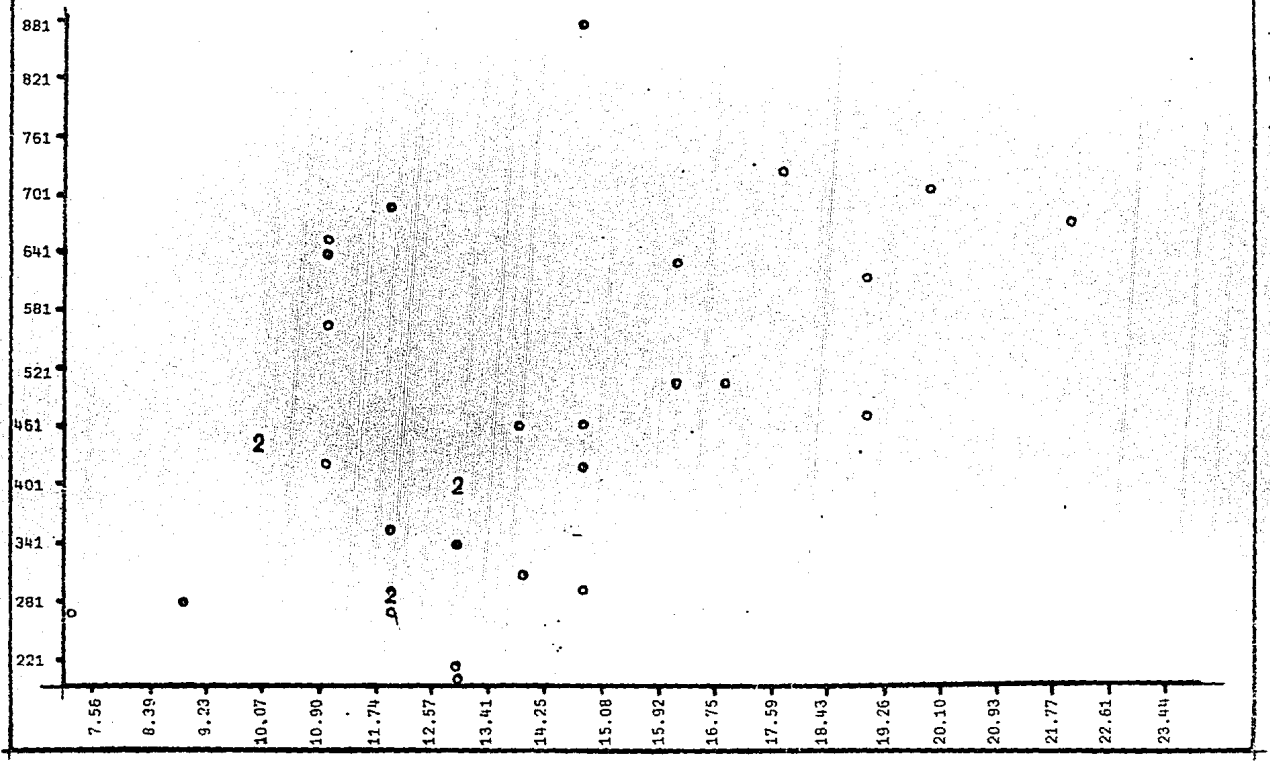
GRAFICA Nº 1

Y CONTRA Z COLUMNA 2



GRAFICA Nº 2

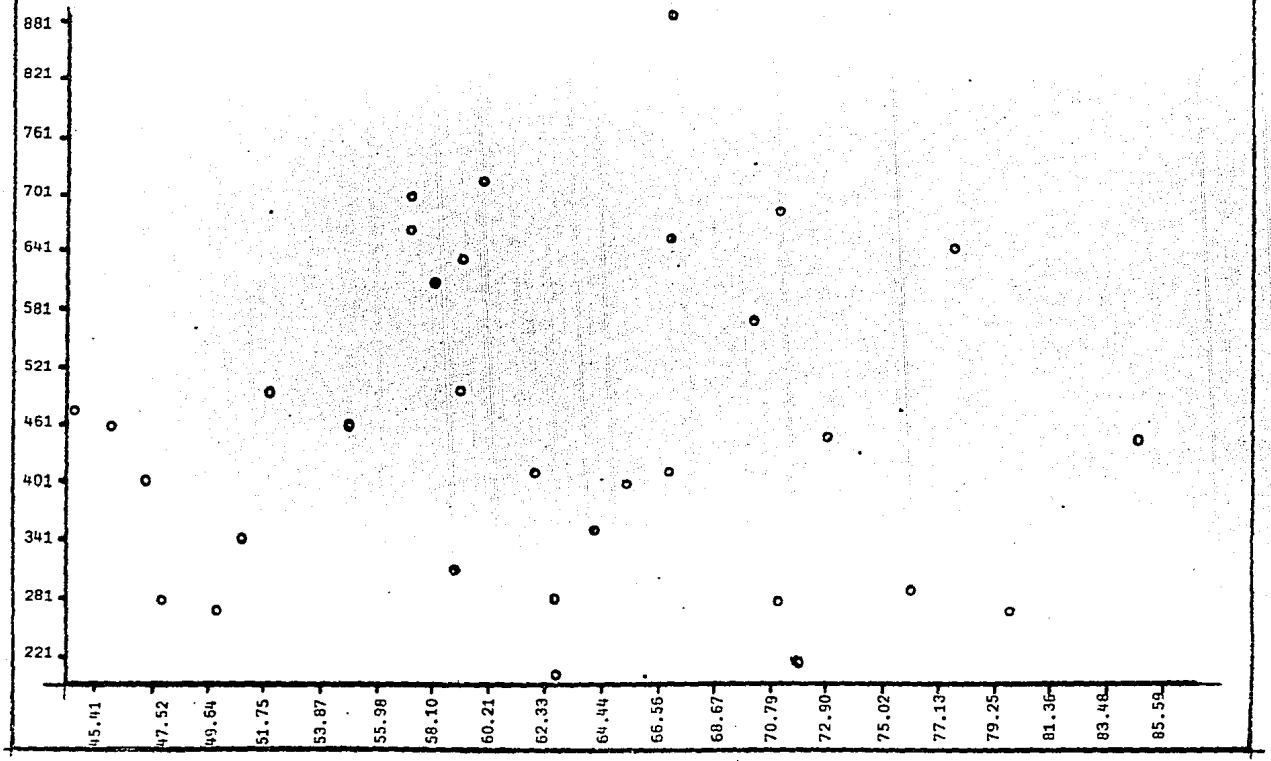
Y CONTRA Z COLUMNA 3



GRAFICA Nº 3

Y CONTRA Z COLUMNA 4

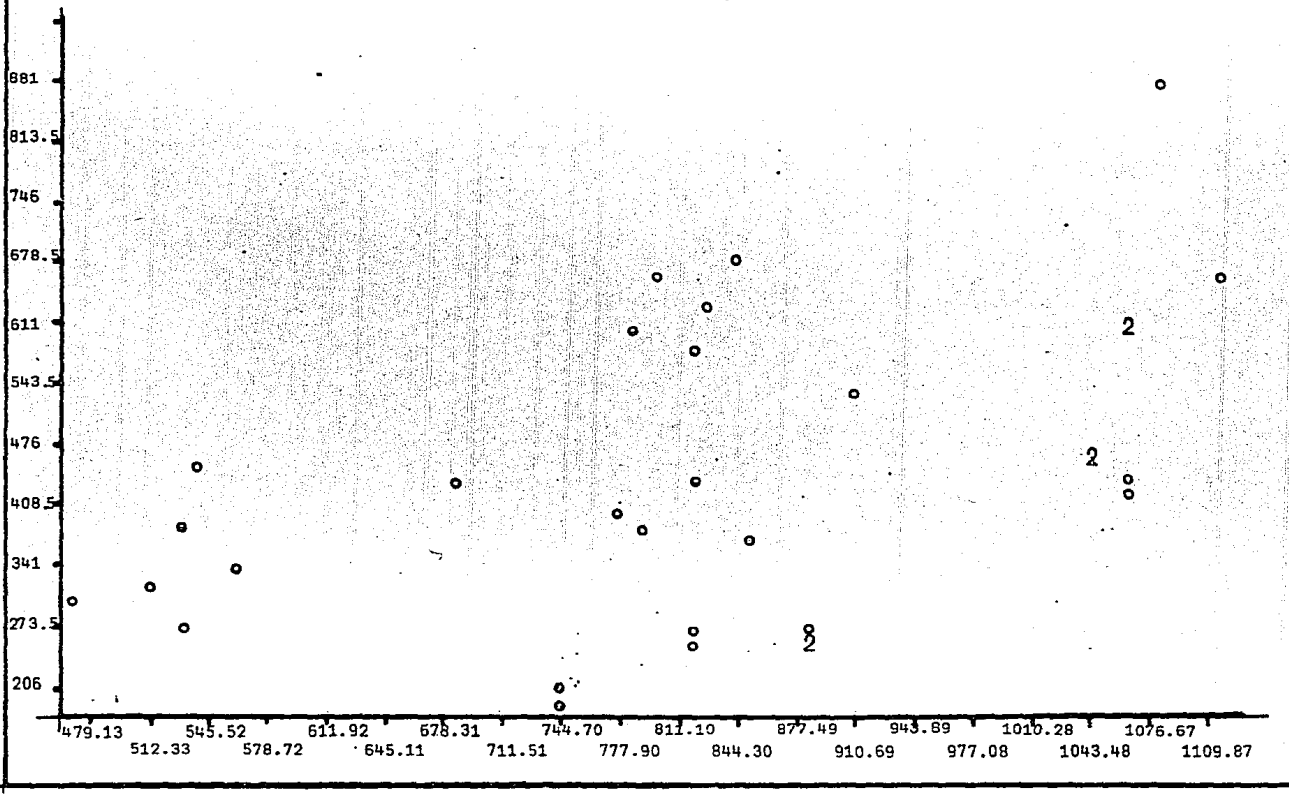
GM 86C



GRAFICA Nº 4

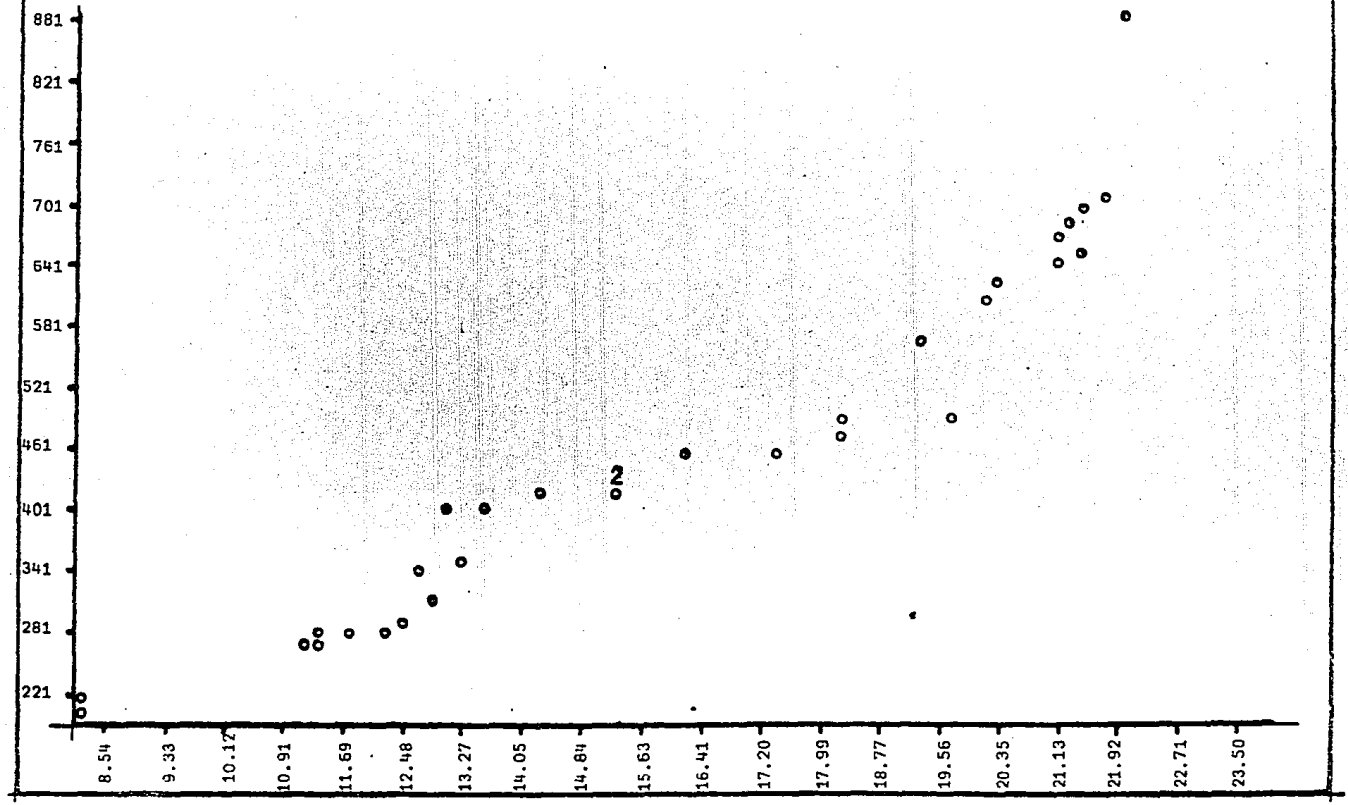
Y CONTRA Z COLUMNA 5

GM 86C



GRAFICA Nº 5

Y CONTRA Z COLUMNA 10



diferente de variables independientes, se puede decir que se contrasta H_0: $\beta_1=\beta_2=\ldots=\beta_p=0$ vs. H_α: al menos un $\beta_j\neq 0$ $j=1,2\ldots p$ ó también se puede realizar con un sub conjunto de β-*etas* y contrastarlas con el mismo tipo de hipótesis.

La tabla (2.3) de análisis de varianza del modelo 1.1 es:

Fuente	GL	SC	CM	F	P
H_0: $\beta = 0$	9	841569.469	93507.719	36.986	0.00000
Residual	22	55619.808	2528.173		
Total	31	897189.277			

Como F calculada es 36.986 mayor que una F (.05,9,22) = 2.34 rechazamos la hipotesis nula H_0: $\beta_1 = \beta_2 = \ldots. \beta p = 0$ por un margen bastante grande. La *P* significa que se trabaja con un nivel de significancia descriptivo; si $\alpha \geq P$ entonces rechazamos la hipótesis nula.

El valor del coeficiente de determinación cuya expresión está expuesta en el punto (2.13) es igual al coeficiente de correlación múltiple al cuadrado, a continuación se tiene que:

El coeficiente de correlación al cuadrado = .938

→ El coeficiente de correlación es .968504

es decir, que cerca del 94% de la variabilidad observada en la variable respuesta es explicada por las variables independientes.

Ahora tratamos de checar si el modelo inicial 1.1 es el correcto, o no, para lo cual es necesario realizar la prueba de hipótesis H_0: no hay carencia de ajuste vs. H_1: hay carencia de ajuste.

Realizando los cálculos pertinentes se tiene:

Tabla (2.4) **De Analisis de Varianza**

Fuente	GL	SC	CM	F
Residual	22	55619.808	2528.173	
Carencia De A.	15	50720.593	3381.373	3.882
Error Puro	7	4899.215	699.888	

De acuerdo a la tabla (2.4), se tiene una $F_{cal} = 3.882$ y por otro lado, se tiene una F de tablas $F(.05,15,7)=3.51$ es decir $F_{cal} > F_{tab}$, por lo tanto para una $\alpha = .05$ se rechaza H_0: No hay carencia de ajuste. Esto indica que el modelo es inadecuado para este nivel de significancia.*

Con lo anterior la tabla (2.3) es irrelevante y por lo tanto es necesario parar el análisis del modelo ajustado y pasar a examinar residuales. Debido a que si se tiene evidencia de carencia es obsoleto realizar in-

* Es de interés aclarar que la matriz X no tiene repeticiones, por lo cuál se aplicó la prueba arco iris (UTTS 1982).

tervalos de confianza, ya que las suposiciones en las cuales esos cálculos están basadas, son no verdaderos si hay carencia de ajuste en el modelo ajustado.

NOTA: Se observan sugerencias de transformaciones en las gráficas 1, 2, 3, 4, y 5.

Analisis De Residuales

Como se tiene evidencia que el modelo es el inadecuado los residuales exhibirán un comportamiento incongruente con las suposiciones hechas al ajustar el modelo.

En el análisis de residuales se tienen varias formas de graficación que nos sirven para realizar un diagnóstico y proponer si es necesario cambios a el modelo.

Se cuenta con graficación:

1. Directa
2. En orden de obtención en el tiempo (si esto se conoce)
3. Contra los valores ajustados $\hat{y}$
4. Contra las variables independientes $X_1, X_2 \ldots X_n$
5. Contra alguna variable explicativa

El diagnóstico más importante es el diagrama de

$\hat{e}_i$ contra $\hat{Y}_i$, es decir la graficación Nº 3. Cuando el diagrama es nulo*el modelo es correcto, por lo tanto los puntos tienden a caer en una banda horizontal, sin tener un comportamiento periodico. Pasamos a analizar la gráfica Nº 6 en lo cual observamos directamente que existe no lineabilidad, por lo cual, al parecer, se requiere de una transformación.

Los principales remedios para la nolineabilidad son las transformaciones, el incluir un nuevo término y a veces el producto de términos cruzados (los términos cruzados son variables explicativas, que están en producto).

En la gráfica Nº 6 residuales estudentizados contra "y" estimada, se observa una varianza no constante, hay necesidad de usar mínimos cuadrados ponderados o transformar las observaciones.

Tanto la gráfica Nº 8 como la Nº 9 (Residuales estudentizados contra X_2 y X_3 respectivamente). Exhiben un comportamiento cuadrático. Por lo tanto al parecer se requiere de una transformación de grado dos.

La gráfica Nº 10 al parecer no tiene ninguna visible tendencia, entonces hasta el momento a la variable X_4 no se le tiene que aplicar ninguna modificación.

* Diagrama nulo es aquél que no exhibe comportamiento alguno.

La gráfica Nº 11 tiene "a primera vista" un comportamiento cuadratico, con las puntas abiertas hacia arriba.

Es curioso ver que la gráfica de los residuales estudentizados contra tiempo (gráfica Nº 12) tiene una secuencia periódica y que en el último valor tiene un "brinco", por lo cual consideramos que ese valor va a causar "ruido", es decir ese dato debe de discrepar de los demás, causando una varianza elevada.

Ahora observamos la gráfica Nº 13 residuales menos contra residuales, la cual es de suma importancia puesto que en ella se visualizan tanto puntos discrepantes como influyentes.

Encontramos en esta un punto que sospechamos que es discrepante, cuyas coordenadas son (160.444,217.21), es decir que el punto tiende a dispararse del resto de los datos. (En el capítulo III se realizará la prueba para puntos discrepantes).

Es usual que en el análisis de regresión se asuma que los errores son distribuidos normalmente. El problema de no-normalidad de los errores es muy difícil de diagnosticar al examinar los residuales. Una técnica que ayuda es lo graficación de residuales estudentizados en papel normal. (Su respectiva teoría esta mostrada en

en el Apendice I). Los puntos se espera que caigan aproximadamente sobre una línea recta.

En la gráfica Nº 14 se observa la no-normalidad y para remediar el problema se requiere de transformar a la variable respuesta. A menudo ocurre que al transformar se tiene poca lineabilidad y varianza constante, pero la hipótesis de normalidad es bastante importante en el análisis, más no necesaria.

La tabla (2.5) muestra el cálculo del vector de residuales, del vector de residuales estudentizados, de la distancia de COOK y del vector de puntos discrepantes *T(Outlier)* .

Se aprecia directamente que la observación Nº 32 (4.88) tiene el valor máximo de la distancia de COOK, si este caso fuera omitido del análisis el estimador del vector beta sufriría un movimiento equivalente a trasladar el estimador inicial a un elipsoide de 11.99% de confianza. Por lo anterior nos inclinamos a concluir con certeza que no hay observaciones particulares que presenten una influencia contundente en el modelo.

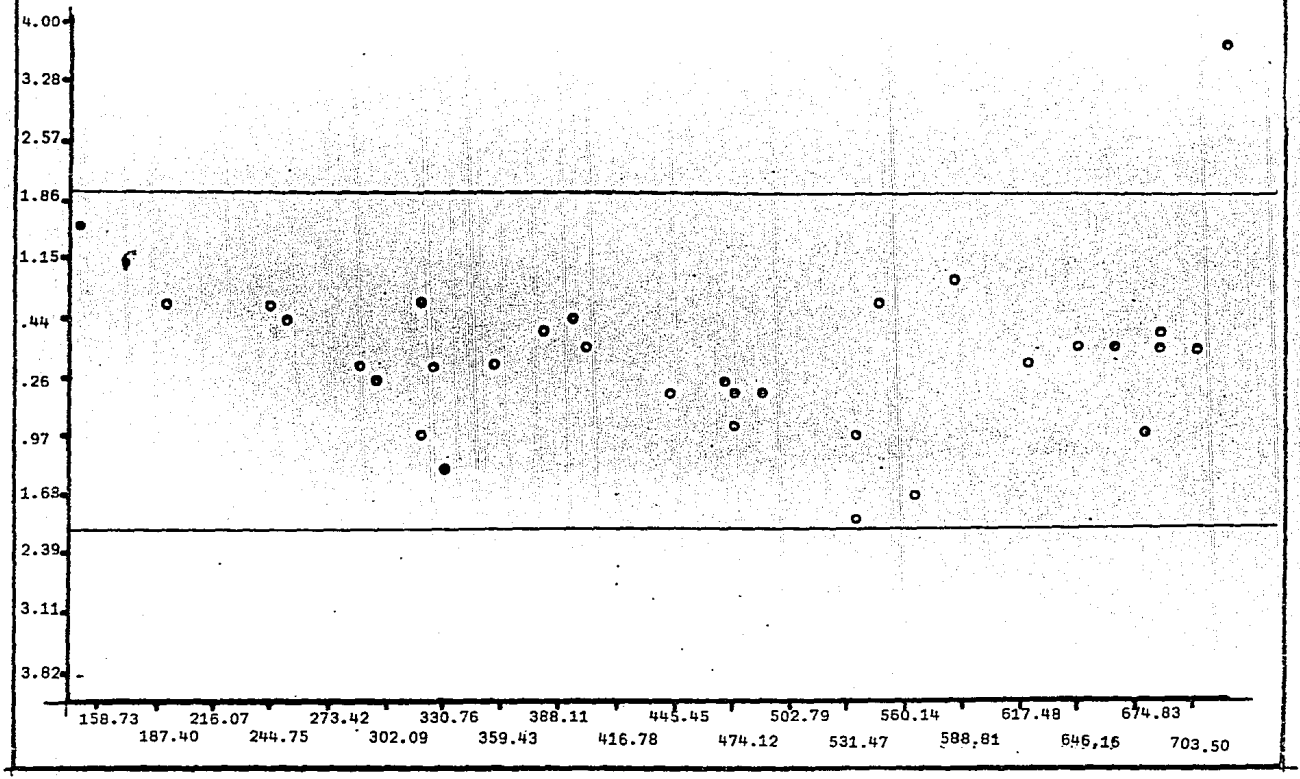
Como se puede observar en la tabla (2.5), el valor máximo de un *T(Outlier)* es el que pertenece a la observación Nº 32 (5.936252). Los candidatos para ser puntos discrepantes, son aquellos que tengan residuales ele

vados respecto a los demás. Por lo tanto, tenemos gran sospecha de que la observación número 32 es un punto discrepante, puesto que se tiene 3.71 de residual studentizado, siendo éste un valor extremo.

Al parecer, es necesario transformar al modelo inicial 1.1, debido a todas las características no linea bles que presenta.

GRAFICA Nº 6

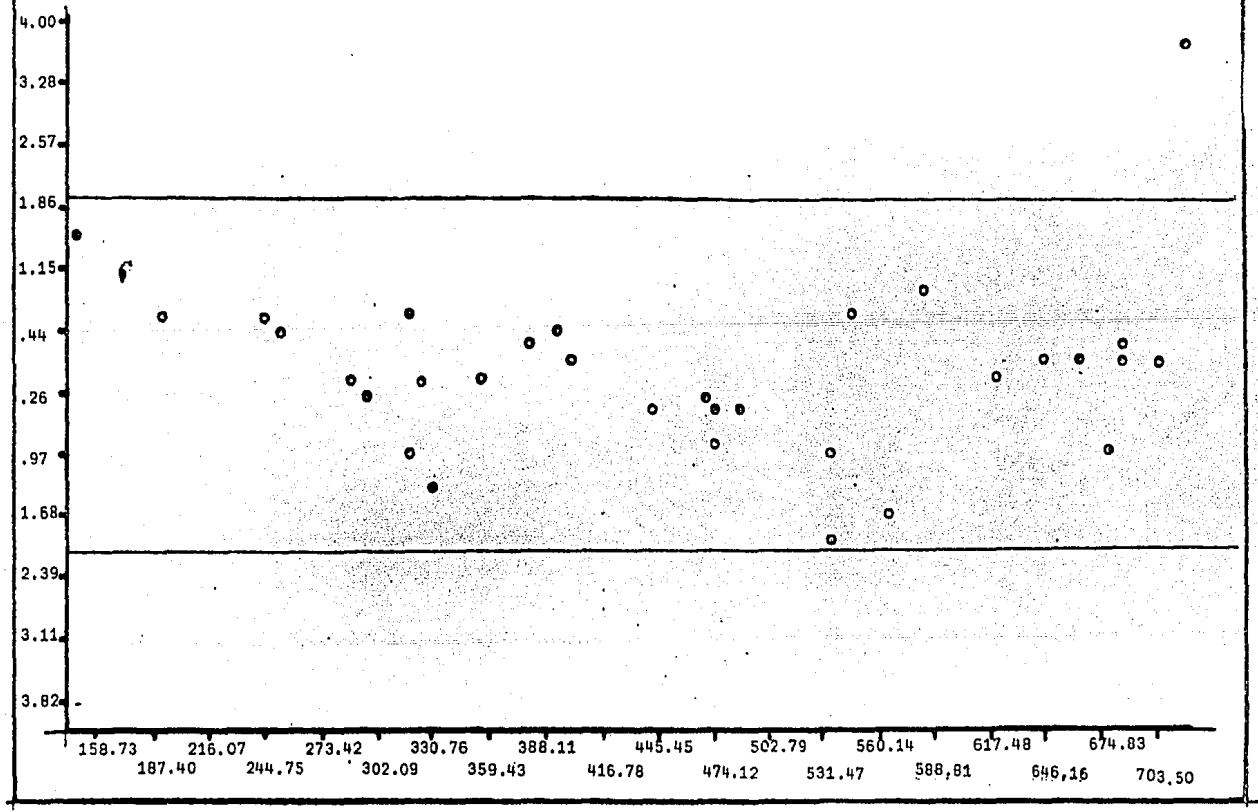
RESIDUALES . ESTUDIANTIZADOS CONTRA Y ESTIMADA



GRAFICA Nº 6

RESIDUALES . ESTUDIANTIZADOS CONTRA Y ESTIMADA

GM 86C

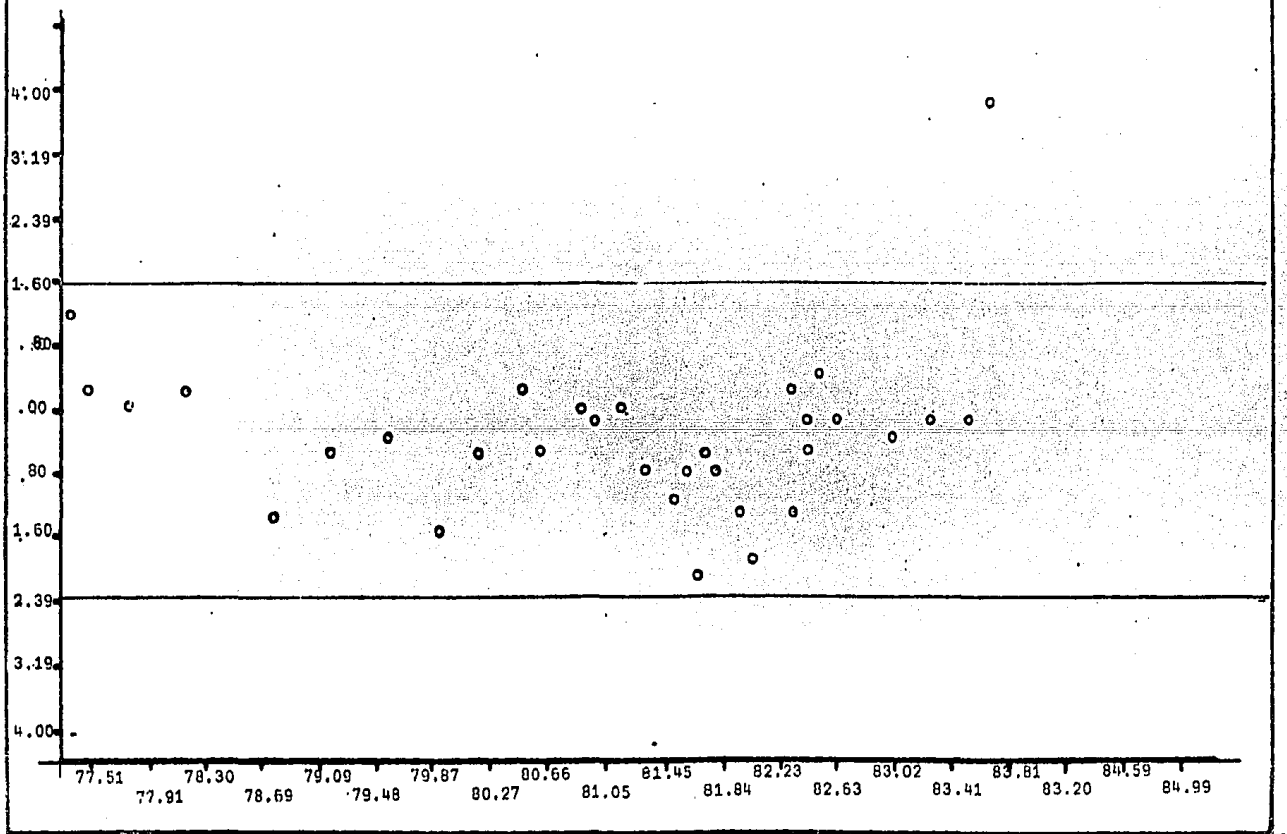


GRAFICA Nº 7

RESIDUALES . ESTUDIANTIZADOS CONTRA Z

COLUMNA 2

GM 86C

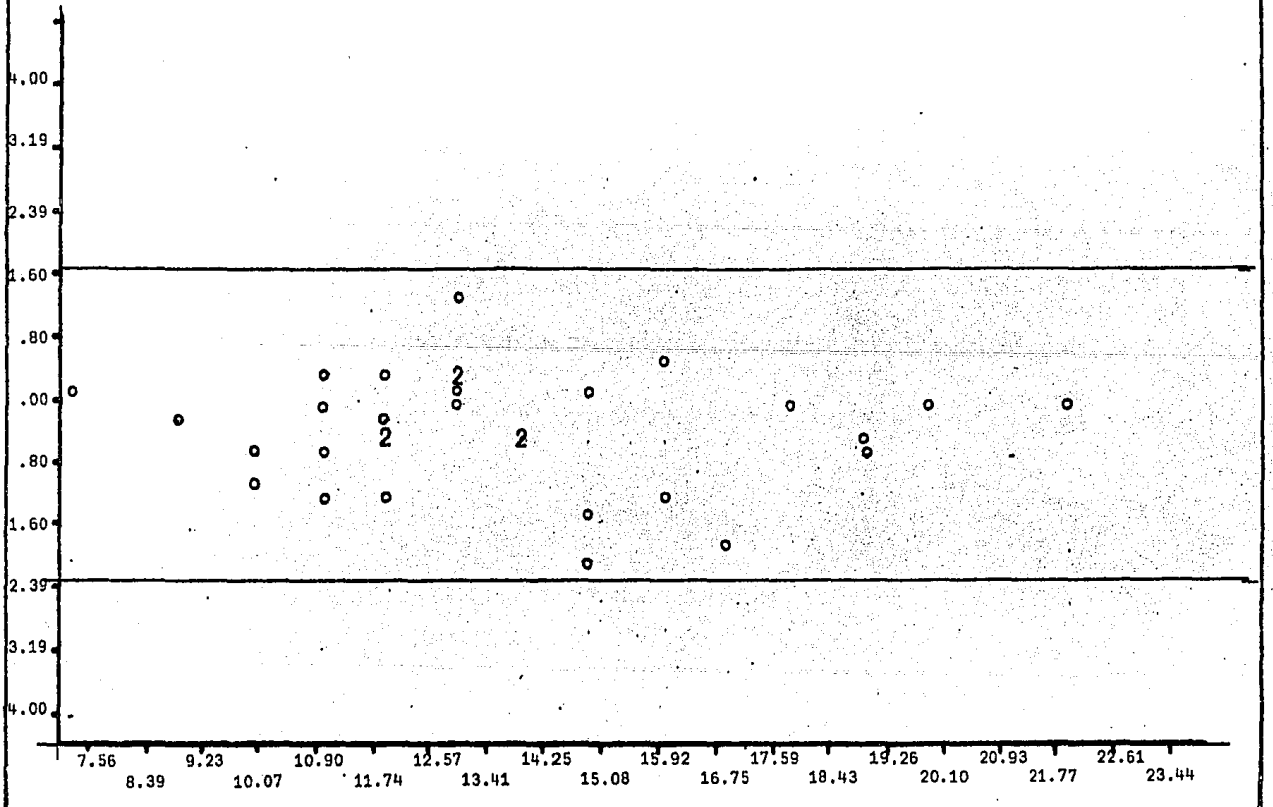


GRAFICA Nº 8

RESIDUALES . ESTUDIANTIZADOS CONTRA Z

COLUMNA 3

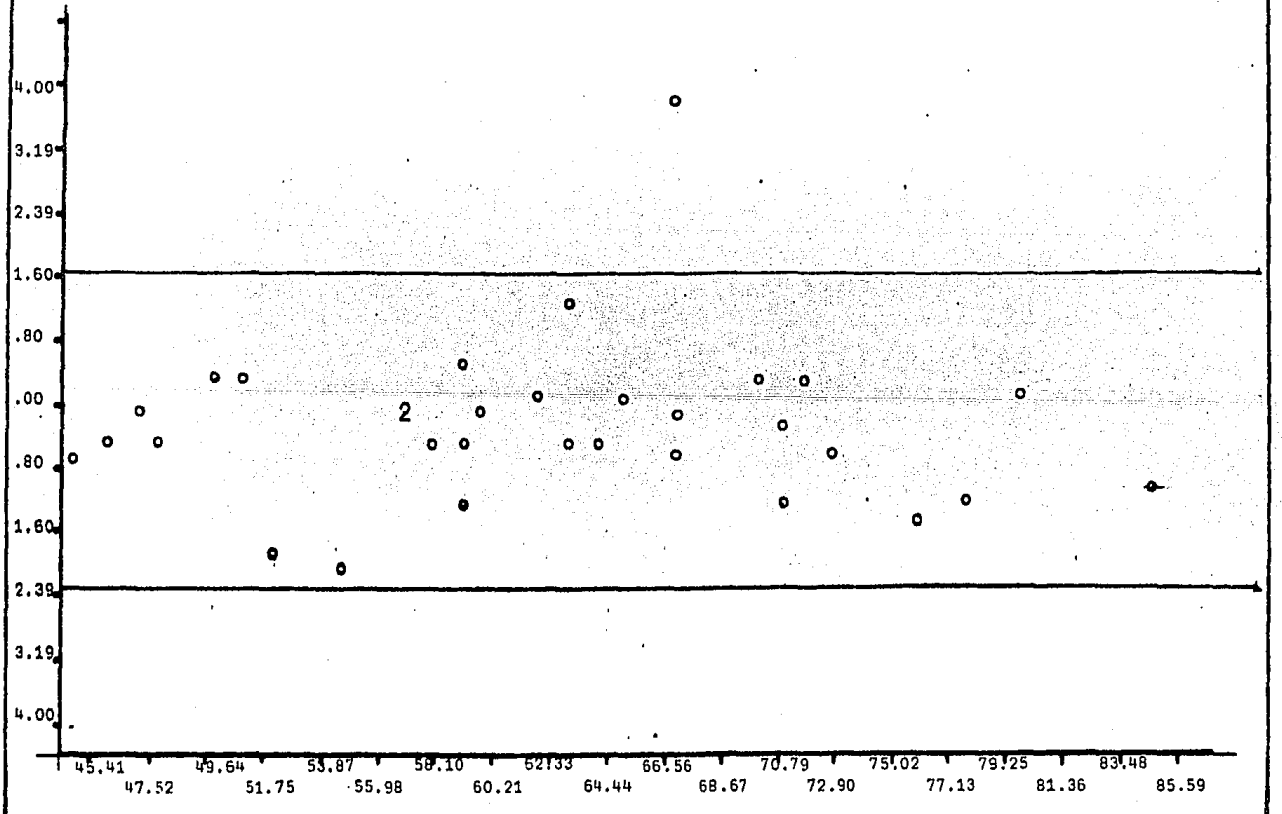
GM 86C



GRAFICA Nº 9

RESIDUALES ESTUDIANTIZADOS CONTRA Z

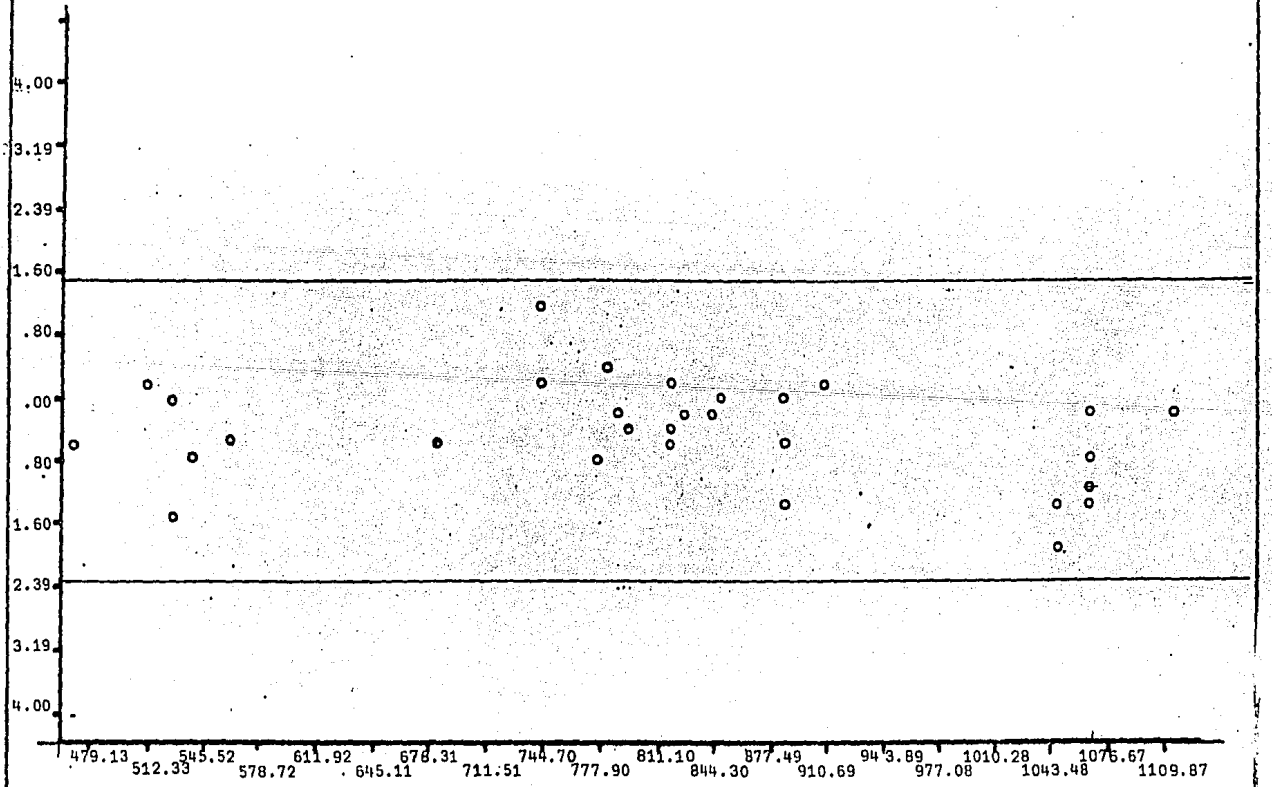
COLUMNA 4



GRAFICA Nº 10

RESIDUALES . ESTUDIANTIZADOS CONTRA Z

COLUMNA 5

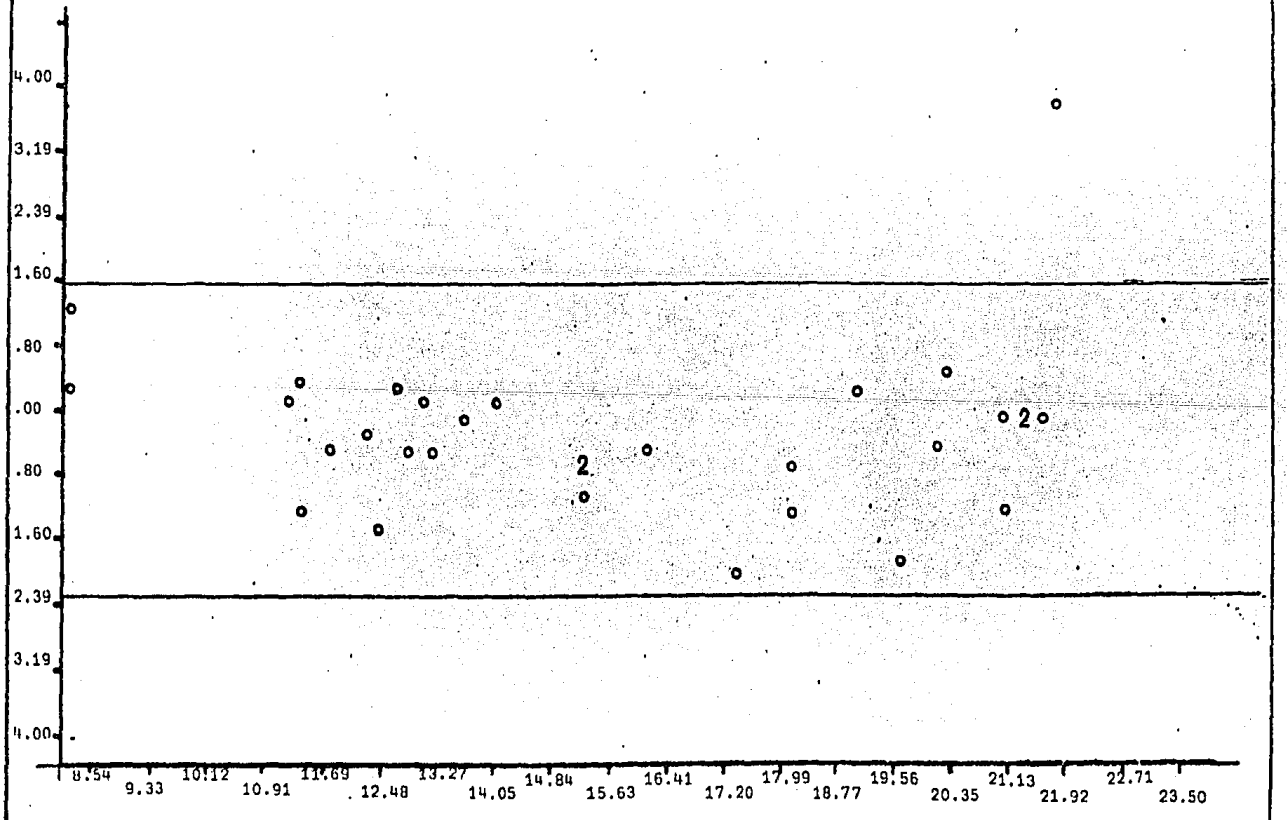


GRAFICA Nº 11

RESIDUALES . ESTUDIANTIZADOS CONTRA Z

COLUMNA 10

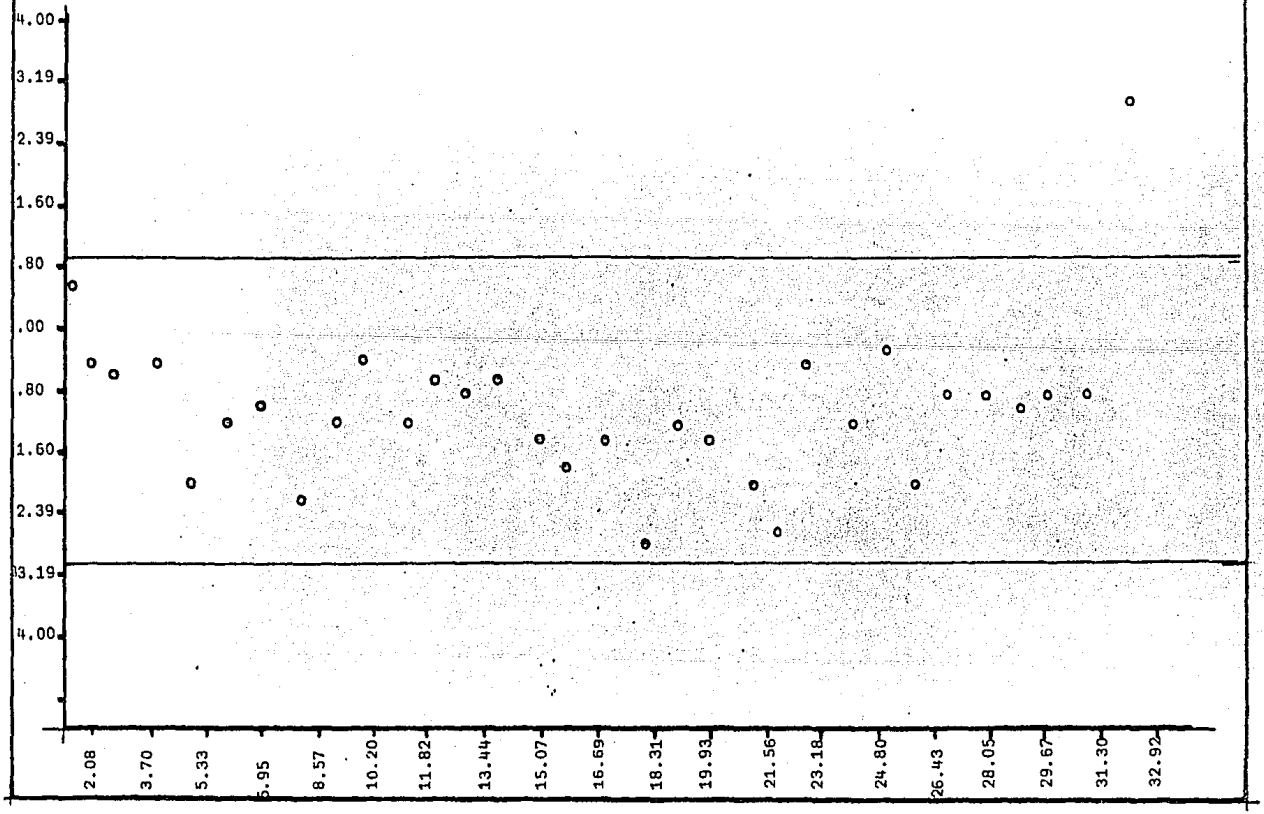
GM 86C



GRAFICA Nº 12

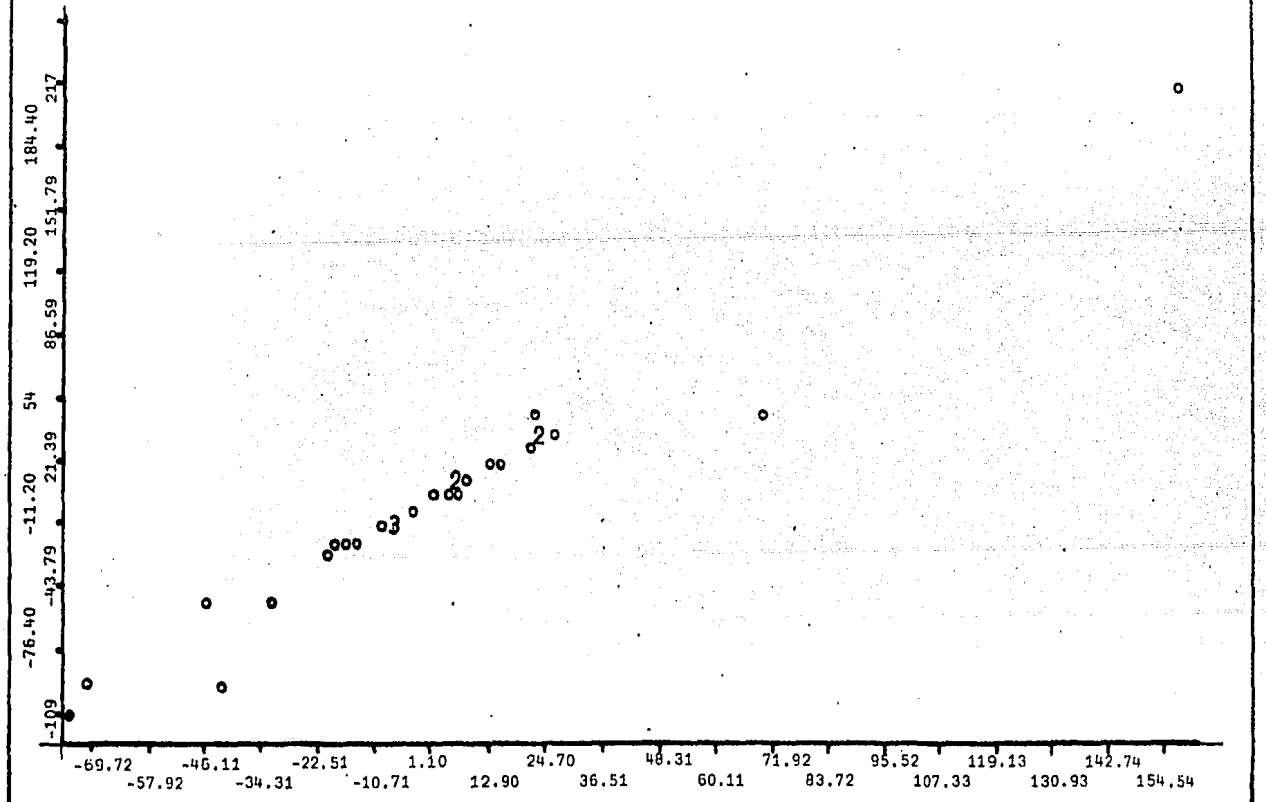
RESIDUALES . ESTUDIANTIZADOS CONTRA TIEMPO

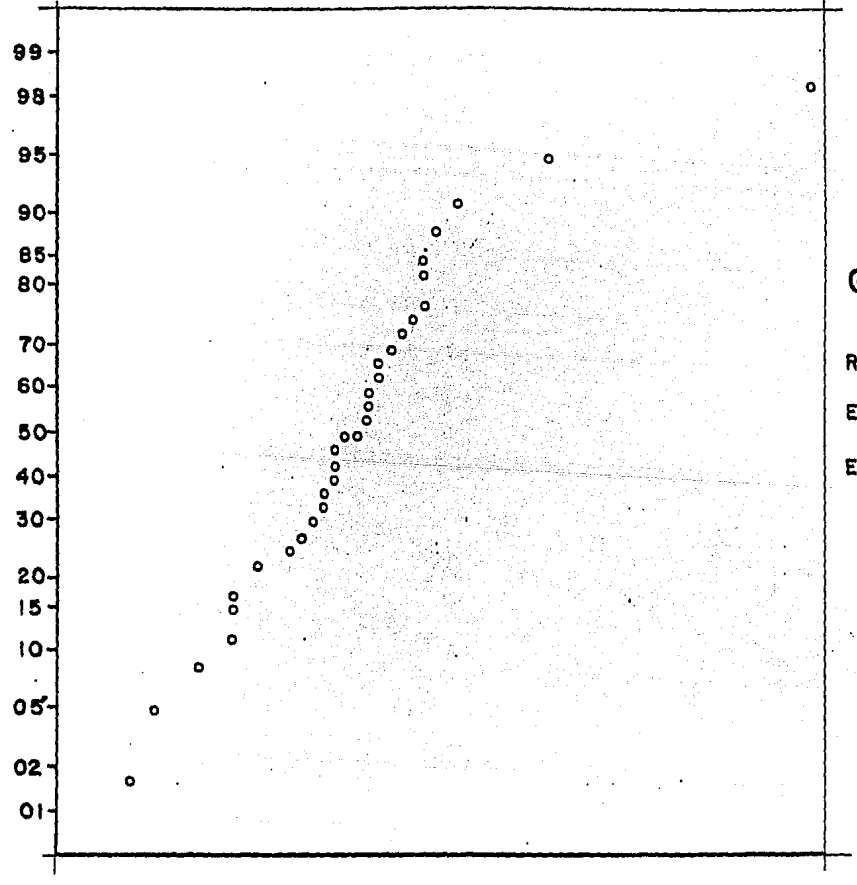
GM 86C



GRAFICA Nº 13 RESIDUALES MENOS CONTRA RESIDUALES

GM 86C





GM
86C

GRAFICA Nº 14

RESIDUALES

ESTUDIANTIZADOS

EN PAPEL NORMAL

T A B L A 2.5

VECTOR *YT*	VECTOR *Y*-ESTIMADA	VECTOR RESIDUAL	RESIDUALES STUDENTIZADOS	D(COOK)	T(OUTLIER)
207.51000	139.614164	67.895836	1.599437	0.103493	1.662297
217.38000	194.026572	23.353428	0.575229	0.017665	0.566278
270.71000	272.370000	14.039803	0.362974	0.009088	0.355695
272.37000	246.247206	26.122794	0.622151	0.016800	0.613265
280.36000	321.138882	-40.778882	-0.982775	0.045240	-0.981977
284.88000	296.708203	-11.828203	-0.271248	0.002425	-0.265456
288.48000	291.390629	-2.910629	-0.085081	0.000840	-0.083139
289.66000	332.456451	-42.796451	-1.249763	0.180554	-1.266826
317.21000	324.096238	-6.886238	-0.154688	0.000660	-0.151213
345.39000	322.623551	22.766449	0.599154	0.026961	0.590213
350.63000	356.735858	-6.105868	-0.135447	0.000448	-0.132388
394.36000	380.709405	13.650595	0.341914	0.006852	0.334944
402.59000	396.087832	6.502168	0.160718	0.001407	0.157115
412.18000	391.575837	20.604163	0.461163	0.005668	0.452754
423.32000	441.488154	-18.168154	-0.391841	0.002702	-0.384175
443.22000	473.677293	-30.457293	-0.770774	0.036781	-0.763431
452.05000	472.860570	-20.810570	-0.477948	0.007619	-0.469403
457.12000	534.710341	-77.590341	-1.817925	0.128179	-1.926731
460.05000	469.533372	-9.483372	-0.216806	0.001511	-0.212048
473.64000	488.795832	-15.155832	-0.363486	0.006001	-0.356200
490.88000	536.225804	-45.345804	-0.993840	0.021178	-0.993549
495.58000	566.152230	-70.572230	-1.596979	0.075134	-1.659406
567.79000	544.849370	22.940630	0.651688	0.044178	0.642941
608.80000	616.446779	-7.646779	-0.167722	0.000608	-0.163971
621.45000	583.068282	38.381718	0.854444	0.018465	0.849005
642.23000	682.322640	-40.092640	-0.997295	0.056127	-0.997166
652.32000	647.499964	4.820036	0.118852	0.000759	0.116157
665.99000	659.158687	6.831313	0.176979	0.002183	0.173033
690.19000	688.213944	1.976056	0.053105	0.000233	0.051887
697.14000	687.520060	9.619940	0.229496	0.002311	0.224489
712.27000	705.590011	6.679989	0.147578	0.000510	0.144257
881.24000	720.795633	160.444367	3.712825	0.487769	5.936252

CAPITULO III

AJUSTE DE DIFERENTES SUPERFICIES AL PROBLEMA

i) Aspectos Importantes antes de Transformar al Modelo

El arreglo X es una matriz de 32 X 10 por lo que se requiere analizar la multicolinealidad en el modelo, antes de realizar cualquier transformación.

Puede ocurrir que algunas de las variables explicativas $X_1, X_2, X_3, \ldots X_p$ esten en función exacta de otras, es decir, $X_9 = X_3 \times X_4$, por ejemplo.

Se trata de usar un número pequeño de variables independientes para explicar a la variable dependiente. En muchos casos, modelos basados en subconjuntos dan resultados más eficientes que modelos basados en más variables.

Un modelo empírico que usa pocas variables es más útil que uno con muchas variables.

Con lo anterior se presenta entonces el problema de selección de variables. Este tema está íntimamente relacionado con el conocido como multicolinealidad.

ii) **Multicolinealidad**

Multicolinealidad es la dependencia lineal que existe entre varias variables (al menos aproximada).

Sea S_{px1} un vector dirección tal que $S^1 S = 1$, diremos que existe colinealidad entre las columnas de la matriz X_{nxp} si el producto XS está muy cercano a un vector cero (*).

Esto es, al menos una columna de X puede ser explicada muy cercanamente por medio de una combinación lineal de las otras columnas de X.

Para medir la colinealidad es conveniente considerar la siguiente expresión cuadrática $(XS)^1 XS = S^1 X^1 XS$; existe colinealidad sí $S^1 (X^1 X) S$ se aproxima bastante a cero para algún S.

Se puede mostrar que para algún S, el valor de $S^1 (X^1 X) S$ es mayor o igual que el más pequeño eigenvalor de $X^1 X$.

Por lo cual la colinealidad puede ser diagnosticada si el eigenvalor más pequeño es relativamente "pequeño" respecto a los otros eigenvalores.

---

* Silvey 1969.

Una medida de colinealidad es el llamado número de condición K, definido como:

$$K = \left(\frac{\lambda g}{\lambda c}\right)^{1/2} \geq 1 \qquad (3.1)$$

donde λ_c es el eigenvalor más chico y λ_g es el eigenvalor más grande de X^1X. Con valores grandes sugiere colinealidad.

Sin embargo, este criterio no es invariante ante cambios de escala; por lo tanto su interpretación es difícil, depende de la precisión de la computadora y de los algoritmos usados.

Otra manera de analizar la colineabilidad entre dos variables puede ser por diagnóstico de la matriz de correlación muestral entre las variables explicativas potenciales, aunque solamente dependencias entre pares de variables podrán diagnosticarse de esta forma.

iii) **Selección de Variables**

Para llevar a cabo la selección de variables se usarán tres criterios: *R-cuadrada* (R^2_p), *R-Ajustada* ($\bar{R}^2_p$) y C_p-*Mallows*.

Recordando que R^2 es el cuadrado de la correlación entre la variable dependiente Y y la mejor combi-

nación de un conjunto de variables explicativas.

La forma para calcular R^2 en un modelo de p-*parámetros* es:

$$R_p^2 = 1 - \frac{SCR}{SYY} \; ; \quad 0 \leq R_p \leq 1 \qquad (3.2.)$$

La cual es la proporción de variabilidad en Y explicada por la regresión en las X's.

La $\bar{R}_p^2$ se calcula mediante la siguiente expresión, que toma en consideración los grados de libertad.

$$\bar{R}_p^2 = 1 - \left(\frac{n-1}{n-p}\right)(1-R_p^2) \qquad (3.3)$$

Buenos modelos serán aquello que tengan valores grandes de $\bar{R}_p^2$. Es importante hacer notar que $\bar{R}^2$ puede llagar a ser negativa.

Un estadístico alternativo que actualmente es muy -- usado, es el conocido como C_p de Mallows*. Supone que el objetivo de regresión es la estimación de valores ajustados. La precisión de la estimación de valores ajustados depende - de dos elementos: la varianza y el sesgo de los estimadores. Vamos a preferir aquel subconjunto de modelos que tomen el - más pequeño error cuadrático medio, posible.

---

* Mallows (1973)

Un criterio importante es el de la siguiente forma:

$$\zeta_p = \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} \{E(y_i) - E(\hat{y}_i)\}^2 + \sum_{i=1}^{n} V(\hat{y}_i) \right) \quad (3.4)$$

el cual se debe tomar el más pequeño.

C_p de Mallows es un estimador de ζ_p que tiene la siguiente forma

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

Suponiendo que

$$E(\hat{\sigma}^2) = \sigma^2 \text{ y que } \quad E\left(\frac{SCR}{\hat{\sigma}^2}\right) \doteq \frac{(n-p)\sigma^2}{\sigma^2}$$

$$\rightarrow \quad E\{C_p \mid \text{sesgo} = 0\} \doteq \frac{(n-p)\sigma^2}{\sigma^2} + 2p - n$$

$E\{C_p \mid \text{sesgo} = 0\} \doteq p$ Para un modelo adecuado. La gráfica de C_p contra p seleccionará los modelos "adecuados"

Para un sesgo despreciable C_p es aproximadamente p. Entonces el procedimiento de selección involucra identificar el subconjunto de variables explicativas el cual da a la vez mínimo C_p y $C_p \doteq p$.

Mallows sugiere que los buenos modelos serán

aquellos que tengan $C_p - p$ negativa o muy pequeña.

Para el modelo inicial 1.1 se tienen 9 predictores, por lo tanto hay $2^9 = 512$ posibles subconjuntos de modelos incluyendo al modelo 1.1.

Al parecer el trabajo computacional involucrado es grande (todas las regresiones posibles) pero puede "remediarse" la aparente dificultad, debido a las relaciones que existen entre R^2_p, $\bar{R}^2_p$ y C_p, por lo que sólo es necesario calcular directamente uno de ellos.

La relación entre el coeficiente de correlación multiple y el coeficiente de correlación multiple ajustado es:

$$\bar{R}^2_p = 1 - \frac{(n-1)\ (1-R^2_p)}{n-p}$$

y la relación con C_p es:

$$C_p = \frac{(n-K)\ (1-R^2_p)}{(1-R^2_K)} + 2p - n$$

$$= (n-p)\ \frac{(1-\bar{R}^2_p)}{(1-\bar{R}^2_{K_1})} + 2p - n \qquad (*)$$

Donde K^1 es el número de predictores del Modelo

---

* Mary L. Thompson (1978) International Statistical Review p.p. 1-19 y 129-148.

inicial 1.1 Aplicando SAREG SELECCION y además seleccionado los mejores subconjuntos obtenemos lo siguiente.*

Subconjuntos con 1 Variable

R-cuadrada	*R-ajustada*	C_p-*Mallows*	
.912755	.909847	2.961312	X_0, X_9
.208592	.182211	252.852286	X_0, X_2

Subconjuntos con 2 Variables

R-cuadrada	*R-ajustada*	C_p-*Mallows*	
.920958	.915507	2.050159	X_5, X_9, X_0
.918139	.912494	3.050475	X_0, X_3, X_9
.916947	.911220	3.473450	X_0, X_1, X_9
.914541	.908647	4.32740	X_0, X_7, X_9
.913786	.907840	4.59526	X_0, X_6, X_9

Subconjuntos con 3 Variables

R-cuadrada	*R-ajustada*	C_p-*Mallows*	
.928082	.910376	1.522074	X_0, X_1, X_5, X_9
.923512	.915317	3.143638	X_0, X_3, X_5, X_9
.922729	.914450	3.421698	X_0, X_2, X_5, X_9
.922610	.914318	3.463982	X_0, X_5, X_7, X_9
.921619	.913221	3.815596	X_0, X_5, X_6, X_9

* El criterio que utiliza SAREG es el propuesto por G.M. Furnival y R.W. Wilson (Technometrics, Vol. 16, 1974, pp. 499-511.

Subconjuntos con 4 Variables

R-cuadrada	*R-ajustada*	C_p-*Mallows*	
.933555	.923711	1.579901	X_0, X_1, X_4, X_5, X_9
.931293	.921114	2.382642	X_0, X_1, X_3, X_5, X_9
.931043	.920827	2.471147	X_0, X_1, X_2, X_5, X_9
.931043	.920827	2.471244	X_0, X_1, X_5, X_7, X_9
.928515	.917924	3.368496	X_0, X_1, X_5, X_8, X_9

Subconjuntos con 5 Variables

R-cuadrada	*R-ajustada*	C_p-*Mallows*	
.935297	.922854	2.961580	$X_0, X_1, X_4, X_5, X_2, X_9$
.934857	.922330	3.117552	$X_0, X_1, X_4, X_5, X_6, X_9$
.934440	.921832	3.265811	$X_0, X_1, X_3, X_4, X_5, X_9$
.934264	.921622	3.328316	$X_0, X_1, X_2, X_4, X_5, X_9$
.934012	.921322	3.417737	$X_0, X_1, X_3, X_5, X_7, X_9$

Subconjuntos con 6 Variables

R-cuadrada	*R-ajustada*	C_p-*Mallows*	
.936819	.921656	4.421349	$X_0, X_2, X_7, X_4, X_5, X_1, X_9$
.936674	.921476	4.472894	$X_0, X_6, X_7, X_4, X_5, X_1, X_9$
.935383	.919875	4.930921	$X_0, X_2, X_6, X_7, X_5, X_1, X_9$
.935177	.919620	5.004010	$X_0, X_2, X_6, X_4, X_5, X_1, X_9$
.935164	.919604	5.008684	$X_0, X_3, X_2, X_7, X_5, X_1, X_9$

Subconjuntos con 7 Variables

R-cuadrada	*R-ajustada*	C_p-*Mallows*	
.937886	.919769	6.042926	$X_0, X_3, X_6, X_7, X_4, X_5, X_1, X_9$
.936932	.918538	6.381254	$X_0, X_2, X_3, X_7, X_4, X_5, X_1, X_9$
.936850	.911843	6.410501	$X_0, X_8, X_6, X_7, X_4, X_5, X_1, X_9$
.936820	.918392	6.421151	$X_0, X_8, X_2, X_7, X_4, X_5, X_1, X_9$
.936283	.917698	6.611778	$X_0, X_8, X_3, X_7, X_4, X_5, X_1, X_9$

Subconjuntos con 8 Variables

R-cuadrada	*R-ajustada*	C_p-*Mallows*	
.938006	.916443	8.000103	$X_0, X_2, X_3, X_6, X_7, X_4, X_5, X_1, X_9$
.937892	.916289	8.040655	$X_0, X_8, X_3, X_6, X_7, X_4, X_5, X_1, X_9$
.937607	.915905	8.141748	$X_0, X_8, X_2, X_6, X_7, X_4, X_5, X_1, X_9$
.936933	.914997	8.380884	$X_0, X_8, X_2, X_3, X_7, X_4, X_5, X_1, X_9$
.935972	.913702	8.721979	$X_0, X_8, X_2, X_3, X_6, X_4, X_5, X_1, X_9$

Subconjuntos con 9 Variables

R-cuadrada	*R-ajustada*	C_p-*Mallows*	
.912646	.938007	10.0000	$X_0, X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$

X_0 Representa el término independiente y X_i representa a el predictor *i-ésimo*

Analizando los subconjuntos, decidimos escoger al conjunto que tiene como elementos X_0, X_1, X_5, X_9, el cual

tiene una *R-cuadrada* igual a .928082, una *R-ajustada* igual a .910376 y un C_p-*Mallows* igual a 1.522074; tanto la *R-cuadrada* como la *R-ajustada* se aproximan bastante a uno y tenemos una diferencia entre C_p y P de 1.477926. De acuerdo con la sugerencia de *Mallows* el modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_5 X_5 + \beta_9 X_9 + \varepsilon \qquad (3.5)$$

es bueno.

La expresión (3.5) asume que

$$E(\varepsilon) = 0 \;\; ; \;\; v(\varepsilon) = \sigma^2 In \;\; y \;\; Cov(\varepsilon_i, \varepsilon_j) = 0 \;\; \forall \; i,j \;\; i \neq j$$

Tenemos que la *R-cuadrada* es igual a .928082, es decir cerca del 92.8% de la variabilidad observada en la variable respuesta es explicada por las variables independientes del Modelo (3.5) y la *R-cuadrada* del modelo inicial 1.1 es de .938, por lo cual la diferencia entre .938 y 928082 es de .009918. Que nos lleva a concluir que la pérdida es cercana al .99%, la cual es bastante pequeña, y esto se debe a la reducción en el No. de variables.

Se puede decir que la matriz reducida proporciona bastante información y el Modelo (3.5) es uno de los más indicados para proseguir con el análisis.

Iremos eliminando las columnas: X_8, X_7, X_6, X_4,

X_3 y X_2 una a una y analizaremos que es lo que ocurre en las varianzas, en los Betas estimados, en su tabla de análisis de varianza, en su varianza estimada, en su coeficiente de correlación y en la condición K.

Eliminando la columna 9 (X_8) tenemos que:

La condición K no converge.*

Betas	Estimación de Betas	Varianzas
β_0	-2553.98	1379010.04
β_1	31.71	246.20
β_2	-1.01	22.96
β_3	0.48	1.55
β_4	0.05	0.00
β_5	39.84	666.89
β_6	15.20	580.51
β_7	-25.09	834.12
β_9	22.84	52.80

La varianza estimada es 2418.26

Tabla de análisis de Varianza

F.V.	G.L	S.C	C.M	F	P
Regresión	8	841569.208	105196.15	43.501	0.000
Residual	23	55620.06	2418.26		
Total	31	897189.27			

* La razón por la cual no converge es por que el Eigenvalor más chico tiende a ser cero.

Coeficiente de correlación al cuadrado es = .938

Eliminando la columna 8 (X_7) tenemos que:

La condición K no converge

Betas	Estimación de Betas	Varianzas
β_0	-2468.83	1355423.72
β_1	30.07	240.16
β_2	0.27	20.55
β_3	0.66	1.49
β_4	0.07	0.00
β_5	35.94	640.15
β_6	17.22	569.26
β_9	23.11	52.16

La varianza estimada es 2393.574

Tabla de Análisis de Varianza

F.V	G.L.	S.C.	C.M.	F	P
Regresión	7	839743.495	119963.356	50.119	0.000
Residual	24	57445.782	2393.574		
Total	31	897189.227			

El coeficiente de correlación al cuadrado es = .936

Eliminando la columna 7 (X_6) tenemos que:

La condición K no converge

Betas	Estimación de Betas	Varianzas
β_0	-2463.72	1329410.54
β_1	30.37	235.39
β_2	-0.96	17.26
β_3	0.40	1.33
β_4	0.06	0.00
β_5	41.21	575.53
β_9	24.10	49.35

La varianza estimada es 2347.72

Tabla de Análisis de Varianza

F.V.	G.L.	S.C.	C.M.	F	P
Regresión	6	838496.18	139749.36	59.525	0.000
Residual	25	58693.09	2347.72		
Total	31	897189.27			

El coeficiente de correlación al cuadrado es = .935

Eliminando la columna 5 (X_4) tenemos que:

La condición K no converge

Betas	Estimación de Betas	Varianzas
β_0	-2121.23	1218394.59
β_1	26.10	218.16
β_2	-1.89	16.46
β_3	0.63	1.28
β_5	40.68	576.13
β_9	27.41	38.86

La varianza estimada es 2351.27

Tabla de Análisis de Varianza

F.V.	G.L.	S.C.	C.M.	F	P
Regresión	5	836056.07	167211.21	71.11	0.000
Residual	26	61133.20	2351.27		
Total	31	897189.27			

El coeficiente de correlación al cuadrado es = .932

Eliminando la columna 4 (X_3)

La condición K no converge

Betas	Estimación de Betas	Varianzas
β_0	-2081.96	1182542.17
β_1	26.30	212.48
β_2	-3.33	9.57
β_5	47.46	418.07
β_9	27.51	37.83

La varianza estimada es 2291.381

Tabla de Análisis de Varianza

F.V.	G.L.	S.C.	C.M.	F	P
Regresión	4	835321.98	208830.49	91.13	0.000
Residual	27	61867.29	2291.38		
Total	31	897189.27			

El coeficiente de correlación al cuadrado es = .931

Eliminando la columna 3 (X_2) tenemos que:

La condición K = 9.8535

Betas	Estimación de Betas	Varianzas
β_0	-1944.71	1172942.57
β_1	24.10	209.51
β_5	40.50	378.53
β_9	27.34	38.02

La varianza estimadas es 2304.436

Tabla de Análisis de Varianza

F.V.	G.L.	S.C.	G.M.	F	P
Regresión	3	832665.05	277555.01	120.44	0.000
Residual	28	64524.22	2304.43		
Total	31	897189.27			

El coeficiente de correlación al cuadrado es = .928

iv) **Comparaciones de Interés**

Comparaciones entre los estimadores de los modelos 1.1 y 3.6, los contrastes que a continuación se llevarán a cabo, son respecto a las varianzas de los estimadores, a los análisis de varianza y a sus respectivos coeficientes de correlación, con el propósito de que más adelante se analicen sus pérdidas o ganancias al cancelar variables explicativas.

	Modelo 1.1		Modelo 3.6	
	Estimación de Betas	Varianzas	Estimación de Betas	Varianzas
β_0	-2555.72	141004.79	-2553.98	1379010.04
β_1	31.74	262.29	31.71	246.20
β_2	-1.00	24.83	-1.01	22.96
β_3	0.48	1.65	0.48	1.55
β_4	0.05	0.00	0.05	0.00
β_5	39.83	698.39	39.84	666.89
β_6	15.22	608.27	15.20	580.51
β_7	-25.10	872.88	-25.09	834.12
β_9	22.83	56.05	22.84	52.80

Como se puede observar la varianza de $\hat{\beta}_0$ en el Modelo 3.6 aumentó aproximadamente 10 veces respecto a la varianza de $\hat{\beta}_0$ del Modelo 1.1 teniendo en cuenta que las estimaciones de $\hat{\beta}_0$ en ambos modelos se estabilizan. Sin embargo obtenemos ganancia en las varianzas de los $\hat{\beta}_i$, $i = 1,9$, $i \neq 8$ del Modelo 3.6, puesto que disminuye y los Betas estimados permanecen casi invariables.

Por otro lado la varianza estimada del Modelo 1.1 es de 2528.17, mientras que la varianza estimada del Modelo 3.6 es de 2418.26, obteniéndose una ganancia de 109.9

Respecto a la tabla de análisis de varianza, en ambos Modelos se sigue rechazando H_0: $\beta = 0$

El coeficiente de correlación no sufre ninguna modificación hasta el momento.

Cancelando la columna 8 (X_7) tenemos el siguiente Modelo.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_9 X_9 + \varepsilon \qquad (3.7)$$

Bajo los supuestos $E(\varepsilon) = 0$; $V(\varepsilon) = \sigma^2 I_n$ y

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall \ i, \ j \quad i \neq j$$

Comparaciones entre los estimadores de los Modelos (3.6) y (3.7)

	Modelo 3.6		**Modelo 3.7**	
	Estimación de Betas	Varianzas	Estimación de Betas	Varianzas
β_0	-2553.98	1379010.04	-2468.83	1355423.72
β_1	31.71	246.20	0.27	240.16
β_2	-1.01	22.96	0.27	20.55
β_3	0.48	1.55	0.66	1.49
β_4	0.05	0.00	0.07	0.00
β_5	39.84	666.89	35.94	640.15
β_6	15.20	580.51	17.22	569.26
β_9	22.84	52.80	23.11	52.16

Tanto en el Modelo 3.6 como en el Modelo 3.7 se

observa una estabilidad en la estimación y varianza de las Betas estimadas.

Las tablas de análisis de varianza en ambos modelos rechazan H_0: $\beta = 0$.

A lo que respecta al coeficiente de correlación se tiene que se han perdido dos milésimas, pero se ha ganado en la varianza estimada 24.68.

Pasamos a cancelar la columna 7 (X_6), obteniendo el Modelo sig.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_9 X_9 + \varepsilon \qquad (3.8)$$

Bajos los supuestos $E(\varepsilon) = 0$; $Var(\varepsilon) = \sigma^2 I_n$

Comparaciones entre los estimadores de los Modelos (3.7) y (3.8).

Modelo 3.7		Modelo 3.8	
Estimación de Betas	Varianzas	Estimación de Betas	Varianzas
β_0 -2468.83	1355423.72	-2463.72	1329410.54
β_1 30.07	240.16	30.37	235.39
β_2 0.27	20.55	-0.96	17.26
β_3 0.66	1.49	0.40	1.33
β_4 0.07	0.00	0.06	0.00
β_5 35.94	640.15	41.21	575.53
β_9 23.11	52.16	24.10	49.35

Como se puede observar la varianza de $\hat{\beta}_0$ disminuye en 30,000 aproximadamente al pasar el Modelo 3.7 al 3.8, las demás varianzas están disminuyendo, mientras que los estimadores de Betas siguen estables.

La varianza estimada se reduce de 2393.57 a 2347.72 obteniéndose una ganancia de 45.85.

Las tablas de análisis de varianza satisfactoriamente siguen rechazando $H_0: \beta = 0$

Respecto al coeficiente de correlación se ha perdido otra milésima puesto que en el Modelo 3.7 se tiene .936 y en el Modelo 3.8 tenemos .935.

Cancelando la columna 5 (X_4) tenemos el siguiente Modelo.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_5 + \beta_9 X_3 + \varepsilon \qquad (3.9)$$

Bajo los supuestos $E(\varepsilon) = 0$; $Var(\varepsilon) = \sigma^2 I_n$ y $Cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall, i, j \quad i \neq j$.

Realizamos las comparaciones entre los estimadores de los Modelos (3.8) y (3.9).

Modelo 3.8		Modelo 3.9	
Estimación de Betas	Varianzas	Estimación de Betas	Varianzas
β_0 -2463.72	13929410.54	-2121.23	1218394.59
β_1 30.37	235.39	26.10	218.16
β_2 -0.96	17.26	-1.89	16.46
β_3 0.40	1.33	0.63	1.28
β_5 41.21	575.53	40.68	576.13
β_9 24.10	49.35	27.41	38.86

Se observa que la varianza de $\hat{\beta}_0$ del Modelo 3.9 difiere en más de 100 000 respecto a la del Modelo 3.8, sin alterarse en gran medida la estimación de $\hat{\beta}_0$. Las varianzas de las demás Betas sufren un aumento o disminución muy pequeño al igual que sus respectivas estimaciones de parámetros.

La varianza estimada sufre un aumento de 3.55 al pasar del Modelo 3.8 al 3.9.

Tanto la tabla de análisis de varianza del Modelo 3.8 como la del Modelo 3.9 rechazan H_0: $\beta = 0$

El coeficiente de correlación disminuye en 3 milésimas, sin afectarlo de manera determinante.

Cancelando la columna 4 (X_3) tenemos el siguiente modelo.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_5 X_5 + \beta_9 X_9 + \varepsilon \qquad (3.10)$$

Bajo los supuestos $E(\varepsilon) = 0$; $Var(\varepsilon) = \sigma^2 In$ y $Cov(\varepsilon i, \varepsilon j) = 0 \; \forall \; i, \; j \quad i \neq j$.

Comparaciones entre los estimadores de los Modelos (3.9) y (3.10)

	Modelo 3.9		**Modelo 3.10**	
	Estimación de Betas	Varianzas	Estimación de Betas	Varianzas
β_0	-2121.23	1218394.59	-2081.96	1182542.17
β_1	26.10	218.16	26.30	212.48
β_2	-1.89	16.46	-3.33	9.57
β_5	40.68	576.13	47.46	418.07
β_9	27.41	38.86	27.51	37.63

Directamente observamos que la varianza $\hat{\beta}_0$ del Modelo 3.10 difiere en más de 30,000 unidades respecto a la varianza de $\hat{\beta}_0$ del Modelo 3.9. Las varianzas de las demás Betas estimados disminuyen en menor grado, mientras que las estimaciones de los parámetros permanecen bastante estables.

Las tablas de análisis de varianza en los dos Modelos rechazan la hipótesis nula H_0: $\beta = 0$

El coeficiente de correlación del Modelo 3.10 es de .931 mientras que el del Modelo 3.9 es de .931, por

lo cual se tiene aproximadamente una perdida de una milésima. Sin embargo tenemos que la varianza estimada disminuye de 2351 a 2291.38 obteniéndose una ganancia de 59.88 unidades.

Pasamos por último a cancelar la columna $3(X_2)$ y tenemos el siguiente Modelo.

$$Y = \beta_0 + \beta_1 X_1 + \beta_5 X_5 + \beta_9 X_9 + \varepsilon \qquad (3.5)$$

Bajo los supuestos $E(\varepsilon) = 0$; $Var(\varepsilon) = \sigma^2 In$

$$Cov(\varepsilon i, \varepsilon j) = 0 \quad \forall \; i, \; j \quad i \neq j.$$

NOTA: Este modelo es el mismo, que encontramos después de haber realizado la selección de variables, es por eso que tiene el número 3.5.

Comparaciones entre los estimadores de los modelos (3.10) y (3.5)

	Modelo 3.10		Modelo 3.5	
	Estimación de Betas	Varianzas	Estimación de Betas	Varianzas
β_0	-2081.96	1182542.17	-1944.71	1172942.57
β_1	26.30	212.48	24.10	209.51
β_5	47.46	418.07	40.50	378.53
β_9	27.51	37.63	27.34	38.02

Todas las varianzas de los estimadores de los parámetros disminuyen pero también se reducen las estimaciones de las Betas, por lo cual consideramos que existe pequeña transformación en los valores.

La varianza estimada aumenta en más de 10 unidades siendo menos de .4% del total de varianza estimada.

Las tablas de análisis de varianza finalmente siguen rechazando la hipótesis nula $H_0: \beta = 0$.

En lo que respecta al coeficiente de correlación al cuadrado se obtuvo una disminución de aproximadamente 3 milésimas, el cual consideramos pequeño.

Lo más interesante que se obtiene es que la condición del número $K = (\lambda g / \lambda c)^{1/2} = 9.8535$*, es decir ahora contamos con una Multicolinealidad modesta. Esto sugiere que se puedan eliminar ese número de variables $(X_2, X_3, X_4, X_6, X_7 \text{ y } X_8)$ sin perder mucha información.

Con el modelo 3.5 tenemos que las variables X_1, X_5 y X_9 cuentan con una dependencia lineal bastante pequeña.

Después de haber realizado el análisis y la selección de variables, podemos afirmar con certeza que la matriz "Z" que a continuación se expone proporciona

* Este número representa multicolinealidad modesta por los estudios realizados por Stewart (1974).

información relevante acerca de la variable respuesta "Y".

MATRIZ Z

X_0	X_1	X_5	X_9
1	77.25	0.000	8.02
1	77.50	1.000	8.12
1	77.83	0.000	11.35
1	78.17	0.000	11.45
1	78.83	1.000	11.45
1	79.20	0.000	11.78
1	79.63	0.000	12.41
1	79.95	1.000	12.42
1	80.23	0.000	12.81
1	80.54	0.000	12.73
1	80.58	0.000	13.30
1	80.92	0.000	13.16
1	81.00	0.000	13.60
1	81.21	0.000	14.20
1	81.34	0.000	15.30
1	81.50	0.000	15.34
1	81.60	0.000	15.42
1	81.67	1.000	17.50
1	81.72	0.000	16.14
1	81.79	0.000	18.19
1	81.97	0.000	18.28

1	82.03	0.000	19.70
1	8232	0.000	19.15
1	82.40	1.000	20.18
1	82.47	0.000	20.30
1	82.32	1.000	21.12
1	82.43	0.000	21.42
1	82.63	1.000	21.20
1	83.00	1.000	21.30
1	83.25	1.000	21.40
1	83.53	1.000	21.70
1	83.70	1.000	21.90

Por lo cual el Modelo 3.5 se puede expresar como:

$$Y = Z\beta + \varepsilon \qquad (3.11)$$

$$E(\varepsilon) = 0 \;\; ; \;\; Var(\varepsilon) = \sigma^2 I_n \;\; y \;\; Cov(\varepsilon_i, \varepsilon_j) = 0 \;\; \forall \; i,j \;\; i \neq j$$

Ahora bajo el modelo 3.5 pasamos a analizar la prueba de carencia de ajuste, la cual es la siguiente:

F.V.	G.L.	S.C.	C.M.	F	P
Error	28	64524.221	2304.436		
Carencia A	15	18474.998	1231.667	0.347	1.000
Error P	13	46049.223	3542.248		
Total	32	7713535.078			

Los resultados que arrojan lo anterior quieren decir, que no hay argumentos suficientes para rechazar H_0: no hay carencia de ajuste, puesto que el estadístico de prueba "F" no es mayor que la distribución $F^{(1-K)}_{(15,32)}$, por lo tanto no se rechaza la hipótesis nula con probabilidad 1 de que $F_{tab} > F_{cal}$

Ahora estamos en condiciones de proseguir el análisis sin ningún problema. Debido a que se tienen argumentos para decir que el ajuste es bastante bueno, se pueden llevar a cabo tanto intervalos de confianza como pruebas de hipótesis.

Pasamos a analizar los residuales studentizados contra Z, columna dos teniendo una banda de confiabilidad de 95%. La gráfica No. 15 tiene un diagrama nulo a excepción de la observación No. 32 la cual se dispara bastante del conjunto de observaciones; además puede verse que hay indicios de existir una varianza no constante. Como remedio para corregir la heterogeneidad de varianza se cuenta tanto con la técnica de mínimos cuadrados ponderados asi como el uso de transformaciones.

En la Gráfica No. 16 (Residuales estudentizados contra Z, columna 4) se observa una tendencia periódica dentro de la banda de confianza, existe un punto fuera de ella bastante alejado de los demás (Residual estudentizado No. 32). Por lo cual tenemos la sospecha de una observación discrepante. Pudiéndose corregir la varianza no constante y la posible existencia del punto discrepante por medio de una tranformación.

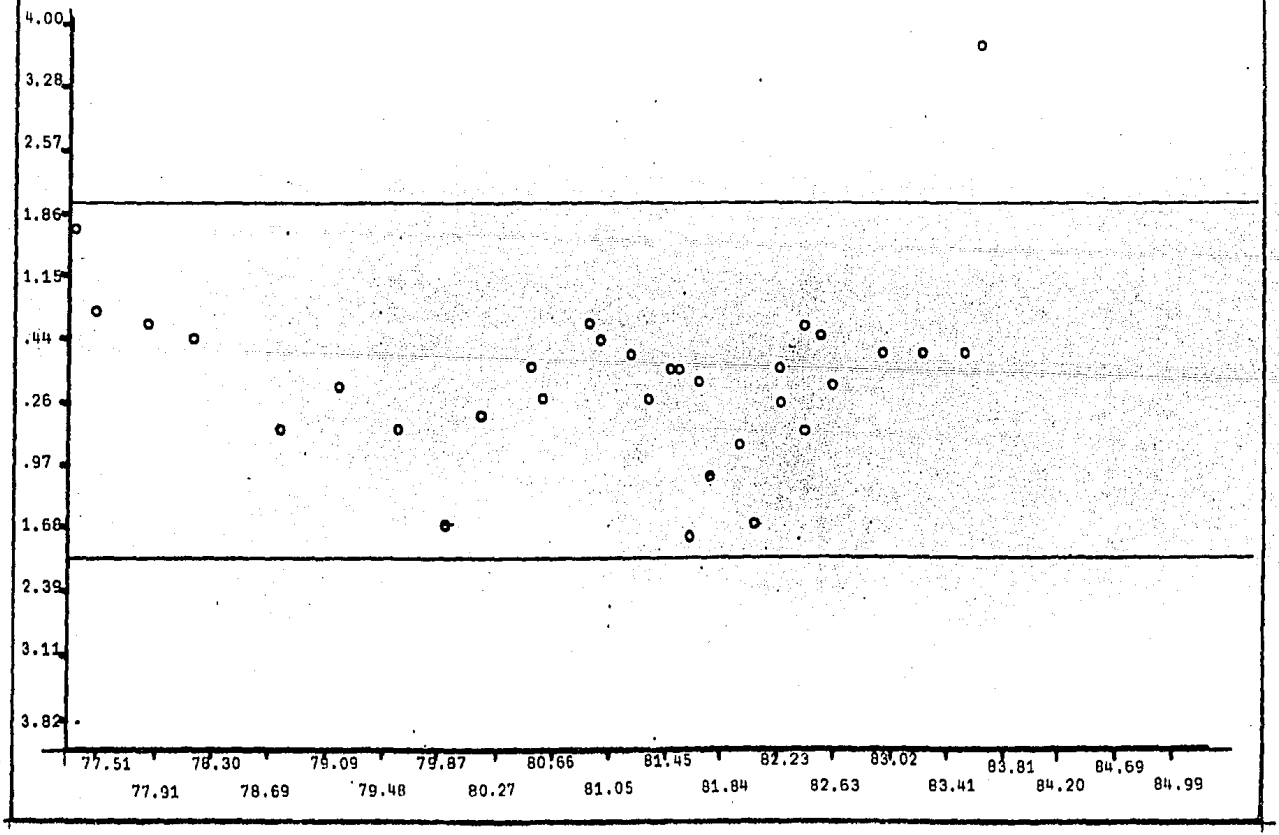
La Gráfica No. 17, es el diagrama más importante dentro de los diagnósticos individuales, en ella observamos una no lineabilidad y una varianza no constante, pudiéndose corregir la primera mediante transformaciones. En este diagrama se muestran tendencias semejantes a las de las gráficas 15 y 16 por lo cual se sospecha que las causas y correcciones sean las mismas.

GRAFICA Nº 15

RESIDUALES . ESTUDIANTIZADOS CONTRA Z

COLUMNA 2

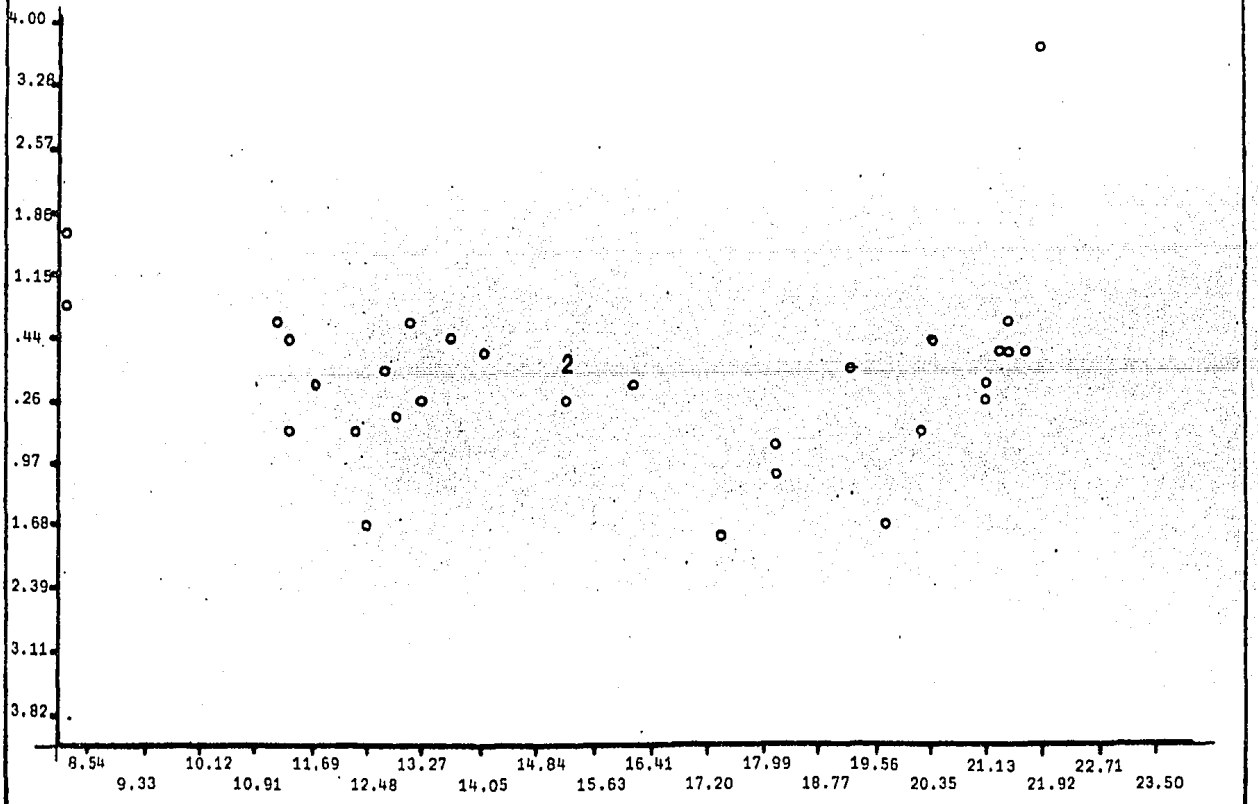
GM 86C



GRAFICA Nº 16

RESIDUALES ESTUDIANTIZADOS CONTRA Z

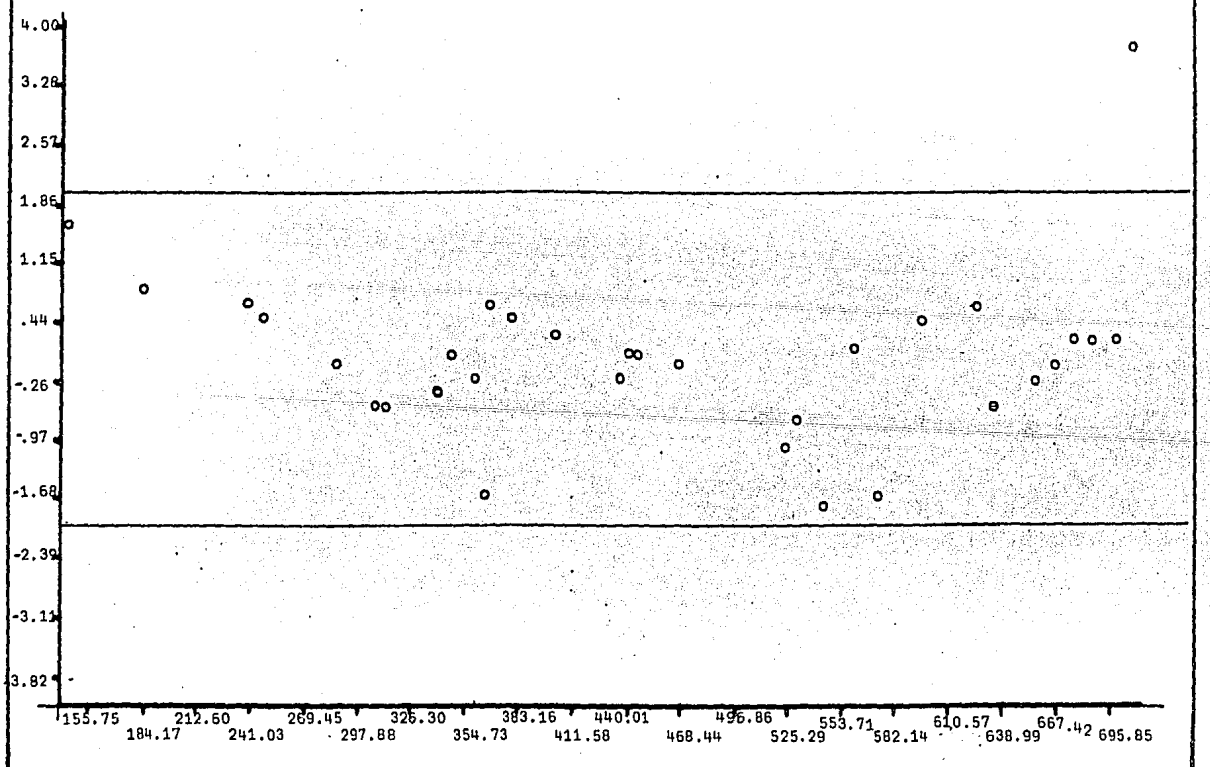
COLUMNA 4



GRAFICA Nº 17

RESIDUALES ESTUDIANTIZADOS CONTRA Y ESTIMADA

GM 86C



v) **Residuales Parciales** (*)

Antes de llevar a cabo el análisis de residuales parciales en nuestro problema, exponemos su utilidad y el por qué de esta técnica.

Los residuales parciales son una útil adición a las técnicas de post-ajuste.

Def: $r_{pi} = y_i - (\hat{y}_i - \hat{B}_j x_{ij}) = r_i + \hat{B}_j x_{ij}$ (*)

donde $\hat{y}_i - \hat{B}_j x_{ij}$ predice la i-ésimo respuesta usando todas las variables explicativas excepto xj.

Gráficar r_p contra xj permite examinar la relación entre y y xj habiendo tomado en cuenta los efectos de las restantes variables explicativas.

La pendiente de la gráfica es $\hat{B}_j$ y sus residuales (con regresión por el orígen) son dados por r.

Este método proporciona resultados de suma importancia, puesto que muestra la tendencia sistemática relativa a X_j ajustada por las otras variables, pero la dispersión es la relativa a X_j sin ajustar.

Los residuales parciales indican grado y direc-

---

* Larsen W.A. and S.A. McCleary (1972)

ción de la lineabilidad, permiten además establecer la importancia de cada variable explicativa en presencia de las otras variables explicativas y evaluar el grado de no lineabilidad que exista en la relación. También puede ayudar en la selección de transformaciones.

Para llevar a cabo el análisis de las gráficas 18 y 19 se consultó a JASA, 19, 431-455, Technometrics, 14, 781-790 y Larsen & Mc. Cleary.

En la Gráfica No. 18 Residuales parciales contra la matriz Z, columna dos (X_1), al parecer es necesaria una transformación, posiblemente exponencial. También se puede pensar en una observación discrepante.

La Gráfica No. 19 residuales parciales contra la matriz Z, columna cuatro, se ve una fuerte correlación entre r_p y X_9. No hay indicios de no lineabilidad, pero se sospecha que existe una observación discrepante.

vi) **Distribución de los errores**

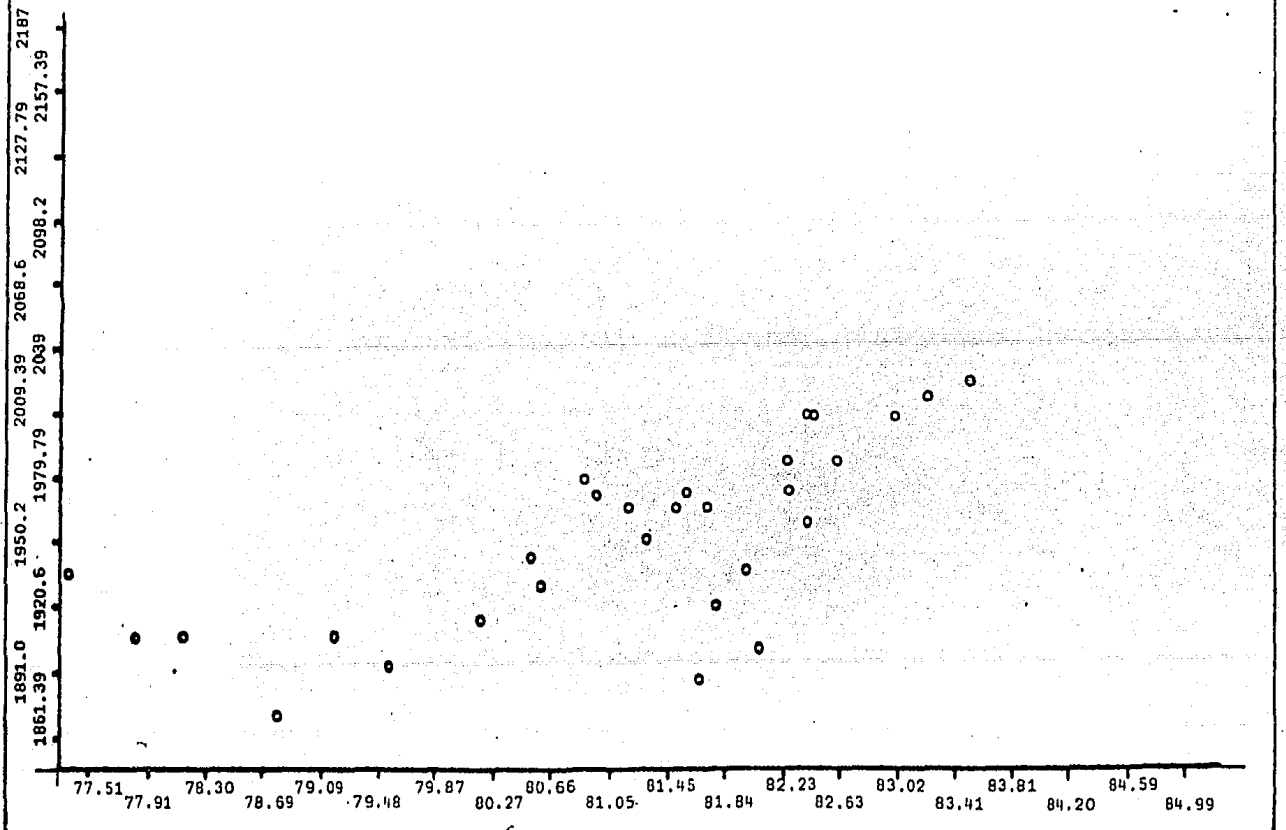
Es común que se asuma que los errores son distribuidos normalmente, y en caso de no cumplirse, constituya una falla usual, puesto que el problema de no normalidad de errores es muy difícil de diagnosticar por análisis de residuales.

GRAFICA Nº 18

RESIDUALES PARCIALES CONTRA LA MATRIZ Z

COLUMNA 2.

GM 86C

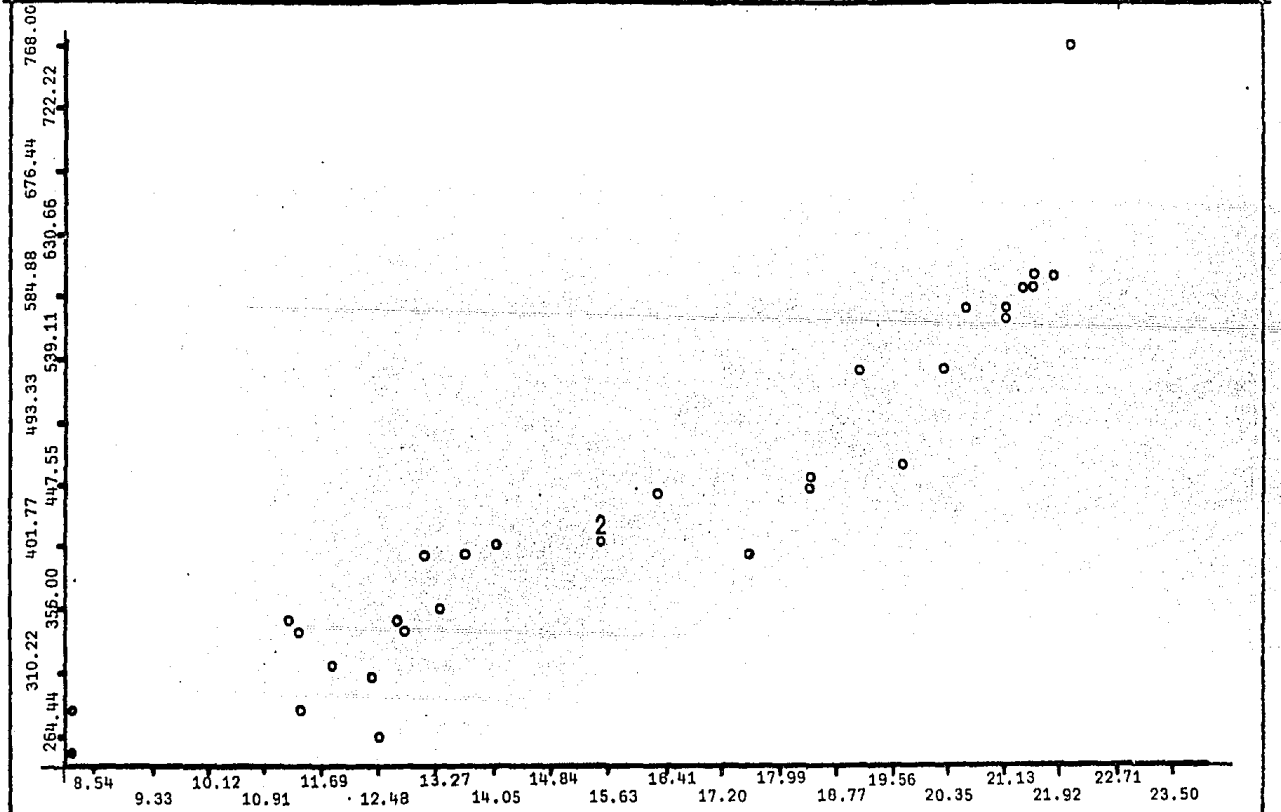


GRAFICA Nº 19

RESIDUALES PARCIALES CONTRA LA MATRIZ Z

COLUMNA 4

GM 86C



Una técnica que se usa actualmente es la graficación de residuales estudentizados en papel normal, que nos dice que tan errados estamos en esta suposición.

En la gráfica No.20 de residuales estudentizados en pâpel normal, se observa directamente que los puntos del diagrama se apegan bastante a una recta, por lo cual la hipótesis de Normalidad se puede tomar, sin caer en un error de peso.

vii) **Correlación de errores**

Otra de las suposiciones del modelo es la no correlación de los errores. Para saber si están o no correlacionados, contamos con la técnica de Durbin-Watson que fue originalmente propuesta por Von Neumann en otro contexto.

La prueba detecta solo correlación de primer orden, pues se basa en el supuesto de que los errores constituyen una serie autoregresiva de 1er. orden:

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \qquad |\rho| < 1$$

donde $u_t \sim N(0, \theta^2) \forall\ t$

Esta suposición es solo una aproximación a la distribución real de los errores.

El objeto es probar $H_0: \rho_\delta = 0$ $H_1: \rho_\delta = \rho^\delta$ y el estadístico de prueba es:

$$d = \sum_{t=2}^{n} (r_t - r_{t-1})^2 \Big/ \sum_{t=1}^{n} r_t^2$$

Para nuestro poblema tenemos que $d = 1.094$, como d^α_L es mayor que d entonces d es significante a un nivel $\alpha = .05$, por lo tanto se cree que exista correlación positiva.

NOTA: d^α_L fue tomado de tablas D y W Biometrica 1971 $d^\alpha_L = 1.18$

Otro método de suma importancia es la prueba de Rachas Wald-Wolfowitz test , que es un procedimiento no paramétrico y su idea básica es: si los errores no están correlacionados esto ha de reflejarse en los residuales y así, si registramos los signos de estos últimos podemos examinar las rachas obtenidas para detectar algún patrón anómalo (muy pocas o muchas).

La distribución exacta del número de rachas puede encontrarse de la misma manera que para las estadísticas basadas en rangos.

Si se tienen n_1 signos positivos y n_2 negativos al número de formas en que pueden arreglarse es

$(n_{1}{}_{n_2} + n_2)$, pues al seleccionar n_2 lugares para los elementos de signo negativo, los otros deben asignarse a los de signo positivo

Así, solo resta calcular el número de maneras en que pueden obtenerse digamos rachas y efectuar el cociente sobre el total para obtener la probabilidad correspondiente.

Para nuestro problema tenemos la siguiente sucesión de signos de los residuos.

+ + + + - - - - - - + - + + + - + + - - - - - + - + - + - + + + +

Por lo cual contamos con 17 residuos + y 15 residuos -, se tienen 15 rachas bajo este modelo.

Al consultar las tablas encontramos que no hay valores tabulados par $n_1 = 17$ y $n_2 = 15$ por lo cual requerimos de aproximar mediante una normal con parámetros μ y σ^2 a fin de aplicar esta técnica, sabiendo que

$$\mu = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

$$\sigma^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

$$\mu = \frac{2(17)(15)}{17 + 15} + 1 = 16.9375$$

$$\sigma^2 = \frac{2(17)(15)\{2(17)(15) - 17 - 15\}}{(17+15)^2 (17+15-1)} = \frac{243780}{31744} = 7.6795$$

Como:

$$u \sim N(\mu, \sigma^2)$$

$$\rightarrow \quad u \sim N(16.94,\ 7.68)$$

$$\rightarrow \quad \frac{u - 16.94}{7.68} \sim N(0,1)$$

$$\rightarrow \quad F(X) = p(X \quad 15) = \Phi\ \left(\frac{15 - 16.94 + .5}{7.68}\right)$$

$$= \Phi\ (-.521) = 1 - \Phi\ (.521) = 1 - .5985$$

$$= .4015 = 40.15\%$$

Un evento no raro ocurre en un 40.15% de los casos, por lo cual se rechaza que haya pocas rachas en los datos.

Ahora consideramos el otro enfoque con las siguientes hipótesis H_0: No $\exists$ un número demasido grande de rachas H_1: $\exists$ un número demasido grande de rachas.

$$\rightarrow \quad p(X \geq 15) = 1 - p(X \leq 15) = 1 - .4015 = .5985$$

→ Si rechazo H_0 tengo una probabilidad de error de 59.85%, si determino un nivel de significancia de .05 → $\therefore$ se rechaza que haya muchas rachas en los datos.

Conclusión

No se rechaza aleatoriedad de los residuos, por lo que no hay indicio de correlación serial, de acuerdo a esta prueba.

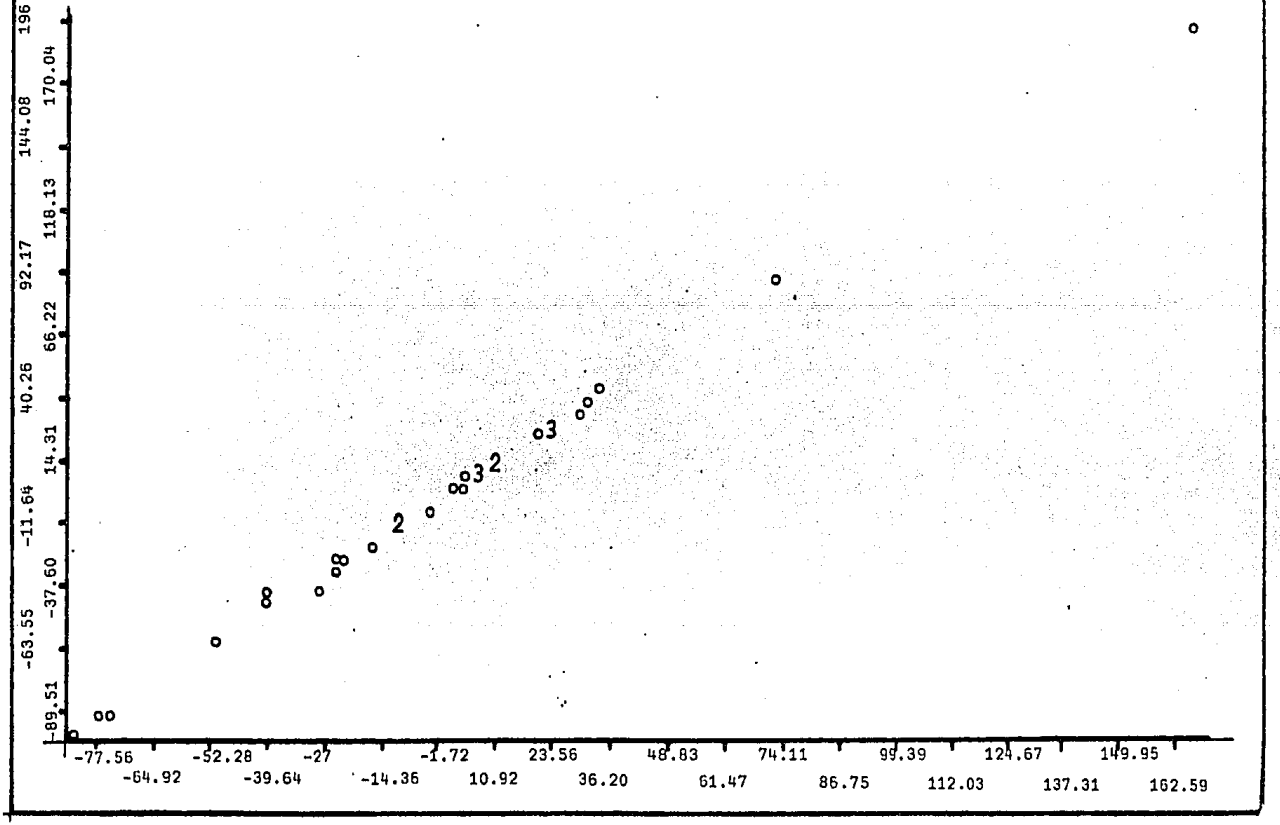
Pasamos a analizar los puntos discrepantes e incluyentes mediante la gráfica de residuales menos contra residuales (gráfica No. 21), como se puede observar tenemos al menos un punto que discrepa bastante de los demás dentro del enfoque geométrico. Pero además contamos con una técnica analística para observar puntos discrepantes.

viii) **Observaciones Discrepantes e Influyentes del Modelo**

En bastantes problemas se tienen casos individuales que tienen impacto sobre la regresión, los cuales pueden ser verdaderos o falsos. Muchas veces debido a un descuido o a una mala medición, los valores de las variables están bastante alejados de lo real, por lo tanto requerimos analizar esas observaciones que se disparan del conjunto de datos. Es importante examinar en detalle tales casos pues pueden brindar gran información no sólo del ajuste, sino también de la situación experimental ya que puede ser indicio de la necesidad de un modelo más complejo. En un principio podemos considerar cuatro diferentes situaciones:

GRAFICA Nº 21 RESIDUALES MENOS CONTRA RESIDUALES

GM 86C



a) Una observación improbable pero perfectamente conformable fue realizada

b) La observación se debe a fallas del instrumento de medición

c) Se ha observado un evento excepcional

d) El caso es perfectamente legítimo y posiblemente de los más importantes en estudio, que puede representar información nueva e inesperada.

En b), c)y d) la estimación se debe hacer excluyendo los casos identificados como observaciones discrepantes (outliers), ya que una estimación insesgada requiere su identificación y eliminación en la estimación.

Los candidatos para ser puntos discrepantes son aquellos con valores extremos de residuales estudentizados.

Necesitamos recordar que la identificación de un caso discrepante es relativa al modelo lineal específico. Si la forma del modelo es modificada, los casos individuales catalogados como discrepantes pueden cambiar. El contraste de hipótesis es $H_0: E(y_i - \tilde{y}_i) = 0$ vs. $H: E(y_i - \tilde{y}_i) \neq 0$ y la estadística de prueba de casos discrepantes está dada por:

$$t_i = \frac{y_i - \tilde{y}_i}{\hat{\sigma}_{-i}\,(1+z_i^1\,(Z_{-i}^1 Z_{-i})^{-1}\,z_i)^{1/2}} \sim t_{n-p^1-1}$$

donde Z_{-i} es la matriz Z sin el renglon i ; $\hat{\beta}_{-i}$, $\hat{\sigma}_{-i}$ son estimaciones de β y σ sin utilizar el i-ésimo caso; $\tilde{y}_i = z^1_i\,\hat{\beta}_{-i}$ es decir la predicción de y_i, sin considerar z_i.

Bajo el enfoque computacional la expresión de t_i se puede calcular como (*)

$$t_i = e_i \left(\frac{n-p^1-1}{n-p^1-e_i^2}\right)^{1/2} \qquad (3.11)$$

En nuestro análisis encontramos que las observaciones 1 y 32 tienen el valor de t_i 1.695599 y 5.341531 respectivamente siendo estos dos candidatos a ser observaciones descrepantes (ver tabla 3.1), ahora realizaremos las pruebas relevantes para tener evidencia de si son o no puntos discrepantes.

Tenemos que t_1= 1.695599, este estadístico tiene 32-4-1=27 grados de libertad, nosotros encontramos el valor crítico para una α =.01 es decir $t_{27}(.01) = 2.77$, claramente t_1 no excede este valor, por lo cual no tenemos evidencia de que t_1 sea una observación discrepante.

---

* Demostración en Apendice I

T A B L A 3.1

VECTOR Y_T	VECTOR Y-ESTIMADA	VECTOR RESIDUAL	RESIDUALES STUDENTIZADOS	D(COOK)	T(OUTLIER)
207.510000	136.796735	70.713265	1.641526	0.162902	1.695599
217.380000	186.066570	31.313430	0.778145	0.064042	0.772522
270.710000	241.844315	28.865685	0.705569	0.046899	0.699098
272.370000	252.775075	19.594925	0.456250	0.012976	0.449704
280.360000	309.193654	-28.833654	-0.665380	0.025144	-0.658618
284.880000	286.628828	-1.748828	-0.037888	0.000029	-0.037207
288.480000	314.223132	-25.743132	-0.553654	0.005050	-0.546678
289.660000	362.719141	-73.059141	-1.676510	0.149997	-1.735721
317.210000	339.625606	-22.415606	-0.483499	0.004217	-0.476781
345.390000	344.910693	0.479307	0.010522	0.000003	0.010332
350.630000	361.462824	-10.832824	-0.234648	0.001118	-0.230646
394.360000	365.830261	28.529739	0.635595	0.014518	0.628694
402.590000	379.791499	22.798501	0.502934	0.007679	0.496118
412.180000	401.262066	10.917934	0.239947	0.001627	0.235866
423.320000	434.477741	-11.157741	-0.240573	0.001031	-0.236482
443.220000	439.428589	3.791411	0.082359	0.000148	0.080884
452.050000	444.026963	8.023037	0.175039	0.000742	0.171979
457.120000	543.105126	-85.985126	-1.882007	0.092071	-1.977389
460.050000	466.609651	-6.559651	-0.141930	0.000397	-0.139422
473.640000	524.358783	-50.718783	-1.098327	0.024326	-1.102547
490.880000	531.159110	-40.279110	-0.870637	0.014528	-0.866762
495.580000	571.438460	-75.858460	-1.691851	0.104658	-1.753403
567.790000	563.388255	4.401745	0.095833	0.000212	0.094122
608.800000	633.992943	-25.192943	-0.554082	0.008803	-0.547105
621.450000	598.453408	22.996592	0.511048	0.009013	0.504196
642.230000	657.770808	-15.540808	-0.350219	0.005222	0.344664
652.320000	628.118005	24.201995	0.565864	0.020793	0.558872
665.990000	667.431444	-1.441444	-0.032049	0.000036	-0.031472
690.190000	679.085383	11.104617	0.245591	0.001917	0.241426
697.140000	687.846602	9.293398	0.206102	0.001417	0.202542
712.270000	702.800437	9.469563	0.211501	0.001673	0.207856
881.240000	712.367893	168.872107	3.792916	0.584464	5.341531

Sin embargo t_{32} es evidente que sobrepasa a $t_{27}(.01)$ por lo cual bajo este nivel de significancia t_{32} es una observación discrepante. (Ver tabla 3.1).

Como se expuso en el capitulo II entre lo que se puede hacer para obtener "mejores resultados" se encuentra el utilizar una transformación de la variable respuesta (y) en vez de la orginalmente usada en el análisis.

(x) Transformaciones Box y Cox.

Por medio de los diagramas de dispersión de los residuales estudentizados contra "y" estimada es posible detectar la transformación conveniente a la variable respuesta; Sin embargo, existen métodos más objetivos para determinar la transformación apropiada.

Si se detecta que se debe tranformar a la variable respuesta y y si sabemos que solo puede tomar valores positivos es conveniente ajustar al modelo (*)

$$y^{(\lambda)} = Z\beta + \varepsilon$$

Donde

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ Ln(y) & \lambda = 0 \end{cases}$$

(*) Box y Cox 1964.

Un gran número de transformaciones que usualmente se utilizan están incluidas en esta familia; por ejemplo, para $\lambda = 1/3$ corresponde a la transformación raiz cúbica. Cuando $\lambda = 0$ la transformación logaritmo natural será la adecuada debido a que $y^{(\lambda)} = \ell n y$

Como la λ apropiada es desconocida, podemos estimarla al mismo tiempo que β y σ^2. Box y Cox nos sugieren utilizar el estimador máximo verosímil $(\hat{\beta}, \hat{\sigma}^2, \hat{\lambda})$ bajo la suposición de distribución de los errores. En particular si suponemos normalidad en los errores podemos calcular, para un rango razonable de valores de λ, el valor del logaritmo de la verosimilitud dado por la siguiente expresión:

$$L(\lambda) = \frac{n}{2} \ell n(\lambda^2) - \frac{n}{2} SCR_\lambda + (\lambda - 1) \sum_1^n \ell n y_i$$

donde $$SCR_\lambda = (\underline{y}^\lambda)' (I - H) \underline{y}^\lambda$$

Generalmente en la práctica no es necesario conocer el valor exacto de $\hat{\lambda}$, ya que es posible que tenga mayor sentido práctico el utilizar un valor más simple para simplificar la interpretación, por decir si $\hat{\lambda} = .9983$, es más conveniente utilizar $\hat{\lambda} = 1$.

La función $L(\lambda)$ depende, únicamente de λ y entonces como método aproximado, justificado por el párrafo anterior, se puede graficar para un conjunto de valores de es-

te parámetro para obtener una aproximación a $\hat{\lambda}$. Este ciclo se puede repetir cerca del máximo tantas veces como sea necesario para lograr un cierto grado de aproximación.

Para nuestro problema, tenemos que los valores para lambda y L Max son:

LAMBDA	L MAX
-0.60	-101.999
-0.50	-101.714
-0.40	-101.852
-0.30	-102.385
-0.20	-103.266
-0.10	-104.437
-0.00	-105.836
0.10	-107.406
0.20	-109.099
0.30	-110.874
0.40	-112.701
0.50	-114.558

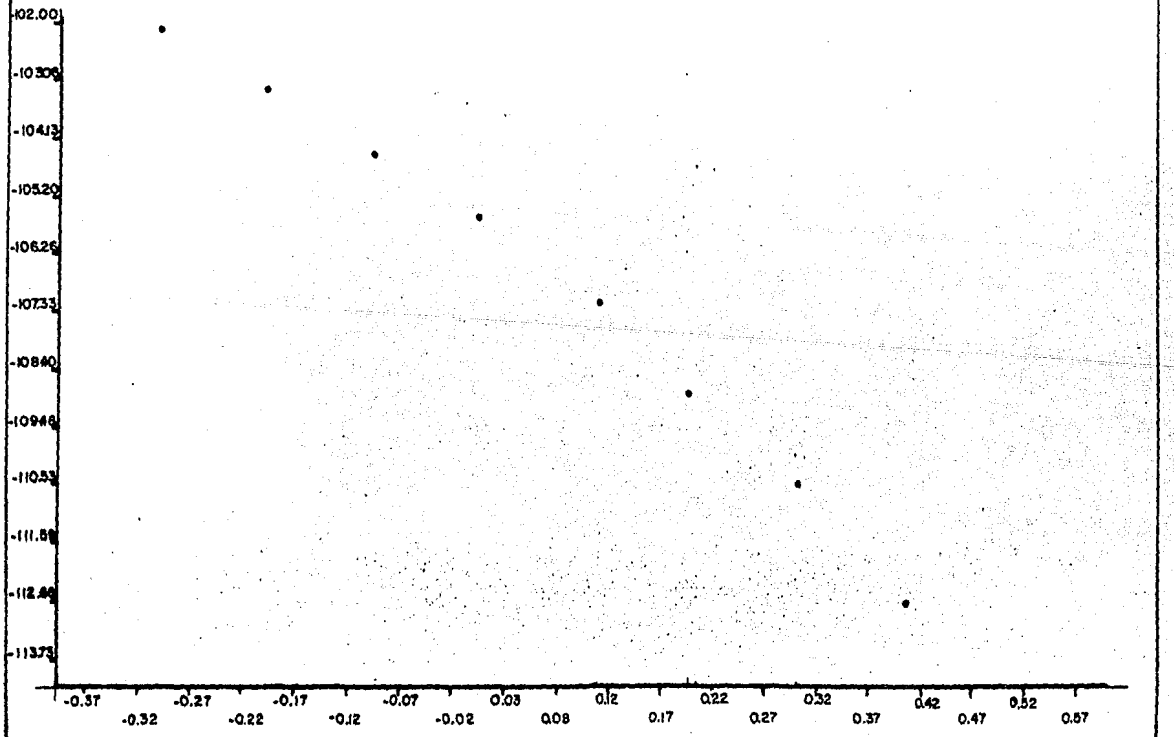
Como se puede observar en la tabla de arriba el valor máximo está cercano a $\lambda = -.50$, por lo cual la transformación que se sugiere para la variable respuesta y es:

$$(y^{\lambda}) = \frac{y^{-.5} - 1}{-0.5} \quad \text{para} \quad \lambda = -0.5 \qquad (3.12)$$

GRAFICA Nº22

FUNCION DE VEROSIMILITUD (BOXCOX)

GM 86C



Dentro del modelo 3.5 checaremos por último la distancia de Cook, puesto que se tienen indicios de que la observación 32 se dispara bastante del conjunto de datos. (Ver tabla 3.1).

En la tabla nombrada arriba se observa que efectivamente la observación No. 32 tiene el máximo valor en distancia de Cook (.584464), si este caso fuera omitido del análisis el estimador del vector beta sufriría un movimiento equivalente a trasladar el estimador inicial a un elipsoide de 32.35% de confianza.

Desde el capitulo II se tenian indicos que esta observación tiene influencia sobre el modelo, pero eliminarla, siendo verdadera puede alterar substancialmente las conclusiones, por lo cual consideramos mantenerla en el conjunto de datos y realizar la transformación 3.12, dandole importancia a esta observación en el siguiente desarrollo.

La expresión más indicada, desde el punto de vista computacional es la siguiente (esta expresión es equivalente a la presentada en 3.12).

$$yT = (2 - (2/(y^{**}-0.5))) \qquad (3.13)$$

El Modelo es $"yT" = Z\beta + \varepsilon$ (3.14)

bajo $E(\varepsilon)=0$: $Var(\varepsilon)=\sigma^2 I$ y $Cov(\varepsilon_i, \varepsilon_j)=0 \ \forall \ i, \ j \ \ i \neq j$

x) **Transformación del Modelo**

Hasta el momento, hemos realizado una transformación a la variable respuesta "y" (3.13) sugerida al aplicar la técnica de Box y Cox.

Lo siguiente a efectuar son los diagramas de dispersión de las variables explicativas contra la variable respuesta yT ("y" transformada).

La importancia de realizar estos diagramas, es que se pueden observar sus tendencia y posibles inadecuaciones.

Por lo tanto pasamos a graficar la variable "yT" contra Z, columna 2, (X_1) y se observa directamente que los valores se comportan de una manera casi lineal, sin haber puntos que se disparen de los demás, lo cual es un comportamiento favorable. Ahora en la gráfica No. 24 de "yT" contra Z, columna 4, (X_3) se visualizan dos puntos fuera del conjunto de datos, a los cuales tendremos cuidado especial dentro del análisis del problema.

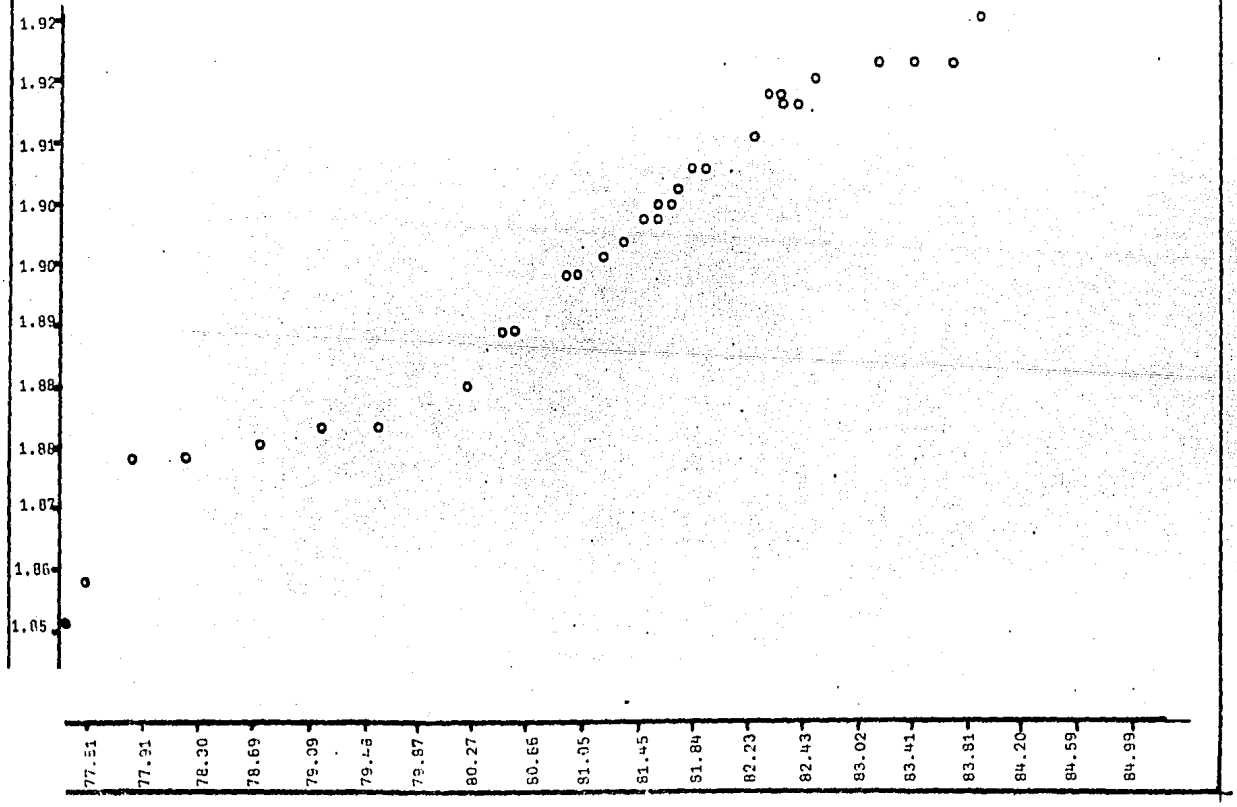
Aplicamos a continuación la técnica de mínimos cuadrados, con la variable respuesta ya transformada bajo las siguientes suposiciones:

$E(\varepsilon)=0$, $Var(\varepsilon)=\sigma^2 In$ $Cov(\varepsilon i, \varepsilon j)=0 \ \forall \ i, \ j, \ i \neq j$. Los resultados obtenidos son los siguientes:

GRAFICA Nº 23

Y CONTRA Z COLUMNA 2

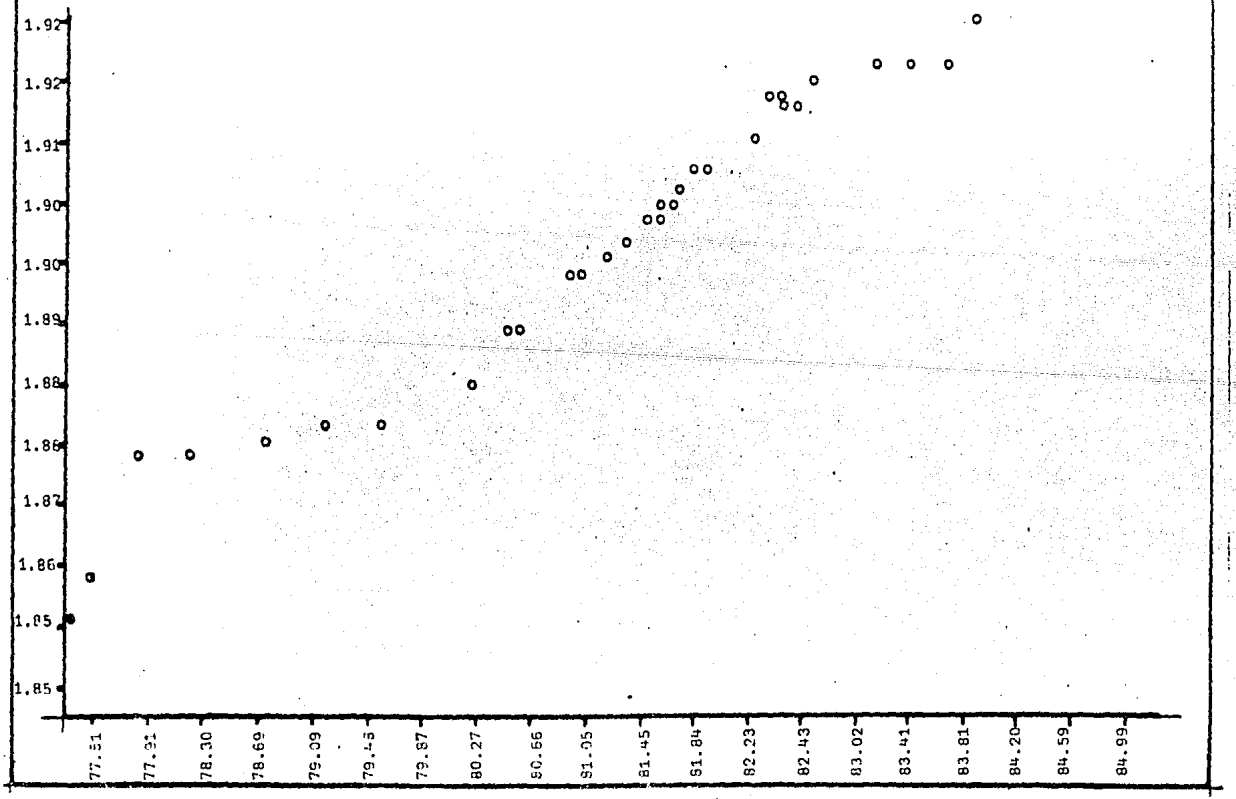
GM 86C



GRAFICA Nº 23

Y CONTRA Z COLUMNA 2

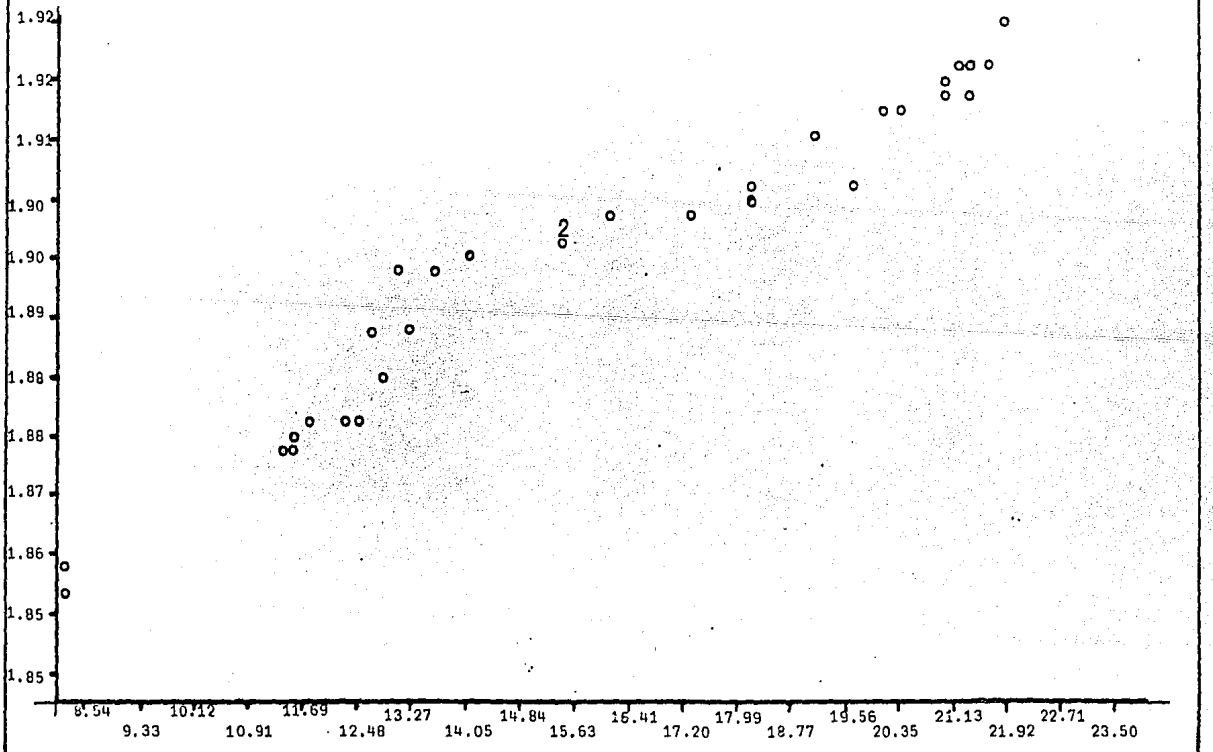
GM 86C



GRAFICA Nº 24

Y CONTRA Z COLUMNA 4

GM 86C



$$\hat{\underline{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_5 \\ \hat{\beta}_9 \end{bmatrix} = \begin{bmatrix} 1.3931 \\ 0.0059 \\ 0.0008 \\ 0.0020 \end{bmatrix}$$

$$\rightarrow \quad \hat{Y}T = Z\hat{\beta} \qquad (3.15)$$

Las varianzas correspondientes a los betas "gorros" son las siguientes:

$$\hat{Var}\,(\hat{\beta}_0) = 0.0037$$

$$\hat{Var}\,(\hat{\beta}_1) = 6.5281 \times 10^{-7}$$

$$\hat{Var}\,(\hat{\beta}_5) = 1.1794 \times 10^{-6}$$

$$\hat{Var}\,(\hat{\beta}_9) = 1.1849 \times 10^{-7}$$

Como se expuso en el capítulo anterior estos calculos son de suma importancia para contrastar modelos y visualizar ganancias o pérdidas al realizar transformaciones. Se cuenta además con el estadístico $\hat{\sigma}^2$ que nos sirve también para comparar modelos, (ver punto 2.6) obteniendo:

$$\hat{\sigma}^2 = 7.180 \times 10^{-6}$$

A continuación seguimos con la tabla de análisis de varianza, que nos proporciona información acerca de los parámetros que están expuestos en el modelo transfor

mado (3.14), lo que en sí realiza esta tabla es el contraste de H_0: $\beta_1 = \beta_5 = \beta_9 = 0$ contra H_1: al menos un $\beta_j \neq 0$ $j = 1, 5, 9$.

Tabla de Análisis de Varianza

Fuente	G.L.	S.C.	C.M.	F.	P
Regresión	3	0.011	0.004	494.877	0.0000
Residual	28	2.010^{-4}	7.180×10^{-6}		
Total	31	0.011			

Al tener una F calculada igual a 494.877 se ve claro que se rechaza H_0: $\beta_1 = \beta_5 = \beta_9 = 0$ por un margen bastante amplio, con probabilidad 0.0000 de que $F_{tab} > F_{cal.}$

El coeficiente de determinación en el modelo 3.14 es igual .981 es decir que el 98.1% de la variabilidad observada en Y_T es explicada por las variables independientes. El coeficiente de correlación múltiple es la raíz cuadrada del coeficiente de determinación, entonces tenemos que:

$$r = \sqrt{R^2} = \sqrt{.981} = .9904544$$

el cual es altamente satisfactorio.

A continuación, realizamos la prueba de carencia de ajuste, la cual es una de las técnicas más relevantes dentro del análisis de Regresión, puesto que si se tienen argumentos para rechazar H_0: No hay carencia de

ajuste, el modelo es no conveniente y además es obsoleta la tabla de análisis de varianza anterior, por otro lado las pruebas de hipótesis y los intervalos de confianza carecerían de importancia, dentro del análisis, si el modelo es pobre en ajuste.

Tabla de Análisis de Varianza

Fuente	G.L.	S.C.	C.M.	F	P
Residual	28	2.010×10^{-4}	7.180×10^{-6}		
Carencia de A	15	7.460×10^{-5}	4.974×10^{-6}	0.511	1.0000
Error Puro	13	1.264×10^{-4}	9.726×10^{-6}		

De acuerdo con la tabla de arriba la $F_{cal} = 0.511$ y la $F_{(.05,\ 15,\ 13)} = 2.53$, es decir $F_{cal} \leq F_{tab}$ entonces no se tienen argumentos suficientes para rechazar H_0: No hay carencia de ajuste. Es importante hacer notar que no se rechaza la hipótesis nula, con probabilidad 1 de que $F_{tab} > F_{cal}$.

Los cálculos obtenidos en la tabla de arriba nos dan el fundamento para llevar a cabo, tanto la construcción de intervalos de confianza como la realización de pruebas de hipótesis, además valida la tabla donde se contrasta $H_0: \beta_i = 0$ vs $H_1: \beta_i \neq 0$ $i = 1, 5, 9$.

Las gráficas de residuales estudentizados contra

Z se llevaran a cabo con dos confiabilidades, una al 90% de confianza y la otra al 95% de confianza, con el propósito de proporcionar más información.

La Gráfica No. 25: residuales estudentizados contra Z, columna 2 tiene una banda de confiabilidad del 95% este diagrama es nulo (no existe ningún residual fuera de la banda) sin embargo sospechamos que existe una pequeña heterogeneidad de varianza, puesto que los residuales estudentizados muestran un comportamiento periódico.

En la Gráfica No. 26: residuales estudentizados contra Z, columna 4, se tiene la misma confiabilidad que la gráfica anterior y se visualiza un comportamiento bastante semejante a la gráfica No. 25, por lo cual se puede afirmar que es un diagrama nulo, con sospecha de mínima variabilidad.

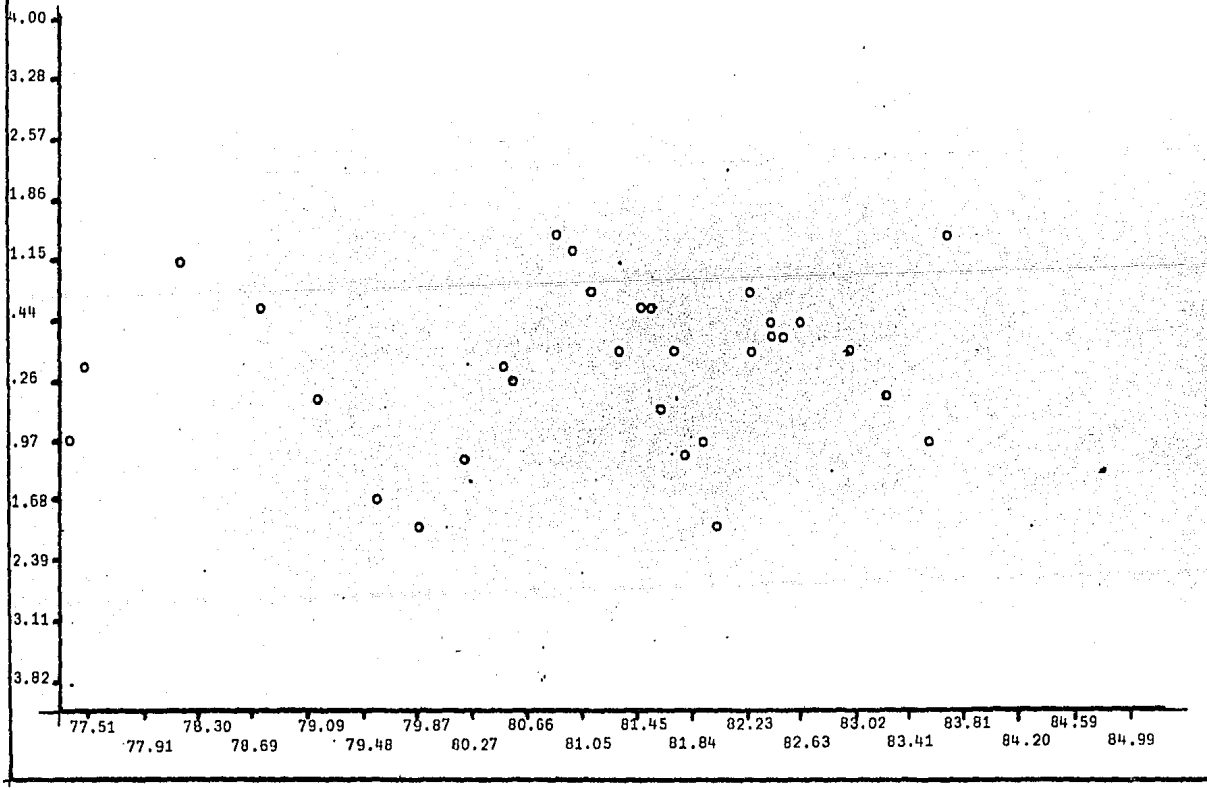
En la Gráfica No. 27: residuales estudentizados contra y estimada, como se ha dicho anteriormente, es el diagrama más importante y nuestro diagnóstico respecto a este esquema, es que es nulo su comportamiento, pensamos que a escalas crecientes su tendencia podría ser periódica con sospecha de nolineabilidad. Se muestra en este diagrama que más del 95% de los residuales caen en el rango (+ 2, -2), por lo que esto es un indicio de que la forma del modelo es correcta.

GRAFICA Nº25

RESIDUALES . ESTUDIANTIZADOS CONTRA Z

COLUMNA 2

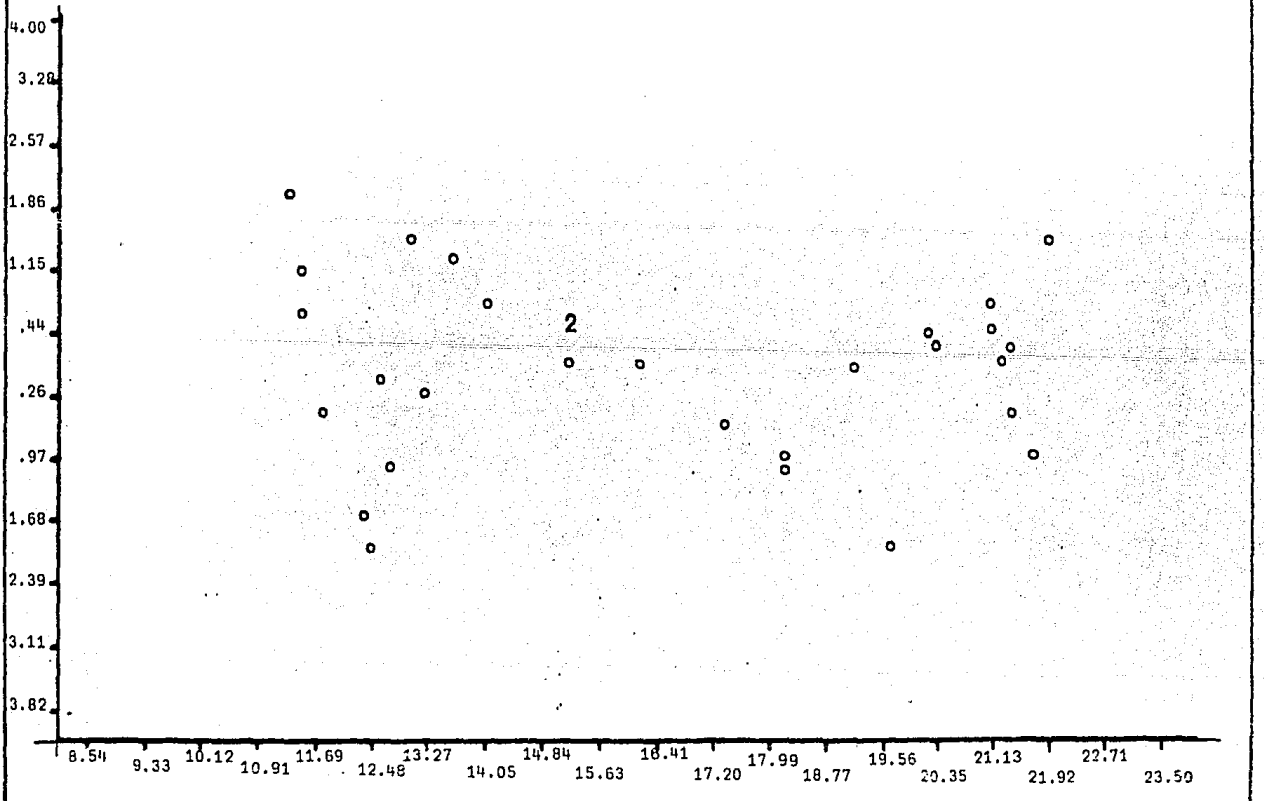
GM 86C



GRAFICA Nº 26

RESIDUALES . ESTUDIANTIZADOS CONTRA Z

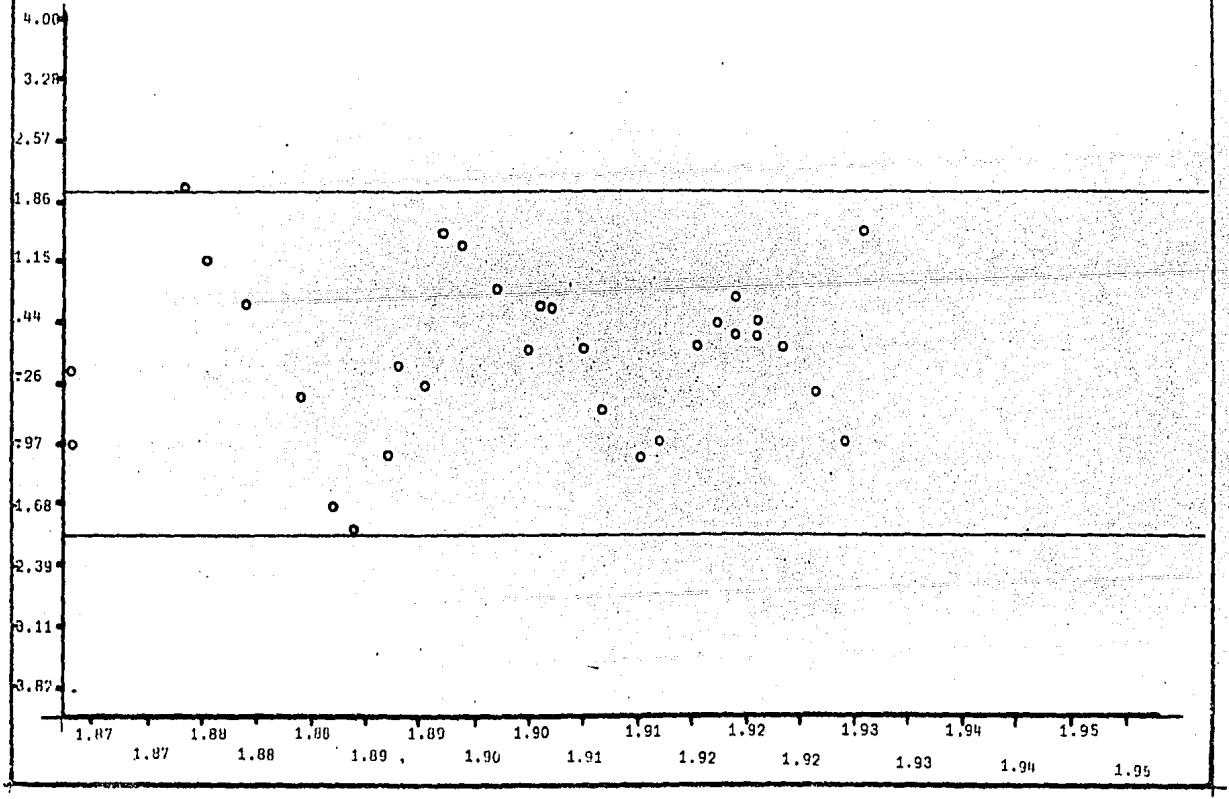
COLUMNA 4



GRAFICA Nº 27

RESIDUALES . ESTUDIANTIZADOS CONTRA Y ESTIMADA

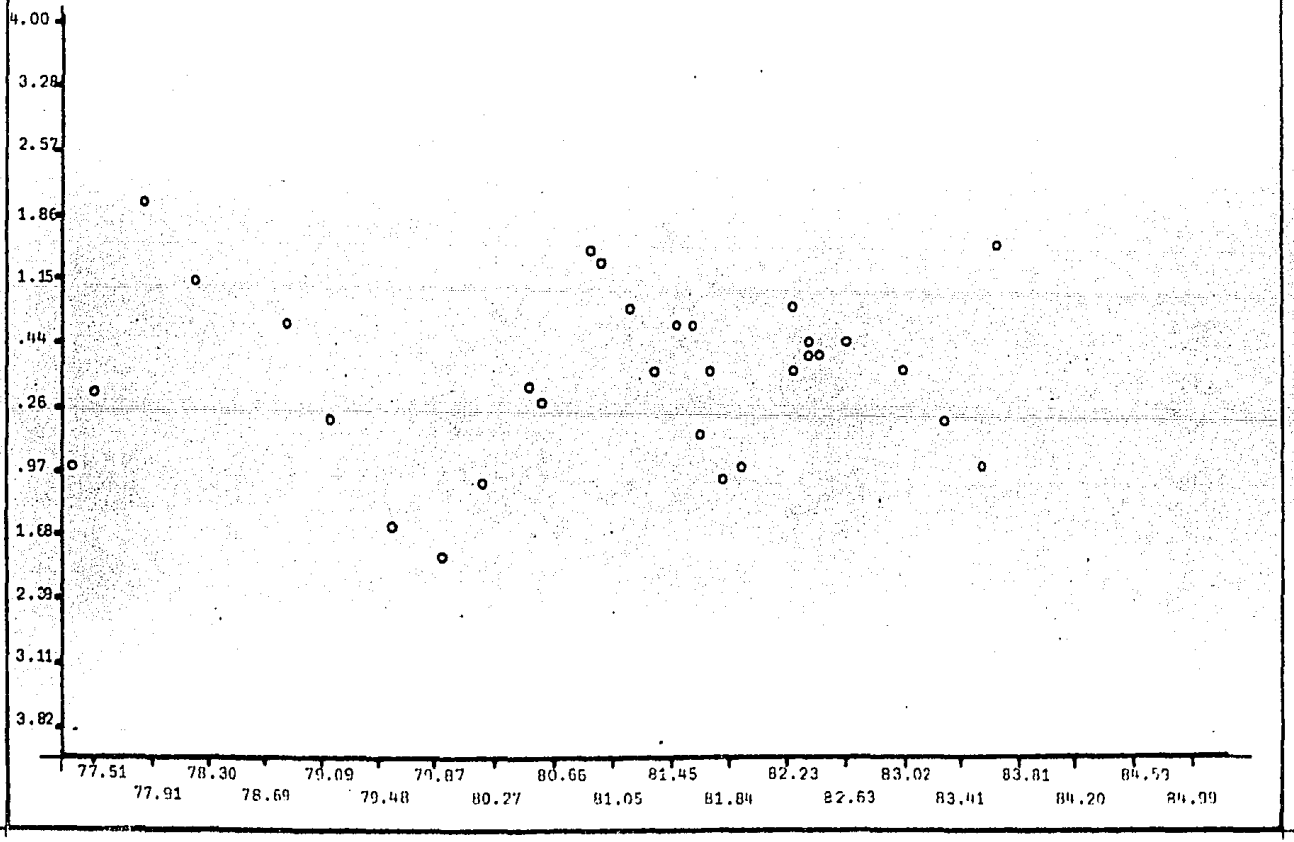
GM 86C



GRAFICA Nº 28

RESIDUALES . ESTUDIANTIZADOS CONTRA Z
COLUMNA 2

GM 86C

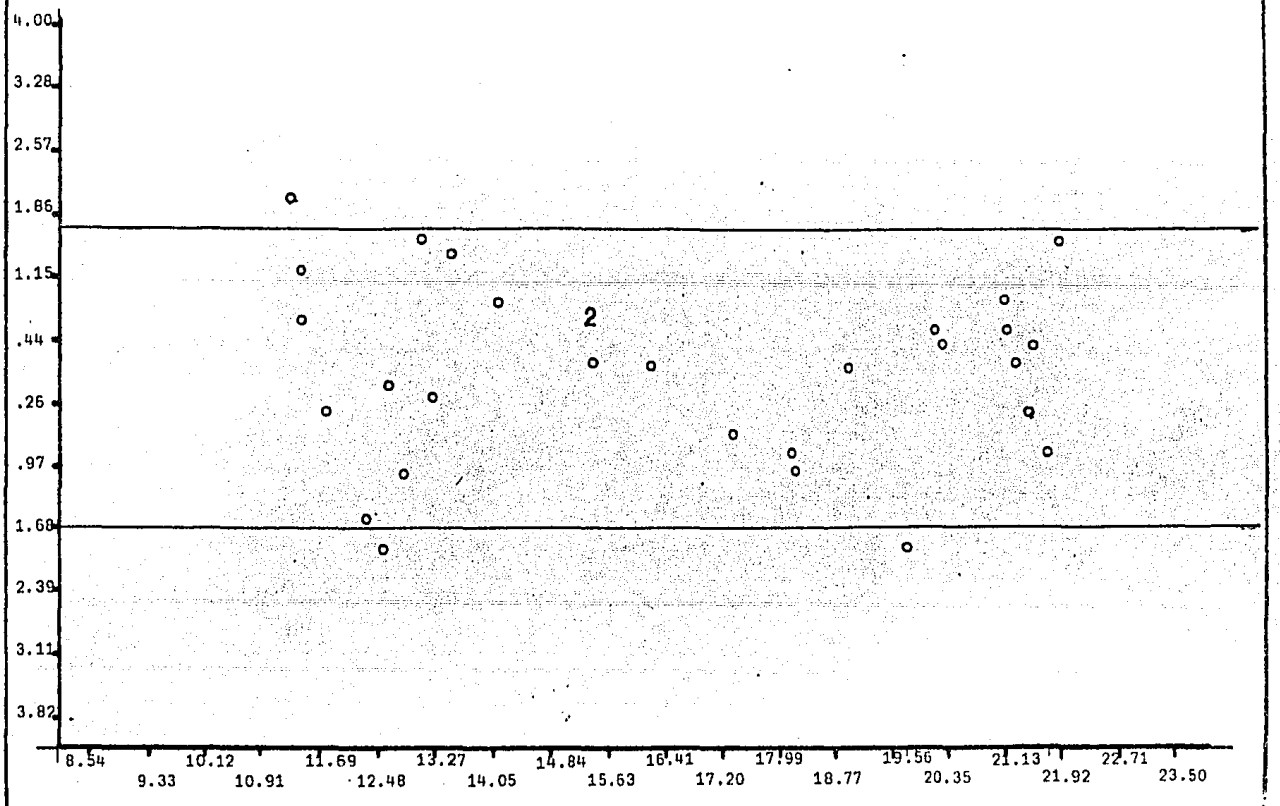


GRAFICA Nº 29

RESIDUALES ESTUDIANTIZADOS CONTRA Z

COLUMNA 4

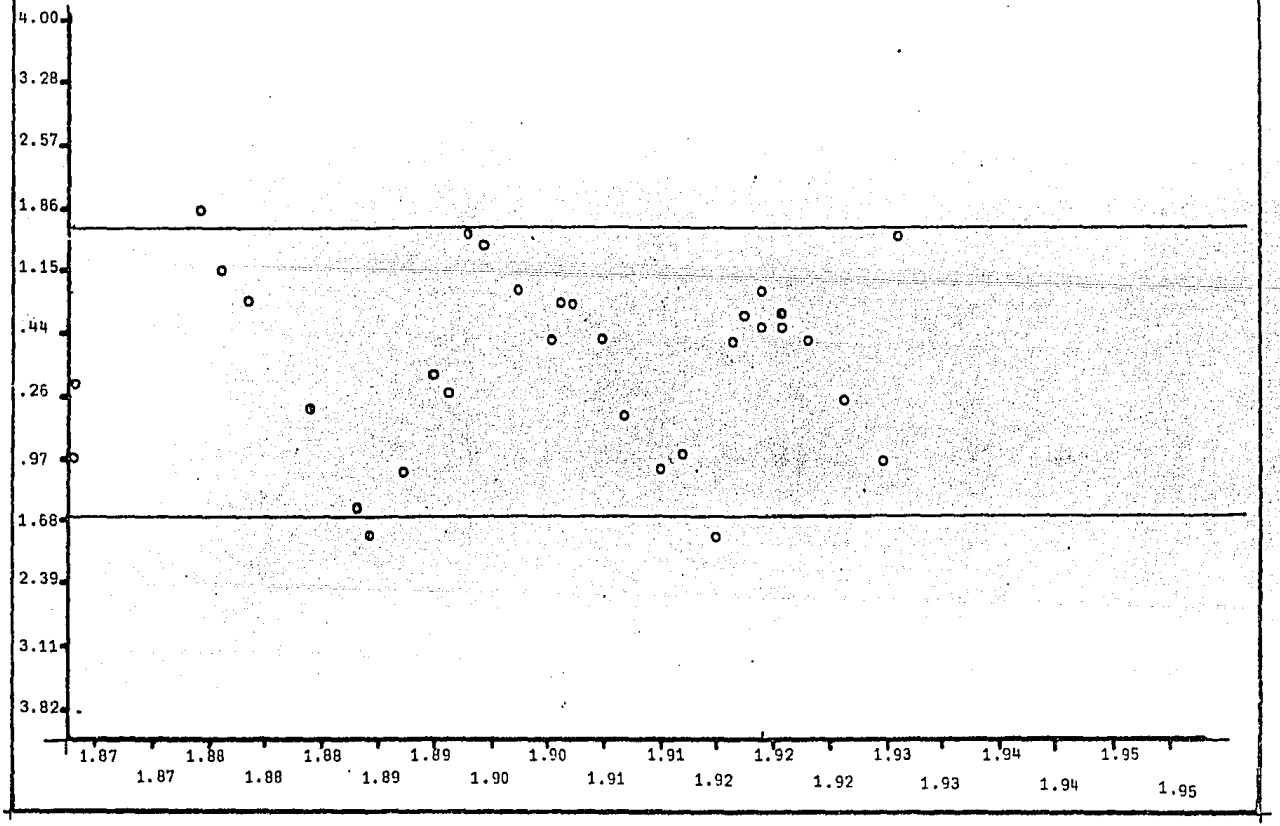
GM 86C



GRAFICA Nº 30

RESIDUALES . ESTUDIANTIZADOS CONTRA Y ESTIMADA

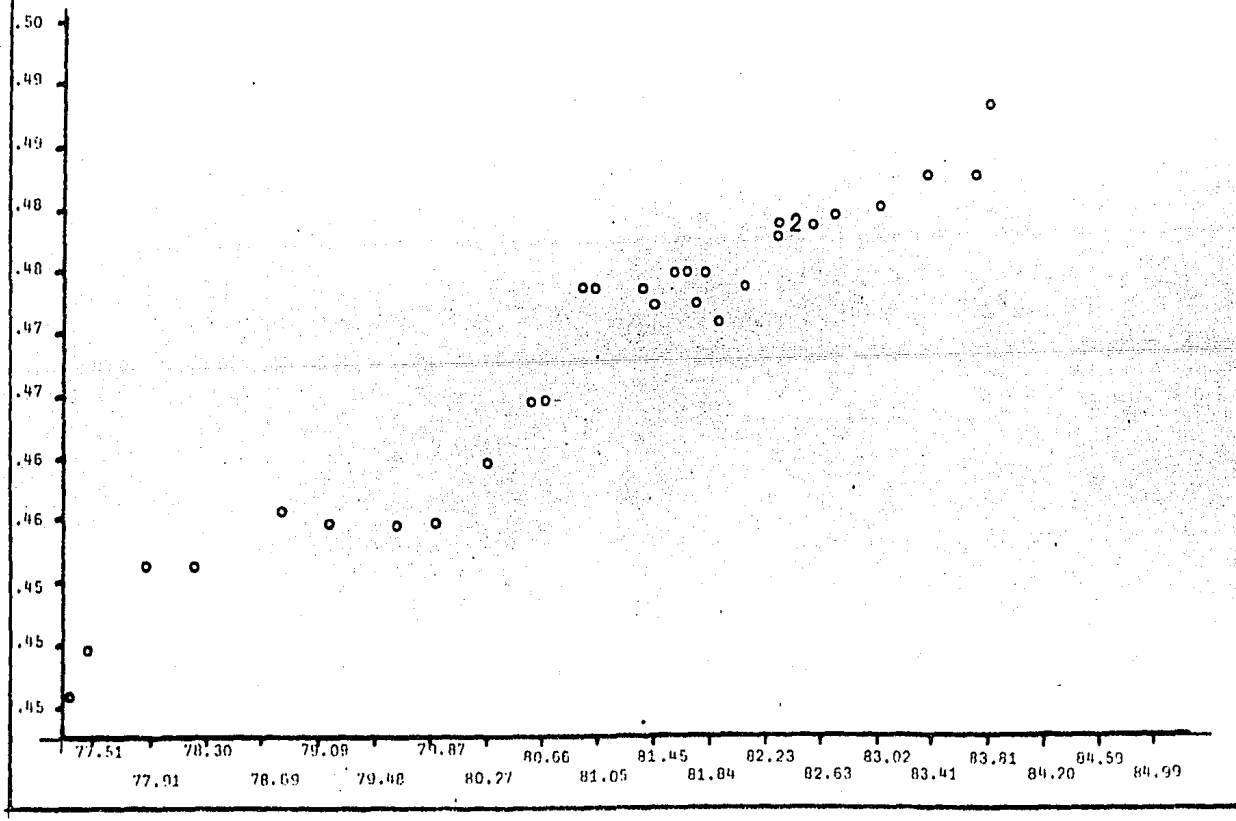
GM 86C



GRAFICA Nº 31

RESIDUALES PARCIALES CONTRA LA MATRIZ Z

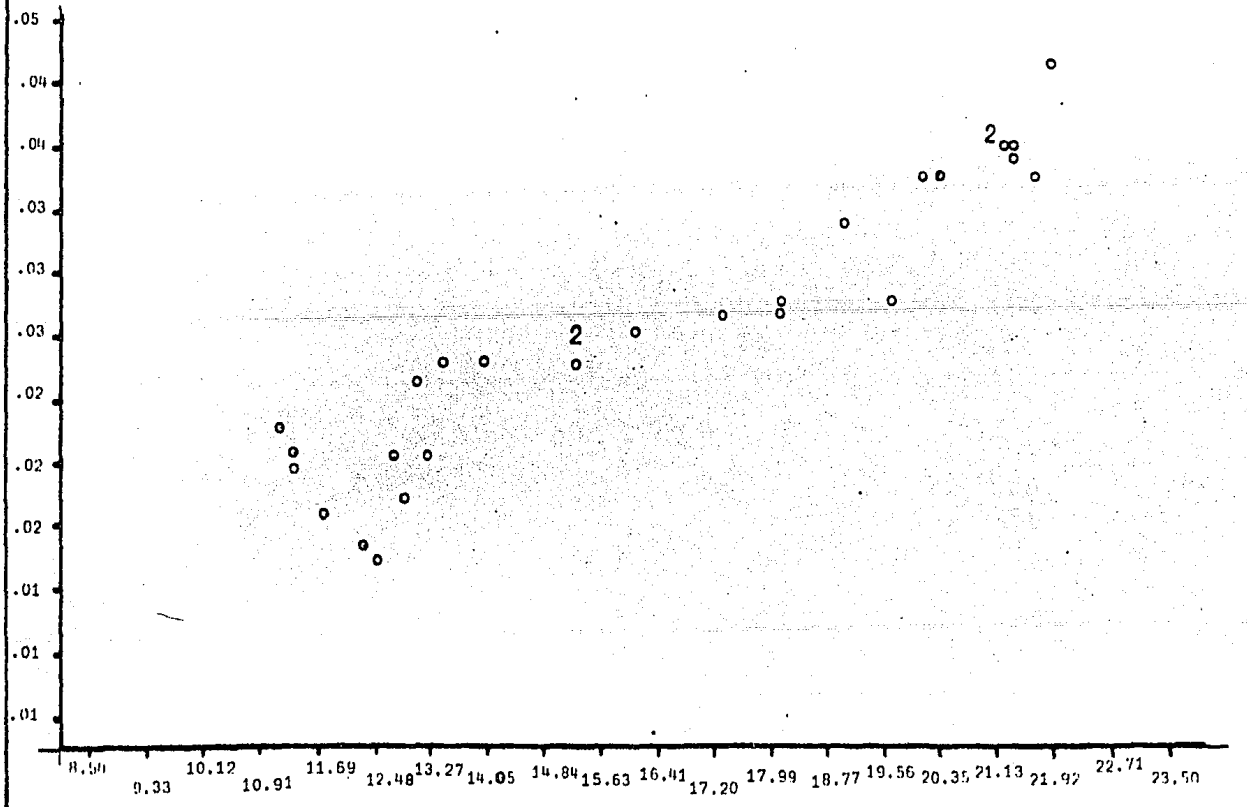
COLUMNA 2



GRAFICA Nº 32

RESIDUALES PARCIALES CONTRA LA MATRIZ Z
COLUMNA 4 .

GM
86C



Como era de esperarse las gráficas 28, 29 y 30 tienen los mismos comportamientos que las gráficas 25, 26 y 27, también es bastante lógico que al tener un 90% de confiabilidad las bandas se cierren, quedando fuera de ellas puntos que antes se encontraban en la frontera de ella. La confiabilidad que se requiera está en función del riesgo que se quiera correr en cualquier modelo estadístico que se trabaje.

Ahora llevaremos a cabo el análisis de las gráficas de residuales parciales, los cuales se expusieron en el inciso "v" de este capítulo.

La gráfica No. 31: residuales parciales contra la matriz Z, columna 2, se observa que hay una fuerte correlación de r_p y X_1, también se ve que no hay indicios de no linealidad, ni observaciones discrepantes.

La gráfica No. 32: residuales parciales contra la matriz Z, columna 4, también se visualiza que existe una fuerte correlación de r_p y X_3, además no hay indicios de no lineabilidad ni observaciones discrepantes. La diferencia entre estos dos diagramas es que en la No. 31 se sospecha de una periodicidad ascendente, mientras que en la No. 32 se podría pensar, ya en un caso extremo de una transformación cuadratica.

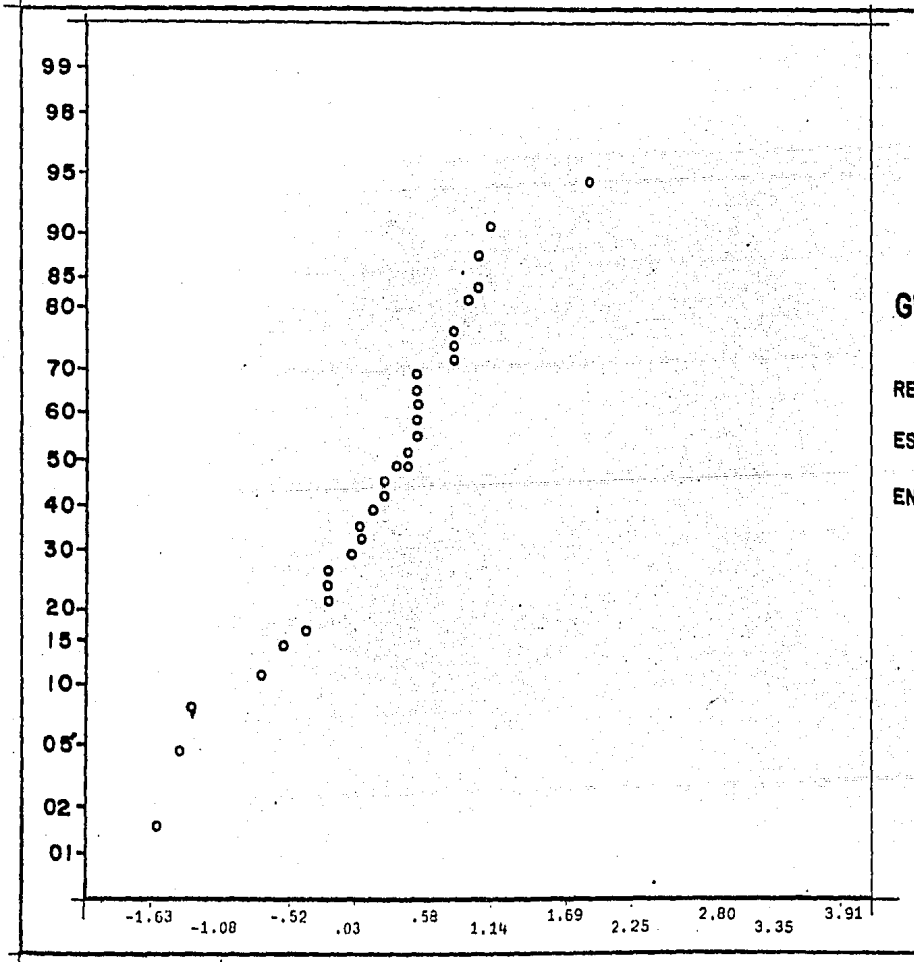
Pasamos a examinar a la gráfica No. 33 residuales estudentizados en papel normal, en este diagrama se observa perfectamente que el comportamiento de los residuales se ajusta bastante a una recta cuya pendiente es positiva, por lo tanto la hipótesis de normalidad se puede tomar, sin temor a incurrir en un error de "peso".

Ahora verificaremos otra de las suposiciones del modelo, con el objeto de afirmar con un nivel de significancia si los errores están correlacionados o no. Para ello requerimos de la técnica Durbin-Watson. Después de efectuar los cálculos correspondientes obtenemos que $D\text{-}w = 0.995$ y como $d_L^{\alpha} = 1.18$ con $\alpha = .05$, se tiene entonces d_L^{α} es mayor que $D\text{-}w$, por lo tanto "$D\text{-}w$" es significante bajo el nivel de significancia dado, es decir se sospecha que exista correlación positiva.

Aplicaremos la prueba de rachas de Wald-Wolfowitz, siendo un análisis no paramétrico. Para efectos de este modelo, se observó la siguiente sucesión de signos de los residuos.

\- - + + + - - - - - - - + + + + + + - + + - - - + + + + + + + - - +

Por lo tanto contamos con 18 residuos + y 14 residuos menos, se observaron 10 rachas bajo este modelo transformado.



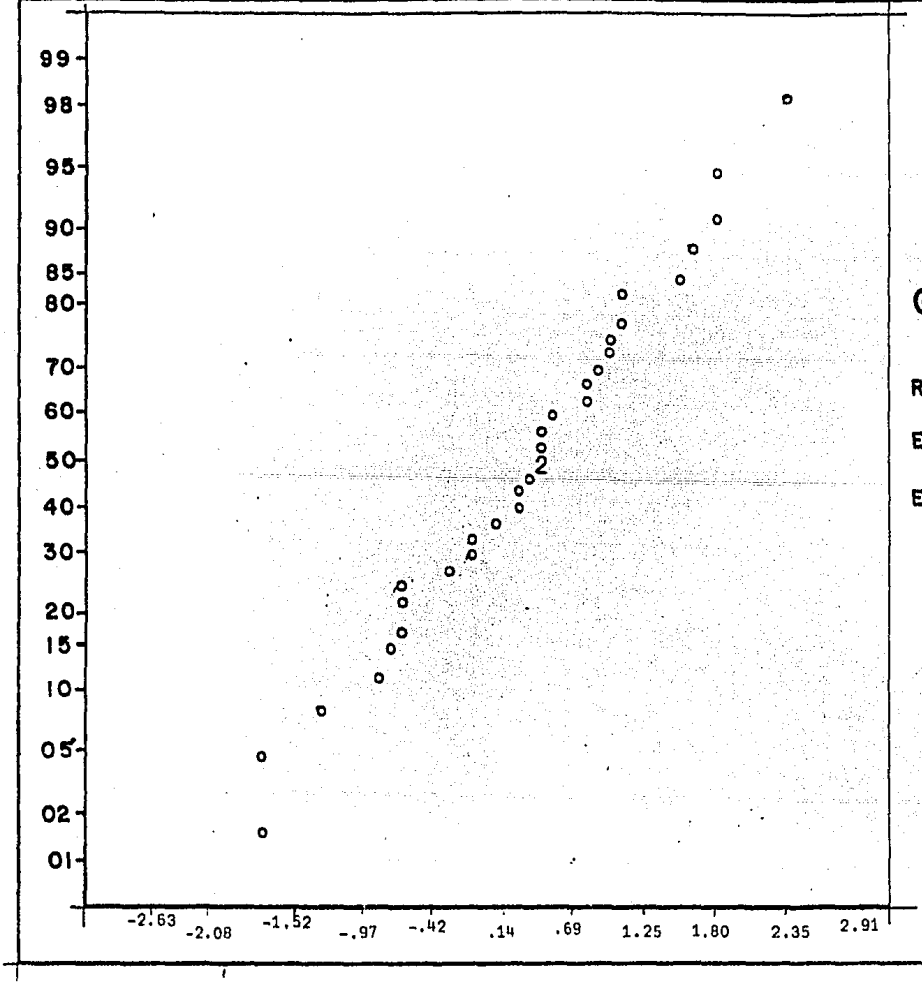
GM
86C

GRAFICA Nº 20

RESIDUALES

ESTUDIANTIZADOS

EN PAPEL NORMAL



GM
86C

GRAFICA Nº 33

RESIDUALES

ESTUDIANTIZADOS

EN PAPEL NORMAL

Como ocurrió en el modelo anterior no se cuenta con tablas que nos proporcionen un valor tabulado para los valores de $n_1 = 18$ y $n_2 = 14$, por lo cual tenemos que aproximar mediante una normal con parámetros μ y σ^2, con el propósito de verificar si existen muchas rachas o muy pocas.

$$\mu = \frac{2n_1n_2}{n_1+n_2} + 1$$

$$\therefore \quad \mu = \frac{2(18)(14)}{18+14} = 1$$

$$\mu = 16.75$$

$$\sigma^2 = \frac{2n_1n_2(2n_1n_2 - (n_1+n_2))}{(n_1+n_2)^2(n_1+n_2-1)}$$

$$\therefore \quad \sigma^2 = \frac{2(18)(14)\{2(18)(14) - (18+14)\}}{(18+14)^2(18+14-1)}$$

$$\sigma^2 = 7.4939$$

Como

$$u \sim N(\mu, \sigma^2)$$

$$\frac{u-\mu}{\sigma^2} \sim N(0,1)$$

Sustituyendo obtenemos lo siguiente:

$$\frac{u - 16.75}{7.499} \sim N(0,1)$$

$$F(X) = P(X \leq 10) = \underline{\Phi}\left(\frac{10-16.75+.5}{7.494}\right) = \underline{\Phi}(-2.2830)$$

Como $\underline{\Phi}(-y) = 1 - \underline{\Phi}(y)$

$\therefore$ $F(x) = 1 - \underline{\Phi}(2.2830)$

$F(x) = 1 - .9887 = .0113$

$\therefore$ No se rechaza que haya pocas rachas en los datos bajo un nivel de .05, pero bajo un nivel de .01 se rechazaría lo anterior.

Pasamos a considerar el otro enfoque complementario teniendo las siguientes hipótesis.

H_0: No $\exists$ un número demasiado grande de rachas

H_1: $\exists$ un número demasiado grande de rachas

$P\{X > 10\} = 1 - P\{X \leq 10\} = 1 - .0113 = .9887$

$\therefore$ Si rechazo H_0 tengo una probabilidad de error de 98.8% de otra manera, si determino un $\alpha = .05$, entonces no rechazo la hipótesis nula.

Por lo cual se rechaza que haya muchas rachas en los datos.

Concluimos que: Si tomamos un nivel del .01, no se rechaza aleatoriedad de los residuos, por lo que no hay in-

dicio de correlación serial, pero si tomamos un nivel del .05 tendríamos correlación serial positiva y se rechazaría la aleatoriedad de los residuos.

Continuamos el análisis con el examen de los puntos discrepantes e influyentes mediante la gráfica No. 34 residuales menos contra residuales, se puede visualizar directamente en este diagrama que los puntos se apegan bastante a una recta, sin sospechar hasta el momento que exista algún punto discrepante o influyente. Para verificarlo pasamos analizar la tabla 3.2, en la cual tenemos que .403 es el valor máximo de la distancia de Cook este valor corresponde al caso No. 3, si este caso fuera omitido del análisis el estimador del vector beta sufriría un movimiento equivalente a trasladar el estimador inicial a un elipsoide de 19.54% de confianza. Hasta el momento sospechamos que no es un punto influyente, pero es necesario verificarlo mediante la función influencia, la cual fue expresada en el capítulo II.

Tenemos que $D_3 = .403$ $F(.05, 4, 28) = 2.71$ observando esta escala familiar nos inclinamos a pensar en que no hay observaciones particulares que presenten una influencia contundente en el modelo.

Pasemos a verificar si existen puntos discrepantes en este modelo, para lo cual tomamos al mayor valor del estadistico de prueba, correspondiendole a la observación No. 22, $t_{22} = 2.2421$ (mayor en valor absoluto) indicandonos lo siguiente:

Eligiendo un nivel de significancia representativo para el contraste de Hipótesis:

$$H_0 : E(y_{22} - \tilde{y}_{22}) = 0 \quad \text{vs} \quad H_1 : E(y_{22} - \tilde{y}_{22}) \neq 0$$

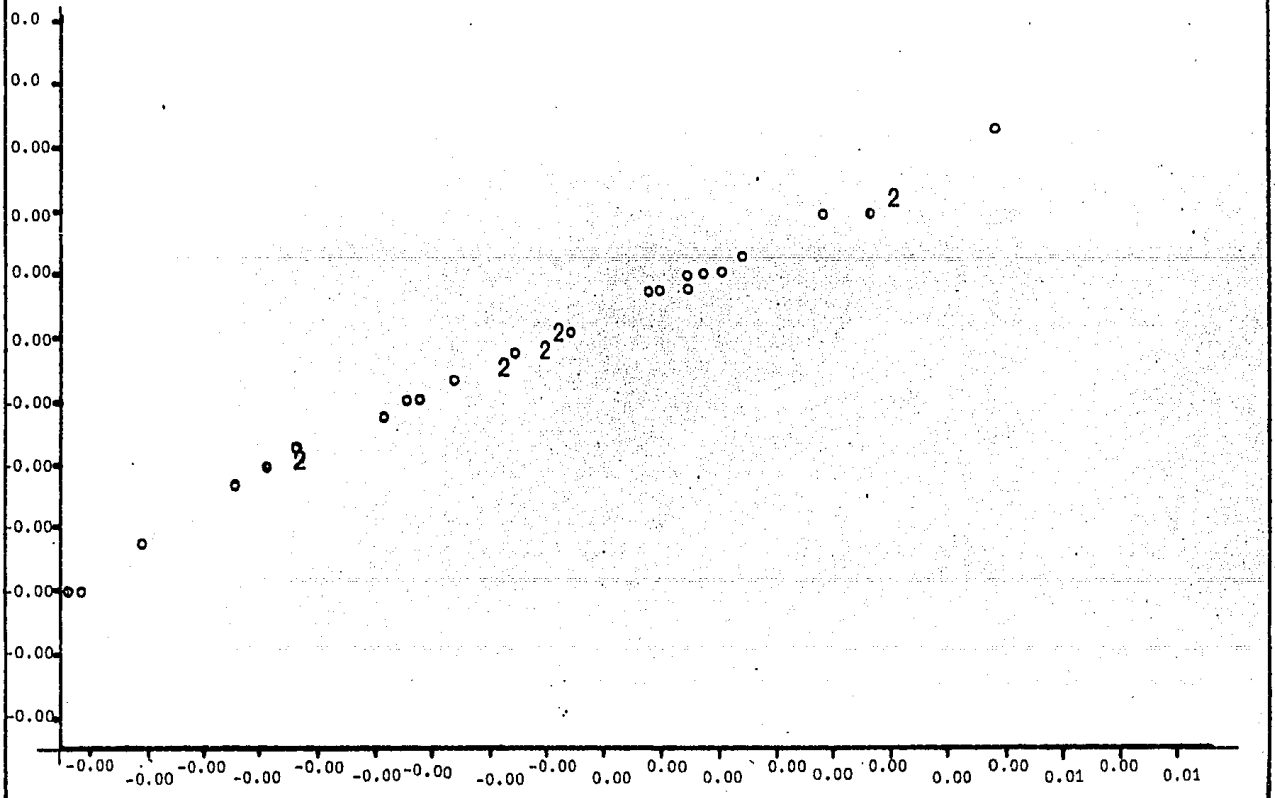
Tenemos que $t_{22} = 2.2421$ en valor absoluto y el estadístico $t_{22}(.01) = 2.77$, por lo cual $t_{22} < t_{22}(.01)$, por lo tanto no se cuenta con argumentos, para rechazar la hipótesis nula, es decir consideraríamos que la observación veintidós es no discrepante.

xi) Intervalos de Confianza

Puesto que en este modelo no hay evidencia de cancia de ajuste, estamos facultados para poder realizar estimación por intervalos.

GRAFICA Nº 34 RESIDUALES MENOS CONTRA RESIDUALES

GM 86C



T A B L A 3.2

VECTOR Y_T	VECTOR Y-ESTIMADA	VECTOR RESIDUAL	RESIDUALES STUDENTIZADOS	D(COOK)	T(COUTLIER)
1.861161	1.863531	-0.002370	-0.985475	0.058711	-0.984848
1.864350	1.864391	-0.000041	-0.018442	0.000036	-0.018110
1.878444	1.873719	0.004725	2.068981	0.403271	2.207432
1.878815	1.875921	0.002893	1.206938	0.090807	1.217277
1.880554	1.878988	0.001566	0.647235	0.023791	0.640381
1.881505	1.882648	-0.001143	-0.443591	0.004016	-0.437137
1.882247	1.886458	-0.004212	-1.622654	0.043382	-1.674069
1.882487	1.887547	-0.005060	-2.080175	0.230925	-2.221551
1.887706	1.890800	-0.003094	-1.195585	0.025783	-1.205207
1.892384	1.892460	-0.000075	-0.029589	0.000024	-0.029057
1.893192	1.893855	-0.000663	-0.257443	0.001346	-0.253104
1.899287	1.895569	0.003719	1.484116	0.079155	1.518314
1.900322	1.896935	0.003387	1.338689	0.054406	1.358768
1.901489	1.899391	0.002098	0.825997	0.019283	0.821179
1.902793	1.902394	0.000400	0.154329	0.000424	0.151612
1.905001	1.903416	0.001585	0.616721	0.008309	0.609763
1.905933	1.904167	0.001766	0.690410	0.011544	0.683814
1.906456	1.907999	-0.001543	-0.604906	0.009512	-0.597926
1.906755	1.906338	0.000417	0.161609	0.000515	0.158771
1.908102	1.910922	-0.002820	-1.093920	0.024131	-1.097926
1.909730	1.912163	-0.002433	-0.942123	0.017011	-0.940168
1.910159	1.915406	-0.005247	-2.096413	0.160695	-2.242106
1.916066	1.915992	0.000075	0.029126	0.000020	0.028601
1.918943	1.917746	0.001197	0.471666	0.006379	0.465017
1.919772	1.919214	0.000558	0.221977	0.001701	0.218170
1.921080	1.919188	0.001892	0.763811	0.024838	0.757986
1.921693	1.921259	0.000434	0.181952	0.002150	0.178779
1.922501	1.921174	0.001327	0.528635	0.009723	0.521719
1.923872	1.923553	0.000319	0.126474	0.000508	0.124231
1.924252	1.925226	-0.000974	-0.386831	0.004990	-0.380880
1.925061	1.927483	-0.002422	-0.968922	0.035101	-0.967826
1.932627	1.928889	0.003738	1.504208	0.091924	1.540664

La estimación puntual de un parámetro no resulta de mucho valor si no se posee alguna medida del posible error cometido en la estimación. Toda estimación de un parámetro debe acompañarse de cierto intervalo que incluya al estimador. La importancia de los intervalos de confianza radica en la eficiencia de determinar un intervalo alcatorio en el que puede esperarse que este contenido el verdadero valor de la incógnita con probabilidad muy cercana a uno.

Calcularemos primero el intervalo de confianza para la *i-ésima* componente de $\underline{\beta}$, para ello requerimos la suposición de normalidad en los errores (esta conjetura puede ser aceptada pues anteriormente se verificó que los errores tienden a distribuirse normalmente es decir, que $\beta \sim N(\beta, \sigma^2 (Z^1 Z)^{-1})$, además tomando en cuenta la independencia de $\hat{\beta}$, y S^2 el intervalo está dado por:

$$\{\hat{\beta}_i - t_{(\alpha/2,\ n-p)}\ S\ C_{ii} \leq \beta_i < \hat{\beta}_i + t_{(\alpha/2,\ n-p)}\ S\ C_{ii}\}$$

donde, C_{ii} es la *i-ésima* componente de la diagonal de $(Z^1 Z)^{-1}$ y $t_{\alpha/2}$ es el cuantil $(1 - \alpha/2)$ de la distribución *t-student*.

Por otro lado tenemos que:

$$Pr\ (\hat{\beta}_i - (t_{\alpha/2},\ n-p)\ S\ C_{ii} \leq \beta_i \leq \hat{\beta}_i + (t_{\alpha/2},\ n-p)\ S\ C_{ii}) = 1-\alpha$$

donde 100 x $(1-\alpha)$ es la confiabilidad del intervalo.

Aplicando lo arriba señalado al modelo 3.15, calculamos el intervalo de confianza para la beta cero al 95% de confianza, el cual es:

$$\{1.269 \leq \beta_0 \leq 1.517\}$$

Ahora para la beta uno al 95% de confianza es:

$$\{.004 \leq \beta_1 \leq 0.008\}$$

Para la beta cinco al 95% de confianza es:

$$\{-0.003 \leq \beta_5 \leq 0.001\}$$

Y finalmente para la beta nueve al 95% de confianza es:

$$\{0.001 \leq \beta_9 \leq 0.003\}$$

Otro de los intervalos de confianza de interés es el de la varianza, para calcularlo requerimos de la siguiente expresión:

$$\frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p}$$

Dado lo anterior es posible construir un intervalo de confianza para σ^2, nuestro problema se reduce a encontrar valores A y B tales que:

$$Pr\{A \leq \frac{(n-p)S^2}{\sigma^2} \leq B\} = 1 - \alpha$$

Nuestro deseo es que estos valores produzcan un intervalo cuya longitud fuese lo más pequeña posible. Por lo anterior el intervalo tiene la siguiente forma:

$$\{\frac{(n-p)S^2}{B} \leq \sigma^2 \leq \frac{(n-p)S^2}{A}\}$$

Comunmente, no es sencillo seleccionar A y B que cumplan las siguientes restricciones.

$$P\{u \leq A\} = P\{B < u\} = \alpha/2$$

donde $u \sim \chi^2_{n-p}$

Afortunadamente existe otra manera de seleccionar A y B, la cual es proporcionada por Tate y Klett (1959).

Como fue expuesto anteriormente, deseamos encontrar valores A y B de modo tal que el intervalo para σ^2 sea de longitud minima. Aplicamos el procedimiento desarrollado por Tate Klett (1959), el cual es un procedimiento iterativo que produce los valores deseados. Los intervalos tienen la siguiente forma:

$$\{ \frac{1}{b_n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \leq \sigma^2 \leq \frac{1}{a_n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \}$$

donde a_n y b_n son valores tabulados y la n representa los grados de libertad (la tabla está anexa en el apéndice I).

En caso de que existan repeticiones es necesario cambiar los numeradores del intervalo de arriba y poner la siguiente expresión:

$$\sum_{i=1}^{K} \sum_{u=1}^{n_i} (y_{iu} - \bar{y}_i)^2 \text{ con } \sum_{i=1}^{K} n_i - K \text{ grados de libertad.}$$

donde

ni es el número de repeticiones en Z_i

Z_i vectores de observaciones de 1 x p

El intervalo de confianza para la varianza al 95% de confianza es:

$$\{4.787 \times 10^{-6} \leq \sigma^2 \leq 1.915 \times 10^{-5}\}$$

Los siguientes intervalos son para las combinaciones lineales de las componentes de β.

La forma de obtención del intervalo de confianza para combinaciones lineales de β es por medio de un elipsoide de confianza dado por:

$$\frac{(\hat{\psi} - \psi)^1 \; B^{-1} \; (\hat{\psi} - \psi)}{S^2} \sim F(q, n-p)$$

donde $\psi = C\beta$, C_{qxp} y $B = C(Z'Z)^{-1}C^1$

Además sabemos que:

$$\hat{\beta} \sim N(\beta, \sigma^2(Z^1Z)^{-1})$$

$$c\hat{\beta} \sim N(c^1\beta, \sigma^2 c^1(Z^1Z)^{-1}C)$$

$$M = \frac{c\hat{\beta} - c\beta}{\sigma A^{1/2}} \sim N(0,1)$$

$$\rightarrow N = \frac{n-p}{\sigma^2} S^2 \sim X^2_{n-p} \qquad T = \frac{M}{N/n-p} \sim t(n-p)$$

donde c' es un vector de $(1 \times p)$ y $A = c'(Z'Z)c^{-1}$.

Entonces el intervalo de confianza para una combinación lineal de β está dado por:

$$c^1\hat{\beta} \pm t^{(\alpha/2)}_{n-p} \; S \; \{c^1(Z^1Z)^{-1}c\}^{1/2}$$

Por lo anterior en nuestro problema el intervalo de confianza al 95% para $c'\beta$ es:

$$\{1.278,\ 1.523\}$$

donde $c^1 = (1,\ 1,\ 1,\ 1)$

El siguiente intervalo de confianza al 95% para $c^1\beta$ es:

$$\{1.269,\ 1.516\}$$

Para este intervalo $c^1 = (1,\ 0,\ 1,\ 0)$

Ahora realizaremos el intervalo de confianza para y dado x_0. A éste también se le llama intervalo de predicción, es decir, si nosotros deseamos un intervalo de confianza para el valor de la variable respuesta en x_0, es lo mismo que encontrar el intervalo de predicción.

El intervalo de predicción se obtiene de la diferencia $\hat{y}_0 - y_0$, donde $\hat{y}_0 = \hat{\beta}^1 x_0$ y $y_0 = \beta^1 x_0 + \varepsilon$, por lo tanto $\hat{y}_0 - y_0 \sim N(0,\ \sigma^2(x_0^1(Z^1Z)^{-1}x_0 + 1))$, así, el intervalo está dado por:

$$\left\{\hat{y}_0 - t_{\alpha/2}\, S(x_0^1(Z^1Z)^{-1}x_0+1)^{1/2} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2}\, S(x_0^1(Z^1Z)^{-1}x_0+1)^{1/2}\right\}$$

Con lo expuesto arriba pasemos a calcular el in-

tervalo de confianza para y/x_0 con una confiabilidad de .95

{1.935, 1.948} y cuyo centro está en 1.941

el vector x_0^1 = (1,85.7, 1, 22.3)

Ahora calculamos otro intervalo de predicción, combinando unicamente el vector x_0, el cual es: x_0 = (1,90.94, 0, 30.3) y el intervalo de confianza para y/x_0 es:

{1.98, 1.999} centrado en 1.989

Por último calculamos el intervalo de confianza para $E(Y/x_0)$. Para un vector x_0 el estimador de la media de y_0 esta dado por:

$$E(y/x_0) = x_0^1 \beta \qquad (x_0 \text{ vector de } Px1);$$

Por lo que el intervalo de confianza corresponde al intervalo de confianza de una media; en este caso el intervalo está dado por:

$$\{x_0^1\hat{\beta} - t_{\alpha/2} S(x_0^1(Z^1Z)^{-1}x_0)^{1/2} < x_0^1\beta < x_0\hat{\beta} + t_{\alpha/2} S(x_0^1(Z^1Z)^{-1}x_0)^{1/2}\}$$

El intervalo de confianza correspondiente para $E(y/x_0)$ al .95 es:

{1.936, 1.945} centrado en 1.941

Los vectores X_0 fueron simulados de acuerdo a las características y tendencias que presenta la información proporcionada por la Cía. Aeroméxico. La simulación fue necesaria debido a que los responsables del área no propusieron ningún vector.

donde $x_0^1 = (1.85.7,\ 1,\ 22.3)$

Calculamos además otro intervalo de confianza para $E(y/x_0)$ al .95, cambiando solamente al vector x_0, el cual es:

(1.982, 1.997) centrado en 1.989

donde $x_0^1 = (1, 90.94,\ 0,\ 30.3)$

xii) **Transformación de Variables Explicativas**

Box y Tidwell (1962) proponen una familia de transformaciones definida por:

$$W_i = \begin{cases} X_i^{\alpha i} & \alpha_i \neq 0 \\ \ln X_i & \alpha i = 0 \end{cases}$$

Para $i = 1 \ldots p$ donde $X_1, X_2 \ldots Xp$ son las variables independientes originales.

La estimación directa en un modelo $y = Z\beta + \varepsilon$ implica análisis de un modelo no lineal.

Box y Tidwell proponen un método iterativo que evita estimación no lineal y es como sigue:

1. Se inicia con valores $\alpha_1 = \alpha_2 = \ldots \alpha p = 1$

2. Se ajusta un modelo lineal aumentado para mejorar la estimación de α. Este proceso se repite hasta lograr un nivel de convergencia adecuado

Primero se ajusta

$$\hat{y} = \hat{\beta}_0 + \sum_1^p \beta_j w_j$$

con $w_j = X_j^{\alpha_j} = X_j \qquad j = 1, \ldots p$

Por lo cual se forman p nuevas variables $Z_1, \ldots Z_p$ definidas como $Z_j = w_j \ell n(X_j) \qquad j = 1 \ldots p$ que se añaden al modelo.

El método iterativo para realizar la transformación de la variable explicativa se basa en ajustar

$$y = \beta_0 + \beta_1 X^{\alpha_1} + \varepsilon$$

con $\alpha_1 = 1$ como valor inicial

Se ajusta ahora un modelo aumentado para obtener una estimación mejorada de α, han de crearse un número de variables auxiliares igual al de variables explicativas por transformar, en este caso.

$$Z_1 = X^{\alpha_1} \ell n (X^{\alpha_1})$$

se define entonces

$$y = \beta_0^* + \beta_1^* X^{\alpha_1} + \gamma_1 Z_1 + \xi$$

Tal que si (γ_1) es grande hay necesidad de transformar

$$\hat{\alpha}_1^* \left(\frac{\hat{\gamma}_j}{\beta_j} + 1 \right) \times (\text{valor actual de } \hat{\alpha}_1)$$

Al realizar los cálculos correspondientes obtenemos que la transformación adecuada para nuestro modelo es:

$$yT = \beta_0 + \beta_1 X_1 + \beta_5 X_5 + \beta_9 X_9^{1.5} \qquad (3.16)$$

$$E(\varepsilon) = 0, \; Var(\varepsilon) = \sigma^2 In \; y \; Cov(\varepsilon_i, \varepsilon_j) = 0 \; \forall \, i,j \quad i \neq j$$

Nosotros realizamos una transformación a la variable respuesta X_9 sugerida al aplicar el Método de Box y Tidwell, esta transformación puede mejorar la situación, pero esto no se garantiza totalmente. Es necesario llevar a cabo un análisis completo de los datos transformados.

Debido al comentario expuesto arriba continuaremos el análisis, calculando primero los estimadores respectivos de los parámetros.

Para encontrar los estimadores, aplicaremos la técnica de Mínimos Cuadrados, bajo las siguientes suposiciones $E[\varepsilon] = 0$ y $Var(\varepsilon) = \sigma^2 I_n$, los resultados obtenidos se exponen a continuación.

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_5 \\ \hat{\beta}_9 \end{bmatrix} = \begin{bmatrix} 1.3380 \\ 0.0068 \\ -0.0015 \\ 5.36 \times 10^{-5} \end{bmatrix}$$

Las varianzas correspondientes para las betas estimados son:

$$\hat{Var}(\hat{\beta}_0) = 0.0030$$

$$\hat{Var}(\hat{\beta}_1) = 5.0048 \times 10^{-7}$$

$$\hat{Var}(\hat{\beta}_5) = 1.3662 \times 10^{-6}$$

$$\hat{Var}(\hat{\beta}_9) = 9.3635 \times 10^{-11}$$

Los cálculos de las varianzas estimadas de los betas gorros tienen bastante importancia para el contraste de modelos, pues se observan las ganancias ó perdidas que se obtuvieron al transformar.

Pasamos a calcular otro estadístico de interés, que sirve para comparar modelos (expresión 2.6), este tiene el nombre de varianza estimada.

$$\hat{\sigma}^2 = 7.691 \times 10^{-6}$$

Deseamos hacer notar que para este modelo transformado la condición del número $K = 5.3476$, implica que ahora contamos con una multicolinealidad más modesta, pues en el modelo 3.5 se tenía como condición $K = 9.8535$.

Proseguimos el análisis del Modelo al examinar la tabla de análisis de varianza, la cual nos proporciona información acerca de los parámetros en cuestión. La prueba de hipótesis consiste en contrastar H_0: $B_1 = B_5 = B_9 = 0$ contra H_1: al menos un $B_j \neq 0$ $j = 1, 5, 9$

Tabla de Análisis de Varianza

Fuente	**G.L.**	**S.C.**	**C.M.**	**F**	**p**
Regresión	3	0.011	0.064	461.402	0.0000
Residual	28	2.153×10^{-4}	7.691×10^{-6}		
Total	31	0.011			

Al examinar la tabla de arriba observamos que tenemos una F calculada bastante grande, por lo cual se cuenta con argumentos para rechazar H_0: $B_1 = B_5 = B_9 = 0$ con probabilidad 0.0000 de cometer un error.

El coeficiente de determinación para el modelo 3.16) es .980, es decir el 98% de la variabilidad observada en YT es explicada por las variables indepen-

dientes y el coeficiente de correlación es simplemente la raíz cuadrada de este, por lo cual

$$r = \sqrt{R^2} = \sqrt{.98} = .9899$$

La prueba de carencia de ajuste consiste en contrastar H_0: no hay carencia de ajuste contra H_1: existe carencia de ajuste, en esta prueba lo más importante es estimar la variación debido al modelo, ya que si ésta es grande el modelo es erróneo.

En el caso en que no se cuente con repeticiones, se aplicará una prueba conocida como "Prueba Arcoiris" (rainbow test) para carencia de ajuste, esta prueba fue desarrollada inicialmente por UTTS (1982).

La "Prueba Arcoiris" está basada en el cociente de dos estimadores de la varianza, ambos son insesgados si las suposiciones son válidas y sesgadas en otro caso.

En nuestros modelos no existen repeticiones en las observaciones, por lo cual las pruebas de carencia de ajuste, fueron calculadas mediante la "Prueba Arcoiris" para carencia de ajuste.

Tabla de Análisis de Varianza

Fuente	G.L.	S.C.	C.M.	F	P
Residual	28	2.153×10^{-4}	7.691×10^{-6}		
Carencia de A	15	6.619×10^{-5}	4.413×10	0.385	1.000
Error Puro	13	1.491×10^{-4}	1.147×10^{-5}		

De acuerdo con la tabla de arriba la $F_{cal} = 0.385$ y la $F_{(.05, 15, 13)} = 2.53$, es decir $F_{cal} < F_{tab}$, por lo cual no se tienen argumentos para rechazar H_0: no hay carencia de ajuste, es decir no se rechaza la hipótesis nula con probabilidad 1.0000.

Los cálculos obtenidos arriba nos proporcionan bases suficientes para continuar nuestro trabajo elaborando intervalos de confianza y pruebas de hipótesis, además de validar la tabla donde se contrasta $H_0: B_i = 0$ vs $H_1: B_i \neq 0$ para algún $i = 1,5,9$.

Continuando con el análisis pasamos a graficar los residuales estudentizados contra la matriz Z, las cuales se llevarán a cabo con el 95% de confianza.

La gráfica No. 35 residuales estudentizados contra Z columna 2 al parecer tiene un diagrama de dispersión nulo, sin embargo sospechamos que una pequeña heterogeneidad de - varianza, puesto que, los residuales estudentizados muestran

un comportamiento periódico.
Es importante hacer notar que existe una observación que se dispara un poco del conjunto de datos, por lo cual ese dato es candidato a ser discrepante o influyente.

La gráfica No. 36: residuales estudentizados contra Z, columna 4, muestra en su diagrama también un comportamiento periódico, bastante similar al de la gráfica No. 35 por lo cual podemos sospechar que es un diagrama nulo, con pequeña heterogeneidad de varianza. Por supuesto las tendencias de ambos diagramas son diferentes, aunque ambas tienen una observación que discrepa de las demás.

En la gráfica No. 37: residuales estudentizados contra "Y" estimada, como se expuso en el capítulo anterior, es el diagrama más importante y nuestro diagnóstico respecto a él, es que su comportamiento es nulo al igual que las gráficas anteriores, sospechamos que existe heterogeneidad de varianza. Se observa en este diagrama que más del 95% de los residuales caen en el rango (+2, -2,) por lo cual sospechamos que la forma del modelo es correcta.

Por otro lado, los residuales parciales contra la matriz Z, al ser graficados proporcionan las tendencias sistemáticas de cada una de las variables X.

En la gráfica No. 38 residuales parciales contra la matriz Z columna 2, se observa una fuerte correlación entre ηp y X_1. No hay indicios de no lineabilidad, ni de observaciones discrepantes.

La gráfica No. 39 residuales parciales contra la matriz Z, columna 4, muestra una fuerte correlación entre ηp y X_3, hay indiciones de dos observaciones discrepantes y se puede decir que existen indicios de lineabilidad.

Ahora aplicamos la técnica de graficación de residuales estudentizados en papel normal. (Gráfica No. 40), la cual muestra que aproximadamente el 95% de los puntos del diagrama se apegan a una recta, por lo cual la hipótesis de normalidad de los errores, se puede tomar sin caer en un error de "peso".

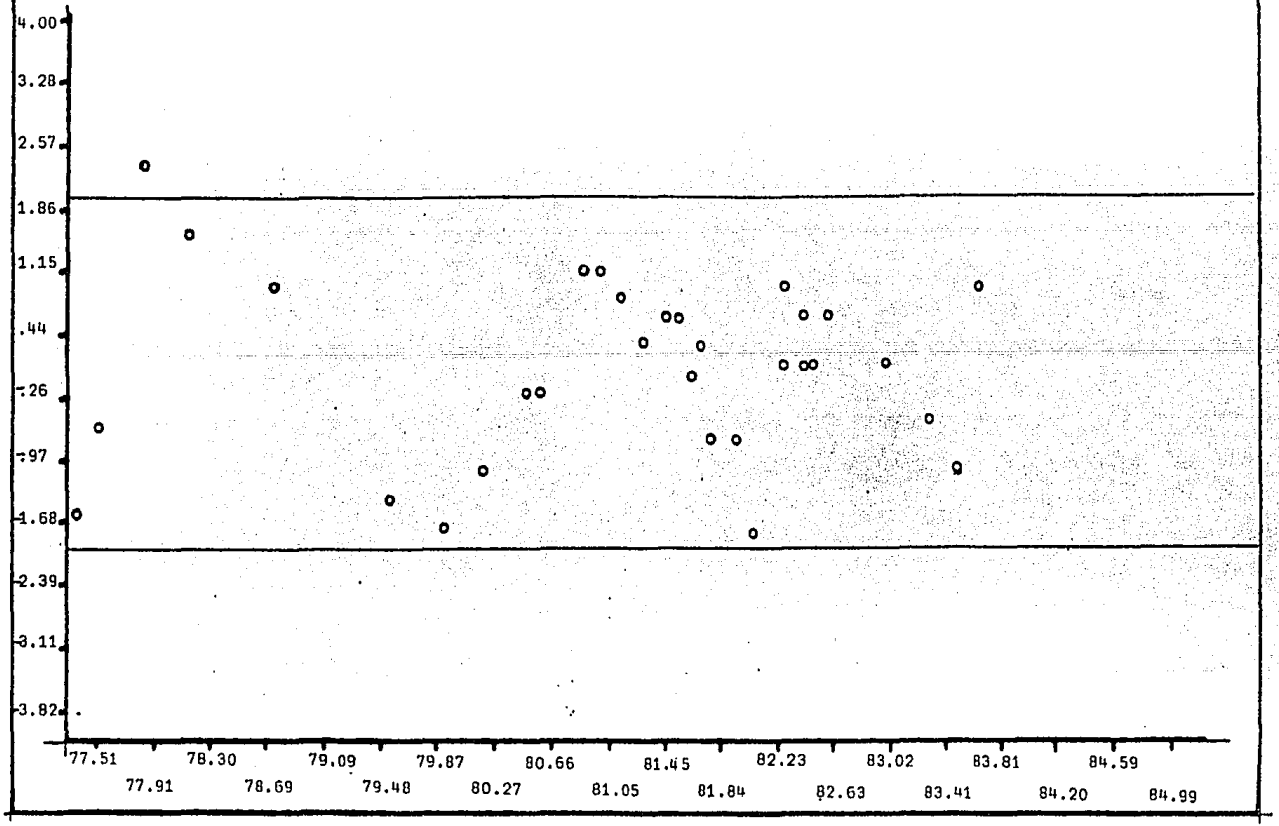
Para verificar la no-correlación de los errores, utilizaremos la técnica Durbin-Watson.

El modelo en cuestión tiene asociado un valor para D-W igual a 1.022, es decir $d = 1.022$, y como $d_L^{\alpha} = 1.18$, $\alpha = 05$, entonces $d_L^{\alpha} > d$, por lo cual es significante a un nivel $\alpha = .05$, por lo tanto se cree que exista correlación positiva.

GRAFICA Nº 35

RESIDUALES ESTUDIANTIZADOS CONTRA Z
COLUMNA 2

GM
86C

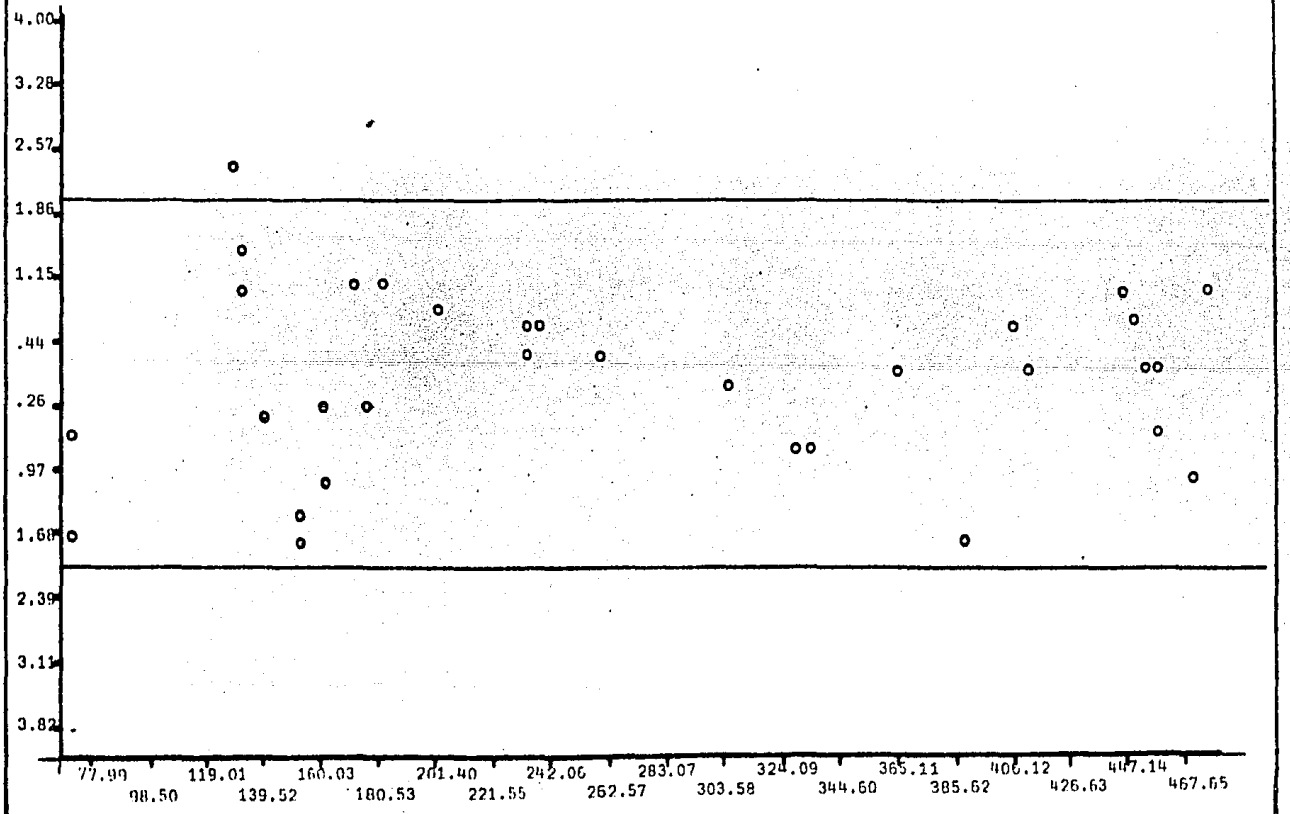


GRAFICA Nº 36

RESIDUALES ESTUDIANTIZADOS CONTRA Z

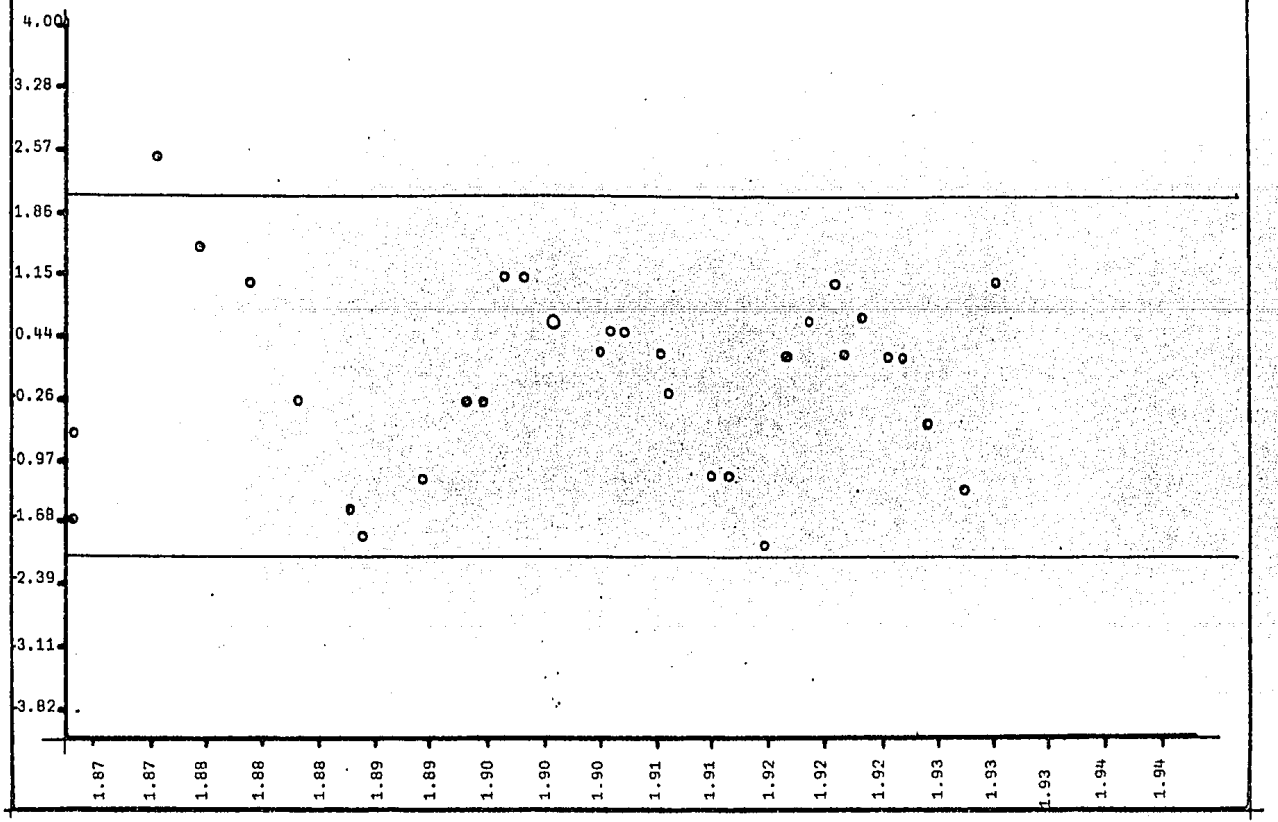
COLUMNA 4

GM 86C



GRAFICA Nº 37 RESIDUALES ESTUDIANTIZADOS CONTRA Y ESTIMADA

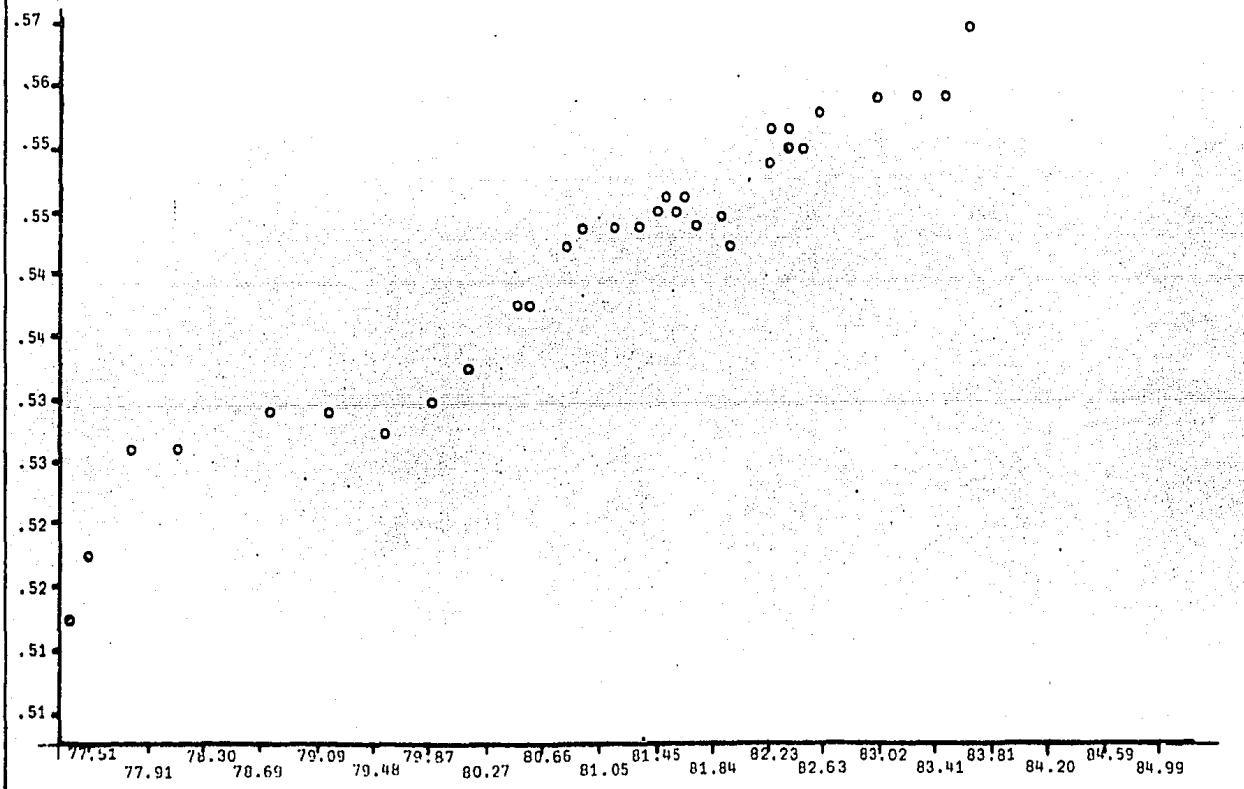
GM 86C



GRAFICA Nº 38

RESIDUALES PARCIALES CONTRA LA MATRIZ Z
COLUMNA 2 .

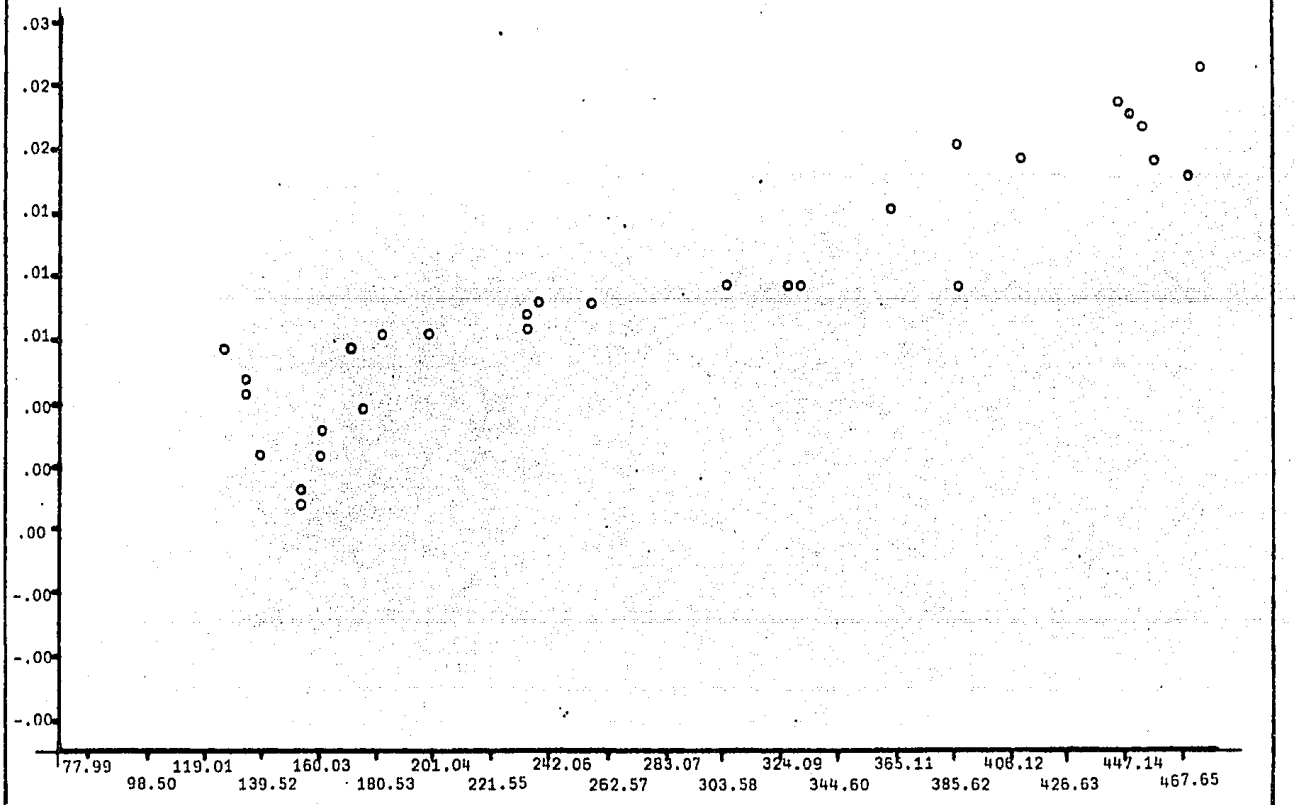
GM
86C

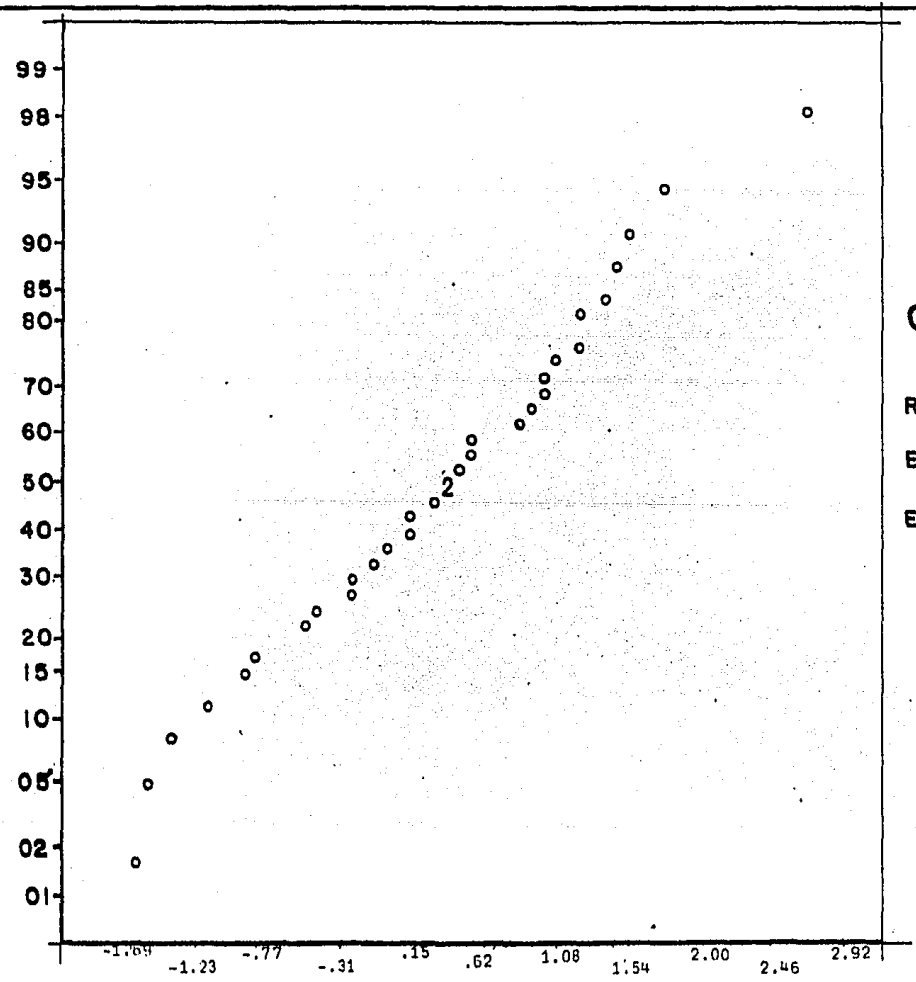


GRAFICA Nº 39

RESIDUALES PARCIALES CONTRA LA MATRIZ Z
COLUMNA 4 .

GM 86C





GM
86C

GRAFICA Nº 40

RESIDUALES

ESTUDIANTIZADOS

EN PAPEL NORMAL

También utilizaremos la prueba de Wald-Wolfowitz. Para nuestro problema contamos con 32 signos de residuos, los cuales son:

- - + + + - - - - - - + + + + + + - + - - - + + + + + + + - - +

La sucesión presenta 18 residuos positivos y 14 residuos negativos, teniendo 10 rachas.

Como anteriormente se explicó, no se cuenta con tablas para valores n_1 $n_2 > 10$, por lo tanto explicaremos la aproximación, a la normal. Los estimadores son:

$$\mu = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

$$\sigma^2 = \frac{2n_1 n_2 \ (2n_1 n_2 - n_1 - n_2)}{(n_1+n_2)^2 \quad (n_1+n_2 - 1)}$$

$$\mu = \frac{2\ (18)\ (14)}{18 + 14} + 1$$

$$\sigma^2 = \frac{2(18)\ (14)\ \{\ 2\ (18)\ (14) - 18 - 14\}}{(18+14) \quad (18+14 - 1)}$$

entonces

$$\mu = 16.75$$

$$\sigma^2 = 7.49$$

$$F(X) = p(X \leq 10) = \Phi\left(\frac{10-16.75+.5}{7.49}\right)$$

$$= \Phi(-2.28) = 1 - (2.28) = 1 - .9887$$

$$= .0113$$

Fijando un nivel de significancia del .05, entonces rechazamos la idea de que el arreglo es aleatorio, pero si fijamos un nivel de .005, entonces no tenemos argumentos para rechazar, esto está en función del riesgo que se quiera correr.

Por otro lado contrastaremos las siguientes hipótesis:

H_0: No existe un número demasiado grande rachas.

H : Existe un número demasiado grande de rachas.

$$P(X > 10) = 1 - P(X \leq 10) = 1 - .0113 = .9887$$

entonces si rechazamos H_0 tenemos una probabilidad de error de 98.8%, si además determinamos un nivel de significancia de .05 entonces no rechazaremos la hipótesis nula.

Por lo tanto se rechaza que haya muchas rachas en los datos.

Continuando analizaremos los puntos influyentes y discrepantes, mediante la gráfica No. 41 de los residuales

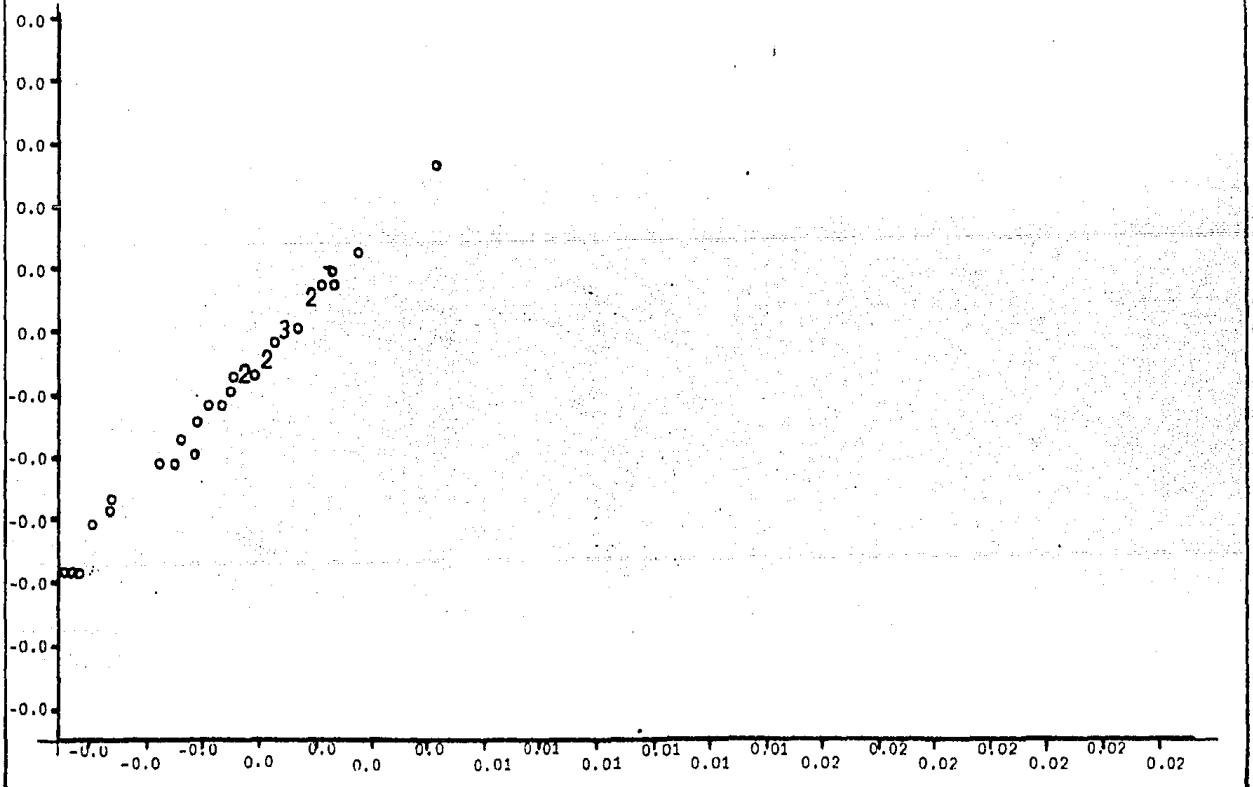
menos contra residuales, en ella se observa que existen un punto cuyo comportamiento se aleja bastante de todo el conjunto, por lo cual tenemos sospecha de que existe una observación discrepante y para verificarlo aplicaremos la técnica analítica que ya fue expuesta en este capítulo.

La observación No. 3 tiene $t_3 = 2.597684$, siendo este el valor más grande dentro del grupo, por lo cual es candidato a ser una observación discrepante, este estadístico tiene 27 grados de libertad y en tablas encontramos que el valor crítico para un $\alpha = .01$ es 2.77, claramente t_3 no excede este valor, por lo tanto no se tiene evidencia de que la observación número 3 sea discrepante.

El mayor valor en distancia de Cook es el de la observación número tres, si este caso fuera omitido del análisis el estimador del vector beta sufriría un movimiento equivalente a trasladar el estimador inicial a un elipsoide de 21.81% de confianza. Tomando en cuenta que $F(.05,4,28) = 2.71$, entonces nos inclinamos a pensar que existen observaciones que presenten una influencia elevada en el modelo.

Se concluye este capítulo con la necesidad de analizar y comparar todos los modelos expuestos anteriormente para seleccionar el que sea más conveniente.

GRAFICA Nº 41 RESIDUALES MENOS CONTRA RESIDUALES



T A B L A 3.3

VECTOR *YT*	VECTOR *Y*-ESTIMADA	VECTOR RESIDUAL	RESIDUALES STUDENTIZADOS	*D*(COOK)	*T*(COUTLIER)
1.861161	1.865316	-0.004155	-1.709540	0.220572	-1.773860
1.864350	1.865641	-0.001292	-0.552740	0.031212	-0.545766
1.878444	1.872713	0.005731	2.366146	0.435335	2.597684
1.878815	1.875141	0.003674	1.460965	0.115396	1.492659
1.880554	1.878160	0.002394	0.960325	0.054682	0.958945
1.881505	1.882537	-0.001032	-0.386935	0.003041	-0.380983
1.882247	1.886271	-0.004024	-1.498593	0.037277	-1.534410
1.882487	1.886998	-0.004511	-1.814803	0.201625	-1.897169
1.887706	1.890882	-0.003176	-1.185557	0.025199	-1.194562
1.892384	1.892874	-0.000490	-0.185314	0.000857	-0.182086
1.893192	1.893942	-0.000751	-0.281446	0.001603	-0.276766
1.899287	1.896049	0.003238	1.240165	0.049194	1.252708
1.900322	1.897224	0.003098	1.179514	0.039847	1.188154
1.901489	1.899543	0.001945	0.739402	0.015209	0.733272
1.902793	1.902167	0.000626	0.234318	0.001055	0.230322
1.905001	1.903318	0.001683	0.634000	0.009230	0.627093
1.905933	1.904128	0.001805	0.682668	0.011694	0.676016
1.906456	1.906822	-0.000366	-0.139880	0.000600	-0.137407
1.906755	1.906162	0.000592	0.222374	0.001031	0.218560
1.908102	1.910415	-0.002313	-0.864185	0.013701	-0.860162
1.909730	1.911812	-0.002082	-0.777731	0.011078	-0.772102
1.910159	1.915115	-0.004955	-1.906917	-0.126301	2.007439
1.916066	1.915934	0.000132	0.049872	0.000057	0.048976
1.918943	1.917194	0.001748	0.663929	0.012052	0.657159
1.919772	1.919387	0.000385	0.148655	0.000815	0.146034
1.921080	1.918736	0.002344	0.909278	0.032491	0.906376
1.921693	1.921624	0.000069	0.028279	0.000060	0.027770
1.922501	1.921020	0.001481	0.569336	0.011108	0.562342
1.923872	1.923758	0.000114	0.043752	0.000062	0.042965
1.924252	1.925682	-0.001430	-0.549200	0.010131	-0.542232
1.925061	1.928275	-0.003214	-1.241402	0.056729	-1.254031
1.932627	1.929896	0.002731	1.059276	0.044000	1.061679

CAPITULO IV

ANALISIS DE LOS RESULTADOS

i) Aspectos Relevantes del Tratamiento de los Modelos Ajustados

En el modelo inicial 1.1 se cuenta con nueve variables independientes y una dependiente las cuales las vamos a abreviar de la siguiente manera:

Costo de Operación	= *CO*
Fecha de Evaluación	= *FE*
Seguro y Renta de Aviones	= *SR*
Sobrecargos y Pilotos	= *SP*
Manteniemiento (materiales)	= *MM*
Nuevas Rutas	= *NR*
Gas subsidiado	= *GS*
Compra de nuevas naves	= *CN*
Renta de al menos un Avión	= *RA*
Salarios Variables	= *SV*

Estas nueve variables explicativas fueron trabajadas todas en conjunto, calculando los estimadores de los parámetros en cuestión. Las hipótesis fuertes en el modelo 1.1 ($E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2 I_n$) son de suma impor

tancia, para la estimación insesgada de los parámetros.

Es importante aclarar que en todas las ecuaciones de regresión múltiple, el $\hat{\beta}_j$ $j = 0,\ldots,9$ está ajustado por todas las otras variables del modelo.

El primer modelo ajustado es de la siguiente forma:

$$Co = -2555.73 + 31.74\,FE - SR + .48\,SP + .05\,MM + 39.83\,NR + 15.22\,GS - 25.10\,CN + .26\,RA + 22.83\,SV$$

Las variables explicativas tienden a modificarse en el momento que alguna variable fundamental es cambiada, es decir, si se compra una nave, las variables SV, MM, SP y SR sufren cambios fuertes o débiles, dependiendo del tipo de Aeronave que se adquiera.

Al calcular las varianzas estimadas de los parámetros estimados, se observa directamente que la varianza estimada en beta cero estimado difiere en más de 100 000 unidades respecto a la mayor varianza estimada de cualquier beta estimado. El hecho de esta varianza tan grande podría sugerir algúna transformación a la variable respuesta y.

La varianza estimada en el modelo 1.1 es de 2528.173, la cual depende de las suposiciones concernientes a

los errores, es decir, si cualquiera de las hipótesis (Independencia, media cero, varianza común) no se cumplen, entonces este estadístico no debe de ser usado como escala de dispersión. Por ejemplo, la no independencia puede propiciar un estimador demasiado pequeño o demasiado grande.

Antes de examinar la tabla 2.3 de análisis de varianza, examinaremos la tabla 2.4, en la cual se contrastan las siguientes hipótesis H_0: no hay carencia de ajuste contra H_1: hay carencia de ajuste.

Es importante aclarar que en las observaciones no existen repeticiones por lo cual se aplicó la prueba Arco Iris (UTTS 1982). En esta tabla a un nivel de significancia de .05 se rechazó la hipótesis nula, sin embargo a un nivel de .01 no se cuenta con argumentos suficientes para rechazar H_0: no hay carencia de ajuste. Para este contraste de hipótesis se concluye que el rechazo o el no rechazo depende del riesgo que se requiera correr.

Bajo el nivel de significancia .05 es obsoleta la tabla 2.3 de análisis de varianza, puesto que se tiene evidencia que existe carencia de ajuste en el modelo ajustado.

El modelo ajustado 1.1 tiene de coeficiente de correlación .9685 el cual consideramos que es bastante elevado. Aplicando la relación que existe con el coeficiente de determinación podemos afirmar que el 93.8% de la variabilidad observada en la variable dependiente es explicada por el conjunto de variables independientes.

Los residuales son útiles dentro del análisis de regresión, ya que al ser graficados contra la variable independiente estimada, nos proporcionan información acerca de las fallas en las suposiciones, de las transoformaciones al parecer adecuadas, y si existe varianza no constante. También los residuales se grafican generalmente contra cada una de las variables independientes proporcionando información acerca de sus tendencias, sugiriendo algúna transformación factible.

En las gráficas 6, 8, 9, 10, 11 y 12 del capítulo II se observan comportamientos periódicos y la necesidad de transformar algunas variables.

La gráfica No. 6 al parecer requiere de una transformación senoidal o una de cuarto grado. Las gráficas No. 8 y No. 9, se les puede aplicar una transformación exponencial con el propósito de eliminar la no-lineabilidad y estabilizar la varianza. La gráfica No. 11 no exhibe ningún comportamiento claro y sería bastante aventurado -

proponer algúna modficación.

En todas las gráficas se observan puntos que son candidatos a ser discrepantes. Por otro lado la carencia de ajuste se puede deber a que existen puntos discrepantes en el modelo.

Las gráficas de residuales estudentizados contra las variables explicativas, nos proporcionan también información acerca de las suposiciones hechas a los errores en el planteamiento del modelo propuesto.

La gráfica No. 13:muestra una distribución de puntos muy cercanos a la línea $r_m = r$, lo cual indica no tener indicios de observaciones discrepantes, salvo una observación - que consideramos se dispara del conjunto de datos.

La gráfica No. 14: presenta una distribución de puntos que nos conducen a sospechar que los errores no se distribuyen normalmente.

En resumen el modelo 1.1 no es adecuado y requiere de alguna transformación, debido a todas las características que se presentan en el análisis desarrollado.

En el capítulo número tres, visualizamos que la matriz X tiene el problema de multicolinealidad, es decir, se tiene dependencia lineal (al menos aproximada) entre dos o más variables independientes.

Es importante hacer notar que existen dos tipos de dependencia lineal, la numérica y la estadística. La dependencia lineal estadísticamente severa puede serlo mucho menos que la dependencia lineal que determina el mal condicionamiento numérico, el mal-condicionamiento depende de la -- precisión de la computadora y de los algoritmos usados.

Al tener la matriz X el problema de multicolinealidad, propicia que los predictores o las varianzas de los estimadores sean muy grandes, por lo que se procedió a hacer selección de variables y de acuerdo a los criterios R^2 $\overline{R^2}$ y C_p se escogió el modelo 3.5 $Y = \beta_0 + \beta_1 X_1 + \beta_5 X_5 + \beta_9 X_9 + \varepsilon$

Esta expresión asume que:

$$E(\varepsilon) = 0 \; ; \; Var(\varepsilon) = \sigma^2 I n$$

Compararemos ahora el modelo 1.1 con el modelo 3.5, respecto a:

- La condición K
- Estimadores de los parámetros
- Varianzas de los estimadores
- Varianza estimada
- Tablas de análisis de varianza
- Tablas de análisis de varianza (carencia de ajuste)
- Coeficiente de determinación
- Puntos influyentes y discrepantes

MODELO 1.1		MODELO 3.5	
Condición K = No converge		Condición K = 9,8535	
Estimación de Betas	Varianzas	Estimacion de Betas	Varianzas
-2555.72	1471004.79	-1944.71	1172942.67
31.74	262.29	24.10	209.51
39.83	698.39	40.50	378.53
22.83	56.05	27.34	38.02
Varianza estimada = 2528.173		Varianza estimada = 2304.436	
Coeficiente de Determinación = .938		Coeficiente de Determinación = .928	

En el modelo 1.1 la condición K no converge mientras que en el modelo 3.5, la condición K = 9.85, lo que indica una multicolinealidad modesta.

La varianza estimada de $\hat{\beta}_0$ en el modelo (3.5) es bastante menor, sin embargo es aproximadamente 600 veces el estimador, mientras que en el modelo (1.1) la varianza estimada es aproximadamente 575 veces el estimador.

La varianza estimada de $\hat{\beta}_5$ en el modelo (3.5) es menor, y además es 9 veces el estimador, mientras que en el modelo (1.1) la varianza estimada es 18 veces el estimador.

Los estimadores de β_1 y β_9 no sufren cambios significativos en relación a sus respectivas varianzas estimadas.

La varianza estimada del modelo (1.1) difiere en 224 unidades respecto a la varianza estimada del modelo reducido (3.5).

El coeficiente de correlación se reduce en .517% al eliminar las variables independientes X_8, X_7, X_6, X_4, X_3 y X_2 siendo una pérdida bastante pequeña.

En ambos modelos se rechaza la hipótesis nula -- H_o: $\underline{\beta} = 0$, sin embargo la F_c del modelo 1.1 es 83 unidades menor que la F_c del modelo 3.5, por lo cual a niveles de significancia bastante reducidos se seguirá rechazando la hipótesis H_o: $\underline{\beta} = 0$ del modelo 3.5.

Bajo el modelo 1.1 se obtuvo evidencia de ca-- rencia de ajuste en el modelo ajustado, mientras que en - el modelo (3.5) no se tienen argumentos para rechazar - H_o: no hay carencia de ajuste. La F_c del modelo 3.5 es 13 veces menor que la del otro modelo.

En ambos modelos (1.1 y 3.5) se tiene evidencia de que existen puntos discrepantes (observación No. 32). Bajo el modelo 3.5 se tiene que la observación -- No. 32 es influyente puesto que si este caso fuera omitido del análisis el estimador del vector beta sufriría un movimiento equivalente a trasladar el estimador inicial a un elipsóide de 32.35% de confianza, mientras que bajo -

el modelo 1.1 el eliminar dicha observación propiciaría un cambio en los estimadores del 11.99%.

Al aplicar las técnicas (gráficas de residuales estudiantizados en papel normal, Durbin-Watson, Wald-Wolfwitz) a los residuales del modelo 3.5 se obtuvo que se puede tomar la hipótesis de normalidad en los errores, -- además bajo la prueba D-W se obtuvo que existía correlación positiva, sin embargo en la prueba de Wald-Wolfowitz se concluyó que no existían indicios de correlación se--- rial, este tipo de contradicciones se puede deber a la robustez de los métodos.

Con todo lo anterior se ve que el modelo 3.5 se comporta mucho mejor que el modelo 1.1. Sólo llama - la atención el caso No. 32, que resulta moderadamente influyente.

La necesidad de transformar al modelo 3.5 es - motivada por los siguientes argumentos: las varianzas de los estimadores se disparan mucho entre sí, tanto residuales estudentizados como residuales parciales proporcionan información acerca de transformaciones y existencia de -- una observación influyente.

Al aplicar el método Box y Cox se encontró que la transformación a la variable dependiente más adecuada era:

$$YT = (2 - (2/(Y^{**}0.5)))$$

dando lugar al modelo transformado 3.14, el cual compararemos con el modelo 3.5 respecto a todas sus características.

MODELO 3.5		MODELO 3.14	
Condición $K = 9.8535$		Condición $K = 9.8535$	
Estimacion de Betas	Varianzas	Estimación de Betas	Varianzas
-1944.71	1172942.67	1.3931	0.0037
24.10	209.51	0.0059	$6.5281X10^{-7}$
40.50	378.53	-0.0008	$1.1794X10^{-6}$
27.34	38.02	0.0020	$1.1849X10^{-7}$
Varianza estimada = 2304.436		Varianza estimada = 7.180×10^{-6}	
Coeficiente de determinación = .928		Coeficiente de determinación = .981	

La condición K para ambos modelos es la misma, debido a que no se realizó ningúna transformación a las variables explicativas.

La varianza estimada de $\hat{\beta}_0$ del modelo 3.14 es .0027 veces el estimador, mientras que en el modelo 3.5 la varianza estimada es 603 veces el estimador, es decir se obtuvo ganancia al transformar el modelo, en lo que respecta a este estimador.

La varianza estimada de $\hat{\beta}_1$ del modelo 3.5 es 8.69 veces el estimador, mientras que en el modelo 3.14 la varianza estimada es .0001 veces el estimador.

La varianza estimada de $\hat{\beta}_5$ del modelo 3.5 es 9.35 veces el estimador, mientras que en el modelo 3.14 la varianza estimada es .0015 veces el estimador.

La varianza estimada de $\hat{\beta}_9$ del modelo 3.5 es 1.39 veces el estimador, mientras que en el modelo 3.14 la varianza estimada es .00006 veces el estimador.

Es claro que al transformar se obtuvieron ganancias considerables, en lo que respecta a estimadores.

La varianza estimada del modelo 3.14 es ------- .000000003 veces la varianza estimada del modelo 3.5, siendo una ganancia estupenda.

Al transformar el modelo 3.5 se obtiene el ---- 2.74% de aumento en el coeficiente de correlación. Por lo que el modelo 3.14 cuenta con un coeficiente de correlación bastante elevado (r = .99).

Los modelos 3.5 y 3'14 en lo que respecta a la tabla de análisis de varianza, siguen rechazando la hipótesis nula H_o: $\underline{\beta} = 0$, más aún la Fc del modelo 3.5 es 374 unidades menor que la Fc del modelo 3.14, --- Es decir, si el nivel de significancia es sumamente --

pequeño, la hipótesis nula se seguirá rechazando, bajo el modelo 3.14.

En ambos modelos no se tienen argumentos para rechazar H_o: no hay carencia de ajuste, aunque la Fc de la tabla de análisis de varianza del modelo 3.14 aumentó .16 unidades.

En lo que respecta al análisis de residuales del modelo 3.14 se tiene que éstos se comportan bastante bien, sin embargo se tienen sospechas que la variable X_9 puede requerir de alguna transformación.

Al graficar a los residuales studentizados en papel normal, se obtuvieron resultados positivos, por lo cual se puede llevar a cabo el cálculo de intervalos de confianza.

Al aplicarle la prueba de rachas al modelo 3.14 se obtuvo bajo un nivel de significancia de .01 argumentos suficientes para no rechazar la aleatoriedad de los residuos, por lo que se puede suponer no-correlación en los errores.

Es de interés hacer notar que ningún valor del vector de residuales se dispara de los demás y esto puede sugerir que las suposiciones concernientes a los errores son correctas.

En este modelo (3.14) se tiene evidencia que no existen puntos influyentes y discrepantes.

Por todo lo anterior se puede afirmar que el modelo 3.14 es mejor que el modelo 3.5.

Para el modelo 3.5 se calcularon diferentes intervalos de confianza, los cuales fueron expuestos con cierto cuidado en el capítulo III, es de interés hacer notar que en el intervalo de confianza para Y/X_0 (Intervalo de predicción) el valor del centro es la predicción.

Aun siendo el modelo 3.5 uno de los mejores modelos, se le aplicó la técnica de Box y Tidwell, la cual propone una familia de transformaciones para las variables explicativas. Encontrándose que era conveniente --- transformar la variable $X_9^{1.5}$.

Por último realizaremos comparaciones entre los modelos 3.14 y 3.16 respecto a todas sus características al ser ajustados para determinar la ganancia al cambiar de modelo.

MODELO 3.14		MODELO 3.16	
Estimación de Betas	Varianzas	Estimación de Betas	Varianzas
1.3931	0.0037	1.3380	0.0030
0.0059	6.528×10^{-7}	0.0068	5.0048×10^{-7}
-0.0008	1.1794×10^{-6}	-0.0015	1.3662×10^{-6}
0.0020	1.1849×10^{-7}	5.36×10^{-5}	9.3635×10^{-11}
Varianza estimada = 7.18×10^{-6}		Varianza estimada = 7.691×10^{-6}	
Condición K = 9.8535		Condición K = 5.3476	
Coeficiente de determinación = .981		Coeficiente de determinación = .98	

Las varianzas estimadas de los betas "gorros" del modelo 3.16 son muy semejantes a las del modelo 3.14. La varianza estimada en los modelos 3.14 y 3.16 son prácticamente iguales.

La condición K disminuyó aproximadamente en 4.5 unidades al transformar al modelo 3.14, es decir, se cuenta ahora con una multicolinealidad más modesta, pero con un modelo más complejo.

En lo que respecta al coeficiente de determinación, en ambos modelos es relativamente el mismo.

Los modelos (3.14 y 3.16) en sus respectivas tablas de análisis de variaza, rechazan la hipótesis nula H_0: $\beta_1 = \beta_5 = \beta_9 = 0$ con una F calculada muy similar.

En el modelo 3.16 los diagramas de residuales estudentizados contra cada una de las variables se comportan bastante bien, sin embargo se sospecha que existen -- puntos discrepantes.

La gráfica de residuales estudentizados en papel normal tiene un comportamiento bueno, por lo cual se puede aceptar la hipótesis de que los errores se distribuyen normalmente.

Bajo el modelo 3.16 se aplicó la prueba de Wald-Welfowitz y considerando niveles de significancia pequeños se obtuvieron argumentos suficientes para no rechazar la - aleatoriedad de los residuos, por lo que se puede afirmar que existe evidencia de no-correlación en los errores.

También en el modelo (3.16), se encontró que no - hay observaciones discrepantes o que presenten una influencia contundente.

Después de haber realizado el análisis de los diferentes modelos, por el que más nos inclinamos es el 3.14, considerándolo uno de los mejores, más no el mejor, por ser de gran sencillez y fácil de interpretar.

CONCLUSIONES

Las conclusiones que a continuación se presentan será divididas en cuatro incisos puesto que en este trabajo se analizaron varios modelos, con el propósito de encontrar el más adecuado.

Para dictaminar cuál modelo es uno de los más convenientes, se requirió de la aplicación de las siguientes técnicas.

- Mínimos cuadrados.
- Tablas de Análisis de varianza
- Tablas de análisis de varianza (carencia de ajuste)
- Coeficiente de determinación y correlación
- Condición del número K
- Varianza estimada
- Gráficas de residuales estudentizados, parciales, estudentizados en papel normal, residuales menos y además gráficas de máxima verosimilitud y de dispersión.
- Pruebas de autocorrelación
- Transformaciones
- Prueba sobre puntos discrepantes e influyentes

Todas estas son argumentos para definir la expresión que explique convenientemente a la variable respues

ta (Costo de Operación).

a) Para el modelo 1.1 se encontraron deficiencias tales como: excesivo número de variables aleatorias;evidencia de carencia de ajuste en el modelo; necesidad de realizar transformaciones, problemas de multicolinealidad y existencia de observaciones discrepantes. Como consecuencia se tuvo que continuar con el modelaje (Inciso b).

b) En el modelo 3.5 se encontraron ventajas al aplicar selección de variables, tales como: obtención de multicolinealidad modesta, argumentos para no rechazar la no carencia de ajuste, no rechazo de aleatoriedad de los residuos y el trabajar con un modelo reducido. Las desventajas que encontramos son: la necesidad de transformar a la variable dependiente, comportamientos periódicos al visualizar las gráficas de residuales estudentizados contra variables explicativas y la evidencia de que la observación No. 32 es discrepante.

c) El modelo 3.14 tiene una expresión explicativa que proporciona bastante información acerca de la variable respuesta transformada. la tranfor

mación que se efectuó al modelo 3.5 fue sugerida al aplicar el método de Box y Cox, obteniéndose significativas ganancias:
en la estimación de los parámetros respecto a sus varianzas, en el coeficiente de correlación, en las tablas de análisis de varianza, e indicios de que el modelo es correcto.

Por otro lado se encontró que el modelo transformado al .05 de nivel de significancia se rechaza la hipótesis de que los errores están correlacionados positivamente, por lo tanto el error Tipo I (Rechazar H_o siendo verdadera) se comete en proporción de 1 a 20. Sin embargo, si tomamos un nivel de significancia de .01, no hay argumentos para rechazar aleatoriedad en los residuos, cometiendo un error Tipo I de 1 en 100, es decir rechazaríamos acertadamente la hipótesis alternativa 99 veces de 100. es de interés hacer notar que la potencia de la prueba decrece cuando el nivel de significancia disminuye. En nuestro enfoque particular es mas peligroso rechazar la hipótesis nula cuando es verdadera. Además, para este modelo se obtuvo evidencia de no existir puntos influyen

tes ni discrepantes.

En lo que respecta a las variables explicativas, se observó que en particular la variable X_9 presenta, al parecer, un comportamiento de segundo grado. En general este modelo se comporta bastante bien.

d) Por último llevamos a cabo una transformación a la variable X_9, la cual fue sugerida al aplicar el método de Box y Tidwell (Modelo 3.16). Al efectuar la transformación se obtuvo ganancia en lo que respecta a la multicolinealidad del modelo y pérdidas muy ligeras en el coeficiente de correlación y en los diagramas de residuales estudentizados contra variables explicativas.

Finalmente, analizando todos los resultados y características obtenidas en los modelos, nos inclinamos por el siguiente:

$$\frac{(Co)^{-1/2} - 1}{-\frac{1}{2}} = \beta_0 + FE\beta_1 + NR\beta_5 + SV\beta_9$$

Bajo

$$E(\varepsilon) = 0 \; ; \; Var(\varepsilon) = \sigma^2 I_n \; y \; Cov(\varepsilon i, \varepsilon j) = 0 \quad i, j \; i \neq j$$

APENDICE I

DEMOSTRACIONES

1.2 Demuestre que $Var(\hat{\beta}) = \sigma^2(XX)^1$ bajo las hipótesis

$$Var(\varepsilon) = \sigma^2 In \text{ y } E(\varepsilon) = 0$$

Para demostrar lo anterior es necesario mostrar que

$$Var(\ AX + B\ ) = A\ Var\ A'$$

$Var(AX+B) = D(AX+B) = D(AX) + D(B) = D(AX)$

$Var(AX+B) = Cov(AX,AX) = A\ Cov(X,X)\ A' = A\ D(X)\ A'$

$Var(AX+B) = A\ Var\ A'$

donde D(X) es la matriz de dispersión.

Con esta demostración paso a probar la expresión - inicial donde es el vector estimado de entonces tomando $\hat{\beta} = (X'X)^{-1}X'Y$

Sustituyendo

$$Var(\hat{\beta}) = Var[\ (X'X)^{-1}X'Y\ ]$$

Ahora tomando la ecuación (2,3)

$Var(\hat{\beta}) = Var[\ (X'X)^{-1}X'Y\ ] = Var[\ (X'X)^{-1}X'X + (X'X)^{-1}X'\varepsilon]$

$= Var[In + (X'X)^{-1}X'\ ] = Var[\ In\beta] + Var[(X'X)^{-1}X'\varepsilon]$

$= 0 + Var[\ (X'X)^{-1}X'\varepsilon]$

Por la demostración anterior obtenemos que

$$\text{Var}(\hat{\beta}) = (X'X)^{-1}X' \text{ Var}(\varepsilon) X(X'X)^{-1}$$

$$= \sigma^2 (X'X)^{-1}X'X (X'X)^{-1}$$

$$= \sigma^2 (X'X)^{-1}$$

2.2 Demostraremos que $\hat{\sigma}^2$ es un estimador insesgado de σ^2

Si $E(Y) = X\beta$ X de rango P

y $\text{Var}(Y) = \sigma^2 I_n$ entonces

$\hat{\sigma}^2 = \frac{SCR}{n-p}$ es el estimador insesgado

$$\hat{\sigma}^2 = \frac{(Y-X\beta)'(Y-X\beta)}{n-p}$$

Demostración. Puesto que $\hat{\beta} = (X'X)^{-} X'Y$

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

Sea $H = X(X'X)^{-1}X'$

$$Y - X\hat{\beta} = Y - HY = (I_n - H)Y \quad \ldots\ldots\ldots\ldots (1)$$

Pasamos primero a demostrar que H es idempotente, para ser idempotente tiene que ocurrir que HH = H

$$H = X(X'X)^{-1}X'$$

$$HH = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = H$$

Ahora pasamos a demostrar que $(I_n - H)$ es idempo -

tente

$\rightarrow$ $(In-H)\ (In-H) = In^2 - InH - HIn + H^2 = In - 2H + H...$

Puesto que H es idempotente

$$In - 2H + H = In - H \qquad (2)$$

Después de haber demostrado lo anterior, pasamos a probar que $\hat{\sigma}^2$ es un estimador insesgado de σ^2

$SCR = (Y-X\beta)^1(Y-X\hat{\beta}) = Y^1(In-H)^1(In-H)Y$ Por (1)

$= Y^1(In-H)Y$ Por (2)

$E(SCR) = E[Y^1(In-H)Y] = \sigma^2 Tr(In-H) + \underbrace{\beta^1X(In-H)X\beta}_{\sim 0}$

$\rightarrow E(SCR) = \sigma^2(n-p^1)$

$\rightarrow E[SCR/n-p^1] = \sigma^2$

3.2 Probar que $\frac{\hat{\sigma}^2}{\sigma^2} \sim X^2(n-p^1)$

La hipótesis necesaria es que $\varepsilon \sim N(0,\sigma^2 In)$ Además se requiere de la demostración de que $E(\hat{\beta}) = \beta$, es decir, que $\hat{\beta}$ sea un estimador insesgado de β

$E(\hat{\beta}) = E[(X^1X)^{-1}X^1Y] = (X^1X)^{-1}X^1\ E(Y) = (X^1X)^{-1}X^1X\ \beta$

$\rightarrow E(\hat{\beta}) = In\ \beta = \beta \qquad (3)$

Con la demostración anterior, paso a demostrar la expresión inicial $\frac{\hat{\sigma}^2}{\sigma^2} \sim X^2(n-p^1)$

$SCR = (Y - X\hat{\beta})^1\ (Y - X\hat{\beta}) = Y^1\ (In - H)\ Y$

$E(Y-X\hat{\beta}) = E(Y) = E(X\hat{\beta}) = X\beta - X\beta = 0$ Por (3)

$Var\ (Y-X\hat{\beta}) = Var\ (Y-X\hat{\beta}) = Var\ [(In-H)Y]$

$= (In-H)\ Var\ (Y)\ (In-H)'$

$= \sigma^2\ (In-H)\ In\ (In-H)'$

$= \sigma^2\ (In-H)$

bajo la suposición de normalidad de los errores

$\therefore (Y-X\hat{\beta}) \sim N\ (0, \sigma^2 In-H)$

$\frac{\hat{\sigma}^2}{\sigma^2} \sim X\ (n-p^1)$

5.2 Demostraremos que:

$esa(\hat{y}/x) = \hat{\sigma}\ [x^1(X^1X)^{-1}x]^{\frac{1}{2}}$

Como $\hat{y} = x^1\hat{\beta}$

$\therefore Var\ (\hat{y}) = Var\ (\beta) = x^1\ Var\ (\hat{\beta})\ x$

Por la demostración número uno, obtenemos

$Var\ (\hat{y}) = \sigma^2 x^1(X^1X)^{-1}x$

La varianza estimada de $\hat{y}$ es

$\hat{Var}\ (\hat{y}) = \sigma^2 x^1(X^1X)^{-1}\ x$

Por lo anterior y por definición de error estándar.

$esa\ (\hat{y}/x) = [\hat{Var}\ (\hat{y})]^{\frac{1}{2}}$

Suponemos que $E(\varepsilon) = 0$ $V(\varepsilon) = \sigma^2$ y que ε es independ-

diente de $x^1\hat{\beta}$, además necesitamos calcular la varianza de $\hat{y}$

Sea $\tilde{y} = x^1\beta + \tilde{\varepsilon}$

$\rightarrow$ $Var\ (\tilde{y}) = Var(x^1\hat{\beta}+\tilde{\varepsilon}) = Var\ (x^1\hat{\beta}) + Var\ (\tilde{\varepsilon})$

$Var\ (\tilde{y}) = \sigma^2 x^1(X^1X)^{-1}x + \sigma^2$

$= \sigma^2[x^1(X^1X)^{-1}x + 1] = \sigma^2[1+x^1(X^1X)^{-1}x]$

$\rightarrow$ $Var\ (\tilde{y}) = \sigma^2[1 + x^1(X^1X)^{-1}x]$

Por definición de error estándar tenemos que

$esp\ (\tilde{y}/x) = \hat{Var}\ (\tilde{y})^{\frac{1}{2}} = \hat{\sigma}[1 + x^1(X^1X)^{-1}x]^{\frac{1}{2}}$

7.2 Demostrar que $R^2 = r^2$

$$R^2 = \frac{SCreg}{SYY} = \frac{SYY - SCR}{SYY}$$

donde $SYY = Y^1Y - n\bar{Y}^2$ $\quad SCR = Y^1Y - \hat{\beta}^1X^1Y$

como $SCR = SYY - \frac{(SXY)^2}{SXX}$

$$R^2 = \frac{SYY - SYY + \frac{(SXY)^2}{SXX}}{SYY} = \frac{(SXY)^2}{SXXSYY}$$

donde SXY es la covarianza de $(Y,\hat{Y})$ y SXX es la varianza de $\hat{Y}$.

TABLA DE DIVISORES PARA INTERVALOS DE CONFIANZA DE EXTENSION MINIMA.

Coeficiente de Confianza = , Grados de Libertad = n, a_n es el número superior y b_n el inferior.

n	ε .900	.950	.990	.995	.999	n	ε .900	.950	.990	.995	.999
2	.2104	.1025	.0201	.0100	.0020	16	9.0440	7.8043	5.7559	5.1040	3.9248
	18.0017	21.4812	29.1362	32.3240	39.5708		31.5125	34.6197	41.1710	43.7951	49.5766
3	.5821	.3513	.1148	.0717	.0244	17	9.7883	8.4947	6.3425	5.6523	4.3954
	17.6381	20.7437	27.5102	30.3027	36.5959		32.7139	35.8560	42.4728	45.1206	50.0511
4	1.0561	.7082	.2969	.2069	.0908	18	10.5385	9.1932	6.9402	6.2128	4.8800
	18.1062	21.0632	27.4603	30.0848	35.9845		33.9148	37.0919	43.7748	46.4465	52.3245
5	1.5938	1.1392	.5534	.4113	.2102	19	11.2947	9.8991	7.5481	6.7846	5.3786
	18.9081	21.8001	28.0209	30.5697	36.2654		35.1148	38.3271	45.0765	47.7723	53.6990
6	2.1750	1.6233	.8760	.6747	.3806	20	12.0563	10.6119	8.1654	7.3666	5.8882
	10.8739	22.7410	28.8928	31.3966	36.9947		36.3137	39.5611	46.3772	49.0774	55.0743
7	2.7883	2.1473	1.2350	.9871	.5979	21	12.8230	11.3310	8.7915	7.9580	6.4085
	20.9303	23.7944	29.9229	32.4106	37.9541		37.5112	40.7936	47.6767	50.4216	56.4507
8	3.4262	2.7027	1.6397	1.3406	.8560	22	13.5946	12.0561	9.4259	8.5588	6.9406
	22.0405	24.9147	31.0507	33.5358	39.0631		38.7070	42.0243	48.9736	51.7426	57.8190
9	4.0840	3.2836	2.0775	1.7288	1.1499	23	14.3706	12.7868	10.0679	9.1679	7.4824
	23.1844	26.0769	32.2397	34.7308	40.2631		39.9011	43.2532	50.2686	53.0616	59.1857
10	4.7584	3.8855	2.5434	2.1469	1.4755	24	15.1508	13.5227	10.7169	9.7845	8.0322
	24.3498	27.2662	33.4675	35.9714	41.5223		41.0935	44.4802	51.5619	54.3793	60.5545
11	5.4467	4.5054	3.0334	2.5906	1.8287	25	15.0351	14.2636	11.3728	10.4088	8.5919
	25.5294	28.4733	34.7240	37.2430	42.8238		42.2840	45.7051	52.8521	55.6935	61.9157
12	6.1472	5.1409	3.5447	3.0573	2.2078	26	16.7230	15.0090	12.0348	11.0396	9.1580
	26.7180	29.6920	35.9963	38.5330	44.1445		43.4728	46.9281	54.1407	57.0065	63.2808
13	6.8583	5.7899	4.0744	3.5439	2.6086	27	17.5145	15.7587	12.7024	11.6774	9.7293
	27.9126	30.9184	37.2869	39.8378	45.4880		44.6598	48.1491	55.4277	58.3186	64.6514
14	7.5788	6.4510	4.6205	4.0483	3.0296	28	18.3095	16.5128	13.3767	12.3211	10.3146
	29.1109	32.1497	38.5733	41.1517	46.8441		45.8446	49.3675	56.7096	59.6230	65.9955
15	8.3078	7.1227	5.1813	4.5685	3.4676	29	19.1076	17.2706	14.0554	12.9699	10.9003
	30.3113	33.3842	39.8715	42.4732	48.2150		47.0279	50.5813	57.9914	60.9295	67.3589

APENDICE II

El programa Simulación y Análisis de Regresión es interactivo y fue diseñado para brindar apoyo en docencia e investigación aplicada. SAREG posee una gran habilidad para el manejo de datos, y no requiere que el usuario tenga un conocimiento profundo en cómputo. El programa realiza el ajuste de un modelo de regresión lineal mediante técnicas numéricas modernas, por lo que proporciona precisión en los resultados.

La forma de comunicarse con el programa es mediante -- COMANDOS, uno de los cuales explica cuantos están disponibles para el usuario, y que operaciones realiza cada uno de ellos.

El programa está escrito en lenguaje ALGOL, y fue diseñado para ser usado por personas con conocimientos de análisis de regresión; sin importar el campo específico de su actividad.

BIBLIOGRAFIA

- Anscombe, F.J. (1973) "Graphs in Statistical Analysis" Amer. Statist., 27, 17-21

- Box, G.E.P. and D.P. Cox (1964) "An analysis of transformations (with discussion)". J.R. statist. Soc. Ser. B, 26, 211-246.

- Box, G.E.P. and P.W. Tidwell (1962). "Transformation of the independient variables". Technometrics, 4, 531-550.

- Draper, N.R. and H. Smith (1967). "Applied Regression Analysis". New York: Wiley TP. 278.

- Francisco Aranda Ordaz y Silvia Ruíz Velasco A. (1984) "Sareg". Un programa interatictivo para simulación y análisis de problemas de regresión. Comunicación Técnicas. IIMAS.

- Folleto, "Hoy compromiso con México", Aeroméxico 1984.

- Folleto, "50 Aniversario", Aeroméxico 1934-1984.

- G.M. Furnival y R.W. Wilson, Techometrics, Vol. 16 1974, pp. 499-511.

- Larsen, W.A. and S.A., McCleary (1972) "The use of partial residual plots in regresion analysis "Technometrics, 14, 781-790.

- Mallows C.L. (1973) "Some comments on Cp". Technometrics 15, 661-676.

- Mary L. Thompson (1978), International Stadistical Review, pp. 1-19 y 129-148.

- Regression By Leaps and Bounds". Technometrics, Vol. 16, 1974, pp. 499-511.

- Sanford Weisberg (1980), "Applied Linear Regression" New York: John Wiley &Sons Tp. 324.

- Scheffe, H. (1957), "The Analysis of Variance", New York: Wiley TP. 324.

- Searle S.R. (1966), "Matriz Algebra for the Biological Scioences", New York: Wiley TP. 295.

- Searle S.R. (1971), "Lineal Models", New York: John & Sons Inc.

- Seber, G.A.F. (1977), "Linear Regression Analysis", New York: Wiley TP. 295.

- Silvey, S.A. (1969), "Multicolinearity and Imprecise Estimation". J.R. Statist. soc. ser. B., 31, 539-552

- Stewart, G.W. (1979), "Assessing The Efects of Variable Error in Linear Regression", Computer Science Technical, Reporte No. 818.

- Stewart, G.W. (1974), Introduction to Madrix Computationes, New York Academic.

- Tate, R.F. and Klett, G.W (1959), Optimom Confidence Intervals for the variance of a normal distribution, J.A. mer. Statist. Assoc. 54, 674-682.

- Utts, J.M. Tje rainbow test for lack of fit in regression co mmun, statist-theor. meth, 11 (24), 2801-2815.

- Wood, F.S. (1973), "The Use of Individual Efects and Residuals in Fitting Equations to Data", Technometrics, 15, 677-695.