36
29

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

"DESARROLLO DE UN PAQUETE ESTADISTICO EN APOYO A LOS CURSOS DE ANALISIS DE REGRESION".

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I O

P R E S E N T A

CLAUDIA JIMENEZ VILLASEÑOR

MEXICO, D. F. SEPTIEMBRE 1991.

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

Contenido : página

Introducción.

Las técnicas estadísticas hoy en día recurren a paquetes de cómputo, ya que las operaciones que en ellas intervienen llegan a ser tan extensas que la presencia de errorres es muy frecuente.

El Análisis de Regresión, como técnica estadística, no escapa a este tipo de problemas por lo cual se ve en la necesidad de recurrir a paquetes de cómputo, sin embargo el uso de estos tiene ciertas desventajas ya que muchas veces se usan de una manera tan automática que se llegan a olvidar las técnicas a través de las cuales se obtiene la información estadística, por lo que surge la preocupación de crear un paquete de cómputo que proporcione los elementos necesarios para obtener resultados sin perder de vista la manera mediante la cual se llega a ellos.

En este trabajo se elaboró un paquete de cómputo en apoyo a los cursos de Análisis de Regresión que se imparten en la Facultad de Ciencias de la UNAM. Fue adaptado para ser ejecutado en microcomputadoras PC o compatibles con el objeto de que todos los alumnos de la facultad que esten interesados en la materia tengan acceso a él.

El paquete de cómputo, proporciona los cálculos necesarios facilitando de esta manera una rápida interpretación y tratando de que el alumno no pierda de vista las técnicas a través de las cuales llega a los resultados deseados.

Debido a la naturaleza de las cálculos utilizados en esta tecnica se utilizó el lenguaje de programación PASCAL (*Turbo PASCAL versión 5*), el cual es muy apegado a este tipo de operaciones.

El presente trabajo se encuentra estructurado de la siguiente forma :

Los tres primeros capítulos estan dedicados a la teoria en la que se basa el Análsis de Regresión.

El capítulo 1 describe en forma detallada la construcción del modelo de Regresión así como cada uno de los supuestos que lo componen con el objeto de relacionarlo con los elementos estadísticos que se conocen para después tratarlo como una técnica de inferencia estadística.

En el capítulo 2 se presenta el modelo de Regresión Lineal Múltiple el cual servirá de base teórica para la implantación del paquete de cómputo.

El capítulo 3 se refiere a selección de variables. Se presentan algunos métodos muy comentados en la literatura, mismos que podrán ser utilizados a través del paquete de cómputo.

En el capítulo 4 se presenta en forma detallada el diseño de cada uno de los programas que conforman el paquete así como los algoritmos de cada uno de ellos.

Por último en el capítulo 5 se muestra la aplicación del paquete mediante ejemplos.

Es importante mencionar que no se pretende competir con los paquetes estadísticos comerciales y menos aún comparar su utilidad. Como se mencionó anteriormente el objetivo central del presente trabajo es proporcionar una herramienta didáctica a los cursos de Análisis de Regresión.

Capitulo 1. Construcción Del Modelo.

Al conjunto de métodos cuyo objetivo es la formulación de modelos matemáticos que muestran relaciones modeladas con el propósito de predecir y hacer inferencias estadísticas se le conoce como *Análisis de Regresión*.

La aplicación de Análisis de Regresión es muy amplia y variada .Como ejemplo se pueden citar las siguientes situaciones:

Explicar el rendimiento de alfalfa (en toneladas por acre) de una granja experimental a través de la cantidad de agua aplicada (en pulgadas por acre),si las variables guardan cierta relación sería posible tratar de predecir el rendimiento a través de la cantidad de agua.

Una cadena de tiendas esta interesada en abrir una nueva unidad para lo cual considera importante saber cómo se explican las ventas semanales en cada tienda, a través del número de empleados y el tamaño de cada tienda. Si existe una relación razonable entre las variables, sería posible predecir las ventas semanales en cada tienda a través del número de empleados y el tamaño de tienda para poder así evaluar si es conveniente abrir o no una nueva tienda.

La palabra *regresión* la utilizó por primera vez *Sir Francis Galton* quien analizó la altura de hijos en terminos de la altura promedio de sus padres .De sus observaciones concluyó que hijos de padres muy altos eran generalmente más altos que el promedio pero no tan altos como sus padres y lo mismo para el caso de hijos de padres muy bajos .Este estudio se publicó en 1885 bajo el título de "*Regresión hacia la mediocridad en la herencia de estatura*".En este caso, el termino regresión se usó en el sentido de que las alturas de los hijos tendían al promedio más que a valores extremos.

En este trabajo se denotará por "y" a la característica de interes (manifestación del fenómeno) y se le denominará variable de respuesta .A la(s) característica(s) a través de la(s) cual(es) se construirá la relación con "y", se le(s) conocerá como variable(s) explicativa(s) (o factores de interes) y se denotarán por $x_1, x_2, \ldots, x_p$.

Para establecer relaciones funcionales entre variables hay que tomar en cuenta el tipo de relaciones que pueden existir. Por ejemplo, Mendez(1977), menciona las siguientes:

1.*Asociación Simple*. Es cuando un fenómeno o modalidad de fenómeno "A" se presenta frecuentemente acompañado de "B", otro fenómeno o modalidad .Sin embargo puede darse "A" sin que ocurra "B" y "B" sin que ocurra "A" .Como ejemplo considérese el hecho de fumar (A) y el cáncer pulmonar (B), los dos aspectos ocurren frecuentemente juntos, sin embargo, hay fumadores a los que no les da cáncer y hay personas con cáncer que no son o han sido fumadores.

Este tipo de asociación puede servir de base para la construcción de hipótesis que ligan los fenómenos en términos de causalidad.

Se llama causa al fenómeno o conjunto de fenómenos que preceden a otro y le dan orígen.

Se llama efecto el fenómeno que sigue a otro y es originado por él.

2.*Causalidad Probabilística o Relación producto-productor.* Este es el caso en que un fenómeno "B" es necesario para que ocurra otro "A" .Pero "B" no es suficiente para que ocurra "A", es decir dada "B" es probable que ocurra "A"; y si ocurre "A" debe haber ocurrido "B".Un ejemplo de ésto es observar que el agua de un recipiente esta hirviendo como fenómeno "B" e inferir que la temperatura de ese recipiente es de 100 grados centigrados o más, fenómeno "A".Sin embargo puede suceder que el agua esté hirviendo sin que se eleve la temperatura, sino únicamente por un efecto de vacio, fenómeno "A1", sobre el recipiente considerado . Se dice que "A" es una entre varias de las causas posibles de "B".

3.*Causalidad Determinística o Relación de causa y efecto.* En este caso un fenómeno "A" es necesario y suficiente para que ocurra otro fenómeno "B".Si "A" ocurre, ocurrirá o ha ocurrido "B" y si "B" ocurre, ocurrirá o ha ocurrido "A".Se habla en este caso de que "A" es la causa única de "B" o viceversa. Como ejemplo de esta relación considérese la ocurrencia de un cromosoma 21 por triplicado en un ser humano (trisonomia 21) como fenómeno "A" y el síndrome de down o mongolismo, fenómeno "B".

Se puede retomar el ejemplo del hecho de fumar, "A", y el cáncer pulmonar, "B", podria considerarse "A" como causa y "B" como efecto, aquí existe una probabilidad de que una persona que fuma enferme de cáncer, sin embargo pueden existir otras causas que originen que una persona enferme de cáncer.

De acuerdo al tipo de relación que se tenga y al marco de referencia sobre el fenómeno se puede saber si tiene sentido establecer una relación funcional e intentar el ajuste de funciones entre las variables.

Establecer la relación funcional puede ser un problema complicado sin embargo en ocasiones puede aproximarse por medio de una función matemática lo más simple posible.

Como se vió anteriormente se trata de construir una función que relacione de la mejor manera posible la manifestación del fenómeno bajo estudio con lo factores de interés.

Al llevar a cabo un experimento con el fin de producir información existen factores que pueden ser de dos tipos : controlados y no-controlados.

Factores Controlados. Variables de interés bajo control que afectan la manifestación del fenómeno y que toman diferentes modalidades .Dentro de esta categoría están las condiciones experimentales que son los factores que permanecen constantes a lo largo del experimento.

Factores No-controlados. Este tipo de factores, como su nombre lo indica no se controlan debido a diversas causas : puede ser que dichos factores se consideraron no relevantes al fenómeno bajo estudio, o ni siquiera se ha percatado su existencia o bien no se pueden manipular.

Sean

y : Variable de respuesta

$x_1,\ldots,x_p$: variables explicativas, es decir los factores de interés que se controlan y toman diversas modalidades.

$A_1,A_2,\ldots A_n$: condiciones experimentales.

$B,C,D,F,\ldots$: factores no controlados.

La manifestación y queda descrita por la relación

$$y = f_o(x_1,\ldots,x_p,A_1,\ldots,A_n,B,C,D,F,\ldots)$$

Como se puede observar los factores controlados junto con las condiciones experimentales constituyen el dominio de f_o.

Dado que $A_1,\ldots,A_n$ son factores que toman la misma modalidad a lo largo de todo el experimento, se restringe el dominio a :

$$y = f_1(x_1,\ldots,x_p,B,C,D,F,\ldots) \quad \ldots(1.1)$$

Sin embargo tanto f_o como f_1 son funciones cuyo dominio es infinito y habrá factores que difícilmente se identificaran por lo que no es posible construir una regla de correspondencia bajo estas condiciones.

Para delimitar el argumento de f_2 es necesario considerar algunos supuestos. Seria deseable inducir una partición de tal forma que

$$P_1 = \{ x_1, \ldots, x_p \} \quad \ldots \text{ factores controlados}$$

$$P_2 = \{ B, C, D, F, \ldots \} \quad \ldots \text{factores no-controlados}$$

Para estos conjuntos sean f_2 y f_3 tales que

$$y = f_2(x_1, \ldots, x_p) + f_3 (B, C, D, F, \ldots) \ldots . (1.2)$$

De lo anterior se desprende el supuesto estructural básico de separabilidad aditiva : *Los efectos que produce el conjunto* P_1 *en la variable "y" y los efectos que produce el conjunto* P_2 *en la misma variable, son separables en forma aditiva.*

Así el dominio de f_2 es finito, pero el de f_3 sigue siendo infinito.

Sin embargo hay que reconocer que aunque f_3 no podrá determinarse lo que interesa es su magnitud.

Como f_3 representa la función que agrupa a todos los factores que no se controlan y que afectan a la manifestación del fenómeno en mayor o menor grado, es usual que se le denomine "error" y cuya magnitud se denota por ε_1

$$y = f_2(x_1, \ldots, x_p) + \varepsilon_1 \ldots (1.3)$$

Nótese que si ε_1 es muy grande podría pensarse que no se ha considerado un factor importante o bien no se tomó en cuenta otro factor como factor experimental o bien la ralación entre x y y no es muy fuerte.

De (1.3) se obtiene que

$$\varepsilon_1 = y - f_2(x_1, \ldots, x_p)$$

Por facilidad supóngase que se cuenta con un sólo factor de interés, x, con $x \in P_1$, entonces

$$\varepsilon_1^* = y - f_2(x)$$

ε_1^* es desconocido y representa la magnitud del error .Sería deseable que esta cantidad fuese lo más pequeña posible.

Entonces el problema es determinar f tal que ε_1^* sea mínimo.

Nótese que la variable x puede ser cualitativa (nominal) o cuantitativa (discreta o continua).Si x es cualitativa se cuenta con una colección de posibles valores (o etiquetas) .Por ejemplo si la variable a estudiar es sexo, sólo admite dos posibles valores: hombre y mujer (h,m) o bien "0,1" o bien "1,2", etc. Los valores que se asignan son arbitrarios y por tanto no tiene ningún sentido establecer una función .La solución a este problema se enmarca dentro de las tecnicas conocidas como *Diseño de Experimentos*.

Sí x es cuantitativa, se tiene que definir una familia de funciones y escoger f_2 en alguna forma.

Sea $f_2 \in L$, con L = { g|g es función }

Hacer una selección en L resulta sumamente difícil debido a la diversidad de funciones que existen.Sería conveniente delimitar una familia de funciones como por ejemplo :

C = { g|g es función continúa en [a,b] }

Se sabe que toda función contínua en un intervalo cerrado y acotado puede ser aproximada uniformemente por un polinomio (teorema de aproximación de *Stone-Weierstrass*).

Tomando en cuenta este resultado, se puede restringir aún más el espacio de funciones a elegir como :

F = { f : [a,b] [c,d] | f es polinomio }

Aún restringiendose a F ¿Cómo elegir F ∈ F?

Sea h ⊂ F, lo cual quiere decir que h será una aproximación polinomial de f, y el error de aproximación se sumará a ε_1^* de tal forma que

$$y = h(x) + \varepsilon \qquad \text{con} \quad \varepsilon = \varepsilon_1^* + (f(x) - h(x))$$

De donde $\varepsilon = y - h(x)$

El problema es encontrar una función h(x) tal que minimice a ε .

Sean (x_1, y_1), (x_2, y_2),....,(x_n, y_n) n parejas de datos y

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ los errores correspondientes.

Se debe determinar un criterio Δ para elegir h tal que minimice los errores.

Sea $\Delta : F \to \mathbb{R}^n \to \mathbb{R}$ donde $h(x) \in F$ y $\underline{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$

Una medida de la magnitud de ε podría ser su norma definida en general como

$$N = \left[\sum_{i=1}^{n} |\varepsilon_i| \right]^{1 \setminus p}$$

El criterio Δ consistiría en minimizar la norma de $\underline{\varepsilon}$.

Restringiendose a aquellos polinomios de grado k, se tiene que:

$$h(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \ldots + \alpha_k x^k$$

entonces

$$y_i = \alpha_0 + \alpha x_1 + \alpha_2 x_i^2 + \ldots + \alpha_k x_i^k + \varepsilon_i \quad i=1,\ldots n$$

entonces

$$\varepsilon_i = y_i \ (\alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \ldots + \alpha_k x_i^k) \qquad i=1,\ldots n$$

Elegir $h(x)$ tal que minimice al error, es equivalente a elegir los coeficientes $(\alpha_0, \alpha_1, \ldots, \alpha_k)$ que minimicen a ε_i (ya que el valor de x está fijo).

Sea $z_{ji} = x_i^j$ entonces

$$y_i = \alpha_0 + z_{1i} + \alpha_1 z_{2i} + \ldots + \alpha_k z_{k\,i} + \varepsilon_i \qquad i=1,\ldots n$$

Que es una combinación lineal en los parámetros (desconocidos) y en las z's, que son funciones conocidas y totalmente especificadas de la variable explicativa.

Hay que notar que esta modalidad permite escribir en forma de combinación lineal a las potencias sucesivas de x, pero aún más, permite que z_{ji} no sólo sean potencias sino funciones de la x en cuestión e incluso diferentes funciones de varias variables explicativas, siempre y cuando estas esten completamente especificadas.

Por ejemplo :

$$w = \alpha_0 + \alpha_1 f_1 (x_1) + \alpha_2 f_2 (x_2) + \alpha_3 f_3 (x_3) + \varepsilon$$

Podría reescribirse como :

$$y = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \varepsilon_i$$

Bajo esta modalidad incluso funciones no lineales pero susceptibles de linealizarse se tomarán en cuenta.

Supóngase entonces que :

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{p-1} + \varepsilon_i \qquad i=1,\ldots,n$$

en donde

$\{ x_1, x_2, \ldots, x_p \}$ es el conjunto de variables explicativas o bien funciones de ellas completamente especificadas.

y es la variable de respuesta y

$\beta_0, \beta_2, \ldots, \beta_{p-1}$ son coeficientes desconocidos

entonces

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{p-1})$$

Para minimizar $\underline{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ es razonable pensar en minimizar la norma de $\underline{\varepsilon}$. La norma p se define como

$$N = \left[\sum_{i=1}^{n} |_i \varepsilon \, | \right]^{1/p}$$

y el criterio Δ consistiría en minimizar N_p , es decir

$$\min_{\{\alpha_0 \ldots \alpha_{p-1}\}} \left[\sum_{i=1}^{n} |y_i - \alpha_1 x_{1i} - \ldots - \alpha_{p-1} x_{p-1i}| \right]^{1/p}$$

Por facilidad sea $p = 2m$ con lo cual $| \varepsilon_i |^{2m} = \varepsilon_i^{2m}$

La función $(\sum_{i=1}^{n} \varepsilon_i)^{1/2m}$ es monótona no decreciente por lo que basta minimizar

$$\Delta' (\beta_0, \beta_1, \ldots, \beta_{p-1}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_{p-1} x_{p-1})^{2m}$$

Derivando

$$\frac{\partial \Delta'}{\partial \beta_0} = \sum_{i=1}^{n} z_m \left(y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_{p-1} x_{p-1} \right)^{z_{m-1}} (-1)$$

$$\vdots$$

$$\frac{\partial \Delta'}{\partial \beta_{p-1}} = \sum_{i=1}^{n} z_m \left(y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_{p-1} x_{p-1} \right)^{z_{m-1}} \left(-x_{p-1i} \right)$$

Igualando a cero :

$$\frac{\partial \Delta'}{\partial \beta_1} = 0$$

$$\vdots$$

$$\frac{\partial \Delta'}{\partial \beta_{p-1}} = 0$$

El sistema de p-ecuaciones no es sencillo de resolver ya que queda un sistema de ecuaciones en potencias de (z_{m-1}). Sin embargo si m=1, se obtiene un sistema de ecuaciones lineales que preserva las condiciones originales :

$$-2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \dots - \beta_{p-1} x_{p-1} \right) = 0$$

$$\vdots$$

$$-2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \dots - \beta_{p-1} x_{p-1} \right) = 0$$

Por tanto basta minimizar

$$\Delta'(\beta_0, \beta_1, \dots, \beta_{p-1}) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \dots - \beta_{p-1} x_{p-1} \right)^2$$

$$= \sum_{i=1}^{n} e_i$$

Al criterio anterior se le conoce como *Mínimos Cuadrados* .

En el caso de una variable se tendrá :

$$\sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1)^2$$

Derivando con respecto a β_0 y β_1 e igualando a cero se obtienen las ecuaciones :

$$\sum_{i=1}^{n} y_i = n \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} y_i x_i = \hat{\beta}_0 + \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

De las cuales se obtiene

$$\hat{\beta}_1 = \sum_{i=1}^{n} y_i x_i \qquad \text{y} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Como se vió anteriormente el error ε es desconocido pues involucra a los factores no controlados. En este sentido existe incertidumbre en torno a él. Es razonable suponer que el error es una variable aleatoria.

Como tal la manera natural de describir su comportamiento será a traves de una función de distribución.

Como es natural suponer que $\varepsilon \in \mathbb{R}$ entonces la distribución que se considere debe ser continua.

Dada la construcción del modelo, es razonable suponer que :

$$E(\varepsilon) = 0 \qquad \text{y además que} \qquad P(\varepsilon > 0) = P(\varepsilon < 0)$$

Lo cual lleva a que la distribución será simétrica centrada en cero.

Dado que las fuentes de variación son en principio, las mismas para cada observación (ya que los factores no controlados son los mismos de observación a observación), se puede suponer que

$$V(\varepsilon_i) = \sigma^2 \qquad i = 1, \ldots, n$$

Además los factores controlados junto con las condiciones experimentales generan los datos por lo que adicionalmente se hace el siguiente supuesto :

$$\text{Cov}\ (\varepsilon_i , \varepsilon_j) = 0 \quad i \neq j$$

Con los supuestos anteriores sobre la esperanza, varianza y covarianza de los errores, se obtienen resultados importantes de los estimadores por mínimos cuadrados. Estos resultados se resumen en el llamado *Teorema de Gauss-Markov* el cual grantiza que β_0 y β_1 son los mejores estimadores lineales insesgados en el sentido de varianza mínima.

Aún cuando este resultado es poderoso, se requiere de un supuesto adicional para poder explicar tecnicas de inferencia, es decir, se requiere suponer una distribución para los errores.

Bajo las propiedades antes descritas es razonable pensar en la dn Normal para explicar el comportamiento de los ε_i , pues esta dn posee características que permiten hacer uso de elementos estadísticos con mayor facilidad que con otra dn.

Ahora bien, tomando el vector $\underline{\varepsilon} = (\varepsilon_1 , \ldots , \varepsilon_n)$ se tiene entonces un vector aleatorio tal que

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

Y la matriz de varianzas y covarianzas Σ es de la forma $\sigma^2 I$. Entonces

$$\underline{\varepsilon} \sim N_n (\underline{0} , \sigma^2 I)$$

(En el apéndice 1 se presenta una sección sobre esta distribución).

Como $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$

Entonces $\underline{Y} \sim N_n (X\underline{\beta} , \sigma^2 I)$

CAPITULO 2 : MODELO DE REGRESION LINEAL MULTIPLE.

2.1. El modelo.

Cuando se tiene más de una variable explicativa el modelo de regresión lineal múltiple (RLM) queda representado de la siguiente manera :

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_p + \varepsilon_i \qquad 2.1.1$$

En *RLM* resulta natural utilizar notación matricial

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

es decir

$$\underset{nx1}{Y} = \underset{nxp-1}{X}\underset{px1}{\beta} + \underset{nx1}{\varepsilon}$$

en donde

$\underset{nx1}{Y}$ Es el vector de nx1 observaciones (variables de respuesta).

$\underset{nxp-1}{X}$: Es una matriz de valores fijos a la cual se le denomina matriz de datos o de diseño. de rango p.

$\underset{px1}{\beta}$: Vector de parámetros desconocidos.

$\underset{nx1}{\varepsilon}$ Vector aleatorio que agrupa a los factores no controlados.

De 2.1.1 se tiene que

$$\varepsilon = Y - X\beta$$

Generalizando los resultados del capítulo 1, se tiene que

i) $E(\varepsilon) = 0$

ii) $V(\varepsilon) = \sigma^2 I$

De acuerdo a lo visto en el capítulo 1, para encontrar β tal que minimice el error es necesario minimizar $\Sigma\varepsilon_i^2$ es decir encontrar $\min(\varepsilon^t\varepsilon) = \sum_{i=1}^{n} \varepsilon_i^2 = \min \{ (\underline{Y} - X\underline{\beta})^t (\underline{Y} - X\underline{\beta}) \}$

Al diferenciar con respecto a β, las ecuaciones normales resultantes son :

$$(X^tX)\underline{\beta} = X^t\underline{Y}$$

y se obtiene que :

$\hat{\beta} = (X^tX)^{-1} X^t\underline{Y}$ y $(X^tX)^{-1}$ existe ya que el rango de X = p.

El estimador β por mínimos cuadrados es insesgado y su varianza es $\sigma^2(X^tX)^{-1}$. Además el teorema de *Gauss Markov* asegura que el mejor estimador lineal insesgado de $l^t\beta$ en el sentido de varianza mínima es $l^t\hat{\beta}$ con $\hat{\beta}$ el estimador por mínimos cuadrados y $l \in \mathbb{R}^p$ vector de constantes conocidas.

Aunque el Teorema de Gauss Markov proporciona una propiedad importante sobre β, el método de *Mínimos Cuadrados* no proporciona directamente un estimador para σ^2. Además sería deseable poder hacer uso de tecnicas de inferencia. Incorporando el supuesto de Normalidad,

$\underline{Y} \sim N_n(X\underline{\beta}, \sigma^2 I)$, los estimadores máximo verosímiles de β y σ^2 se obtienen maximizando la función de verosimilitud:

$$L(\sigma^2, \underline{\beta} | \underline{Y}) = (2\pi\sigma^2)^{-n/2} \exp \left| \frac{-1}{2\sigma^2} (\underline{Y}-X\underline{\beta})^t (\underline{Y}-X\underline{\beta}) \right|$$

Derivando e igualando a 0 se tiene $\hat{\underline{\beta}} = (\underline{X}^t\underline{X})^{-1}\underline{X}^t\underline{Y}$ que coincide con el obtenido por mínimos cuadrados.

El estimador máximo verosímil de σ^2 es :

$$\hat{\sigma}^2 = \frac{1}{n} (\underline{Y} - X\hat{\underline{\beta}})^t (\underline{Y} - X\hat{\underline{\beta}}).$$

El vector de observaciones ajustadas es : $\hat{Y} = X\hat{\beta}$.
Si se desea predecir el valor de Y para un vector X^t_* (con $\underset{*}{X}_{px1}$ de valores conocidos) se tiene que:

$$\hat{Y}_* = \hat{\beta}_0 + \hat{\beta}_1 X_{1*} + \hat{\beta}_2 X_{2*} + \ldots + \hat{\beta}_{p-1} X_{p-1*} = X^t_* \hat{\beta}$$

Por otra parte, el valor esperado de Y cuando $\underset{\sim}{X} = \underset{\sim}{X}_*$ es :

$$\mu_{\underset{\sim}{X}_*} = \mathbb{E}\ (\ Y_* \mid \underset{\sim}{X} = \underset{\sim}{X}_*\ ) = \beta_0 + \beta_1 X_{1*} + \ldots + \beta_{p-1} X_{p-1*}$$

El estimador máximo verosimil de $\mu_{\underset{\sim}{X}_*}$ es :

$$\hat{\mu}_{\underset{\sim}{X}_*} = \hat{\beta}_0 + \hat{\beta}_1 X_{1*} + \ldots + \hat{\beta}_{p-1} X_{p-1*}$$

2.2. Intervalos de confianza y pruebas de hipótesis.

Ahora bien en base a los supuestos distribucionales se tiene que:

$$\hat{\underset{\sim}{\beta}} \sim N\left(\underset{\sim}{\beta}\ ,\ \sigma^2 (\ X^t X\ )^{-1}\right)$$

y

$$\underset{\sim}{1}^t \hat{\underset{\sim}{\beta}} \sim N_p\left(\underset{\sim}{1}^t \underset{\sim}{\beta}\ ,\ \underset{\sim}{1}^t \sigma^2 (\ X^t X\ )^{-1} \underset{\sim}{1}\right)$$

Resulta interesante construir intervalos de confianza para $\underset{\sim}{1}^t\underset{\sim}{\beta}$ con $\underset{\sim}{1} \in \mathbb{R}^p$ vector de constantes conocidas.

En particular si $\underset{px1}{1}$ es el i-esimo vector canónico se estará construyendo un intervalo de confianza para la componente β_{i+1} del vector β.

El intervalo para $1^t\beta$ al $(1-\alpha)$x100% de confianza es :

$$I = \left[\ \underset{\sim}{1}^t\hat{\underset{\sim}{\beta}} \pm t^{1-\alpha/2}_{n-p}\ \tilde{\sigma}\ \sqrt{\underset{\sim}{1}^t (\ X^t X\ )^{-1}\ \underset{\sim}{1}}\ \right]$$

Con

$$\tilde{\sigma}^2 = \frac{(\underline{Y} - X\hat{\beta})^t (\underline{Y} - X\hat{\beta})}{n - p}$$

el estimador insesgado de σ^2.

Tambien se construye el intervalo para σ^2 de la forma

$$\sigma^2 \in \left[0, \frac{(n-p)\tilde{\sigma}^2}{\chi^2_{n-p\ \alpha}} \right]$$

Este intervalo proporciona información de que tan grande puede llegar a ser σ^2.

Si se quiere un intervalo para μ_{y_*}, este es :

$$\left[\hat{\mu}_{x_*} \pm t^{1-\alpha/2}_{n-2}\ \tilde{\sigma} \sqrt{\frac{1}{n} + \frac{(X_* - \bar{X})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \right]$$

El intervalo de predicción para Y_* es :

$$\hat{\mu}_{x_*} \pm t^{1-\alpha/2}_{n-2}\ \tilde{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_* - \bar{X})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Prueba de Hipótesis Lineal General.

Dado que los valores del vector β son desconocidos resulta natural hacer conjeturas sobre sus posibles valores.

En particular, una prueba de hipótesis que proporciona mucha información es la llamada *hipótesis de significancia global* quepropone probar si los elementos $\beta_1, \beta_2, \ldots, \beta_{p-1}$ son todos ceros.

Es decir

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \quad vs \quad H_a : \text{alguna } \beta_i = 0\ \forall i=1,p-1$$

En caso de rechazar H_o , se tendría que $Y_o = \beta + \varepsilon$ lo cual implica que ninguna de las variables explicativas se relacionan con Y, al menos en forma lineal, y permitiría deducir que el modelo no es apropiado. Por el contrario, si se rechaza H_o se tiene que al menos una variable explicativa guarda relación con Y.

Es posible tambien estar interesado en probar si una parte del vector β toma ciertos valores ó incluso si uno solo de los componentes, digamos β_i , es igual a un valor predeterminado.

Todas las hipótesis antes mencionadas se engloban en lo que se conoce como *Hipótesis Lineal General* que toma la forma:

$$H_o : C\underline{\beta} = \underline{\gamma} \qquad vs \qquad H_a : C\underline{\beta} \neq \underline{\gamma}$$

Donde $\underset{r\times p}{C}$ de rango r, $r \leq p$, $\underset{r\times 1}{\gamma}$ completamente especificadas.

Se tiene $\underline{Y} \sim N(X\underline{\beta}, \sigma^2 I)$

La región crítica utilizando Cociente de Verosimilitudes es :

$$C = \left\{ \frac{\sup_{H_o} L\ (\underline{\beta}, \sigma^2) \mid \underline{Y})}{\sup_{H_o \cup H_a} L\ (\underline{\beta}, \sigma^2) \mid \underline{Y})} \leq K_1 \right\}$$

$$\sup_{H_o \cup H_a} L\ (\underline{\beta}, \sigma^2 \mid \underline{Y}) = L\ (\hat{\underline{\beta}}, \hat{\sigma}^2 \mid \underline{Y})$$

Con $\hat{\beta}$ y $\hat{\sigma}^2$ los estimadores maximos verosímiles de β y σ^2.

Ahora

$$\sup_{H_o} L(\beta, \sigma^2 \mid Y) = L\ (\tilde{\beta}, \hat{\sigma}_o^{\,2} \mid Y)$$

En donde $\tilde{\beta}$, $\hat{\sigma}_o^2$ son los estimadores máximo verosímiles bajo la restricción $C\beta = \gamma$ con

$$\tilde{\underline{\beta}} = \hat{\underline{\beta}} - (X^tX)^{-1}C^t(C(X^tX)^{-1}C^t)(C\hat{\underline{\beta}} - \underline{\gamma})$$

$$\sigma_*^2 = \frac{1}{n}(\underline{Y} - X\underline{\beta})^t(\underline{Y} - X\underline{\beta})$$

El cociente de verosimilitudes resulta ser :

$$\Delta = \frac{2\pi \hat{\sigma}_0^2 \exp\{-n/2\}}{2\pi \hat{\sigma}^2 \exp\{-n/2\}}$$

$$\Delta = \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}$$

Se rechaza H_0 si $\Delta \leq K$ es decir

$$\leftrightarrow \quad \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \leq K$$

$$\leftrightarrow \quad \frac{(\underline{Y} - X\underline{\beta})^t (\underline{Y} - X\underline{\beta})}{(\underline{Y} - X\hat{\underline{\beta}})^t (\underline{Y} - X\hat{\underline{\beta}})} \leq K$$

Sea $SCE_{MC} = (\underline{Y} - X\underline{\beta})^t (\underline{Y} - X\underline{\beta})$

SCE_{MC} representa la variabilidad de Y con respecto a $\hat{Y}$ cuando se ajuste el modelo completo $\underline{Y} = X\underline{\beta} + \underline{e}$

Puede interpretarse como el error que se comete al ajustar el MC.

Sea $SCE_{MR} = (Y - X\tilde{\beta})^t (Y - X\tilde{\beta})$

SCE_{MR} representa la variabilidad de Y con respecto a $\hat{Y}$ cuando el modelo es el modelo reducido, el modelo se reduce dependiendo de la forma de C y γ.

Entonces

$$\Delta = \frac{SCE_{MC}}{SCE_{MR}}$$

Este cociente, por construcción es menor o igual que uno lo que implica que $SCE_{MR} \geq SCE_{MC}$

De hecho puede demostrarse que $SCE_{MR} = SCE_{MC} + SCE_{H_o}$ es decir

$$(\underline{Y} - X\tilde{\underline{\beta}})^t (\underline{Y} - X\tilde{\underline{\beta}}) = (\underline{Y} - X\hat{\underline{\beta}})^t (\underline{Y} - X\hat{\underline{\beta}}) + (C\tilde{\underline{\beta}} - \underline{\gamma})^t (C (X^tX) C^t)^{-1} (C\tilde{\underline{\beta}} - \underline{\gamma})$$

La cual se conoce como *La Partición Fundamental en Sumas De cuadrados.*

Sea $SCE_{HO} = (C\tilde{\underline{\beta}} - \underline{\gamma})^t (C (X^tX) C^t)^{-1} (C\tilde{\underline{\beta}} - \underline{\gamma})$

La SCE_{HO} es la magnitud en variabilidad, que hace falta para que el modelo completo tenga la misma variabilidad que el reducido.

$$SCE_{HO} = SCE_{MR} - SCE_{MC}$$

Se rechaza Ho si $\dfrac{SCE_{MC}}{SCE_{MR}} \leq K$

$$\leftrightarrow \quad \frac{SCE_{MR}}{SCE_{MC}} \geq K_2$$

$$\leftrightarrow \quad \frac{SCE_{HO} + SCE_{MC}}{SCE_{MC}} \geq K_2$$

$$\leftrightarrow \quad \frac{SCE_{HO}}{SCE_{MC}} + 1 \geq K$$

$$\frac{SCE_{HO}}{SCE_{MC}} \geq K'$$

Entonces $\mathbb{C} = \{ \underline{Y} \mid \dfrac{SCE_{HO}}{SCE_{MC}} \geq K \}$

$$Pr \left(\frac{SCE_{HO}}{SCE_{MC}} \geq K \mid Ho \text{ es cierta} \right) = \alpha$$

Para determinar K es entonces necesario determinar la Distribución de $\frac{SCE_{Ho}}{SCE_{MC}}$

Ahora

$$\frac{SCE_{MC}}{\sigma^2} \sim X^2_{(n-p)} \qquad \text{(ver apendice 2)}$$

y $$\frac{SCE_{Ho}}{\sigma^2} \sim X^2_r$$

Bajo Ho.

Las dos variables en cuestion son independientes (ver apéndice 2), y su cociente, afectado por ciertas constantes, resulta ser una variable aleatoria con distribución " F " :

$$\frac{r \quad SCE_{MC}/\sigma^2}{n-p \quad SCE_{HO}/\sigma^2} \rightarrow F\ (\ r, n-p\ )\ \text{ bajo } H_0$$

Entonces :

$$\frac{SCE_{HO}/\ r}{SCE_{MC}/\ n-p} \rightarrow F\ (\ r\ ,\ n-p\ )\ \text{bajo } H_0$$

Se denotan los cuadrados medios del error de la siguiente manera

$$\frac{SCE_{HO}}{r} = CME_{HO} \quad ; \quad \frac{SCE_{MC}}{n-p} = CME_{MC}$$

entonces la región critica está dada por

$$C = \{\ \frac{CME_{HO}}{CME_{MC}} \geq F^{1-\alpha}_{r,n-p}\ \}$$

La tabla de análisis de varianza (*ANOVA*) para probar las hipótesis :

$H_0 : C\underline{\beta} = \underline{\gamma}$ VS. $H_a : C\underline{\beta} \neq \underline{\gamma}$ en el modelo $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$

con $\underline{\varepsilon} \rightarrow N_n\ (\underline{0}, \sigma^2 I)$ es

Fuente de variación	g.l.	Suma de Cuad.	Cuad. Medios	F_c F calcula
H_0	r	$(C\hat{\underline{\beta}}-\underline{\gamma})[C(X'X)^{-1}C']^{-1}(C\hat{\underline{\beta}}-\underline{\gamma})$	$CME_{HO} = \frac{SCE_{HO}}{r}$	$\frac{CME_{HO}}{CME_{MC}}$
Error MC	n - p	$(\underline{Y}-X\hat{\underline{\beta}})'(\underline{Y}-X\hat{\underline{\beta}})$	$CME_{MC} = \frac{SCE_{MC}}{n-p}$	
Total MR	n-p-r	$(\underline{Y}-X\bar{\underline{\beta}})(\underline{Y}-X\bar{\underline{\beta}})$		

Como se mencionó anteriormente existen casos particulares para esta prueba entre los que cabe resaltar :

1) $H_o : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$ VS. $H_a : \beta_i \neq 0 \quad i=1,\ldots,p-1$

Es decir

$$H_o : C\underline{\beta} = \underline{0} \quad VS. \quad H_a : C\underline{\beta} \neq \underline{0}$$

$$\underset{(p-1)\times p}{C} = \begin{bmatrix} 0\ 1\ \ldots 0 \\ 0\ 0\ 1\ldots 0 \\ \ldots\ldots\ldots \\ 0\ldots\ldots 1 \end{bmatrix} \qquad \underset{p\times 1}{\underline{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_{p-1} \end{bmatrix} \qquad \underset{(p-1)\times 1}{\underline{\gamma}} = \begin{bmatrix} 0 \\ \cdot \\ 1 \\ \cdot \\ \cdot \end{bmatrix}$$

A esta prueba se le conoce como *Prueba de Significancia Global* lo que indica que existe al menos una $\beta_i \neq 0$ $i=1,\ldots,p-1$ entonces el modelo tiene sentido.

Por lo anterior bajo H_o cierta el modelo reducido es

MR : $Y_i = \beta_1 + \varepsilon_i \qquad i=1,\ldots,n \qquad con \quad Y_i \rightarrow N(\beta_o, \sigma^2)$

y la suma de cuadrados del error para el MR queda como

$SCE_{MR} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \underline{Y}^t\underline{Y} - n\bar{Y}^2$ que es la variabilidad total de las y's. El modelo completo es

MC : $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$ y la suma de cuadrados del error para el MC es

$SCE_{MC} = \underline{Y}^t\underline{Y} - \underline{\beta}^t X^t \underline{Y}$

Por la partición fundamental en sumas de cuadrados se tiene

$SCE_{MR} = SCE_{MC} + SCE_{HO}$

De donde

$SCE_{HO} = SCE_{MR} - SCE_{MC}$

$SCE_{HO} = \underline{\hat{\beta}}^t X^t \underline{Y} - n\bar{Y}^2$

Por tanto la tabla Anova es :

Fuente de variación	g.l.	Suma de Cuad.	Cuad. Medios	F_c F calculada
H_o	p - 1	$SCE_{HO} = \underline{\hat{\beta}}^t X^t \underline{Y} - n\bar{Y}^2$	$CME_{HO} = \frac{SCE_{HO}}{p-1}$	$\frac{CME_{HO}}{CME_{MC}}$
Error MC	n - p	$SCE_{MC} = \underline{Y}^t\underline{Y} - \underline{\hat{\beta}}^t X^t \underline{Y}$	$CME_{MC} = \frac{SCE_{MC}}{n-p}$	
Total MR	n - 1	$SCE_{MR} = \underline{Y}^t\underline{Y} - n\bar{Y}^2$		

donde la región crítica es :

$C = \{ \underline{Y} \mid F_c \geq F^{1-\alpha}_{r,n-p} \}$

2)

$H_0 : l^t\beta = m$ con $m \in \mathbb{R}$ conocida donde l $_{p\times 1}$ vector de constantes conocidas.

Un caso particular de esta situación es cuando l es canónico y m = 0, es decir l es el i-esimo vector canónico, entonces

$$l^t\beta = \beta_{i+1} \quad \text{y la prueba será} \quad H_0 : \beta_{i+1} = 0$$

Esta prueba se conoce como prueba *F-parcial*, sirve para probar si una β_i toma el valor cero, lo cual equivale a decir que la x asociada a ese coeficiente no entra al modelo.

A continuación se presentan algunos conceptos importantes

Una herramienta que sirve para medir el grado de asociación lineal entre dos variables x y Y es el coeficiente de correlación lineal que esta dado por

$$r_{xy} = \frac{\Sigma (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\Sigma (x_i - \bar{x})^2 \Sigma (Y_i - \bar{Y})^2}} \quad \text{con } -1 < r_{xy} < 1$$

si r_{xy} = 1 se tiene una asociación lineal directa.
si r^{xy} = -1 se tiene una asociación lineal indirecta.
si r^{xy}_{xy} = 0 no se tiene una asociación lineal, lo cual no significa que x y Y no guarden una relación.

Una medida de la bondad del ajuste de un modelo de regresión es la estadística R^2. Se define como

$$R^2 = \frac{SCE_{Ho}}{SCE_{MR}} \qquad 0 < R^2 < 1 \quad .$$

Con SCE_{MR} de significancia global y se interpreta como la proporción de la variabilidad de la Y que quedó explicada por el modelo.

Residuales. Los residuales se definen como $e_i = Y_i - \hat{Y}_i$. Sirven para hacer un análisis de los supuestos distribucionales.

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

$E(\underline{e}) = \underline{0}$

$V(e) = \sigma^2 (I - X(X^tX)^{-1}X^t)$

Residuales Studentizados. Se definen como $r_i = e_i / \hat{\sigma}\sqrt{1 - h_{ii}}$

donde h_{ii} es el componente de la matriz $(X(X^tX)^{-1}X^t$

Estos residuales son más utilizados que los e_i estandarizados, pues las pruebas en las que se involucran, como Papel Normal, entre otras, son más precisas.

Para probar el supuesto de Normalidad es usual utilizar estos residuales o los e_i, lo cual consiste en tomar los residuales ordenados de menor a mayor y graficar contra $K_i=(i-1/2)/n$, en papel Normal, si la gráfica presenta un comportamiento aproximadamente lineal, esto querra decir que el supuesto de Normalidad es correcto.

Otras gráficas de gran utilidad es tomar los residuales contra las y_i estimadas y contra las x_i, las cuales sirven para analizar entre otras cosas, la variabilidad de los e_i.

Una estadística que sirve para analizar la correlación entre los residuales es la propuesta por Durbin y Watson. Se define como:

$$d_i = \sum_{i=2}^{n} (e_i - e_{i-1})^2 / \sum_{i=1}^{n} e_i^2 \qquad i=1,\ldots,n.$$

Capítulo 3 : Selección de Variables.

En el *análisis de regresión* surge un problema que consiste en incluír un conjunto de variables de tal manera que para este conjunto el modelo de regresión sea óptimo, esto es, que el modelo refleje al fenómeno bajo estudio. Como se vió anteriormente el modelo es de la forma :

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon_i$$

Donde las x_i todas las variables independientes que se piensa ayudarán a explicar el fenómeno lo mejor posible. Cuando se plantea el modelo, se trata de involucrar el mayor número de variables explicativas de tal manera que existan menos factores no controlados y asi tener un mejor control, sin embargo, es importante notar que el hecho de incluir en el modelo todas las variables explicativas que se tengan, no implica que el modelo sea el mejor pues puede ocurrir que existan variables que no aporten al modelo suficiente información, o que unas esten explicadas a traves de otras, además de que el modelo sería más complicado y su interpretación sería menos clara. Por lo anterior surge el problema de seleccionar un subconjunto de q variables($q < p-1$) a travs de las cuales sea posible explicar el fenómeno, donde dicha selección se llevará a cabo mediante tecnicas estadísticas que serán discutidas más adelante, con lo que el nuevo modelo sería de la forma :

$$y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ji} + \varepsilon_i$$

Problema de multicolinealidad.

Cuando el modelo presenta variables tales que unas están explicadas a traves de otras, lo cual dificultará en un momento determinado precisar la relación que existe entre la manifestación del fenómeno con respecto a alguna variable explicativa, este último hecho es lo que se conoce como *multicolinealidad* (correlación entre variables explicativas).

El tener efectos de *multicolinealidad* en un modelo origina algunos problemas. Por ejemplo, si se desea realizar una interpretación de los coeficientes parciales de regresión, el efecto de multicolinealidad puede conducir a contradicciones. Estos problemas se pueden evitar mediante la eliminación de alguna de las variables que presentan tal efecto, proceso que puede llevarse a cabo mediante tecnicas de selección de variables.

Selección de Variables.

- *Todas las regresiones posibles.*
- *Backward.*
- *Forward.*
- *Stepwise.*

Cabe aclarar que ninguno de estos métodos garantiza que el subconjunto de variables explicativas que intervienen en el nuevo modelo sea el mejor.

Todas las regresiones posibles.

Bajo este criterio se requiere ajustar todas las posibles combinaciones de las x_i. Para cada x_i existen dos posibilidades de estar o no estar en el modelo de regresión, por lo tanto existen 2^p modelos posibles a examinar.

Cabe mencionar que bajo este criterio es claro que si se tiene un conjunto de variables explicativas grande, se obtendrán un gran número de posibles modelos, lo cual en términos de cálculo y costo tiene grandes limitaciones.

Suponiendo que no existen limitaciones, el análisis para seleccionar el mejor subconjuto de x_i puede hacerse a través de diferentes criterios entre los cuales se puede mencionar a los siguientes :

1. R^2.
2. R^2 ajustada.
3. Cp-Mallows.

Criterio de la R^2. R^2 mide la proporción de la variabilidad de la Y que quedo explicada por el modelo. R^2 es usada como una medida de a traves de la cual se compara la validez de los resultados que se obtienen bajo especificaciones alternativas de las variables explicativas en el modelo.

El uso de la R^2 acarrea varios problemas, en primer lugar es sensible al número de variables explicativas incluidas en el modelo. La inclusión de más variables explicativas nunca hara que R^2 decrezca, por lo que es posible añadir más y más variables, lo que no necesariamente significa que todas esas variables tengan que estar en el modelo.

Bajo este criterio se calculan las R^2 para cada modelo obtenido dentro de todas las posibles combinaciones, así para todos los modelos con q variables se elige a la R^2 mayor y se toma ese modelo como el más significativo. En el caso de existan empates entre variables, se puede calcular su correspondiente coeficiente de correlación parcial, mismos que se definen más adelante, pues puede suceder que una de estas variables quede reflejada por otra, en ese caso se desecha la variable que esta siendo redundante para el ajuste del modelo.

R^2 ajustada. La dificultad con R^2 radica en que contiene la variación de Y explicada por el modelo y la no explicada pero no toma en cuenta los grados de libertad del problema. Una solución natural es pensar en las varianzas (muestrales) o cuadrados medios y no en las variaciones (o sumas de cuadrados). Se define como :

$$\bar{R}^2 = 1 - \frac{\text{Var}(e_i)}{\text{Var}(Y)}$$

donde $$\text{Var}(e_i) = \hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i) / n-p$$

$$\text{Var}(y) = \hat{\sigma}_o^2 \quad \sum_{i=1}^{n} (y_i - \bar{y}_i) / n-1$$

con p = número de parámetros.

Aún cuando la SCE_{MC} decrezca, a medida que se añaden más variables, la varianza de los e_i no necesariamente decrece. Hay que notar que tanto el numerador como el denominador de $\hat{\sigma}^2$ cambian al incluir una nueva variable en el modelo. Es fácil checar que

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

y resulta inmediato que

1) si $p = 1$ entonces $R^2 = \bar{R}^2$

2) si $p > 1$ entonces $R^2 \geq \bar{R}^2$

3) $\bar{R}^2$ puede ser negativa

Estas propiedades hacen que $\bar{R}^2$ resulte una medida más adecauda que R^2.

Cp-Mallows. Mallows sugiere una estadística para elegir al mejor subconjunto de variables, la cual es comparada con el número de parámetros del modelo y se recurre a una gráfica en la que intervienen la Cp y p, número de parámetros.

$$Cp = SCEMC_p / CMEMC(n - 2p)$$

donde p se refiere a la SCE_{MC} que resulta al tomar todas las variables explicativas.

El valor de cada Cp es muy importante porque es un estimador de la suma de cuadrados global de las discrepancias entre el modelo ajustado y el modelo verdadero (desconocido).

Para un modelo adecuado se tiene que E(Cp)=p . Se sigue que una gráfica de Cp vs. p mostrará modelos adecuados en la medida en que los puntos se acerquen a la recta Cp=p. Ajustes que adolezcan de carencia de ajuste, esto es, ecuaciones sesgadas, produciran puntos muy por encima de la recta Cp=p. Debido a variaciones aleatorias, puntos que representen modelos adecuados pueden estar por debajo de la recta Cp=p.

En la gráfica se toma el siguiente criterio, si Cp se aproxima a p querrá decir que el modelo que corresponda a esa Cp será el más significativo.

Backward. (*Eliminación hacia atrás*). El método consiste en los siguientes pasos :

1. Se ajusta el modelo de regresión con todas las variables.

2. Para cada variable se calcula la F-parcial correspondiente.

3. De todas las F-parciales calculadas se toma la de menor valor, la cual se denotara como F*, y se compara con una F-tablas seleccionada de antemano.

3.1.Si F* < F-tablas, entonces la variable que corresponda a esa F-parcial se excluye del modelo de regresión, ajustando un nuevo modelo con las variables restantes.

Es importante aclarar que una vez que sale una variableyano puede volver a entrar al modelo, después regresar al paso número 2.

3.2.Si F* > F-tablas, no sale ninguna variable y se toma el modelo que se ajusto para predecir a Y.

Este criterio de selección en términos de cálculo y de costo resulta mucho más adecuado que el anterior, pues sólo se ajustan los modelos que sugieren las F-parciales escogidas.

Forward. (*Eliminación hacia adelante*).

En este criterio usualmente se utilizan los coeficientes de correlación de las x_i y Y, siguiendose los siguientes pasos :

1. Se calculan los coeficientes de correlación de todas las x_i con Y. Se selecciona a la x_i más correlacionada con Y, de este modo la x_i seleccionada será la primer variable incluida en el modelo.

2. Se calculan los coeficientes de correlación de Y con las variables que aún no han entrado al modelo de regresión. La variable con coeficiente de correlación parcial más alto es la que entra al modelo.

El coeficiente de Correlación parcial mide la correlación parcial entre la variable y y x_i en presencia de un conjunto de variables exolicativas y se define como :

$$r_{y_i \cdot 12 \ldots p} = \frac{-R_{yi}}{\sqrt{R_{yy} R_{ii}}}$$

Donde cada R son los cofactores que se determinan mediante los menores que resultan de obtener una submatriz de correlaciones misma que se compone a partir de la eliminación del renglón donde se encuentra la y columna i en la matriz de correlaciones.

$$R_{ij} = (-1)^{i+j} \alpha_{ij}$$

3. Se calcula la F-parcial correspondiente y se compara con una F-tablas.

Si F-parcial > F-tablas la variable entra al modelo y se regresa a 2.

Si F-parcial < F-tablas la variable no entra al modelo termina el proceso.

Bajo este criterio el cálculo es mucho menor que en los anteriores pues sólo se consideran las variables necesarias.

Stepwise. (*Eliminación paso a paso*). Este método es una mezcla de los dos anteriores y consiste en lo siguiente :

1. Se seleccionan las variables siguiendo los mismos pasos de Forward. La diferencia es que una vez incluida una variable se calcula la F-parcial y se compara con una F-tablas.

Si F-parcial < F-tablas sale del modelo.

El proceso continúa hasta que ya no puedan entrar ni salir variables.

Es importante notar que en este criterio se cuestiona la permanencia de todas las variables incluidas en cada etapa.

Capítulo 4 : Paquete de cómputo.

4.1.Introducción.

En este capítulo se presenta el paquete de cómputo desarrollado para apoyar los cursos se Análisis de Regresión impartidos en la Facultad de Ciencias.

El paquete ha sido desarrollado de manera que el alumno pueda obtener en forma clara y sencilla la información necesaria para interpretarla bajo el contexto estadístico y no perder de vista los puntos importantes de la tcnica que generalmente se pierde al hacer uso de paquetes muy conocidos y ya implantados.

Como todo sistema de cómputo, es necesario analizar la manera a traves de la cual será implantado, tratando que esta sea la óptima.

La manera en la que fué desarrollado el paquete consistió basicamente en tres etapas :

1) Análisis de Datos.
2) Análisis de Manipulación de Datos.
3) Análisis de Tecnicas de Cálculo Estadístico.

En la primer etapa, se analizó qué tipos de datos serían utilizados así como su forma de entrada, llegando a las siguientes conclusiones :

a) Datos de entrada de tipo real.
b) Entrada de Datos a traves de la creación de un archivo.

En la segunda etapa, se analizó que una vez creado un archivo de datos, antes de aplicarles cualquir técnica estadística, es necesario verificar que estos sean correctos permitiendo visualizarlos y modificarlos, si es necesario.

En la etapa tres, se analizó la naturaleza de los cálculos estadísticos utilizados en las técnicas del Análisis de Regresión, con lo cual se llegó a la conclusión, junto con el análisis de las dos etapas anteriores, que es necesario emplear un lenguaje de programación lo más estructurado posible y apegado al desarrollo de cálculos matemáticos, por lo cual se decidió aplicar el lenguaje de programación PASCAL.

4.1.1.Estructura del paquete de Cómputo.

La estructura del paquete de cómputo consta de varios módulos :

1. Módulo de Captura de Datos.
2. Módulo de Matrices.
3. Módulo de Cálculos Estadísticos.
4. Módulo de graficación.

5. Módulo de Ayuda.
6. Módulo Principal.

Los módulos 1 a 5 se identificarán como módulos auxiliares al módulo principal.

1. **Módulo de Captura de Datos.** Se compone de varios submódulos :

1.1. Creación de archivo de datos.
1.2. Abertura de archivo de datos.
1.3. Modificación de archivo de datos.
1.4. Visualización de archivo de datos.
1.5. Selección del conjunto de columnas donde se encuentran las variables explicativas y variable de respueta a partir del archivo de datos.
1.6. Ayuda.

A continuación se describen brevemente cada uno de los sumbmódulos anteriores.

1.1. Se refiere a la creación del archivo de datos, el cual tendrá máximo 20 columnas y 50 renglones.

1.2. Una vez creado el archivo de datos, siempre será necesario abrirlo para despues hacer cualquier manipulación con el mismo, es importante aclarar que si el archivo no es abierto no se podrá proceder a los siguientes submódulos.

1.3. Una vez creado y abierto el archivo de datos será posible hacerle las modificaciones que se deseen para evitar errores en los resultados.

1.4. La visualización del archivo de datos permitirá darse cuenta de posibles errores en el mismo, pudiéndose modificar en el momento que se desee.

1.5. Una vez que se tienen listos los datos, es decir sin errores, se procederá a seleccionar a traves de columnas, la(s) variable(s) explicativa(s) y la variable dependiente que intervendrán en el modelo de regresión, este submódulo permitirá hacer tantas regresiones como se deseen, hecho que facilitará el empleo de métodos de selección de variables.

1.6. En este submódulo unicamente se presenta un texto en el cual se describen los pasos que deben seguirse en el módulo de Captura.

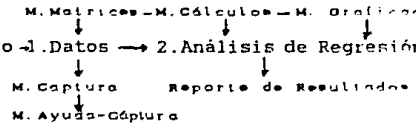
2. **Módulo de Matrices.** Este módulo consta unicamente de operaciones aplicables a matrices. Como se vió en capítulos anteriores, para la obtención de resultados en el Análisis de Regresión se involucran diversas operaciones con matrices tales como :

Cabe señalar que en este paquete de cómputo se utiliza una *unidad* creada fuera del mismo por otro programador cuya función es la de leer y verificar números reales, enteros y caractéres.

Para visualizar más claramente la estructura del paquete se presentan a continuación los siguientes diagrámas :

		Módulos Auxiliares
	<---------	Módulo Captura
	<---------	Módulo Matrices
Módulo Principal	<---------	Módulo Cálculos
	<---------	Módulo Graficación
	<---------	Módulo Ayuda
	<---------	Módulo Especial

El módulo principal está estructurado de la siguiente forma:



4.2. Instrucciones y Especificaciones.

4.2.1. Módulos Auxiliares. A continuación se describen los algoritmos de cada uno de los módulos auxiliares, así como los algoritmos de los submódulos que los componen.

Módulo Captura. Este módulo está formado por varios submódulos, estos a su vez estan desarrollados en *procedimientos* que son instrucciones especiales del lenguaje PASCAL.

Descripción de Procedimientos.
Cada submódulo tiene asociado un procedimiento en PASCAL, a continuación se describen cada uno de ellos.

Submódulo : Creación de Archivo de DAtos.

Asociación en PASCAL: *Procedure Creo_Arch.*

Algoritmo :

Inicio
1. Nombre del Archivo.
2. Lee el nombre del archivo y verifica que efectivamente se le d un nombre (cadena de caracteres).
3. Abre y prepara el archivo para ser creado.

4. Lee las dimensiones del archivo a crear, por renglón y columna.

5. Captura los datos según las dimensiones leídas.
6. Graba el archivo de datos en disco.

Fin.

Submódulo : Abertura de Archivo de Datos.

Asociación en PASCAL : *Procedure Abrir_Arch.*

Algoritmo :

Inicio
1. Lee nombre del archivo que se desea abrir.
2. Verifica que el archivo ya haya sido crado.
3. Abre el archivo.

Fin.

Submódulo : Modificación de Archivo de Datos.

Asociación en PASCAL : *Procedure Modifica_Arch.*

Algoritmo :

Inicio
1. Lee nombre del archivo a modificar.
2. Verifica que el archivo exita y esté abierto.
3. Lee del archivo.
4. Pregunta tipo de modificación.
   4.1.En renglón.
   4.2.En Columna.
   4.3.En Celda.
5. Lee tipo de modificación.
6. Modifica el tipo seleccionado.
7. Graba la modificación en archivo.
8. Cierra el archivo de Datos.

Cabe señalar que si se entra en este módulo será necesario volver a abrir el archivo pues como se pudo observar se cerró al final del algoritmo.

Submódulo : Visualización de Archivo de Datos.

Asociación en PASCAL : *Procedure Listar_Arch.*

Algoritmo :

Inicio
1. Lee nombre del archivo a listar.
2. Verifica que el archivo exista y este abierto.
3. Despliega el archivo de datos.
4. Cierra el archivo.

Fin.

Al igual que en *Modifica_Arch*, si se entra a este submódulo será necesario volver a abrir el archivo para después poder ejecutar cualquier otro submódulo.

Submódulo : Selección del conj. de variables explicativas y variable dependiente.

Asociación en PASCAL : *Procedure Selec_var*.

Algoritmo :

Inicio
1. Verifica que el archivo este abierto.
2. Lee número de variables explicativas.
3. Asigna 1's a la primer columna de la matriz de diseño.
4. Lee columna(s) de la(s) variable(s) explicativa(s).
5. Lee columna de la variable dependiente.

Fin.

Submódulo : Ayuda.

Asociación en PASCAL : *Procedure MenuDatos*(llamado de la unidad Ayuda)

Algoritmo :

Inicio
1. Despliega texto de ayuda para el módulo de Captura.

Fin.

El módulo Captura contiene otro submódulo auxiliar, cuya asociación en PASCAL es *Procedure Cap_Dat*, cuya función es preguntar qué submódulo de los antes descritos se desea ejecutar, permitiendo así ejecutar el submódulo seleccionado.

Algoritmo :

Inicio
1. Pregunta Submódulo a ejecutar.
2. Despliega menú de opciones.
   2.1. Captura.
   2.2. Abrir.
   2.3. Modificar.
   2.4. Listar.
   2.5. Selección de variable(s) explicativa(s) y var. dpendiente.
   2.6. Ayuda.
3. Lee submódulo seleccionado.
4. ejecuta submódulo seleccionado.

Fin.

Módulo Matrices. Al igual que Captura, este módulo está formado por varios submódulos, cada uno asociado a un procedimiento definido en PASCAL.

Descripción de Procedimientos.

Submódulo : Inversa.

Asociación en PASCAL : *Procedure Inversa.*

Algoritmo :

Inicio
1.Inicializa.
1.1.Verifica que la matríz A sea invertible.
2.Invierte (Eliminación Gaussiana).
2.1.Calcula pivote (renglón de referencia).
2.2.Cambia renglones.
2.3.Divide renglones en matríz A y A^{-1}.
2.4.Multiplica y suma renglones en A y A^{-1}.
2.5.Obtiene inversa.
Fin.

este submódulo será utilizado para cualquier matríz que sea invertible,la cual será pasada como parámetro.

Submódulo : Transpuesta.

Asociación en PASCAL : *Procedure Trans.*

Algoritmo :

Inicio
1.Cambia renglones por columnas.
2.Obtiene la transpuesta de una matríz (pasada como parametro).
Fin.

Submódulo : Solución de Sistemas Ax = b.

Asociación en PASCAL : *Bgorro.*

Algoritmo :

Inicio
1.Calcula Bgorro utilizando los procedimentos *inversa* y *Trans.*
Fin.

Dentro del Módulo de Matrices se desarrolló otro submódulo cuya asociación en PASCAL es *Escribe*, cuyo objeto es mandar a escribir las matrices que se requieran.

Módulo Cálculos. Este módulo igual que los anteriores, está formado por varios submódulos asociados a procedimientos y funcionesen PASCAL.

Descripción de Procedimientos.

Submódulo : Media y Desviación.

Asociación en PASCAL : *Procedure Media_Desv.*

Algoritmo :

Inicio
1.Verifica Datos.
2.Calcula suma de Y_i.
3.Calcula suma de Y_i^2.
4.Calcula suma de Y_i entre total de datos : Media de las Y_i.
5.Calcula varianza de las Y_i.
6.Calcula suma de x_i para cada i.
7.Calcula suma de x_i^2 para cada i.
8.Calcula suma de x_i entre total de datos para cada i obteniendo la media de x_i para cada i.
9.Calcula varianza de x_i para cada i.
Fin.

Las medias y las varianzas se imprimen a través de los procedimientos *Esc_medes* y *Esc_var*, desarrollados exclusivamente para escritura.

Submódulo : Matriz de Correlaciones.

Asociación en PASCAL : *Procedure Matcorre*

Algoritmo :

Inicio
1.Verifica Datos.
2.Calcula coeficiente de Correlación de Y con cada x_i.
3.Calcula coeficiente de correalación de x_i con x_j.
Fin.

La matriz de correlaciones es mandada a escribir en el módulo principal a través del submódulo *Escribe* desarrollado en Matrices.

Submódulo : Residuales, SCEMC, Coeficiente de Determinación.

Asociación en PASCAL : *Procedure Calresid_est.*

Algoritmo :

Inicio
1.Calcula Y_i estimadas.
2.Calcula residuales de la forma $Y_i - \hat{Y}_i$.
3.Calcula SCEMC.
4.Calcula residuales studentizados.
5.Calcula coeficiente de Determinación.
Fin.

Los residuales estimados y studentizados son escritos a través de un procedimiento de escritura llamado *resid_est.*

La SCEMC es escrita a través del procedimiento *Esc_SCEMC.*

Submódulo : Suma de las Y_i^2.

Asociación en PASCAL : *Function Sc_de_Y.*

Algoritmo :

Inicio
1.Calcula suma de cuadrados de las y_i al cuadrado.
Fin.

Submódulo : Suma de cuadrados de las Y_i corregidas por la media.

Asociación en PASCAL : *Funtion SC_de_Ycorr.*

Algoritmo :

Inicio
1.Calcula suma de las Y_i al cuadrado corr. por la media.
Fin.

Submódulo : Estadística de Durbin-Watson.

Asociación en PASCAL : *Procedure Durbin_Watson.*

Algoritmo :

Inicio
1.Calcula sumade $(\varepsilon_i - \varepsilon_{i-1})^2$ $i = 2,\ldots,n$.
2.Calcula sumade ε_i^2 $i = 1,\ldots,n$.
3.Divide 1./2.obteniendo a la estadística deDurbin-Watson.
Fin.

Submódulo : Tabla de Análisis de Varianza. (Para Sig. Global y $C\beta=\gamma$).

Asociación en PASCAL : *Procedure Tanova.*

Algoritmo :

Inicio
1.Genera Matriz $C_{r\times p}$.
2.Genera vector $\gamma_{r\times 1}$.
3.Calcula $(X^tX)^{-1}C^t$.
4.Calcula Inversa de $C(X^tX)^{-1}C^t$.
5.Calcula $C\hat{\beta} - \gamma$.
6.Calcula $(C\hat{\beta} - \gamma)^t$.
7.Calcula $(C\hat{\beta} - \gamma)^t[C(X^tX)^{-1}C][(C\hat{\beta} - \gamma)$
8.Calcula Suma de Cuadrados del Error bajo la Hip. Ho.
9.Calcula Suma de Cuadrados del Error bajo el modelo reducido.
10.Calcula Tabla de Análisis de Varianza TANOVA.
Fin.

Submódulo F's Parciales.

Asociación en PASCAL : *Procedure F-Parcial.*

El algoritmo es el mismo que en *Tanoma*, la única diferencia radica en las dimensiones de la matriz C y el vector γ, es decir C_{kxp} y γ_{kx1}.

En este módulo se desarrolló el procedimiento *escribe_vector*, para escribir los vectore que se necesiten posteriormente.

Submódulo : Cofactores.

Asociación en PASCAL : *Procedure Cofactores.*

Algoritmo :

Inicio.
1. Lee i,j para calcular i,j.
2. Obtiene matriz auxiliar de Matcorre.
3. Calcula determinante de Matriz auxiliar.
4. reporta cofactor.

Fin.

Módulo Graficación. Esta formado por varios submódulos asociados a procedimientos en PASCAL.

Descripción de Procedimientos.

Submódulo : Ordenamiento.

Asociación en PASCAL : *Procedure Ordena.*

Este submódulo se encarga de ordenar una serie de datos, lo cual es necesario en el caso de graficación de Papel Normal,el algoritmo utilizado para el ordenamiento fu el **Quick Sort**, tomado de *Tenenbaum and Augenstein, (1981).*

Submódulo : Cuantiles.

Asociación en PASCAL : *Procedure Cuantil.*

Algoritmo :

Inicio
1.Calcula cuantiles : 100*(i - 0.5)/n.
Fin.

Submódulo : Gráfica de Cp-Mallows.

Asociación en PASCAL : *Procedure Cp_Mallows.*

Algoritmo :

Inicio
1.Lee número de regresiones (p).
2.Lee valores de Cp.
3.Lee número de variables explicativas.
4.Calcula máximos y mínimos de Cp y p.
5.Calcula abscisas y oordenadas.

```
          6.Calcula escalas para los ejes X y Y.
          7.Grafica Cp vs p.
        Fin.
```

Submódulo : Papel Normal.

Asociación en PASCAL : *Procedure PNormal*

```
Algoritmo :
          Inicio
           1.Pregunta gráfica que desea.
             1.1.Ri vs Ki. (donde Ki son los cuantiles)
             1.2.Ri vs Xi.
             1.3.Ri vs Yi.
           2.Lee gráfica seleccionada.
           3.Calcula máximos y minimos de Ri  y  Ki,  o  xi  o  Yi,
             según sea el caso.
           4.Calcula abscisas y oordenadas.
           5.Calcula escalas para los ejes X y Y.
           6.Grafica Ri vs Ki, xi o Yi según sea el caso.
          Fin.
```

Submódulo : Graficación.

Asociación en PASCAL : *Procedure GrRivsKi*

Este submódulo fué creado con el fin de llamar a cualquiera de los anteriores.

```
Algoritmo :
          Inicio
           1.Pregunta tipo de Gráfica.
             1.1.Ki vs Ri.
             1.2.Ri vs xi.
             1.3 Ri vs Yi
             1.4.Cp-Mallows.
           2.Lee tipo de gráfica.
           3.Si tipo = 1 2 ó 3 pregunta :
             3.1.Ri Yi - Yi.
             3.2.Ri studentizados.
             3.3.Lee tipo de Ri.
             3.4.Despliega Gráfica.
            Si tipo = 4
               Despliega gráfica de Cp vs p.
Fin.
```

Módulo Principal. Este módulo agrupa a todos los anteriores a los cuales llama a ejecutar en el momento que se requiera.

Algoritmo del módulo Principal :

```
Inicio
 1.Verifica Datos.
 2.Despliega menu Principal :
     1.Datos.
     2.Análisis de Regresión.
     3.Ayuda.
     4.Salir.
 3.Lee opción.
  3.1.Si 1 : ejecuta Cap_dat (Módulo de Captura).
  3.2.Si 2 : Calcula :
                        medias
                        varianzas
                        Transpuestas
                        Matríz de correlación
                        Inveras
                        B-gorro
                        Residuales
                        Residuales studentizados
                        Ejecuta_menu de Regresión :
                         Ejecuta Presenta_Menu :
                          1.Matríz de Datos y vector Y
                          2.B-gorro
                          3.Cofactores
                          4.Matríz de Correlación
                          5.Hipótesis Cβ = γ tabla Anova
                          6.F-Parciales
                          7.Residuales
                          8.Matríz inversa de X'X
                          9.SCE bajo el modelo completo
                          10.Suma de cuadrados de Y
                          11.Suma de cuadrados de Y
                             corregida por su media
                          12.Medias
                          13.Varianzas
                          14.Graficación
                          15.Durbin_Watson
                          16.Coeficiente de Determinación
                          17.Fin
                        Lee opción
                         Si opción es :
                          1 : Ejecuta Escribe Datos y escribe
                              vector Y.
                          2 : Ejecuta Esc_vector B-gorro.
                          3 : Ejecuta Cofactores.
                          4: Ejecuta Escribe Matríz  de
                              Correlación.
                          5 : Ejecuta Tanova.
                          6 : Ejecuta F-Parcial.
                          7 : Ejecuta Resid_est.
```

8 : Ejecuta *Escribe* inversa de X^tX.
9 : Ejecuta *Esc_SCEMC*.
10 : Escribe *Sc_y*.
11: Escribe *Sc_de_ycorr*.
12: Ejecuta *esc_medes*.
13: Ejecuta *esc_var*.
14: Ejecuta *GrRiusKi*.
15: Ejecuta *Durbin_Watson*.
16: Escribe coef. de Determinación.
17: Sale de *Ejecuta_menu*.

3.3. Si 3 :Despliega texto de ayuda en el cual se encuentran las instrucciones del paquete.
3.4. Si 4 : Sale de menu principal y del paquete.
Fin.

En el módulo principal se desarrollaron dos procedimientos, *presenta_menu* y *ejecuta_menu*, el primero despliega el menú de 1 a 16 opciones, y el segundo llama a los procedimientos necesarios para la ejecución de cada una de las opciones de *presenta_menu*.

Capítulo 5 : Aplicación del paquete mediante los datos de HALD.

5.1.Hald.

Para ilustrar el uso del paquete, se escogieron los datos de *Hald*, presentados en *Industrial and Engineering Chemistry*, 24, 1932, 1207-14, tabla I, por *H. Woods, H.H. Steinour and H.R. Starke*.

En estudios hechos sobre el calentamiento generado durante el endurecimiento del cemento de Portland (Oregon,EUA), se supone una función de composición química. Las variables medidas son :

x_1 = Cantidad de Tricalcio de Aluminio.

x_2 = Cantidad de Tricalcio de Silicato.

x_3 = Cantidad de Tetracalcio de Aluminio Férrico.

x_4 = Cantidad de Dicalcio de Silicato.

x_5 = Y = Calentamiento desarrollado en calorías por gramo de cemento.

x_i	x_1	x_2	x_3	x_4	x_5
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

Datos de Hald.

Las variables x_1 a x_4 son medidas como porcentajes de hulla en el cemento producido.

Se supone que el calor generado durante el endurecimiento es una función lineal de las cuatro variables x_1 a x_4, el modelo de RLM es

$$Y_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i}+\beta_4 x_{4i}+\varepsilon_i \qquad i = 1,\ldots,n$$

Como se vió en capítulos anteriores el error ε es un vector aleatorio con Dn Normal $(0, I\sigma^2)$.

Los métodos de Regresión son usados para estimar β_i y probar hipótesis acerca de ellas.

5.2.Aplicación del paquete a los datos de Hald.

A continuación se exponen los pasos a seguir para ejecutar el paquete mediante los datos de Hald.

1.Creación del archivo de datos. En esta parte se crea el archivo correspondiente a los datos de Hald los cuales podrán ser manipulados de acuerdo a las necesidades del usuario. Los pasos a seguir son los siguientes :

i) Dentro del menú principal se debe entrar a la opción Datos.
ii) En el menú de Datos se debe entrar a la opción Crear.
iii) Dentro de la opción Crear, se leen el nombre del archivo así como el número de renglones y el de columnas, los cuales no deberán ser mayores a 50 y 20 respectivamente, una vez que se proporcionaron estos datos se procede a la captura, la cual terminará automaticamente cuando se llegue al número de renglones leído. En ese momento, se despliega una lista de los datos para que el usuario verifique que estan coorectos, de no estarlos puede modificarlos en la opción de Modificar que se presenta más adelante, sino hubo errores el archivo queda listo para ser abierto, cabe recordar que como se mencionó en el capítulo 4 una vez que se crea un archivo, ya no es necesario volverlo a crear cuando se vuelva a correr el paquete .

2.Abriendo el archivo. Una vez que se creó el archivo de datos se debe abrir el archivo, para poder pasar a cualquiera de las siguientes opciones, recordando que si se entra a Listar o a Modificar se debe regresar a abrir el archivo, una vez que se abrió el siguiente paso es escoger las columnas que representan las variables explicativas y la variable de respuesta.

3.Listando el archivo. Si se desea ver los datos, basta con dar el nombre del arcvhivo para poderlo ver, recordando que después de entrar a esta opción es necesario regresar a abrir el archivo.

4.Modifcando el archivo. Si existe algún error en los datos a traves de esta opción se pueden hecer las modificaciones que sean necesarias sólo basta con dar el nombre del archivo, posteriormente aparecerá otro menú el cual se peueden hacer modoficaciones por renglón, columna o celda. Una vez que se hicieron los cambios en el archivo se debe regresar a abrirlo nuevamente para asegurar que se tome el archivo ya modoficado con el cual se trabajará a lo largo del paquete.

5.Selección de Columnas. Una vez que se tiene el archivo creado y abierto se procede a escoger las columnas donde se encuentran las variables explicativas y la variable dependiente, en el caso de Hald, la variable de respuesta se metió en la columna 5 y las variables explicativas x_1 a x_4 en las columnas 1 a 4, respectivamente,con esta opción se da por terminado lo referente a la opción Datos del menú principal. Cabe señalar que si se presentan dudas de esta opción dentro de la misma existe una de ayuda para facilitar el segumiento de la misma.

Una vez que se llevó a cabo la opción Datos del menú principal el siguiente paso es pasar a la opción 2.Análisis de Regresión, en la cual antes de entrar al menú presenta algunos avisos con respecto a la impresión de resultados así como la salida de cada una de las opciones del menú de regresión. Inmediatamente después, se presenta en menú de Análisis de Regresión toda la información que se puede obtener a través del mismo, en el caso de Hald se obtuvieron los resultados que a continuación se describen.

5.3.Resultados obtenidos a través del paquete de Cómputo.
Opción : Matriz de Diseño y vector de variables de respuesta.

$$\begin{bmatrix} 1.0 & 7.0 & 26.0 & 6.0 & 60.0 \\ 1.0 & 1.0 & 29.0 & 15.0 & 52.0 \\ 1.0 & 11.0 & 56.0 & 8.0 & 20.0 \\ 1.0 & 11.0 & 31.0 & 8.0 & 47.0 \\ 1.0 & 7.0 & 52.0 & 6.0 & 33.0 \\ 1.0 & 11.0 & 55.0 & 9.0 & 22.0 \\ 1.0 & 3.0 & 71.0 & 17.0 & 6.0 \\ 1.0 & 1.0 & 31.0 & 22.0 & 44.0 \\ 1.0 & 2.0 & 54.0 & 18.0 & 22.0 \\ 1.0 & 21.0 & 47.0 & 4.0 & 26.0 \\ 1.0 & 1.0 & 40.0 & 23.0 & 34.0 \\ 1.0 & 11.0 & 66.0 & 9.0 & 12.0 \\ 1.0 & 10.0 & 68.0 & 8.0 & 12.0 \end{bmatrix} \quad \begin{bmatrix} 78.5 \\ 74.3 \\ 104.3 \\ 87.6 \\ 95.9 \\ 109.2 \\ 102.7 \\ 72.5 \\ 93.1 \\ 115.9 \\ 83.8 \\ 113.3 \\ 109.4 \end{bmatrix}$$

Opción :Vector de Párametros Estimados $\hat{\beta}$. De la opción 2.B-gorro,se obtuvo :

$$\underset{\sim}{\hat{\beta}} = \begin{bmatrix} 62.4054 \\ 1.5511 \\ 0.1019 \\ -0.1441 \end{bmatrix}$$

De donde $\underset{\sim}{\hat{Y}} = X\underset{\sim}{\hat{\beta}}$

Opción : Matríz de Correlaciones. ($\underline{Y}$ con x_1, x_2, x_3, y x_4)

	Y	X_1	X_2	X_3	X_4
Y	1.0000	0.7907	0.8163	-0.5347	-0.8213
X_1	0.7907	1.0000	0.2286	-0.5347	-0.2454
X_2	0.8163	0.2286	1.0000	-0.1392	-0.9730
X_3	-0.5347	-0.8741	-0.1392	1.0000	0.0295
X_4	-0.8213	-0.2454	-0.9730	0.0295	1.0000

De la matriz de Coorrelaciones se puede observar que las variables más coorrealcionadas con Y son x_2 y x_1, mientras que x_4 presenta una correlación negativa con Y, así también x_4 esta altamente coorrelacionada con x_2, estos resultados son de gran importancia en selección de variables.

Es importante mencionar que la primera columna y el primer renglón de esta matriz es la variable Y.

Opción : Tabla Anova

Prueba de Significancia Global.

Se verá si $\beta_1, \ldots, \beta_4 = 0$ mediante la prueba de significancia global :

$$Ho : C\underline{\beta} = 0 \qquad vs. \qquad Ha : C\underline{\beta} \neq 0$$

con

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Tabla de Análisis de Varianza

f.var	g.l.	SCE	CME	fcalc.
Ho	4	2667.9008	666.9752	111.4823
MC	8	47.8623	5.9828	
MR	12	2715.7631		

No.de datos : 13 No.de parámetros : 5
Coeficiente de Determinación : 0.9824

De la tabla de Análisis de Varianza se obtuvo una F-calculada de 111.29 y comparándola con una F de tablas con 4 y 8 grados de libertad con un α de 0.05 se obtuvo que :

$$F(4,8) = 3.84 < \text{F-Calculada} = 111.29 \quad \alpha = 0.05$$

Del resultado anterior se puede afirmar que existe suficiente evidencia en contra de alguna βi sea igual a cero por lo tanto el modelo de regresión es significativo.

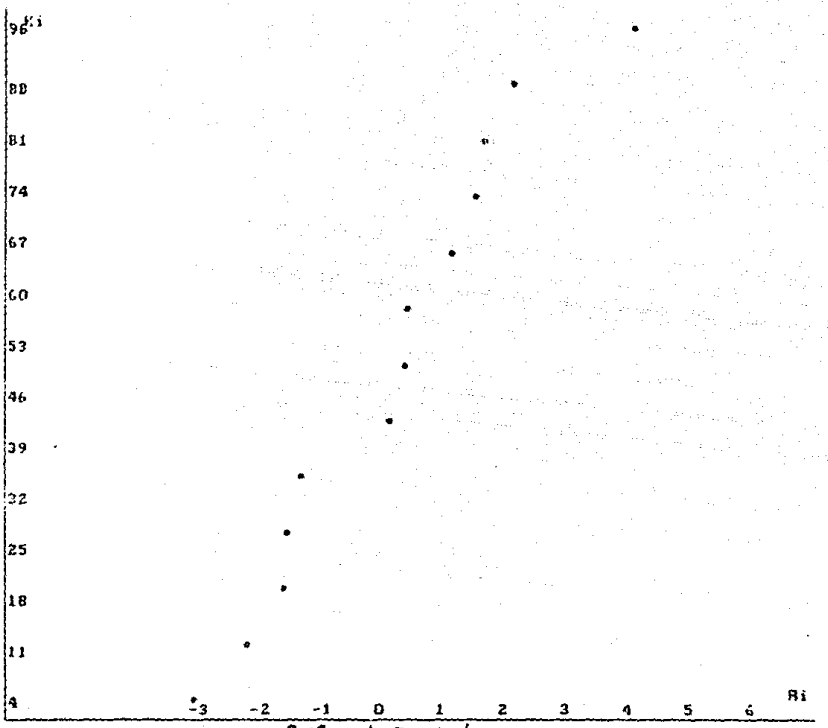
Opción :Coeficiente de Determinación.

$R^2 = 0.9824$

El coeficiente de Determinación está indicando que más del 98% de la variabilidad total del as Yi's queda explicada por el modelo.

Opción : Residuales. Aquí se presentan los residuales, de la forma $Y_i - Y_i$ y los studentizados. Posteriormente se puede pasar a la opción de **Graficación,** mediante la cual se pueden obtener diferentes gráficas de residuales con las cuales se puede hacer un análisis de supuestos.

I	Y	YESTIMADA	RESIDUAL	RESID.STUD
1	78.5000	78.4952	0.0048	0.0029
2	74.3000	72.7888	1.5112	0.7566
3	104.3000	105.9709	-1.6709	-1.0503
4	87.6000	89.3271	-1.7271	-0.8411
5	95.9000	95.6492	0.2508	0.1279
6	109.2000	105.2746	3.9254	1.7148
7	102.7000	104.1487	-1.4487	-0.7445
8	72.5000	75.6750	-3.1750	-1.6878
9	93.1000	91.7217	1.3783	0.6708
10	115.9000	115.6185	0.2815	0.2103
11	83.8000	81.8090	1.9910	1.0739
12	113.3000	112.3270	0.9730	0.4634
13	109.4000	111.6943	-2.2943	-1.1241



Gráfica de Residuales.

Selección de Variables.

Todas las regresiones posibles.

Para poder aplicar este método mediante el paquete se deben ir eligiendo las columnas que vayan involucrando cada uno de los modelos, como el número de parámetros es igual a 5 se tendrán 2^4 modelos de regresión.

Para los modelos que involucran un cierto número de variables explicativas se tendrá que seleccionar en el módulo de Datos las columnas de la variable de respuesta y la(s) correspondiente(s) a la (s) variable(s) explicativa(s), así por ejemplo si se va a ajustar el modelo con x_1 se eligirá la columna donde se encuentra esta variable junto con la columna de la variable y, después de esto se regresa al menú principal para ir a la opción de Análisis de Regresión, donde se obtendrá la información requerida para ese modelo.

Cabe aclarar que cada vez que se dese ajustar un modelo será necesario regresar al módulo de Datos para dar el nuevo conjunto de columnas que contienen a las variables de interés.

Tabla 1. Todas las regresiones posibles.

variables	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2	$\bar{R}^2$
x_1	81.48	1.86				0.583	0.491
x_2	57.42		0.78			0.666	0.636
x_3	110.20			-1.25		0.285	0.220
x_4	889.9				-0.78	0.674	0.644
x_1 x_2	52.58	1.46	0.66			0.978	0.974
x_1 x_3	72.85	2.31		0.40		0.548	0.457
x_1 x_4	109.10	1.44			-0.61	0.972	0.967
x_2 x_3	72.07		0.73	-1.00		0.847	0.816
x_2 x_4	94.16		0.31		-0.45	0.680	0.616
x_3 x_4	191.28			-1.70	-0.72	0.995	0.992
x_1 x_2 x_3	48.19	1.69	0.65	0.25		0.982	0.978
x_1 x_2 x_4	71.65	1.45	0.41		-0.23	0.982	0.976
x_1 x_3 x_4	111.68	1.05		-0.41	-0.64	0.981	0.975
x_2 x_3 x_4	208.64		-0.97	-1.44	-1.55	0.972	0.965
x_1 x_2 x_3 x_4	62.98	1.55	0.51	0.10	-0.14	0.982	0.978

Criterio de la R^2.

De la tabla 1 se puede observar que del conjunto de modelos que contienen una sola variable explicativa la R^2 mayor es la correspondiente a la variable x_4, por lo que se escoge el modelo que contiene a esta.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_4 x_4 \qquad R^2 = .674$$

Del conjunto de modelos que contienen dos variables explicativas, existe un empate para los modelos que contienen a las variables x_1, x_2, x_3 y x_1, x_2 y x_4 ya que presentan una R^2 de .982, en este caso habría que analizar la correlación entre esta variables.

A partir de la R^2 se pueden ir calculando las R^2 ajustadas y tomar el modelo que tenga la R^2 ajustada mayor.

Opción :Graficación : Cp-Mallows.

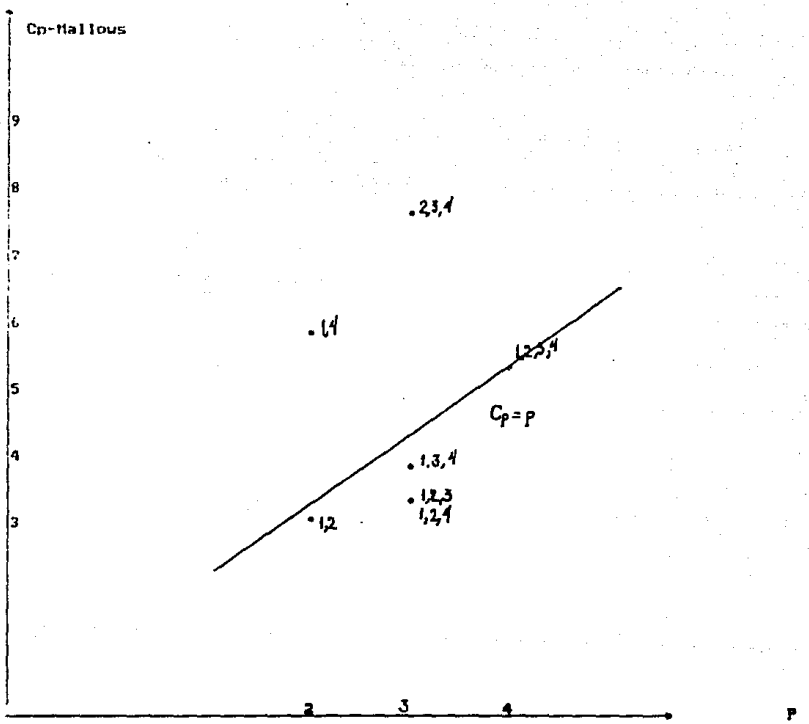
De acuerdo a la definición de la Cp vista en el capítulo 3, se tomaron algunos modelos para describir el uso del paquete, obtienendo los siguientes valores :

SCE_{MC} = 5.983

Para x_1, x_4	SCE_{MC} = 1265.6	Cp = 5.5
x_1, x_2	SCE_{MC} = 57.9	Cp = 2.7
x_2, x_3, x_4	SCE_{MC} = 73.8	Cp = 7.3
x_1, x_3, x_4	SCE_{MC} = 50.8	Cp = 3.5
x_1, x_2, x_4	SCE_{MC} = 48.0	Cp = 3.0
x_1, x_2, x_3	SCE_{MC} = 48.1	Cp = 3.0

Para calcular cada uno de los valores de las Cp, se debe ajustar el modelo de acuerdo a las variables que se estén involucrando, es decir se debe ir al módulo de Datos y dar las columnas adecuadas de acuerdo a las variables que contenga cada modelo pues para cada uno se reportan en el paquete las SCE_{MC} correspondientes.

En esta opción el alumno deberá calcular las Cp para cada uno de los modelos, pues el paquete aporta la información necesaria para hacerlo. Bajo este criterio se llega a lo siguiente :


Cp-Mallous
9
8
7
6
5
4
3
2
3
4
P
2,3,4
1,4
1,2,3,4
$C_p = P$
1,3,4
1,2,3
1,2,4
1,2

De la gráfica (opción Graficación-Cp Mallows) se observa que los mejores subconjuntos son :

$(x_1\ x_2)$ y el $(x_1, x_2, x_3, y\ x_4)$

Para aplicar métodos de selección de variables es necesario ir haciendo regresiones con diferentes subconjuntos de variables, mediante el paquete se pueden obtener estas regresiones para lo cual se deben seguir los siguentes pasos, primero se regresa al menú principal dentro del cual se deberá entrar de nuevo a la opción Datos, una vez que se encuentra en dicha opción, se procede a hacer una nueva selección de las columnas donde se encuentran las variables explicativas según el subconjunto que se haya seleccionado, la variable de respuesta , en este caso x_5, sigue estando en la columna 5 del archivo, si por ejemplo se desea hacer una regresión con las variables x_1, x_7 y x_8 se deberán dar las columnas 1,2 y 3 respectivamente.

Backward. :

Opción : F-parcial. Con esta opción será posible aplicar métodos de selección de variables, pues según el modelo que se tenga se reportan las respectivas F's parciales facilitando así la aplicación del método de selección.

En el caso de Backward, se debe empezar con el modelo completo, para lo cual se deben de proporcionar las columnas correspondientes que contiene dicho modelo para después pasar al menú de Análisis de Regresión y en este obtener las F's parciales adecuadas.

Paso 1.

Modelo Completo : $\hat{Y} = 67.41 + 1.55\ X_1 + 0.51\ X_2 + 0.10\ X_3 - 0.14\ X_4$

Variable	F-parcial ($\beta_j = 0$)
x_1	4.33
x_2	0.50
x_3	0.02
x_4	0.04

La F-parcial de menor valor es F = 0.02
Comparando con una F-Tablas

$$F^{.05}_{(4,8)} = 3.84 > 0.02 \quad \therefore$$

Se elimina la variable x_3

En este momento se debe regresar al módulo de Datos para dar el nuevo subconjunto de variables (x_1, x_2 y x_4).

Paso 2.

Nuevo Modelo : $\hat{Y} = 71.64 + 1.45\ X_1 + 0.41\ X_2 - 0.23\ X_4$

Variable	F-parcial ($\beta_j = 0$)
x_1	154.01
x_2	5.03
x_4	1.86

La F-parcial de menor valor es F = 1.86
Comparando con una F-tablas

$$F^{.05}_{(3,9)} = 3.86 > 1.86 \quad \therefore$$

Se elimina la variable x_4

$R^2 = 0.982$

Despues de haber dado las columnas del nuevo subconjunto de variables se procede con el siguiente paso.

Paso 3.

Nuevo Modelo : $\hat{Y} = 52.57 + 1.46\ X_1 + 0.66\ X_2$

Variable	F-parcial ($\beta_j = 0$)
x_1	146.52
x_2	208.58

La f-parcial de menor valor es F = 146.52
Comparando con una F-Tablas se tiene que

$$F^{0.05}_{(2,10)} = 4.86 < 146.52 \quad \therefore$$

No sale ninguna variable.

$R^2 = .97$

Para este subconjunto de variables :

$$Y_i = \beta_0 + \beta_1 \; x_{1i} + \beta_2 \; x_{2i} + \varepsilon_i$$

El vector de parámetros estimados es :

$$\hat{\beta} = \begin{bmatrix} 32.37 \\ 1.40 \\ 0.00 \end{bmatrix}$$

Obteniendo el nuevo modelo :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Este modelo se tomará como el más significativo.

Cabe mencionar que para aplicar otros métodos de Selección de Variables, como Stepwise y Forward, el paquete proporciona los cofactores a traves de los cuales se calculan los coeficientes de correlación parcial, mismos que son utilizados bajo estos metodos, por ejemplo si se desea calcular $r_{y\, x_1 x_4}$ el paquete proporciona los cofactores correspondientes, recordando que para este conjunto de variables la matriz de correlaciones será :

$$\begin{bmatrix} 1.0 & 0.78 & -0.82 \\ 0.78 & 1.0 & -0.24 \\ -0.82 & -0.24 & 1.0 \end{bmatrix}$$

$R_{yx_1} = (-1)^3 a_{12}$ eliminando el renglón 1 y columna 2

donde $a_{12} = \begin{vmatrix} 0.78 & -0.24 \\ -0.82 & 1.0 \end{vmatrix} = -0.5291$

Se llega a que

$$R_{yx_1} = R_{12} = -(-0.5291) = 0.5291$$

De la misma forma se obtiene

$$R_{22} = 0.32 \quad , \quad R_{11} = 0.93$$

$$r_{y\, x_1 .\, x_4} = - R_{12} / \sqrt{R_{11} R_{22}} = 0.9566$$

Conclusiones.

En este paquete se intentó proporcionar elementos a través de los cuales se llegara de una manera más rápida a los cálculos necesarios para el Análisis de Regresión, tratando de no perder de vista la teoria en la que se basa esta tecnica estadistica.

El paquete de cómputo se elaboró abarcando lo mínimo que se utiliza en un curso de Análisis de Regresión aunque cabe señalar que aún puede ampliarse mucho más. Sin embargo con lo tratado en este trabajo se puede llegar a realizar numerosas aplicaciones.

Este trabajo puede tomarse como material didáctico por quienes imparten los cursos de Análisis de Regresión así como tambien puede ser de utilidad para quienes tienen interes en la materia.

A través de este paquete los alumnos de Análisis de Regresión e incluso de otras materias relacionadas con la materia podrán obtener información, interpretarla y no perder de vista la forma en la que se llega a la misma, ya que proporciona las herramientas necesarias para calcular las estadisticas relevantes permitiendo así, no perder de vista los métodos de cálculo que frecuentemente se olvidan cuando se utilizan paquetes reconocidos, que más bien son recomendables para quienes dominan la materia y no para los alumnos pertenecientes a un curso de la misma.

En resumen se cubrieron tres objetivos principalmente :

- La presentación muy general de la teoria del Análisis de Regresión, lo cual puede servir en momento dado como material de consulta para quienes esten interesados en la materia. Si se desea profundizar más en la misma, la bibliografía de este trabajo puede servir de referencia.

- La aplicación de un paquete de cómputo que permita a los alumnos de los cursos de Análisis de regresión, obtener resultados perdiendo menos tiempo que si tuvieran que hacer todos los calculos manualmente y sobre todo sin perder de vista la teoria tanto matemática como estadística, lo cual es una de las inquietudes de los profesores que imparten dichos cursos.

- La aplicación de un paquete de cómputo a una de las técnicas estadisticas más importantes que se conocen,a través del cual se obtiene la información sufuciente para hacer una inferencia acerca del fenómeno que se esté estudiando en un momento dado.

Cabe señalar que hoy en dia la teoria de la computación crece a pasos agigantados por lo cual este trabajo puede ser retomado en el futuro y ser mejorado con las nuevas inovaciones que esta teoría computacional ofrezca.

Por último, no se pretende competir con ningún paquete estadistico conocido y mucho menos cubrir todos los casos que podrían enumerarse en Regresión.

APENDICE I.

Apéndice 1.

1. La Normal Multivariada.
La función de densidad de probabilidad Normal, en el caso multivariado, se expresa como sigue :

$$f(Y,\underline{\theta},\Sigma) = |2\pi\Sigma|^{-1/2}\exp[-1/2(Y-\theta)^t\Sigma^{-1}(Y-\theta)]$$

$\forall\ Y \in \mathbb{R}^p$, $\underline{\theta} \in \mathbb{R}^p$ y $_{p\times p}\Sigma$ definida positiva.

Resultados respecto a la Normal Multivariada.

1. $E(\underline{Y}) = \underline{\theta}$ y $V(\underline{Y}) = \Sigma$

2. $(\underline{Y} - \underline{\theta})^t\Sigma^{-1}(\underline{Y} - \underline{\theta}) \quad X^2_{(p)}$

3. Si $\underline{Y}\ N_p(\underline{\theta}, \Sigma)$ entonces $(\underline{Y} - \underline{\theta}) \sim N_p(\underline{0}, \Sigma)$

2. Formas Cuadráticas.

Si Y es un vector de $_{p\times 1}$ y A es una matriz de $_{p\times p}$ simétrica entonces la forma cuadrática

$$\underline{Y}^t A\ \underline{Y} = \sum_{i=1}^{n}\sum_{j=1}^{n} Y_i Y_j\ a_{ij}$$

2.1.Se dice que $\underline{Y}^t A\ \underline{Y}$ es positiva definida (p.d.) $\leftrightarrow$ $\underline{Y}^t A\ \underline{Y} > 0 \quad \forall\ \underline{Y} \neq 0$

2.2.Se dice que $\underline{Y}^t A\ \underline{Y}$ es positva semidefinida $\leftrightarrow$ $\underline{Y}^t A\ \underline{Y} \geq 0 \forall\ \underline{Y}$ y $\exists\ \underline{Y} \neq 0$ tal que $\underline{Y}^t A\ \underline{Y} = 0$

2.3.A es positiva definida si $\underline{Y}^t A\ \underline{Y} > 0 \quad \forall\ \underline{Y}$
A es positiva semidefinida si $\underline{Y}^t A\ \underline{Y} \geq 0 \quad \forall\ \underline{Y}$ y $\exists\ Y \neq 0$ tal que $\underline{Y}^t A\ \underline{Y} = 0$

2.4.Si P es no singular y A es positiva definida entonces $P^t A\ P$ es p.d.

2.5.Una condición necesaria y suficiente para que A (simétrica) sea p.d. es que exista P, una matriz no singular tal que $A = P^t P$.

2.6.Una condición necesaria y suficiente para que A sea p.d. es que los menores principales sean positivos (Det A > 0).

2.7.A es definida negativa si los menores principales alternan el signo.

2.8.Si $A_{n\times m}$ es una matriz de rango $_{n\leq m}$ entonces A^tA es positiva definida y AA^t es positiva semidefinida.

2.9.Si A es positiva semidefinida entonces $Tr(A) \geq 0$.

2.10.Si A es positiva definida A^{-1} tambien lo es.

2.11. Los elementos diagonales de una matriz positiva definida son todos positivos.

3. Distribución de Formas Cuadráticas.

3.1. Si $Y \sim N_n(0, I)$ una condición necesaria y suficiente para que la forma cuadrática $\underline{Y}^t A \underline{Y} \sim \chi^2_{(k)}$ es que A sea una matriz idempotente de rango k.

3.2. Si $Y \sim N_n(\underline{\theta}, I)$ entonces $\underline{Y}^t A \underline{Y} \sim \chi^2_{(r,\lambda)}$ con $\lambda = \underline{\theta}^t A \underline{\theta}/2 \leftrightarrow$ A es una matriz idempotente de rango k.

3.3. Si $Y \sim N_n(\underline{\theta}, \sigma^2 I)$ entonces $Y^t A Y/\sigma^2 \sim \chi^2_{(k,\lambda)}$ con $\lambda = \underline{\theta}^t A \underline{\theta} \leftrightarrow$ A es una matriz idempotente de rango k.

3.4. Si $Y \sim N_n(0, \Sigma)$, entonces $\underline{Y}^t B \underline{Y} \sim \chi^2_{(k)} \leftrightarrow B\Sigma_r^+$ es idempotente de rango k.

3.5. Si $Y \sim N_n(\underline{\theta}, \Sigma_r^+)$ entonces $\underline{Y}^t B \underline{Y} \sim \chi^2_{(k,\lambda)}$ con $\lambda = \underline{\theta}^t B \underline{\theta}/2$ y B de rango k $\leftrightarrow B\Sigma_r^+$ es idempotente.

3.6. Distribución de Ji-cuadrada no central (χ^2)

3.6.1. Si $Y \sim N_n(0, I)$ entonces $\underline{Y}^t \underline{Y} = \sum_{i=1}^{n} Y_i^2 \sim \chi^2_{(n)}$

3.6.2. Si $Y \sim N_n(\underline{\theta}, I)$ entonces $\underline{Y}^t \underline{Y} \sim \chi^2_{(n,\lambda)}$, es decir, $\underline{Y}^t \underline{Y}$ es una χ^2 no central con n grados de libertad y parámetro de no centralidad $\lambda = \underline{\theta}^t \underline{\theta}/2$

3.6.3. Si $w_1, w_2, \ldots, w_k$ son independientes y si $w_i \sim \chi^2_{(p_i, \lambda_i)}$ $(i=1,\ldots,k)$, entonces $w = \sum_{i=1}^{k} w_i \sim \chi^2_{(p,\lambda)}$ con $p = \sum_{i=1}^{k} p_i$ y $\lambda = \sum_{i=1}^{k} \lambda_i$

3.6.4. Si $Y \sim N_n(\underline{\theta}, \sigma^2 I)$, entonces $\underline{Y}^t \underline{Y}/\sigma^2 \sim \chi^2_{(n,\lambda)}$ con $\lambda = \underline{\theta}^t \underline{\theta}/2\sigma^2$

3.6.5. Si $Y \sim N_n(\underline{\theta}, D)$ con D diagonal, entonces $Y^t D^{-1} Y \sim \chi^2_{(n,\lambda)}$ con $\lambda = \underline{\theta}^t D^{-1} \underline{\theta}/2$

4. Independencia de Formas Cuadráticas.

4.1. Si $\underline{Y} \sim N_p(\underline{\theta}, \Sigma)$ entonces $\underline{Y}^t A \underline{Y}$ y $\underline{Y}^t B \underline{Y}$ son independientes $\leftrightarrow A \Sigma B = 0$

4.2. Si $Y \sim N(\underline{\theta}, \sigma^2 I)$ la forma cuadrática positiva semidefinida $\underline{Y}^t A \underline{Y}$ y $\underline{Y}^t B \underline{Y}$ son independientes si $Tr(AB) = 0$.

APENDICE 2.

Lo cual significa que la variable y y x_1 dado x_4 estan altamente correlacionadas.

Aplicando los métodos de selección antes mencionados se toma la variable más correlacionada con Y, y se continua con el proceso según lo indiquen dichos métodos .

En el menú de Analisis de Regresión también se reporta la estadística de Durbin y Watson la cual toma el valor de 2.0526.

Apéndice 2.

Independencia entre las variables aleatorias SCE_{MC} y SCE_{H_0}.

Nótese que

$SCE_{H_0} = (C\hat{\beta} - \gamma)^t [C (X^tX)^{-1}C^t)](C\hat{\beta} - \gamma)$ con $\hat{\beta} = (X^tX)^{-1}X^tY$

entonces

$SCE_{H_0} = [C(X^tX)^{-1}X^tY-\gamma)]^t [C(X^tX)^{-1}C^t]^{-1} [C(X^tX)^{-1}X^tY-\gamma]$

ahora

$[C(X^tX)^{-1}X^t]XC^t = CC^t \rightarrow [C(X^tX)^{-1}X^t]XC^t(CC^t)^{-1} = I$

en donde $(CC^t)^{-1}$ existe pues C es de rango completo.

La SCE_{H_0} puede entonces expresarse como :

$SCE_{H_0}=[C(X^tX)^{-1}X^t(Y - XC^t(CC^t)^{-1}\gamma)]^t [C(X^tX)^{-1}C]^{-1} [C(X^tX)^{-1}X [Y-XC^t(CC^t)^{-1}]]$

$=[Y-XC^t(CC^t)^{-1}\gamma] X(X^tX)^{-1}C^t [C(X^tX)^{-1}C^t] C(X^tX)^{-1}X^t [Y-XC^t(CC^t)^{-1}\gamma]$

Sea $\underline{Y}_* = \underline{Y} - XC^t(CC^t)^{-1}\gamma$

Dado que $\underline{Y} \rightarrow N_n(X\beta, \sigma^2 I)$ se sigue que $\underline{Y}_* \sim N_n(X\beta-XC^t(CC^t)^{-1}\gamma, \sigma^2 I)$

Sea $\underline{\theta} = X\beta - XC^t(CC^t)^{-1}\gamma$ entonces $\underline{Y}_* \sim N_n(\underline{\theta}, \sigma^2 I)$

$SCE_{H_0} = \underline{Y}_*^t X(X^tX)^{-1}C^t [C(X^tX)^{-1}C^t]^{-1}C(X^tX)^{-1}X^t\underline{Y}_*$

Sea $A = X(X^tX)^{-1}C^t [C(X^tX)^{-1}C^t]^{-1}C(X^tX)^{-1}X^t$

Con respecto a SCE_{MC} se tiene que

$SCE_{MC} = Y^t [I - X(X^tX)^{-1}X^t]Y$

además $X\,[I - X(X^tX)^{-1}X^t] = [I - X(X^tX)^{-1}X^t]X = 0$

La SCE_{MC} se puede expresar tambien como

$$SCE_{MC}=(Y-XC^t(CC^t)^{-1}\gamma)\,[I-X(X^tX)^{-1}X^t](Y-XC^t(CC^t)^{-1}\gamma$$
$$=Y_*^t[I - X(X^tX)^{-1}X^t]Y_* \quad \text{con } Y_* \text{ definida anteriormente}$$

Sea $B = [I - X(X^tX)^{-1}X^t]$

Entonces se quiere demostrar que Y_*^t A Y_* (SCE_{HO}) es independiente de Y_*^t B Y_* (SCE_{MC}) donde $Y_* \rightarrow N_n(\theta, \sigma^2 I)$, i.e. $\Sigma = \sigma^2 I$

Se debe demostrar de acuerdo al resultado 4.1 del apendice 1 que $A\Sigma B = 0$

Sustituyendo se tiene que

$$\sigma^2X(X^tX)^{-1}C^t\,[C(X^tX)^{-1}C^t]^{-1}C(X^tX)^{-1}X^t\,[I-X(X^tX)^{-1}X^t] = 0$$

De donde $X^t[I-X(X^tX)^{-1}X^t] = 0$

Por lo tanto SCE_{MC} y SCE_{HO} son independientes.

Distribución de SCE_{MC}

$$SCE_{MC} = (\underline{Y} - X\hat{\beta})^t(\underline{Y} - X\hat{\beta}) = (\underline{Y}-X(X^tX)^{-1}X^t\underline{Y})^t(\underline{Y}- X(X^tX)^{-1}X^t\underline{Y})$$
$$[(I-X(X^tX)^{-1}X^t)\underline{Y}][(I-X(X^tX)^{-1}X^t)\underline{Y}]$$
$$\underline{Y}'\underbrace{[(I-X(X^tX)^{-1}X)(I-X(X^tX)^{-1}X^t)]}_{*}\underline{Y}$$

forma cuadrática con matriz asociada *

Ahora sea $P=X(X^tX)^{-1}X^t$ entonces la forma cuadrática se puede escribir como

$$\underline{Y}^t[(I-P)(I-P)]\underline{Y}^t$$

(I-P) es idempotente ya que

(I-P)(I-P) = (I-P) lo cual es facil de checar sustituyendo P

P.D. que $\dfrac{(n-p)\tilde{\sigma}}{\sigma^2} \quad \chi^2_{(n-p)}$

es decir

$$\frac{Y^t(I-P)Y}{\sigma^2} \qquad \chi^2_{(n-p)} \qquad (I-P) \text{ es idempotente}$$

De acuerdo al resultado 3.5 del apéndice 1, en este caso se tiene

$Y \sim N(X\beta, \sigma^2 I)$ entonces $\theta = X\beta$ y $\Sigma = \sigma^2 I$ en este caso $\frac{Y^t B Y}{2} \sim \chi^2_{(k,\lambda)}$

con $\lambda = \theta^t B \theta/2$ y B de rango k $\leftrightarrow$ $B\Sigma$ es idempotente

Sea $B = (I-P)/\sigma^2$

$B\Sigma = \frac{(I-P)}{\sigma^2} \sigma^2 I = (I-P)$ que es idempotene

Recordar que X de rango p, $(X^t X)_{p \times p}$ de rango p.

$(X^t X)^{-1}_{p \times p}$ de rango p, $X(X^t X)^{-1}_{n \times p}$ de rango p, $X(X^t X)^{-1} X^t$ de rango p.

(I-P) es simetrica e idempotente, es decir es una matriz proyección

entonces Tr(A)=rango(A)

$$\text{rango}(I-P) = Tr(I-P) = Tr(I) - Tr(P)$$
$$= n - Tr(P)$$

pero $Tr(P) = Tr(X(X\ X)\ X) = Tr\ (X\ X(X\ X)\ ) = Tr\ (I\ ) = p$

por lo tanto el rango de (I-P)=n-p

entonces el rango de B es n-p por lo que los grados de libertad de la X son n-p

Por otro lado $\lambda = \theta\ B\ \theta/2$

$\theta = X\beta$, $B = \frac{(I-P)}{\sigma^2}$

$$\lambda = \frac{1}{2}(X\beta)^t \frac{(I-P)}{\sigma^2}(X\beta)$$
$$= \frac{1}{2\sigma^2}(\beta^t X^t)(I-P)(X\beta)$$
$$= \frac{1}{2\sigma^2}[(\beta^t X^t - \beta^t X^t P)(X\beta)$$
$$= \frac{1}{2\sigma^2}[\beta^t X^t - \beta^t \underbrace{X^t X(X^t X)^{-1}}_{I} X^t] X\beta$$
$$= \frac{1}{2\sigma^2}[\beta^t X^t - \beta^t X^t]\ X\ \beta = 0$$ por lo tanto $\lambda = 0$

por lo tanto $\frac{Y^t(I-P)Y}{\sigma^2} \sim \chi^2_{(n-p)}$

Distribución de SCE_{Ho}

Del resultado 3.6.5 del apendice 1, se tiene que

$$SCE_{Ho} = (C\hat{\beta} - \gamma)^{t} \left[\frac{C(X^{t}X)^{-1}C^{t}}{\sigma^{2}} \right]^{-1} (C\hat{\beta} - \gamma) = Q$$

como $(C\hat{\beta} - \gamma) \sim N_{n}(C\hat{\beta} - \gamma)$, $C(X^{t}X)^{-1}C^{t}$ y según el resultado 3.6.5. $Y = (C\hat{\beta} - \gamma)$ basta probar que $B\Sigma$ es idempotente

$$\underbrace{\left[\frac{C (X^{t}X) C^{t}}{\sigma^{2}} \right]^{-1}}_{B} \cdot \underbrace{\sigma^{2} C(X^{t}X)^{-1}C^{t}}_{\Sigma} = I \quad \text{idempotente}$$

entonces

$Q = SCE_{Ho} \sim \chi^{2}_{(r,\lambda)}$ Se tienen r grados de libertad ya que $B = \frac{1}{\sigma^{2}} [C(X^{t}X)^{-1}C^{t}]^{-1}$ es de rango r.

Dado que $SCE_{Ho}/\sigma^{2} \sim \chi^{2}_{(r)}$ bajo Ho.

$SCE_{MC}/\sigma^{2} \sim \chi^{2}_{(n-p)}$

Se tiene que

$$\frac{SCE_{Ho} / r\sigma^{2}}{SCE_{MC} / (n-p)\sigma^{2}} \sim F(r, n-p) \quad \text{bajo Ho}$$

$$\frac{SCE_{Ho} / r}{SCE_{MC} / (n-p)} \sim F(r, n-p) \quad \text{bajo Ho}$$

Bibliografía.

Bernhard,F. Riedwil,H. (1988).*Multivariate Statistics "A practical Approach"*. Chapman & hall. London.

Chaterjee,S. and Price, B.(1977). *Regression Analysis by Example*. Wiley & Sons. New York.

Dongarra,J.J. (1979). *Linpack Users' Guide*. Siam. Philadelphia.

Draper,N,R, and Smith,H. (1981). *Applied Regression Analysis*. Wiley & Sons. New York.

Freund,R.J. and Minton, P.M., (1979). *Regression Methods*. Dekker,Inc.New york.

Hennefeld,J. (1989). *Turbo Pascal con aplicaciones 3.0, 4.0 y 5.0*. Grupo Editorial Iberoamérica. México,D.F.

Koffman,E.B. (1985). *Introducción al lenguaje y resolución de problemas con programación estructurada*. Fondo Educativo Interamericano. México,D.F.

Lang,S. (1976). *Algebra lineal*. Fondo Educativo Interamericano, S.A. México,D.F.

Mendez Ramirez Ignacio. (1981). *Modelos Estadisticos Lineales*. CONACYT. México,D.F.

Pindyck,R.S. and Rubinfeld,D.L. (1981). *Econometric Models and Economic Forecasts*. Mc.Graw. Hill. Tokio.

Tenenbaum,A.M.and Augenstein,M.J. (1981). *Estructura de Datos en Pascal*. Prentice-Hall. México,D.F.

Younger,M.S. (1979). *A handbook for linear Regression*. Duxbury Press. Massachusetts.