UNIVERSIDAD DE PANAMÁ

VICERRECTORÍA DE INVESTIGACIÓN Y POSTGRADO

PROGRAMA CENTROAMERICANO DE MAESTRÍA EN MATEMÁTICA

EL MODELO DE REGRESIÓN LINEAL MULTIVARIADO Y SU APLICACIÓN.

POR:

ALBERTO CASTILLO PORTUGAL

TESIS PRESENTADA COMO UNO DE LOS REQUISITOS PARA OPTAR POR EL TITULO DE MAESTRO EN CIENCIAS CON ESPECIALIZACIÓN EN ESTADÍSTICA MATEMÁTICA.

PANAMÁ, REP. DE PANAMÁ

2002

APROBADO POR:

M. en C. GLADYS E. SEGURA
PRESIDENTE

M. en C. JOSE OCHOA
MIEMBRO

M. en C. ADELA ABAD
MIEMBRO

REPRESENTANTE DE LA VICERRECTORIA
DE INVESTIGACION Y POSTGRADO

FECHA: 28 - II - 2002

DEDICATORIA

Dedico este trabajo de graduación, permanentemente **a Dios Todopoderoso**, como muestra de gratitud por la oportunidad que me ha brindado de superarme en mi vida profesional, a mis padres y esposa por su apoyo y comprensión, a mis hijos **Katherine**, **Jesús** y **Alexis**, como ejemplo de la constancia en el esfuerzo por ser mejor cada día.

AGRADECIMIENTO

Agradezco infinitamente a **Dios**, por darme la oportunidad, fuerza y conocimiento para culminar mis estudios; a mi asesora la **Profesora Gladys Segura** por sus empeño y constancia durante la realización de este trabajo.

INDICE GENERAL

Pág.

INDICE DE CUADROS

Pág.

RESUMÉN

En el presente trabajo definimos el modelo de regresión lineal multivariado, en su forma matricial, demostrándose algunas de las propiedades de los estimadores, tanto en el insesgamiento como en la verosimilitud.

Se plantean las pruebas de hipótesis referente a la matriz de parámetros, para comprobar la correlación existente entre las variables explicativas y las explicadas, con el uso de la distribución **Wishart**, donde se demuestran algunas proposiciones de esta distribución. También consideramos dentro de este modelo la prueba de razón de verosimilitud y el **estadístico de Wilks**, utilizado en la realización de pruebas de hipótesis.

Contemplamos los intervalos de confianza para un valor numérico, con el apoyo de las distribución **Wishart**, $\mathbf{T^2}$ y **F**. Además se consideró el **Coeficiente de Correlación multivariado**, para la base de datos proporcionada por el **M.I.P.P.E.** (Ministerio de Planificación y Política Económico), sobre los niveles de vida en Panamá, en la que se determinó la ecuación de regresión multivariada, se realizan pruebas de hipótesis para verificar las influencia de las variables explicativas.

SUMMARY

Present work defined the model of multivaried lineal regression, in their matricial form, demonstrating some of the properties of the estimators, so much in the unbiasing like in the verisimilitude.

We are expounded the hypothesis taste with respect to the womb of parameters, in order to check the existent correlation between the explanatory variables and explained variables, with the use of the **Wishart Distribution**, where some propositions of these distribution are demonstrated. We also considered within this model taste reason of verisimilitude and the **Statistic of Wilk's**, utilized in the realization of you taste hypothesis.

We contemplated the intervals of trust for a numeric courage, with the support of the **Wishart Distribution**, $\mathbf{T^2}$ and **F**. were Also considered the multivaried correlation coefficient, for the base of data proportioned by the **M.I.P.P.E.** (Ministry of Planning and Economical Politics), on the levels of life in Panama, they in the one which was determined the equation of multivaried regression, are carried out taste hypothesis in order to verify the influence of the explanatory variables.

INTRODUCCIÓN

Es normal encontrar en la realización de investigaciones muchas variables explicativas y explicadas, que influyen en cierto grado dentro de un fenómeno dado, por tal razón es de gran utilidad trabajar con todas ellas.

Una forma de estudiar estas variables en conjunto, es a través del Modelo de Regresión Lineal Multivariado y de la correlación de las mismas con las que se puede hacer las interrelaciones, estimando la matriz de los parámetros de todos las variables explicativas y explicadas, además ver las interrelaciones de dos a dos .

La regresión fue utilizada por primera vez en el año 1880 por el científico inglés **Sir Francis Galton**, dedicado a investigaciones genéticas, quien trataba de establecer las características trasmitidas de padres a hijos a través de sus estaturas.

El modelo de regresión nos permite eliminar aquellas variables que producen poco o ningún efecto en la regresión, esto por medio de las pruebas de hipótesis o de la correlación. Realizar estimaciones es esencial en el modelo de regresión, donde se ha de comprobar la significancia de la ecuación por medio del coeficiente de determinación.

Los intervalos de confianza también son considerados en la regresión como un elemento que ayuda a fortalecer los resultados con el coeficiente de determinación como una forma de evaluar la ecuación de regresión por medio de la proximidad del ajuste a los valores observados.

Este trabajo contempla tres capítulos con los siguientes contenidos:

En el primer capítulo se considera la ecuación de regresión, matrices del modelo, algunas propiedades del modelo y de la matriz P, además de los estimadores de las matrices B y Σ.

El segundo capítulo hace referencia a las pruebas de hipótesis según las matrices, la distribución Wishart centrada con sus propiedades, se definen algunos estadísticos, además de los intervalos de confianza, la correlación múltiple y el coeficiente de determinación, con lo que termina la parte teórica.

El capítulo tercero considera el análisis e interpretación de los resultados, en el que se determina la ecuación de regresión, se hacen estimaciones, además se realizan pruebas de hipótesis y se calcula el coeficiente de correlación sobre una base de datos, referente a una encuesta de niveles de vida en Panamá, en el año 1997.

CAPÍTULO I

EL MODELO DE REGRESIÓN LINEAL MULTIVARIADO

La Ecuación de Regresión.

Consideramos el modelo definido por $Y = XB + E$, donde las matrices Y, X, B y E son tales que $Y_{(nxp)}$ es una matriz observada, de p variables respuestas en cada uno de los n vectores de variables aleatorios; $X_{(nxq)}$ una matriz de valores fijos, q variables independientes observadas en cada uno de los n vectores filas; donde estos vectores filas son mutuamente independientes, cada una con matriz de media cero (0) y matriz de varianza covarianza común Σ; $B_{(qxp)}$ es una matriz de parámetros desconocidos, afectados por las filas de X y $E_{(nxp)}$ es una matriz aleatoria de valores desconocidos (matriz de error).

La ecuación lineal $Y = XB + E$ (1) es llamada Modelo de Regresión Lineal Multivariado. En el caso de que X sea una matriz aleatoria, entonces la distribución de E se asume que no está relacionada con X.

En particular las variables $X_1, X_2, \ldots, X_q$ predicen cada una de las Y's.

Las columnas de la matriz Y representan variables dependientes que están explicadas en términos de las variables independientes o explicativas dadas por las columnas de X.

Matrices del Modelo.

Así, el modelo lineal Y = XB + E en su forma matricial es:

$$\begin{bmatrix} y_{11} & y_{12} \cdots & y_{1p} \\ y_{21} & y_{22} \cdots & y_{2p} \\ \cdot & \cdot \cdots & \cdot \\ \cdot & \cdot \cdots & \cdot \\ \cdot & \cdot \cdots & \cdot \\ y_{n1} & y_{n2} \cdots & y_{np} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \cdots & x_{1q} \\ 1 & x_{21} & x_{22} \cdots & x_{2q} \\ \cdot & \cdot & \cdot \cdots & \cdot \\ \cdot & \cdot & \cdot \cdots & \cdot \\ \cdot & \cdot & \cdot \cdots & \cdot \\ 1 & x_{n1} & x_{n2} \cdots & x_{nq} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ \beta_{11} & \beta_{12} \cdots & \beta_{1p} \\ \cdot & \cdot \cdots & \cdot \\ \cdot & \cdot \cdots & \cdot \\ \cdot & \cdot \cdots & \cdot \\ \beta_{q1} & \beta_{n2} \cdot & \beta_{qp} \end{bmatrix} + \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1p} \\ e_{21} & e_{22} & \cdots & e_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ e_{n1} & e_{n2} & \cdots & e_{np} \end{bmatrix}$$

$$nxp = nx(q+1) * [(q+1)xp] + nxp$$

Aquí, cada columna de la matriz X representa un vector de variables unitarias; cada uno de los n vectores filas de Y contiene los valores de las p variables medidas en un sujeto.

Cada una de las columnas de Y consiste de las observaciones en cada una de las p variables que corresponden a un vector unitario $Y_{(i)}$. Para cada columna de Y tenemos una columna de parámetros de β's. Las columnas de β's forman una matriz a la que llamamos B.

El modelo también puede ser expresado en términos de los vectores columnas, en el caso de que los vectores de la matriz respuesta estén dados en columna; para la i-ésima respuesta, el modelo se puede escribir como $Y_{(i)}=XB_{(i)} + \varepsilon_{(i)}$; $1 \leq i \leq p$.

donde $Y_{(i)}$, $B_{(i)}$ y $\varepsilon_{(i)}$ son vectores columnas.

En este caso el modelo así definido recibe el nombre de modelo de Regresión Lineal Múltiple.

Por otro lado los n vectores de orden (px1) de la matriz E, están distribuidos normalmente con matriz de media (0) y matriz de covarianza Σ.

Definición I.1

Diremos que la matriz de error $E_{(nxp)} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ en donde para cada $1 \leq i \leq n$ el ε_i corresponde a un vector fila de orden 1xp que representa el i-ésimo vector aleatorio de error, tiene distribución normal con matriz de media (0) y matriz de varianza covarianza Σ. Además $E \sim N_p(0, I \otimes \Sigma)$, donde I es la matriz identidad de orden (nxn) y $I \otimes \Sigma$ denota el producto de Kronecker de la matriz I y la matriz Σ.

Aquí las filas de E son normalmente independientes.

Definición I.2

Diremos que el producto de Kronecker definido por $I \otimes \Sigma$ representa el producto de los n vectores columnas de una matriz que son mutuamente independientes, cada uno con matriz de varianza covarianza Σ y el producto de Kronecker dado por $\Sigma \otimes I$, se refiere a la matriz de covarianza del vector X^v de orden $[(nq)x\ 1]$ obtenido por la colocación de los vectores uno sobre el otro.

Proposición I.1

En el modelo Y se tiene que

$Tr[(Y-XB)\Sigma^{-1}(Y-XB)']$ es igual a $Tr[\Sigma^{-1}(Y-XB)'(Y-XB)]$.

Demostración:

Sea Y la variable aleatoria del modelo dada por la ecuación (1) y $f(Y)$ su función de densidad, consideremos a $Y = [Y_1, Y_2, \ldots, Y_n]'$, donde las filas son mutuamente independientes.

Por otro lado como E tiene una distribución normal con matriz de media (0) y matriz de varianza covarianza $I \otimes \Sigma$, entonces para cada $1 \leq i \leq n$, los n vectores filas tienen una distribución igualmente normal con media $B'X_i$ y matriz de varianza covarianza Σ, por lo tanto como $Y_i = B'X_i + \varepsilon_i$ es una combinación lineal de ε_i que también tiene una distribución normal con media (0) y matriz de varianza covarianza Σ, esto es $Y_i \sim Np\ (B'X_i, \Sigma)$.

Sabemos además que $Y - XB = [Y_1 - B'X_1,\ Y_2 - B'X_2, \dots Y_n - B'X_n]'$.

Realizando el producto $(Y - XB)'(Y - XB)$ tenemos que

$$(Y - XB)'(Y - XB) = \sum_{i=1}^{n} (Yi - B'Xi)(Y - B'Xi)'$$

por otro lado tenemos que $\sum_{i=1}^{n} (Y_i - B'Xi)'\sum^{-1} (Yi - B'Xi)$

Luego:

$$Tr\left[\sum_{i=1}^{n}(Y_i - B'X_i)'\sum^{-1}(Y_i - B'X_i)\right]^{(1)}$$

$$= \sum_{i=1}^{n} Tr\left[(Y_i - B'X_i)'\sum^{-1} (Y_i - B'X_i)\right]$$

$$= \sum_{i=1}^{n} Tr\left[\sum^{-1} (Yi - B'Xi)(Yi - B'Xi)'\right]$$

por propiedad de traza

$$= Tr\left[\sum^{-1} (Y - XB)'(Y - XB)\right]$$

por lo tanto la $Tr[(Y - XB)\sum^{-1}(Y - XB)'] = Tr[\sum^{-1}(Y - XB)'(Y - XB)]$

---

(1) Richard y Wichern (1982).

La función de verosimilitud del Modelo.

Definición I.3

La función de verosimilitud de Y, está dada por

$$f(B,\Sigma) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(Y_i - B'X_i)'\Sigma^{-1}(Y_i - B'X_i)}$$

donde Y_i es un vector fila con media $B'X_i$ y matriz de varianza covarianza Σ, $1 \leq i \leq n$.

En el modelo lineal, tenemos que la función de densidad de Y es $f(Y) = \frac{1}{(2\pi)^{np/2}} \frac{1}{|\Sigma|^{n/2}} e^{-\frac{1}{2}(Y-XB)'\Sigma^{-1}(Y-XB)}$ y considerando las suposiciones de que la matriz X tiene rango q y que la covarianza de $(X'X)$ existe.

Definición I.4

La función log de verosimilitud para Y en términos de los parámetros B y Σ, está dado por

$$L\,f(B,\Sigma) = -\frac{1}{2} n \log|2\pi\Sigma| - \frac{1}{2} Tr(Y - XB)\Sigma^{-1}(Y - XB)'$$

donde XB es la media de Y.

Propiedades del Modelo.

Definición I.5

En el modelo de regresión lineal Y, se cumplen las propiedades

a. $E(E)=0$

b. $E(Y_i)=B'X_i \qquad 1 \le i \le n.$

c. $COV\,(Y_i)=\Sigma \qquad \forall\ i=1,2,\ldots,n$

donde Y_i es la i-ésima fila de Y

d. $COV(Y_i,Y_j)=0 \qquad \forall i \neq j$

siendo Y_i,Y_j filas de la matriz Y.

Propiedades de la Matriz P.

Proposición 1.2

La matriz $P_{(nxn)}$ definida por $P=[I-X\,(X'X)^{-1}X']$ es simétrica e idemponente y de rango (n-q), donde las matrices I y $X(X'X)^{-1}X'$ son de orden (nxn).

Demostración:

Sea P una matriz de orden (nxn) tal que

$P=[I-X\,(X'X)^{-1}X']$, consideramos primero si P es simétrica, esto es $P=P'$

Si $P = [I - X(X'X)^{-1}X']$, luego su traspuesta P' es:

$$P' = [I - X(X'X)^{-1}X']'$$

$$P = I - [X(X'X)^{-1}X']$$

por definición de traspuesta

$$P = [I - X(X'X)^{-1}X']$$

dado que $(X'X)^{-1}$ es una matriz cuadrada de orden (q x q), que es simétrica.

Luego como $P = P'$, entonces la matriz P es simétrica.

Veamos ahora la idempotencia de P, esto es $PP' = P$. Así este producto es

$[I-X(X'X)^{-1}X'][I - X(X'X)^{-1}X']'$, puesto que $PP' = P$ entonces

$$[I - X(X'X)^{-1}X'][I - X(X'X)^{-1}X']$$

realizando los productos

$$I - IX(X'X)^{-1}X' - X(X'X)^{-1}X'I + X(X'X)^{-1}(X'X)(X'X)^{-1}X'$$

$$I - X(X'X)^{-1}X' - X(X'X)^{-1}X' + XI(X'X)^{-1}X'$$

$$P = [I - X(X'X)^{-1}X']$$

por lo tanto P es una matriz simétrica y luego P es idempotente.

Calculemos ahora el rango de P

Ran $[I - X(X'X)^{-1}X']$= $Tr[I - X(X'X)^{-1}X']$ puesto que P es idempotente y como I es de orden (nxn) y $X(X'X)^{-1}X'$ es de orden (nxn) de rango q.

Se tiene que $Tr[I - X(X'X)^{-1}X']$ = $Tr(I) - Tr[X(X'X)^{-1}X']$ por propiedad de traza.

Luego $Tr(I) = n$ y $Tr[X(X'X)^{-1}X'] = q$

Por lo tanto la $Tr[I - X(X'X)^{-1}X'] = (n-q)$.

Así, P[I - X(X'X)-1X'] es una matriz simétrica indempotente y de rango (n-q).

Estimadores de Máxima Verosimilitud de B y Σ.

Proposición 1.3

En el modelo multivariado de rango completo q donde $E \sim N_P(O, I \otimes \Sigma)$ y $Y \sim N_P(XB; I \otimes \Sigma)$, los estimadores de máxima verosimilitud de B y Σ son $\hat{B} = (X'X)^{-1}X'Y$ y $\hat{\Sigma} = n^{-1}\ Y'PY$, o bien $\hat{\Sigma} = n^{-1}\ \hat{E}'\hat{E}$ donde $P = [I - X(X'X)^{-1}X']$.

Demostración:

Si en el modelo de rango completo, $E \sim N_P(0; I \otimes \Sigma)$ y de acuerdo al modelo de regresión $Y = [Y_1, Y_2, \ldots, Y_n]'$, donde las filas de Y son independientes con distribución $Y_i \sim N_P(B'X_i, \Sigma)$, consideremos ahora

$$(Y-XB) = [Y_1 - B'X_1, Y_2 - B'X_2, \ldots, Yn - B'Xn]'$$

así, $$(Y-XB)'(Y-XB) = \sum_{i=1}^{n} (Yi - B'X_i)\ (Yi - B'X_i)'$$

y también se tiene que

$$\sum_{i=1}^{n} (Yi - B'Xi)' \Sigma^{-1} (yi - B'Xi) = \sum_{i=1}^{n} Tr\left[\Sigma^{-1}(Yi - B'Xi)'(Yi - B'Xi)\right]$$

$$= \sum_{i=1}^{n} Tr\left[\Sigma^{-1}(Yi - B'Xi)(Yi - B'Xi)'\right]$$

$$= Tr\left[\Sigma^{-1}(Y - XB)'(Y - XB)\right]$$

Por **Proposición I.1**, obtenemos que

Sí $\hat{E} = Y - X\hat{B}$, entonces $\hat{E} + X\hat{B} = Y$

$$\hat{E} + \left(X\hat{B} - XB\right) = Y - XB$$

$$\hat{E} + X\left(\hat{B} - B\right) = Y - XB$$

Si la función de verosimilitud de Y es

$$f(B,\Sigma) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}Tr(Yi-B'Xi)'\Sigma^{-1}(Yi-B'Xi)}$$

$$= \frac{1}{(2\Pi)^{np/2}} \frac{1}{|\Sigma|^{n/2}} e^{-\frac{1}{2}Tr(Yi-B'Xi)'\Sigma^{-1}(Yi-B'Xi)}$$

desarrollando el producto (Y-XB)′ (Y-XB)

$$= \left[\hat{E} + X(\hat{B} - B)\right]' \left[\hat{E} + X(\hat{B} - B)\right]$$

$$= \left[\hat{E} + (\hat{B} - B)' X'\right] \left[\hat{E} + X(\hat{B} - B)\right]$$

$$= \left[\hat{E}'\hat{E} + \hat{E}'X(\hat{B} - B) + (\hat{B} - B)'X'\hat{E} + (\hat{B} - B)'X'X(\hat{B} - B)\right]$$

$$= \hat{E}'\hat{E} + (\hat{B} - B)'X'X(\hat{B} - B) + \hat{E}X'(\hat{B} - B) + (\hat{B} - B)'X'\hat{E}$$

Consideremos $\hat{E}'X(\hat{B} - B)$ sí $\hat{E} = PY$ que es igual a $\hat{E}' = Y'P$

puesto que P es simétrica entonces $\hat{E}'X(\hat{B} - B) = Y'PX(\hat{B} - B)$, pero PX es la matriz cero (0). Por lo tanto $\hat{E}'X(\hat{B} - B) = 0$

De igual forma $(\hat{B}-B)'X'\hat{E}=0$

entonces $f(B,\Sigma)=\dfrac{1}{(2\pi)^{np/2}}\dfrac{1}{|\Sigma|^{n/2}}e^{-\frac{1}{2}\mathrm{Tr}\left[\Sigma^{-1}(\hat{E}'\hat{E}+(\hat{B}-B)'X'X(\hat{B}-B))\right]}$

a través de la función log., obtenemos que

$$Lf(B,\Sigma)=-\frac{1}{2}np\log 2\pi-\frac{1}{2}n\log|\Sigma|-\frac{1}{2}\mathrm{Tr}\left[\Sigma^{-1}(\hat{E}'E+(\hat{B}-B)'X'X(\hat{B}-B)\right]$$

esta función alcanza su valor máximo cuando $B=\hat{B}$,

luego $Lf(\hat{B},\Sigma)=-\frac{1}{2}np\log 2\pi-\frac{1}{2}n\log|\Sigma|-\frac{1}{2}\mathrm{Tr}\Sigma^{-1}\hat{E}'\hat{E}$

donde $\hat{E}'\hat{E}=n\hat{\Sigma}$, entonces $\hat{\Sigma}=n^{-1}\hat{E}'\hat{E}$

$$Lf(\hat{B},\Sigma)=-\frac{1}{2}np\log 2\pi-\frac{1}{2}n\left(\log|\Sigma|+\mathrm{Tr}\,\Sigma^{-1}\hat{\Sigma}\right) \quad (2) \quad y$$

la expresión (2) alcanza su valor máximo cuando $\Sigma=\hat{\Sigma}$

$$\text{Así, } Lf(\hat{B},\hat{\Sigma})=-\frac{1}{2}n\log\left|2\pi\hat{\Sigma}\right|-\frac{1}{2}n\mathrm{Tr}\,\hat{\Sigma}^{-1}\hat{\Sigma}$$

$$=-\frac{1}{2}n\log\left|2\pi\hat{\Sigma}\right|-\frac{1}{2}n\mathrm{Tr}(I)$$

$$Lf(\hat{B},\hat{\Sigma})=-\frac{1}{2}n\log\left|2\pi\hat{\Sigma}\right|-\frac{1}{2}np$$

Tendremos que el valor máximo de la función de verosimilitud, se obtiene cuando $\Sigma = \hat{\Sigma}$ por lo tanto $\hat{B}$, $\hat{\Sigma}$ son los estimadores de máxima verosimilitud de B y Σ.

Como E tiene una distribución normal, entonces Y = XB + E; como combinación lineal de E, también tiene distribución normal, con matriz de media XB y matriz de covarianza, $I \otimes \Sigma$, esto es $Y \sim Np(XB, I \otimes \Sigma)$.

Propiedades de los Estimadores de B y Σ.

Proposición I.4

Para el modelo Y = XB + E con una distribución normal multivariada, donde la matriz de error E se distribuye $Np \sim (O, I \otimes \Sigma)$, se cumple que:

a) $\hat{B}$ es un estimador insesgado de B

b) $\hat{\Sigma}$ no es un estimador insesgado de Σ

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

d) $\hat{B}$ y $\hat{E}$ son matrices con distribuciones normales y multivariado

e) La matriz $\hat{B}$ es estadísticamente independiente $\hat{E}$ y también de $\hat{\Sigma}$.

Demostración:

Dado que $\hat{B} = (X'X)^{-1}X'Y$ y si $Y = XB + E$ el modelo multivariado, al reemplazar Y en $\hat{B}$ tenemos que:

$$\hat{B} = (X'X)^{-1}X'(XB+E)$$

$$\hat{B} = (X'X)^{-1}X'XB + (X'X)^{-1}X'E$$

$$\hat{B} = B + (X'X)^{-1}X'E$$

El valor esperado de ambas expresiones

$$E(\hat{B}) = E[B + (X'X)^{-1}X'E]$$

$$= E(B) + E(X'X)^{-1}X'E)$$
$$= B + (X'X)^{-1}X'E(E)$$

$$= B + O \text{ puesto que } E(E) = 0$$

$$= B$$

Tenemos que $\hat{B}$ es un estimador insesgado de B y por la **Proposición I.3.** $\hat{B}$ es un estimador de máxima verosimilitud de B, con lo que se demuestra (a). El estimador de máxima verosimilitud de Σ es, $\hat{\Sigma}$ el cual puede ser expresado como $n^{-1}E'PE$ o bien $n^{-1}\hat{E}'\hat{E}$ donde E es una matriz de datos distribuidos $N_p \sim (O,\ I \otimes \Sigma)$.

Para $\hat{\Sigma}$ así definido, resulta no ser un estimador insesgado de Σ; para que este ocurra se debe hacer un arreglo sobre $\hat{\Sigma}$ dado por $\dfrac{\hat{\Sigma}}{n-(q+1)}$ el cual es el estimador insesgado de $\hat{\Sigma}$, esto es $E\left[\dfrac{\hat{\Sigma}}{n-(q+1)}\right]=\Sigma$.

Con lo que se demuestra que $\hat{\Sigma}$ no es un estimador insesgado de Σ, demostrándose la parte (b).

En el siguiente caso:

Si $\hat{E}=Y-X\hat{B}$, consideramos $E(\hat{E})$ entonces

$$
\begin{aligned}
E(\hat{E}) &= E(Y-X\hat{B}) \\
&= E(Y)-E(X\hat{B}) \\
&= XB-XB \\
&= 0
\end{aligned}
$$

por lo tanto $E(\hat{E})=0$, en consecuencia $E(\hat{\varepsilon}_i)=0$, para cada $\hat{\varepsilon}_i$, $i\leq n$.

En la parte (d), mostramos que $\hat{B}$ y $\hat{E}$ sonnormales multivariados.

Si tenemos que $\hat{E}=Y-X\hat{B}=PY$, además $PY=PE$ y $PE=[I-X(X'X)^{-1}X']E$.

En consecuencia ambos estimadores $\hat{B}$ y $\hat{E}$ son funciones lineales de E. Luego tenemos que E tiene una distribución normal y como $\hat{B}$ y $\hat{E}$ son funciones lineales de E, entonces $\hat{B}$ y $\hat{E}$ tienen distribuciones normales multivariadas.

Para la parte (c), tenemos que

$\hat{B}$ es estadísticamente independiente de $\hat{E}$ y por lo tanto de $\hat{\Sigma}$.

Por la parte (e) obtenemos que $\hat{E} = PY$ y

$\hat{B} = B + (X'X)^{-1}X'E$.

Si X es otra matriz tal que tiene una distribución de $N_q(\mu,\Sigma)$ y si $Y = AXB$ y $Z = CXD$, entonces los elementos de Y son independientes de los elementos de Z si y solamente si $AC' = 0$.

Así, $\hat{B} = (X'X)^{-1}X'YI$, $\hat{E} = PYI$, además Y es una matriz de datos, donde $Y \sim N_p(XB, I \otimes \Sigma)$, luego $(X'X)^{-1}X' = A$, y $P = C$, donde $P = P' = C'$.

Entonces
$$\begin{aligned} AC' &= (X'X)^{-1}X'[I - X(X'X)^{-1}X']' \\ &= (X'X)^{-1}X'I - (X'X)^{-1}X'(X'X)(X'X)^{-1}X' \\ &= (X'X)^{-1}X' - (X'X)^{-1}X' \\ &= 0 \end{aligned}$$
así, $\hat{B}$ y $\hat{E}$ son estadísticamente independientes.

Consideremos ahora $\hat{B}$ y $\hat{\Sigma}$, dado que $\hat{B} = (X'X)^{-1}X'YI$ y $\hat{\Sigma} = n^{-1}Y'PY$.

También $\hat{\Sigma} = n^{-1}(XB + E)'PYI$

Ahora multiplicando AC' [2]

tenemos que $AC' = (X'X)^{-1}X'n^{-1}P(XB + E)$

$$= n^{-1}(X'X)^{-1}X'P(XB + E)$$

$$= n^{-1}(X'X)^{-1}X'[I - X(X'X)^{-1}](XB + E)$$

$$= n^{-1}[(X'X)^{-1}I - (X'X)^{-1}X'X(X'X)^{-1}X'](XB + E)$$

$$= n^{-1}0(XB + E)$$

$$= 0$$

En consecuencia, $\hat{B}$ es estadísticamente independiente de $\hat{\Sigma}$ lo que demuestra la parte (d).

Proposición I.5:

En la matriz de parámetros estimados de $\hat{B}$, si $\hat{\beta}_{(i)}$ es un vector columna, entonces el valor esperado de $\hat{\beta}_{(i)}$ es $\beta_{(i)}$ para cada $i \leq p$.

Demostración:

Sea $\hat{B}$ el estimador insesgado de B, si, $\hat{\beta}_{(i)} = (X'X)^{-1}X'Y_{(i)}$ donde $Y_{(i)} = X\beta_{(i)} + \varepsilon_{(i)}$, por cada $i \leq p$.

Luego $\hat{\beta}_{(i)} = (X'X)^{-1}X'\left[X\beta_{(i)} + \varepsilon_{(i)}\right]$

---

[2] Mardia et al (1979).

$$= (X'X)^{-1}(X'X)\beta_{(i)} + (X'X)^{-1}X'\varepsilon_{(i)}$$

$$= \beta_{(i)} + (X'X)^{-1}X'\varepsilon_{(i)}$$

Así $E(\hat{\beta}_{(i)}) = E[\beta_{(i)} + (X'X)^{-1}X'\varepsilon_{(i)}]$

$$= E[\beta_{(i)}] + E[(X'X)^{-1}X'\varepsilon_{(i)}]$$

$$= \beta_{(i)} + (X'X)^{-1}X'E(\varepsilon_{(i)})$$

$$= \beta_{(i)} + (X'X)^{-1}X'(0)$$

$$= \beta_{(i)}$$

Por lo tanto $E(\hat{\beta}_{(i)}) = \beta_{(i)}$ para cada $i \leq p$ demostrando que el vector columna estimado $\hat{\beta}_{(i)}$ es un estimador insesgado del vector columna $\beta_{(i)}$.

Proposición I.6:

En el Modelo de Regresión Lineal Multivariado de rango completo, la covarianza entre dos vectores columnas de $\hat{B}$ es cov $(\hat{\beta}_{(i)}, \hat{\beta}_{(k)}) = \sigma_{ik}G$, para $\forall$ i = k, donde $G = (X'X)^{-1}$; $i, k = 1, 2, 3, \ldots, n$ y cov $(\hat{\beta}_{(i)}, \hat{\beta}_{(k)}) = 0$ para todo $i \neq k$.

Demostración:

Sean $\hat{\beta}_{(i)}$ y $\hat{\beta}_{(k)}$ dos vectores columnas de $\hat{B}$, tales que $\hat{\beta}_{(i)} = (X'X)^{-1}X'Y_{(i)}$, donde $E(\hat{\beta}_{(i)}) = \beta_{(i)}$, para cada $i \leq p$

Si $Y_{(i)} = X\beta_{(i)} + \varepsilon_{(i)}$ entonces

$$\hat{\beta}_{(i)} = (X'X)^{-1}X'[X\beta_{(i)} + \varepsilon_{(i)}]$$

$$= \beta_{(i)} + (X'X)^{-1}X'\varepsilon_{(i)}$$

$$\hat{\beta}_{(i)} - \beta_{(i)} = (X'X)^{-1}X'\varepsilon_{(i)},$$

Sea $A = (X'X)^{-1}X'$ una matriz (qxn) así: $\hat{\beta}_{(i)} - \beta_{(i)} = A\varepsilon_{(i)}$.

Consideremos la $\operatorname{cov}(\hat{\beta}_{(i)}, \hat{\beta}_{(k)}) = E[(\hat{\beta}_{(i)} - \beta_{(i)})(\hat{\beta}_{(k)} - \beta_{(k)})']$ por definición de covarianza

$$E[(A\varepsilon_{(i)})(A\varepsilon_{(k)})'] = E[(A\varepsilon_{(i)})(\varepsilon'_{(k)}A')]$$

$$= AE[(\varepsilon_{(i)}\varepsilon'_{(k)})]A'$$

$$= \delta_{ik}\sigma_{ik}IAA'$$

$$= \delta_{ik}\sigma_{ik}(X'X)^{-1}$$

entonces, para cada $1 \le i \le p, \quad 1 \le k \le p$.

$$\operatorname{Cov}(\hat{\beta}_{(i)}, \hat{\beta}_{(k)}) = \begin{cases} \sigma_{ik}(X'X)^{-1}, & \text{si} \quad i = k \\ 0 & , \text{si} \quad i \ne k \end{cases}$$

Donde $\delta_{ik} = \begin{cases} 1 & \text{si} \quad i = k \\ 0 & \text{si} \quad i \ne k \end{cases}$

Es conocido como el delta de Kronecker, por lo tanto la covarianza de $\hat{\beta}_{(i)}$ y $\hat{\beta}_{(k)}$ es:

$$\text{cov}(\hat{\beta}_{(i)},\ \hat{\beta}_{(k)}) = \sigma_{ik}(X'X)^{-1}, \qquad \text{si} \quad i = k, 1 \leq i \leq p, \quad 1 \leq k \leq p \text{ y}$$

$$\text{cov}(\hat{\beta}_{(i)},\ \hat{\beta}_{(k)}) = 0, \qquad \text{si} \quad i \neq k$$

no correlacionado. Para $1 \leq i \leq p$, $1 \leq k \leq p$.

Proposición I.7:

La covarianza entre dos filas $\hat{\beta}_i, \hat{\beta}_k$ de la matriz B está dada por $g_{ik}\Sigma$, donde $g_{ik} = (X'X)^{-1}$ para $i \leq p$, $k \leq p$.

Demostración:

Sean $\hat{\beta}_i, \hat{\beta}_k$ dos filas cualesquiera de la matriz de parámetros estimados $\hat{B}$, sea además $\hat{\beta}_i = (X'X)^{-1} X'Y_i$, donde $Y_i = X\beta_i + \varepsilon_i$

Así, $\hat{\beta}_i = (X'X)^{-1} X'(X\beta_i + \varepsilon_i)$

$= \beta_i + (X'X)^{-1} X'\varepsilon_i$

entonces $\quad \hat{\beta}_i - \beta_i = (X'X)^{-1} X'\varepsilon_i$,

sea $A = (X'X)^{-1}X'$ una matriz de orden (qxn) y $(X'X)^{-1}$ es simétrica de orden (qxq),

luego $\hat{\beta}_i - \beta_i = A\varepsilon_i$

por otro lado $E(\hat{\beta}_i) = E[\beta_i + (X'X)^{-1}X'\varepsilon_i\ ]$

$$= \beta_i + (X'X)^{-1}X'E(\varepsilon_i)$$

$$= \beta_i$$

Consideremos ahora la $\operatorname{cov}\left(\hat{\beta}_i, \hat{\beta}_k\right)$

$$= E\left[(\hat{\beta}_i - \beta_i)(\hat{\beta}_k - \beta_k)'\right]$$

$$= E\left[\left(A\varepsilon_i\right)\left(A\varepsilon_k\right)'\right]$$

$$= E\left[\left(A\varepsilon_i\right)\left(\varepsilon'_k A'\right)\right]$$

$$= AE\left(\varepsilon_i \varepsilon'_k\right)A'$$

$$= A\Sigma A'$$

$= AA'\Sigma,$ pero $AA' = (X'X)^{-1}X'[X(X'X)^{-1}]$

$$= (X'X)^{-1}(X'X)(X'X)^{-1}$$

$AA' = (X'X)^{-1}$ que es una matriz simétrica

por lo que $AA'\Sigma = g_{ik}\Sigma,$ donde $g_{ik} = (X'X)^{-1}_{ik}$.

En consecuencia, $\operatorname{cov}(\hat{\beta}_i, \hat{\beta}_k) = g_{ik}\Sigma$

$$= (X'X)^{-1}_{ik}\Sigma \qquad 1 \le i \le q, \quad 1 \le k \le q.$$

Proposición I.8.

La covarianza entre un elemento de $\hat{\beta}_{(i)}$ y otro del $\hat{\beta}_{(k)}$, esto es **cov**$\left(\hat{\beta}_{ij},\hat{\beta}_{kl}\right)$ es $\sigma_{jl}g_{ik}$, donde g_{ik} es un elemento de la matriz $G=(X'X)^{-1}_{ik}$.

Demostración:

Si $\hat{\beta}_{ij}$, $\hat{\beta}_{kl}$ son dos elementos de B, si además $\hat{\beta}_{i}$, es insesgado, esto es $E(\hat{\beta}_{ij})=\beta_{ij}$, tendremos entonces que su covarianza es $\mathrm{Cov}(\hat{\beta}_{ij},\hat{\beta}_{kl})=E\left[(\hat{\beta}_{ij}-\beta_{ij})(\hat{\beta}_{kl}-\beta_{kl})\right]$ puesto que $\hat{B}-B=AE$, donde $A=(X'X^{-1})X$ luego $\hat{\beta}_{ij}-\beta_{ij}=a'_{i}\,\varepsilon_{(j)}$.

Por lo tanto la $\mathrm{cov}(\hat{\beta}_{ij},\hat{\beta}_{kl})=E\,[\{a'_{i}\,\varepsilon_{(j)}\}\,\{a'_{k}\,\varepsilon_{(l)}\}]'$

$$=E\,[\{a'_{i}\,i\varepsilon_{(j)}\}\,\{\varepsilon'_{(l)}\,a_{k}\}]$$

$$=a_{i}'E\,[\,\varepsilon_{(j)}\varepsilon'_{(l)}\,]\,a_{k}$$

$$=a_{i}'\sigma_{j}\varepsilon_{i}Ia_{k}$$

$$=a_{i}'\sigma_{il}Ia_{k}$$

$$=\sigma_{jl}I(AA')_{ik}\quad \text{donde}\quad AA'=(X'X)^{-1}$$

Si además g_{ik} es el i-ésimo elemento correspondiente a la k-ésima columna de $G = (XX)^{-1}$; por lo que $g_{ik} = (X'X)^{-1}_{ik}$, esto demuestra que

$$cov\ (\hat{\beta}_{ij}, \hat{\beta}_{kl}) = \sigma_{jl} g_{ik} \qquad j \leq q, \quad l \leq p$$

Distribución de $\hat{B}$.

Proposición I.9:

En el modelo Y, si $E \sim Np\,(0, I \otimes \Sigma)$ y además $\hat{Y} = X\hat{B}$, se tiene entonces que $\hat{B} \sim Np\,[B, (X'X)^{-1} \otimes \Sigma\,]$.

Demostración:

Sea Y el modelo, con la matriz de error $E \sim Np\,(0, I \otimes \Sigma)$, dado que $\hat{Y} = X\hat{B}$ entonces por la **Proposición I.5**; en su parte (a) tenemos que $\hat{B}$ es un estimador insesgado de B, esto es $E(\hat{B}) = B$.

Por otro lado veamos la $cov(\hat{B})$, por definición la $cov(\hat{B}) = E\left[(\hat{B} - B)(\hat{B} - B)'\right]$.

Si $\hat{B} = (X'X)^{-1}X'$ y entonces en términos de B y E está dada por $\hat{B} = B + (X'X)^{-1}X'E$.

Luego $\hat{B} - B = (X'X)^{-1}X'E$

Por lo tanto,

$$E\left[\left(\hat{B} - B\right)\left(\hat{B} - B\right)'\right]$$

$$= E\left\{\left[(X'X)^{-1}X'E\right]\left[(X'X)^{-1}X'E\right]'\right\}$$

$$= E\left\{\left[(X'X)^{-1}X'E\right]\left[E'X(X'X)^{-1}\right]\right\}$$

$$= (X'X)^{-1}X'E(EE')X(X'X)^{-1}$$

$$= (X'X)^{-1}X'(I \otimes \Sigma)X(X'X)^{-1}$$

$$= (X'X)^{-1}(X'X)(X'X)^{-1}(I \otimes \Sigma)$$

$$= (X'X)^{-1}(I \otimes \Sigma),$$

$$= (X'X)^{-1} \otimes \Sigma$$

Dado que $(X'X)^{-1}I = (X'X)^{-1}$, por propiedad de matriz identidad.

Por lo tanto la covarianza de $\hat{B}$ es $(X'X)^{-1} \otimes \Sigma$.

Como E se distribuye normalmente, entonces $\hat{B} = B + (X'X)^{-1}X'E$, también se distribuye normalmente con matriz de media B y matriz de covarianza $(X'X)^{-1} \otimes \Sigma$.

Esto es $\hat{B} \sim N_p\left[B, (X'X)^{-1} \otimes \Sigma\right]$.

Por el resultado anterior para cada uno de los vectores $\beta_{(i)}$, $1 \leq i \leq p$, estos son independientes entre si, con matriz de covarianza Σ y con distribución normal, coincidiendo éste con los resultados anteriores.

C A P I T U L O   I I

PRUEBAS DE HIPÓTESIS

Pruebas de Hipótesis.

El análisis de regresión lineal multivariado, contempla varios tipos de pruebas de hipótesis con las que se pueden hacer inferencias de investigaciones.

Estas hipótesis estarán compuestas por el producto de matriz de la forma $C_1BM_1 = D$, donde $C_{1(gxq)}$, $M_{1(`pxr)}$, y $D_{(gxr)}$, son tales que C_1 y M_1, tienen rango g y r respectivamente; y C_1 , es una partición de C, esto es $C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}$ y $C' = (C_1', C_2')$. La hipótesis así planteada es llamada la hipótesis lineal general, con varias alternativas para las matrices M_1 y D_1 estas alternativas incorporan las posibles hipótesis de interés en el modelo lineal multivariado para la realización de las inferencias estadísticas.

Tipos de Prueba según las Matrices

Los tipos de prueba, dependerán primeramente de lo que se tenga interés de comprobar en la investigación, luego se hace la escogencia del producto de matrices adecuadas que formarán dichas hipótesis.

De esta forma se presentan los siguientes casos.

a). $Ho: C_1B = 0$ Contra $Ha: C_1B \neq 0$

donde $M_1 = I$, y $D = 0$.

Esta hipótesis nula es equivalente a probar la hipótesis Ho: $B = 0$.

En este caso se considera el hecho, que el efecto de la combinación lineal no está afectada por la matriz M_1 y que la misma es una matriz identidad, además que no hay influencia de las variables independientes en el modelo.

b). $\text{Ho}: C_1B = D$ Contra $\text{Ha}: C_1B \neq D$

Aquí $M_1 = I$, no hay influencia de esta matriz, pero si hay influencia de las variables independientes en el modelo, indicando que existe por lo menos una variable independiente que está influyendo en el modelo, y que la matriz B es distinta de la matriz cero.

c). $\text{Ho}: C_1BM_1 = D$ Contra $\text{Ha}: C_1BM_1 \neq D$

En este caso $M_1 \neq I$, y la combinación lineal indica que si hay influencia de las variables independientes en el modelo generando la matriz D; que la matriz B no es matriz nula.

En estos casos las filas de la matriz C_1, influyen sobre el efecto en las combinaciones lineales de la regresión de las variables independientes, las columnas de M_1, son focos de atención en particular de las combinaciones lineales de las variables independientes.

Distribución Wishart Centrada.

Definición II.1

Si E es una matriz de datos tal que $E \sim Np(0, I \otimes \Sigma)$ y P es una matriz idempotente, si M es una matriz que puede ser escrita como $M = E'PE$, diremos que M tiene un distribución Wishart centrada con matriz escalar Σ, y (n-q) grados de libertad, esto es $M \sim Wq(\Sigma, n-q)$.

Algunas Propiedades de la Distribución Wishart.

Proposición II.1

Si (n-q) es el número de columnas independiente de la matriz simétrica e independiente P, y $\hat{\Sigma}$ si es el estimador de máxima verosimilitud de Σ, entonces $n\hat{\Sigma} \sim Wq(\Sigma, n-q)$.

Demostración:

Sea P una matriz simétrica e idempotente y $\hat{E}$ el estimador de máxima verosimilitud de Σ, si $\hat{E} = PY$, además $PY = PE$, donde $\hat{\Sigma} = n^{-1}\hat{E}'\hat{E}$ por otro lado $n\hat{\Sigma} = \hat{E}'\hat{E}$ y como $\hat{E}'\hat{E} = E'PE$ donde $P = \left[I - X(X'X)^{-1}X'\right]$ y de rango (n-q), por **Proposición I.2**.

Por definición $E \sim Np(0, I \otimes \Sigma)$, como P es matriz idempotente y E es normal, multivariada, además como la $TrP = \text{rango } P = n - q$ por la **Proposición I.2**, entonces $E'PE \sim Wp(\Sigma, n-q)$, pero como $n\hat{\Sigma} = \hat{E}'P\hat{E}$ entonces $n\hat{\Sigma} \sim Wp(\Sigma, n-q)$ lo que se quería demostrar.

Proposición II.2

Si $E \sim Np(0, I \otimes \Sigma)$, entonces la matriz de datos $Z = EM_1 \sim Np(0, I \otimes M_1'\Sigma M_1)$, donde $M_{1(pxr)}$, y además E y M_1 son matrices independientes.

Demostración:

Sea $E \sim Np(0, I \otimes)$, y M_1 una matriz de orden (pxr) de constantes.

Como E y M_1 son matrices independientes entonces $E(Z) = E(EM_1) = E(E)E(M_1)$, pero E tiene matriz de media cero (0).

Luego $E(E)E(M_1) = 0(E)(M_1) = 0$

Por lo tanto E(Z)=0.

Por otro lado podemos escribir

$EM_1 = IEM_1$

Entonces $\text{cov}(EM_1) = \text{cov}(IEM_1)$

$$= E\left[(IEM_1 - 0)(IEM_1 - 0)'\right]$$
$$= E[(IEM_1 - 0)(M_1'E' - 0)]$$
$$= E[(IEM_1M_1'E'I)]$$

$$= I \otimes IM_1'\Sigma M_1{}^{(3)}$$

$$= I \otimes IM_1'\Sigma M_1$$

por lo tanto la matriz de covarianza de Z es $I \otimes M_1'\Sigma M_1$. Como E se distribuye normal, entonces $Z = EM_1$, también se distribuye normal con matriz de media cero (0) y matriz de covarianza $I \otimes M_1'\Sigma M_1$, esto es $Z \sim Np(0, I \otimes M_1'\Sigma M_1)$.

Lo cual se quería demostrar.

(3) Mardía et al (1979).

Proposición II.3

Si una matriz $X_{(nxq)}$ tiene distribución multivariante $Np(0, I \otimes M_1'\Sigma M_1)$ y **a** es un n-vector fijo, entonces $X'a \sim Nq[0,(a'a)\Sigma]$, donde $V = X'a$.

Demostración:

Consideremos $E(V) = E(X'a)$.

$$
\begin{aligned}
&= E(X')a \\
&= [E(X)]'a \\
&= 0.a \\
&= 0
\end{aligned}
$$

por lo tanto $E(V) = 0$.

Además si $V = X'a$ entonces $V = \sum_{i=1}^{n} X_{ji} a_i, \quad j = 1,2,\ldots,q$

La $\operatorname{cov}(Z) = E[(X'a - 0)(X'a - 0)']$

$$
\begin{aligned}
&= E[(X'a)(X'a)'] \\
&= E[(X'a)(a'X)'] \\
&= E\left[\left(\sum_{i=1}^{n} X_{ji} a_i\right)\left(\sum_{i=1}^{n} a' X_{ij}\right)\right] \\
&= E\left[\sum_{i=1}^{n} a'_i a_i (X_{ji} X_{ij})\right] \\
&= \sum_{i=1}^{n} a'_i a_i E(X_{ji} X_{ij}) \\
&= \sum_{i=1}^{n} a'_i a_i \Sigma' \\
&= (a'a)\Sigma
\end{aligned}
$$

por lo tanto la matriz de covarianza de V es $(a'a)\Sigma$ como X se distribuye normal, entonces $V = X'a$ también se distribuye normal con matriz de media cero (0) y matriz de covarianza $(a'a)\Sigma$ esto es $V \sim Np(0, (a'a)\Sigma)$.

Proposición II.4

Si M_1 es una matriz de orden (pxr) y P es idempotente de rango n-q, entonces la forma $M_1'Y'PYM_1 \sim Wp(M_1'\Sigma M_1, n-q)$, donde Y está dado por la ecuación (1).

Demostración:

Si P es una matriz idempotente y de rango n-q y además $Y'PY = E'PE$, donde $E \sim Np(0, I \otimes \Sigma)$, $R = M_1'Y'PYM_1$ y por otro lado tenemos que $EM_1 \sim Np(0, I \otimes M_1'\Sigma M_1)$ por **Proposición II.2**.

Por lo tanto como P es idempotente y EM_1 tiene distribución normal $Np(0, I \otimes M_1 \Sigma M_1)$, entonces $M_1'E'PEM_1 \sim Wp(M_1'\Sigma M_1, n-q)$ pero $M_1'E'PEM = M_1'Y'PYM_1$, luego $R \sim Wp(M_1'\Sigma M_1, n-q)$.

Proposición II.5

En el modelo $Y_+ = Y - XB_0$, donde B_0 es una matriz de orden (qxp), si además, $C_1B = D$ y $C_1B_0 = D$ se cumple que

$Y_+ ' P_2 ' Y_+ = E' P_2 E \sim Wp(\Sigma, g)$ donde P_2 es simétrica idempotente y de rango g.

Demostración:

Sea $Y_+ = Y - XB_0$, el modelo dado, y

$P_2 = X(X'X)^{-1} C_1 ' [C_1 (X'X)^{-1} C_1 ']^{-1} C_1 (X'X)^{-1} X'$ una matriz simétrica idempotente y de rango g.

Por hipótesis $C_1 B = D$ y $C_1 B_0 = D$ consideramos ahora $Y_+ ' P_2 Y_+$, donde $Y_+ = Y - XBo$.

Así: $Y_+ = XB + E \quad - XBo$

$$= X(B - Bo) + E$$

Luego $Y_+ ' P_2 Y_+ = [X(B - Bo) + E]' P_2 [X(B - Bo) + E]$

$$= [(B - Bo)' X' + E] P_2 [X(B - B0) + E]$$
$$= (B - Bo)' X' P_2 X(B - Bo) + (B - Bo)' X' P_2 E + E' P_2 X(B - Bo) + E' PE$$

puesto que $C_1 B = D$

$C_1 B = D$ entonces $C_1 (B - B_0) = 0$

$$(B - B_0) C_1 ' = 0$$

Luego tendremos que:

$$
\begin{aligned}
Y_+'P_2Y_+ &= (B-Bo)'C_1'\left[C_1(X'X)^{-1}C_1'\right]^{-1}C_1(B-Bo) \\
&\quad + (B-Bo)'C_1'\left[C_1(X'X)^{-1}C_1'\right]^{-1}C_1(X'X)^{-1}X'E \\
&\quad + E'X(X'X)^{-1}C_1'\left[C_1(X'X)^{-1}C_1'\right]^{-1}C_1(B-Bo) + E'P_2E \\
&= E'P_2E.
\end{aligned}
$$

Dado que los demás términos son cero, con lo que se demuestra que $Y_+'P_2Y+ = E'P_2E$.

Por otro lado como $E \sim Np(0, I \otimes \Sigma)$ y también que P_2 es una matriz idempotente de rango g.

Entonces $E'P_2E \sim Wp(\Sigma, g)$ lo que es equivalente a $Y_+'P_2Y_+ \sim Wp(\Sigma, g)$.

Distribución de $M_1'Y_+'P_2Y_+M_1$.

Proposición II.6

Si $H = M_1'Y_+'P_2Y_+M_1$ donde M_1 es una matriz de orden (pxr), $Y_+ = Y - XBo$, Bo una matriz de orden (qxp) y $Y_+'P_2Y_+ \sim Wp(\Sigma, g)$, entonces $M_1'Y_+'P_2Y_+M_1 \sim Wp(M_1'\Sigma M_1, g)$.

Demostración:

Sea $H = M_1'Y_+'P_2Y_+M_1$ y M_1 es una matriz (pxr), además $Y_+'P_2Y_+ \sim Wp(\Sigma', g)$ por **Proposición II.5**, puesto que $Y_+'P_2Y_+ = E'P_2E$, donde $E \sim Np(0, I \otimes \Sigma)$ entonces $M_1'Y_+'P_2Y_+M_1 = M_1'E'P_2EM_1$ también $EM_1 \sim Np(0, I \otimes M_1'\Sigma M_1)$ por **Proposición II.2** y como P_2 es idempotente, de rango g, se tiene entonces que $M_1'E'P_2EM_1 \sim Wp(M_1'\Sigma M_1, g)$, por lo tanto como $M_1'E'P_2EM_1 = M_1'Y_+'P_2Y_+M_1$ entonces $M_1'Y_+'P_2Y_+M_1 \sim Wp(M_1\Sigma M_1, g)$ lo que se requiere demostrar.

Proposición II.7

En el modelo $Y_+ = Z\Delta + E$, donde $Y_+ = Y - XBo$, $Z = XC^{-1}$ y $\Delta = (\Delta_1'\Delta_2')'$, probar la hipótesis $C_1B_0 = D$ es equivalente a demostrar que $\Delta_1 = 0$.

Demostración:

Sea $Y_+ = Z\Delta + E$, donde $Z = XC^{-1}$ y $\Delta = (\Delta_1'\Delta_2')'$ por lo tanto reemplazando estas expresiones en el modelo obtendremos que:

$$XB - XBo - E = XC^{-1}(\Delta_1'\Delta_2')' + E$$

$X(B - Bo) = XC^{-1}(\Delta_1'\Delta_2')'$ puesto que X tiene inversa, entonces $X^{-1}X(B-Bo)=X^{-1}XC^{-1}(\Delta_1'\Delta_2')'$

$$
\begin{aligned}
(B-Bo) &= -C^{-1}(\Delta_1'\Delta_2')' \\
C(B-Bo) &= CC^{-1}(\Delta_1'\Delta_2')' \\
C(B-Bo) &= I(\Delta_1'\Delta_2')' \\
C(B-Bo) &= \Delta
\end{aligned}
$$

luego para la hipótesis $C_1B=D$

dado que $C^1=(C_1^{'},C_2^{'})$ y además $C^{-1}=(C_1^{(1)},C_2^{(2)})$ por otro lado tenemos que

$$C(B-Bo)=\begin{pmatrix} \underset{gxq}{C_1} \\ \underset{(q-g)xq}{C_2} \end{pmatrix}(B-Bo)$$

y también $\Delta=(\Delta_1'\Delta_2')'=C(B-Bo)=\begin{pmatrix} C_1 \\ C_2 \end{pmatrix}(B-Bo)$

$$=\begin{pmatrix} C_1(B-Bo) \\ C_2(B-Bo) \end{pmatrix}$$

para la hipótesis $C_1B_o=D$, entonces

$$\Delta=(\Delta_1',\Delta_2')'=\begin{pmatrix} C_1(B-B_0) \\ C_2(B-B_0) \end{pmatrix}=\begin{pmatrix} C_1(B-D) \\ C_2(B-B_0) \end{pmatrix}$$

$$\Delta'=\begin{pmatrix} \Delta_1 \\ \Delta_2 \end{pmatrix}=\begin{pmatrix} C_1(B-D) \\ C_2(B-B_0) \end{pmatrix}$$

lo que equivale a:

$\Delta_1 = C_1B - D$ y $\Delta_2 = C_2(B - B_o)$

pero $\Delta_1 = 0$ Luego, $C_1B = D$

por lo tanto la hipótesis $C_1B = D$ es equivalente a $\Delta_1 = 0$.

La Prueba de Razón de Verosimilitud.

Si se selecciona una muestra aleatoria de una distribución y $L(\theta)$ es una función tanto de parámetros inconvenientes desconocidos como de θ. Si la hipótesis nueva Ho especifica que θ cae en un conjunto particular de posibles valores de Ω_0, esto es $H_o : \Theta = \Theta$, y la hipótesis alternativa especificada que θ cae en otro conjunto de posibles valores de Ω_a, tal como $H_a : \Theta > \Theta$, donde $\Omega_{0,}$ no se traslapa con Ω_a y $\Omega_0 \cup \Omega_a = \Omega$.

Sea $L(\Omega_0)$ la función de verosimilitud con todos los parámetros desconocidos reemplazados por sus estimadores de máxima verosimilitud, sujetos a la restricción de que $\theta \in \Omega_0$.

De manera similar se tiene $L(\hat{\Omega})$ pero con la restricción de que $\theta \in \Omega_0$. La prueba de razón de verosimilitud se basa entonces en la razón

$$\frac{L(\Omega_0)}{L(\hat{\Omega})} = \lambda.$$

Esta prueba utiliza el estadístico λ como estadístico de prueba y la región de rechazo se determina por $\lambda \leq k$, donde el valor de **k** se escoge de tal forma que α queda a un nivel predeterminado.

Definición II.2

La prueba de razón de verosimilitud (P.R.V.) de tamaño α para probar la hipótesis H_o contra H_a tiene como región de rechazo $K = \{w / \lambda(w) < c\}$, donde c es determinado por

$$\operatorname{Sup} P_0(w \in k) = \alpha \qquad \theta \in \Omega o$$

Estadístico de Wilk's.

Definición II.3

Si $M \sim Wp(I, m)$ y $N \sim Wp(I, n)$ son dos Wishart independientes, $m \geq p$, decimos que $\Delta = \dfrac{|M|}{|M+N|} = |I + M^{-1}N|^{-1} \sim \Delta(p, m, n)$, tiene una distribución Lambda Wilk's con parámetro p, m y n.

Donde m representa los grados de libertad del error, n los grados de libertad de la hipótesis; por lo que $m+n$ representan el total de los grados de libertad.

El Estadístico de Prueba dado por $\mathrm{Ln}\,\Delta$.

Corolario II.1

El estadístico de prueba, para la (P.R.V.) de la hipótesis nula $Ho : C_1B_2 = 0$, bajo el modelo $Y = XB + E$, esta dado por,

$\ln \Delta = \frac{n}{2} \ln \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|} \right)$, donde $\hat{\Sigma}$ y $\hat{\Sigma}_1$ son los estimadores de máxima verosimilitud de Σ_1 y Σ que tienen distribuciones **Wishart**.

Demostración:

La (P.R.V) de la hipótesis H_o: $C_1B=0$, es equivalente a probar la hipótesis H_o: $B_{(2)} = 0$, donde

$$B = \begin{bmatrix} B_{(1)\,(r+1)xp} \\ B_{(2)\,(q-r)xp} \end{bmatrix} \quad \text{para } X = \underset{nx(r+1)\ nx(q-r)}{[X_1 \vdots X_2]}$$

El modelo general puede ser escrito como $Y = X_1B_{(1)} + X_2B_{(2)} + E$, para la hipótesis $B_{(2)} = 0$ se tiene , $Y = X_1B_{(1)} + E$.

Esta prueba de razón de verosimilitud de la hipótesis nula está basada en la suma de cuadrados y productos cruzados, el cual es

$$(Y - X_1\hat{B}_{(1)})'(Y - X_1\hat{B}_{(1)}) - (Y - X\hat{B})'(Y - X\hat{B})$$

donde $\hat{B}_{(1)} = (X_1'X_1)^{-1}X_1'Y$;

$$\hat{\Sigma}_1 = \frac{(Y - X_1\hat{B}_{(1)})'(Y - X_1\hat{\Sigma}_{(1)})}{n}$$

y $$\hat{\Sigma} = \frac{(Y - XB)'(Y - XB)}{n}$$

que son los estimadores de máximo verosimilitud de Σ_1 y Σ, además tenemos que

$$L(\hat{B},\hat{\Sigma}) = -\frac{1}{2}np\log 2\pi - \frac{p}{2}\log\left|\hat{\Sigma}\right| - \frac{1}{2}np$$

$$= -\frac{1}{2}n\log\left|2\pi\hat{\Sigma}\right| - \frac{1}{2}np$$

$$L\left(\hat{B},\hat{\Sigma}\right) = \log\left|2\pi\hat{\Sigma}\right|^{-n/2} - \frac{1}{2}np$$

y también:

$$L(\hat{B}_{(1)},\hat{\Sigma}_1) = \log\left|2\pi\hat{\Sigma}_1\right|^{-n/2} - \frac{1}{2}np$$

por lo tanto

si $$L\left(\hat{B},\hat{\Sigma}\right) = \frac{e^{-np/2}}{(2\pi)^{-np/2}\left|\hat{\Sigma}\right|^{p/2}}$$

y $L\left(\hat{B}_1, \hat{\Sigma}_1\right) = \dfrac{e^{-np/2}}{(2\pi)^{-np/2}\left|\hat{\Sigma}_1\right|^{p/2}}$, la razón de verosimilitud Δ puede ser expresada en términos de la varianza generalizada.

Así,

$$\Delta = \frac{\max\limits_{B_{(1)},\Sigma} L\left(\hat{B}_{(1)}, \hat{\Sigma}_1\right)}{\max\limits_{B,\Sigma} L(\hat{B}, \hat{\Sigma})}$$

$$\Delta = \frac{L\left(\hat{B}_{(1)}, \hat{\Sigma}_1\right)}{L\left(\hat{B}, \hat{\Sigma}\right)}$$

$$\Delta = \frac{e^{-np/2} \Big/ (2\pi)^{np/2}\left|\hat{\Sigma}_1\right|^{-n/2}}{e^{-np/2} \Big/ (2\pi)^{np/2}\left|\hat{\Sigma}\right|^{-n/2}}$$

$$\Delta = \frac{\left|\hat{\Sigma}_1\right|^{-n/2}}{\left|\hat{\Sigma}\right|^{-n/2}} \text{ o bien } \left(\frac{\left|\hat{\Sigma}\right|}{\left|\hat{\Sigma}_1\right|}\right)^{n/2}$$

donde Δ será el estadístico

$$\Delta^{2/n} = \frac{\left|\hat{\Sigma}\right|}{\left|\hat{\Sigma}_1\right|}$$

$$2\ln\Delta = n\ln\left(\frac{\left|\hat{\Sigma}\right|}{\left|\hat{\Sigma}_1\right|}\right)$$

$$\ln\Delta = \frac{n}{2}\ln\left(\frac{\left|\hat{\Sigma}\right|}{\left|\hat{\Sigma}_1\right|}\right)$$

lo que se quería demostrar.

Proposición II.8

La (P.R.V) de la hipótesis nula Ho: $C_1B = D$ para el modelo $Y_+ = Z\Delta + E$ tiene como estadística de prueba

$$\lambda^{\frac{2}{n}} = \frac{\left|Y'PY\right|}{\left|Y'PY + Y'_+P_2Y_+\right|} = \frac{\left|\hat{\Sigma}\right|}{\left|\hat{\Sigma} + \hat{\Sigma}_2\right|}$$

El cual tiene distribución $\Delta(p, n-q, g)$, bajo la hipótesis nula.

Demostración:

Sea Ho: C_1B=D la hipótesis considerada, $Y_+ = Z\Delta + E$ el modelo.

Si $L\left(\hat{B},\hat{\Sigma}\right)=\frac{1}{2}\log\left|2\pi\hat{\Sigma}\right|^{-n/2}-\frac{1}{2}np$ es el valor máximo de la función de verosimilitud de B y Σ.

y $L\left(\hat{B}_2,\hat{\Sigma}_2\right)=\frac{1}{2}\log\left|2\pi\hat{\Sigma}_2\right|-\frac{1}{2}np$ es el valor de máximo de la función de verosimilitud; $\hat{B}_2$ y $\hat{\Sigma}_2$ estimadores de máximo verosimilitud de B_2 y Σ_2 respectivamente.

La razón de verosimilitud puede expresarse en términos de la varianza generalizada como:

$$\lambda=\frac{\max\limits_{B_{(2)},\Sigma_2} l\left(\hat{B}_2,\hat{\Sigma}_2\right)}{\max\limits_{B,\Sigma} l\left(\hat{B},\hat{\Sigma}\right)}$$

$$=\frac{e^{-pn/2}/(2\pi)^{pn/2}\left|\hat{\Sigma}_2\right|^{n/2}}{e^{-pn/2}/(2\pi)^{pn/2}\left|\hat{\Sigma}\right|^{n/2}}$$

por lo tanto

$$\lambda=\frac{\left|\hat{\Sigma}\right|^{n/2}}{\left|\hat{\Sigma}_2\right|^{n/2}}$$

$$\lambda^{2/n}=\frac{\left|\hat{\Sigma}\right|}{\left|\hat{\Sigma}_2\right|}$$

además Σ_2 es independiente de Σ.

Si $P_2 = P_1 - P$, entonces $P_2 + P = P_1$, luego se entiende que $Y'_+ P_2 Y_+$, $Y'PY$ se distribuye $W_p(\Sigma_2, g)$ y $W_p(\Sigma, n-q)$ respectivamente, por lo tanto

$$\lambda = \frac{\left|\hat{\Sigma}\right|^{n/2}}{\left|\hat{\Sigma}_1\right|^{n/2}}$$

$$\lambda = \left(\frac{\left|\hat{\Sigma}\right|}{\left|\hat{\Sigma}_1\right|}\right)^{n/2}$$

$$\lambda = \left(\frac{\left|n\hat{\Sigma}\right|}{\left|n\hat{\Sigma} + n\hat{\Sigma}_2\right|}\right)^{n/2}$$

$$\lambda^{2/n} = \left(\frac{\left|n\hat{\Sigma}\right|}{\left|n\hat{\Sigma} + n\hat{\Sigma}_2\right|}\right)$$

puesto que $n\hat{\Sigma} = Y'PY$ y también

$$n\hat{\Sigma}_2 = Y'_+ P_2 Y_+$$

tenemos entonces que

$$\lambda^{2/n} = \frac{|Y'PY|}{|Y'PY + Y_+'P_2Y_+|}$$

el cual se distribuye como $\Delta(p, n-q, g)$ por definición II.3, con lo cual se demuestra esta proposición.

Proposición II.9

El estadístico de prueba para la (P.R.V.) de la hipótesis nula Ho: $C_1B = D$ bajo el modelo de la ecuación (1), está dado por

$$\lambda^{2/n} = \frac{|R|}{|H+R|} \sim \Delta(r, n-q, g),$$ donde $R = M_1'Y'PYM_1$ y $H = M_1'Y'_+P_2Y_+M_1$ tiene distribución Wishart.

Demostración:

Sea la hipótesis nula Ho: $C_1B = D$, $Y = XB + E$ el modelo; $H = M_1'Y'_+P_2Y_+M_1$ y $R = M_1'Y'PYM_1$

consideramos la función de máximo verosimilitud de B y Σ, dada por

$L\left(\hat{B}, \hat{\Sigma}\right) = \left|2\pi n^{-1}\hat{\Sigma}\right|^{-n/2} \exp\left(-\frac{1}{2}np\right)$ para R será

$$\left|2\pi n^{-1}M'Y'PYM_1\right|^{-n/2} \exp\left(-\frac{1}{2}np\right)$$

lo que podemos expresar directamente como

$$|2\pi n^{-1}R|^{-n/2}\exp\left(-\frac{1}{2}np\right)$$

y para H, se tiene

$$|2\pi n^{-1}M_1'Y_+'P_2Y_+M_1|^{-n/2}\exp\left(-\frac{1}{2}np\right)$$

por otro lado como $R \sim Wr(M_1'\Sigma M_1, n-q)$

$H \sim Wr\ (M_1'\Sigma M_1, g)$, además como R y H son Wishart independientes, luego $(R+H) \sim Wr(M_1'\Sigma M_1, n-q+g)$.

Se tiene también que P y P_2 son proyecciones independientes $PP_2 = 0$.

Por un lado la máxima verosimilitud de $\hat{B}$ y $\hat{\Sigma}$ para (R+H) está dada por $|2\pi n^{-1}(R+H)|^{-n/2}\exp\left(-\frac{1}{2}np\right)$

luego el cociente de R y $(R+H)$ será

$$\lambda^{\frac{2}{n}} = \frac{|2\pi n^{-1}R|^{-n/2}\exp\left(-\frac{1}{2}np\right)}{|2\pi n^{-1}(R+H)|^{-n/2}\exp\left(-\frac{1}{2}np\right)}$$

$$\lambda^{\frac{2}{n}} = \frac{|2\pi n^{-1}R|^{-n/2}}{|2\pi n^{-1}(R+H)|^{-n/2}}$$

$$\lambda^{\frac{2}{n}} = \frac{|R|}{|R+H|} \sim \Delta(r, n-q, g)$$

además

$$\frac{|R|}{|R+H|} = \left|\frac{R}{R+H}\right|$$

$$= \left|\frac{1}{(R+H)R^{-1}}\right|$$

$$= \left|\frac{1}{I+HR^{-1}}\right|$$

$$= \left|I+HR^{-1}\right|^{-1}$$ [4]

y también

$$\left|I+HR^{-1}\right|^{-1} = \prod_{i=1}^{r}(I+\lambda i)^{-1}$$

$$= \prod_{i=1}^{K}(I+\lambda i)$$

donde $k = \min(g, r)$ y $\lambda_1 >_1 \lambda_2 >_1 \dots >_1 \lambda r$ denota los valores propios de HR^{-1}.

Pruebas de Unión e Intersección.

Una prueba de unión e intersección (P.U.I) para una hipótesis nula Ho es una prueba donde la región de rechazo correspondiente puede ser escrita como $R = URa$, siendo la región de rechazo correspondiente de la hipótesis Hoa y donde Ho se escribe como $Ho = \cap Hoa$.

---

[4] Mardia et al (1979).

La prueba de unión e intersección de la hipótesis $Ho: B_1 = 0$, utiliza como estadístico de prueba el estadístico de Roy's, dado por

$$\Theta = \frac{\lambda_1}{1+\lambda_1}$$

donde $1+\lambda_1$ es el mayor valor propio de HE^{-1}, comparado con el punto de mayor porcentaje Θ_2, dado de valores de las tablas.

Los parámetros utilizados son

$$s = \min(p,q), m = \frac{1}{2}(|q-p|-1)$$
$$v = \frac{1}{2}(n-q-p)-2)$$

la hipótesis es rechazada si $\Theta > \Theta_\lambda$, s, m, n.

Definición II.4

Si Ho es verdadera para todo vector **a** p-dimencional, diremos que Ho puede ser escrita como la intersección del conjunto de hipótesis univariantes Hoa que es $Ho = Hoa$.

De la definición anterior la hipótesis multivariada $C_1BM_1 = D$ es verdadera si y sólo si $b'C_1BM_1a = b'Da$, para todo vector **a** y **b** de dimensiones $(rx1)$ y $(gx1)$.

Podemos reemplazar las matrices C_1 y M_1 por $b'C_1$ y M_1a en las expresiones anteriores y obtenemos:

$H = M_1'Y_+'P_2Y_+M_1$ y $R = M_1'Y'YM_1$ obtenemos que la razón

$\frac{H}{R} = \frac{a'M_1'Y_+'PY_+M_1a}{a'M_1'Y'PYM_1a}$ tiene una distribución $\left[\frac{g}{(n-q)}\right]F_{g,n-q}$, cada una de estas hipótesis univariantes pueden ser probados por el

$$\delta = \frac{\left\{b'C_1(X'X)^{-1}XY_+Y_+M_1a\right\}^2}{\left\{b'C_1(X'X)^{-1}C_1b\right\}\left\{a'M'Y'PYM_1a\right\}} \sim (n-q)^{-1}F_{1,n-q}$$

bajo la hipótesis nula, para a y b fijos.

Proposición II.9

El cociente de $\frac{H}{R}$ genera al estadístico

$$\delta = \frac{\left\{b'C_1(X'X)^{-1}X^1Y_+M_1a\right\}^2}{\left\{b'C_1(X'X)^{-1}C_1'b\right\}\left\{a'M_1'Y'PM_1a\right\}}$$

que tiene una distribución $(n-q)^{-1}F_{1,n-q}$, para **a** y **b** fijos, en el modelo $Y_+ = Z\Delta + E$.

Demostración:

Si $H = M_1'Y_+'P_2Y_+M_1$, donde

$$P_2 = X(X'X)^{-1}C_1'\left[C_1(X'X)^{-1}C_1'\right]^{-1}C_1(X'X)^{-1}X' \quad \text{y}$$

$Y_+ = Z\Delta + E$, por lo cual

$$H = M_1'Y_+'X(X'X)^{-1}C_1'\left[C_1(X'X)^{-1}C_1'\right]^{-1}C_1(X'X)^{-1}X' \quad \text{y}$$

$R = M_1'Y'PYM_1$, entonces $R = M_1'Y'\left[I - X(X'X)^{-1}X'\right]YM_1$

sustituyendo $C_1 = b'C_1$ y $M_1 = M_1a$ se tiene

$$R = a'M_1'Y'\left[I - X(X'X)^{-1}X'\right]YM_1a \quad \text{y}$$

$H = a'M_1'Y_+'X(X'X)^{-1}C_1'b\left[b'C_1(X'X)^{-1}C_1'b\right]^{-1}bC_1(X'X)^{-1}X'Y_+M_1a$ por lo tanto rechazando el cociente de $\frac{H}{R}$ tenemos que

$$\frac{H}{R} = \frac{a'M_1'Y'_+X(X'X)^{-1}C_1'b\left[b'C_1(X'X)^{-1}C_1'b\right]^{-1}b'C_1'(X'X)^{-1}X'Y_+M_1a}{a'M_1'Y'\left[I - X(X'X)^{-1}X'\right]YM_1a}$$

$$= \frac{\left\{a'M_1'Y_+'X(X'X)^{-1}C_1'b\right\}\left\{b'C_1(X'X)^{-1}X'Y_+M_1a\right\}}{\left\{b'C_1(X'X)^{-1}C_1'b\right\}\left\{aM_1Y'PYM_1a\right\}}$$

$$= \frac{\left\{b'C_1(X'X)^{-1}X'Y_+M_1a\right\}^2}{\left\{b'C_1(X'X)^{-1}C_1'b\right\}\left\{a'M_1'Y'PYM_1a\right\}}$$

si maximizamos la expresión anterior sobre b tenemos que

$$\delta = \frac{a'Ha}{a'Ra} \sim \left(\frac{g}{n-q}\right)F_{\alpha,g,n-q}$$

para un **a** fijo.

Dado que H y R tienen distribuciones Wishart, independientes, para **a** y **b** fijos, entonces $\frac{H}{R} \sim (n-q)^{-1} F_{\alpha,1,n-q}$.

Proposición II.10

Si $H \sim Wp(M_1'\Sigma M_1, g)$ y $a \neq o$ es cualquier vector no aleatorio de orden (px1), tal que $a'M_1'\Sigma M_1 a \neq o$, entonces $\frac{a'Ha}{a'M_1'\Sigma M_1 a} \sim X^2 g$.

Demostración:

Sea $H \sim Wp(M_1'\Sigma M_1, g)$ y $a \neq o$ es cualquier vector no aleatorio de orden (px1), tal que $a'M_1'\Sigma M_1 a \neq o$, entonces por definición existe una matriz de datos multivariados X(pxn) que sigue una distribución normal, con matriz de media cero (0) y de matriz de covarianza Σ esto es $X \sim Np(0, I \otimes \Sigma)$, tal que $H = XX'$ además $a'Ha = a'XX'a$, que es igual a $a'X \sim Np(0, a'\Sigma a \otimes I)$(5)

$Np(0, I \otimes a'\Sigma a)$

Como $a'\Sigma a$ es unidimencional, el producto de $a'\Sigma a \otimes I$ (6)

$$= (a'\Sigma a)I$$

por lo tanto $a'X \sim Np[0, (a'\Sigma a)I]$

luego $\frac{a'Ha}{a_1'M_1'\Sigma M_1 a} \sim X_g^2$ (7)

tiene distribución de un cuadrado con g-grados de libertad.

$$\frac{a'Ra}{a'M_1'\Sigma M_1 a} \sim X^2_{n-q}$$

dado que $R \sim Wp(M_1'\Sigma M_1, n-q)$.

De lo cual obtenemos que el cociente de dos Chi-cuadrado tiene una distribución F.

Esto es $\frac{X_g^2}{X^2_{(n-q)}} \sim \left[\frac{g}{(n-q)}\right] F_{g,(n-q)}$ (8)

Intervalos de Confianza y la Correlación Múltiple.

El análisis de Regresión Lineal Multivariado, también contempla los intervalos simultáneos de confianzas basados en matrices de orden n y rango completo, como elemento que ayuda a complementar la inferencia estadística, de la regresión.

Para generar estos intervalos, consideramos como verdadero valor de la matriz de parámetros desconocidos B y, dado el modelo $Y_+ = Y - XB$, por el principio de unión e intersección tenemos que:

---

(7) Graybil (1961).
(8) Hogg Kraiz.

$$P\left\{\frac{\left[b'C_1(X'X)^{-1}X'Y_{+}M_1a\right]^2}{\left[b'C_1(X'X)^{-1}C_1'b\right]\left[a'M_1'Y'PYM_1a\right]}\leq\frac{\theta_\alpha}{(1-\theta_\alpha)}\right\}=1-\alpha$$

donde θ_α denota el valor crítico superior de la distribución $\theta(r, n-q, g)$.

Definición II.5

Sea $L \sim Wp(I, m)$ independiente de $T \sim Wp(I, n)$, donde $m > p$.

Entonces el mayor valor propio $\theta = \dfrac{\lambda_1}{(1+\lambda_1)}$ de $(L+T)^{-1}T$ es llamado la mayor raíz estadística, donde su distribución es denotada por $\theta(p, m, n)$.

Tenemos entonces que un intervalo simultáneo de confianza para $b'C_1BM_1a$ está dado por:

$$b'C_1(X'X)^{-1}X'YM_1a-\left\{\frac{\theta_\alpha}{(1-\theta_\alpha)}(a'Ra)\left[b'C_1(X'X)^{-1}C_1'b\right]\right\}^{1/2}$$

$$\leq b'C_1BM_1a \leq b'C_1(X'X)^{-1}X'YM_1a+\left\{\frac{\theta_\alpha}{(1-\theta_\alpha)}(a'Ra)\left[b'C_1(X'X)^{-1}C_1'b\right]\right\}^{1/2} \quad (10)$$

Proposición II.11

Sean $Y \sim Np(XB, I \otimes \Sigma)$ y el modelo definido en la ecuación (1), entonces el término $b'C_1(X'X)^{-1}X'YM_1a$ es un estimador insesgado de $b'C_1BM_1a$.

Demostración:

Consideremos $b'C_1(X'X)^{-1}X'YM_1a$, al reemplazar el valor de Y obtenemos que $b'C_1(X'X)^{-1}X'(XB+E)M_1a$. Luego si calculamos su valor esperado, esto es $E\left[b'C_1(X'X)^{-1}(XB+E)M_1a\right]$

$$
\begin{aligned}
&= E\left[b'C_1(X'X)^{-1}(X'X)BM_1a + b'C_1(X'X)^{-1}EM_1a\right] \\
&= E\left[b'C_1BM_1a + b'C_1(X'X)^{-1}X'EM_1a\right] \\
&= E\left[b'C_1BM_1a\right] + E\left[b'C_1(X'X)^{-1}X'EM_1a\right] \\
&= b'C_1E(B)M_1a + b'C_1(X'X)^{-1}X'E(E)M_1a
\end{aligned}
$$

por ser estos vectores y matrices no aleatoria

$$
\begin{aligned}
&= b'C_1BM_1a + b'C_1(X'X)^{-1}X'(0)M_1a \\
&= b'C_1BM_1a
\end{aligned}
$$

puesto que $E(B) = B$ y $E(E) = 0$. Luego concluimos que

$b'C_1(X'X)^{-1}X'YM_1a = b'C_1\hat{B}M_1a$ es un estimador insesgado de $b'C_1BM_1a$.

Proposición II.12

Si en la expresión $$\frac{\left[b'C_1(X'X)^{-1}X'Y_*M_1a\right]^2}{\left[b'C_1(X'X)^{-1}C_1'b\right]\left[a'M_1'Y'PYM_1a\right]} = L$$

el vector **b** es fijo, $C_1B = D$ y $C_1Bo = D$, entonces

$$L \sim (n-q)^{-1} T^2_{\alpha,r,n-q}$$
$$= \left[\frac{r}{(n-q-r+1)}\right] F_{\alpha,r,n-q-r+1}$$

Demostración:

Sea la expresión $\dfrac{\left[b'C_1(X'X)^{-1}X'Y_+M_1a\right]^2}{\left[b'C_1(X'X)^{-1}C_1'b\right]\;\left[a'M_1'Y'PYM_1a\right]}$

Y **b** un vector fijo.

La expresión anterior es equivalente a

$$\frac{\left\{a'M_1'Y'_+X(X'X)^{-1}C_1'b\left[b'C_1(X'X)^{-1}C_1'b'\right]^{-1}b'C_1(X'X)^{-1}X'Y_+M_1a\right\}}{\left[a'M_1'Y'PYM_1a\right]}$$ que

también podemos escribir como

$$\frac{\left\{a'M_1'Y'_+X(X'X)^{-1}C_1'b\left[a'M_1'Y'PYM_1a\right]^{-1}b'C_1(X'X)^{-1}X'Y_+M_1a\right\}}{\left[b'C_1(X'X)^{-1}C_1'b\right]}$$

puesto que **b** es fijo este denominador es un número.

Sea $k = b'C_1(X'X)^{-1}X'Y_+M_1a$

consideramos la distribución de k, así:

$$E(k) = E\left\{b'C_1(X'X)^{-1}X'Y_+M_1a\right\}$$

$$= E\left\{b'C_1(X'X)^{-1}X'\left[X(B-Bo)+E\right]M_1a\right\}$$
$$= E\left\{b'C_1(X'X)^{-1}(X'X)(B-Bo)M_1a + b'C_1(X'X)^{-1}X'EM_1a\right\}$$
$$= E\left\{b'C_1(B-Bo)M_1a + b'C_1(X'X)^{-1}X'EM_1a\right\}$$

si por hipótesis $C_1B = D$ y $C_1Bo = D$

entonces $C_1B=D$ menos $C_1Bo = D$

$$C_1B - C_1Bo = 0$$

$$C_1(B - Bo) = 0$$

El resultado de $b'C_1(B - Bo)M_1a = 0$

Se tendrá que $E\left[b'C_1(X'X)^{-1}X'EM_1a\right]$

$$b'C_1(X'X)^{-1}X'E(E)M_1a$$
$$= 0$$

puesto que $E(E) = 0$

Por otro lado la cov (k) = $E[(k-0)(k-0)']$

$$= E\left[\left(b'C_1(X'X)^{-1}X'Y_+M_1a\right)\left(a'M_1'Y_+'X(X'X)^{-1}C_1'b\right)\right]$$
$$= E\left[\left(b'C_1(X'X)^{-1}X'EM_1a\right)\left(a'M_1'E'X(X'X)^{-1}C_1'b\right)\right]$$
$$= E\left[b'C_1(X'X)^{-1}X'EM_1'aa'M_1'E'X(X'X)^{-1}C_1'b\right]$$

$$= b'C_1(X'X)^{-1}X'E[EM_1aa'M_1'E']X(X'X)^{-1}C_1'b$$
$$= b'C_1(X'X)^{-1}I \otimes a'M_1'\Sigma M_1aX(X'X)^{-1}C_1'b$$
$$= b'C_1(X'X)^{-1}(X'X)(X'X)^{-1}C_1'bI \otimes a'M_1'\Sigma M_1a$$
$$= b'C_1(X'X)^{-1}C_1'bI \otimes a'M_1'\Sigma M_1a$$

dado que $b'C_1(X'X)^{-1}C_1'bI \otimes a'M_1'\Sigma M_1a$ es un número

$$b'C_1(X'X)^{-1}c_1'bI = d$$

se tiene que $k \sim Np(O, dI \otimes a'M_1'\Sigma M_1a)$.

Si: $N = [a'M_1'Y'PYM_1a]$, por **Proposición II.9**, se tiene que $N \sim Wp(a'M_1'\Sigma M_1a, g)$.

Los límites de confianza para la ecuación (10) en el caso de que **a** y/o **b** estén dados a priori en

$$P\left\{ bC_1M_1Ba \in b'C_1(XX)^{-1}X'YM_1a \pm \left\{ \frac{\Theta_\alpha}{(1-\Theta_\alpha)}(a'Ra) \right\} \right.$$

$$\left. [b'C_1(X'X)^{-1}C_1'b]^{1/2} \right\} = 1-\alpha \qquad \forall a, b$$

pueden ser expresados de la siguiente manera para el vector a fijo, entonces el intervalo para b se obtiene sustituyendo $\frac{\Theta_\alpha}{1-\Theta_\alpha}$ por la **F de Fisher**

$$\left[\frac{g}{n-q}\right]F_{\alpha;g,n-q}.$$

Para el vector b fijo, entonces el intervalo de confianza se obtiene al sustituir $\frac{\Theta_\alpha}{1-\Theta_\alpha}$ por la **T^2 de Hotelling**

$(n-q)^{-1}T^2_{\alpha,r,n-q} = \left[\frac{r}{(n-q-r+1)}\right]F_{\alpha;r,n-q-r+1}$. Luego, éste será $bMBa \in b'$.

Correlación Múltiple.

Este mide la proporción de la variación total en las Y's que pueden ser atribuidas a la regresión en las X's y se denota por R^2.

Se utiliza para eliminar variables cuando la situación lo permita, además de que se puede llegar a una ecuación de predicción más fácil de trabajo, se debe encontrar la mejor ecuación de regresión que sólo involucre variables que sean predictorias útiles.

El criterio que se utiliza para obtener el mejor modelo de regresión ajustado es el coeficiente de determinación múltiple, el cual está dado por:

$$R^2 = \frac{ssR}{ssT}$$

$$= \frac{\sum_{i=1}^{n}\left(\hat{y}i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(yi - \bar{y}\right)^2}$$

$$= \frac{\hat{B}X'Y - n\bar{y}^2}{Y'Y - n\bar{Y}^2}$$

Esta cantidad indica que proporción de la variación total de la respuesta Y se explica mediante el modelo ajustado.

Si multiplicamos $R^2 x100\%$ el resultado se interpretará como el porcentaje de una variación explicada por el modelo.

Coeficiente de Correlación Múltiple.

La raíz cuadrada de R^2 es el coeficiente de correlación múltiple entre Y y el conjunto $X_1, X_2, \ldots, X_q$, dada por

$$R = \frac{\sqrt{SSR}}{\sqrt{SST}}$$

Proposición II.13

Si $W = E - 1'E'$ es la matriz de error centrada, entonces el vector columna $w^v \sim N_{np}(0, \Sigma \otimes H)$, donde $H = I - n^{-1}11'$.

Demostración:

Sea $W = E - 1'E'$ la matriz de error centrada, y E la matriz de error y w^v es un vector obtenido por ubicación de las columnas de W.

Consideramos $E(W^v)$

$$\begin{aligned}
&= [E(W)]^v \\
&= [E(E - 1\bar{e}')]^v \\
&= [E(E) - 1E(\bar{e}')]^v \; E(e^{-1}) = (E(e))^{-1} \\
&= (0 - 0)^v \\
&= 0
\end{aligned}$$

por lo tanto el valor esperado de W^v es la matriz nula.

por otro lado la $\mathrm{cov}(w^v)$

$$\begin{gathered}
E\left[(W^v - 0)(W^v - 0)'\right] \\
E\left[(W^v)(W^v)'\right] \\
E\left[(W^v)(W^v)'\right] \\
E\left[(W)^v (W)^v\right] \\
E\left[(E - 1e^{-1})(E - 1e')'\right] \\
E\left[(E - 1e^{-1})(E - \bar{e}1')\right] \\
E\left[(I - 11')EE'(I - 11')\right] \\
(I - 11')E(EE')(I - 11') \\
(EE')(I - 11')(I - 11') \\
(I \otimes \Sigma)
\end{gathered}$$ [9]

Por lo tanto, $w^v \sim N_p(0, \Sigma \otimes H)$.

Si $P = 1$, en la ecuación $Y = XB + E$, tendremos que $Y = 1\mu' + XB + E$,

el cual se convierte en $y - \bar{y}1 = X\beta + v$, donde $V \sim Np(O, T^2H)$, además donde $H = I - n - 11'$ las columnas de X están cada una centrada con media cero (0).

Se define $S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = n - 1\begin{pmatrix} y' - \bar{y}1' \\ X' \end{pmatrix}(y - \bar{y}1, X)$

---

[9] Dillon, 1965

Que es la matriz de covarianza de la muestra para **Y** y las columnas de **X**.

Dado que R denota la correlación de la matriz X, entonces:

$$\hat{\beta} = (X'X)^{-1}X'y$$
$$\hat{\beta} = (X'X)^{-1}X'(y - y1)$$
$$\hat{\beta} = S_{22}^{-1}S_{21}$$

luego la correlación de la muestra entre Y y $X\hat{B}$ está dada por:

$$Ryx = \operatorname{corr}\left(y, x\hat{\beta}\right)$$

$$= \frac{S_{12}S_{22}^{-1}S_{21}}{\{S_{11}S_{12}S_{22}^{-1}S_{21}\}^{1/2}}$$

$$= \frac{S_{12}\hat{\beta}}{\left\{S_{11}S_{12}\hat{\beta}\right\}^{1/2}}$$

$$= \left\{\frac{S_{12}S_{22}^{-1}}{S_{11}}\right\}^{1/2}$$

Por lo que Ryx es llamado el coeficiente de correlación múltiple de y con X; el cual puede ser expresado en términos de las correlaciones

$$R^2_{yx} = R_{12}R_{22}^{-1}R_{21} = \left\{\frac{S_{12}S_{22}^{-1}S_{21}}{S_{11}}\right\}^{1/2}$$

Correlación para Muestras Grandes.

Proposición II.14

La correlación para muestras grandes, entre **y** y una combinación lineal de las columnas de X, está dada por Ryx, y es obtenido por la combinación lineal de $X\hat{\beta}$, donde $\hat{\beta} = (X'X)^{-1}X'Y$ y que es llamado el coeficiente de regresión de y en X.

Demostración:

Sea y un vector columna de la matriz Y, y Xb una combinación lineal de las columnas de X.

Si $\quad Ryx = \operatorname{corr}(y, X\beta)$,

entonces $\operatorname{corr}(y, X\beta) = (\beta' S_{21})/(S_{11}\beta' S_{22}\beta)^{1/2}$

por lo que

$$\operatorname{corr}^2(y, X\beta) = (\beta' S_{21})^2 / (S_{11}\beta' S_{22}\beta)$$

esta función es una razón de formas cuadráticas en B y, por corolario el máximo de $A'X$ sujeto a $X'BX = 1$ es obtenido cuando el vector propio de $B^{-1}A$ corresponde al mayor valor propio.

Esto si λ_1 y λ_2 son los mayores valores propios de $B^{-1}A$, entonces sujeto a $X'BX = 1$.

$$\text{El máx } X'AX = \lambda_1 \quad , \quad \text{mín } X'AX = \lambda p$$

Más aun

$$\max \left\{ \frac{(a'X)^2}{(X'BX)} \right\} = a'B^{-1}a$$

y el máximo se obtiene cuando $X = \dfrac{B^{-1}a}{(a'B^{-1}a)^{1/2}}$

donde $(X'AX) = (a'X)^2 = X'(aa')X$.

El máximo está dado por $S_{12}S_{22}^{-1}/S_{11}$ y es obtenido cuando

$$\beta = S_{22}^{-1}S_{21} = (X'X)^{-1}X'y.$$

$$\text{Así la } \operatorname{corr}^2(y, X\beta) = \frac{(\beta' S_{21})^2}{(S_{11}\beta' S_{22}\beta)}$$

La que podemos escribir como

$$\operatorname{corr}^2(y, X\beta) = \frac{(S_{12}S_{22}^{-1}S_{21})^2}{\{S_{11}S_{12}S_{22}^{-1}S_{21}\}}$$

$$= S_{21}S_{22}^{-1}S_{21}S_{11}^{-1}$$

con lo que se demuestra que la corr(y, XB) es

$$\frac{S_{12}S_{22}^{-1}S_{21}}{\{S_{11}S_{12}S_{22}^{-1}S_{21}\}^{1/2}}$$

Definición II.6

La correlación múltiple de las columnas $X_{(1)}$ con respecto a $X_{(2)}, X_{(3)}, \cdots, X_{(p)}$ de una matriz X(nxq) está dada por $R_{1 \cdots 2 \cdots q}$.

Una correlación entre $Y_{(1)}$ y $Y_{(2)}$ eliminando después el efecto de X, puede ser definido como:

$$r_{12X} = \frac{\varepsilon'_{(1)} \varepsilon_{(2)}}{\left\{ \left\| \varepsilon_{(1)} \right\| \left\| \varepsilon_{(2)} \right\| \right\}}$$

El coeficiente r_{12X} es llamado el coeficiente de correlación parcial muestral entre $Y_{(1)}$ y $Y_{(2)}$ dado X.

Para calcular este coeficiente se particiona la matriz de covarianza de Y y X como

$$S = \begin{pmatrix} S_{11} & S_{12} & S_1' \\ S_{12} & S_{22} & S_2^1 \\ S_1 & S_2 & S_{22} \end{pmatrix} = n^{-1} \begin{pmatrix} Y'_{(1)} \\ Y'_{(2)} \\ X \end{pmatrix} \left(Y_{(1)}, Y_{(2)}, X \right)$$

el conjunto $S_{ij}^* = S_{ij} - S_i' S_{22}^{-1} S_j \quad i, j = 1,2$

entonces $r_{12X} = \dfrac{S_{12}}{\left\{ S_{11}^* S_{22}^* \right\}^{1/2}}$.

CAPÍTULO III

ANÁLISIS E INTERPRETACIÓN DE LOS RESULTADOS

Mejorar la calidad de vida de la población, disminuyendo los niveles de pobreza, a través de una adecuada distribución de los ingresos y servicios, son elementos fundamentales para el gobierno. En tal situación el Ministerio de Planificación y Política Económica (M.I.P.P.E.) y su Dirección de Políticas Sociales (D.P.S.) realizan la tarea de actualizar y ampliar las fuentes de información sobre el bienestar y la pobreza, en el año 1997, realizando la primera Encuesta de Niveles de Vida en Panamá, donde se entrevistaron alrededor de 5 591 viviendas en las áreas urbana, rural, indígena y de difícil acceso.

A continuación presentamos los resultados obtenidos para una base de datos de 3007 y 439 individuos en el área urbana e indígena, entre los 25 y 64 años, como una submuestra referente a la Encuesta de niveles de vida, donde se consideran las variables: Edad (Var1), ocupación (Var2), años de ocupación (Var3), número de personas (Var4), salario e ingreso; las cuales se distribuyen normalmente.

Análisis de los Resultados Obtenidos en el Área Urbana.

En el Cuadro 1 (Pág. 78) se muestran los valores promedios y las desviaciones estándar de cada una de las variables explicativas y explicadas que se están considerando en el estudio sobre bienestar y pobreza en el área urbana.

Observamos que el promedio del salario es aproximadamente de Bl. 362.92 y que el promedio del ingreso es de Bl. 453.69.

Para las variables explicativas las mayores dispersiones relativas se dan en las variables ocupación, con 66.8%: y años de ocupación con 95.6%, respectivamente. En el cuadro II, se presenta el análisis de varianza para el modelo de regresión multivariado de las variables explicativas con respecto a la variable salario, observamos que la probabilidad dada resultante es de (0.0001).

Análisis de Regresión Multivariado de la Variable Salario, para el Área Urbana.

En el cuadro III obtenemos los resultados para la realización de las pruebas de hipótesis del parámetro para la variable dependiente **salario** con respecto a las independientes.

Hipótesis

Ho: No existe influencia o relación lineal de las variables explicativas edad, ocupación, años de trabajo, número de personas, sobre la variable **salario**. (Ho: B = 0)

Ha: Las variables explicativas (Edad, ocupación, años de trabajo, números de personas) tienen influencia o relación lineal sobre la variable **salario**. (Ha: B ≠ 0)

En base a los resultados del cuadro III, concluimos que en forma general, las cuatro variables influyen en su forma conjunta sobre la variable salario, puesto que $P = 0.0001$, una probabilidad menor de 5%. ($P < 0.05$). Analizando por separado los resultados observamos que solamente la variable número de personas no influye linealmente sobre la variable dependiente salario. En el cuadro IV se presenta el análisis de varianza para el modelo de regresión multivariado de la variable explicada ingreso con respecto a las variables explicativas, obteniendo una probabilidad de 0.0001.

Análisis de Regresión Multivariado de la Variable Ingreso, para el Área Urbana.

Para el cuadro V, ésta presenta los resultados referente al análisis de regresión multivariado, donde verificamos las siguientes hipótesis. (Paquete S.A.S.)

Ho: No existe influencia o relación lineal de las variables explicativas edad, ocupación, años de trabajo, número de personas, sobre la variable **Ingreso**. (Ho: B = 0)

Ha: Las variables explicativas (Edad, ocupación, años de trabajo, números de personas) tienen influencia o relación lineal sobre la variable **Ingreso**. (Ha: $B \neq 0$)

Para la variable ingreso, obtenemos que el efecto conjunto de las variables explicativas sobre la variable ingreso se considera significativa al 5%, puesto que $p < 0.05$ ($p = 0.0001$). Cuando consideramos este efecto individual, vemos que en algunas variables existe una significancia estadística; para la ocupación y años de trabajo, ya que $p < 0.05$.

Concluimos que el ingreso está relacionado linealmente con las variables años de trabajo y ocupación. Podemos construir una ecuación de regresión para la variable ingreso en función de estas dos variables explicativas.

$$Y_{in} = 537.16 - 7.15V2 + 8.78V3$$

Observamos que el coeficiente de determinación R^2 respecto a la variable salario en función de las variables explicativas es de 0.0846, muy próximo a cero, lo que nos dice que sólo el 8.46% de la variación total en los valores de Y son explicados por el plano de regresión ajustado o bien que las variables explicativas no están influyendo sobre la variable salario, por lo que no es recomendable predecir sobre esta variable.

Para el ingreso encontramos que el coeficiente de determinación es también de 8.46%, un valor muy próximo cero al igual que las variables salario, encontramos que en este caso las variables edad y número de personas, no influyen en el ingreso de las personas.

La matriz de parámetros estimados, está dado por

$$\hat{B} = \begin{bmatrix} 537.16 & -428.73 \\ -4.47 & -3.57 \\ 2.11 & 1.68 \\ -7.15 & -5.72 \\ 8.78 & 7.02 \end{bmatrix}$$

Y la ecuación de estimación será $\hat{Y} = X\hat{B}$ que en forma matricial es:

$$\begin{bmatrix} \hat{Y}_{11} & \hat{Y}_{11} \\ \hat{Y}_{21} & \hat{Y}_{22} \\ \vdots & \vdots \\ \hat{Y}_{n1} & \hat{Y}_{n2} \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots X_{16} \\ 1 & X_{21} & X_{22} & \cdots X_{26} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots X_{n6} \end{bmatrix} \begin{bmatrix} 537.16 & -428.73 \\ -4.47 & -3.57 \\ 2.11 & 1.68 \\ -7.15 & -5.72 \\ 8.78 & 7.02 \end{bmatrix}$$

donde n = 3007.

Consideremos una matriz X_{1x4} y realizamos la estimación para un individuo, tomando en cuenta las variables edad, ocupación, año de ocupación, número de personas, donde se obtuvo la estimación de Bl. 353.36 de salario y Bl.441.75 de ingreso.

Se determinaron los valores propios de $E^{-1}H$ los cuales fueron de 0.4962 y 0.4001.

Los cuadros que a continuación presentamos contienen los resultados del análisis estadístico para el modelo de regresión lineal multivariado.

CUADRO I. ESTADÍSTICA DESCRIPTIVA, MUESTRA LOS VALORES PROMEDIOS Y LAS DESVIACIONES ESTÁNDAR DE CADA UNA DE LAS VARIABLES EN EL ESTUDIO SOBRE NIVELES DE VIDA EN EL ÁREA URBANA, 1997.

Variable	Promedio	Desviación Estándar
Edad	39.99	10.1152
Ocupación	32.70	21.8597
Annocup	9.96	9.5226
Npers.	4.76	2.2873
Salario	362.95	528.0718
Ingreso	453.69	660.0898

Fuente: Datos proporcionados por Economía y Finanzas del Ministerio de Planificación y Política Económica.

CUADRO II. ANÁLISIS DE VARIANZA PARA EL MODELO DE REGRESIÓN LINEAL MULTIVARIADO DE LA VARIABLE EXPLICADA SALARIO CON RESPECTO A LAS VARIABLES EXPLICATIVAS EN EL ESTUDIO DE NIVELES DE VIDA EN EL ÁREA URBANA, 1997.

Variable	G.l.	s.s.	c.m.	f.	p.
Modelo	4	70954090.205	17738522.55	69.401	0.0001
Error	3002	7672986924	255595.833		
c.total	3006	83825278261			

Fuente: Datos proporcionados por Economía y Finanzas del Ministerio de Planificación y Política Económica

CUADRO III. ESTIMACIÓN DE PARÁMETROS, EL ESTADÍSTICO PARA LA UNIFICACIÓN DE LA HIPÓTESIS NULA Y LAS RESPECTIVAS PROBABILIDADES EN CADA UNO DE LOS CASOS PARA LAS VARIABLES EN ESTUDIO CON RESPECTO A LA VARIABLE SALARIO EN EL ÁREA URBANA, 1997.

Parámetro	Estimación	T para la Ho Parámetro = 0	Prob > |T|
Intercepto	– 40.80197	9.248	0.0001
Var.1	1.6882	1.578	0.1147
Var.2	– 5.7270	– 13.447	0.0001
Var.3	– 2.0293	– 6.175	0.0001
Var.4	– 3.5784	– 0.881	0.3783

Fuente: Datos proporcionados por Economía y Finanzas del Ministerio de Planificación y Política Económica

CUADRO IV. ANÁLISIS DE VARIANZA PARA EL MODELO DE REGRESIÓN LINEAL MULTIVARIADO DE LA VARIABLE EXPLICADA INGRESO CON RESPECTO A LAS VARIABLES EXPLICATIVAS EN ESTUDIO SOBRE NIVELES DE VIDA EN EL ÁREA URBANA, 1997.

Variable	G.L.	s.s.	c.m.	f.	p.
Model	4	110865220.48	27716442.62	69.401	0.0001
Error	3002	119804190.00	39368.48		
c.total	3006	1309769960.5			

Fuente: Datos proporcionados por Economia y Finanzas del Ministerio de Planificación y Política Económica

CUADRO V. ESTIMACIÓN DE PARÁMETROS, EL ESTADÍSTICO T PARA LA VERIFICACIÓN DE LA HIPÓTESIS NULA Y LAS RESPECTIVAS PROBABILIDADES EN CADA UNO DE LOS CASOS PARA LAS VARIABLES EN ESTUDIO CON RESPECTO A LA VARIABLE INGRESO EN EL ÁREA URBANA, 1997.

Parámetro	Estimación	T para la Ho Parámetro = 0	Prob > |T|
Intercepto	537.165124	9.248	0.0001
Var.1	2.110364	1.578	0.1147
Var.2	– 7.158852	– 13.447	0.0001
Var.3	8.786674	6.175	0.0001
Var.4	– 4.473071	– 0.881	0.383

Fuente: Datos proporcionados por Economía y Finanzas del Ministerio de Planificación y Política Económica.

Análisis de los Resultados Obtenidos en el Área Indígena.

Referente al área indígena presentamos los siguientes resultados:

El cuadro VI muestra los valores promedios y las desviaciones estándar de cada una de las variables explicativas y explicadas que se están considerando en el estudio sobre bienestar y pobreza en el área indígena.

Observamos que el promedio del salario es de aproximadamente Bl. 116.75 y que el ingreso promedio es de Bl. 145. Para las variables explicativas las mayores dispersiones relativas se dan en las variables años de trabajo y número de personas, con 78.8% y 50%, respectivamente.

En el cuadro VII, se presenta el análisis de varianza para el modelo de regresión multivariado de las variables explicativas con respecto a la variable salario, observamos que la probabilidad dada resultante es de (0.0018.

Análisis de Regresión Multivariado de la Variable Salario, para el Área Indígena.

Para el cuadro VIII obtenemos los resultados para la realización de las pruebas de hipótesis del parámetro B para la variable dependiente **salario** con respecto a las variables independientes.

Hipótesis

Ho: No existe influencia o relación lineal de las variables explicativas edad, ocupación, años de trabajo, número de personas, sobre la variable **salario**. (Ho: $B = 0$)

Ha: Las variables explicativas (Edad, ocupación, años de trabajo, números de personas) tienen influencia o relación lineal sobre la variable **salario**. (Ha: $B \neq 0$)

En base a los resultados del cuadro VIII, concluimos que en forma general, las cuatro variables influyen en forma conjunta sobre la variable salario, puesto que $P = 0.0001$, una probabilidad menor de 5%. ($P < 0.05$)

Analizando por separado los resultados observamos que las variables edad y años de trabajo no influyen linealmente en la variable dependiente **salario**.

En el cuadro IX se presenta el análisis de varianza para el modelo de regresión multivariado de la variable explicada ingreso con respecto a las variables explicativas, donde obtenemos una probabilidad baja de 0.0018.

Análisis de Regresión Multivariado de la Variable Ingreso, para el Área Indígena.

En el cuadro X, se presentan los resultados referentes al análisis de regresión multivariado, donde verificamos las siguientes hipótesis.

Ho: No existe influencia o relación lineal de las variables explicativas edad, ocupación, años de trabajo, número de personas, sobre la variable **Ingreso**. (Ho: $B = 0$)

Ha: Las variables explicativas (Edad, ocupación, años de trabajo, números de personas) tienen influencia o relación lineal sobre la variable **Ingreso**. (Ha: $B \neq 0$)

Para la variable ingreso, obtenemos que el efecto conjunto de las variables explicativas sobre la variable ingreso se considera significativa al 5%, puesto que $p < 0.05$ ($p = 0.0018$). Cuando consideramos este efecto individual, vemos que en algunas variables existe una significancia estadística; para la ocupación y años de trabajo, ya que $p < 0.05$.

Concluimos que el ingreso está relacionado linealmente con las variables ocupación y número de personas. Podemos construir una ecuación de regresión para la variables ingreso en función de estas dos variables explicativas.

$$Y_{in} = 263.18 - 2.24V2 - 5.04V4$$

Observamos que el coeficiente de determinación R^2 respecto a la variable salario en función de las variables explicativas es de 0.0462, muy próximo a cero, lo que nos dice que sólo el 4.62% de la variación total en los valores de Y son explicados por el plano de regresión ajustado o bien que las variables explicativas no están influyendo en la variable salario, por lo que no es recomendable predecir sobre esta variable.

Para el ingreso encontramos que el coeficiente de determinación es también de 8.46%, un valor muy próximo cero al igual que las variables salario, encontramos que en este caso las variables edad y número de personas, no influyen en el ingreso de las personas.

La matriz de parámetros estimados, está dado por

$$\hat{B} = \begin{bmatrix} 263.13 & 210.54 \\ -0.49 & -0.39 \\ -2.24 & -1.79 \\ 0.29 & 0.23 \\ -5.03 & -4.03 \end{bmatrix}$$

Y la ecuación de estimación será $\hat{Y} = X\hat{B}$ que en forma matricial es:

$$\begin{bmatrix} \hat{Y}_{11} & \hat{Y}_{11} \\ \hat{Y}_{21} & \hat{Y}_{22} \\ \vdots & \vdots \\ \hat{Y}_{n1} & \hat{Y}_{n2} \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots X_{16} \\ 1 & X_{21} & X_{22} & \cdots X_{26} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots X_{n6} \end{bmatrix} \begin{bmatrix} 263.13 & 210.54 \\ -0.49 & -0.39 \\ -2.24 & -1.79 \\ 0.29 & 0.23 \\ -5.03 & -4.03 \end{bmatrix}$$

donde n = 439.

Consideremos una matriz $X_{1\times 4}$ y realizamos la estimación para un individuo, tomando en cuenta las variables edad, ocupación, año de ocupación, número de personas, donde se obtuvo la estimación de Bl. 133.43 de salario y Bl. 166.66 de ingreso.

Se determinaron los valores propios de $E^{-1}H$ los cuales fueron de 0.4002 y 0.3901.

Los cuadros que a continuación presentamos contienen los resultados del análisis estadístico para el modelo de regresión lineal multivariado.

CUADRO VI. ESTADÍSTICA DESCRIPTIVA, MUESTRA LOS VALORES PROMEDIOS Y LAS DESVIACIONES ESTÁNDAR DE CADA UNA DE LAS VARIABLES EN EL ESTUDIO SOBRE NIVELES DE VIDA EN EL ÁREA INDÍGENA, 1997.

Variable	Promedio	Desviación Estándar
Edad	39.82	10.0455
Ocupación	28.45	11.3580
Annocup	16.48	12.9816
Npers.	7.63	3.8223
Salario	116.75	140.9336
Ingreso	145.95	176.1670

Fuente: Datos proporcionados por Economía y Finanzas del Ministerio de Planificación y Política Económica.

CUADRO VII. ANÁLISIS DE VARIANZA PARA EL MODELO DE REGRESIÓN LINEAL MULTIVARIADO DE LA VARIABLE EXPLICADA SALARIO CON RESPECTO A LAS VARIABLES EXPLICATIVAS EN EL ESTUDIO DE NIVELES DE VIDA EN EL ÁREA INDÍGENA, 1997.

Variable	G.l.	s.s.	c.m.	f.	p.
Model	4	336976.6337	84244.1584	4.372	0.0018
Error	435	8382569.0953	19270.2738		
c.total	439	8719545.7289			

Fuente: Datos proporcionados por Economía y Finanzas del Ministerio de Planificación y Política Económica

CUADRO VIII. ESTIMACIÓN DE PARÁMETROS, EL ESTADÍSTICO PARA LA UNIFICACIÓN DE LA HIPÓTESIS NULA Y LAS RESPECTIVAS PROBABILIDADES EN CADA UNO DE LOS CASOS PARA LAS VARIABLES EN ESTUDIO CON RESPECTO A LA VARIABLE SALARIO EN EL ÁREA INDÍGENA, 1997.

Parámetro	Estimación	T para la Ho Parámetro = 0	Prob > |T|
Intercepto	210.5423	6.178	0.0001
Var.1	– 0.3969	– 0.508	0.6119
Var.2	– 1.7884	– 2.943	0.0034
Var.3	0.2305	0.372	0.7100
Var.4	– 4.0301	– 2.300	0.0219

Fuente: Datos proporcionados por Economía y Finanzas del Ministerio de Planificación y Política Económica

CUADRO IX. ANÁLISIS DE VARIANZA PARA EL MODELO DE REGRESIÓN LINEAL MULTIVARIADO DE LA VARIABLE EXPLICADA INGRESO CON RESPECTO A LAS VARIABLES EXPLICATIVAS EN ESTUDIO SOBRE NIVELES DE VIDA EN EL ÁREA INDÍGENA, 1997.

Variable	G.l.	s.s.	c.m.	f.	p.
Modelo	4	526526.0951	131631.5237	4.372	0.0018
Error	435	13097759.308	30109.7915		
c.total	439	13624285.403			

Fuente: Datos proporcionados por Economía y Finanzas del Ministerio de Planificación y Política Económica

CUADRO X. ESTIMACIÓN DE PARÁMETROS, EL ESTADÍSTICO T PARA LA VERIFICACIÓN DE LA HIPÓTESIS NULA Y LAS RESPECTIVAS PROBABILIDADES EN CADA UNO DE LOS CASOS PARA LAS VARIABLES EN ESTUDIO CON RESPECTO A LA VARIABLE INGRESO EN EL ÁREA INDÍGENA, 1997.

Parámetro	Estimación	T para la Ho Parámetro = 0	Prob > |T|
Intercepto	263.1780	6.178	0.0001
Var.1	– 0.4961	– 0.508	0.6119
Var.2	– 2.2355	– 2.943	0.0034
Var.3	0.2882	0.372	0.7100
Var.4	– 5.0377	– 2.300	0.0219

Fuente: Datos proporcionados por Economía y Finanzas del Ministerio de Planificación y Política Económica.

CONCLUSIONES Y RECOMENDACIONES

1. El análisis de regresión multivariado nos permite estudiar un gran número de variables explicativas y explicadas, eliminando aquellas variables que no influyen en el modelo.
2. El modelo de regresión multivariado nos permite hacer estimaciones de varias variables explicativas a la vez y obtener varias variables respuestas.
3. La distribución de la matriz del modelo Y_{nxp} es $(XB, I \otimes \Sigma)$ lo que implica que cada uno de los vectores $Y_{(i)}$ son mutuamente independientes con matriz de media $B'X_i$ y matriz de varianza covarianza Σ.
4. La matriz $\hat{B}$ de parámetros estimados, tienen distribución normal con media B y matriz de varianza covarianza $(X'X)^{-1} \otimes \Sigma$.
5. La matriz $\hat{B}$ es un estimador insesgado y de máxima verosimilitud de B.
6. Eliminadas las variables que no influyen en el modelo, se puede ajustar el modelo con las variables significativas para realizar estimaciones.
7. En el estudio realizado sobre bienestar y pobreza en diferentes áreas, recomendamos que las variables consideradas sean analizadas por área.
8. Observamos que las variables: ocupación y años de trabajo son las que más influyen en el modelo para la variable ingreso.
9. Para el área indígena, observamos que la variable ocupación y número de personas son las que más influyen en el modelo.

BIBLIOGRAFÍA

1. **ANDERSON**, T.W. 1958. An introduction to multivariate analysis. 1ª ed. JOHN WILEY & SONS, Inc. Nueva York. 374 págs.

2. **MANLY**, BRYAN F.J. 1970. Multivariate Statistical methods A primer. 2ª ed. PERGANON PRESS Ltd. 350 págs.

3. **MARDIA**, K.V., KENT . Multivariate Analysis,. 6ª ed.

4. J.T. y BIBAY, J.M. 1988. ACADEMIC PRESS. HARCOURT BRACE JOVANOVICH, PUBLISHERS, London, San Diego. 521 págs.

5. **LINARES**, G.; ACOSTA. Estadística Multivariada. Unidad Administrativa

6. **L.; SISTACH**, V. 1986. Santiago. Establecimiento imprenta. 318 págs.

7. **SEARLE**, S. R. 1971. Linear Models. 1ª ed. JOHN WILEY & SONS, Inc. Nueva York. 527 págs.

8. **NETER**, J.; WASSWEMAN. Applied Lineal Regression Models. 1ª ed.

9. **W.; KUTNER**, M. H. 1983. RICHARD IRWIN, Inc. Homewood, Illinois. 547 págs.

10. **KERLINGER**, F.N.; Multiple Regression Behaviral Research. 1ª ed.

11. **PEDHAZAR**, E. J. 1983. ADDISON WESLEY PUBLISHING COMPANY. 378 págs.

12. **MORRISON**, D. F. 1976. Multivariate Statistical Methods. 2ª ed. MC. GRAW HILL BOOK COMPANY, New York. 415 págs.

13. **MYERS**, D. 1975. Classical and Modern Regression With Applications. 1ª ed. JOHN WILEY & SONS, Inc. 475 págs.

14. **DRAPER,** N.R y **SMITH** H. 1966. Aplicad Regression Analysis. JOHN WILEY & SONS. New York. 405 págs.

15. **KSHIRZAGAR**, A. 1972. Multivariate Analysis. 2ª ed. MARCEL DEKKER, Inc. 435 págs.

16. **LINARES y VASQUEZ.** 1975. Estadística Multivariada. FACULTAD DE PSICOLOGÍA. La Habana, Cuba. Págs. 378.

17. **MONTGOMERY y PECK.** 1982. Regression Analysis. 2ª ed. FRICKER VIEWEG & SONS BRAURSCHWEIG. 485 págs.

18. **DILLON**, R. 1965. Multivariate Analysis Methods and Applications. 3ª ed. ADDISON WESLEY PUBLISHING COMPANY. 365 págs.

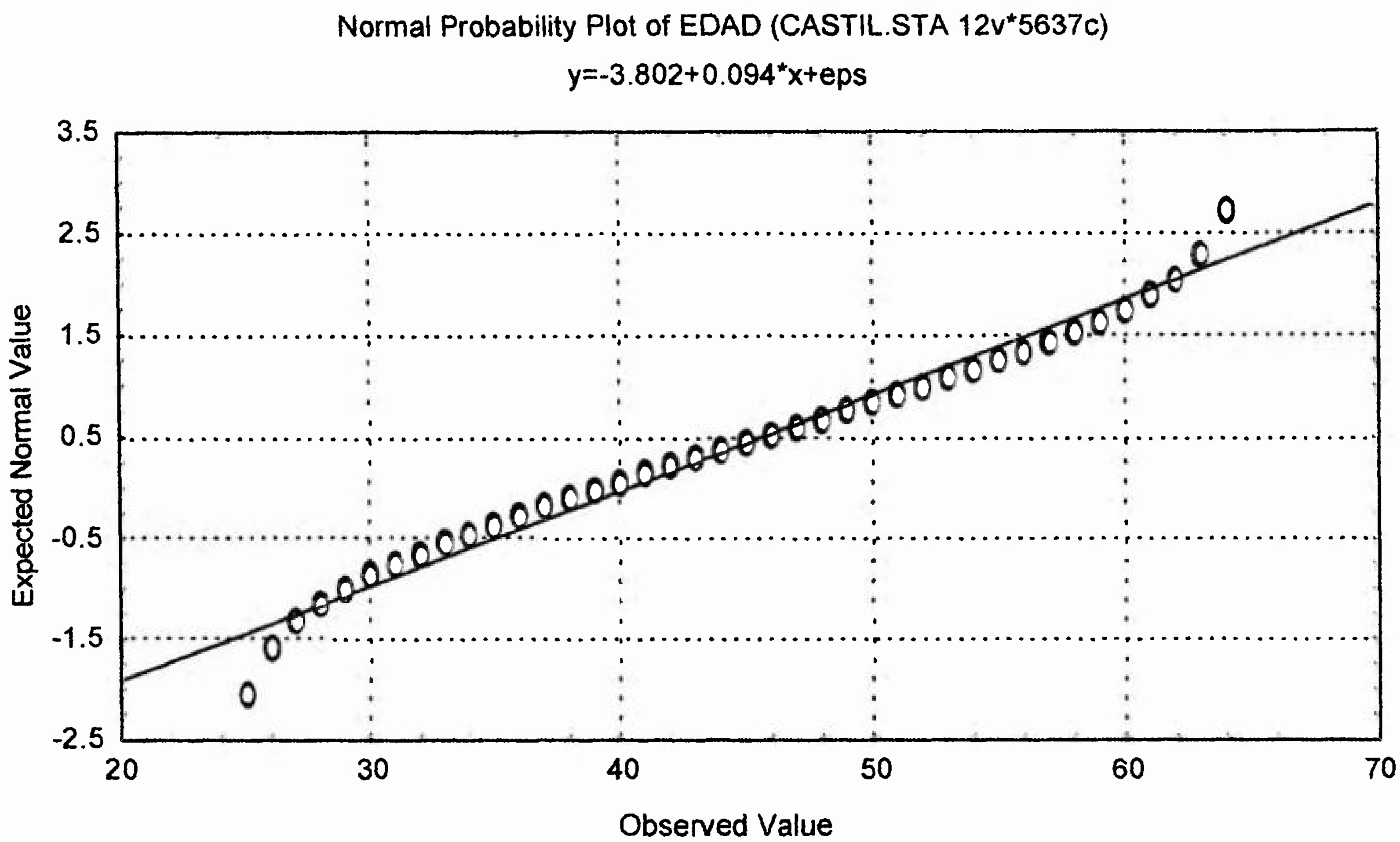
19. **EVERITT**, Q. 1988. Applied Multivariate Data Analysis. 1ª ed. JOHN WILEY & SONS, Inc. 420 págs.

20. **MUIRHEAD J.**, ROBB. 1982. Aspects of Multivariate Statistical Theory. JOHN WILEY & SONS, Inc. 675 págs.

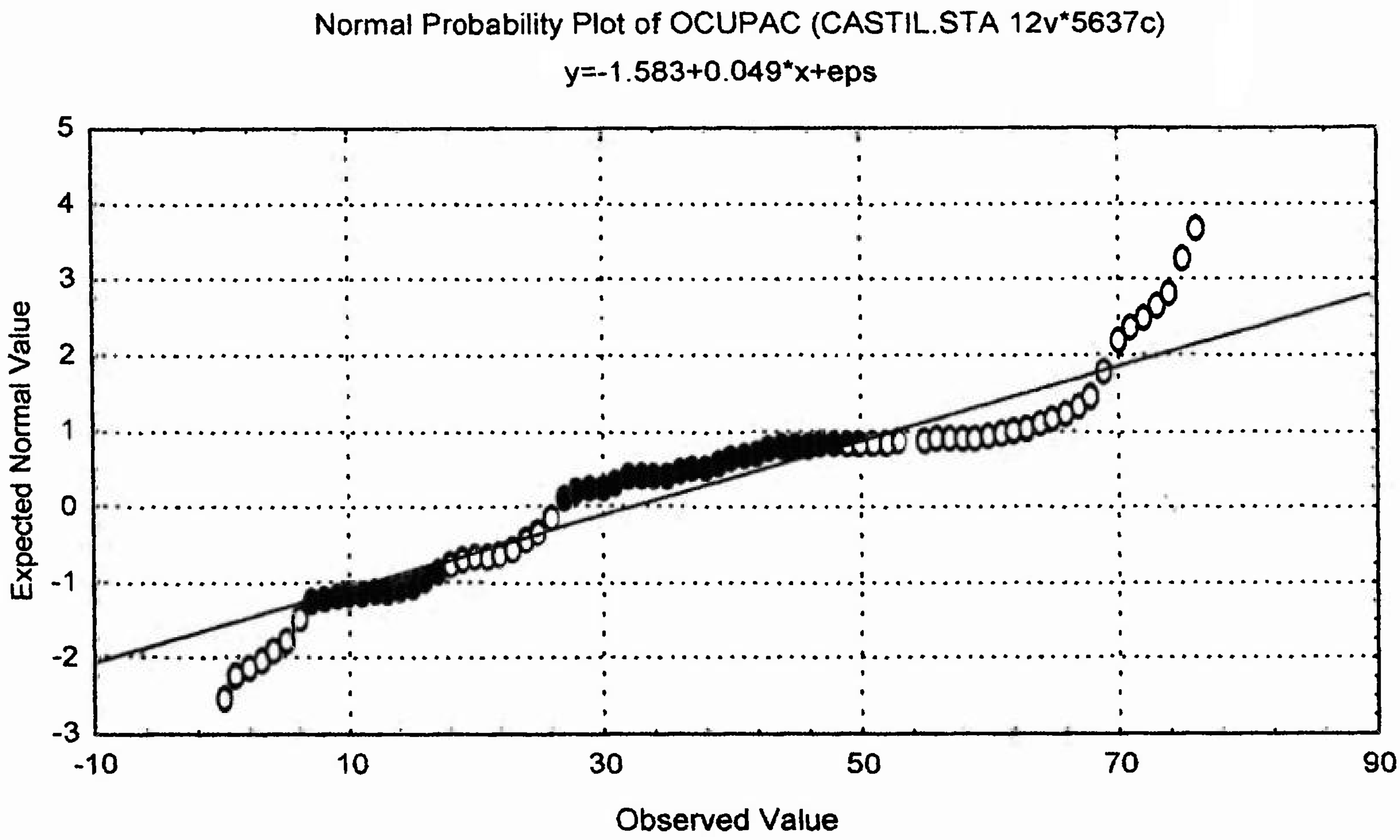
21. **GRAYBILL A.**, F. 1961. An Introduction to Linear Statis-tical Models. Vol. 1. MC. GRAW HILL BOOK COMPANY, Inc. New York. 462 págs.

22. **PRESS S.**, J. 1982. Applied Multivariate Analysis. 2ª ed. ROBERT E, KRIEGER PUBLISHING COMPANY. Malabar, Florida. 630 págs.

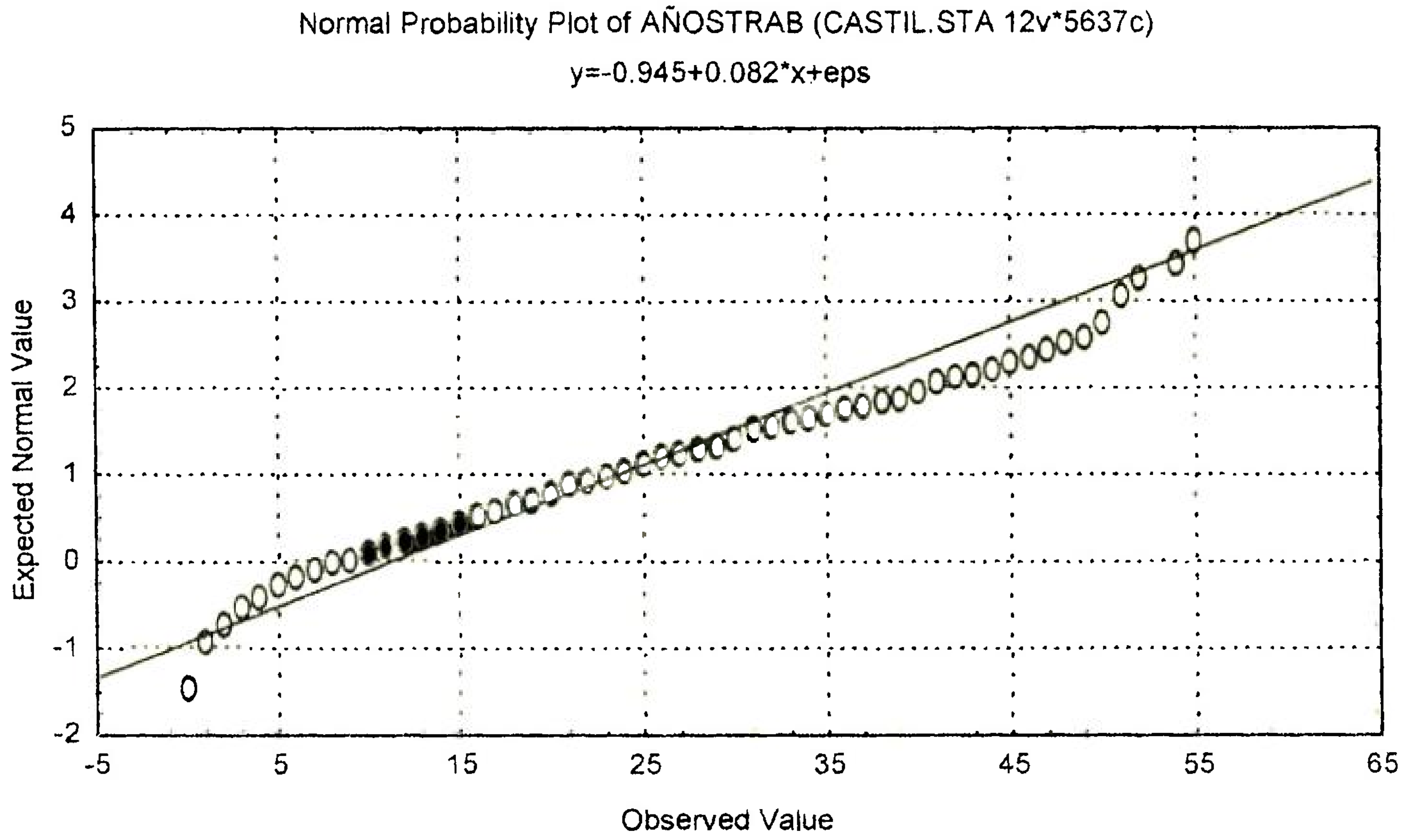
ANEXO



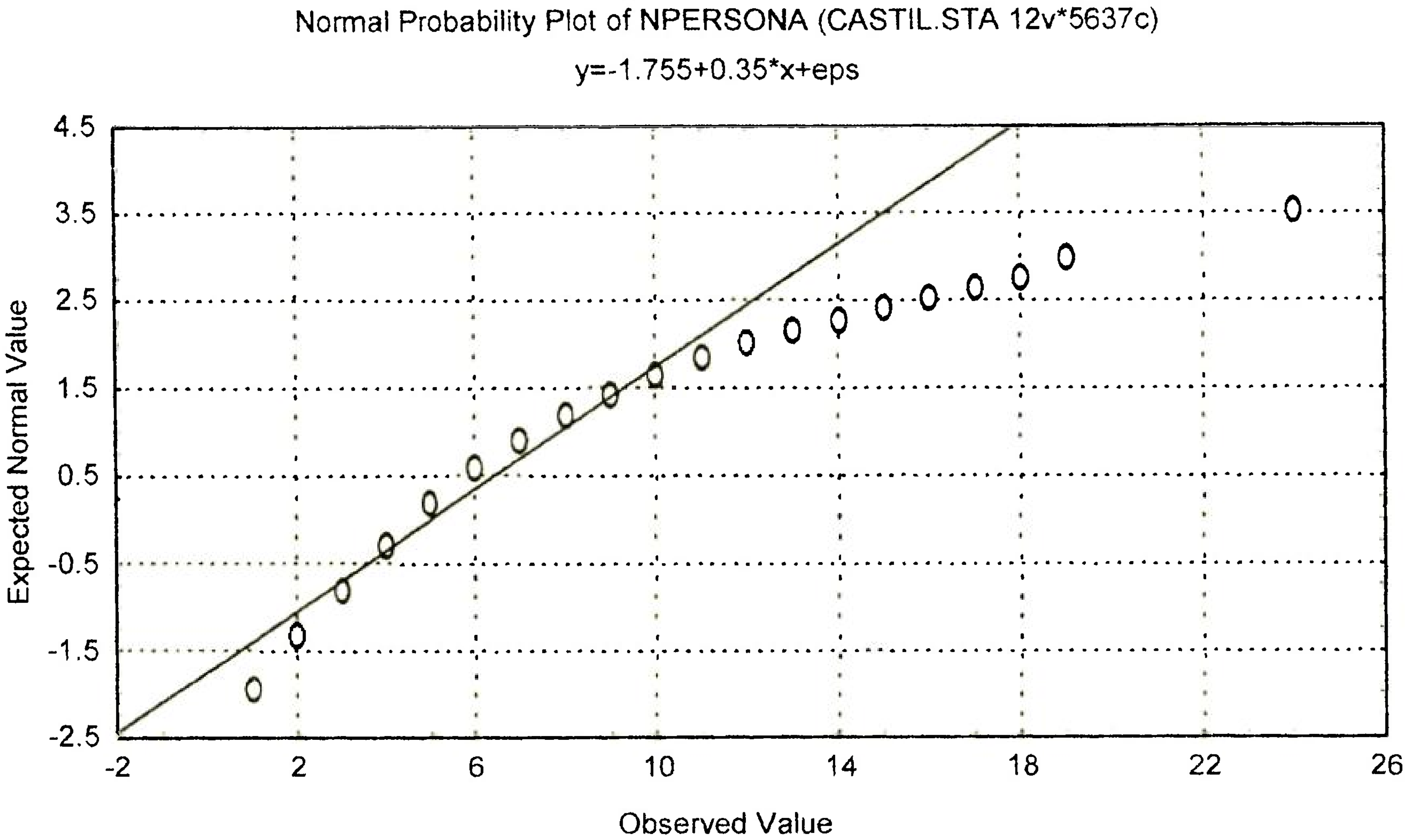
Fuente: Datos proporcionados por Economía y Finanzas del Ministerio de Planificación y Política Económica.



Fuente: Datos proporcionados por Economía y Finanzas del Ministerio de Planificación y Política Económica.



Fuente: Datos proporcionados por Economía y Finanzas del Ministerio de Planificación y Política Económica.



Fuente: Datos proporcionados por Economía y Finanzas del Ministerio de Planificación y Política Económica.