ANEXO 1

Fundamentos Teóricos de Regresión Lineal Múltiple

FUNDAMENTOS TEÓRICOS DE REGRESIÓN LINEAL MÚLTIPLE

Los fundamentos teóricos expuestos en este anexo fueron extraidos de una versión tradu[illegible] libro "Econometría Básica" del economista Damodar Guajarati. En 1995, McGraw-Hill p[illegible] la segunda edición de este libro y en 2002, McGrall-Hill y Irwin publicaron la cuarta y [illegible] reciente edición. El economista Víctor Manuel Mayorga Torrado tradujo la segunda edición, la cual fue publicada por McGraw-Hill en 1996.

1. DEFINICIÓN

La regresion lineal múltiple se puede definir como una función de estadística de dependencia lineal entre una variable aleatoria dependiente, explicada o endógena (Y) y varias variables aleatorias independientes, explicativas o exógenas (X):

$$Y = f_{LINEAL}(X) = X\beta$$

Y = Variable explicada
X = Variables explicativas
β = Parámetros de regresión

2. METODOLOGÍA DE ANÁLISIS

Para analizar regresiones lineales múltiples se recomienda aplicar los siguientes seis pasos:

- **Formulación de la hipótesis estadística**: La hipótesis estadística es un enunciado lógico que se quiere aceptar o rechazar. Esta hipótesis debe ser los más consistente posible con la realidad que representa.

- **Formulación de la regresión estadística**: Toda regresión lineal múltiple se puede formular matemáticamente, usando una notación matricial en donde se define claramente, la variable explicada, las variables explicativas y los parámetros de regresión.

- **Estimación de los parámetros de regresión**: En este anexo, los parámetros de regresión lineal se estiman mediante el método de mínimos cuadrados, el cual minimiza la suma de cuadrados de los residuos de regresión. Más adelante se explica con detalle este método.

- **Pruebas estadísticas de bondad de ajuste**: Existen varias pruebas estadísticas que se usan para verificar la bondad de ajuste de una regresión lineal múltiple. Entre más pruebas se usen mejor será el ajuste.

- **Pronóstico de las variables explicativas**: Por lo general se definen y justifican tres escenarios de pronóstico (bajo, medio y alto) para cada variable explicativa. A cada escenario se le puede asignar una probabilidad de ocurrencia.

- **Pronóstico de la variable explicada**: El pronóstico de la variable explicada depende de los escenarios de proyección de las variables explicativas y del modelo de regresión lineal desarrollado, ajustado y validado.

3. PROGRAMAS ESTADÍSTICOS

En el mercado existen una amplia gama de programas estadísticos, con precios que varían entre $300 y $1300 USD. Estos precios incluyen una licencia corporativa comercial y profesional. Los programas estadísticos más populares son:

PROGRAMA	INTERNET	PRECIO
SPSS 12.0	www.spss.com	$1145 USD
PCGIVE 10.3	www.pcgive.com	$850 USD
SYSTAT 10.2	www.systat.com	$1300 USD
SHAZAM 9.0	econometrics.com	$490 USD
STATA 8.0	www.stata.com	$1300 USD
LIMDEP 8.0	www.limdep.com	$895 USD
XLSTAT 6.1 (add-in)	www.xlstat.com	$295 USD
STATISTICA 6.0	www.statsoftinc.com	$959 USD
SAS/JMP 5.1	www.jmp.com	$995 USD
RATS 5.0	www.estima.com	$500 USD
TSP 4.5	www.tspintl.com	$900 USD
EVIEWS 4.1	www.eviews.com	$1195 USD
REALSTAT 4.3	www.realstat.com	$500 USD

La selección del programa estadístico depende de las necesidades y presupuestos del usuario. Si las necesidades se limitan a un análisis de regresión lineal múltiple, un complemento (add-in) como XLSTAT es suficiente. Si se desea análisis más complejos y especializados, EVIEWS, LIMDEP o SHAZAM son la solución. Para análisis robustos y complejos, STATISTICA, STATA o SPSS son los indicados.

4. FORMULACIÓN DE LA HIPÓTESIS ESTADÍSTICA

Una hipótesis estadística es un enunciado lógico que se quiere aceptar o rechazar. Esta hipótesis debe ser consistente con la realidad que representa. Por ejemplo, no es lógica correlacionar las ventas de carros con el consumo de alcohol. En el sector eléctrico, se usan las dos siguientes hipótesis para correlacionar la demanda de energía eléctrica con parámetros socioeconómicos:

- **Primera hipótesis ilustrativa**: La demanda de energía eléctrica en el sector residencial depende linealmente de la población y del ingreso promedio disponible.
- **Segunda hipótesis ilustrativa**: La demanda de energía eléctrica en el sector industrial depende linealmente del producto interno bruto y del precio promedio de la energía eléctrica.

5. FORMULACIÓN DE LA REGRESIÓN ESTADÍSTICA

La formulación matemática de una regresión lineal múltiple se representa con la siguiente notación matricial:

$$Y = X\beta \qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1..x_{1j}..x_{1m} \\ \vdots \quad \vdots \quad \vdots \\ 1..x_{ij}..x_{im} \\ \vdots \quad \vdots \quad \vdots \\ 1..x_{nj}..x_{nm} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_m \end{bmatrix}$$

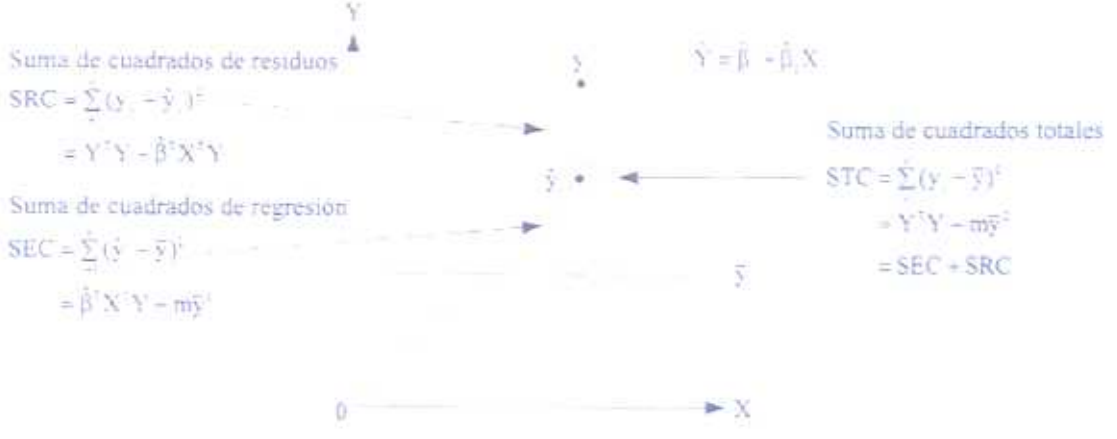
Y = Variable explicada
X = Variables explicativas
β = Parámetros de regresión
n = Número de observaciones
m = Número de variables explicativas
i = Subindice de la i-ésima observación
j = Subindice de la j-ésima variable explicativa

Un modelo de regresión lineal simple involucra solamente una variable explicativa y dos parametros de regresión. Por tal motivo, su formulación matricial se reduce a una ecuación escalar bastante conocida y sencilla de resolver ($Y = \beta_2 + \beta_1 X$). Cuando se involucra más de una variable explicativa, la formulación matricial simplifica las ecuaciones matemáticas. Sin embargo, se requiere un dominio de las operaciones matriciales, tales como matriz identidad, inversa y transpuesta de una matriz, y suma y multiplicación de matrices.

3. ESTIMACIÓN DE PARÁMETROS DE REGRESIÓN

Existen varios métodos para resolver regresiones estadisticas: minimos cuadrados ordinarios (OLS); minimos cuadrados ponderados (WLS); método generalizado de momentos (GMM); y heterocedasticidad autoregresiva condicional (ARCH). El método de minimos cuadrados ordinarios es el más indicado para regresiones lineales múltiples. Más adelante se pueden incorporar métodos más avanzados.

Para entender el método de minimos cuadrados, es necesario definir tres estadisticos que miden la dispersión de las estimaciones con respecto a las observaciones: suma de cuadrados de residuos (SRC); suma de cuadrados de regresión (SEC); y suma de cuadrados totales (STC).



Y = Variable explicada
Y^T = Matriz transpuesta de Y
X = Variables explicativas
β = Parámetros de regresión
y_i = i-ésima observación
$\hat{y}_i$ = i-ésima estimación de regresión
$\bar{y}$ = Promedio de las observaciones
n = Número de observaciones
m = Número de variables explicativas

- **Suma de cuadrados de regresión (SEC)**: Este estadístico mide la dispersión entre las estimaciones de regresión ($\hat{y}_i$) y el promedio de las observaciones ($\bar{y}$):

$$SEC = \hat{\beta}^T X^T Y - m\bar{y}^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- **Suma de cuadrados de residuos (SRC)**: Este estadístico mide la dispersión entre las observaciones (y_i) y las estimaciones de regresión ($\hat{y}_i$):

$$SRC = Y^T Y - \hat{\beta}^T X^T Y = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- **Suma de cuadrados totales (STC)**: Este estadístico mide la dispersión entre las observaciones (y_i) y el promedio de estas observaciones ($\bar{y}$):

$$STC = SEC + SRC = Y^T Y - m\bar{y}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

En una regresión estadística, la dispersión de las observaciones con respecto al promedio se puede dividir en dos componentes: la dispersión de las observaciones con respecto a las estimaciones; y las dispersiones de las estimaciones con respecto al promedio. El método de mínimos cuadrados ordinarios (OLS) minimiza la suma de cuadrados de residuos (SRC), es decir la dispersión de las estimaciones con respecto a las observaciones:

$$\min SRC = Y^T Y - \hat{\beta}^T X^T Y = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Con este método se busca que las estimaciones incluyan todos los factores no aleatorios que explican el comportamiento de las observaciones. De esta forma, los residuos ($e_i = y_i - \hat{y}_i$) depende exclusivamente de factores aleatorios. Para que una regresión lineal múltiple este bien ajustada, se requiere entonces que los residuos tengan una distribución Normal con un valor esperado de cero y con una mínima desviación:

$$e_i = y_i - \hat{y}_i \sim N(0,\sigma^2) \qquad \sigma^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-m}$$

$e_i = y_i - \hat{y}_i$: Residuos de regresión

$N(0;\sigma^2)$: Distribución Normal con media 0 y desviación σ

σ^2: Varianza muestral de los residuos de regresión

n = Número de observaciones

m = Número de variables explicativas

Aplicando el método de mínimos cuadrados ordinarios, se pueden estimar los parámetros de regresión, usando la siguiente fórmula matricial:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \sim N(\beta;(X^T X)^{-1}\sigma^2)$$

$N(\beta;(X^T X)^{-1}\sigma^2)$: Distribución Normal con media β y desviación $\sqrt{(X^T X)^{-1}}\,\sigma$

$E(\hat{\beta}) = \beta$: Valor esperado de β

$V(\hat{\beta}) = (X^T X)^{-1}\sigma^2$: Varianza muestral de β

4. PRUEBAS ESTADÍSTICAS DE BONDAD DE AJUSTE

Existen varias pruebas estadísticas para verificar la bondad de ajuste de una regresión lineal múltiple. En este anexo se enumeran las pruebas usadas con mayor frecuencia, para mayor información se recomienda consultar el manual del usuario de EVIEWS 4.1:

- Correlación de variables
- Aurocorrelación de observaciones
- Distribución normal de residuos
- Prueba estadística colectiva
- Prueba estadística individual

4.1. Correlación de variables

La correlación entre la variable explicada y las variables explicativas se mide con el coeficiente de correlación (R^2), es decir con el cociente entre la suma de cuadrados de regresión y la suma de cuadrados totales. Entre más cercano a uno, menor es la suma de cuadrados de residuos (SRC) y mejor es el ajuste de la regresión:

$$R^2 = \frac{SEC}{STC} = \frac{SEC}{SEC + SRC} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Para garantizar un mejor ajuste, se calcula el coeficiente de correlación ajustado (R^2_{aj}), el cual es menor pero más exacto pues tiene en cuenta el número de variables explicativas y el número de observaciones disponibles:

$$R^2_{aj} = 1 - \frac{\frac{SRC}{n-m}}{\frac{STC}{n-1}} = \frac{\frac{\sum_{i=1}^{n} e_i^2}{n-m}}{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = 1 - (1 - R^2)\frac{n-1}{n-m}$$

4.2. Autocorrelación de observaciones

La autocorrelación de los residuos de observaciones sucesivas se mide con el coeficiente de autocorrelación Durbin-Watson (d). Si es cercano a dos no hay autocorrelación, si es cercano a cero o a cuatro hay autocorrelación. Sin embargo, hay dos zonas intermedias en donde no es posible llegar a una conclusión. Puede existir una autocorrelación de orden superior, es decir entre observaciones desfasadas dos o más periodos. En este caso, conviene usar un método de regresión más avanzado, tal como la heterocedasticidad autoregresiva condicional (ARCH).

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Autocorrelación positiva	Autocorrelación indeterminada	Sin autocorrelación	Autocorrelación indeterminada	Autocorrelación Negativa
0	d_l d_U	2	$4 - d_U$	$4 - d_l$ 4

En caso de haber correlación positiva o negativa, es necesario incorporar las observaciones de la variable explicada como una variable explicativa, desfasa en uno o más periodos, dependiendo del grado de correlación.

4.3. Distribución normal de residuos

Como se mencionó anteriormente, el método de mínimos cuadrados ordinarios, se fundamenta en el principio de que los residuos de regresión tienen una distribución Normal, con un valor esperado de cero y con una mínima desviación. En otras palabras, la diferencia entra las estimaciones y las observaciones debe depender exclusivamente de factores aleatorios. Para verificar esta aleatoriedad de los residuos, se usa el estadístico Jarque Bera ($J\chi$) el cual mide el ajuste normal de los residuos de regresión:

$$J\chi = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

$S = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{\hat{\sigma}}\right)^3$: Skewness mide la asimetría de las observaciones con respecto al promedio

$K = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{\hat{\sigma}}\right)^4$: Kurtosis mide la amplitud y el ancho de la distribución de las observaciones

4.4. Prueba estadística colectiva

La prueba estadística colectiva, también denominada análisis de varianzas (ANOVA), verifica que los parámetros de regresión no sean simultáneamente nulos. En otras palabras, esta prueba verifica que las variables explicativas sean simultáneamente relevantes dentro de la regresión estadística. Para esta prueba se usa el estadístico F que mide el cociente entre la suma de cuadrados de regresión y la suma de cuadrados de residuos, ajustados por el número de variables explicativas y el número de observaciones disponibles. Si este estadístico es superior a un valor crítico determinado para un nivel de confianza dado $(1 - \alpha)$ entonces se rechaza la hipótesis nula. En otras palabras, si la probabilidad de que el estadístico calculado sea inferior a un valor crítico es muy pequeña (inferior o igual a α) entonces se rechaza la prueba nula.

$$H_0 : \beta_1 = .. = \beta_i = .. = \beta_m = 0 \text{ (prueba nula que se quiere rechazar)}$$
$$H_1 : \beta_1 \neq .. \neq \beta_i \neq .. = \beta_m \neq 0 \text{ (prueba alternativa)}$$

VARIABLE	GRADOS LIBERTAD	SUMA CUADRADOS	PROMEDIO CUADRADOS	ESTADÍSTICO F	PROBABILIDAD $P(F_\alpha > F)$
STC	$n-1$	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	STC $(n-1)$		
SEC	$m-1$	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	SEC $(m-1)$	$\frac{SEC/(m-1)}{SRC/(n-m)}$	$P(F_\alpha > F) \leq \alpha$
SRC	$n-m$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	SRC $(n-m)$		

$$F > F_\alpha \text{ o } P(F_\alpha > F) \leq \alpha \Leftrightarrow \text{se rechaza } H_0$$

Raramente se acepta la prueba nula de que todas las variables explicativas sean simultáneamente irrelevantes dentro del modelo. Por eso, esta prueba estadística no determinante por sí sola, debe estar complementada con otras pruebas.

4.4. Prueba estadística colectiva

La prueba estadística individual, también denominada pruebas de intervalos de confianza, verifica que el parámetro de regresión de una variable explicativa no sea nulo. En otras palabras, esta prueba verifica que cada variable explicativa sea relevante dentro de la regresión estadística.

Para esta prueba se usa el estadístico t-student que mide el cociente entre el valor esperado del parámetro de regresión y su desviación estándar. Si el valor absoluto de este estadístico es superior a un valor crítico determinado para un nivel de confianza dado $(1 - \alpha)$ entonces se rechaza la hipótesis nula. En otras palabras, si la probabilidad de que el valor absoluto del estadístico sea inferior a un valor crítico es muy pequeña (inferior o igual a α) entonces se rechaza la prueba nula.

$$H_0 : \beta_i = 0 \text{ (prueba nula que se quiere rechazar)}$$
$$H_1 : \beta_i \neq 0 \text{ (prueba alternativa)}$$

VARIABLE	GRADOS LIBERTAD	PROMEDIO $\hat{\beta}$	DEVIACIÓN $\sqrt{V(\hat{\beta})}$	ESTADISTICO t	PROBABILIDAD $P(\lvert t_{n-2}\rvert > t)$
X	$n-m$	$(X^TX)^{-1}X^TY$	$\text{diag}\sqrt{(X^TX)^{-1}\sigma^2}$	$\frac{\hat{\beta}_i}{\sqrt{V(\hat{\beta}_i)}}$	$P(\lvert t_{n-2}\rvert > t) \leq \alpha$

$$|t| > t_{\alpha,2} \text{ o } P(|t_{\alpha,2}| > t) \leq \alpha \Leftrightarrow \text{se rechaza } H_0$$

Con esta prueba estadística individual, se determina de igual forma un intervalo de confianza para cada parámetro de regresión. Es decir un intervalo entre el cual puede variar cada parámetros de regresión, con un nivel de confianza dado $(1 - \alpha)$:

$$\left[\hat{\beta} - t_{\alpha,2}V(\hat{\beta}); \hat{\beta} + t_{\alpha,2}V(\hat{\beta})\right]$$

5. PRONÓSTICO DE LAS VARIABLES EXPLICATIVAS

Antes de entrar en esta etapa de pronóstico, se debe haber seleccionado la regresión lineal múltiple que mejor correlacione la variable explicada con las variables explicativas. Para pronosticar la variable explicada se debe pronosticar primero las variables explicativas seleccionadas (x_j). Por lo general, se definen y justifican tres escenarios de pronóstico (bajo, medio y alto) para cada variable explicativa:

$$x_{i+1,j} = (1 + \lambda_{i+1,j})x_{i,j} \quad \lambda_{i+1,j} \begin{cases} \text{Escenario de crecimiento bajo} \\ \text{Escenario de crecimiento medio} \\ \text{Escenario de crecimiento alto} \end{cases}$$

$x_{i,j}$: Valor de la j-ésima variable explicativa para el periodo i

$x_{i+1,j}$: Pronóstico de la j-ésima variable explicativa para el periodo i+1

$\lambda_{i+1,j}$: Escenario de variación porcentual de la j-ésima variable entre los periodos i e i+1

A cada escenario se le puede asignar una probabilidad de ocurrencia, con el objeto de obtener el escenario más probable, calculado a partir de un promedio ponderado:

$$x_{i+1,j} = \frac{P_L x_{i+1,j,L} + P_M x_{i+1,j,M} + P_H x_{i+1,j,H}}{3}$$

P_L / P_M / P_H: Probabilidad de ocurrencia del escenario bajo (L), medio (M) y alto (H)

$x_{i+1,j,L}$ / $x_{i+1,j,M}$ / $x_{i+1,j,H}$: Pronóstico de la j-ésima variable explicativa para el periodo i+1 y para el escenario bajo (L), medio (M) y alto (H).

$x_{i+1,j}$: Pronóstico más probable de la j-ésima variable explicativa para el periodo i+1

6. PRONÓSTICO DE LA VARIABLE EXPLICADA

Una vez seleccionado y validado los escenarios de pronóstico de las variables explicativas seleccionadas y la regresión lineal múltiple de mejor ajuste, se procede finalmente a pronosticar la variable explicada. De esta forma se concluye el pronóstico estadístico.