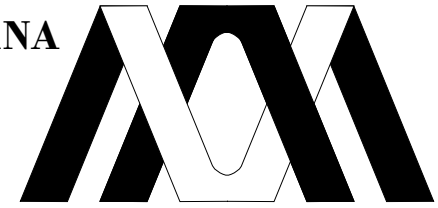
Universidad Autónoma Metropolitana
Casa abierta al tiempo
Iztapalapa



Análisis del Cambio Estructural en los Modelos Lineales

Trabajo terminal de la Maestria en Ciencias

Matemáticas Aplicadas e industriales

que presenta:

María Guadalupe García Salazar

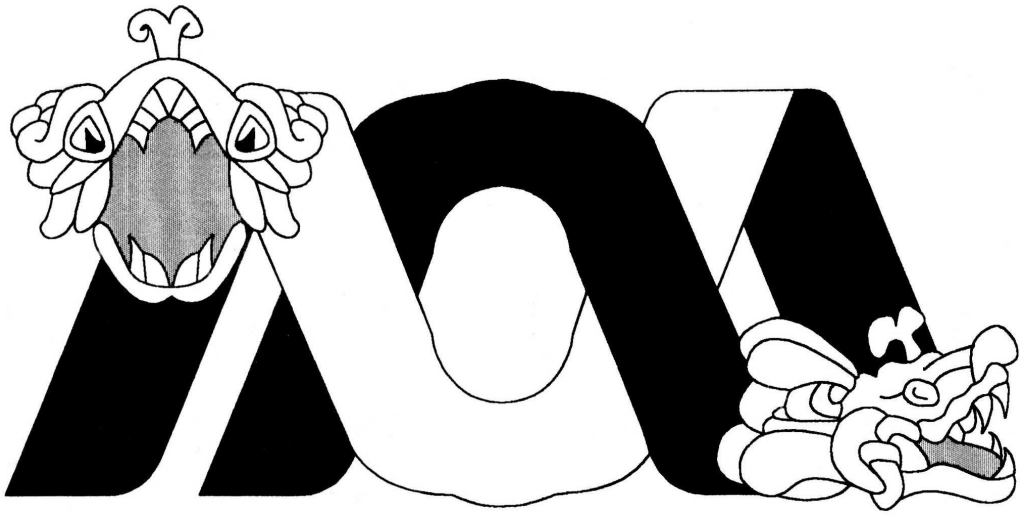
Asesora:

Blanca Rosa Pérez Salvador

Sinodales:

Lucía Atzimba Ruíz Galindo
Alberto Castillo Morales
Juan González Hernández

13 de Julio de 2009



IN KALLI IXKAHUIKOPA

A mis padres y hermanas
por todo el amor que me han dado.
A mi asesora y sinodales por su paciencia.
A mis amigos por su mágica y sublime amistad.

Índice General

Introducción

¿Te has preguntado si existe una cantidad en el ingreso personal de un individuo que le permite cambiar la tendencia en lo que ahorra? O ¿A qué edad, en qué nivel de la presión sanguínea, o cuántos cigarros fumados por un individuo, producen un cambio en la probabilidad de sufrir un infarto cardiaco? Para responder a estas preguntas suele utilizarse un modelo de regresión lineal que relaciona una variable respuesta Y con una o más variables explicativas $X_1, X_2, \ldots, X_p$, considerando que los parámetros del modelo son diferentes antes y después de un punto m.

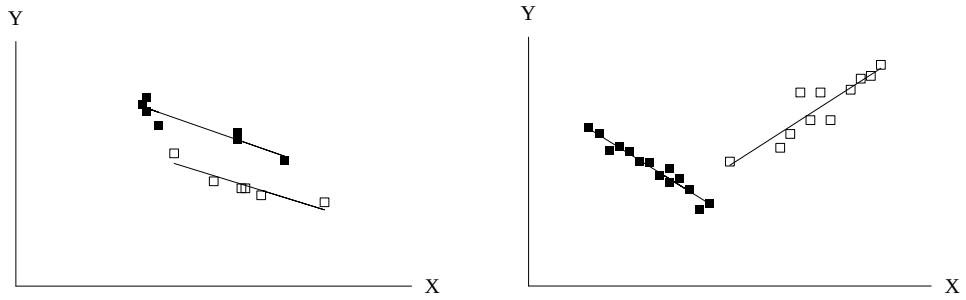
Suponga que se obtiene una muestra de n observaciones y se sospecha que las primeras m observaciones de la muestra siguen un modelo de regresión, y que las $n - m$ observaciones restantes siguen un modelo de regresión diferente, es decir, que existe un cambio estructural en el periodo de observación, entendiéndose como cambio estructural a aquella alteración o modificación de los parámetros en un modelo de regresión.

El modelo de regresión lineal asociado a la existencia de cambio estructural puede ser representado como

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_{1i} + \varepsilon_i & \text{si} \quad i \leq m \\ \beta_0^* + \beta_1^* X_{1i} + \varepsilon_i & \text{si} \quad i > m \end{cases}$$

con $\beta_j \neq \beta_j^*$ al menos para una j, $0 \leq j \leq 1$. Al número m se le conoce como el punto donde se da el cambio. El problema del cambio estructural en modelos de regresión lineales, viene a modelar problemas en los cuales se presenta un cambio en la tendencia de Y dentro de la región de observación.

Por ejemplo, en los gráficos (a) y (b) se puede observar la existencia del cambio estructural, en donde se considera que los puntos en negro son las primeras m observaciones y los puntos en blanco representan a las $n - m$ observaciones restantes.



(a) (b)

Cabe señalar que aunque en ambos conjuntos de datos existe el cambio estructural, en (a) la variable X no está correlacionada con el tiempo; mientras que en (b) si lo esta.

El objetivo de este trabajo es determinar para un modelo de regresión lineal con p variables explicativas si en el periodo de observación, n, hay un cambio estructural, para ello se propone una prueba de hipótesis en la cual se formula la existencia o no del cambio estructural.

Este trabajo fue motivado por el articulo de Muggeo(2003) quien abordó el problema de estimación del punto de cambio estructural a partir de un modelo no lineal y propuso estimar el punto de cambio aplicando una simple técnica de linealización. Él considera que la técnica de verosimilitud es inaplicable para estimar los parámetros del modelo por el hecho de que el logaritmo de la función de verosimilitud no es diferenciable en el punto de cambio. Con base a esta idea sobre la verosimilitud, se propone un estadístico de prueba para resolver los problemas de estimación y pruebas de hipótesis sobre el punto de cambio estructural, aplicando la razón de verosimilitud para encontrar la región crítica y el método de máxima verosimilitud para la estimación del punto de cambio.

A partir del estadístico de prueba encontrado a través de aplicar la razón de verosimilitud, se especifica su distribución de probabilidad y a continuación se procede a calcular la región crítica de la prueba de hipótesis propuesta, pero dado que el cálculo de las probabilidades de la distribución encontrada no se pueden obtener analíticamente, se recurre a una integración estocástica y/o a una integración numérica. Estas dos formas de integración son aplicadas a 6 ejemplos, de los cuales en dos de ellos se ha probado la existencia del cambio estructural por medio de otros estadísticos y después de aplicar el estadístico propuesto aquí a dichos ejemplos, los resultados que se obtienen son muy similares. Por ultimo, se propone un estimador del punto de cambio estructural, $\hat{m}$, y para los ejemplos en donde se encuentre evidencia de cambio estructural, se estima el punto de cambio y se analiza su sesgo y varianza, encontrando para cada ejemplo que el estimador presenta un pequeño sesgo.

Uno de los métodos que se ha utilizado para probar la ocurrencia de un cambio estructural es el de la prueba Chow (Chow, 1960). La característica principal de la prueba Chow es que el posible momento en que ocurre el cambio está bien determinado, es decir, se sospecha el punto m para el cual los datos muestrales antes de m siguen un modelo de regresión lineal diferente al modelo de regresión lineal que siguen los datos después de m. La prueba Chow es inaplicable si se desconoce el punto donde pudo producirse el cambio. Una posible solución es utilizar la prueba CUSUM

(Greene, 1999:309) la cual no necesita del conocimiento del momento del cambio estructural, esta prueba esta basada en la suma acumulada de residuos recursivos y la determinación del rechazo o no rechazo de cambio estructural se lleva a cabo por medio de un análisis del comportamiento de éstos.

Otros estudios sobre la prueba de hipótesis para determinar la existencia del cambio estructural han sido efectuados por Beckman y Cook (1979), Horvath y Shao (1993), Antoch y Hušcová (2001), quienes formularon como hipótesis nula el que no existe cambio en la región o periodo de observación y encontraron la región crítica asintótica. Por otro lado, la estimación del punto de cambio estructural fue abordado por Muggeo (2003), quien consideró que el método de máxima verosimilitud era inaplicable en este problema por que el logarirmo de la función de verosimilitud no es diferenciable en el punto de cambio.

La importancia de este trabajo es que se propopone una metodología no asintótica para determinar la función de la distribución de probabilidad exacta del estadístico de prueba propuesto y estimar el punto de cambio, la cual contempla el planteamiento de una hipótesis estadística, la propuesta de un estadístico de prueba y la determinación de su distribución; de manera que si mediante este procedimiento se rechaza la hipótesis nula de no cambio estructural, se procede a estimar el punto de cambio. De esta manera, el trabajo se distingue de los que se han mencionado por la propuesta que se realiza y se diferencia de ellos, por su carácter no asintótico.

El trabajo se encuentra conformado por 4 capítulos. En el capítulo 1 se introducen algunos conceptos básicos de la estimación de parámetros y de la prueba estadística de hipótesis, en este sentido se estudia el método de máxima verosimilitud, se revisan las propiedades de los estimadores y se estudia el lema de Neyman Pearson, entre otras cosas. En el capítulo 2 se presentan algunos resultados de algebra matricial y de la teoría de la distribución Normal, los cuales se aplican en los capítulos 3 y 4, en el capítulo 3 se habla sobre el modelo de regresión lineal en varias variables y el capítulo 4 esta dedicado al estudio del punto de cambio estructural, principalmente se busca determinar si en un modelo de regresión lineal en varias variables el cambio ha ocurrido o no y cuándo. Por último se presentan las conclusiones del trabajo.

Capítulo 1

Inferencia estadística

Frecuentemente se esta interesado en conocer un valor relacionado con una población, pero en la mayoría de los casos resulta difícil o imposible examinarla en su totalidad, por lo cual, para lograr el objetivo suele estudiarse sólo una pequeña parte de la misma y a partir de los resultados encontrados deducir el valor del parámetro de interés para esta. La inferencia estadística es la herramienta principal que se utiliza para dar solución a este tipo de problemas.

La aplicación de la estadística se puede realizar en dos formas diferentes y complementarias; siendo estas:

La estadística descriptiva que se enfoca a analizar y a presentar, visualmente, la información obtenida de la muestra extraída de una población y,

La estadística inferencial que se emplea para hacer afirmaciones sobre la población tomando como base la información contenida en una muestra aleatoria de la misma. Estas afirmaciones se pueden abordar en dos sentidos, la estimación de parámetros y la prueba de hipótesis estadística.

En este capítulo se presentan algunos conceptos de la estimación de parámetros. Se introduce la notación y terminología básica de estimación puntual, y se finaliza con las ideas elementales de razón de verosimilitud en el contexto de prueba de hipótesis.

1.1 Estimación puntual y propiedades de los estimadores

Cuando no se conoce el valor de un parámetro asociado a la población, se puede tomar la información de una muestra de la misma para dar un valor o un conjunto de valores que puedan representar al parámetro en estudio; si la estimación consta de un sólo valor se dice que esta es una *estimación puntual* para el parámetro, pero si en cambio se proporcionan dos valores extremos o límite para el parámetro poblacional, permitiendo que el estimador pueda tomar una infinidad de valores, entonces se habla de una *estimación por intervalos*.

Estimación Puntual.
Resulta natural pensar que el promedio de los valores de una muestra de datos estima bien al promedio real de los datos de la población, entonces se puede decir que un estimador para la media poblacional es la media muestral, esto es

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Un *estimador* es una regla que indica como calcular el valor de una estimación con base en los datos de una muestra. En este caso, $\overline{X}$ es una regla que indica la suma de todos los datos de la muestra dividida entre su tamaño n. Cabe mencionar que, cada estimador representa una única regla para obtener una sola estimación del parámetro desconocido. Algunas de las propiedades matemáticas de los estimadores puntuales son:

i) **Estimador Insesgado.**
Se dice que un estimador $\hat{\theta}$ es un estimador insesgado del parámetro θ cuando el valor esperado del estimador es igual al valor del parámetro poblacional, es decir,

$$E(\hat{\theta}) = \theta.$$

Ejemplo 1.1.1. *Sea $Y_1, Y_2, ..., Y_n$ una muestra aleatoria con $E(Y_i) = \mu$. Demuestre que*

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

es una estimador insesgado de μ.

Solución.

$$E(\overline{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i)$$

dado que $E(Y_i) = \mu$,

$$E(\overline{Y}) = \frac{1}{n}\ n\mu = \mu$$

por lo tanto, se ve que $\overline{Y}$ es un estimador insesgado de μ.

ii) **Estimador Eficiente.**
Se dice que un estimador es eficiente si es insesgado, de varianza finita y si no existe un estimador insesgado con varianza menor que él, esto es, de todos los estimadores insesgados con varianza finita que haya para un parámetro, aquel que tenga mínima varianza será el estimador eficiente.

Partiendo del hecho de que el estimador es insesgado, un método que se puede aplicar para ver si el estimador es eficiente es utilizar la desigualdad de Cramér Rao; desigualdad que indica que si la varianza del estimador es igual a un estadístico dado, entonces el estimador es eficiente.

Teorema 1.1.1 (**Desigualdad de Cramér-Rao**). *Sea $Y_1, Y_2, ..., Y_n$ una muestra aleatoria de una función de densidad de probabilidad $f(y)$, con un parámetro de población θ cuyo valor se desconoce. Si $\hat{\theta}$ es un estimador insesgado de θ, entonces, en condiciones generales:*

$$V(\hat{\theta}) \geq I(\theta) \qquad donde \qquad I(\theta) = \frac{1}{nE\left(\frac{-\partial^2 \ln f(Y)}{\partial \theta^2}\right)}$$

esta expresión se conoce como la desigualdad de Cramér-Rao. *Cuando se cumple que $V(\hat{\theta}) = I(\theta)$ se dice que el estimador $\hat{\theta}$ es un estimador insesgado de varianza mínima de θ, es decir, eficiente (Mendenhall, 2002:420).*

Hay dos posibles usos para esta desigualdad; Primero, proporciona una cota mínima para la varianza de un estimador insesgado. Un estimador insesgado cuya varianza esté "muy cerca" de la cota mínima de Cramér-Rao es un buen estimador. Segundo, un estimador insesgado cuya varianza coincida con la cota de Cramér-Rao es un estimador de varianza mínima, por lo tanto, eficiente.

Otra forma de ver que tan eficiente es un estimador, es compararlo con otro y ver su eficiencia relativa.

Definición 1.1.1. *Para dos estimadores insesgados, $\hat{\theta}_1$ y $\hat{\theta}_2$, de un parámetro θ cuyas varianzas son $V(\hat{\theta}_1)$ y $V(\hat{\theta}_2)$ respectivamente, la eficiencia (ef) de $\hat{\theta}_1$ con respecto de $\hat{\theta}_2$, la cual se denota mediante $ef(\hat{\theta}_1, \hat{\theta}_2)$, está dada por:*

$$ef(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}.$$

Si la eficiencia es menor que 1, entonces $V(\hat{\theta}_2) < V(\hat{\theta}_1)$ y se dice que $\hat{\theta}_2$ es mejor estimador que $\hat{\theta}_1$ (Mendenhall, 2002:417).

En el siguiente ejemplo se revisa la propiedad de insesgamiento y se calcula la eficiencia relativa.

Ejemplo 1.1.2. *Sean $Y_1, Y_2, ..., Y_n$ una muestra aleatoria de una distribución uniforme en el intervalo $(0, \theta)$. Dos estimadores insesgados para θ son*

$$\hat{\theta}_1 = 2\overline{Y} \quad y \quad \hat{\theta}_2 = \left(\frac{n+1}{n}\right) Y$$

donde $Y = \max\{Y_1, Y_2, ..., Y_n\}$. Encuentre la eficiencia de $\hat{\theta}_1$ respecto a $\hat{\theta}_2$ (Mendenhall, 2002:418-419).

Solución. *Cada Y_i tiene una distribución uniforme en el intervalo $(0, \theta)$,*
$\mu = E(Y_i) = \frac{\theta}{2}$ y $\sigma^2 = V(Y_i) = \frac{\theta^2}{12}$. Por lo tanto,

$$E(\hat{\theta}_1) = E(2\overline{Y}) = 2E(\overline{Y}) = 2\mu = 2\frac{\theta}{2} = \theta$$

y $\hat{\theta}_1$ *es insesgado. Además,*

$$V(\hat{\theta}_1) = V(2\overline{Y}) = 4V(\overline{Y}) = 4\left(\frac{V(Y_i)}{n}\right) = \left(\frac{4}{n}\right)\left(\frac{\theta^2}{12}\right) = \frac{\theta^2}{3n}$$

para determinar la media y la varianza de $\hat{\theta}_2$ *se debe tomar en cuenta que la función de densidad de* Y *esta determinada por*

$$t(y) = \begin{cases} n\left(\frac{y}{\theta}\right)^{n-1}\left(\frac{1}{\theta}\right), & si\ 0 \leq y \leq \theta \\ 0, & en\ cualquier\ otro\ punto \end{cases}$$

ya que las variables Y_i *son independientes y* $P(Y_i \leq y) = F(y)$ *para* $i = 1, 2, \ldots, n$*, la función de distribución de* Y *está dada por*

$$T(y) = [F(y)]^n = \left(\frac{y}{\theta}\right)^n$$

de manera que

$$\begin{aligned} E(Y) &= \int_0^\theta yt(y)dy = \int_0^\theta yn\left(\frac{y}{\theta}\right)^{n-1}\left(\frac{1}{\theta}\right)dy \\ &= \frac{n}{\theta^n}\int_0^\theta y^n dy = \frac{n}{\theta^n}\left[\frac{y^{n+1}}{n+1}\right]_0^\theta = \frac{n}{\theta^n}\left[\frac{\theta^{n+1}}{n+1}\right] = \left(\frac{n}{n+1}\right)\theta \end{aligned}$$

y por lo tanto $E(\hat{\theta}_2) = E\left(\frac{n+1}{n}Y\right) = \theta$*; es decir,* $\hat{\theta}_2$ *es un estimador insesgado para* θ*. Como*

$$\begin{aligned} E(Y^2) &= \int_0^\theta y^2 t(y)dy = \int_0^\theta y^2 n\left(\frac{y}{\theta}\right)^{n-1}\left(\frac{1}{\theta}\right)dy \\ &= \frac{n}{\theta^n}\int_0^\theta y^{n+1}dy = \left(\frac{n}{n+2}\right)\theta^2 \end{aligned}$$

se obtiene

$$V(Y) = E(Y^2) - E(Y)^2 = \left(\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2\right)\theta^2$$

y

$$V(\hat{\theta}_2) = V\left(\frac{n+1}{n}Y\right) = \left(\frac{n+1}{n}\right)^2 V(Y) = \left[\frac{(n+1)^2}{n(n+2)} - 1\right]\theta^2 = \frac{\theta^2}{n(n+2)}$$

por lo tanto, la eficiencia de $\hat{\theta}_1$ *con respecto a* $\hat{\theta}_2$ *esta dada por*

$$ef(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)} = \frac{\frac{\theta^2}{n(n+2)}}{\frac{\theta^2}{3n}} = \frac{3}{n+2}$$

nótese que la eficiencia es menor que 1 si $n > 1$*, es decir,* $\hat{\theta}_2$ *tiene una varianza menor que* $\hat{\theta}_1$ *y, por ende se debe preferir* $\hat{\theta}_2$ *sobre* $\hat{\theta}_1$ *como estimador de* θ*.*

iii) **Estimador Consistente.**
Si se denota $\hat{\theta}_n$ como el estimador calculado en base a una muestra de tamaño n, entonces se dice que $\hat{\theta}_n$ es un estimador *consistente* de θ si, para cualquier número positivo ε,

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| \leq \varepsilon\right) = 1$$

o

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = 0.$$

Un resultado sobre estimadores (Mendenhall, 2002:421) muestra que un estimador insesgado $\hat{\theta}_n$ de θ es un *estimador consistente* de θ si

$$\lim_{n\to\infty} V\left(\hat{\theta}_n\right) = 0.$$

Un estimador consistente se muestra en el siguiente ejemplo.

Ejemplo 1.1.3. *Sea $Y_1, Y_2, ..., Y_n$ una muestra aleatoria de una distribución Normal con media μ y varianza $\sigma^2 < \infty$, $N(\mu, \sigma^2)$. Demuestre que $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ es un* estimador consistente *de μ (Mendenhall, 2002:422).*

Solución. *Dado que $E(\overline{Y}) = \mu$ y $V(\overline{Y}) = \frac{\sigma^2}{n}$ entonces $\overline{Y}$ es insesgado para μ y $V(\overline{Y}) \to 0$ conforme $n \to \infty$, por lo tanto $\overline{Y}$ es un estimador consistente de μ.*

iv) **Estimador Suficiente.**
Cuando se toma una muestra aleatoria $Y_1, Y_2, \ldots, Y_n$ de una distribución de probabilidad con un parámetro θ desconocido, entonces un estadístico U que está en función de la muestra se dice que es *suficiente* para θ si la distribución condicional de la muestra dado U no depende de θ.

Definición 1.1.2. *Sea $Y_1, Y_2, \ldots, Y_n$ una muestra aleatoria de una distribución con un parámetro θ cuyo valor se desconoce. Se dice que el estadístico $U = u(y_1, y_2, \ldots, y_n)$ es* suficiente *para θ si la distribución condicional de $Y_1, Y_2, \ldots, Y_n$, dado U, no depende de θ (Mendenhall, 2002:430).*

Esto se trata de aclarar en el siguiente ejemplo:

Ejemplo 1.1.4. *Considere los resultados de n ensayos de un experimento Bernoulli, Y_1, Y_2, $Y_3, \ldots, Y_n$, donde*

$$Y_i = \begin{cases} 1, & \textit{si el i-ésimo ensayo es un éxito,} \\ 0, & \textit{si el i-ésimo ensayo es un fracaso.} \end{cases}$$

Si p es la probabilidad de éxito en cualquier ensayo, entonces, para $i = 1, 2, .., n$

$$Y_i = \begin{cases} 1, & \textit{con probabilidad p,} \\ 0, & \textit{con probabilidad q=1-p.} \end{cases}$$

Suponga que se tiene un estadístico $X = \sum_{i=1}^{n} Y_i$, *el número de éxitos en los* n *ensayos. Considere la distribución condicional de* $Y_1, Y_2, ..., Y_n$ *dado* X

$$P(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n \mid X = x) = \frac{P(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n, X = x)}{P(X = x)}.$$

El numerador del miembro derecho de esta expresión es cero si $\sum_{i=1}^{n} y_i \neq x$, *y es la probabilidad de una sucesión de ceros y unos, con la cantidad total* x *de unos y* $(n-x)$ *ceros si* $\sum_{i=1}^{n} y_i = x$. *De la misma forma, el denominador es la probabilidad de exactamente* x *exitos en* n *ensayos. Por lo tanto, si* $x = 0, 1, 2, ..., n$

$$P(Y_1 = y_1, ..., Y_n = y_n \mid X = x) = \begin{cases} \frac{p^x(1-p)^{n-x}}{\binom{n}{x} p^x(1-p)^{n-x}} = \frac{1}{\binom{n}{x}}, & si \sum_{i=1}^{n} y_i = x, \\ 0, & en\ otro\ caso, \end{cases}$$

esta distribución condicional $Y_1, Y_2, \ldots, Y_n$ *dado* X *no depende de* p. *Esto quiere decir que una vez que se conoce el valor de* X *ninguna otra función de* $Y_1, Y_2, \ldots, Y_n$ *proporcionará más información sobre el posible valor de* p. *En este sentido* X *contiene toda la información respecto a* p *y por lo tanto se dice que el estadístico* X *es suficiente para* p *(Mendenhall, 2002:430).*

La definición de suficiencia nos indica cuando un estimador es suficiente, pero no indica como calcularlo, una forma de hacerlo es mediante la *función de verosimilitud*:

Definición 1.1.3. *Sean* y_1, y_2,...,y_n *los valores de una muestra de* p *variables aleatorias,* $Y_1, Y_2, ..., Y_n$, *cuya distribución depende del parámetro* θ. *La* **función de verosimilitud** *de la muestra se define como el valor de la distribución de probabilidad conjunta de las* n *variables aleatorias, es decir*

$$\begin{aligned} L(y_1, y_2, ..., y_n; \theta) &= f(y_1, y_2, ..., y_n; \theta) \\ &= f_1(y_1, \theta) f_2(y_2; \theta) ... f_n(y_n; \theta) = \prod_{i=1}^{n} f_i(y_i; \theta). \end{aligned}$$

Teorema 1.1.2. *Si* U *es un estadístico basado en la muestra aleatoria* $Y_1, \ldots, Y_p$, *entonces* $U = u(y_1, y_2, \ldots, y_n)$ *es un* **estadístico suficiente** *para la estimación de un parámetro* θ *si y sólo si la verosimilitud* $L(y_1, y_2, ..., y_n; \theta)$ *se puede factorizar en dos funciones no negativas*

$$L(y_1, y_2, ..., y_n; \theta) = g(u; \theta) h(y_1, y_2, ..., y_n)$$

donde $g(u, \theta)$ *es una función de* u *y* θ, *mientras que* $h(y_1, y_2, ..., y_n)$ *no es una función de* θ *(Mendenhall, 2002:431).*

Ejemplo 1.1.5. *Sean $Y_1, Y_2, ..., Y_n$ una muestra aleatoria en la que Y_i posee la función de densidad de probabilidad*

$$f(y \mid \theta) = \begin{cases} \frac{2y}{\theta} e^{-y^2/\theta}, & si \;\; y > 0 \\ 0, & en \; cualquier \; otro \; punto. \end{cases}$$

Demuestre que $U = \sum_{i=1}^{n} Y_i$ es un estadístico suficiente para la estimación de θ.

Solución. *La verosimilitud de la muestra es*

$$\begin{aligned} L(y_1, y_2, ..., y_n; \theta) &= f(y_1, y_2, ..., y_n; \theta) \\ &= f_1(y_1, \theta) f_2(y_2; \theta) ... f_p(y_n; \theta) \\ &= (\frac{2y_1}{\theta} e^{-y_1^2/\theta})(\frac{2y_2}{\theta} e^{-y_2^2/\theta}) \cdots (\frac{2y_n}{\theta} e^{-y_n^2/\theta}) \\ &= \left(\frac{2}{\theta}\right)^n (y_1 y_2 \cdots y_n) e^{-\frac{1}{\theta}\sum_{i=1}^{n} y_i^2} \\ &= \underbrace{\left(\frac{2}{\theta}\right)^n e^{-\frac{1}{\theta}\sum_{i=1}^{n} y_i^2}}_{g(\sum_{i=1}^{n} Y_i, \theta)} \underbrace{(y_1 y_2 \cdots y_n)}_{h(y_1, y_2, ..., y_n)}. \end{aligned}$$

Por lo tanto, $U = \sum_{i=1}^{n} Y_i$ es el estadístico suficiente para la estimación de θ.

Cuando se busca un estimador se desea que cumpla con las propiedades de *insesgabilidad, varianza mínima, consistencia, suficiencia*, pues ello permite concluir que se tiene un buen estimador. Algunos de los métodos que generan estimadores con algunas de estas propiedades son:[1]

a) El método de máxima verosimilitud.

b) El método de mínimos cuadrados.

c) El método de momentos.

d) La estimación Bayesiana.

[1] En el presente trabajo sólo se estudia el método de máxima verosimilitud

1.2 Método de máxima verosimilitud

Este método consiste en encontrar para la función de verosimilitud, $L(x_1, x_2, ..., x_n; \theta)$, el valor del parámetro desconocido θ para el cual el valor de la función de verosimilitud sea el más alto. La máxima verosimilitud se puede obtener tomando la derivada de $L(x_1, x_2, ..., x_n; \theta)$ con respecto a θ e igualándola a cero. Como el logaritmo natural es una función monótonamente creciente, el $\ln L(x_1, x_2, ..., x_n; \theta)$ y $L(x_1, x_2, ..., x_n; \theta)$ alcanzan su máximo en el mismo punto. Por tanto, se utiliza el $\ln L(x_1, x_2, ..., x_n; \theta)$ en lugar de $L(x_1, x_2, ..., x_n; \theta)$ para derivar por ser más fácil de trabajar. De aquí en adelante se denota a la función de verosimilitud como $L(\theta)$ en vez de $L(x_1, x_2, ..., x_n; \theta)$. A continuación se presentan algunos ejemplos de este método.

Ejemplo 1.2.1 (Distribución de Poisson)**.** *Sea una muestra aleatoria $X_1, ..., X_n$ de una distribución Poisson, encontrar el estimador de máxima verosimilitud del parámetro μ, la media de X_i.*

Solución. *La función de distribución de probabilidad Poisson está dada por*

$$f(x) = \frac{\mu^x}{x!} e^{-\mu} \qquad (x = 0, 1, ...)$$

de manera que la función de verosimilitud queda de la siguiente forma

$$L(\mu) = \frac{\mu^{x_1}}{x_1!} e^{-\mu} \frac{\mu^{x_2}}{x_2!} e^{-\mu} ... \frac{\mu^{x_n}}{x_n!} e^{-\mu}$$

o bien

$$L(\mu) = \frac{1}{x_1!...x_n!} \mu^{x_1+...+x_n} e^{-n\mu} = \frac{1}{x_1!...x_n!} \mu^{n\overline{x}} e^{-n\mu}$$

aplicando logaritmo natural a la función

$$\ln L(\mu) = -\ln(x_1!...x_n!) + n\overline{x} \ln \mu - n\mu$$

derivando con respecto a μ e igualando a cero

$$\frac{\partial \ln L(\mu)}{\partial \mu} = \frac{n\overline{x}}{\mu} - n = 0$$

se obtiene el estimador

$$\hat{\mu} = \overline{x}.$$

Puesto que

$$\frac{\partial^2 L(\mu)}{\partial \mu^2} = -\frac{n\overline{x}}{\mu^2}$$

sutituyendo el valor de μ,

$$\frac{\partial^2 L(\mu)}{\partial \mu^2} = -\frac{n}{\overline{x}} < 0$$

por tanto, hay un máximo.

Ejemplo 1.2.2 (Distribución binomial)**.** *Suponga que en cierto experimento un evento A tiene probabilidad desconocida q y que en n repeticiones independientes del experimento, el evento A ocurre k veces. Estimar p utilizando máxima verosimilitud.*

Solución. *Sea X el número de veces que ocurre A en una sola repetición del experimento. Entonces X sólo puede tomar los valores 1 o 0 si A ocurre o no respectivamente. Por lo tanto, la función de probabilidades $f(x)$ toma los valores*

$$f(0) = P(X = 0) = 1 - q \quad y \quad f(1) = P(X = 1) = q$$

así, con la muestra de n ensayos $X_1, X_2, \ldots, X_n$ donde A ocurre k veces se tiene la función de verosimilitud

$$L(q) = q^k(1-q)^{n-k}$$

cuyo logaritmo natural es

$$\ln L(q) = k \ln q + (n-k)\ln(1-q)$$

por lo tanto, al derivar e igualar a cero

$$\frac{\partial \ln L(q)}{\partial q} = \frac{k}{q} - \frac{n-k}{1-q} = 0$$

al resolver para q se obtiene

$$\hat{q} = \frac{k}{n}.$$

Este estimador coincide con la frecuencia relativa de A. Donde

$$\frac{\partial^2 \ln L(q)}{\partial q^2} = \frac{k}{q^2} - \frac{n-k}{(1-q)^2}$$

sutituyendo el valor de q

$$\frac{\partial^2 \ln L(q)}{\partial q^2} = \frac{-n^3}{k(n-k)} < 0$$

por lo que se tiene un máximo.

Ejemplo 1.2.3 (Distribución Normal)**.** *Sea una variable aleatoria X la cual se distribuye Normal con media desconocida μ y varianza desconocida σ^2. Aplicar el método de máxima verosimilitud para encontrar los estimadores de μ y σ^2, respectivamente.*

Solución. *La densidad de X esta dada por*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2} \qquad (\sigma > 0)$$

por lo que la función de verosimilitud está dada por

$$L(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma^2}\right)^2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_2-\mu}{\sigma^2}\right)^2} \ldots \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_n-\mu}{\sigma^2}\right)^2}$$

haciendo operaciones

$$L(\mu,\sigma^2)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\left(\frac{1}{\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2}$$

al tomar logaritmos se tiene

$$\ln L(\mu,\sigma^2)=-n\ln\sqrt{2\pi}-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2$$

como se quieren estimar dos parámetros, se obtienen las derivadas parciales respecto a cada una de ellas y se igualan a cero

$$\frac{\partial \ln L(\mu,\sigma^2)}{\partial\mu}=\frac{1}{\sigma^2}\sum_{k=1}^{n}(x_k-\mu)=0$$

de donde se obtiene

$$\sum_{k=1}^{n}(x_k-\mu)=\sum_{k=1}^{n}x_k-n\mu=0$$

por lo tanto, el estimador para μ queda

$$\hat{\mu}=\frac{1}{n}\sum_{k=1}^{n}x_k=\overline{x}$$

y

$$\frac{\partial \ln L(\mu,\sigma^2)}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{k=1}^{n}(x_k-\mu)^2=0$$

ahora sustituyendo μ por $\hat{\mu}=\overline{x}$ resulta la condición

$$-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{k=1}^{n}(x_k-\overline{x})^2=0$$

de donde

$$\frac{n}{2\sigma^2}=\frac{1}{2\sigma^4}\sum_{k=1}^{n}(x_k-\overline{x})^2$$

y finalmente

$$\hat{\sigma^2}=\frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})^2$$

este último estimador no es insesgado, pero en base a él se puede obtener un estimador insesgado. Ahora sólo falta verificar que el valor estimado de μ y σ^2 sean máximos

$$\frac{\partial^2 \ln L(\mu,\sigma^2)}{\partial\mu^2} = -\frac{n}{\sigma^2}$$

$$\frac{\partial^2 \ln L(\mu,\sigma^2)}{\partial(\sigma^2)^2} = \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{k=1}^{n}(x_k-\mu)^2$$

sutituyendo el valor estimado de μ y σ^2 en la derivadas

$$\frac{\partial^2 \ln L(\mu,\sigma^2)}{\partial \mu^2} = -\frac{1}{\sum_{k=1}^{n}(x_k-\overline{x})^2} < 0$$

$$\frac{\partial^2 \ln L(\mu,\sigma^2)}{\partial (\sigma^2)^2} = -\frac{n^3}{\left(\sum_{k=1}^{n}(x_k-\overline{x})^2\right)^2} < 0$$

por tanto, hay un máximo en ambos casos.

Hasta el momento se ha hablado de estimadores puntuales, pero ¿cómo saber hasta que tanto un estimador puede alejarse del valor del parámetro real? Tema de estudio de la siguiente sección.

1.3 Estimación por intervalos

Un estimador de intervalo se define como una regla de cómo utilizar los datos de una muestra, $x_1, x_2, \ldots, x_n$, para calcular dos valores extremos θ_1 y θ_2 de un intervalo, en el que con un alto grado de confianza, el valor del parámetro estimado θ esta incluido en él. θ_1 y θ_2 se calculan a partir de los valores de la muestra, los cuales se pueden considerar como los valores de p variables aleatorias $X_1, X_2, \ldots, X_n$. Entonces, Θ_1 y Θ_2 son funciones de estas las variables aleatorias y, por lo tanto, también son variables aleatorias. Por consiguiente, se buscan Θ_1 y Θ_2 que produzcan un intervalo $[\Theta_1, \Theta_2]$ tal que

$$P(\Theta_1 \leq \theta \leq \Theta_2) = \gamma,$$

al número γ se le conoce como *nivel de confianza.*

Definición 1.3.1. *El intervalo con valores extremos θ_1 y θ_2 se llama intervalo de confianza para el parámetro desconocido θ y se representa por*

$$conf_\gamma \{\theta_1 \leq \theta \leq \theta_2\}$$

los valores θ_1 y θ_2 se llaman limites de confianza inferior y superior, respectivamente para θ (Kreysig, 1981:245).

Por ejemplo, si $\gamma = 95\%$ se dice que existe el 95% de confianza de que el intervalo que se de contenga el valor real del parámetro. Debido a que es común considerar que las observaciones de la muestra tienen una distribución Normal, se revisa el algoritmo para obtener un intervalo de confianza para la media μ de una distribución Normal con varianza conocida σ^2. Los pasos a seguir son:

1. Elegir el nivel de confianza γ (95%, 99%).
2. Determinar el valor crítico Z_c correspondiente (Z_c mide la cantidad de desviaciones estándar a ambos lados de la media que delimitan un porcentaje de área bajo la curva Normal (de gauss) igual a γ), ver la siguiente tabla como ejemplo.

γ	0.90	0.95	0.99	0.999
Z_c	1.645	1.960	2.576	3.291

3. Calcular la media $\overline{X}$ de la muestra $X_1, X_2, ..., X_n$.

4. Calcular
$$k = Z_c \frac{\sigma}{\sqrt{n}}$$
El intervalo de confianza para la media μ de la población es
$$conf\left\{\overline{X} - k \leq \mu \leq \overline{X} + k\right\}$$

Ejemplo 1.3.1. *Determinar un intervalo con 95% de confianza para la media de una distribución Normal con varianza $\sigma^2 = 9$, usando una muestra de $n = 100$ valores con media $\overline{x} = 5$.*

Solución. -

1. *Se requiere $\gamma = 0.95$*
2. *La c correspondiente es igual a* 1.960
3. *$\overline{X} = 5$ es un dato.*
4. *Se necesita* $k = 1.960\frac{(3)}{\sqrt{100}} = 0.588$, *en donde,*
$$\overline{X} - k = 4.412 \quad y \quad \overline{X} + k = 5.588.$$
El intervalo de confianza es
$$conf\left\{4.412 \leq \mu \leq 5.588\right\}.$$

1.4 Prueba de hipótesis

Las pruebas de hipótesis son un tipo de inferencia estadística y su objetivo es ayudar a tomar una decisión sobre la población de interés al examinar sólo una muestra de ella. Es decir, ayudan a decidir, a partir de los datos del muestreo, problemas tales como si un sistema de enseñanza es mejor que otro, si una moneda esta cargada o no esta cargada, si un nuevo medicamento es efectivo o no en la cura de alguna enfermedad, etc. Antes de continuar es preciso que se defina el término hipótesis.

Definición 1.4.1. *"Una hipótesis se define como una proposición acerca de una o más poblaciones"(Daniel, 1999:245).*

Definición 1.4.2. *"Una hipótesis estadística es una afirmación o conjetura acerca de la distribución de una o más variables aleatorias". Es decir, se refiere a una aseveración sobre los parámetros de las distribuciones de los datos, al comportamiento de las variables aleatorias. "Si una hipótesis estadística especifica completamente la distribución se conoce como hipótesis simple, si no, se conoce como hipótesis compuesta"(Freund, 2000:384).*

Las pruebas de hipótesis consideran dos hipótesis estadísticas, la primera de ellas es conocida como hipótesis nula o de no diferencia (H_0), puesto que es una afirmación que se considera cierta dentro de la población de interés; mientras que la segunda se llama hipótesis alternativa (H_1) y se refiere a cualquier hipótesis que difiera de una hipótesis dada (H_0). Por lo regular, lo que se busca es invalidar o rechazar a la hipótesis nula y la conclusión a la que se espera llegar se expresa por medio de la hipótesis nula.

Por ejemplo, cuando se quiere decidir si la vida promedio de una máquina es diferente de 5 años, la hipótesis nula, dado que es una proposición de no diferencia, es que la vida promedio de la máquina es de 5 años y la hipótesis alternativa, la conclusión a la que se espera llegar, es que la vida promedio de la maquina fuera diferente de 5 años (mayor o menor a 5 años). En términos formales se tiene:

$$H_0 : \theta = 5 \quad \text{versus} \quad H_1 : \theta \neq 5$$

donde el parámetro θ indica el valor del parámetro que se quiere explicar, en este caso, la vida promedio de la máquina. La hipótesis nula se refiere siempre a un valor especificado del parámetro de población, no a un estadístico de la muestra.

Las hipótesis estadísticas pueden clasificarse en simples o compuestas. En términos generales, se dice que se tiene una hipótesis simple cuando, aparte de especificar la forma de la distribución de los datos, el parámetro de suposición asume sólo un valor específico, por ejemplo, $H : \theta = 5$. Y se esta ante una hipótesis compuesta cuando la hipótesis no asigna un valor específico al parámetro de suposición, por ejemplo, $H : \theta \leq 5$, $H : \theta \geq 5$, $H : \theta \neq 5$.

La finalidad de la pruebas de hipótesis es decidir si la hipótesis nula (H_0) es rechazada o no. Si H_0 no se rechaza se dice que los datos muéstrales sobre los cuales se hace la prueba arrojan suficiente evidencia para no rechazar H_0. En caso contrario, si se rechaza H_0, se dice que los datos proporcionan evidencia suficiente para no rechazar alguna otra hipótesis (H_1). Para decidir el rechazo o no rechazo de la hipótesis nula se necesitan establecer criterios que nos ayuden a tomar tal decisión. Si se parte del supuesto de que ya se formularon las hipótesis estadísticas (H_0 y H_1) y de que se tiene alguna idea de cómo se distribuyen los datos de una muestra, lo primero es calcular un estadístico de prueba[2] en base a los datos de la muestra.

Suponga que se tiene una variable aleatoria $\overline{Y}$ con distribución Normal con media μ y varianza σ^2. Se dispone de una muestra aleatoria de tamaño n de la variable Y, $Y_1, Y_2, \ldots, Y_n$, se desea probar la hipótesis H_0: $\mu = \mu_0$, el estadístico de prueba a utilizar será:

$$z = \frac{\overline{y} - \mu}{\sigma/\sqrt{n}}.$$

[2]Un estadístico de prueba es una función que depende de los valores de la muestra.

Lo segundo es identificar los valores del estadístico de prueba calculado para los cuales no se rechaza H_0 o se rechaza H_0. Los valores posibles que toma el estadístico de prueba se pueden dividir en dos regiones, las cuales se conocen como región de aceptación y región de rechazo (también conocida como región crítica).

Definición 1.4.3. *Una región rechazo, C, es una región en el espacio n-dimensional que consta de todos los puntos de muestra para los que el estadístico de prueba nos lleva al rechazo de la hipótesis (Kreysig, 245).*

Si el valor calculado del estadístico de prueba cae en la región de rechazo se rechazará H_0 y si el valor del estadístico cae en la región de aceptación no se rechazará H_0. Lo que conduce a preguntar cómo se determina la región de rechazo C, pero antes de responder a esta pregunta se necesitan algunos conceptos más.
La identificación de las regiones de aceptación o rechazo de H_0 puede llevar a dos tipos de errores, conocidos con el nombre de error tipo I y error tipo II.

Definición 1.4.4. *Se comete un error tipo I cuando se rechaza H_0 siendo H_0 verdadera y la probabilidad de cometer este error*[3] *se denota por α.*
Se comete un error tipo II cuando se acepta H_0 siendo H_0 falsa y la probabilidad de cometer este error se denota por β (Freund, 2000:386).

Gráfica 1.1: Errores en la Prueba de Hipótesis

		Realidad	
		***H*₀ Verdadera**	***H*₁ Verdadera**
Decisión	**Aceptar *H*₀**	Decisión Correcta	Error Tipo II
	Rechazar *H*₀	Error Tipo I	Decisión Correcta

Fuente: Kreysig E. (1981). Introducción a la Estadística Matemática, p.225

Por ejemplo, suponga que se quiere probar la hipótesis nula de que la media de una población Normal con varianza igual a uno es μ_0 versus la hipótesis alternativa de que es μ_1, donde $\mu_0 < \mu_1$.

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu = \mu_1$$

el estadístico de prueba es $\overline{Y}$, la media muestral. Por tanto, la región de aceptación de H_0 vendrá dada por $\overline{Y} \leq k$, mientras que la región de rechazo será $\overline{Y} > k$. La probabilidad de cometer un

[3]La probabilidad de cometer el error tipo I también se conoce con el nombre de nivel de significancia de la prueba o como tamaño de la región de rechazo.

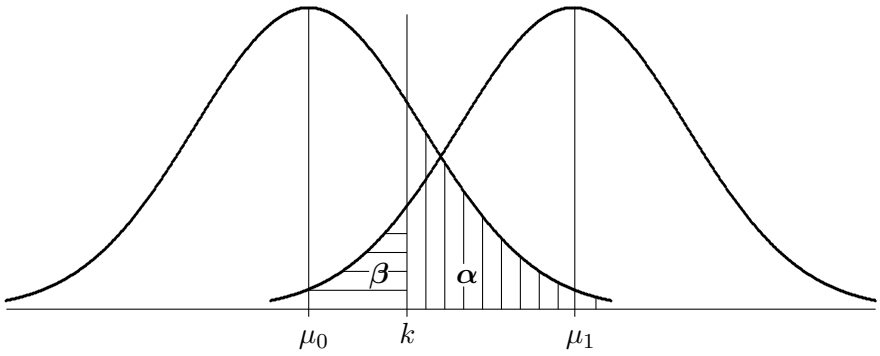
error tipo I será igual a $\alpha = P(\overline{Y} > k \mid \mu = \mu_0)$ y la probabilidad de cometer un error tipo II será $\beta = P(\overline{Y} \leq k \mid \mu = \mu_1)$.
Las probabilidades de error se ven en la gráfica siguiente.

Gráfica 1.2: Probabilidades de Error

$\boldsymbol{\alpha}$ Probabilidad de cometer el Error Tipo I

$\boldsymbol{\beta}$ Probabilidad de cometer el Error Tipo II

$\boldsymbol{\overline{Y} > k}$ Región de Rechazo(C)



Fuente: Kreysig E. (1981). Introducción a la Estadística Matemática, p.226

Definición 1.4.5. *Sea X el estadístico de prueba y C la región de rechazo de una prueba de hipótesis respecto al valor de un parámetro θ. Se conoce como* ***potencia de la prueba*** *a la probabilidad de que la prueba indique rechazar H_0 cuando el valor real del parámetro es θ. Es decir,*

$$potencia(\theta) = P(\text{Rechazar } \ H_0 \ \text{ cuando el valor del parámetro es } \ \theta) = P(X \in C \mid \theta).$$

Suponga que se quiere probar

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1.$$

La potencia de la prueba para H_0, $potencia(\theta_0)$, será igual a la probabilidad de rechazar H_0 cuando H_0 es verdadera. Es decir

$$potencia(\theta_0) = P(\text{rechazar } H_0 \text{ cuando } \theta = \theta_0) = \alpha.$$

Para el caso de la hipótesis alternativa $\theta = \theta_1$, la potencia de la prueba, $potencia(\theta_1)$, será igual a la probabilidad de rechazar H_0 cuando H_1 es verdadera. Es decir

$$potencia(\theta_1) = P(\text{rechazar } H_0 \text{ cuando } \theta = \theta_1).$$

Si se toma en cuenta que β es la probabilidad de no rechazar H_0 cuando H_0 es falsa, se tiene

$$\beta = P(\text{no } \ \text{rechazar } H_0 \text{ cuando } \theta = \theta_1),$$

la *potencia*(θ_1) se puede expresar como

$$potencia(\theta_1) = 1 - P(\text{no rechazar } H_0 \text{ cuando } \theta = \theta_1) = 1 - \beta.$$

En resumen

$$\begin{aligned} potencia(\theta_0) &= \alpha, \\ potencia(\theta_1) &= 1 - \beta. \end{aligned}$$

Ejemplo 1.4.1. *Un profesor aplica una prueba de 10 preguntas falso-verdadero y desea ensayar la hipótesis nula de que el estudiante acierta por casualidad. El estadístico de prueba es X, el número de preguntas contestadas correctamente. El profesor aceptará la hipótesis nula si $X < 7$, de otra forma la rechazará.*
a) Encuentre el error tipo I (α).
b) Halle la probabilidad de aceptar la hipótesis nula $p = 0.5$ cuando realmente es $p = 0.8$.

Solución. *La probabilidad de contestar correctamente x de las* 10 *preguntas esta dada por $\binom{n}{x} p^x (1-p)^{1-x}$, donde p es la probabilidad de que una pregunta sea contestada correctamente ($p = 0.5$). La región de rechazo esta dada por $x \geq 7$,*

$$\begin{aligned} \alpha &= P(x \geq 7 \mid p = 0.5) \\ &= \tbinom{10}{7} (0.5)^7 (0.5)^3 + \tbinom{10}{8} (0.5)^8 (0.5)^2 + \tbinom{10}{9} (0.5)^9 (0.5)^1 + \tbinom{10}{10} (0.5)^{10} (0.5)^0 \\ &= 0.1719 \end{aligned}$$

por tanto el error tipo I es igual a 0.1719.

La probabilidad de aceptar $p = 0.5$ cuando en verdad es $p = 0.8$ es:

$$\beta = P(x < 7 \mid p = 0.8) = 1 - P(x \geq 7 \mid p = 0.8) = 0.121.$$

La potencia de la prueba bajo esta hipótesis es:

$$1 - \beta = P(x \geq 7 \mid p = 0.8) = 0.879.$$

Por lo regular en las pruebas de hipótesis el investigador decide el nivel de significancia de la prueba. En otras palabras, decide el tamaño de la región crítica. En consecuencia, si se puede fijar el nivel de significancia (la probabilidad de cometer el error tipo I) y dado que nos interesa minimizar los errores, con la finalidad de tomar la mejor decisión, se buscará minimizar la probabilidad de cometer un error tipo II (β) o análogamente maximizar la potencia de la prueba (1-β).

Suponga que se quiere probar

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1$$

Para tomar la decisión de rechazar o no rechazar la H_0, se tiene que determinar la mejor región de rechazo C, la más potente, la cual será aquella donde la potencia de la prueba en $\theta = \theta_1$ se encuentre en un máximo.

La construcción de la prueba que maximiza la potencia(θ_1) se basa en la función de verosimilitud, en particular, se utiliza el Lema de Neyman Pearson, el cual se aplica sólo sobre hipótesis simples.

Teorema 1.4.1 (**Lema de Neyman Pearson**). [4] *"Sea $L(\theta)$ la verosimilitud de la muestra cuando el valor del parámetro es θ. Entonces para un valor dado de α, la prueba que maximiza la potencia en θ_1 tendrá un región de rechazo C determinada por:*

$$\frac{L(\theta_0)}{L(\theta_1)} < \lambda$$

El valor de λ se elige de tal manera que la prueba tenga el valor de α que se busca. Dicha prueba es la prueba de nivel α más potente" (Mendenhall, 2002:510).

En conclusión, la respuesta a la pregunta de cómo determinar la región de rechazo esta dada por la aplicación del lema anterior, el cual encuentra un estadístico de prueba que determina para que valores del mismo se rechaza la hipótesis nula, es decir, se determina la región de rechazo.

Ejemplo 1.4.2. *Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño 20, de una distribución Bernoulli. Dadas las hipótesis $H_0 : p = 0.3$ versus $H_1 : p = 0.8$, encuentre la mejor región crítica C por medio del Lema de Neyman Pearson.*

Solución. *Dado que X_i se distribuye Bernoulli sólo puede tomar el valor de cero o uno. La función de densidad esta dada por $f(x) = p^x(1-p)^{1-x}$. La probabilidad asociada a H_0 es $p_0 = 0.3$ y la asociada a H_1 es $p_1 = 0.8$. Aplicando el Lema:*

$$\frac{L(p_0)}{L(p_1)} = \frac{p_0^{\sum x_i}(1-p_0)^{n-\sum x_i}}{p_1^{\sum x_i}(1-p_1)^{n-\sum x_i}} < \lambda$$

para despejar los términos en función de la muestra se aplica la función logaritmo a ambos lados de la desigualdad,

$$\sum x_i \ln p_0 + (n - \sum x_i)\ln(1-p_1) - \sum x_i \ln p_1 - (n - \sum x_i)\ln(1-p_1) < \ln \lambda$$

agrupando términos y utilizando las propiedades de logaritmos,

$$\sum x_i(\ln p_0 - \ln(1-p_0) - \ln p_1 + \ln(1-p_1)) + n(\ln(1-p_0) - \ln(1-p_1)) < \ln \lambda$$

$$\sum x_i \ln \frac{p_0(1-p_1)}{p_1(1-p_0)} < \lambda^*$$

$$\sum x_i > \lambda^{**}.$$

Por tanto

$$C = \left\{(X_1, X_2, ..., X_n) \mid \sum x_i > \lambda^{**}\right\}.$$

[4] El lema describe un procedimiento para el cual, dado un nivel de significación, la probabilidad de cometer el error tipo II es mínima.

Hasta aquí se ha hablado de como calcular la región de rechazo cuando las hipótesis de prueba son simples, pero cuando las hipótesis son compuestas la prueba que proporciona su región de rechazo se encuentra dada por la razón de verosimilitud.

Suponga que Θ es el conjunto de todos los valores que el parámetro de suposición θ puede asumir y que se quiere probar las hipótesis

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

donde Θ_0 es un subconjunto de Θ y Θ_1 es el complemento de Θ_0 con respecto a Θ.

El estadístico de prueba es,

$$k = \frac{\max\limits_{H_0:\theta\in\Theta_0} L(\theta)}{\max\limits_{H_1:\theta\in\Theta_1} L(\theta)}$$

donde $\max\limits_{H_0:\theta\in\Theta_0} L(\theta)$ y $\max\limits_{H_1:\theta\in\Theta_1} L(\theta)$ son los valores máximos de la función de verosimilitud para todos los valores θ en Θ_0 y θ en Θ_1, respectivamente, y la región crítica está determinada por

$$k \leq \lambda$$

donde $0 < \lambda < 1$, define una **prueba de razón de verosimilitud** de la hipótesis nula $\theta \in \Theta_0$ versus la hipótesis alternativa $\theta \in \Theta_1$. (Freund, 2000: 402).

Definición 1.4.6. *Dadas las hipótesis $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, la región crítica C se dice que es uniformemente más potente si para cada $\theta_0 \in \Theta_0$ y $\theta_1 \in \Theta_1$, C es la mejor región crítica más potente de la prueba*

$$H_0 : \theta = \theta_0 \quad \textit{versus} \quad H_1 : \theta = \theta_1.$$

En conclusión, los pasos que se necesitan llevar a cabo para utilizar una prueba de hipótesis en la toma de decisiones son:

1. Formular la H_0 y la H_1, y especificar el nivel de significación de la prueba (α).
2. En base a la distribución de los datos y a la información con que se cuente sobre estos, utilizar una estadística de prueba apropiada, y determinar una región crítica de tamaño α.
3. Determinar el valor del estadístico de prueba a partir de los datos de la muestra.
4. Comprobar si el estadístico de prueba cae en la región crítica y, en base a esto, emitir una opinión sobre si se rechaza o no la hipótesis nula.

Ejemplo 1.4.3. *Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de una población con distribución Poisson, $p(x) = \frac{\theta^x e^{-\theta}}{x!}$. Dadas las hipótesis $H_0 : \theta = 0.3$ versus $H_1 : \theta < 0.3$, encuentre la región crítica C por medio de la prueba de razón de verosimilitud.*

Solución. *La región crítica se encuentra como:*

$$C = \left\{ (X_1, X_2, ..., X_n) | \frac{L(\theta = 3)}{\max\limits_{H_0:\theta<3} L(\theta)} \leq \lambda \right\}$$

Para el numerador se tiene que,

$$L(\theta = 3) = \frac{3^{\sum\limits_{i=1}^{n} x_i} e^{-3n}}{X_1!X_2!...X_n!}$$

Mientras que para el denominador se tiene que encontrar el

$$\max_{H_0:\theta<3} L(\theta) = \frac{\theta^{\sum\limits_{i=1}^{n} x_i} e^{-n\theta}}{x_1!x_2!...x_n!}.$$

Aplicando logaritmos y derivando con respecto a θ se obtiene que,

$$\ln L(\theta) = \sum_{i=1}^{n} x_i \ln \theta - n\theta - \ln(x_1!x_2!...x_n!))$$

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \frac{\sum\limits_{i=1}^{n} x_i}{\theta} - n = 0$$

lo que implica que,

$$\theta = \frac{\sum\limits_{i=1}^{n} x_i}{n} = \overline{x}$$

A continuación se tiene que verificar que $\theta = \overline{x}$ sea un máximo,

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = -\frac{\sum\limits_{i=1}^{n} x_i}{\theta^2}$$

sutituyendo el valor de θ,

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = -\frac{n^2}{\sum\limits_{i=1}^{n} x_i} < 0$$

por tanto, hay un máximo.

Finalmente, la razón de verosimilitud queda

$$\frac{L(\theta = 3)}{\max\limits_{H_0:\theta<3} L(\theta)} = \frac{\frac{3^{\sum\limits_{i=1}^{n} x_i} e^{-3n}}{x_1!x_2!...x_n!}}{\frac{\overline{x}^{\sum\limits_{i=1}^{n} x_i} e^{-n\overline{x}}}{x_1!x_2!...x_n!}} \leq \lambda$$

simplificando términos

$$\frac{3^{\sum_{i=1}^{n} x_i} e^{-3n}}{\overline{x}^{\sum_{i=1}^{n} x_i} e^{-n\overline{x}}} = \left(\frac{3}{\overline{x}}\right)^{\sum_{i=1}^{n} x_i} e^{n(\overline{x}-3)} \leq \lambda$$

Por lo tanto

$$C = \left\{ (x_1, x_2, ..., x_n) \mid \left(\frac{3}{\overline{x}}\right)^{\sum_{i=1}^{n} x_i} e^{n(\overline{x}-3)} \leq \lambda \right\}.$$

Se ha dado un panorama general de lo que aborda la estimación de parámetros y el concepto de la máxima verosimilitud; puntos, que se aplicaran en el tercero y cuarto capítulos.

Capítulo 2

Matrices y distribuciones multivariadas

Cuando se habla de modelos de más de una variable explicativa, en ocasiones surge la necesidad de emplear matrices, por ello, en el presente capítulo se proporcionan algunos conceptos y resultados de algebra matricial que se vinculan con el desarrollo de una función de distribución de p variables aleatorias, en particular se revisa el caso cuando la función sigue una distribución Normal.

2.1 Algebra matricial

En este sección se dan algunos conceptos sobre vectores y matrices, tales como su definición, propiedades y características.

Definición 2.1.1. *Un vector es un arreglo ordenado de n números (Grossman, 45).*

$$v = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Definición 2.1.2. *Una matriz A de $m \times n$ es una colección rectangular de $m \times n$ números ordenados en m renglones y n columnas.*

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mj} & \ldots & a_{mn} \end{pmatrix} \qquad i = 1, 2, ...m; \quad j = 1, 2, ..., n$$

donde a_{ij} indica un elemento dentro de la matriz. Es decir, a_{ij} es el elemento que aparece en el renglón i y en la columna j de A.

La dimensión de una matriz se encuentra determinado por el número de renglones y columnas que la conforman. En el caso de la matriz A su dimensión es de m renglones por n columnas $(m \times n)$.[1] Puesto que un vector puede entenderse también como una matriz, entonces la dimensión del vector v es de n renglones por 1 columna $(n \times 1)$.

Ejemplo 2.1.1. *Sea* $B = \begin{pmatrix} 3 & 5 & 7 \\ 4 & 8 & 0 \end{pmatrix}$, *esta matriz tiene* 2 *renglones y* 3 *columnas, por tanto su dimensión es de* 2×3.

Definición 2.1.3 (**Matriz Cuadrada**). *Si A es una matriz de $m \times n$ con $m = n$, entonces A se llama matriz cuadrada. Es decir, A tiene el mismo número de columnas y renglones $(m = n)$.*

Definición 2.1.4 (**Igualdad de matrices**). *Dos matrices A y B son iguales, si y sólo si, tienen la misma dimensión y cada elemento de A es igual a cada elemento de B.*

$$A = B, \text{ si y sólo si } \ a_{ij} = b_{ij} \text{ para todo } i \text{ y } j.$$

Definición 2.1.5 (**Suma de matrices**). *Sean A y B dos matrices de $m \times n$. Entonces la suma de A y B es la matriz de $n \times m$, $A + B$ esta dada por,*

$$A + B = \begin{pmatrix} a_{11} + b_{11} & \dots & a_{1j} + b_{1j} & \dots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} + b_{i1} & \dots & a_{ij} + b_{ij} & \dots & a_{in} + b_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \dots & a_{mj} + b_{mj} & \dots & a_{mn} + b_{mn} \end{pmatrix}$$

$A + B$ se obtiene al sumar los elementos correspondientes de A y B (Grossman, 51).[2]

Definición 2.1.6 (**Multiplicación de una matriz por un escalar**). *Si A es una matriz de $m \times n$ y α es un escalar, entonces la matriz $m \times n$, αA, esta dada por,*

$$\alpha A = \begin{pmatrix} \alpha a_{11} & \dots & \alpha a_{1j} & \dots & \alpha a_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \alpha a_{i1} & \dots & \alpha a_{ij} & \dots & \alpha a_{2n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \alpha a_{m1} & \dots & \alpha a_{mj} & \dots & \alpha a_{mn} \end{pmatrix}$$

Definición 2.1.7 (**Producto Punto**). *Sean* $a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$ *y* $b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$ *dos vectores de igual dimensión,* $1 \times n$. *Entonces el producto punto de a y b, denotado por $a \cdot b$ esta dado por*

$$a \cdot b = a_1 b_1 + a_2 b_2 + \cdots + a_n a_n.$$

[1] A partir de ahora, se designa a una matriz mediante una letra mayúscula y a un vector mediante una letra minúscula.

[2] La suma de matrices sólo esta definida cuando las matrices tienen la misma dimensión, puesto que la suma realiza la operación elemento a elemento, $a_{ij} + b_{ij}$.

Observe que el producto de estos vectores es un escalar (Grossman, 62).

Definición 2.1.8 (Producto de matrices). *Sean A una matriz de $n \times m$ y B una matriz de $m \times p$. Entonces el producto de A y B es una matriz C de $n \times p$, en donde* [3]

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{im}b_{mj} = (\textit{renglón } i \textit{ de } A) \cdot (\textit{columna } j \textit{ de } B)$$

Es decir,

$$AB = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{im} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} \begin{pmatrix} b_{11} & \dots & b_{1j} & \dots & b_{1p} \\ b_{21} & \dots & b_{2j} & \dots & b_{2p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{m1} & \dots & b_{mj} & \dots & b_{mp} \end{pmatrix} = \begin{pmatrix} c_{11} & \dots & c_{1j} & \dots & c_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{i1} & \dots & c_{ij} & \dots & c_{ip} \\ \vdots & \vdots & \ddots & \vdots & \\ c_{n1} & \dots & c_{nj} & \dots & c_{np} \end{pmatrix} = C.$$

La multiplicación de AB sólo esta definida si y sólo si el número de columnas de A es igual número de renglones de B.

Ejemplo 2.1.2. *Sea* $A = \begin{pmatrix} 3 & 6 \\ 1 & 2 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 5 & 2 \\ 1 & 0 & 3 \end{pmatrix}$, *y* $C = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$.

Calcule AC, CA y BC.[4]

Solución.

$$AC = \begin{pmatrix} \begin{pmatrix} 3 & 6 \end{pmatrix}\begin{pmatrix} 1 \\ 3 \end{pmatrix} & \begin{pmatrix} 3 & 6 \end{pmatrix}\begin{pmatrix} 2 \\ 2 \end{pmatrix} \\ \begin{pmatrix} 1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 3 \end{pmatrix} & \begin{pmatrix} 1 & 2 \end{pmatrix}\begin{pmatrix} 2 \\ 2 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} (3)(1)+(6)(3) & (3)(2)+(6)(2) \\ (1)(1)+(2)(3) & (1)(2)+(2)(2) \end{pmatrix}$$

$$= \begin{pmatrix} 21 & 18 \\ 7 & 6 \end{pmatrix}.$$

$$CA = \begin{pmatrix} \begin{pmatrix} 1 & 2 \end{pmatrix}\begin{pmatrix} 3 \\ 1 \end{pmatrix} & \begin{pmatrix} 1 & 2 \end{pmatrix}\begin{pmatrix} 6 \\ 2 \end{pmatrix} \\ \begin{pmatrix} 3 & 2 \end{pmatrix}\begin{pmatrix} 3 \\ 1 \end{pmatrix} & \begin{pmatrix} 3 & 2 \end{pmatrix}\begin{pmatrix} 6 \\ 2 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} (1)(3)+(2)(1) & (1)(6)+(2)(2) \\ (3)(3)+(2)(1) & (3)(6)+(2)(2) \end{pmatrix}$$

$$= \begin{pmatrix} 5 & 10 \\ 11 & 22 \end{pmatrix}.$$

Como se observa $AC \neq CA$.
La multiplicación BC no esta definida dado que el número de renglones de B es distinto del

[3] El renglón i de A y la columna j de B tienen que ser de la misma dimensión.
[4] El producto de matrices no es conmutativo, es decir, $AB \neq BA$.

número de columnas de C.

$$BC = \begin{pmatrix} \begin{pmatrix}1 & 5 & 2\end{pmatrix}\begin{pmatrix}1\\3\end{pmatrix} & \begin{pmatrix}1 & 5 & 2\end{pmatrix}\begin{pmatrix}2\\2\end{pmatrix} \\ \\ \begin{pmatrix}1 & 0 & 3\end{pmatrix}\begin{pmatrix}1\\3\end{pmatrix} & \begin{pmatrix}1 & 0 & 3\end{pmatrix}\begin{pmatrix}2\\2\end{pmatrix} \end{pmatrix}.$$

Definición 2.1.9 (**Matriz Identidad**). *La matriz identidad I_n es una matriz de $n \times n$ cuyos elementos de la diagonal principal son unos en la diagonal principal y todos los demás son cero. Por ejemplo,*

$$I_2 = \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}.$$

Definición 2.1.10 (**Matriz Inversa**). *Sean A y B dos matrices de $m \times n$. Suponga que $AB = BA = I$. Entonces B se llama la inversa de A y se denota por A^{-1}, donde se tiene que*

$$AA^{-1} = A^{-1}A = I.$$

Si A tiene inversa se dice que A es invertible.

Definición 2.1.11 (**Matriz Transpuesta**). *Sea A una matriz de $m \times n$. La transpuesta de A, la cual de denota como A^T, es la matriz de $n \times m$ obtenida al intercambiar los renglones por las columnas de A. Es decir,*

$$A^T = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix}.$$

Simplemente se coloca el renglón i de A como la columna i de A^T y la columna j de A como el renglón j de A^T (Grossman, 122).

Ejemplo 2.1.3. *Encuentre la transpuesta de la siguiente matriz*

$$A = \begin{pmatrix}3 & 6 & 7\\5 & 8 & 2\end{pmatrix}.$$

Solución. *Al intercambiar los renglones por las columnas se tiene:*

$$A^T = \begin{pmatrix}3 & 5\\6 & 8\\7 & 2\end{pmatrix}.$$

Definición 2.1.12 (**Matriz simétrica**). *La matriz cuadrada A de $n \times n$,se llama simétrica si $A^T = A$ renglones. Es decir, las columnas de A son también los renglones. Por ejemplo,*

$$A = \begin{pmatrix}3 & 5\\5 & 8\end{pmatrix}, \qquad A^T = \begin{pmatrix}3 & 5\\5 & 8\end{pmatrix}.$$

Como $A^T = A$, entonces A es simétrica.

Teorema 2.1.1. *Sea A una matriz de $m \times n$ y B una matriz de $n \times p$ (Grossman, 1996:122). Entonces*

1. $(A^T)^T = A$.
2. $(AB)^T = B^T A^T$.
3. *Si A y B son de igual dimensión, $n \times n$, implica que* $(A + B)^T = A^T + B^T$.
4. *Si A es simétrica,* $A = A^T$.
5. *Si A es invertible,* $(A^T)^{-1} = (A^{-1})^T$.

Definición 2.1.13 (**Matriz diagonal**). *Sea A una matriz cuadrada $n \times n$, entonces A se llama matriz diagonal si todos sus elementos fuera de la diagonal principal son cero. Por ejemplo,*

$$D = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}.$$

Definición 2.1.14 (**Matrix triangular**). *Sea A una matriz cuadrada $n \times n$, entonces A se llama matriz diagonal si tiene ceros por encima o por debajo de su diagonal principal.*

Definición 2.1.15 (**Matrix J**). *La matriz $J_{m\times n}$ es una matriz donde todos sus elementos son uno. Por ejemplo,*

$$J_{n\times 1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad J_{n\times n} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}.$$

Definición 2.1.16 (**Menor**). *Sea A una matriz de $n \times n$ y sea M_{ij} una matriz de dimensión $(n-1) \times (n-1)$ la cual se obtiene de eliminar el renglón i y la columna j de A. M_{ij} se conoce como el menor de A(Groosman, p.175).*

Definición 2.1.17 (**Cofactor**). *Sea A una matriz de $n \times n$. El cofactor ij de A esta dado por* $A_{ij} = (-1)^{i+j} det(M_{ij})$.

Definición 2.1.18 (**Determinante**). *Sea A una matriz cuadrada $n \times n$. El determinante de A, el cual se denota como $det(A)$, está dado por:*

$$det(A) = a_{11}A_{11} + a_{12}A_{12} + \cdots + a_{1n}A_{1n} = \sum_{t=1}^{n} a_{1t}A_{1t}$$

donde A_{ij} es el cofactor ij de A.

Ejemplo 2.1.4. *Sea* $A = \begin{pmatrix} 1 & 3 & 5 \\ 0 & -1 & 3 \\ 2 & 1 & 9 \end{pmatrix}$. *Calcule su determinante.*

Solución. *Primero se encuentran los menores de a A. Eliminando el primer renglón y la primera columna de A se obtine* $M_{11} = \begin{pmatrix} -1 & 3 \\ 1 & 9 \end{pmatrix}$. *De manera similar eliminando el primer renglón y la segunda columna, el primer renglón y la tercera columna, respectivamente, se obtiene* $M_{12} = \begin{pmatrix} 0 & 3 \\ 2 & 9 \end{pmatrix}$ *y* $M_{13} = \begin{pmatrix} 0 & -1 \\ 2 & 1 \end{pmatrix}$. *Los cofactores de A son*[5] $A_{11} = (-1)^{1+1} det \begin{pmatrix} -1 & 3 \\ 1 & 9 \end{pmatrix} = -12$, $A_{12} = (-1)^{1+2} det \begin{pmatrix} 0 & 3 \\ 2 & 9 \end{pmatrix} = 6$ *y* $A_{13} = (-1)^{1+3} det \begin{pmatrix} 0 & -1 \\ 2 & 1 \end{pmatrix} = 2$. *Por tanto, el determinate de*

$$det(A) = a_{11}A_{11} + a_{12}A_{12} + +a_{1n}A_{13} = (1)(-12) + (3)(6) + (5)(2) = 16.$$

Definición 2.1.19 (**Rango**). *El rango de una matriz A de dimensión* $m \times n$ *es el número máximo de vectores columna (o vectores renglón) linealmente independientes dentro de la matriz. Si* $m > n$ *el número máximo de vectores que pueden ser linealmente independientes es n. El rango de la matriz A se denota como:* $rango(A)$ *(Peña, 2002:25).*

Algunas de las propiedades del $rango(A)$ son:

- $rango(A) \leq min(m, n)$.
- Si $rango(A) = min(m, n)$, se dice que A tiene rango completo.
- $rango(A + B) \leq rango(A) + rango(B)$.
- $rango(AB) \leq min(rango(A), rango(B))$.
- $rango(AA^T) = rango(A^T A) = rango(A)$.

Definición 2.1.20 (**Traza**). *La traza de una matriz cuadrada A de* $n \times n$ *es la suma de los elementos de la diagonal principal (Greene, 35):*

$$tr(A) = \sum_{i=1}^{n} a_{ii}.$$

Algunas propiedades de la traza son:

- $tr(cA) = c(tr(A))$, donde c es un escalar.
- $tr(A^T) = tr(A)$.
- $tr(A + B) = tr(A) + tr(B)$.
- $tr(AB) = tr(BA)$.
- Si A es simétrica, $tr(A^2) = tr(AA) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j}^2$.

[5]El determinante de una matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ de 2×2 se define como $det(A) = a_{11}a_{22} - a_{12}a_{21}$.

Definición 2.1.21 (**Matriz idempotente**). *Una matriz idempotente es aquella que es igual a su cuadrado, es decir,* $A^2 = AA = A$ *(Greene, 1999:13).*

Algunos resultados importantes acerca de las matrices idempotentes son:

- Si A es una matriz idempotente simétrica, entonces $A^TA = A$.
- Si A es una matriz idempotente, el $rango(A) = tr(A)$.
- La traza de una matriz idempotente simétrica es igual a la suma de los valores propios (λ) de A.

$$tr(A) = \sum_{i=1}^{n} \lambda_i.$$

- Los valores propios de una matriz idempotente son cero o uno.

Ejemplo 2.1.5. *Sea* $F = I_n - \frac{1}{n}J_{n\times n}$. *Muestre que es idempotente,* $FF = F$.

Solución.

$$(I_n - \frac{1}{n}J_{n\times n})(I_n - \frac{1}{n}J_{n\times n}) = I_n - \frac{1}{n}J_{n\times n} - \frac{1}{n}J_{n\times n} + \frac{1}{n^2}J^2_{n\times n}$$

donde $J^2_{n\times n} = nJ_{n\times n}$

$$\begin{aligned}(I_n - \frac{1}{n}J_{n\times n})(I_n - \frac{1}{n}J_{n\times n}) &= I_n - \frac{1}{n}J_{n\times n} - \frac{1}{n}J_{n\times n} + \frac{1}{n}J_{n\times n} \\ &= I_n - \frac{1}{n}J_{n\times n}\end{aligned}$$

por lo tanto F *es idempotente.*

Definición 2.1.22 (**Valor y Vector Propio**). *Sea* A *una matriz de* $n \times n$ *y* λ *un número. Entonces* λ *se denomina el valor propio de* A *si existe un vector diferente de cero* $\boldsymbol{u}$ *tal que*

$$A\boldsymbol{u} = \lambda\boldsymbol{u}.$$

El vector $\boldsymbol{u}$ *se conoce como el vector propio de* A *correspondiente al valor propio* λ *(Grossman, 535).*

Teorema 2.1.2. *Sea* A *una matriz de* $n \times n$. *Entonces* λ *es un valor propio de* A *si y sólo si*

$$p(\lambda) = det(A - \lambda I) = 0$$

donde $p(\lambda)$ *se llama el polinomio característico de* A.

Hasta aquí se presentaron conceptos esenciales de álgebra matricial requeridos para el desarrollo de las siguientes secciones. A continuación se estudia la función de distribución Normal multivariada.

2.2 Distribución de probabilidad multivariada

En esta sección se presentan ciertos conceptos de la distribución de probabilidad para el caso n variables aleatorias, en particular se revisa el caso cuando la función sigue una distribución Normal.

2.2.1 Función de distribución y función de densidad

Definición 2.2.1. *La función de distribución conjunta de p variables aleatorias,* $X_1, X_2, \ldots, X_n$, *se define como:*

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

para $-\infty < x_i < \infty, \quad i = 1, 2, \ldots, n.$

A) En el caso de que las n variables aleatorias sean discretas,

$$F(x_1, x_2, \ldots, x_n) = \sum_{u_1 \leq x_1} \sum_{u_2 \leq x_2} \cdots \sum_{u_n \leq x_n} f(u_1, u_2, \ldots, u_n)$$

para $-\infty < x_i < \infty, \quad i = 1, 2, \ldots, n.$

Donde $f(x_1, x_2, \ldots, x_n)$ es la función de densidad de probabilidad conjunta,[6] la cual esta dada por

$$f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

B) Mientras que en el caso continuo

$$F(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_n} \int_{-\infty}^{x_{n-1}} \cdots \int_{-\infty}^{x_1} f(u_1, u_2, \ldots u_n) du_1 du_2 \ldots du_n$$

para $-\infty < x_i < \infty, \quad i = 1, 2, \ldots, n.$ Donde[7]

$$f(x_1, x_2, \cdots, x_n) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}.$$

En ambos casos la función de distribución, $F(x_1, x_2, \ldots, x_n)$ se encuentra dentro del intervalo $[0, 1]$.

[6] La función de densidad conjunta de n variables aleatorias discretas debe satisfacer:

(a) $f(x_1, x_2, \ldots, x_n) \geq 0$, para cada n-esíma de valores $(x_1, x_2, \ldots, x_n)$ dentro de su dominio.

(b) $\sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty} \cdots \sum_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) = 1.$

[7] La función de densidad conjunta de n variables aleatorias continuas debe satisfacer:

(a) $f(x_1, x_2, \ldots, x_n) \geq 0$, para $-\infty < x_i < \infty, \quad i = 1, 2, \ldots, n.$

(b) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_n) dx_1 \cdots dx_n = 1.$

2.2.2 Función de densidad marginal e independencia

Definición 2.2.2. *La función de densidad marginal de $F(x_1, x_2, \ldots, x_n)$ se define con respecto a una variable aleatoria individual, se obtiene sumando (caso discreto) o integrando (caso continuo) sobre las demás variables (Greene, 1999:66).*

A) En el caso discreto se tiene

$$f_i(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} f(x_1, \ldots, x_i, \ldots, x_n)$$

B) Mientras que en el caso continuo

$$f_i(x_i) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_i, \ldots, x_n) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n$$

Definición 2.2.3. *Sean $X_1, X_2, \ldots, X_n$, n variables aleatorias con función de densidad de probabilidad conjunta $f(x_1, x_2, \ldots x_n)$ y $f_i(x_i)$ la función de distribución marginal de X_i para $i = 1, 2, \ldots, n$, entonces las n variables aleatorias son independientes si y sólo si*

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdot \cdots \cdot f_n(x_n)$$

para toda $(x_1, x_2, \ldots x_n)$.[8]

2.2.3 Valor esperado y matriz de varianza covarianza

Definición 2.2.4. *Sean $X_1, X_2, \ldots, X_n$, n variables aleatorias, entonces una matriz aleatoria es un vector cuyos elementos son variables aleatorias.*

Definición 2.2.5. *Sea $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$ un vector aleatorio. Entonces, el valor esperado de X es el vector de valores esperados, esto es*

$$E(X) = E \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \mu.$$

Definición 2.2.6. *La matriz de varianza covarianza de un vector aleatorio X se define como:*

$$V(X) = E[(X - E(X))(X - E(X))^T].$$

[8] Mendenhall, 2002:234.

Es decir,

$$V(X) = E\left[\begin{pmatrix} X_1 - E(X_1) \\ X_2 - E(X_2) \\ \vdots \\ X_n - E(X_n) \end{pmatrix} \begin{pmatrix} X_1 - E(X_1), & X_2 - E(X_2), & \ldots, & X_n - E(X_n) \end{pmatrix}\right]$$

$$= E\begin{pmatrix} (X_1 - E(X_1))^2 & (X_1 - E(X_1))(X_2 - E(X_2)) & \cdots \\ (X_1 - E(X_1))(X_2 - E(X_2)) & (X_2 - E(X_2))^2 & \cdots \\ \vdots & \vdots & \cdots \\ (X_n - E(X_n))(X_1 - E(X_1)) & (X_n - E(X_n))(X_2 - E(X_2)) & \cdots \end{pmatrix}$$

$$= \begin{pmatrix} E(X_1^2) - \mu_1^2 & E(X_1X_2) - \mu_1\mu_2 & \cdots \\ E(X_2X_1) - \mu_2\mu_1 & E(X_2^2) - \mu_2^2 & \cdots \\ \vdots & \vdots & \cdots \\ E(X_nX_1) - \mu_n\mu_1 & E(X_nX_2) - \mu_n\mu_2 & \cdots \end{pmatrix} = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots \\ Cov(X_2, X_1) & Var(X_2) & \cdots \\ \vdots & \vdots & \cdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \cdots \end{pmatrix}$$

donde $Var(X_i)$ representa la varianza de la variable aleatoria X_i y los elementos fuera de la diagonal $Cov(X_i, X_j)$ representan la covarianza que existe entre X_i y X_j.

Teorema 2.2.1. *Sea X un vector aleatorio con media $\mu = E(X)$, entonces la matriz de varianza covarianza es*

$$V(X) = E[(X - E(X))(X - E(X))^T] = E(XX^T) - \mu\mu^T.$$

Demostración.

$$\begin{aligned} V(X) &= E[(X - E(X))(X - E(X))^T] \\ &= E[(X - E(X))(X^T - E(X)^T)] \\ &= E[(XX^T - XE(X)^T - E(X)X^T + E(X)E(X)^T)] \end{aligned}$$

dado que $E(X) = \mu$, y aplicando el valor esperado, se tiene

$$\begin{aligned} V(X) &= E(XX^T) - \mu\mu^T - \mu\mu^T + \mu\mu^T \\ &= E(XX^T) - \mu\mu^T. \end{aligned}$$

■

Teorema 2.2.2. *Si X_i y X_j son variables aleatorias independientes, entonces*

$$E(X_iX_j) = E(X_i)E(X_j) \quad y \quad Cov(X_i, X_j) = 0.$$

Demostración.

$$E(X_iX_j) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_ix_jf(x_i,x_j)dx_idx_j.$$

Dado que X_i y X_j son independientes, $f(x_i,x_j) = f_i(x_i)f_j(x_j)$

$$E(X_iX_j) = \int_{-\infty}^{\infty} x_if_i(x_i)dx_i \int_{-\infty}^{\infty} x_jf_j(x_j)dx_j = E(X_i)E(X_j).$$

$$\begin{aligned} Cov(X_i,X_j) &= E(X_i - E(X_i))(X_j - E(X_j)) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_i - E(X_i))(x_j - E(X_j))f(x_i,x_j)dx_idx_j \end{aligned}$$

Dado que X_i y X_j son independientes, $f(x_i,x_j) = f_i(x_i)f_j(x_j)$

$$\begin{aligned} Cov(X_i,X_j) &= \int_{-\infty}^{\infty} (x_i - E(X_i))f_i(x_i)dx_i \int_{-\infty}^{\infty} (x_j - E(X_j))f_j(x_j)dx_j \\ &= \left(\int_{-\infty}^{\infty} x_if_i(x_i)dx_i - E(X_i)\int_{-\infty}^{\infty} f_i(x_i)dx_i\right)\cdot \\ &\qquad \left(\int_{-\infty}^{\infty} x_jf_j(x_j)dx_j - E(X_j)\int_{-\infty}^{\infty} f_j(x_j)dx_j\right) \\ &= \left(\int_{-\infty}^{\infty} x_if_i(x_i)dx_i - E(X_i)\right)\cdot\left(\int_{-\infty}^{\infty} x_jf_j(x_j)dx_j - E(X_j)\right) \end{aligned}$$

$$Cov(X_i,X_j) = (E(X_i) - E(X_i))\,(E(X_j - E(X_j)) = 0$$

■

Definición 2.2.7. *Sea X un vector aleatorio de $n \times 1$ con función de densidad de probabilidad $f(x_1,\ldots,x_n)$ y a^T un vector de $1 \times n$. Se tiene que*

$$Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n \tag{2.1}$$

es una combinación lineal.

La ecuación 2.1 se puede expresar en forma matricial como

$$Y = a^TX$$

donde $a^T = (a_1, a_2, \cdots, a_n)$ y $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$.

Teorema 2.2.3. *Sea X un vector aleatorio con $E(X) = \mu$, $V(X) = V$ y $Y = a^T X$, con a^T constante, entonces*

1. *El valor esperado de Y es $E(Y) = a^T \mu$.*
2. *La matriz de varianza covarianza de Y es $V(Y) = a^T V a$.*

El valor esperado de Y es

$$E(Y) = E(a^T X) = a^T E(X)$$

dado que $E(X) = \mu$, se tiene que

$$E(Y) = a^T \mu.$$

Mientras que la matriz de varianza covarianza de Y es

$$\begin{aligned} V(Y) = V(a^T X) &= E((a^T X - E(a^T X))(a^T X - E(a^T X))^T) \\ &= E((a^T (X - E(X)))(a^T (X - E(X)))^T) \\ &= E(a^T (X - E(X))(X - E(X))^T a) = a^T E((X - E(X))^2) a \\ &= a^T E((X - E(X))(X - E(X))^T) a \end{aligned}$$

dado que $V(X) = E[(X - E(X))(X - E(X))^T] = V$, se tiene que

$$V(Y) = a^T V a.$$

2.2.4 Forma lineal y forma cuadrática

Definición 2.2.8. *Sea X un vector aleatorio y sea A una matriz . Entonces, una **forma lineal** es una expresión de la forma*

$$AX.$$

De forma analoga, el producto de A con X es una combinación lineal de las columnas de A.

$$AX = X_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{m1} \end{pmatrix} + X_2 \begin{pmatrix} a_{12} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{m2} \end{pmatrix} + ... + X_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{in} \\ \vdots \\ a_{mn} \end{pmatrix}.$$

La forma lineal AX también se puede escribir como

$$AX = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{in} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n \\ \vdots \\ a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_n \\ \vdots \\ a_{m1}X_1 + a_{m2}X_2 + \cdots + a_{mn}X_n \end{pmatrix}$$

donde el i-ésimo elemento de AX es de la forma $Y = a_i^T X$, por ende, por el teorema 2.2.3 se tiene que AX es un vector aleatorio con:

$$E(AX) = A\mu \qquad \text{y} \qquad V(AX) = AVA^T.$$

Definición 2.2.9. *Sea X un vector aleatorio y sea B una matriz simétrica. Entonces, una* ***forma cuadrática*** *es una expresión de la forma*

$$X^T BX.$$

En este apartado se dieron conceptos básicos de la teoria de distribución cuando se tienen n variables aleatorias. A continuación se presentan algunos resultados de la función de distribución Normal multivariada.

2.3 Distribuciones conocidas

2.3.1 Normal multivariada

Definición 2.3.1. *Se dice que el vector aleatorio X, $X^T = (X_1, X_2, \ldots, X_n)$, se distribuye Normal con media $E(X) = \mu$ y matriz de varianza covarianza $V(X) = V$, $X \sim N(\mu; V)$. Su función de densidad conjunta esta dada por,*

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} |V|^{1/2}} e^{-\frac{1}{2}(X-\mu)^T V^{-1}(X-\mu)}, \quad \textit{para toda } x \in \mathbb{R}^n.$$

donde $\mid V \mid$ es el determinante de V.

Definición 2.3.2. *Sea X un vector aleatorio con función de distribución acumulada $F(X)$. La función generatriz de momentos de X es,*[9]

$$M_X(t) = E(e^{t^T X}).$$

Un resultado importante es que si X_1 y X_2 son variables aleatorias con función generatriz de momentos $M_{X_1}(t)$ y $M_{X_2}(t)$ respectivamente. Entonces, X_1 y X_2 tienen la misma distribución de probabilidad sólo si $M_{X_1}(t) = M_{X_2}(t)$ idénticamente (R. Spiegel, 1976:80).

Teorema 2.3.1. *La función generatriz de momentos de la distribución Normal multivariada está dada por*

$$M_X(t) = e^{t^T \mu + \frac{1}{2} t^T V t}$$

Demostración. *Por definición*

$$M_X(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{t^T X} \frac{1}{(2\pi)^{\frac{n}{2}} \mid V \mid^{1/2}} e^{-\frac{1}{2}(X-\mu)^T V^{-1}(X-\mu)} dx_1 \cdots dx_n$$

[9] Casella, 2002:62.

$$M_X(t) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid V \mid^{1/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}[(X-\mu)^T V^{-1}(X-\mu) - 2t^T X]} dx_1 \cdots dx_n$$

completando el cuadrado en el exponente del integrando se obtiene

$$(X-\mu)^T V^{-1}(X-\mu) - 2t^T X = (X-(\mu+Vt))^T V^{-1}(X-(\mu+Vt)) - 2t^T\mu - t^T Vt$$

por tanto

$$M_X(t) = e^{t^T\mu + \frac{1}{2}t^T Vt} \underbrace{\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{n}{2}} \mid V \mid^{1/2}} e^{-\frac{1}{2}(X-(\mu+Vt))^T V^{-1}(X-(\mu+Vt))} dx_1 \ldots dx_n}_{= \ uno}$$

dado que las n integrales de $-\infty$ a ∞ es una función de densidad Normal multivariada con media $\mu + Vt$ y varianza V, por ende, las n integrales son iguales a uno y

$$M_X(t) = e^{t^T\mu + \frac{1}{2}t^T Vt}.$$

■

Teorema 2.3.2. *Sea X un vector aleatorio que se distribuye Normal con media $E(X) = \mu$ y matriz de varianza covarianza $V(X) = V$, $X \sim N(\mu, V)$, y A una matriz, entonces la forma lineal AX tiene una distribución igual a $AX \sim N(A\mu, AVA^T)$.*

Demostración. *Por definición*

$$M_{AX}(t) = E(e^{t^T AX})$$

haciendo un cambio de variable $r^T = t^T A$

$$M_{AX}(t) = E(e^{t^T AX}) = E(e^{r^T X}) = M_X(r)$$

y por el resultado anterior se tiene que

$$M_{AX}(t) = M_X(r) = e^{r^T\mu + \frac{1}{2}r^T Vr}$$

sustitutendo el valor de r^T, se tiene que

$$M_{AX}(t) = e^{t^T A\mu + \frac{1}{2}t^T AVA^T t}$$

que es la función generatriz de momentos de una normal con media $A\mu$ y varianza AVA^T. Por lo tanto,

$$AX \sim N(A\mu, AVA^T).$$

■

Teorema 2.3.3. *Si el vector aleatorio X, $X^T = (X_1, X_2, \ldots, X_n)$, tiene una distribución Normal con media $E(X) = \mu$ y matriz de varianza covarianza $V(X) = V$, $X \sim N(\mu, V)$, y si $Cov(X_i, X_j) = 0$, entonces $X_1, X_2, \cdots, X_n$ son independientes.*

Demostración. *La función de densidad Normal multivariada para el vector aleatorio X, con media μ y matriz de varianza covarianza V, se define como*

$$f(x_1, ..., x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid V \mid^{\frac{1}{2}}} e^{-\frac{1}{2}(X-\mu)^T V^{-1}(X-\mu)}$$

Puesto que la $Cov(X_i, X_j) = 0$, la matriz de varianza covarianza es

$$V(X) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}, \quad \textit{y su inversa es} \quad V^{-1}(X) = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{pmatrix},$$

entonces $\mid V \mid^{\frac{1}{2}} = \sigma_1\sigma_2\cdots\sigma_n$. Sustituyendo estos resultados en la función de densidad, se tiene

$$f(x_1, ..., x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_1\sigma_2\cdots\sigma_n} e^{-\frac{1}{2}\left(\frac{(X_1-\mu_1)^2}{\sigma_1^2} + \frac{(X_2-\mu_2)^2}{\sigma_2^2} + \cdots + \frac{(X_n-\mu_n)^2}{\sigma_n^2}\right)}.$$

Acomodando términos

$$\begin{aligned} f(x_1, ..., x_n) &= \frac{1}{(2\pi)^{\frac{1}{2}}\sigma_1} e^{-\frac{1}{2}\frac{(X_1-\mu_1)^2}{\sigma_1^2}} \cdots \frac{1}{(2\pi)^{\frac{1}{2}}\sigma_n} e^{-\frac{1}{2}\frac{(X_n-\mu_n)^2}{\sigma_n^2}} \\ f(x_1, ..., x_n) &= f_1(x_1)\ldots f_n(x_n). \end{aligned}$$

Por tanto, dado que la función de densidad conjunta es igual al producto de las marginales, $X_1, X_2, \ldots, X_n$ son independientes. ■

Ahora se pueden encontrar las condiciones necesarias y suficientes para que dos formas lineales o dos formas cuadráticas de un vector de variables aleatorias Normales sean independientes.

Teorema 2.3.4. *Sea X un vector aleatorio que se distribuye Normal con media $E(X) = \mu$ y matriz de varianza covarianza $V(X) = V$, $X \sim N(\mu, V)$, y sean AX y BX dos formas lineales de X. Entonces, AX y BX son independientes si y sólo si $AVB^T = 0$.*

Demostración. *Por el teorema 2.3.2 $AX \sim N(A\mu, AVA^T)$ y $BX \sim N(B\mu, BVB^T)$. Para que AX y BX sean independientes se debe satifacer que $Cov(AX, BX) = 0$. La covarianza se define como*

$$\begin{aligned} Cov(AX, BX) &= E[(AX - E(AX))(BX - E(BX))^T] \\ &= E[A(X - E(X))(X - E(X))^T B^T] \\ &= AE[(X - E(X))(X - E(X))^T]B^T \end{aligned}$$

donde $V(X) = E[(X - E(X))(X - E(X))^T] = V$

$$Cov(AX, BX) = AVB^T.$$

Puesto que X es un vector aletorio de variables normalmente distribuidas, $X_1, X_2, \ldots, X_n$, AX y BX son independientemente distribuidos y por el teorema 2.3.3 se tiene que

$$Cov(AX, BX) = AVB^T = 0.$$

■

Teorema 2.3.5. *Sea X un vector aleatorio que se distribuye Normal con media $E(X) = \mu$ y matriz de varianza covarianza $V(X) = V$, $X \sim N(\mu, V)$, y sean la forma lineal AX y la forma cuadrática X^TBX, con B simétrica. Entonces, AX y X^TBX son independientes si y sólo si $AVB^T = 0$.*

Demostración. *Se puede expresar a B como: $B = F^TDF$, donde D es una matriz diagonal y F es una matriz de $n \times n$. Como $X^TBX = X^TF^TDFX = (FX)^TDFX$, entonces FX es una forma lineal asociada a la forma cuadrática X^TBX.*

Por lo tanto, para determinar que AX y X^TBX son independientes, basta con probar que AX y FX son independientes. Es decir, que $AVF^T = 0$.

$$\begin{aligned} Cov(AX, FX) &= E((AX - E(AX))(FX - E(FX))^T) \\ &= E(A(X - E(X))(X - E(X))^TF^T] \\ &= AE((X - E(X))(X - E(X))^T)F^T \end{aligned}$$

donde $V(X) = E[(X - E(X))(X - E(X))^T] = V$. Por tanto

$$Cov(AX, FX) = AVF^T = 0.$$

Ya que X es un vector aletorio de variables normalmente distribuidas, $X_1, X_2, \ldots, X_n$, AX y FX son independientemente distribuidos. Consecuentemente, AX y $X^TBX = X^TF^TDFX$ son independientemente distribuidos. La independencia de AX y X^TBX implica que $AVB^T = 0$, por ende

$$AVB^T = AVF^TDF = 0.$$

Por consiguiente, AX y X^TBX son independientes. ■

Teorema 2.3.6. *Sea X un vector aleatorio que se distribuye Normal con media $E(X) = \mu$ y matriz de varianza covarianza $V(X) = V$, $X \sim N(\mu, V)$, y X^TAX y X^TBX son dos formas cuadráticas, entonces X^TAX y X^TBX son independientes si y sólo si $AVB^T = 0$.*

Demostración. *Primero se tiene que determinar la forma lineal asociada a cada forma cuadrática y después se tiene que probar que la covarianza de esas dos formas lineales es igual a cero.*
Se puede expresar a A como $A = L^TPL$ y a B como $B = F^TDF$, donde P y D son matrices diagonales y, L y F son matrices de $n \times n$. Entonces, X^TAX se puede escribir como $X^TAX = (LX)^TD(LX)$ y X^TBX como $X^TBX = (FX)^TD(FX)$, donde LX y FX son las formas lineales asociadas a estas dos formas cuadráticas respectivamente. En consecuencia, para comprobar que X^TAX y X^TBX son independientes, basta probar que $Cov(LX, FX) = 0$

$$\begin{aligned} Cov(LX, FX) &= E((LX - E(LX))(FX - E(FX))^T) \\ &= E(L(X - E(X))(X - E(X))^TF^T) \\ &= LE((X - E(X))(X - E(X))^T)F^T \end{aligned}$$

donde $V(X) = E[(X-(X))(X-E(X))^T] = V$. *Por tanto,*

$$Cov(LX, FX) = LVF^T = 0$$

Como X *es un vector aletorio de variables normalmente distribuidas,* $X_1, X_2, \ldots, X_n$, LX *y* FX *son independientemente distribuidos. Consecuentemente,* $X^TAX = (LX)^TD(LX)$ *y* $X^TBX = X^TF^TDFX$ *son independientemente distribuidos. La independencia de* X^TAX *y* X^TBX *implica que* $AVB^T = 0$,

$$AVB^T = L^TPLVF^TDF = 0.$$

Por tanto, X^TAX *y* X^TBX *son independientes.* ■

Estos resultados pueden ser aplicados para demostrar un hecho ampliamente utilizado y que se trata en el siguiente teorema.

Teorema 2.3.7. *Si* $\overline{X}$ *y* S^2 *son la media y la varianza de una muestra aleatoria de tamaño* n *de una población Normal con media* μ *y varianza* σ^2, *entonces,* $\overline{X}$ *y* S^2 *son independientes (Freud, 281).*

Demostración. *Sea* $X_1, X_2, \cdots X_n$ *una muestra aleatoria, donde* $X_i \sim N(\mu, \sigma^2)$, *con media* $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *y varianza* $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. *Para determinar la independencia de* $\overline{X}$ *y* S^2, *primero se tiene que encontrar la forma lineal asociada a estos dos estadísticos. Expresando* $\overline{X}$ *y* S^2 *en forma matricial, se tiene,*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\underbrace{(1,1,\ldots,1)}_{n \quad unos}\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \frac{1}{n}J_{1\times n}X = AX \tag{2.2}$$

donde $A = \left(\frac{1}{n}J_{1\times n}\right)$ *y* $J_{1\times n}$ *es la matriz* J *(ver definición 2.1.15).*

Por otro lado,

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 = (X - \overline{X})^T(X - \overline{X})$$

donde,

$$X - \overline{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} - \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

factorizando el vector X *se tiene*

$$X - \overline{X} = \left[\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} - \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}\right]\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \left(I_n - \frac{1}{n}J_{n\times n}\right)X = FX$$

donde $F = \left(I_n - \frac{1}{n}J_{n\times n}\right)$, I_n *es la matriz identidad y* $J_{n\times n}$ *es la matriz* J *(ver definición 2.1.15). Con lo que*

$$(n-1)S^2 = (X-\overline{X})^T(X-\overline{X}) = (FX)^TFX = X^TFX \tag{2.3}$$

puesto que F *es una matriz idempotente, ver ejemplo 2.1.5.*
Por tanto, la forma lineal asociada a $\overline{X}$ *y* S^2 *es* AX *y* FX *respectivamente. Ahora sólo se tiene que probar que* $AVF^T = 0$*, donde* $V = \sigma^2 I_n$ *es la matrix de varianza covarianza de* X*.*

$$AVF^T = \left(\frac{1}{n}J_{1\times n}\right)\sigma^2 I_n\left(I_n - \frac{1}{n}J_{n\times n}\right) = \frac{\sigma^2}{n}\left(J_{1\times n} - J_{1\times n}\right) = 0.$$

En conclusión, se tiene que $\overline{X}$ *y* S^2 *son independientes.* ■

2.3.2 Distribuciones χ^2, t y F

Además de la distribución Normal, hay otras distribuciones que se utilizan y se derivan de la Normal. Estas distribuciones son la $\boldsymbol{\chi^2}$, $\mathbf{t}$ y $\mathbf{F}$ y se definen como:

Definición 2.3.3 (Distribución $\boldsymbol{\chi^2}$). *Sea* X *un vector aleatorio de* $n \times 1$*, con* $X \sim N(0, I_n)$*, entonces*

$$Y = X^TX \sim \chi^2_n$$

es decir, Y *se distribuye* $Ji-cuadrado$ *con* n *grados de libertad,* χ^2_n*. La función de densidad de una* χ^2_n *es,*

$$f(y) = \frac{y^{\frac{n-1}{2}}e^{-\frac{y}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \qquad para\ u > 0$$

donde $\Gamma(\frac{n}{2})$ *es la función gamma con argumento* $\frac{n}{2}$ *(Searle, 1971:47).*

Teorema 2.3.8. *La función generatriz de momentos de una variable* $Ji-cuadrado$ *con* n *grados de libertad,* χ^2_n*, es*

$$M(t) = (1-2t)^{-\frac{n}{2}}.$$

Demostración.

$$M_{X^TX}(t) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{tX^TX}\frac{1}{(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}X^TX}dx_1\cdots dx_n$$

$$M_{X^TX}(t) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}(1-2t)X^TX}dx_1\cdots dx_n$$

$$M_{X^TX}(t) = (1-2t)^{-\frac{n}{2}}\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{(1-2t)^{\frac{n}{2}}}{(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}(1-2t)X^TX}dx_1\cdots dx_n}_{=\ uno}$$

dado que las n integrales son una función de densidad Normal multivariada con media 0 *y varianza* I_n*, las n integrales son iguales a uno y se tiene que*

$$M_{X^TX}(t) = (1-2t)^{-\frac{n}{2}}.$$

■

Teorema 2.3.9. *Sea X un vector aleatorio que se distribuye Normal con media* $E(X) = \mu$ *y matriz de varianza covarianza* $V(X) = \sigma^2 I_n$*,* $X \sim N(\mu, \sigma^2 I_n)$ *y B una matriz simétrica positiva definida, entonces la forma cuadrática* X^TBX *tiene una distribución igual a* $X^TBX \sim \chi^2_n$ *si y sólo si:*

1. $BE(X) = 0$.
2. $BV(X)$ *es idempotente.*
3. $rango(B) = n$*, número grados de libertad.*

Dado el resultado anterior, ahora se quiere determinar cuando una forma cuadrática X^TBX, con $X \sim N(\mu, V)$, se distribuye χ^2.

Teorema 2.3.10. *La función generatriz de momentos de una variable* X^TBX *es*

$$M(t) = \mid I - 2tBV \mid^{-1/2} e^{-\frac{1}{2}\mu^T(I-(I-2tBV)^{-1})V^{-1}\mu}.$$

Lemma 2.3.11. *Para algún vector g y una matriz definida positiva W se tiene* [10]

$$(2\pi)^{n/2} \mid W \mid^{1/2} e^{\frac{1}{2}g^TWg} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}X^TW^{-1}X+g^TX} dx_1 \cdots dx_n.$$

Demostración. *La función generatriz de momentos de* X^TBX *es (Searle, 1971:55)*

$$M(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{tX^TBX} \frac{1}{(2\pi)^{n/2} \mid V \mid^{1/2}} e^{-\frac{1}{2}(X-\mu)^TV^{-1}(X-\mu)} dx_1 \cdots dx_n$$

$$M(t) = \frac{1}{(2\pi)^{n/2} \mid V \mid^{1/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}[(X-\mu)^TV^{-1}(X-\mu)-2tX^TBX]} dx_1 \cdots dx_n$$

factorizando términos en el exponente del integrando

$$M(t) = \frac{e^{-\frac{1}{2}\mu^TV^{-1}\mu}}{(2\pi)^{n/2} \mid V \mid^{1/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}X^T(I-2tBV)V^{-1}X+\mu^TV^{-1}X} dx_1 \cdots dx_n.$$

[10] Searle, 1971:54

Haciendo un cambio de constantes, $g^T = \mu^T V^{-1}$ *y* $W = V(I - 2tBV)^{-1}$ *y por el Lemma 2.3.11*

$$\begin{aligned} M(t) &= \frac{e^{-\frac{1}{2}\mu^T V^{-1}\mu}}{(2\pi)^{n/2} \mid V \mid^{1/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}X^T W^{-1} X + g^T X} dx_1 \cdots dx_n \\ &= \frac{e^{-\frac{1}{2}\mu^T V^{-1}\mu}}{(2\pi)^{n/2} \mid V \mid^{1/2}} (2\pi)^{n/2} \mid W \mid^{1/2} e^{\frac{1}{2}g^T W g} \end{aligned}$$

sustituyendo el valor de g^T *y* W

$$M(t) = \frac{e^{-\frac{1}{2}\mu^T V^{-1}\mu}}{\mid V \mid^{1/2}} \mid V(I - 2tBV)^{-1} \mid^{1/2} e^{\frac{1}{2}\mu^T V^{-1} V (I-2tBV)^{-1} V^{-1}\mu}$$

por tanto

$$M(t) = \mid I - 2tBV \mid^{-1/2} e^{-\frac{1}{2}\mu^T (I - (I-2tBV)^{-1}) V^{-1}\mu}.$$

■

Un resultado importante de esta distribución es que si $\overline{X}$ y S^2 son la media y la varianza de una muestra aleatoria de tamaño n de una población normal con media μ y varianza σ^2, entonces la variable aleatoria $\frac{(n-1)}{\sigma^2}S^2$ se distribuye χ^2_{n-1} (Freud, 281).

En la demostración del teorema 2.3.7 (ecuación 2.3) se tiene que la forma cuadrática asociada a $(n-1)S^2$ es $X^T F X$. Analogamente, se tiene que la forma cuadrática asociada a $\frac{(n-1)}{\sigma^2}S^2$ es

$$\frac{(n-1)}{\sigma^2}S^2 = \frac{1}{\sigma^2}(FX)^T FX = \frac{1}{\sigma^2}X^T FX$$

con $X \sim N(\mu, \sigma^2 I_n)$. A continuación, se procede a verificar (a través del teorema 2.3.9) que $\frac{(n-1)}{\sigma^2}S^2$ se distribuye χ^2_{n-1}.

Sea $F = \frac{1}{\sigma^2}\left(I - \frac{1}{n}J_{n\times n}\right)$, $E(X) = \mu$ y $V(X) = \sigma^2 I_n$, entonces

1. $FE(X) = 0$.

$$FE(X) = \frac{\mu}{\sigma^2}\left(I_n - \frac{1}{n}J_{n\times n}\right)\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix} = \begin{pmatrix}\frac{\mu}{\sigma^2}\\\frac{\mu}{\sigma^2}\\\vdots\\\frac{\mu}{\sigma^2}\end{pmatrix} - \begin{pmatrix}\frac{\mu}{\sigma^2}\\\frac{\mu}{\sigma^2}\\\vdots\\\frac{\mu}{\sigma^2}\end{pmatrix} = 0$$

2. $FV(X)$ es idempotente, es decir, $(FV(X))^2 = FV(X)$.

$$FV(X) = \frac{1}{\sigma^2}\left(I_n - \frac{1}{n}J_{n\times n}\right)\sigma^2 I = I_n - \frac{1}{n}J_{n\times n}$$

y por el ejemplo 2.1.5 se sabe que

$$(FV(X))^2 = \left(I_n - \frac{1}{n}J_{n\times n}\right)\left(I_n - \frac{1}{n}J_{n\times n}\right) = I_n - \frac{1}{n}J_{n\times n}.$$

Por lo tanto, $FV(X)$ es idempotente.

3. El *rango*(F)= número de grados de libertad.
   El rango de una matriz idempotente es su traza, por tanto, para determinar los grados de libertad de F sólo se necesita saber cual es su traza.

$$rango(F) = tr\left(I_n - \frac{1}{n}J_{n\times n}\right) = n - 1.$$

En conclusión,

$$\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2_{n-1}.$$

Definición 2.3.4 (**Distribución t**). *Sea X una variable aleatoria tal que $X \sim N(0,1)$ y Y una variable aleatoria con distribución χ^2 con n grados de libertad; Si X y Y independientes, entonces*

$$T = \frac{X}{\sqrt{Y/n}} \sim t_n$$

es decir, t tiene una distribución t con n grados de libertad.

La función de densidad t es

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\ \Gamma(\frac{n}{2})}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \qquad para - \infty < t < \infty,$$

con media cero y varianza $n/(n-2)$ (Searle, 1971:48).

Definición 2.3.5 (**Distribución F**). *Sean Y_1 y Y_2 variables aleatorias independientes con una distribución χ^2 con n_1 y n_2 grados de libertad, respectivamente, entonces*

$$W = \frac{Y_1/n_1}{Y_2/n_2} \sim F_{n_1,n_2}$$

es decir, W tiene una distribución F con n_1 y n_2 grados de libertad.

La función de densidad F es

$$f(w) = \frac{\Gamma(\frac{n_1+n_2}{2})n_1^{\frac{n_1}{2}}n_2^{\frac{n_2}{2}}w^{\frac{n_1}{2}-1}}{\Gamma(\frac{n_1}{2})\ \Gamma(\frac{n_2}{2})(n_2 + n_1 w)^{\frac{n_1+n_2}{2}}} \qquad para\ \ w > 0.$$

con media $n_2/(n_2-2)$ y varianza $2n_2^2[(n_1+n_2-2)/n_1]/(n_2-2)^2(n_2-4)$.

En este capítulo se presentaron conceptos básicos de algebra con matrices y de la distribución Normal cuando se tienen n variables aleatorias. En el siguiente capítulo se aborda el tema de la regresión lineal múltiple.

Capítulo 3

Regresión lineal

En este capítulo se aborda el tema de la regresión lineal, la cual es una técnica estadística que estudia la descripción y evaluación de la posible asociación entre una variable de interés Y (variable dependiente) y una o más variables explicativas o independientes $X_1, X_2, \cdots, X_p$. Esta asociación es representada mediante una función lineal, es decir, la predicción de Y como función de $X_1, X_2, \cdots, X_p$ es de forma lineal.

3.1 Especificación del modelo de regresión lineal

El modelo clásico de regresión lineal más simple es cuando se tiene una variable dependiente, Y, con una única variable explicativa, X, y el modelo de asociación propuesto es

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{con} \quad i = 1, 2, \ldots, n,$$

donde β_0 y β_1 son parámetros desconocidos y ε_i el error aleatorio o perturbación. Este modelo es lineal en los parámetros dado que β_0 y β_1 no se encuentran elevados a una potencia mayor a uno, ni están divididos o multiplicados por ningún otro parámetro, ni son funciones no lineales y ε_i se define como la desviación de Y_i alrededor de su valor esperado, esta variable incluye toda la información de aquellas otras variables que son omitidas o excluidas del modelo, pero que en conjunto afectan a Y_i, errores de medición en Y y la aleatoriedad de las respuestas humanas. El modelo clásico de regresión plantea algunos supuestos acerca del error aleatorio ε_i (Maddala, 1996:73-74):

a) $E(\varepsilon_i) = 0$ para todo i.
b) $V(\varepsilon_i) = \sigma^2$ para todo i.
c) $Cov(\varepsilon_i, \varepsilon_j) = 0$, esto es, ε_i y ε_j son no autocorrelacionados $i \neq j$.
d) $Cov(\varepsilon_i, X_j) = 0$, esto es, ε_i y X_j son independientes para todo ij.
e) Normalidad: ε_i está normalmente distribuido para todo i. Junto con los supuestos a, b y c, esto implica que los ε_i son independientes y tienen una distribución normal, con media cero y varianza común σ^2. Es decir,

$$\varepsilon_i \sim IN(0, \sigma^2) \qquad \text{para todo} \quad i.$$

Una justificación del supuesto de normalidad para ε_i se encuentra en el teorema del límite central en estadística, ya que gracias a él se puede demostrar que si existe un gran número de variables aleatorias independientes e idénticamente distribuidas entonces la distribución de su suma tiende a ser Normal a la medida que el número de tales variables se incrementa indefinidamente (Gujarati, 1997:101). Además, si se quiere hacer alguna inferencia estadística sobre los datos es necesario partir de alguna distribución.

En general, el modelo de regresión se puede extender a p variables explicativas con una variable dependiente, dando lugar al modelo de regresión lineal general de p variables, el cual se puede escribir en forma matricial como

$$Y = X\beta + \varepsilon \tag{3.1}$$

donde

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

El vector del error aleatorio, $\varepsilon = Y - X\beta$, (al igual que en el caso de regresión clásico) se distribuye Normal con media cero y varianza $\sigma^2 I_n$, $\varepsilon \sim N(0, \sigma^2 I_n)$. El 0 de esta distribución se debe entender como un vector de $n \times 1$ donde todos sus elementos son cero.

Dado que el análisis de regresión lineal tiene como finalidad estimar la función de regresión (ecuación 3.1), en la siguiente sección se utiliza el método de máxima verosimilitud para obtener los estimadores de parámetros desconocidos asociados al modelo.

3.2 Estimación por máxima verosimilitud de los parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\sigma^2}$

Considere el modelo $Y = X\beta + \varepsilon$ con $\varepsilon \sim N(0, \sigma^2 I_n)$. La función de verosimilitud de Y está dada por la función de densidad conjunta de las variables aleatorias de Y y de cada X_i, $i = 1, 2, \ldots, p$, donde

$$f(y_1, \ldots, y_n) = \frac{1}{(2\pi)^{n/2}\sigma^n}\, e^{-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)}$$

Si $Y_1, ..., Y_n$ son conocidos, pero se desconoce el valor de β y σ^2, entonces, se deben estimar los valores de β y σ^2. Para estimar los parámetros del modelo se va a utilizar el método de máxima verosimilitud.[1] Aplicando dicho método se tiene

$$\ln L(\beta, \sigma^2) = -\ln(2\pi)^{\frac{n}{2}} - \ln(\sigma^2)^{\frac{n}{2}} - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)$$

[1] Ver ejemplos 1.2.1, 1.2.2 y 1.2.3

derivando parcialmente con respecto a β y σ^2 respectivamente, se obtiene

$$\begin{aligned}\frac{\partial \ln L(\beta,\sigma^2)}{\partial \beta} &= -\frac{1}{\sigma^2}X^T(Y-X\beta)\\ \frac{\partial \ln L(\beta,\sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}(Y-X\beta)^T(Y-X\beta)\end{aligned}$$

igualando a cero dichas parciales para encontrar el punto crítico se encuentra que el estimador para β y σ^2 respectivamente, es

$$\begin{aligned}\hat{\beta} &= (X^TX)^{-1}X^TY\\ \hat{\sigma^2} &= \frac{1}{n}(Y-X\hat{\beta})^T(Y-X\hat{\beta})\end{aligned}$$

donde $\hat{\beta}$ y $\hat{\sigma^2}$ son los valores que maximizan a la función de verosimilitud, ya que

$$\frac{\partial^2 \ln L(\beta,\sigma^2)}{\partial \beta^2} = \frac{1}{\sigma^2}X^TX$$

$$\frac{\partial^2 \ln L(\beta,\sigma^2)}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}(Y-X\beta)^T(Y-X\beta)$$

sutituyendo el valor estimado de β y σ^2 en la derivadas

$$\frac{\partial^2 \ln L(\beta,\sigma^2)}{\partial \beta^2} = -\frac{nX^TX}{(Y-X(X^TX)^{-1}X^TY)^T(Y-X(X^TX)^{-1}X^TY)} < 0$$

$$\frac{\partial^2 \ln L(\beta,\sigma^2)}{\partial (\sigma^2)^2} = -\frac{n^3}{[(Y-X(X^TX)^{-1}X^TY)^T(Y-X(X^TX)^{-1}X^TY)]^2} < 0$$

por tanto, hay un máximo en ambos casos.

Geométricamente, para encontrar el valor de β que maximiza la función de verosimilitud, se busca minimizar el exponente $(Y-X\beta)^T(Y-X\beta)$, puesto que es negativo dentro de la función de densidad conjunta de Y, por tanto, el problema se "reduce" a resolver

$$\min_{\beta}(Y-X\beta)^T(Y-X\beta).$$

Note que $(Y-X\beta)^T(Y-X\beta) = \|Y-X\beta\|^2$. Entonces, se busca el vector β tal que $Y-X\beta$ tenga norma mínima. Observe que

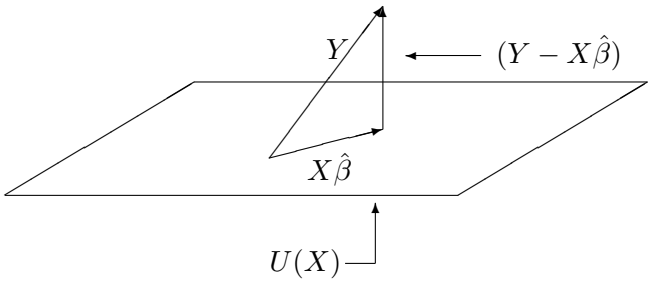
$$U = \{u | u = X\beta,\ \forall\ \beta \in \mathbb{R}^{p+1}\}$$

es el subespacio generado por las columnas de X ya que

$$X\beta = \beta_0\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix} + \beta_1\begin{pmatrix}X_{11}\\X_{21}\\\vdots\\X_{n1}\end{pmatrix} + ... + \beta_n\begin{pmatrix}X_{1p}\\X_{2p}\\\vdots\\X_{np}\end{pmatrix}$$

$X\beta \in U$ y $Y-X\beta$ tiene norma mínima sólo cuando $Y-X\beta$ es ortogonal a U, lo que significa que es ortogonal a las columnas de X.

Plano del subespacio generado por X



Dicha norma se obtiene de una proyección ortogonal (producto interior igual a cero), por lo cual se tiene que

$$X^T(Y - X\beta) = 0$$

o bien

$$\begin{aligned} X^TY - X^TX\beta &= 0 \\ X^TY &= X^TX\beta. \end{aligned}$$

A esta expresión se le conoce como "las ecuaciones normales". Cuando X^TX es no singular se tiene una solución única para el vector β dado por

$$\hat{\beta} = (X^TX)^{-1}X^TY.$$

En conclusión, los estimadores de máxima verosimilitud para β y σ^2 son: [2]

$$\begin{aligned} \hat{\beta} &= (X^TX)^{-1}X^TY \\ \hat{\sigma^2} &= \frac{1}{n}(Y - X\hat{\beta})^T(Y - X\hat{\beta}). \end{aligned}$$

Con estos estimadores se obtiene una función de regresión estimada

$$\hat{Y} = X\hat{\beta} \tag{3.2}$$

donde

$$\hat{Y} = \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}.$$

La ecuación 3.2 se utiliza para pronosticar el comportamiento de la variable Y. Por lo tanto, la ecuación de regresión 3.1 se puede expresar como:

$$Y = \hat{Y} + \hat{\varepsilon}.$$

[2] $E(\hat{\sigma^2}) = E(\frac{1}{n}(Y - X\hat{\beta})^T(Y - X\hat{\beta})) = \frac{1}{n}E((Y - X\hat{\beta})^T(Y - X\hat{\beta})) = \frac{1}{n}\sigma^2$. Por tanto $\hat{\sigma^2}$ no es un estimador insesgado, el estimador insesgado es $\hat{\sigma^2} = \frac{1}{n-p-1}(Y - X\hat{\beta})^T(Y - X\hat{\beta})$.

En la siguiente sección se determina como se distribuyen los estimadores de β y σ^2, ya que ello permite hacer alguna inferencia (a través de intervalos de confianza o pruebas de hipótesis) sobre que tan cerca están estos valores estimados de los valores reales.

3.3 Distribución de $\hat{\beta}$ y $\hat{\sigma^2}$.

Como se observa $\hat{\beta} = (X^TX)^{-1}X^TY$ se puede escribir como una forma lineal

$$\hat{\beta} = AY \tag{3.3}$$

con $A = (X^TX)^{-1}X^T$. Dado que una forma lineal de un vector Normal se distribuye Normal y que $\varepsilon = Y - X\beta$ con $\varepsilon \sim N(0, \sigma^2 I_n)$, se tiene

$$\begin{aligned} E(\hat{\beta}) &= E((X^TX)^{-1}X^TY) = E((X^TX)^{-1}X^T(X\beta + \varepsilon)) \\ &= E((X^TX)^{-1}X^TX\beta) + E((X^TX)^{-1}X^T\varepsilon) \\ &= E(\beta) + (X^TX)^{-1}X^TE(\varepsilon) \\ E(\hat{\beta}) &= \beta \end{aligned}$$

lo que indica que $\hat{\beta}$ es insesgado.

$V(\hat{\beta}) = E(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T$

$$\begin{aligned} (\hat{\beta} - \beta) &= (X^TX)^{-1}X^TY - (X^TX)^{-1}X^TX\beta \\ &= (X^TX)^{-1}X^T(Y - X\beta) \end{aligned}$$

$$\begin{aligned} V(\hat{\beta}) &= E[(X^TX)^{-1}X^T(Y - X\beta)(Y - X\beta)^TX(X^TX)^{-1}] \\ &= (X^TX)^{-1}X^TE[(Y - X\beta)(Y - X\beta)^T]X(X^TX)^{-1} \end{aligned}$$

donde $V(\varepsilon) = E[(Y - X\beta)(Y - X\beta)^T] = E(\varepsilon\varepsilon^T) = \sigma^2 I_n$

$$\begin{aligned} V(\hat{\beta}) &= (X^TX)^{-1}X^T\sigma^2 I_n X(X^TX)^{-1} \\ V(\hat{\beta}) &= \sigma^2(X^TX)^{-1} \end{aligned}$$

entonces se tiene que $\hat{\beta} = AY$ se distribuye Normal con media β y varianza $\sigma^2(X^TX)^{-1}$,

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

Observe que Y también se distribuye Normal con

$$\begin{aligned} E(Y) &= E(X\beta + \varepsilon) \\ &= E(X\beta) + E(\varepsilon) \\ &= X\beta \end{aligned}$$

$$\begin{aligned} V(Y) &= E[(Y - X\beta)(Y - X\beta)^T] \\ &= E(\varepsilon\varepsilon^T) \\ &= \sigma^2 I_n. \end{aligned}$$

Es decir,

$$Y \sim N(X\beta, \sigma^2 I_n).$$

Por otro lado, $\hat{\sigma^2}$ se puede expresar como una forma cuadrática, es decir

$$\begin{aligned} \hat{\sigma^2} &= \frac{1}{n}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) \\ &= \frac{1}{n}Y^T(I - X(X^TX)^{-1}X^T)Y \end{aligned}$$

$$\hat{\sigma^2} = \frac{1}{n}Y^TBY \tag{3.4}$$

con $B = I_n - X(X^TX)^{-1}X^T$ idempotente.

A continuación se verifica si la forma cuadrática Y^TBY cumple con las condiciones del teorema 2.3.9 para distribuirse como una χ^2.

Sea $Y \sim (X\beta, \sigma^2 I_n)$ y $B = I_n - X(X^TX)^{-1}X^T$, entonces

1. $BE(Y) = 0$.

$$\begin{aligned} BE(Y) &= (I_n - X(X^TX)^{-1}X^T)X\beta = 0 \\ &= X\beta - X(X^TX)^{-1}X^TX\beta = 0 \\ &= X\beta - X\beta = 0 \end{aligned}$$

2. $BV(Y)$ idempotente.

$$BV(Y) = \frac{1}{\sigma^2}(I_n - X(X^TX)^{-1}X^T)\sigma^2 I_n = I_n - X(X^TX)^{-1}X^T$$

$$\begin{aligned} (I_n - X(X^TX)^{-1}X^T)(I_n - X(X^TX)^{-1}X^T) &= I_n - X(X^TX)^{-1}X^T \\ (I_n - 2X(X^TX)^{-1}X^T + X(X^TX)^{-1}X^T) &= I_n - X(X^TX)^{-1}X^T \\ I_n - X(X^TX)^{-1}X^T &= I_n - X(X^TX)^{-1}X^T \end{aligned}$$

   Por lo tanto, $BV(Y)$ es idempotente.

3. El $rango(B)$= número de grados de libertad.
   Se sabe que el rango de una matriz idempotente es igual a su traza. Entonces

$$rango(B) = tr(B) = tr(I_n) - tr(X(X^TX)^{-1}X^T)$$

   y como $tr(AB) = tr(BA)$

$$\begin{aligned} rango(B) &= tr(I_n) - tr(X^TX(X^TX)^{-1}) \\ &= tr(I_n) - tr(I_D) \\ &= n - (p+1) \end{aligned}$$

   donde $I_D = X^TX(X^TX)^{-1}$.

Puesto que se cumplen las tres propiedades anteriores se tiene que

$$n\hat{\sigma^2} = \frac{1}{\sigma^2} Y^T B Y \sim \chi^2_{n-p-1}.$$

En conclusión

$$\begin{aligned} \hat{\beta} &= AY \sim N(\beta, \sigma^2 (X^T X)^{-1}) \\ n\hat{\sigma^2} &= \frac{1}{\sigma^2} Y^T B Y \sim \chi^2_{n-p-1}. \end{aligned}$$

3.4 Independencia de $\hat{\beta}$ y $\hat{\sigma^2}$

Por el teorema 2.3.5 se sabe que una forma lineal AX y una forma cuadrática $X^T BX$ de un vector aleatorio X son independientes si y sólo si $AVB^T = 0$, donde A es la matriz asociada a la forma lineal y B la matriz simétrica asociada a la forma cuadrática. En la sección 3.3 se vió que $\hat{\beta}$ y $\hat{\sigma^2}$ se pueden escribir como AY y $\frac{1}{n} Y^T BY$ respectivamente. Por consiguiente, para probar que $\hat{\beta}$ y $\hat{\sigma^2}$ son independientes se debe satisfacer que las formas lineales asociadas a cada estimador satisfagan que $AVB^T = 0$.

Sea $A = (X^T X)^{-1} X^T$, $B = I_n - X(X^T X)^{-1} X^T$ y $Y \sim (X\beta, \sigma^2 I_n)$, donde $V = \sigma^2 I_n$, por ende

$$\begin{aligned} AVB^T &= (X^T X)^{-1} X^T (\sigma^2 I_n)(I_n - X(X^T X)^{-1} X^T) \\ &= (X^T X)^{-1} X^T \sigma^2 I_n - (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} X^T \\ &= \sigma^2 [(X^T X)^{-1} X^T - (X^T X)^{-1} X^T X (X^T X)^{-1} X^T] \\ &= \sigma^2 [(X^T X)^{-1} X^T - (X^T X)^{-1} X^T] = 0 \end{aligned}$$

Por tanto, $AVB^T = 0$, lo que implica que $\hat{\beta}$ y $\hat{\sigma^2}$ son independientes.

3.5 Descomposición de la variación en Y

La variación de los valores observados de Y se puede deber a dos fuentes, una atribuible a la función de regresión estimada y la otra atribuible a la variación del error aleatorio. Es decir, la variación total de Y se puede expresar mediante la siguiente expresión:

$$(Y - \overline{Y})^T (Y - \overline{Y}) = (\hat{Y} - \overline{Y})^T (\hat{Y} - \overline{Y}) + (Y - \hat{Y})^T (Y - \hat{Y})$$

o equivalentemente

$$SCT = SCR + SCE$$

donde

SCT (*Suma de cuadrados totales*): Mide la variación total de los valores observados de la variable Y con respecto a su media $\overline{Y}$.

$$SCT = (Y - \overline{Y})^T (Y - \overline{Y}) = Y^T \left(I_n - \frac{1}{n} J_{n\times n} \right) Y = Y^T F Y \tag{3.5}$$

SCR (*Suma de cuadrados de la regresión*): Mide la variación debida al modelo propuesto de regresión estimado, o sea, la diferencia promedio entre los los valores estimados $\hat{Y}_i$ y la media $\overline{Y}$ del modelo.

$$SCR = (\hat{Y} - \overline{Y})^T (\hat{Y} - \overline{Y}) = Y^T \left(X(X^T X)^{-1} X^T - \frac{1}{n} J_{n\times n} \right) Y = Y^T C Y \tag{3.6}$$

SCE (*Suma de cuadrados del error*): Mide la variación aleatoria, esto es, la diferencia promedio entre los datos observados y los valores estimados del modelo.

$$SCE = (Y - \hat{Y})^T (Y - \hat{Y}) = Y^T \left(I_n - X(X^T X)^{-1} X^T \right) Y = Y^T B Y. \tag{3.7}$$

Estas tres sumas satisfacen la igualdad $SCT = SCR + SCE$,

$$\begin{aligned} Y^T \left(I_n - \frac{1}{n} J_{n\times n} \right) Y &= Y^T \left(X(X^T X)^{-1} X^T - \frac{1}{n} J_{n\times n} \right) Y + Y^T \left(I_n - X(X^T X)^{-1} X^T \right) Y \\ &= Y^T X (X^T X)^{-1} X^T Y - Y^T \frac{1}{n} J_{n\times n} Y + Y^T Y - Y^T X (X^T X)^{-1} X^T Y \\ &= Y^T Y - Y^T \frac{1}{n} J_{n\times n} Y \\ &= Y^T \left(I_n - \frac{1}{n} J_{n\times n} \right) Y \end{aligned}$$

Esta relación se puede interpretar como que la variación total de Y se descompone en la variación dada por el modelo de regresión y la variación aleatoria. Si SCE es pequeña en comparación de SCR, se tiene en consecuencia que el modelo propuesto es bueno, porque da información relevante en la variación de los datos observados.

En este punto, cabe preguntar por la bondad del modelo estimado, es decir, se desea saber qué tan bien la ecuación estimada, $\hat{Y}$, se ajusta a los datos observados de Y, para esta tarea se puede utilizar el *coeficiente de determinación*: es una medida que nos indica que tan bien la ecuación de regresión se ajusta a los datos, efectuar una *prueba de significancia* o una *prueba de falta de ajuste*. Conceptos que se desarrollan a continuación.

3.6 Coeficiente de determinación $\mathbf{R^2}$

Este es un valor que nos indica el porcentaje de variación que se debe al modelo de regresión y se calcula como

$$R^2 = \frac{SCR}{SCT}.$$

R^2 toma valores positivos, menores que uno y de acuerdo a estos valores se puede interpretar que

1. Cuanto R^2 está cerca de uno, mejor será el ajuste.

2. Si R^2 está muy cerca a cero, no se puede dar una conclusión igual al caso anterior, pues esto se puede deber a dos causas diferentes:

   - Primero, puede indicar que la variable Y no esta relacionada con las variables explicativas $X_1, X_2, \cdots, X_p$ y que el modelo correcto es $Y = \beta_0 + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$. Con esto, se puede prescindir de la información dada por $X_1, X_2, \cdots, X_p$.
   - Segundo, puede ser que Y esté relacionada con $X_1, X_2, \cdots, X_p$, pero mediante un modelo diferente al propuesto. Un análisis de residuales puede ayudar a concluir, pues permite ver posibles tendencias entre los valores reales y los arrojados por el modelo.

3.7 Prueba de significancia conjunta.

3.7.1 Región crítica.

Sea el modelo de regresión lineal general con p variables explicativas como se definió en la ecuación 3.1, $Y = X\beta + \varepsilon$ con $\varepsilon \sim N(0, \sigma^2 I_n)$. La prueba de significancia o prueba de la regresión consiste en contrastar las hipótesis

$$\begin{aligned} H_0 &: \beta_1 = 0, \beta_2 = 0, \ldots, \beta_p = 0 \\ H_1 &: \text{alguna } \beta_i \neq 0 \quad \text{para } i = 1, \ldots, p. \end{aligned}$$

Bajo H_0 se tiene el modelo

$$Y = X\beta + \varepsilon \qquad \text{que es equivalente a} \qquad Y = \beta^* + \varepsilon$$

donde β^* es un vector de $n \times 1$ donde todos sus elementos son iguales a β_0, lo que significa que Y no esta asociado a las variables explicativas.
Mientras que el modelo bajo H_1 es

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

La región de rechazo se obtiene utilizando la prueba de razón de verosimilitud, esta prueba consiste en obtener para cada hipótesis el máximo de la función de verosimilitud, $\max_{H_0} L(\beta^*, \sigma^2)$ y $\max_{H_1} L(\beta, \sigma^2)$ respectivamente, y después obtener la razón de estas dos verosimilitudes. Es decir

$$\frac{\max_{H_0} L(\beta^*, \sigma^2)}{\max_{H_1} L(\beta, \sigma^2)} = \frac{\max_{H_0} \frac{1}{(2\pi)^{n/2}\sigma^n}\ e^{-\frac{1}{2\sigma^2}(Y-\beta^*)^T(Y-\beta^*)}}{\max_{H_1} \frac{1}{(2\pi)^{n/2}\sigma^n}\ e^{-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)}} \leq \lambda$$

Para el numerador se debe encontrar

$$\max_{H_0} L(\beta^*, \sigma^2) = \max_{H_0} \frac{1}{(2\pi)^{n/2}\sigma^n}\ e^{-\frac{1}{2\sigma^2}(Y-\beta^*)^T(Y-\beta^*)}.$$

Obteniendo el logaritmo la expresión queda como

$$\ln L(\beta^*, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(Y-\beta^*)^T(Y-\beta^*)$$

derivando parcialmente $\ln L(\beta^*, \sigma^2)$ con respecto a β^* y a σ^2 e igualando a cero, se obtiene el sistema de ecuaciones lineales con dos incógnitas[3]

$$\frac{\partial \ln L(\beta^*, \sigma^2)}{\partial \beta^*} = \frac{1}{\sigma^2}\frac{1}{n}J_{n\times n}(Y-\beta^*) = 0 \tag{3.8}$$

$$\frac{\partial \ln L(\beta^*, \sigma^2)}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y-\beta^*)^T(Y-\beta^*) = 0. \tag{3.9}$$

Despejando a β^* de 3.8 y a σ^2 de 3.9, se tiene que

$$\hat{\beta^*} = \frac{1}{n}J_{n\times n}Y = \overline{Y}$$

$$\hat{\sigma^2} = \frac{1}{n}(Y-\overline{Y})^T(Y-\overline{Y})$$

de lo anterior se obtiene

$$\max_{H_0} L(\beta^*, \sigma^2) = \frac{(2\pi)^{-n/2}\ n^{n/2}\ e^{-n/2}}{\big((Y-\overline{Y})^T(Y-\overline{Y})\big)^{n/2}}.$$

Para el denominador se tiene que

$$\max_{H_1} L(\beta, \sigma^2) = \max_{H_1} \frac{1}{(2\pi)^{n/2}\sigma^n}\ e^{-\frac{1}{2\sigma^2}}(Y-X\beta)^T(Y-X\beta).$$

Aplicando logaritmos

$$\ln L(\beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)$$

derivando parcialmente $\ln L(\beta, \sigma^2)$ con respecto a β y a σ^2 e igualando a cero, se obtiene el sistema

$$\frac{\partial \ln L(\beta, \sigma^2)}{\partial \beta} = \frac{1}{\sigma^2}X^T(Y-X\beta) = 0 \tag{3.10}$$

$$\frac{\partial \ln L(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y-X\beta)^T(Y-X\beta) = 0. \tag{3.11}$$

Despejando a β de 3.10 y a σ^2 de 3.11, se tiene

$$\hat{\beta} = (X^TX)^{-1}X^TY.$$

[3]Note que β^* es igual que decir $\frac{1}{n}J_{n\times n}\beta^*$, por tanto $(Y-\beta^*) = (Y-\frac{1}{n}J_{n\times n}\beta^*)$. Entonces, $\frac{\partial \frac{1}{n}J_{n\times n}\beta^*}{\partial \beta^*} = \frac{1}{n}J_{n\times n}$.

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

Dado que $\hat{Y} = X\hat{\beta}$, entonces

$$\hat{\sigma^2} = \frac{1}{n}(Y - \hat{Y})^T(Y - \hat{Y}).$$

Por tanto,

$$\max_{H_1} L(\beta, \sigma^2) = \frac{(2\pi)^{-n/2}\ n^{n/2}\ e^{-n/2}}{\Big((Y - \hat{Y})^T(Y - \hat{Y})\Big)^{n/2}}.$$

Con lo que finalmente la razón de verosimilitud queda como

$$\frac{(Y - \hat{Y})^T(Y - \hat{Y})}{(Y - \overline{Y})^T(Y - \overline{Y})} \leq \lambda. \tag{3.12}$$

Por las ecuaciónes 3.7 y 3.5, la ecuación 3.12 es equivalente a

$$\frac{SCE}{SCT} \leq \lambda.$$

SCE y SCT no son independientes, pero ya que $SCT = SCR + SCE$ la expresión anterior se puede expresar como

$$\frac{SCE}{SCE + SCR} \leq \lambda$$

$$\begin{aligned} \frac{SCE + SCR}{SCE} &\geq \frac{1}{\lambda} \\ 1 + \frac{SCR}{SCE} &\geq \frac{1}{\lambda} \\ \frac{SCR}{SCE} &\geq \frac{1}{\lambda} - 1 \end{aligned}$$

si definimos a $\dfrac{1}{\lambda} - 1$ como λ^*, entonces

$$\frac{SCR}{SCE} \geq \lambda^*$$

donde SCE y SCR se demuestra más adelante que son independientes, por lo tanto, la región de rechazo de esta prueba esta dada por $\dfrac{SCR}{SCE} \geq \lambda^*$. A continuación se verá como se distribuyen estas dos sumas, con la finalidad de determinar como se distribuye $\dfrac{SCR}{SCE}$.

3.7.2 Distribuciones de las sumas de cuadrados

Debido a que SCR y SCE se pueden espresar como una forma cuadrática

$$SCR = Y^T(X(X^TX)^{-1}X^T - \frac{1}{n}J_{n\times n})Y = Y^TCY$$

con $C = X(X^TX)^{-1}X^T - \frac{1}{n}J_{n\times n}$ y

$$SCE = Y^T(I_n - X(X^TX)^{-1}X^T)Y = Y^TBY$$

con $B = I_n - X(X^TX)^{-1}X^T$ y dado que $Y \sim N(X\beta, \sigma^2 I_n)$, en seguida se inspeccionara si SCE y SCR se distribuyen como una χ^2.

Note que SCE es nada menos que $n\hat{\sigma^2}$ y en la sección 3.3 a partir de la ecuación 3.4 se deduce que $n\hat{\sigma^2} = \frac{1}{\sigma^2}Y^TBY \sim \chi^2_{n-p-1}$. Ahora, sólo se necesita verificar que SCR se distribuye como una χ^2, para dicha tarea se utiliza el teorema 2.3.9. Antes de aplicar el teorema se debe tener presente que SCR tiene como función de densidad conjunta (función de verosimilitud) a la expresada bajo H_0.

Sea $SCR = Y^TCY$ con $C = X(X^TX)^{-1}X^T - \frac{1}{n}J_{n\times n}$, $Y \sim N(X\beta, \sigma^2 I_n)$, $E(Y) = X\beta = \beta^*$ y $V(X) = \sigma^2 I_n$.

1. $CE(Y) = 0$ entonces,

$$\begin{aligned}\left(X(X^TX)^{-1}X^T - \frac{1}{n}J_{n\times n}\right)X\beta &= 0\\ X(X^TX)^{-1}X^TX\beta - \frac{1}{n}J_{n\times n}X\beta &= 0\\ X\beta - \frac{1}{n}J_{n\times n}X\beta &= 0\end{aligned}$$

   Bajo H_0, $X\beta = \beta^*$, recuerde que β^* es un vector de $n \times 1$ donde todos sus elementos son β_0, entonces

$$\begin{aligned}\beta^* - \frac{1}{n}J_{n\times n}\beta^* &= 0\\ \beta^* - \frac{1}{n}n\beta^* &= 0\\ \beta^* - \beta^* &= 0.\end{aligned}$$

2. $CV(Y)$ es idempotente

$$CV(Y) = \frac{1}{\sigma^2}\left(X(X^TX)^{-1}X^T - \frac{1}{n}J_{n\times n}\right)\sigma^2 I = X(X^TX)^{-1}X^T - \frac{1}{n}J_{n\times n}$$

$$[CV(Y)]^2 = CV(Y)$$

$$\begin{aligned}
\left(X(X^TX)^{-1}X^T - \frac{1}{n}J_{n\times n}\right)\left(X(X^TX)^{-1}X^T - \frac{1}{n}J_{n\times n}\right) &= \\
X(X^TX)^{-1}X^T - 2X(X^TX)^{-1}X^T\frac{1}{n}J_{n\times n} + \frac{1}{n^2}J^2_{n\times n} &= \\
X(X^TX)^{-1}X^T - \frac{2}{n}J_{n\times n} + \frac{1}{n}J_{n\times n} &= \\
X(X^TX)^{-1}X^T - \frac{1}{n}J_{n\times n} &= X(X^TX)^{-1}X^T - \frac{1}{n}J_{n\times n}
\end{aligned}$$

Por lo tanto, $CV(Y)$ es idempotente.

3. El $rango(C)$ = número de grados de libertad.
   El rango de una matriz idempotente es igual a su traza.

$$\begin{aligned}
rango(C) &= tr(X(X^TX)^{-1}X^T) - tr(\frac{1}{n}J_{n\times n}) \\
&= tr(X^TX(X^TX)^{-1}) - tr(\frac{1}{n}J_{n\times n}) \\
&= (p+1) - 1 = p
\end{aligned}$$

Por lo tanto, puesto que las tres propiedades se cumplen $\frac{1}{\sigma^2}Y^TCY \sim \chi^2_p$.

En conclusión, SCE y SCR se distribuyen como una χ^2 con $n-p-1$ y p grados de libertad, respectivamente:

$$\frac{1}{\sigma^2}SCE \sim \chi^2_{n-p-1}, \qquad \frac{1}{\sigma^2}SCR \sim \chi^2_p.$$

Dado que SCE y SCR se distribuyen como χ^2, ahora sólo falta verificar que SCE y SCR sean independientes para afirmar que este estadístico tiene una distribución conocida.

3.7.3 Independencia entre las sumas de cuadrados

Utilizando las formas cuadráticas asociadas a estas dos sumas se prueba que la SCR y la SCE son independientes.

Sean B y C las matrices asociadas a SCE y SCR respectivamente, se tiene que estas sumas serán independientes si se cumple que $BVC = 0$, donde $V = \sigma^2 I_n$.

$$(I - X(X^TX)^{-1}X^T)\sigma^2 I_n(X(X^TX)^{-1}X^T - \frac{1}{n}J_{n\times n}) = 0$$

dado que $X(X^TX)^{-1}X^T$ es idempotente, se tiene

$$\sigma^2(X(X^TX)^{-1}X^T - \frac{1}{n}J_{n\times n} - X(X^TX)^{-1}X^T + X(X^TX)^{-1}X^T\frac{1}{n}J_{n\times n}) = 0$$
$$\sigma^2(X(X^TX)^{-1}X^T - \frac{1}{n}J_{n\times n} - X(X^TX)^{-1}X^T + \frac{1}{n}J_{n\times n}) = 0$$

por lo tanto, SCR y SCE son independientes.

Un concepto importante que se deriva de lo anterior es el llamado *cuadrado medio*. Es decir, dado que $\frac{1}{\sigma^2}SCR \sim \chi^2_p$, y puesto que la media de una distribución χ^2 son sus grados de libertad, entonces a la cantidad SCR entre sus grados de libertad se le conoce como el cuadrado medio de la regresión (CMR),

$$CMR = \frac{SCR}{p}$$

dado que

$$E\left(\frac{1}{\sigma^2}SCR\right) = p, \qquad E\left(\frac{SCR}{p}\right) = \sigma^2.$$

Por ende, para el caso de $\frac{1}{\sigma^2}SCE$, el cuadrado medio del error (CME) será,

$$CME = \frac{SCE}{n-p-1}$$

donde

$$E\left(\frac{1}{\sigma^2}SCE\right) = n-p-1, \qquad E\left(\frac{SCE}{n-p-1}\right) = \sigma^2.$$

A partir de la teoría de la distribuciones se sabe que la división entre dos χ^2 independientes se distribuye como una F (ver definición 2.3.5), por lo cual

$$\frac{CMR}{CME} = \frac{SCR/p}{SCE/n-p-1} \sim F_{p,\ n-p-1.}$$

Por tanto, el estadístico de prueba que determina si se rechaza o no H_0 es

$$\frac{SCR/p}{SCE/n-p-1} \sim F_{p,\ n-p-1.}$$

Otra forma de probar la bondad de ajuste del modelo de regresión estimado es utilizar la *prueba de falta de ajuste.*

3.8 Prueba de falta de ajuste.

Esta prueba consiste en separar la suma de cuadrados del error (SCE) en dos distintas fuentes de variación, en la suma de cuadrados del error puro ($SCEP$) y la suma de cuadrados de la falta de ajuste ($SCFA$). Dichas sumas satisfacen la relación

$$SCE = SCEP + SCFA$$

donde $SCEP$ representa la variación entre los datos debida a un error aleatorio y $SCFA$ representa la variación de los datos atribuible a que el modelo propuesto no es el adecuado.

Para calcular la $SCEP$ es necesario que existan datos repetidos, es decir, se tienen datos repetidos cuando para varias observaciones de Y se tiene al menos una misma observación de X. Esta suma de cuadrados representa un error puro porque si una observación de X es idéntica para dos o más observaciones de Y, sólo la existencia de una variación aleatoria en los datos puede influir en los resultados y proporcionar diferencias entre ellos y tiene $n - m$ grados de libertad (Draper, 1966:28).

Por otro lado, se tiene que la suma de cuadrados de la falta de ajuste ($SCFA$) está dada por

$$SCFA = SCE - SCEP$$

con $m - p - 1$ grados de libertad.

Las hipótesis para esta prueba estan dadas por:

H_0**:** El modelo propuesto es el adecuado.

H_1**:** El modelo propuesto no es el adecuado (hay una falta de ajuste).

El modelo de regresión asociado a H_0 es

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

y el modelo de regresión bajo H_1 es

$$Y = f(x_1, \cdots, x_n) + \varepsilon$$

donde $f(x_1, \cdots, x_n) \neq X\beta$.

El estadístico para esta prueba es[4]

$$F_c = \frac{SCFA/k - p - 1}{SCEP/n - k} \sim F_{k-p-1,n-k}$$

y el criterio para rechazar H_0 es

$$F_c > F_{k-p-1,n-k}$$

donde $F_{k-p-1,n-k}$ se busca en tablas para un nivel de significancia α elegido (Acuña, 2006:20-21).

Hasta aquí, se han dado las herramientas necesarias para el desarrollo del objetivo fundamental de este trabajo: el cambio estructural, tema que se aborda a continuación.

[4] k es el número de valores distintos de X para los cuales hay observaciones repetidas de Y, p es el número de variables explicativas y n es el tamaño de la muestra.

Capítulo 4

Cambio Estructural

Este capítulo esta dedicado al estudio del punto de cambio estructural, principalmente se busca determinar si en un modelo de regresión lineal en varias variables el cambio ha ocurrido o no y cuándo. Para ello, primero se revisan algunas de las pruebas que se utilizan para resolver el problema del cambio estructural. Segundo, se propone una prueba de hipótesis alternativa para detectar si ha ocurrido o no un cambio en un modelo de regresión lineal con p variables explicativas; utilizando la razón de verosimilitud se formula un estadístico de prueba para el cual se encuentra su distribución de probabilidad. En seguida, se desarrollan 2 algoritmos para determinar la región crítica de esta prueba y se plantea un posible estimador del punto de cambio utilizando el método de máxima verosimilitud. Por ultimo, se presenta una aplicación para la prueba propuesta sobre cambio estructural.

4.1 Antecedentes

El concepto cambio estructural no tiene una definición única, puesto que depende del enfoque en el cual se utilice,

- Para un económetra dedicado al estudio de series de tiempo, el cambio estructural se define como la modificación que se produce en los parámetros de los modelos de regresión que se utilizan para explicar la evolución temporal de una variable (Broemeling, 1987).

- Desde un análisis insumo-producto, el cambio estructural es un cambio en la estructura productiva, es decir, la estructura de la producción esta reflejada en la matriz de insumo-producto entre sectores y el cambio se basa en las diferencias entre la matriz de coeficientes técnicos en dos momentos diferentes del tiempo, explicando dichas diferencias en función de cambios en las convenciones estadísticas, en los gustos, en la tecnología, en los precios relativos, en la composición de los productos o en el grado de utilización de la capacidad productiva (Pulido, 1993).

- Desde el punto de vista de la Economía del Desarrollo, el cambio estructural es un proceso secuencial por medio del cual distintas estructuras económicas (producción, comercio

internacional, utilización de los factores,...) de un país subdesarrollado se van transformando hasta que el sector industrial desplaza a la agricultura como centro de gravedad de la actividad económica (Chenery, 1980).

Aunque el significado de cambio estructural no se entienda por igual para todos, ya que puede definirse distinto dependiendo del enfoque; las definiciones anteriores buscan un mismo objetivo y eso es: decidir si el cambio ha ocurrido o no. Por ende, en este trabajo el concepto de cambio estructural se define desde el enfoque de un modelo de regresión lineal, entendiéndose como cambio estructural a aquella alteración o modificación de los parámetros del modelo.

4.1.1 Prueba Chow

Uno de los métodos que se ha utilizado para comprobar que ha ocurrido un cambio estructural es conocido como la **Prueba Chow**.

"El test Chow[1] se utiliza cuando el investigador sospecha que el modelo al que responde una parte de la prueba es diferente al que sigue el resto de la muestra." Es decir, esta prueba parte del hecho de que **se conoce** en que momento sucede el cambio estructural, es decir, se supone que se puede dividir una muestra de tamaño n en dos submuestras independientes una de la otra de tamaño m y $n-m$ respectivamente, $n = m + n - m$, en donde el error de ambas submuestras presenta una distribución normal con media cero y varianza σ^2.

El modelo de regresión asociado a la existencia de cambio estructural es:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i \qquad i = 1, \cdots, m \\ Y_j &= \beta_0^* + \beta_1^* X_{1j} + \beta_2^* X_{2i} + \cdots + \beta_p^* X_{pi} + \varepsilon_i \qquad j = m+1, \cdots, n \end{aligned}$$

donde las pruebas de hipótesis son

$$\begin{aligned} H_0 &: \beta_0 = \beta_0^*, \beta_1 = \beta_1^*, \ldots, \beta_p = \beta_p^* \\ H_1 &: \text{alguna } \beta_i \neq \beta_i^* \text{ para } i = 0, 1, \ldots, p. \end{aligned}$$

El procedimiento para llevar a cabo dicha prueba en un modelo con p variables explicativas y n observaciones es:

- Estimar un modelo de regresión para la muestra completa de n observaciones, y denotar su Suma de Cuadrados del Error como SCE, la cual tiene $n - p - 1$ grados de libertad.
- Estimar un modelo para cada submuestra, donde la suma de cuadrados del error en cada caso se denota como SCE_{1_m} y SCE_{2_m} con $m - p - 1$ y $n - m - p - 1$ grados de libertad respectivamente. Los grados de libertad de $SCE_{1_m} + SCE_{2_m}$ son igual a $(m-p-1)+(n-m-p-1) = n-2(p+1)$, por lo que, los grados de libertad de $SCE - (SCE_{1_m} + SCE_{2_m})$ son $p+1$.

[1] Novales, 1997.

- Calcular el estadístico de prueba por medio de

$$F_c = \frac{(SCE - (SCE_{1_m} + SCE_{2_m}))/p+1}{(SCE_{1_m} + SCE_{2_m})/n - 2(p+1)} \sim F_{p+1,\ n-2(p+1)}$$

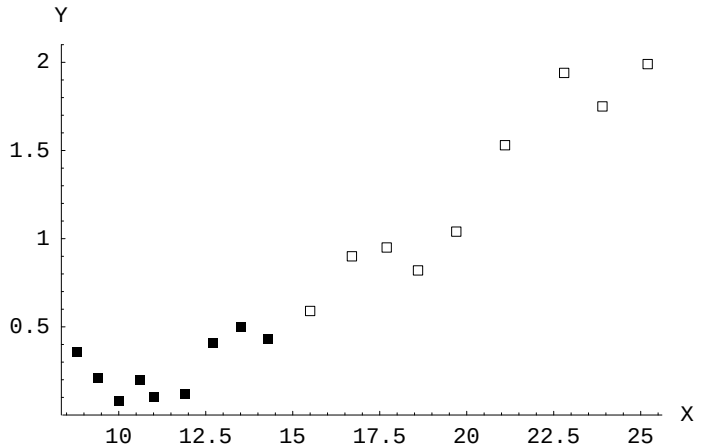
- Determinar la región de rechazo. Es decir, si $F_c > F_{p+1,\ n-2(p+1)}$, entonces se rechazara la hipótesis nula de ausencia de cambio estructural.

Ejemplo 4.1.1. *En la siguiente tabla se tiene información sobre el ahorro (Y) y el ingreso (X) del Reino Unido para el período de 1946 - 1963. Se desea saber si ha habido un cambio significativo en la función de ahorro durante el período posterior a la Segunda Guerra Mundial 1946 - 1954 y 1955 - 1963. Para ver si el cambio ha ocurrido, efectúese la prueba Chow (Gujarati, 1997:258-260).*

Ahorro e Ingreso, Reino Unido
1946 -1963 (millones de libras)

Año	*X*	*Y*	*Año*	*X*	*Y*
1946	*8.8*	*0.36*	*1955*	*15.5*	*0.59*
1947	*9.4*	*0.21*	*1956*	*16.7*	*0.90*
1948	*10.0*	*0.08*	*1957*	*17.7*	*0.95*
1949	*10.6*	*0.20*	*1958*	*18.6*	*0.82*
1950	*11.0*	*0.10*	*1959*	*19.7*	*1.04*
1951	*11.9*	*0.12*	*1960*	*21.1*	*1.53*
1952	*12.7*	*0.41*	*1961*	*22.8*	*1.94*
1953	*13.5*	*0.50*	*1962*	*23.9*	*1.75*
1954	*14.3*	*0.43*	*1963*	*25.2*	*1.99*

Fuente: Gujarati D. N. (1997). Econometría, p. 259



Solución. *Los pasos a seguir de la Prueba Chow son:*

- *Estimar el modelo de regresión para la muestra completa, n=18, y obtener su SCE.*

$$\begin{aligned} \hat{Y} &= -1.0821 + 0.1178X \\ SCE &= 0.5722 \quad \textit{con 16 grados de libertad.} \end{aligned}$$

- *Estimar para cada periodo el modelo de regresión asociado.*

  Período 1946 - 1954 ($m = 9$)

$$\begin{aligned} \hat{Y} &= -0.2622 + 0.0470X \\ SCE_{1_m} &= 0.1396 \quad \textit{con 7 grados de libertad.} \end{aligned}$$

  Período 1955 - 1963 ($n - m = 18 - 9 = 9$)

$$\begin{aligned} \hat{Y} &= -1.7502 + 0.1504X \\ SCE_{2_m} &= 0.1931 \quad \textit{con 7 grados de libertad.} \end{aligned}$$

- *Calcular el estadístico de prueba con $p + 1 = 2$*

$$\begin{aligned} F_c &= \frac{(SCE - (SCE_{1_m} + SCE_{2_m}))/p + 1}{(SCE_{1_m} + SCE_{2_m})/n - 2(p+1)} \\ &= \frac{0.2395/2}{0.3327/14} = 5.04 \end{aligned}$$

- *Determinar la región de rechazo $F_c > F_{p+1,\ n-2(p+1)}$*

  El valor de tablas de la distribución $F_{p+1,\ n-2(p+1)}$ para un nivel de significancia de $\alpha = 0.05$ con $p + 1 = 2$ y $n - 2(p + 1) = 14$ grados de libertad es 3.74.
  Puesto que $5.04 > 3.74$, *se puede concluir que la función de ahorro en los dos periodos es diferente.*

La característica principal de la prueba Chow es que se sospecha el momento en que ocurre el posible cambio, pero qué pasa cuando se sospecha que hubo un cambio, pero **se desconoce** en que momento ocurrió. Una posible solución es utilizar la prueba **CUSUM** la cual no necesita del conocimiento del momento del cambio estructural.

4.1.2 Prueba CUSUM

La prueba CUSUM esta basada en la suma acumulada de los residuos recursivos y éstos se definen como

$$\begin{aligned} r_t &= \frac{y_t - x_t\beta_{t-1}}{\sqrt{1 + x_t(X_{t-1}^T X_{t-1})^{-1}x_t^T}}, \qquad t = p+2, \ldots, n. \\ \beta_{t-1} &= (X_{t-1}^T X_{t-1})^{-1} X_{t-1}^T Y_{t-1} \end{aligned}$$

donde $x_t = (1, X_{t1}, \ldots, X_{tp})$ es el vector correspondiente a la t-íesima observación de los $p+1$ regresores, X_{t-1} es la matriz de datos de los regresores hasta la observación $t-1$ y β_{t-1} es el vector de estimadores mínimo cuadrático basado en las primeras $t-1$ observaciones. Los residuos recursivos siguen una distribución Normal con media cero y varianza constante, $r_t \sim N(0, \sigma^2)$.

La prueba CUSUM se define como

$$R_t = \sum_{j=(p+1)+1}^{t} \frac{r_j}{\hat{\sigma}}$$

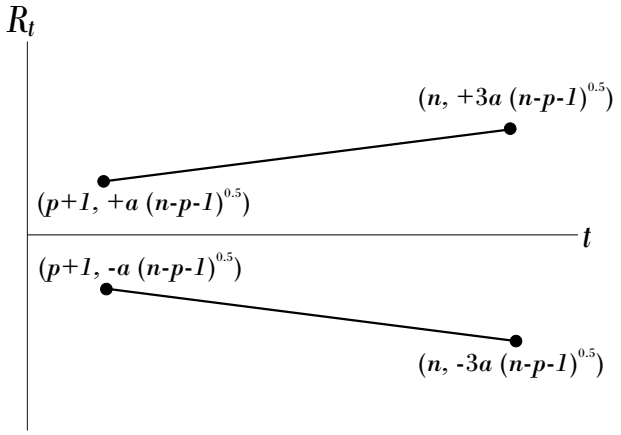
donde $\hat{\sigma^2} = \dfrac{1}{n-p-1} \displaystyle\sum_{j=(p+1)+1}^{n} (r_j - \overline{r}), \quad \overline{r} = \dfrac{1}{n-p-1} \displaystyle\sum_{j=(p+1)+1}^{n} r_j.$

La determinación del rechazo o no rechazo de cambio estructural se lleva a cabo por medio de un análisis visual de la evolución de R_t. La bandas de confianza de esta prueba se calculan como

$$(p+1, \pm a(n-p-1)^{0.5}) \text{ y } (n, \pm 3a(n-p-1)^{0.5}),$$

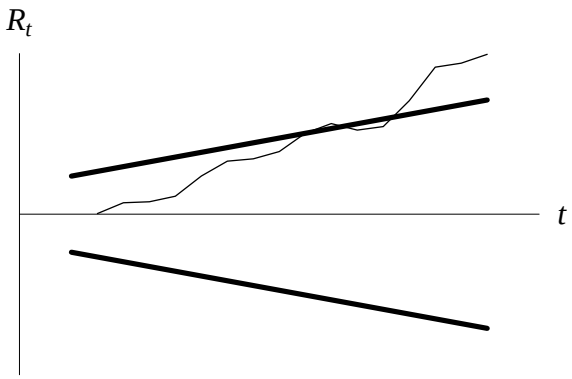
donde el valor de a depende del nivel de α escogido, si:

- α=0.01, a=1.143
- α=0.05, a=0.948
- α=0.10, a=0.850.



Si R_t no se sale de las bandas de confianza, entonces no hay evidencia de cambio estructural (Greene, 1999:309). A continuación se muestra esta prueba para los datos de ahorro ingreso utilizados en la prueba Chow,

Como se observa, la prueba CUSUM indica que en el periodo de observación hay evidencia de un cambio estructural, ya que R_t sale de las bandas de confianza.



4.1.3 Otras pruebas de cambio estructural

Este problema de sospechar que hay un cambio estructural, pero se desconoce el punto donde ocurre, ha sido abordado por diferentes autores, entre los cuales se pueden mencionar a:

Beckman y Cook (1979) consideran el modelo de regresión para una variable explicativa ($p = 1$), para probar la hipótesis $H : d_0 = d_1 = 0$,

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_i & \text{para } i = 1, 2, \ldots, t \geq 2 \\ (\beta_0 + d_0) + (\beta_1 + d_1) X_i & \text{para } i = t+1, \ldots, n \geq 5 \end{cases}$$

donde $X_i < X_j$ para $i < j$, t es desconocido. El estadístico de prueba que proponen es,

$$F = max_m(F_m) \qquad \text{para } \; p+1 \leq m \leq n-p-1,$$

con

$$F_m = \frac{(SCE - (SCE_{1_m} + SCE_{2_m}))/p + 1}{(SCE_{1_m} + SCE_{2_m})/n - 2(p+1)},$$

donde
SCE es la suma de cuadrados del error con las n observaciones,
SCE_{1_m} es la suma de cuadrados del error con las primeras m observaciones y,
SCE_{2_m} es la suma de cuadrados del error con las restantes $n - m$ observaciones.

Por medio de simulación construyen la función de distribución empírica de F y además para el 90 y 95 de los percentiles de esta distribución se obtienen, respectivamente, una tabla de valores críticos para la prueba $H : d_0 = d_1 = 0$.

Worsley (1983), para un modelo de regresión lineal con p variables explicativas

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i, & i = 1, \cdots, k, \\ \beta_0^* + \beta_1^* X_{1j} + \beta_2^* X_{2i} + \cdots + \beta_p^* X_{pi} + \varepsilon_i & i = k+1, \cdots, n, \end{cases}$$

donde $p+1 \leq k \leq n-p-1$, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)$ y ε_i se distribuyen Normal con varianza constante, prueba las hipótesis, $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$, donde $\delta = \beta - \beta^*$, y propone el siguiente estadístico,

$$F = \max_k \frac{U_k/p}{(Q - U_k)/(n-2p)},$$

con $U_k = \hat{\delta}^T \Sigma_k^{-1} \hat{\delta}$ la suma de cuadrados del error bajo H_1, $\hat{\delta} = \hat{\beta} - \hat{\beta}^*$,
$\Sigma_k \sigma^2 = [(X_k^T X_k)^{-1} + (X_k^{*T} X_k^*)^{-1}]\sigma^2$ la varianza bajo H_0, donde X^T y X_k^* son las matrices de dimensión $k \times (p+1)$ y $(n-k) \times (p+1)$ y Q es la suma de cuadrados del error bajo H_0.

El resultado principal de su estudio fue un límite sobre la función de la distribución bajo H_0 del estadístico de prueba, este límite es basado en la desigualdad de Bonferroni mejorada. También, da límites similares para una prueba de hipótesis donde el cambio se da sólo en el término constante de la regresión.

Horvath y Shao (1993) consideran el modelo de regresión lineal

$$Y_i = \begin{cases} X_i\beta + \varepsilon_i, & 1 \leq i \leq m, \\ \\ X_i\beta^* + \varepsilon_i, & m < i \leq n, \end{cases}$$

donde $X_i \in \mathbb{R}^d$, $\beta \neq \beta^*$ y ε_i para $i = 1, \ldots, n$ variables aleatorias independientes e idénticamente distribuidos con media cero y varianza σ^2. Estos autores quieren probar la hipótesis nula de que los coeficientes de regresión permanecen estables en el periodo de observación contra la alternativa la cual indica que los coeficientes cambian en un punto desconocido. Es decir

$$H_0 : k \geq n \quad \text{versus} \quad H_1 : 1 \leq k < n.$$

El estadístico de prueba que proponen es,

$$T_n = \max_{d \leq k \leq n-d} \left\{ \frac{1}{\hat{\sigma^2}} (\hat{\beta}_k - \hat{\beta}_k^*)^T H_k^{-1} (\hat{\beta}_k - \hat{\beta}_k^*) \right\}$$

donde $\hat{\beta}_k = (X_k^T X_k)^{-1} X_k^T Y_k$, $\hat{\beta}^* = (X_k^{*T} X_k^*)^{-1} X_k^{*T} Y_k^*$, $H_k = (X_k^T X_k)^{-1} + (X_k^{*T} X_k^*)^{-1}$ y
$\hat{\sigma_k^2} = \frac{1}{n} \left[\sum_{1 \leq i \leq k} (Y_i - X_i \hat{\beta}_k)^2 + \sum_{k < i \leq n} (Y_i - X_i \hat{\beta}_k^*)^2 \right]$ para $k = 1, \ldots, n$.

Ellos muestran que $(\hat{\beta}_k - \hat{\beta}_k^*) H_k^{-1} (\hat{\beta}_k - \hat{\beta}_k^*)$ alcanza su máximo en un vecindario de k y que en ese vecindario $\hat{\sigma_k^2}$ es asintoticamente consistente bajo H_1 así como bajo H_0. Rechazan H_0 para valores grandes de T_n, además de probar que T_n se distribuye asintótica bajo H_0.

Antoch y Hušcová (2001) consideran el modelo de regresión lineal

$$X_i = \begin{cases} \beta + \varepsilon_i, & \text{si } i \leq m, \\ \\ \beta + \delta_n + \varepsilon_i, & \text{si } i > m, \end{cases}$$

con $1 \leq m \leq n$, β y $\delta_n \neq 0$ y ε_i para $i = 1, \ldots, n$ variables aleatorias independientes e identicamente distribuidos con media cero y varianza positiva. Bajo las hipótesis

$$H_0 : m = n \quad \text{versus} \quad H_1 : m < n,$$

el estadístico de prueba que proponen es,

$$T_{n1}(R) = \max_{1<k<n} \left\{ \sqrt{\frac{n}{k(n-k)}} \frac{1}{\hat{\sigma_n}} \left| \sum_{i=1}^{k} (X_{R_i} - \overline{X}_n) \right| \right\}$$

donde $\hat{\sigma_n^2} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (X_i - \overline{X}_n)^2$ y R indica que el cálculo se hace sobre todas las posibles permutaciones de los datos muéstrales. La propuesta de estos autores consiste en encontrar T_{n1} para la muestra de datos original, después encontrar T_{n1} para cada una de todas las posibles permutaciones de la muestra, ordenar estos valores y extraer el 5% de los valores más pequeños. Se rechaza H_0 si T_{n1} con la muestra original se encuentra dentro del 5% de los valores más pequeños de la muestra permutada. Además de obtener que T_{n1} se distribuye asintóticamente

$$P(\sqrt{2 \log \log n T_{n1}(R)} \leq y + 2 \log \log n - \frac{1}{2} \log \log \log n - \frac{1}{2} \log \pi \mid X_1, \ldots, X_n)$$

$$\longrightarrow exp\{-2exp\{-y\}\}, \quad \text{casi seguramente.}$$

Vito Muggeo(2003) abordó el problema de estimación del punto de cambio, él considero un modelo no lineal y propuso un estimador del punto aplicando una técnica de linealización. El modelo de regresión

$$g(E(Y)) = \eta(X) + \beta h(z; m)$$

donde $\eta(X)$ es la función de variables explicativas, $g(E(Y))$ es la repuesta y $h(z; m)$ es una función no lineal de la variable Z con parámetro m, donde el término no lineal se calcula por medio de la aproximación de Taylor para un punto inicial $m^{(0)}$ conocido

$$h(z; m) \approx h(z; m^{(0)}) + (m - m^{(0)}) h'(z; m^{(0)})$$

con lo que el modelo queda como

$$g(E(Y)) = \eta(X) + \beta h(z; m^{(0)}) + \gamma h'(z; m^{(0)})$$

con $\gamma = \beta(m - m^{(0)})$, después, por medio de ML encuentra $\hat{\beta}$, $\hat{\gamma}$ y propone como estimador para el parámetro no lineal m a

$$\hat{m} = \frac{\hat{\gamma}}{\hat{\beta}} + m^{(0)}.$$

Cuando se obtiene este valor, entonces $h(z; m^{(0)})$, $h'(z; m^{(0)})$ son revaluadas, el modelo es reajustado y se obtiene un nuevo estimador de m; este proceso es iterado hasta que se encuentra una posible convergencia. Menciona que cuando el algoritmo converge, los estimadores ML para

todos los parámetros del modelo ($\hat{\beta}$, $\hat{\gamma}$, m) son obtenidos. En base a esto, Muggeo plantea una parametrización del modelo

$$Y = \alpha(Z) + \beta(Z - m)_+$$

donde m es el punto de cambio y $(Z - m)_+ = (Z - m)I(Z > m)$ siendo $I(A) = 1$ si A es verdad. α es la pendiente del segmento de la línea izquierda para $Z \leq m$, β es la diferencia entre pendientes y $(\alpha+\beta)$ es la pendiente del segmento de línea derecho; entonces si el punto de cambio existe, $|\beta| > 0$, aquí Muggeo hace notar que en $Z = m$ el logaritmo de la verosimilitud no es diferenciable. Dado que por la aproximación de Taylor $(Z - m)_+$ puede linealizarse alrededor de $m^{(0)}$, entonces, se tiene que $m^{(0)}$ es el punto de cambio:

$$(Z - m)_+ = (Z - m^{(0)})_+ + (m - m^{(0)})(-1)I(Z > m^{(0)})$$

donde $(-1)I(Z > m^{(0)})$ es la primera derivada de $(Z - m)_+$ evaluada en $m^{(0)}$. Por lo que, el modelo finalmente queda como

$$Y = \alpha(Z) + \beta U^{(s)} + \gamma V^{(s)}, \quad U^{(s)} = (Z - m^{(s)})_+, V^{(s)} = -I(Z > m^{(s)}).$$

Cabe mencionar que a diferencia de los resultados presentados por estos autores para resolver el problema del cambio estructural, en este trabajo se realiza una prueba de hipótesis, usando la razón de verosimilitud, y se propone un estadístico de prueba para el cual se encuentra su distribución de probabilidad exacta. Además se plantea un posible estimador del punto de cambio, a través del método de máxima verosimilitud.

4.2 Una prueba de hipótesis alternativa para cambio estructural.

En esta sección se efectúa una prueba de hipótesis para el caso donde se sospecha que hubo un cambio estructural, pero se desconoce en que momento ocurrió, utilizando la razón de verosimilitud.

4.2.1 Región crítica.

Sea Y_i y $X_{1i}, X_{2i}, \ldots, X_{pi}$, $i = 1, 2, \ldots, n$, una muestra aleatoria de las variables Y y $X_1, X_2, \ldots, X_p$. Las hipótesis de prueba son

$$H_0 : m = n \quad \text{versus} \quad H_1 : m < n.$$

El modelo asociado a H_1 (cambio estructural) es

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_i & \text{si} \quad i \leq m \\ \beta_0^* + \beta_1^* X_{1i} + \cdots + \beta_p^* X_{pi} + \varepsilon_i & \text{si} \quad i > m \end{cases}$$

con $\beta_j \neq \beta_j^*$ al menos para una j, $0 \leq j \leq p$, esto significa que se manifestó un cambio en la región de observación. ε_i es una variable aleatoria tal que $\varepsilon_i \sim N(0, \sigma^2)$. Al número m se denomina "el punto de cambio estructural".[2]

En forma matricial el modelo bajo H_1 se puede escribir como

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1\beta_1 \\ X_2\beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

donde Y_1 y Y_2 son vectores de dimensión $m \times 1$ y $n - m \times 1$, X_1 y X_2 son las matrices de dimensión $m \times (p+1)$ y $(n-m) \times (p+1)$, β_1 y β_2 son matrices de dimensión $(p+1) \times 1$ tales que $\beta_1^T = (\beta_0, \ldots, \beta_p)$ y $\beta_2^T = (\beta_0^*, \ldots, \beta_p^*)$ y, ε_1 y ε_2 son las matrices de dimensión $m \times 1$ y $(n-m) \times 1$, las cuales se distribuyen Normal con igual varianza.

Mientras que el modelo asociado a H_0 (no cambio estructural) es

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

para toda $i = 1, 2, \ldots, n$, esto es, no existe un punto de cambio en la región de observación.

En forma matricial el modelo bajo H_0 se puede escribir como la ecuación 3.1

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

La región crítica de esta prueba se encuentra con la razón de verosimilitud. La función de densidad conjunta o de verosimilitud de la muestra bajo H_0 es

$$f(x_1, x_2, \ldots, x_n; \beta, \sigma^2) = L(\beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n}\ e^{-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)} \tag{4.1}$$

Mientras que la función de densidad conjunta o de verosimilitud de la muestra bajo H_1 es,

$$L(\beta_1, \beta_2, \sigma^2, m) = \frac{1}{(2\pi)^{n/2}\sigma^n}\ e^{-\frac{1}{2\sigma^2}[(Y_1 - X_1\beta_1)^T(Y_1 - X_1\beta_1) + (Y_2 - X_2\beta_2)^T(Y_2 - X_2\beta_2)]} \tag{4.2}$$

La región crítica es

$$\frac{\max\limits_{H_0} L(\beta, \sigma^2)}{\max\limits_{H_1} L(\beta_1, \beta_2, \sigma^2, m)} \leq \lambda.$$

Es decir

$$\frac{\max\limits_{H_0} \frac{1}{(2\pi)^{n/2}\sigma^n}\ e^{-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)}}{\max\limits_{H_1} \frac{1}{(2\pi)^{n/2}\sigma^n}\ e^{-\frac{1}{2\sigma^2}[(Y_1 - X_1\beta_1)^T(Y_1 - X_1\beta_1) + (Y_2 - X_2\beta_2)^T(Y_2 - X_2\beta_2)]}} \leq \lambda.$$

[2] En realidad el cambio puede haber ocurrido en cualquier punto del intervalo cerrado $[m, m+1]$, pero para efectos de este trabajo se coloca el cambio en m.

Se tiene que los valores de β, y σ^2 que maximizan el numerador son (ver desarrollo del max $L(\beta, \sigma^2)$ en la sección 3.1)

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y \\ \hat{\sigma^2} &= \frac{1}{n}(Y - X\hat{\beta})^T (Y - X\hat{\beta}). \end{aligned}$$

Dado lo anterior

$$\max_{H_0} L(\beta, \sigma^2) = \frac{(2\pi)^{-n/2}\ n^{n/2}\ e^{-n/2}}{[(Y - X\hat{\beta})^T (Y - X\hat{\beta})]^{n/2}}.$$

Siguiendo un proceso similar al de la sección 3.1 y usando el hecho de que $\max_{x,y,z} f(x,y,z) = \max_x(\max_y(\max_z f(x,y,z)))$, los valores de β_1, β_2 y σ^2 que maximizan el denominador para un m fijo son

$$\begin{aligned} \hat{\sigma^2} &= \frac{1}{n}((Y_1 - X_1\hat{\beta}_1)^T (Y_1 - X_1\hat{\beta}_1) + (Y_2 - X_2\hat{\beta}_2)^T (Y_2 - X_2\hat{\beta}_2)) \\ \hat{\beta}_1 &= (X_1^T X_1)^{-1} X_1^T Y_1 \\ \hat{\beta}_2 &= (X_2^T X_2)^{-1} X_2^T Y_2 \end{aligned}$$

por lo que

$$\max_{H_1} L(\beta_1, \beta_2, \sigma^2, m) = \max_m \frac{(2\pi)^{-n/2}\ n^{n/2}\ e^{-n/2}}{[(Y_1 - X_1\hat{\beta}_1)^T (Y_1 - X_1\hat{\beta}_1) + (Y_2 - X_2\hat{\beta}_2)^T (Y_2 - X_2\hat{\beta}_2)]^{n/2}}$$

Reemplazando los resultados obtenidos tanto para el numerador como para el denominador, la razón de verosimilitud queda como

$$\frac{\frac{(2\pi)^{-n/2}\ n^{n/2}\ e^{-n/2}}{(Y - X\hat{\beta})^T (Y - X\hat{\beta})^{n/2}}}{\max_m \frac{(2\pi)^{-n/2}\ n^{n/2}\ e^{-n/2}}{[(Y_1 - X_1\hat{\beta}_1)^T (Y_1 - X_1\hat{\beta}_1) + (Y_2 - X_2\hat{\beta}_2)^T (Y_2 - X_2\hat{\beta}_2)]^{n/2}}} \leq \lambda$$

puesto que

$$\max_m \frac{(2\pi)^{-n/2}\ n^{n/2}\ e^{-n/2}}{[(Y_1 - X_1\hat{\beta}_1)^T (Y_1 - X_1\hat{\beta}_1) + (Y_2 - X_2\hat{\beta}_2)^T (Y_2 - X_2\hat{\beta}_2)]^{n/2}}$$

es equivalente a la expresión

$$\frac{(2\pi)^{-n/2}\ n^{n/2}\ e^{-n/2}}{\min_m [(Y_1 - X_1\hat{\beta}_1)^T (Y_1 - X_1\hat{\beta}_1) + (Y_2 - X_2\hat{\beta}_2)^T (Y_2 - X_2\hat{\beta}_2)]^{n/2}}.$$

Sustituyendo la espresión anterior en la razón de verosimilitud y simplificando terminos se obtiene que

$$\min_m \frac{(Y_1 - X_1\hat{\beta}_1)^T (Y_1 - X_1\hat{\beta}_1) + (Y_2 - X_2\hat{\beta}_2)^T (Y_2 - X_2\hat{\beta}_2)}{(Y - X\hat{\beta})^T (Y - X\hat{\beta})} \leq \lambda^*.$$

Esta expresión se puede escribir también como

$$\lambda_c = \min_m \left\{ \frac{SCE_{1_m} + SCE_{2_m}}{SCE} \right\} \leq \lambda^*.$$

Esta expresión, $\lambda_c \leq \lambda^*$, sólo tiene sentido cuando $p+1 \leq m \leq n-p-1$ debido a que no se puede realizar una regresión lineal con menos observaciones que parámetros.

En la siguiente sección se determina como se distribuye el estadístico de prueba encontrado.

4.2.2 Distribución del estadístico de prueba

Por la sección 3.3 se sabe que la forma cuadrática $\frac{1}{\sigma^2} Y^T B Y = \frac{1}{\sigma^2} SCE \sim \chi^2_{n-p-1}$; de la misma manera, las formas cuadráticas $\frac{1}{\sigma^2} SCE_{1_m} \sim \chi^2_{m-p-1}$ y $\frac{1}{\sigma^2} SCE_{2_m} \sim \chi^2_{n-m-p-1}$.

Dado el hecho de que las sumas de cuadrados asociadas al estadístico de prueba se distribuyen como χ^2, la consecuencia lógica es pensar que el estadístico presenta una función de distribución F, pero el detalle radica en que $SCE_{1_m} + SCE_{2_m}$ no es independiente de SCE. Así que para encontrar la forma en la cual se distribuye el estadístico de prueba se utiliza un cambio de variable adecuado, donde $SCE_{1_m} + SCE_{2_m}$ continua siendo no independiente de SCE.

Dado que la matriz asociada a la forma cuadrática de $SCE = Y^T B Y$ $B = I_n - X(X^T X)^{-1} X^T$ es una matriz simétrica de rango $n-p-1$, existe una matriz P de dimensión $n \times (n-p-1)$ tal que,

$$\begin{aligned} PP^T &= I_n - X(X^T X)^{-1} X^T \qquad y \\ P^T P &= I_{n-p-1} \end{aligned}$$

entonces, SCE se puede expresar como $Y^T B Y = Y^T P P^T Y$, por el teorema 2.3.2 se tiene que $P^T Y \sim N(P^T X\beta, \sigma^2 P^T P)$, y dado que $X^T P$ se puede escribir como $X^T P P^T P$ se tiene

$$\begin{aligned} X^T P P^T P &= X^T (I_n - X(X^T X)^{-1} X^T) P \\ X^T P I_{n-p-1} &= X^T P - X^T X (X^T X)^{-1} X^T P \\ X^T P &= X^T P - X^T P \\ X^T P &= 0 \end{aligned}$$

de donde se sigue $P^T X = 0$ y entonces

$$W = P^T Y \sim N(0, \sigma^2 I_{n-p-1}).$$

De esta manera

$$\frac{1}{\sigma^2} SCE = \frac{1}{\sigma^2} W^T W \sim \chi^2_{n-p-1}.$$

Un resultado importante es que $SCE_{1_m} + SCE_{2_m}$ se puede escribir en términos de W; en efecto, considere a $Y^T N_m Y$ la forma cuadrática asociada a $SCE_{1_m} + SCE_{2_m}$ donde

$$N_m = \left(\begin{array}{c|c} I_m - X_1 (X_1^T X_1)^{-1} X_1^T & 0 \\ \hline 0 & I_{n-m} - X_2 (X_2^T X_2)^{-1} X_2^T \end{array} \right)$$

entonces, si existe una matriz Z_m tal que $SCE_{1_m} + SCE_{2_m} = Y^T N_m Y = W^T Z_m W$, se tiene que la matriz $N_m = P Z_m P^T$, lo que nos lleva a que $Z_m = P^T N_m P$, es decir

$$Z_m = I_{n-p-1} - P^T \left(\begin{array}{c|c} X_1(X_1^T X_1)^{-1} X_1^T & 0 \\ \hline 0 & X_2(X_2^T X_2)^{-1} X_2^T \end{array} \right) P.$$

Por lo que

$$\begin{aligned} W^T Z_m W &= W^T W - W^T P^T \left(\begin{array}{c|c} X_1(X_1^T X_1)^{-1} X_1^T & 0 \\ \hline 0 & X_2(X_2^T X_2)^{-1} X_2^T \end{array} \right) PW \\ &= W^T W - W^T Q_m W \end{aligned}$$

donde

$$Q_m = P^T \left(\begin{array}{c|c} X_1(X_1^T X_1)^{-1} X_1^T & 0 \\ \hline 0 & X_2(X_2^T X_2)^{-1} X_2^T \end{array} \right) P.$$

En consecuencia, la región crítica de la prueba de hipótesis es

$$\min_m \left\{ \frac{W^T Z_m W}{W^T W} \right\} \leq \lambda^*$$

y como

$$\min_m \frac{W^T Z_m W}{W^T W} = \min_m \left\{ 1 - \frac{W^T Q_m W}{W^T W} \right\} = 1 - \max_m \frac{W^T Q_m W}{W^T W} \leq \lambda^*$$

entonces, la región crítica es equivalente a

$$\max_m \frac{W^T Q_m W}{W^T W} \geq 1 - \lambda^*.$$

Ahora, sólo falta encontrar el valor de λ^*, tal que

$$P\left(\max_m \frac{W^T Q_m W}{W^T W} \geq 1 - \lambda^* \right) = \alpha$$

donde α es el nivel de significancia de la prueba. Para ello se va a encontrar la distribución conjunta de las primeras $n-p-2$ coordenadas del vector $\frac{W}{\sqrt{W^T W}}$; esto es, se va utilizar el cambio de variable $V_i = W_i / \sqrt{\sum_{i=1}^{n-p-1} W_i^2}$, el cual satisface la ecuación,

$$V_1^2 + V_2^2 + \ldots + V_{n-p-1}^2 = \frac{W_1^2 + W_2^2 + \ldots + W_{n-p-1}^2}{\sum_{i=1}^{n-p-1} W_i^2} = 1,$$

Por tanto, el último componente de esta sumatoria se puede expresar en función de los demás elementos,

$$V_{n-p-1}^2 = 1 - \sum_{i=1}^{n-p-2} V_i^2,$$

lo que justifica que en el siguiente teorema se considere únicamente $n-p-2$ coordenadas del vector $\frac{W}{\sqrt{W^T W}}$.

Teorema 4.2.1. *Sea el vector* $V^T = (V_1, \ldots, V_{n-p-2})$ *con* $V_i = W_i/\sqrt{\sum_{i=1}^{n-p-1} W_i^2}$, *entonces su función de densidad conjunta bajo* H_0, *es*

$$f_V(v_1, v_2, \ldots, v_{n-p-2}) = \begin{cases} \frac{\Gamma((n-p-1)/2)}{2(\pi)^{(n-p-1)/2}(1-v^Tv)^{1/2}}, & si\ v^Tv < 1 \\ 0, & en\ otro\ caso. \end{cases}$$

Demostración. *La función de densidad conjunta del vector* W *es*

$$f_W(w_1, w_2, \ldots, w_{n-p-1}) = \frac{e^{-w^Tw/2\sigma^2}}{(2\pi)^{(n-p-1)/2}\sigma^{n-p-1}}$$

Por definición

$$v_j^2 = \frac{w_j^2}{\sum_{i=1}^{n-p-1} w_i^2},$$

ordenando terminos

$$v_j^2 \sum_{i=1}^{n-p-1} w_i^2 = w_j^2, \tag{4.3}$$

Resolviendo esta ecuación para w_j^2 *cuando* $j = 1$, *se obtiene*

$$w_1^2 = \frac{v_1^2}{1-v_1^2} \sum_{i=2}^{n-p-1} w_i^2,$$

y luego

$$\sum_{i=1}^{n-p-1} w_i^2 = w_1^2 + \sum_{i=2}^{n-p-1} w_i^2 = \left(\frac{v_1^2}{1-v_1^2} + 1\right) \sum_{i=2}^{n-p-1} w_i^2 = \frac{1}{1-v_1^2} \sum_{i=2}^{n-p-1} w_i^2$$

De esta manera, para la misma ecuación cuando $j = 2$, *se tiene que*

$$\frac{v_2^2}{1-v_1^2} \sum_{i=2}^{n-p-1} w_i^2 = w_2^2.$$

y resolviendo para w_2^2 *se obtiene*

$$w_2^2 = \frac{v_2^2}{1-v_1^2-v_2^2} \sum_{i=3}^{n-p-1} w_i^2,$$

donde

$$\begin{aligned} \sum_{i=1}^{n-p-1} w_i^2 &= \frac{1}{1-v_1^2} \sum_{i=2}^{n-p-1} w_i^2 = \frac{1}{1-v_1^2}(w_2 + \sum_{i=3}^{n-p-1} w_i^2) \\ &= \frac{1}{1-v_1^2-v_2^2} \sum_{i=3}^{n-p-1} w_i^2 \end{aligned}$$

Siguiendo el procedimiento hasta $j = n-p-2$ se tiene:

$$\sum_{i=1}^{n-p-1} w_i^2 = \frac{w_{n-p-1}^2}{1 - v_1^2 - \ldots - v_{n-p-2}^2} = \frac{w_{n-p-1}^2}{1 - v^T v}$$

Reemplazando este resultado en 4.3, se tiene

$$w_j^2 = \frac{w_{n-p-1}^2}{1 - v^T v} v_j^2. \tag{4.4}$$

El siguiente paso ahora es encontrar el Jacobiano de esta transformación.
Las derivadas parciales son:

$$\begin{aligned}
\frac{\partial w_j}{\partial v_j} &= \frac{|w_{n-p-1}|}{\sqrt{1 - v^T v}} + \frac{|w_{n-p-1}| v_j^2}{(1 - v^T v)^{3/2}} \qquad 1 \le j \le n-p-2 \\
\frac{\partial w_j}{\partial v_i} &= \frac{|w_{n-p-1}| v_j v_i}{(1 - v^T v)^{3/2}} \qquad\qquad 1 \le i \ne j \le n-p-2
\end{aligned}$$

y el jacobiano es

$$J = det \begin{pmatrix} \frac{|w_{n-p-1}|}{\sqrt{1-v^T v}} + \frac{|w_{n-p-1}| v_1^2}{(1-v^T v)^{3/2}} & \cdots & \frac{|w_{n-p-1}| v_1 v_{n-p-2}}{(1-v^T v)^{3/2}} \\ \vdots & \ddots & \vdots \\ \frac{|w_{n-p-1}| v_{n-p-2} v_1}{(1-v^T v)^{3/2}} & \cdots & \frac{|w_{n-p-1}|}{\sqrt{1-v^T v}} + \frac{|w_{n-p-1}| v_{n-p-2}^2}{(1-v^T v)^{3/2}} \end{pmatrix}$$

factorizando $\frac{|w_{n-p-1}|}{\sqrt{1-v^T v}}$

$$\begin{aligned}
J &= \left(\frac{|w_{n-p-1}|}{\sqrt{1 - v^T v}} \right)^{n-p-2} det \begin{pmatrix} 1 + \frac{v_1^2}{1-v^T v} & \cdots & \frac{v_1 v_{n-p-2}}{1-v^T v} \\ \vdots & \ddots & \vdots \\ \frac{v_{n-p-2} v_1}{1-v^T v} & \cdots & 1 + \frac{v_{n-p-2}^2}{1-v^T v} \end{pmatrix} \\
J &= \left(\frac{|w_{n-p-1}|}{\sqrt{1 - v^T v}} \right)^{n-p-2} det \left(\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} + \frac{1}{1 - v^T v} \begin{bmatrix} v_1^2 & \cdots & v_1 v_{n-p-2} \\ \vdots & \ddots & \vdots \\ v_{n-p-2} v_1 & \cdots & v_{n-p-2}^2 \end{bmatrix} \right) \\
J &= \left(\frac{|w_{n-p-1}|}{\sqrt{1 - v^T v}} \right)^{n-p-2} det \left(I + \frac{1}{1 - v^T v} v v^T \right).
\end{aligned}$$

El determinante de la matrix $I + \frac{1}{1-v^T v} v v^T$ puede ser obtenido como el producto de sus valores propios. Los vectores propios de la matriz $I + \frac{1}{1-v^T v} v v^T$ son v y cualquier vector z ortogonal a v. El valor propio asociado a v es $\frac{1}{1-v^T v}$, ya que

$$(I + \frac{1}{1 - v^T v} v v^T) v = v + \frac{v^T v}{1 - v^T v} v = \frac{1}{1 - v^T v} v$$

y el valor propio asociado a z ortogonal a v es uno debido a que

$$(I + \frac{1}{1 - v^T v} v v^T) z = z,$$

por lo tanto

$$det\left(I + \frac{1}{1 - v^T v} v v^T\right) = \frac{1}{1 - v^T v}$$

y

$$J = \frac{|w_{n-p-1}^{n-p-2}|}{(1 - v^T v)^{(n-p)/2}}. \tag{4.5}$$

Por la ecuación 4.4 se tiene que $w_1^2 = \frac{w_{n-p-1}^2}{1-v^T v} v_1^2$,..., $w_{n-p-2}^2 = \frac{w_{n-p-1}^2}{1-v^T v} v_{n-p-2}^2$, *y dado que* $w_{n-p-1}^2 = w_{n-p-1}^2$, *se deduce que*

$$w^T w = \frac{w_{n-p-1}^2}{1 - v^T v}. \tag{4.6}$$

Sustituyendo 4.5 y 4.6 en la función de densidad de W se obtiene

$$f_W(w_1, w_2, \ldots, w_{n-p-2}, w_{n-p-1}) = \frac{|w_{n-p-1}^{n-p-2}| \; e^{-\frac{w_{n-p-1}^2}{2\sigma^2(1-v^T v)}}}{(2\pi)^{(n-p-1)/2} \sigma^{n-p-1} (1 - v^T v)^{(n-p)/2}}$$

Entonces, la función de densidad del vector V es

$$f_V(v_1, v_2, \ldots, v_{n-p-2}) = \frac{\displaystyle\int_{-\infty}^{\infty} |w_{n-p-1}^{n-p-2}| \; e^{-\frac{w_{n-p-1}^2}{2\sigma^2(1-v^T v)}} dw_{n-p-1}}{(2\pi)^{(n-p-1)/2} \sigma^{n-p-1} (1 - v^T v)^{(n-p)/2}}$$

y usando el cambio de variable $v = \frac{w_{n-p-1}^2}{2\sigma^2(1-v^T v)}$, *se tiene*

$$\begin{aligned} f_V(v_1, v_2, \ldots, v_{n-p-2}) &= \frac{2^{(n-p-3)/2}}{(2\pi)^{(n-p-1)/2}(1 - v^T v)^{1/2}} \int_{-\infty}^{\infty} |v|^{((n-p-1)/2)-1} e^{-v} dv \\ &= \frac{1}{2(\pi)^{(n-p-1)/2}(1 - v^T v)^{1/2}} \int_{0}^{\infty} |v|^{((n-p-1)/2)-1} e^{-v} dv \\ &= \frac{\Gamma((n-p-1)/2)}{2(\pi)^{(n-p-1)/2}(1 - v^T v)^{1/2}}, \end{aligned}$$

Por tanto,

$$f_V(v_1, v_2, \ldots, v_{n-p-2}) = \begin{cases} \frac{\Gamma((n-p-1)/2)}{2(\pi)^{(n-p-1)/2}(1-v^T v)^{1/2}}, & si \; v^T v < 1 \\ 0, & en \; otro \; caso. \end{cases}$$

■

Con lo cual se tiene que la función $f_V(v_1, v_2, \ldots, v_{n-p-2})$ no depende de ningún parámetro desconocido, por tanto, si se utiliza esta función de densidad la región crítica sólo depende de los valores de la muestra.

Dado que $\max_m \frac{W^T Q_m W}{W^T W} \geq 1 - \lambda^*$ es equivalente a $\max_m u^T Q_m u \geq 1 - \lambda^*$ con $u_i = v_i \quad i = 1, 2, \ldots, n-p-2$ y $u_{n-p-1} = \sqrt{1 - v^T v}$; para encontrar el valor de λ^*, observe que

$$P\left(\max_m u^T Q_m u \geq 1 - \lambda^*\right) = 1 - P\left(\max_m u^T Q_m u < 1 - \lambda^*\right)$$

y que

$$P\left(\max_m u^T Q_m u < 1 - \lambda^*\right) = P\left(u^T Q_{p+1} u < 1 - \lambda^*, \ldots, u^T Q_{n-p-1} u < 1 - \lambda^*\right)$$

entonces, se debe resolver para α y λ^* la ecuación,

$$P\left(\max_m u^T Q_m u < 1 - \lambda^*\right) = \int \cdots \int_{A_{\lambda^*}} f_V(v_1, v_2, \ldots, v_{n-p-2}) dv_1 dv_2 \ldots dv_{n-p-2} = 1 - \alpha$$

con $A_{\lambda^*} = \{v \in \mathbb{R}^{n-p-2} | u^T Q_{p+1} u < 1 - \lambda^*, \cdots, u^T Q_{n-p-1} u < 1 - \lambda^*\}$.

Esta integral es difícil de calcular debido a que la región de integración depende del vector u. Por ende, se procede a encontrar el valor de λ^* para un particular nivel de significancia $\alpha = 0.05$, usando por un lado, el método de Monte Carlo, y por el otro, mediante integración numérica.

4.3 Determinación de la región crítica y estimación del punto de cambio

Esta sección se compone de dos partes, en la primera se muestran dos algoritmos para determinar la región crítica del estadístico de prueba[3] y en la segunda, se propone un estimador el cual indica en donde puede haber ocurrido el cambio, cuando se rechaza la hipótesis nula. Y mediante el método de Monte Carlo se analiza su sesgo y su varianza.

[3] Ambos algoritmos se desarrollaron a partir de un modelo de regresión lineal general con p variables explicativas en el lenguaje de programación Mathematica versión 5.2 y el codigo fuente de ambos algoritmos se encuentran al final de este capítulo.

4.3.1 Algoritmos: simulación por Monte Carlo e integración numérica

a) **Método de simulación por Monte Carlo (MC).**

El algoritmo consiste en los siguientes pasos.

1. Realizar 50000 veces el siguiente proceso:
   a) Generar los datos del vector $W \sim N(0, I\sigma^2)$.
   b) Encontrar el máximo de $u^T Q_m u$, donde $u = W/\sqrt{W^T W}$.
2. Ordenar en forma ascendente los 50000 valores obtenidos
3. Encontrar el valor para el cual se tenga el 5 por ciento de los valores ordenados por arriba de él.
4. El valor encontrado es el valor de $1 - \lambda^*$.

b) **Método de integración numérica (IN).**

Se sabe que la función de distribución del estadístico de prueba esta dada por la ecuación integral

$$P(\max u^T Q_m u < 1 - \lambda^*) = \int \dots \int_{A_\lambda^*} f_V(v_1, v_2, \dots, v_{n-p-2}) dv_1 dv_2 \dots dv_{n-p-2} = 1 - \alpha$$

con $A_{\lambda^*} = \{v \in R^{n-p-2} \mid u^T Q_{p+1} u < 1 - \lambda^*, \dots, u^T Q_{n-p-1} u < 1 - \lambda^*\}$, donde el vector u es de la forma $u = (u_1, u_2, \dots, u_{n-p-1})$ con $u_i = v_i$ si $i = 1, 2, \dots, n-p-2$ y $u_{n-p-1} = \sqrt{1 - v^T v}$); y $f_V(v_1, \dots, v_n) = \frac{\Gamma((n-p-1)/n)}{2\pi^{n-p-1}(1-v^T v)^{1/2}}$.

Por lo tanto, en la región de integración el vector $v = (v_1, v_2, \dots, v_{n-p-2})$ debe satisfacer las restricciones de que $v^T v < 1$ y $\max\{u^T Q_{p+k} u < 1 - \lambda^* \mid k = 1, 2, \dots, n - 2p - 1\}$.

Para la integración numérica se hace una partición del intervalo $[-1, 1]$ dada por

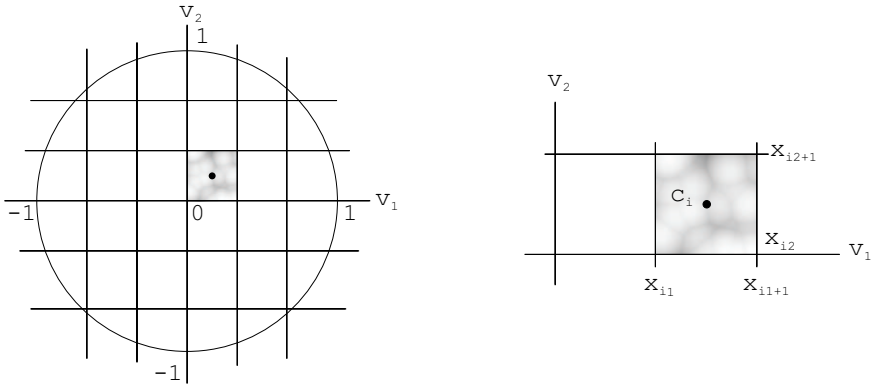
$$x_0 = -1,\ x_1 = -1 + 1/r,\ x_2 = -1 + 2/r, \dots, x_{2r} = 1$$

de esta manera se forman los híper cubitos

$$[x_{i_1}, x_{i_1+1}] \times [x_{i_2}, x_{i_2+1}] \times [x_{i_3}, x_{i_3+1}] \times \dots \times [x_{i_{n-p-2}}, x_{i_{n-p-2}+1}],$$

cuyo volumen está dado por $\Delta(v) = 1/(r)^{n-p-2}$. Enseguida se obtiene el punto central de cada híper cubito c_i.

Para el caso más simple, con dos variables v_1 y v_2, ello se ve como,



Si c_i satisface la restricción $v^T v < 1$, se encuentra para c_i el valor

$$\max(c_i) = \max_{k=1,2,\ldots,n-2p-1}\{u^T Q_{p+k} u < 1-\lambda\} \text{ y el valor de } f_V(c_i)/(r)^{n-p-2};$$

y se guardan en un renglón de una matriz U estos dos valores.

Al final de este proceso, se ordenan los renglones de U de menor a mayor con respecto al valor obtenido de $\max(c_i)$.

Finalmente, se suman los términos $f_V(c_i)\Delta(v_i)$ comenzando con el correspondiente al mayor valor $\max(c_i)$, hasta tener un acumulado en la suma cercano a $\alpha = 0.05$. El último valor agregado a la suma corresponde a $1-\lambda^* = \max(c_i)_{final\ de\ suma}$.

El método de integración numérica sólo se aplicó para ejemplos con un máximo de 13 observaciones con una partición del intervalo de $2r = 6$ partes, puesto que se esta limitado por la capacidad del equipo de computo, es decir, si el número de observaciones es mayor a 13, el número de operaciones que se realizan para calcular los puntos dentro del circulo unitario tiende a ser bastantes elevado, por lo que en algún punto de estos cálculos el equipo no cuenta con la memoria suficiente para seguir con las operaciones del programa, misma razón por la cual el intervalo no se particiono en partes más pequeñas. Cabe señalar que ambos métodos están desarrollados para un modelo de regresión con p variables explicativas.

4.3.2 Estimación del punto de cambio

La función de densidad conjunta o de verosimilitud asociada a la existencia de cambio estructural (H_1) es

$$L(\beta_1, \beta_2, \sigma^2, m) = \frac{1}{(2\pi)^{n/2}\sigma^n}\ e^{-\frac{1}{2\sigma^2}[(Y_1 - X_1\beta_1)^T(Y_1 - X_1\beta_1) + (Y_2 - X_2\beta_2)^T(Y_2 - X_2\beta_2)]}$$

y por la sección 4.2.1 se sabe que la función de verosimilitud alcanza su máximo cuando se encuentra el

$$\min_m[(Y_1 - X_1\hat{\beta}_1)^T(Y_1 - X_1\hat{\beta}_1) + (Y_2 - X_2\hat{\beta}_2)^T(Y_2 - X_2\hat{\beta}_2)] = \min_m(SCE_{1_m} + SCE_{2_m}).$$

Para aclarar el procedimiento de estimación del punto donde se produce el cambio a continuación se detalla el proceso. Para $k = p+1, 2, \ldots, n-p-1$
1) se divide la muestra aleatoria de tamaño n en dos submuestras, una con los primeros k datos y la otra con los $n-k$ datos restantes;
2) para cada submuestra se realiza un análisis de regresión, se encuentra la suma de cuadrados correspondiente y se suman para encontrar el estadístico $SCE_{1_k} + SCE_{2_k}$.
Es decir, se calculan

$$\begin{array}{ll} k = p+1 & SCE_{1_{p+1}} + SCE_{2_{n-p-1}} \\ k = p+2 & SCE_{1_{p+2}} + SCE_{2_{n-p-2}} \\ \vdots & \vdots \\ k = m & SCE_{1_m} + SCE_{2_{n-m}} \\ \vdots & \vdots \\ k = n-(p+2) & SCE_{1_{n-p-2}} + SCE_{2_{p+2}} \\ k = n-(p+1) & SCE_{1_{n-p-1}} + SCE_{2_{p+1}} \end{array}$$

A partir de esto, se propone como estimador de m a

$$\hat{m} = \{m \mid SCE_{1_m} + SCE_{2_m} \leq SCE_{1_k} + SCE_{2_k}, k = p+1, ..., n-p-1.\}$$

El rango de valores de este estimador esta restringido a $p+1 \leq \hat{m} \leq n-p-1$, debido a dos razones técnicas:

1. Si una submuestra tiene menos de $p+1$ datos, entonces se tienen datos insuficientes para ajustar un modelo de regresión con p variables explicativas.
2. Si $k \leq p+1$, implica que $SCE_{1_k} + SCE_{2_k} \geq SCE_{1_{p+1}} + SCE_{2_{p+1}}$ y por lo tanto, k no será elegido como estimador de m. Caso similar ocurre cuando $k \geq n-p-1$. Este resultado se obtiene con el siguiente teorema.

Teorema 4.3.1. *Sea $(Y_1, X_1), ..., (Y_n, X_n)$ una muestra aleatoria y SCE_{1_k} y SCE_{2_k} como se definieron anteriormente, entonces, cuando $k \leq p+1$ implica que $SCE_{1_k} + SCE_{2_k} \geq SCE_{1_{p+1}} + SCE_{2_{p+1}}$.*

Demostración. *Si $k \leq p+1$ se sigue que $SCE_{1_k} = 0$, por lo tanto, $SCE_{1_k} + SCE_{2_k} = SCE_{2_k}$ y para demostrar el teorema basta ver que $SCE_{2_k} > SCE_{2_{p+1}}$ cuando $k < p+1$.*

Considere la matriz de variables explicativas X de $n \times (p+1)$ que se descompone en tres submuestras, X_1 de $k \times (p+1)$, X_2 de $(p+1-k) \times (p+1)$ y X_3 de $(n-p-1) \times (p+1)$ tales que $X^T = (X_1^T, X_2^T, X_3^T)$ de esta manera (con $k < p+1$) se tiene que,

$$\begin{aligned} SCE_{2_{p+1}} &= Y_3^T (I - X_3 (X_3^T X_3)^{-1} X_3^T) Y_3 \\ y \;\; SCE_{2_k} &= \begin{bmatrix} Y_2^T & Y_3^T \end{bmatrix} \left(I - \begin{bmatrix} X_2 \\ X_3 \end{bmatrix} (\begin{bmatrix} X_2^T & X_3^T \end{bmatrix} \begin{bmatrix} X_2 \\ X_3 \end{bmatrix})^{-1} \begin{bmatrix} X_2^T & X_3^T \end{bmatrix} \right) \begin{bmatrix} Y_2 \\ Y_3 \end{bmatrix} \\ &= \begin{bmatrix} Y_2^T & Y_3^T \end{bmatrix} \left(I - \begin{bmatrix} X_2 \\ X_3 \end{bmatrix} (X_2^T X_2 + X_3^T X_3)^{-1} \begin{bmatrix} X_2^T & X_3^T \end{bmatrix} \right) \begin{bmatrix} Y_2 \\ Y_3 \end{bmatrix} \end{aligned}$$

se puede probar que

$$\begin{aligned}(X_2^TX_2 &+ X_3^TX_3)^{-1} \\ &= (X_3^TX_3)^{-1} - (X_3^TX_3)^{-1}X_2^T(I + X_2(X_3^TX_3)^{-1}X_2^T)^{-1}X_2(X_3^TX_3)^{-1}\end{aligned}$$

y que

$$I - \begin{bmatrix}X_2\\X_3\end{bmatrix}(X_2^TX_2 + X_3^TX_3)^{-1}\begin{bmatrix}X_2^T & X_3^T\end{bmatrix} = \left(\begin{array}{c|c}A_{11} & A_{12}\\ \hline A_{21} & A_{22}\end{array}\right)$$

$$\begin{aligned}A_{11} &= I - X_2(X_3^TX_3)^{-1}X_2^T - X_2(X_3^TX_3)^{-1}X_2^T(I + X_2(X_3^TX_3)^{-1}X_2^T)^{-1}X_2(X_3^TX_3)^{-1}X_2^T\\ A_{12} &= -X_2(X_3^TX_3)^{-1}X_3^T - X_2(X_3^TX_3)^{-1}X_2^T(I + X_2(X_3^TX_3)^{-1}X_2^T)^{-1}X_2(X_3^TX_3)^{-1}X_3^T\\ A_{21} &= -X_3(X_3^TX_3)^{-1}X_2^T - X_3(X_3^TX_3)^{-1}X_2^T(I + X_2(X_3^TX_3)^{-1}X_2^T)^{-1}X_2(X_3^TX_3)^{-1}X_2^T\\ A_{22} &= I - X_3(X_3^TX_3)^{-1}X_3^T + X_3(X_3^TX_3)^{-1}X_2^T(I + X_2(X_3^TX_3)^{-1}X_2^T)^{-1}X_2(X_3^TX_3)^{-1}X_3^T\end{aligned}$$

puesto que

$X_2(X_3^TX_3)^{-1}X_2^T(I + X_2(X_3^TX_3)^{-1}X_2^T)^{-1}$
$= (I + X_2(X_3^TX_3)^{-1}X_2^T - I)(I + X_2(X_3^TX_3)^{-1}X_2^T)^{-1} = I - (I + X_2(X_3^TX_3)^{-1}X_2^T)^{-1}$

entonces

$$\begin{aligned}A_{11} &= (I + X_2(X_3^TX_3)^{-1}X_2^T)^{-1}\\ A_{12} &= -(I + X_2(X_3^TX_3)^{-1}X_2^T)^{-1}X_2(X_3X_3^T)^{-1}X_3^T\\ A_{21} &= -X_3(X_3X_3^T)^{-1}X_2^T(I + X_2(X_3^TX_3)^{-1}X_2^T)^{-1}\\ A_{22} &= I - X_3(X_3^TX_3)^{-1}X_3^T + X_3(X_3^TX_3)^{-1}X_2^T(I + X_2(X_3^TX_3)^{-1}X_2^T)^{-1}X_2(X_3^TX_3)^{-1}X_3^T\end{aligned}$$

se tiene que

$$SCE_{2_k} = \begin{bmatrix}Y_2^T & Y_3^T\end{bmatrix}\left(\begin{array}{c|c}A_{11} & A_{12}\\ \hline A_{21} & A_{22}\end{array}\right)\begin{bmatrix}Y_2\\Y_3\end{bmatrix}$$

$$SCE_{2_k} = [(Y_2 - X_2(X_3X_3^T)^{-1}X_3^TY_3)^T(I - X_2(X_3^TX_3)^{-1}X_2^T)^{-1}(Y_2 - X_2(X_3X_3^T)^{-1}X_3^TY_3)]$$

$$+[Y_3^T(I - X_3(X_3^TX_3)^{-1}X_3^T)Y_3]$$

Por consiguiente, la SCE_{2_k} es la suma de dos formas cuadráticas, ambas positivas definidas, en donde la segunda forma cuadrática representa a la $SCE_{2_{p+1}}$.

Por tanto, queda demostrado que $SCE_{2_k} > SCE_{2_{p+1}}$. ■

4.4 Aplicaciones

A continuación, se aplican los dos métodos de integración propuestos en este trabajo para 6 ejemplos en los cuales se quiere saber si existe cambio estructural. En los ejemplos donde se halle evidencia de cambio estructural se estimara el punto en el cual ocurre el cambio y por medio de una simulación por Monte Carlo se sondearan las posibles propiedades de insesgamiento y eficiencia de este estimador. Esta simulación por Monte Carlo propone un modelo donde se conoce al punto de cambio m, después se realizan 1000 corridas, encontrándose el mismo número de estimaciones $\hat{m}$, finalmente, se calculan el promedio y la varianza de estos valores, $E(\hat{m})$ y $V(\hat{m})$, se utilizan 4 posibles valores de σ^2, $\sigma^2 = 0.01, 0.1, 1$ y 10.[4]

El tamaño de muestra de los primeros 4 ejemplos es de 13 observaciones, ya que se desea comparar los resultados de los dos métodos propuestos para calcular la región crítica. Los primeros 2 son ejemplos fueron tomados de Gujarati (1997) en los cuales se demuestra la presencia de cambio estructural, el ejemplo 3 fue tomado al azar de Daniel (1999) y el ejemplo 4 esta construido con datos reales de cifras mexicanas, cabe señalar que para los ejemplos 3 y 4 no se sabe si el cambio ha ocurrido o no.
El ejemplo 5 sólo muestra los resultados de aplicar el método Monte Carlo, ya que el tamaño de muestra es de 35 observaciones y la capacidad del calculo de la computadora que fue utilizada es insuficiente para efectuar la integración numérica. Este ejemplo es utilizado por Steven (2001) en donde él concluye la presencia de cambio estructural. Por ultimo, en el ejemplo 6 se propone un modelo de regresión en donde se conoce el punto de cambio estructural y aquí sólo se muestra como es su media y varianza.

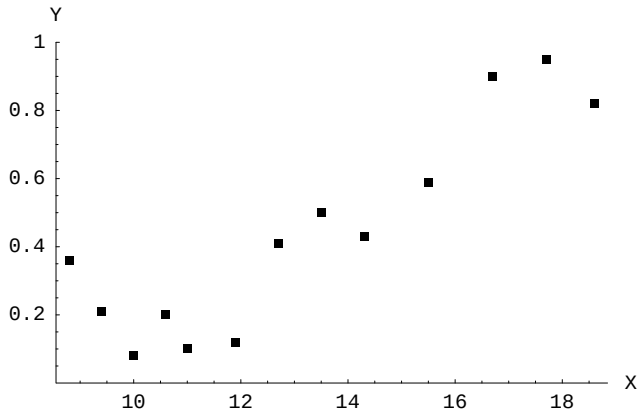
Antes de abordar los ejemplos, la notación que se utiliza para mostrar los resultados son: λ_{MC}, denota el valor de λ^* calculado por Monte Carlo, λ_{IN}, denota el valor de λ^* calculado por integración numérica y, λ_c, denota el valor del estadístico de prueba propuesto en este trabajo.

[4]El código que se utilizó para los algoritmos de estos métodos y para la simulación se hicieron con Mathematica y se encuentran al final del capítulo.

Ejemplo 1: Ahorro-Ingreso
Este ejemplo es el que se utilizo al principio del capítulo en la prueba de Chow, pero aquí el periodo de tiempo sólo abarca de 1946-1958, $n = 13$. Recordando un poco, estos datos fueron tomados en un periodo de reconstrucción y posreconstrucción de la segunda guerra mundial y se desea saber si hubo un cambio significativo en la función de ahorro de Gran Bretaña.

Las variables que se utilizan para ellos son

Y: Ahorro (millones de libras)
X: Ingreso (millones de libras)



Los resultados para este ejemplo son

λ^*_{MC}	λ^*_{IN}	λ_c
0.7754	0.7288	0.3682

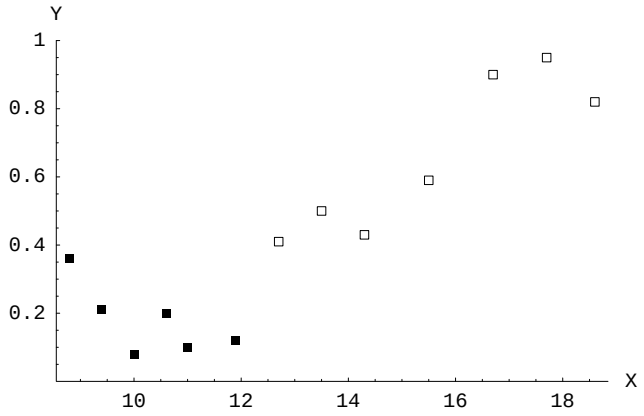
por tanto, se puede concluir que para ambos métodos la relación entre el ahorro y el ingreso si cambia en el periodo de observación, puesto que $\lambda_c \leq \lambda^*$.

Como en $m = 6$ se cumple que $SCE_{1_m} + SCE_{2_m}$ es menor o igual a $SCE_{1_k} + SCE_{2_k}$, se propone a $\hat{m} = 6$ como el punto en donde se da el cambio estructural.

k	SCE_{1_k}	SCE_{2_k}	$SCE_{1_k} + SCE_{2_k}$
2	0.0000	1.0462	1.0462
3	0.0021	0.9064	0.9086
4	0.0024	0.6777	0.6801
5	0.0035	0.3010	0.3044
6	**0.0493**	**0.0766**	**0.1259**
7	0.3002	0.0753	0.3755
8	0.3216	0.0308	0.3524
9	0.3623	0.0066	0.3689
10	0.4743	0.0050	0.4793
11	0.6984	0.0000	0.6984

Puesto que el punto de cambio se encuentra en $m = 6$, se tiene que el modelo asociado al cambio estructural esta dado por

$$Y_i = \begin{cases} 0.853 - 0.065X_i + \varepsilon_i, & \text{si } i \leq 6 \\ -0.796 + 0.093X_i + \varepsilon_i, & \text{si } i > 6 \end{cases} \qquad \text{con} \qquad \varepsilon_i \sim N(0, \sigma^2).$$



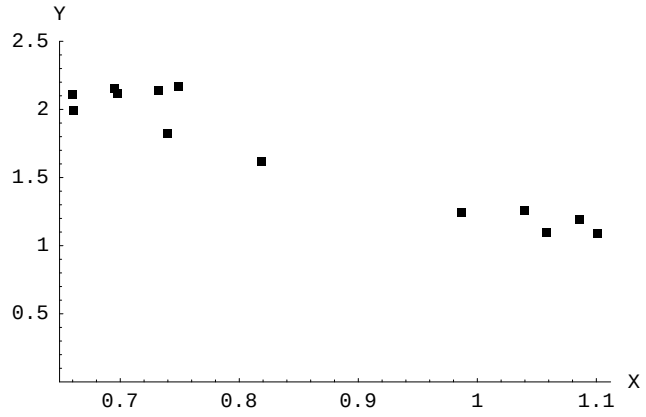
Aplicando la simulación para ver la media y la variaza del estimador, se tiene que

	σ^2			
	0.01	0.1	1	10
$E(\hat{m})$	6	5.639	6.49	6.531
$V(\hat{m})$	0	1.648	5.48	5.71

Ejemplo 2: Desempleo
Este ejemplo fue tomado de la página 503 del libro *"Econometría Básica"* de D. N. Gujarati, el cual tiene información anual trimestral sobre la tasa de desempleo y la tasa de empleos vacantes para el período 1958:IV-1971:II ($n = 51$), donde se muestra que hubo un dezplazamiento en la relación desempleo-vacantes a partir del cuarto trimestre de 1966, pero dada la restricción que se tiene sobre el número de observaciones, el periodo de tiempo que se tomo para este análisis fue de 1965:III-1968:III, $n = 13$.

Las variables consideradas son

Y: Tasa de desempleo, %
X: Tasa de empleos vacantes %



Los resultados de esta simulación son,

λ^*_{MC}	λ^*_{IN}	λ_c
0.7641	0.7222	0.1139

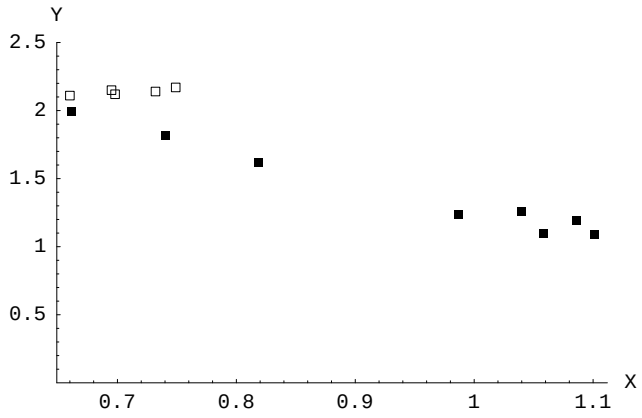
por tanto, el mínimo de λ_c se encuentra en $m = 8$ y ambos métodos sugieren que hay evidencia suficiente para concluir que la relación entre el porcentaje de empleos vacantes y la tasa de desempleo cambia después del segundo trimestre de 1967.

Como en $m = 8$ se tiene que $SCE_{1_m} + SCE_{2_m} \leq SCE_{1_k} + SCE_{2_k}$, se propone a $\hat{m} = 8$ como el punto en donde se da el cambio estructural.

k	SCE_{1_k}	SCE_{2_k}	$SCE_{1_k} + SCE_{2_k}$
2	0.0000	0.1713	0.1713
3	0.0021	0.1627	0.1648
4	0.0124	0.1623	0.1747
5	0.0135	0.1590	0.1724
6	0.0168	0.0938	0.1107
7	0.0193	0.0099	0.0292
8	**0.0205**	**0.0007**	**0.0212**
9	0.0340	0.0007	0.0347
10	0.0612	0.0004	0.0616
11	0.0850	0.0000	0.0850

Entonces, el modelo asociado al cambio estructural es

$$Y_i = \begin{cases} 3.29 - 2.004X_i + \varepsilon_i, & \text{si } i \leq 8 \\ 1.73 + 0.565X_i + \varepsilon_i, & \text{si } i > 8 \end{cases} \qquad \text{con} \qquad \varepsilon_i \sim N(0, \sigma^2).$$



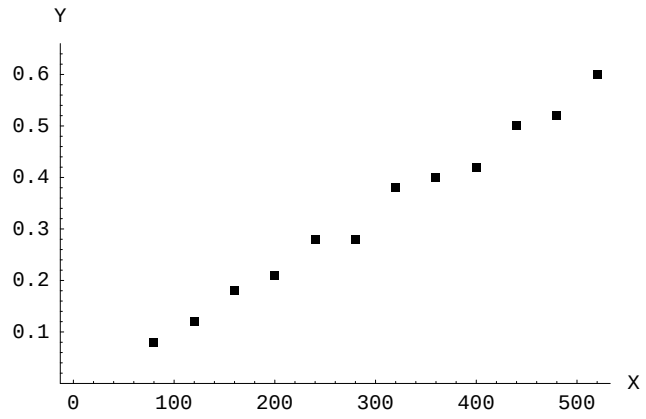
y el resultado de la simulación muestra que,

	σ^2			
	0.01	0.1	1	10
$E(\hat{m})$	8	8	6.389	6.292
$V(\hat{m})$	0	0	5.877	5.928

Ejemplo 3: Densidad
Este ejemplo fue tomado de la página 465 del libro *"Bioestadística: Base para el análisis de las ciencias de la salud"* de W. Daniel, en donde se desea verificar la densidad óptima de cierta sustancia a diferentes niveles de concentración.

Las variables que se utilizan son

Y: Densidad óptima
X: Nivel de concentración



Los resultados para este ejemplo son,

λ^*_{MC}	λ^*_{IN}	λ_c
0.761284	0.7222	0.8029

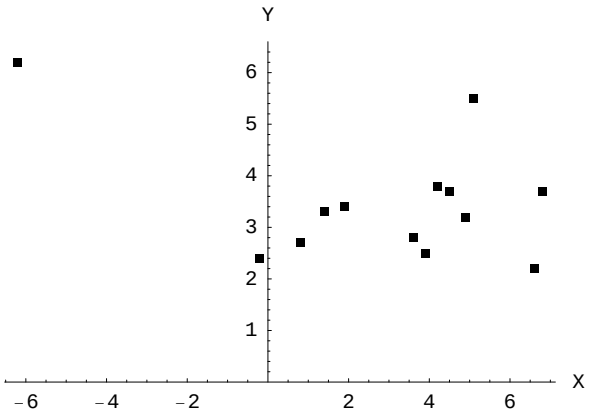
por tanto, para ambos métodos se puede concluir que la relación entre la dosis del medicamento y la reducción del ritmo cardiaco no cambia en el periodo de observación, puesto que $\lambda_c \nleq \lambda^*$.

Puesto que en este ejemplo no hay evidencia de cambio estructural no se procede a calcular el estimador.

Ejemplo 4: Desempleo México
En este ejemplo se desea saber si durante el cambio de presidente en México ha habido un cambio significativo en la tasa de desempleo abierto como función de la tasa de crecimiento del PIB para el período de 1992 - 2004. Los datos para este ejemplo fueron tomados de la pagina del Centro de Estudios de las Finanzas Públicas[5].

Las variables a considerar son

Y: Tasa de desempleo abierto %
X: Tasa de crecimiento del PIB %



Los resultados para este ejemplo son,

λ^*_{MC}	λ^*_{IN}	λ_c
0.8017	0.7633	0.5540

por tanto, ambos métodos indican que hay evidencia suficiente para concluir que la relación entre el crecimiento del PIB y el desempleo se modifican después de 1997, puesto que $\lambda_c \leq \lambda^*$, donde el mínimo de λ_c esta en $m = 6 = 1997$.

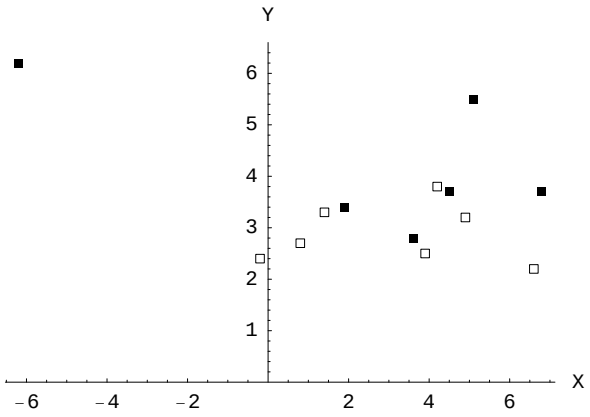
Entonces, como en $m = 6$ se cumple que $SCE_{1_m} + SCE_{2_m} \leq SCE_{1_k} + SCE_{2_k}$, se propone a $\hat{m} = 6$ como el punto en donde se da el cambio estructural.

[5] $http://www.cefp.gob.mx/intr/e-stadisticas/esta001a.xls$

k	SCE_{1_k}	SCE_{2_k}	$SCE_{1_k} + SCE_{2_k}$
2	0.0000	13.3555	13.3555
3	0.4096	13.2105	13.6202
4	0.8364	7.4633	8.2998
5	5.6017	2.4239	8.0256
6	**5.6617**	**1.9933**	**7.6551**
7	5.9834	1.8604	7.8438
8	7.6783	1.7478	9.4261
9	8.4748	0.1182	8.5930
10	12.0451	0.0964	12.1414
11	13.4025	0.0000	13.4025

El modelo asociado al cambio es

$$Y_i = \begin{cases} 4.66 - 0.171X_i + \varepsilon_i, & \text{si } i \leq 6 \\ 2.85 + 0.0051X_i + \varepsilon_i, & \text{si } i > 6 \end{cases} \qquad \text{con} \qquad \varepsilon_i \sim N(0, \sigma^2).$$



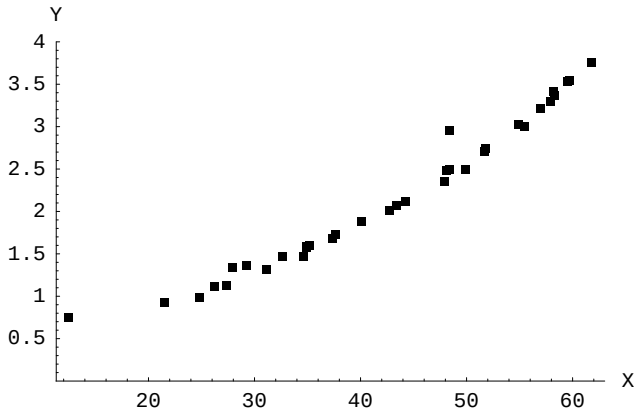
Los resultados de la simulación son,

	σ^2			
	0.01	0.1	1	10
$E(\hat{m})$	6	5.99	6.205	6.499
$V(\hat{m})$	0	0.003	3.946	6.40

Ejemplo 5: Energía
Los datos que se utilizan en este ejemplo fueron tomados del articulo de Steven (2001) para una muestra de 35 observaciones. En donde se desea saber en que momento el tipo de energía que necesita la gente que hace ejercicio cambia de aeróbica a anaeróbica.

Las variables consideradas para un individuo dado son:

Y: El volumen de dióxido de carbono exhalado por minuto
X: El volumen de oxigeno inhalado por minuto



Los resultados de esta simulación son,

λ^*_{MC}	λ_c
0.3530	0.9181

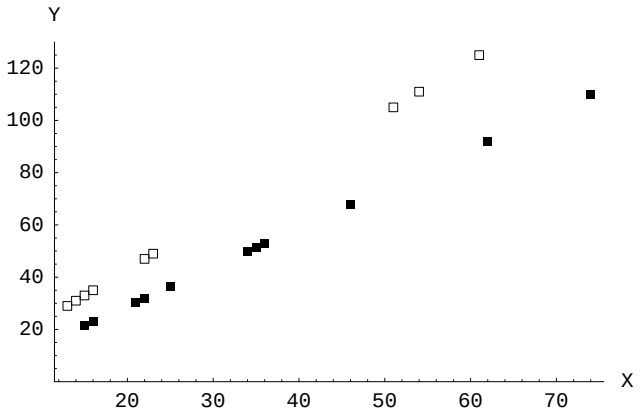
El mínimo de λ_c esta en $m = 15$, donde $X_{15} = 36.3$. Dado que $\lambda_c \leq \lambda^*$ hay evidencia suficiente para concluir que una vez que el oxigeno inhalado esta por encima de los 36.3 litros por minuto, la relación entre el dióxido de carbono y el oxigeno cambia.

Estudio de simulación 6
Sea Y una variable explicativa y X una variable respuesta para una muestra de 20 observaciones relacionadas por el modelo,

$$Y_i = \begin{cases} -1 + 1.5X_i + \varepsilon_i, & \text{si } i \leq 11 \\ 3 + 2X_i + \varepsilon_i, & \text{si } i > 11 \end{cases} \qquad con \qquad \varepsilon_i \sim N(0, \sigma^2).$$

donde el punto de cambio estructural se encuentra en $m = 11$.

Los resultados de la simulación son,



	σ^2			
	0.01	0.1	1	10
$E(\hat{m})$	11	11	11	10.76
$V(\hat{m})$	0	0	0	1.27

En los ejemplos en donde se sabe la ocurrencia del cambio, los métodos propuestos para calcular la probabilidad del estadístico de prueba detectan bien cuando hay cambio. Un resultado sobre estos métodos es que por Monte Carlo el valor de λ^*_{MC} calculado siempre es casi igual, es decir, si el algoritmo se corría 1000, 5000 o 50000 veces λ^*_{MC} variaba muy poco. En cambio, por integración numérica primero se calculo λ^*_{IN} para un intervalo dividido en 5 partes y después se hizo con uno dividido en 6 partes, y lo que se encontró fue que entre mayor sea el número de partes en las que se divide el intervalo, el valor de λ^*_{IN} tiende a ser más pequeño, por lo que se puede inferir que hay una mejor aproximación al valor real λ^*.

Además, como se observa en los ejemplos 1, 2, 4 y 6 donde se estima el punto de cambio, al parecer el estimador es sesgado y el sesgo se carga hacia el lado, con respecto a m, que tiene más datos, esto es, $\hat{m}$ tiende a colocarse en la muestra dividida en dos submuestras por m del lado de la submuestra con más datos.

Codigo para determinar la región crítica mediante el método de Monte Carlo

```
PAQUETES
<<LinearAlgebra‘Orthogonalization
<<LinearAlgebra‘MatrixManipulation
<<Statistics‘ContinuousDistributions

MATRIZ DE DATOS
X;                Matrix de información de las variables explicativas
dim=Dimensions[X];

CALCULO DE LA MATRIZ P (ORTONORMALIZACIÓN DE GRAM SCHMIDT)
B=IdentityMatrix[dim[[1]]]-X.Inverse[Transpose[X].X].Transpose[X];
u=Range[dim[[1]]]; z={}; v={};
If[u[[1]]==Table[0,{dim[[1]]}],z=AppendTo[z,u[[1]]],v=AppendTo[v,u[[1]]]];
dv=Dimensions[v]; For[k=2, k<=dim[[1]],
 u[[k]]=Chop[(B[[k]]-Sum[(N[(v[[i]].B[[k]])])*v[[i]],{i,1,dv[[1]]}])];
  If[u[[k]]==Table[0, {dim[[1]]}], z=AppendTo[z, u[[k]]],
 u[[k]]=u[[k]]/Norm[u[[k]]]; v=AppendTo[v, u[[k]]]];
 dv=Dimensions[v];
k++]

Pt=v; P=Transpose[Pt];

CALCULO DE LA MATRIZ  Rm=X*Inversa[Transpuesta(X)*X]*Transpuesta[X]
calculaRm[g_List, m_Integer]:=
 Module[{d, x1, x2, i, B1, B2, B3, B4},
  d=Dimensions[g]; x1 = {}; x2 = {};
  For[i=1, i<=m, x1=AppendTo[x1, g[[i]]]; i++];
  For[i=m+1, i<=d[[1]], x2=AppendTo[x2, g[[i]]]; i++];
  B2=ZeroMatrix[m, d[[1]]-m];
  B3=ZeroMatrix[d[[1]]-m, m];
  If[m<d[[2]],
   B1=ZeroMatrix[m, m];
   B4=x2.Inverse[Transpose[x2].x2].Transpose[x2],
   If[d[[2]]<=m<=d[[1]]-d[[2]],
    B1=x1.Inverse[Transpose[x1].x1].Transpose[x1];
    B4=x2.Inverse[Transpose[x2].x2].Transpose[x2],
    B1=x1.Inverse[Transpose[x1].x1].Transpose[x1];
    B4=ZeroMatrix[d[[1]]-m, d[[1]]-m]]];
 Rm=BlockMatrix[{{B1, B2}, {B3, B4}}];];

listaRm={};
 For[i=1, i<=dim[[1]]-1,
    calculaRm[X, i];
 listaRm=AppendTo[listaRm, Rm];i++]

CALCULO DE LAMBDA*
iter=50000; EstPrueb={}; T={}; M=Range[iter];
ndist=NormalDistribution[0,1];
 For[h=1, h<=iter,
 W=RandomArray[ndist, dim[[1]]-dim[[2]]];
 For[j=dim[[2]], j<=dim[[1]]-dim[[2]],
```

```
  Qm=Pt.listaRm[[j]].P;
  Sm=W.Qm.W/W.W;
  EstPrueb=AppendTo[EstPrueb, Sm]; j++;];
 M[[h]]=Max[EstPrueb];
 T=AppendTo[T,M[[h]]];
 EstPrueb={};
h++] Tmax=Sort[T]; lambda*=1-Tmax[[2500]];
```

Codigo para determinar la región crítica mediante Integración numérica

```
paso=1/3; dp=6; n=13; p1=1;pp=Pi^((n-p1-1)*0.5); MF={}; MaxEP={};
 For[i1=1, i1<=dp,
  v1=i1*paso-(1+(paso/2));
  For[i2=1, i2<=dp,
   v2=i2*paso-(1 + (paso/2));
   For[i3=1, i3<=dp,
    v3=i3*paso-(1+(paso/2));
    For[i4=1, i4<=dp,
     v4=i4*paso-(1+(paso/2));
     For[i5=1, i5<=dp,
      v5=i5*paso-(1+(paso/2));
      For[i6=1, i6<=dp,
       v6=i6*paso-(1+(paso/2));
       For[i7=1, i7<=dp,
        v7=i7*paso-(1+(paso/2));
        For[i8=1, i8<=dp,
         v8=i8*paso-(1+(paso/2));
         For[i9=1, i9<=dp,
          v9 = i9*paso-(1+(paso/2));
          For[i10=1, i10<=dp,
           v10=i10*paso-(1+(paso/2));
           vv={v1,v2,v3,v4,v5,v6,v7,v8,v9,v10};
           If[Sum[vv[[g]]^2,{g,1,nn}]<1, EstPrueb={};
            For[f=dim[[2]], f<=dim[[1]]-dim[[2]],
              uu=Flatten[{vv, Sqrt[1-Sum[vv[[g]]^2,{g,1,nn}]];
              Qm=Pt.listaRm[[f]].P;  Sm=uu.Qm.uu;
              EstPrueb=AppendTo[EstPrueb, Sm]; uu={};
             f++;];
            MaxEP=AppendTo[MaxEP, Max[EstPrueb]];
            F=Gamma[(n-p1-1)/2]/(2*(pp)*Sqrt[1-Sum[vv[[g]]^2,{g,1,nn}]]
            MF=AppendTo[MF, F]]; resul={MaxEP, MF};
           i10++];
          i9++];
         i8++];
        i7++];
       i6++];
      i5++];
     i4++];
    i3++];
   i2++];
  i1++];
resul1=Transpose[resul]; resul2=Sort[resul1]; inc=paso^nn; dr=Dimensions[resul2];

For[ii=1, ii<=dr[[1]], Suma=Sum[resul2[[tt]][[2]],{tt,1,11}]*inc;
If[Suma>=0.05,Print[Suma]; Print[resul2[[ii]][[1]]]; Break[];];
ii++]
```

Codigo para el punto de cambio.

```
PAQUETES
<<LinearAlgebra‘MatrixManipulation‘
<<Statistics‘ContinuousDistributions‘

X;
dim=Dimensions[X];

calculaError[g_List, h_List, m_Integer]:=
Module[{dim,X1,X2,Y1,Y2,i,SSE1,SSE2},
 dim=Dimensions[g];
 X1={}; X2={}; Y1={}; Y2={};
 For[i = 1, i <= m,AppendTo[X1,X[[i]]];AppendTo[Y1,Y[[i]]];i++];"\\
 For[i = m + 1, i <= dim[[1]], AppendTo[X2, X[[i]]];
 AppendTo[Y2, Y[[i]]]; i++];
 If[m <= dim[[2]],
  SSE1=0;
  SSE2=Y2.(IdentityMatrix[dim[[1]]-m]-
       X2.Inverse[Transpose[X2].X2].Transpose[X2]).Y2,
  If[dim[[2]]<m<dim[[1]]-dim[[2]],
  SSE1=Y1.(IdentityMatrix[m]-X1.Inverse[Transpose[X1].X1].Transpose[X1]).Y1;
  SSE2=Y2.(IdentityMatrix[dim[[1]]-m]-
       X2.Inverse[Transpose[X2].X2].Transpose[X2]).Y2,
  SSE1=Y1.(IdentityMatrix[m]-X1.Inverse[Transpose[X1].X1].Transpose[X1]).Y1;
  SSE2=0];];
Error={SSE1,SSE2};];

Punto de cambio=m;
beta1={a,b};
beta2={c,d};
ndist=NormalDistribution[0, 0.01];
Epocas=1000; SLEPosicion={};
SLEmin={};
 For[z=1, z<=Epocas,
 e=RandomArray[ndist, dim[[1]]-dim[[2]]];
 Y1={};  For[j=1, j<=m, Y=X[[j]].beta1+e[[j]]; AppendTo[Y1, Y]; j++];
 Y2={};  For[j=m+1,j<=dim[[1]],Y=X[[j]].beta2+e[[j]]; AppendTo[Y2,Y];j++];
 Y=Join[Y1, Y2];
 listaError={}; SLE={};
  For[i=1, i<=dim[[1]]-1,
  calculaError[X, Y, i];
  AppendTo[listaError, Error];
  l=(listaError[[i]][[1]] + listaError[[i]][[2]]);
  SLE=AppendTo[SLE,l]; i++];
 SLEPosicion=AppendTo[SLEPosicion, Position[SLE, Min[SLE]]];
 z++]
```

Conclusiones

Este trabajo esta dedicado al estudio del cambio estructural bajo un modelo de regresión lineal en varias variables. Por ello, en el primer capítulo se dio un panorama general de algunos conceptos básicos de la estadística inferencial, principalmente se repasó el método de máxima verosimilitud y el lema de Neyman Pearson que fue utilizado en el capítulo 4. Así mismo, en el capítulo 2 y 3 se expusieron algunos resultados de algebra matricial, de la teoría de la distribución normal y del modelo de regresión lineal para el caso multivariado, todo ello con la finalidad de fincar las bases para el desarrollo de nuestro estudio.

Uno de los objetivos fue determinar si en un periodo de observación la relación entre la variable a explicar y las variables explicativas modifico su comportamiento, concepto que en este trabajo se conoce como cambio estructural.
Para determinar esto, se propuso para un modelo de regresión de una variable dependiente con p variables explicativas la prueba de hipótesis,

$$H_0 : m = n \quad \text{versus} \quad H_1 : m < n$$

con $p+1 \leq m \leq n-p-1$, la cual indica que bajo el supuesto de que la hipótesis nula es cierta, entonces, no existe un punto de cambio en la región de observación, y que bajo el supuesto de que la hipótesis nula sea falsa, entonces, se manifestó un cambio en la región de observación. Aplicando la prueba se obtuvo el estadístico,

$$\max_m \frac{W^T Q_m W}{W^T W} \geq 1 - \lambda^*.$$

Después de hacer algunos cálculos, se encontró que la función de densidad asociada a dicho estadístico es

$$f_V(v_1, v_2, \ldots, v_{n-p-2}) = \begin{cases} \frac{\Gamma((n-p-1)/2)}{2(\pi)^{(n-p-1)/2}(1-v^T v)^{1/2}}, & \text{si } v^T v < 1 \\ 0, & \text{en otro caso.} \end{cases}$$

la cual no depende de ningún parámetro desconocido, por tanto, si se utiliza esta función de densidad la región crítica sólo dependerá de los valores de la muestra.

A continuación se desarrollaron dos programas para calcular la región de rechazo.

El primero de ellos se hizo mediante el método de Monte Carlo y el segundo con integración numérica, ambos métodos se aplicaron a ciertos problemas y el resultado que se obtuvo fue muy

satisfactorio porque en ambos métodos con la región de rechazo que se obtuvo se detectó bien cuando en los datos existe y cuando no existe un cambio estructural en la región de observación.

La integración numérica para encontrar el valor de λ^* se efectuó únicamente para ejemplos con pocos datos, pues el número de operaciones requeridas aumenta considerablemente cuando la muestra crece. Con un equipo de cómputo de mayor capacidad se puede obtener la región crítica para muestras grandes. Problema que se resuelve si se utilizar el método de Monte Carlo puesto que éste no tiene restricción alguna sobre el número de observaciones de la muestra.

Por el método de Monte Carlo el valor de λ^* calculado no varía aunque el algoritmo se corra 1000, 5000 o 50000 veces. Mientras que, por integración numérica entre mayor sea el número de partes en las que se divide el intervalo, el valor de λ^* tiende a disminuir, por lo que se puede inferir que hay una mejor aproximación al valor real λ^*.

En cuanto a la estimación del punto de cambio estructural, se propuso como estimador,

$$\hat{m} = \{m \mid SCE_{1_m} + SCE_{2_m} \leq SCE_{1_k} + SCE_{2_k}, k = p+1, ..., n-p-1\}$$

El cual indica que el punto de cambio se estima como el punto en donde la suma de SCE_{1_m} y SCE_{2_m} sea mínima. Así mismo, para sondear si el estimador es insesgado se efectuó una simulación, la cual se corrió 1000 veces y se encontró que el estimador es sesgado, puesto que $\hat{m}$ tiende a colocarse en la muestra dividida en dos submuestras por m del lado de la submuestra con más datos.

Lo importante de este trabajo fue que se propuso un estadístico de prueba utilizando la razón de verosimilitud y se determino la forma exacta de la distribución de probabilidad del estadístico; además de dar dos formas de obtener una aproximación de la integral asociada a dicha función de distribución. Así mismo, se propuso un estimador del punto de cambio estructural el cual presenta un sesgo pequeño.

Bibliografía

[1] Acuña F. E. (2006). ***Regresión Lineal Multiple*** *www.math.uprm.edu/ edgar/cap2sl.ppt*

[2] Antoch J. and Hušková M. (2001). ***Permutation tests in change point analysis***. Statistics and Probability Letters, Vol. 53, pp. 37-46.

[3] Beckman and Cook (1979). ***Testing for Two-Phase Regressions***. Thechnometrics, Vol. 21, No. 1.

[4] Broemeling, L. D. y Tsumuri, H. (1987). ***Econometric and Structural Change***, Marcel Dekker, Nueva York, pp. 6-24.

[5] Casella G. y Berger R. L. (2002). ***Statistical Inference***, p. 62.

[6] Chow. (1960). ***Tests of equality between sets of coefficients in two linear regressions***. Econometrica, Vol. 52, pp. 211-222.

[7] Daniel W. (1999). ***Bioestadística: Base para el análisis de las ciencias de la salud***, Cap.6.

[8] Draper N. (1966). ***Applied Regression Analysis*** Ed. John Wiley and Sons.

[9] Freund J. E. (2000). ***Estadística Matemática con Aplicaciones***, Cap. 12.

[10] Greene W. H. (1999). ***Análisis Econométrico***. Ed. Prentice Hall.

[11] Grossman S. I. (1996) ***Álgebra Lineal***. Ed. McGraw Hill, pp. 51, 122, 535.

[12] Gujarati D. N. (1997). ***Econometría***, Ed. McGraw Hill, pp. 258-260.

[13] Horvárth L. and Shao Q.M. (1993). ***Limit theorems for the union-intersection test***. Journal of statistical planning and inference, Vol. 44, pp. 133-148.

[14] Kreysig E. (1981). ***Introducción a la Estadística Matemática***, Limusa, p. 245. [1985, pp. 225-226]

[15] Maddala G.S. (1996). ***Introducción a la Econometría***, Thompsom, Cap. 73,74,123.

[16] Mendenhall III, W. ... (2002). ***Estadística Matemática con Aplicaciones***, Thompsom, Cap. 9.

[17] Muggeo, V. M. R. (2003). ***Estimating regression models with unknown breakpoints***. Statistics in Medicine, Vol. 22, pp. 3055-3071.

[18] Novales C. A. (1997). ***Estadística y Econometria***,
Mc Graw Hill, p.572.

[19] Chenery, H. (1980). ***Cambio estructural y política de desarrollo***, Tecnos, Madrid, (1ª ed. 1979).

[20] Peña D. (2002). ***Análisis de datos multivariantes***, Ed. Mc Graw Hill, Cap. 2.

[21] Pulido, A. y Fontela, E. (1993). ***Análisis input-output. Modelos, datos y aplicaciones***, Ed. Pirámide, Madrid, pp. 150-165.

[22] R. Spiegel M. (1976). ***Probabilidad y estadística***. Serie Schaum, Ed. Mc Graw Hill, pp. 80.

[23] Searle S. R. (1971). ***Linear Models***, Ed. John Wiley and Sons, Cap. 2.

[24] Sen A. and Srivastava M.S.(1975). ***On Tests for Detecting Change in Mean***, Annals of Statistics, Vol. 3, pp. 98-108.

[25] Steven A. J. (2001). ***Inference and estimation in a changepoint regression problem***. The Statistician, Vol. 50, Part. 1, pp.51-61.

[26] Worsley (1983). ***Testing for a Two-Phase Multiple Regression***. Thechnometrics, Vol. 25, No. 1, pp. 35-42.