UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

" EL ANALISIS DE LA MULTICOLINEALIDAD Y DE LA HETEROCEDASTICIDAD EN EL MODELO ECONOMETRICO LINEAL "

T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I O

P R E S E N T A :

MIGUEL ANGEL COVARRUBIAS ZUÑIGA

MEXICO, D. F. ENERO DE 1996

VNIVERSIDAD NACIONAL
AVPNOMA DE
MEXICO

M. en C. Virginia Abrín Batule
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
P r e s e n t e

Comunicamos a usted que hemos revisado el trabajo de Tesis:
"EL ANALISIS DE LA MULTICOLINEALIDAD Y DE LA HETEROCEDASTICIDAD EN EL MODELO ECONOMETRICO LINEAL".
realizado por MIGUEL ANGEL COVARRUBIAS ZUÑIGA
con número de cuenta 8018814-4 , pasante de la carrera de ACTUARIA
Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	ACT. YOLANDA SILVIA CALIXTO GARCIA
Propietario	ACT. FRANCISCO SANCHEZ VILLARREAL
Propietario	ACT. GABRIEL NUÑEZ ANTONIO
Suplente	ACT. BENIGNA CUEVAS PINZON
Suplente	ACT. GERARDO NUÑEZ

Consejo Departamental de Matemáticas

FACULTAD DE CIENCIAS
CONSEJO DEPARTAMENTAL
DE
MATEMATICAS

M. EN C. ALEJANDRO BRAVO MOJICA

A mis padres

Prologo

En este trabajo hemos desarrollado el modelo clásico de regresión lineal múltiple bajo los siguientes supuestos: Caso A : **e** está distribuida como $\mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$ donde σ^2 es desconocido. Caso B : e es una variable aleatoria tal que : $E[\mathbf{e}]=0$ y $Cov[\mathbf{e}]=E[\mathbf{ee}^t]=\sigma^2\mathbf{I}$

Para el caso A: obtuvimos para los estimadores de los coeficientes de regresión lineal, intervalos de confianza, además evaluamos las pruebas de hipótesis para los coeficientes de regresión.

Para el caso B: obtuvimos los estimadores de mínimos cuadrados ordinarios de los coeficientes de regresión y además que estos son los mejores linealmente insesgados

En el capítulo cuatro analizamos el modelo de regresión lineal cuando no se cumple que el supuesto de que las variables explicativas no están corrrelacionadas. En este capítulo observamos que existen varias reglas para detectar la multicolinealidad aunque para saber cuál de estas reglas hay que utilizar en la práctica tendremos que considerar la naturaleza de los datos y la severidad de la multicolinealidad.

En el capítulo cinco, analizamos el modelo de regresión lineal cuando no se cumple el supuesto de homocedasticidad. La heterocedasticidad no destruye el insesgamiento de los estimadores por mínimos cuadrados ordinarios , sin embargo estos estimadores no poseen varianza mínima , no siendo por tanto eficientes. Para detectar la heterocedasticidad generalmente se basa en examinar los residuos obtenidos para buscar en ellos patrones sistemáticos.

Contenido

Capítulo 1

Introducción

La econometría es la aplicación de la estadística a los datos económicos con el objeto de proporcionar no solo un apoyo empírico a los modelos construidos por la economía matemática, sino una forma de obtener resultados numéricos.

El trabajo del econometrista consiste en encontrar un conjunto de supuestos que sean suficientemente específicos y realistas, de tal manera que le permitan aprovechar de la mejor manera posible los datos que tiene a su disposición.

La econometría es una mezcla de teoría económica, economía matemática y estadística.

La teoría económica hace afirmaciones o formula hipótesis de naturaleza principalmente cualitativa. La econometría proporciona el contenido empírico a la mayoría de las

teorías económicas.

La economía matemática consiste en expresar la teoría económica en forma matemática, sin prestar atención a la medición ni la verificación empírica de la teoría. La econometría hace uso de las ecuaciones propuestas por el economista, pero de tal forma que éstas puedan estar sujetas a pruebas o comprobaciones de tipo empírico. Esta conversión de ecuaciones matemáticas a ecuaciones econometristas requiere ingenio y destreza.

La estadística centra su atención en la recolección, procesamiento y presentación de cifras económicas. Aunque la estadística proporciona la mayor parte de la herramienta utilizada en econometría, a menudo el econometrista requiere de métodos especiales en virtud del carácter sui generis de la mayor parte de la cifras económicas, debido a que estas no son resultado de un experimento controlado.

1.1 Objetivos de la Econometría

El objetivo de la econometría es la especificación del modelo que prueba la teoría.

Los atributos que deben tener un buen modelo son:

a) Parsimonia. Un modelo nunca puede llegar a ser una descripción completamente exacta de la realidad; para describir la realidad exactamente se tendría que desarrollar un modelo tan complejo que no sería útil en la práctica. Por lo tanto, un modelo debe ser una abstracción de la realidad. El *Principio de Occam o Principio de Parsimonia* enuncia que un modelo debe ser tan simple como sea posible. Lo anterior implica que se deben introducir solamente las variables necesarias en el modelo, relegando todas las

influencias menores y aleatorias al término del error ϵ_i.

b) Consistencia Teórica. Si el objetivo inmediato es tomar una decisión acerca de un problema económico, el modelo debe ser consistente con las partes pertinentes de la teoría económica establecida. Por otra parte, si el objetivo inmediato es buscar la verdad poniendo a prueba la teoría económica, el modelo que se utilice debe de ser consistente con la teoría económica que se trate de poner a prueba, ya que se halle bien establecida o una nueva o muy tentativa. En cualquier caso, se desea manejar un modelo que exprese o al menos sea consistente con las partes de la teoría económica.

c) Poder predictivo. Lo que se busca con un modelo es predecir el futuro.

1.2 Selección de Variables en Regresión Múltiple

En la práctica, aunque se especifica que **Y** depende de $x_1, ..., x_p$, no todos los coeficientes de estas variables pueden estimarse con una precisión razonable. Por ello tenemos que considerar que variables incluir y cuáles excluir. Se han desarrollado algunos procedimientos para añadir y eliminar variables sistemáticamente.

Los procedimientos son:

a) Todas las regresiones.

b) Procedimiento de eliminación hacia atrás.

c) Procedimiento de selección hacia adelante.

d) Regresión por etapas.

1.3 Diferencia Entre Relaciones Estadísticas y Deterministas

La diferencia entre las relaciones estadísticas y las deterministas, las primeras trata con variables aleatorias y en las deterministas no. Por ejemplo, en las primeras la dependencia de los niveles de producción de un cultivo con respecto a la temperatura, la lluvia, el sol, y la fertilidad tienen una relación estadística en el sentido de que las variables explicativas, aunque son muy importantes, no permite al agrónomo predecir la producción de el cultivo en forma exacta, debido a los errores involucrados en la medición de estas variables, como también debido a otra serie de factores, que también afectan la producción pero que pueden ser difíciles de identificar. Por lo tanto, se va a presentar cierta variabilidad aleatoria en la variable dependiente producción del cultivo.

Por otra parte, en las relaciones deterministas tratamos con variables no aleatorias. Por ejemplo, la ley de gravitación universal de Newton, que dice: toda partícula en el universo atrae a otra partícula con una fuerza que es directamente proporcional al producto de sus masas e inversamente proporcional al cuadrado de la distancia que las separa y que se expresa como

$$F = k\frac{m_1 m_2}{r^2}$$

donde F = fuerza, m_1 y m_2 son las masas de las partículas y k = constante de gravitación.

Sin embargo, es preciso aclarar que si existen errores de medición, digamos en el valor de k, entonces la relación que de otra manera sería determinista se convertiría en relación estadística. La razón radica en que bajo esta situación la fuerza puede predecirse únicamente en forma aproximada, a partir de un valor dado de k (m_1 y m_2 y r) que contiene errores. Entonces la variable F, en este caso, se convierte en una variable aleatoria.

1.4 El Análisis de Regresión no Implica que Exista una Relación de Causalidad

Aunque el análisis de regresión tiene que ver con la dependencia de una variable con relación a otras variables, esto no implica necesariamente que exista una relación de causalidad. Por ejemplo, todos sabemos que existe una relación entre la altura y el peso de los seres humanos, pero ¿implica esta relación, que pueda cambiar la altura de una persona si se modifica su peso?

Con lo cual, una relación estadística no puede por si misma implicar en forma lógica una relación de causalidad. Para atribuir causalidad se debe hacer uso de consideraciones teóricas o a priori.

1.5 Diferencia entre el Análisis de Regresión y el de Correlación

Aunque el análisis de correlación esta estrechamente relacionado con el análisis de regresión, conceptualmente los dos son muy diferentes. En el análisis de correlación, el objetivo fundamental es la medición de la fuerza o grado de asociación lineal entre dos o más variables. El coeficiente de correlación, mide esta fuerza de asociación. En el análisis de regresión no estamos fundamentalmente interesados en este tipo de medición. En lugar de ello, se intenta estimar o predecir el valor promedio de una variable con base a los valores fijos de otras variables. Las dos técnicas de regresión y correlación tienen ciertas diferencias fundamentales que vale la pena mencionar. En el análisis de regresión existe una diferencia en la manera como se manejan las variables dependiente y explicativas. Se supone que la variable dependiente es aleatoria y las variables explicativas tienen valores fijos. Por otra parte, en el análisis de correlación las variables se manejan indistintamente; no existe distinción alguna entre las variables dependientes y las explicativas.

Capítulo 2

Antecedentes Matemáticos

2.1 Conceptos Matemáticos

Teorema 1 *Una matriz* $\mathbf{A}$ *de tamaño* $n \times n$ *es singular si y sólo si* $\det(\mathbf{A}) = 0$.

Teorema 2 *Una matriz de tamaño* $n \times n$ *es no singular si y sólo si* $\mathbf{A}$ *no tiene inversa.*

Teorema 3 *Si* $\mathbf{A}$ *es una matriz de tamaño cualquiera entonces la matriz transpuesta de* $\mathbf{A}^t$ *es igual a* $\mathbf{A}$; *es decir,* $(\mathbf{A}^t)^t = \mathbf{A}$.

Teorema 4 *Si* $\mathbf{A}$ *es una matriz no singular de tamaño* $n \times n$ *entonces* $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.

Teorema 5 *Si* $\mathbf{A}$ *es una matriz no singular de tamaño* $n \times n$ *las operaciones de transponer y obtener inversa de una matriz pueden ser permutados, es decir,* $(\mathbf{A}^{-1})^t = (\mathbf{A}^t)^{-1}$.

Teorema 6 *Sean* $\mathbf{A}$ *y* $\mathbf{B}$ *dos matrices cualesquiera tal que el producto* $\mathbf{AB}$ *este definido entonces* $(\mathbf{AB})^t = \mathbf{B}^t\mathbf{A}^t$.

Teorema 7 *Si* $\mathbf{A}$ *y* $\mathbf{B}$ *son dos matrices no singulares de tamaño* $n \times n$ *entonces* $\mathbf{AB}$ *tiene inversa y* $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

Teorema 8 *Si* $\mathbf{A}$ *es matriz cualquiera entonces* $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$, *donde* $\mathbf{I}$ *es la matriz identidad.*

Teorema 9 *Si* $\mathbf{A}$ *es una matriz de tamaño* $n \times n$ *y rango*$(\mathbf{A}) < n$ *entonces las columnas como los renglones de* $\mathbf{A}$ *son linealmente dependientes.*

Teorema 10 *Si* $\mathbf{A}$ *es una matriz de tamaño* $n \times n$ *y rango*$(\mathbf{A}) = m < n$ *entonces el número de columnas o renglones linealmente independientes de* $\mathbf{A}$ *es* m.

Definición 1 *Si* $\mathbf{A}$ *es una matriz de* $n \times n$ *entonces la traza de* $\mathbf{A}$ *es la suma de los elementos de la diagonal; y se denota por* $Tr(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$.

Teorema 11 *Si* $\mathbf{A}$ *y* $\mathbf{B}$ *son dos matrices de tamaño* $n \times m$ *y* $m \times n$ *respectivamente entonces* $tr(\mathbf{AB}) = tr(\mathbf{BA})$.

Teorema 12 $Tr(\mathbf{I}) = n$, *donde* $\mathbf{I}$ *es la matriz identidad de* $n \times n$.

Definición 2 *Una matriz es idempotente si y sólo si y* $\mathbf{A} = \mathbf{A}^2$.

2.2 Conceptos Estadísticos

Teorema 13 *Si* $\mathbf{A}$ *es una matriz no singular y definida positiva entonces existe una matriz* $\mathbf{B}$ *simétrica de tamaño* $n \times n$ *tal que* $\mathbf{A} = \mathbf{B}^t\mathbf{B}$.

Teorema 14 *Si* $\mathbf{Y}$ *esta distribuido como una normal con media* $\mathbf{0}$ *y matriz de covarianza* $\sigma^2\mathbf{I}$ *entonces* $E[\mathbf{Y}^t\mathbf{A}\mathbf{Y}] = \sigma^2 tr(\mathbf{A})$

Teorema 15 *Si* $\mathbf{Y}$ *es un vector de* $p \times 1$ *y es distribuido normalmente con media* μ *y matriz de covarianza* $\mathbf{V}$ *y si* $\mathbf{B}$ *es una matriz de* $q \times p$ *donde* $q \leq p$ *y* $rango(\mathbf{B}) = q$ *entonces el vector* $\mathbf{Z} = \mathbf{B}\mathbf{Y}$ *también se distribuye como una normal.*

Teorema 16 *Si* $\mathbf{Y}$ *es distribuido como* $N(\mu, \sigma^2\mathbf{I})$ *entonces* $\frac{\mathbf{Y}^t\mathbf{A}\mathbf{Y}}{\sigma^2}$ *es distribuido como* $\chi'^2(k, \lambda)$ *donde* $\lambda = \frac{\mu^t\mathbf{A}\mu}{2\sigma^2}$ *si solo si* $\mathbf{A}$ *es una matriz idempotente de rango*$(\mathbf{A})$.

Teorema 17 *Si* $\mathbf{B}$ *es una matriz de* $q \times n$, $\mathbf{A}$ *es una matriz de* $n \times n$ *y* $\mathbf{Y}$ *es un vector distribuido como* $N(\mu, \sigma^2\mathbf{I})$ *entonces la forma lineal* $\mathbf{B}\mathbf{Y}$ *es independiente de la forma cuadratica* $\mathbf{Y}^t\mathbf{A}\mathbf{Y}$ *si* $\mathbf{B}\mathbf{A} = 0$.

Teorema 18 *Sea Y una variable aleatoria distribuida como $N(\mu, \sigma^2 I)$ y*

$$\sum^{k} Y^t A_i Y = \mathbf{Y}^t \mathbf{Y}$$

donde el rango$(A_i) = n_i$ si se cumplen las siguientes condiciones:

i) A_i **es una matriz idempotente para** $i = 1, \ldots, k$

ii) $A_i A_j = 0$ **para** $i \neq j$

iii) $\sum_{i=1}^{k} n_i = n$

entonces

i) $\frac{Y^t A_i Y}{\sigma^2}$ **es distribuida como** $\chi'^2(n, \lambda)$ donde

$$\lambda_i = \frac{\mu A_i \mu}{2\sigma^2}$$

ii) $Y^t A_i Y$ y $Y^t A_j Y$ son independientes.

Definición 3 Distribución Normal. *Si X es una variable aleatoria con función de densidad de probabilidad*

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-(x-\mu)^2/2\sigma^2} \quad \textit{donde} \quad -\infty < \mu < \infty,\ \sigma > 0 \ \ y \ \ -\infty < x < \infty$$

entonces X tiene una función de distribución normal.

Definición 4 Distribución Ji-Cuadrada. *Si X es una variable aleatoria con función de densidad de probabilidad*

$$f_x(x) = \frac{1}{\Gamma(k/2)} \left(\frac{1}{2}\right)^{k/2} x^{k/2-1} \exp(-x/2)\, I_{(0,\infty)}(x)$$

entonces X tiene una distribución Ji-Cuadrada con k grados de libertad.

Teorema 19 *Si X es una variable aleatoria distribuida como una Ji-Cuadrada con k grados de libertad entonces su función generadora de momentos de la distribución Ji-Cuadrada es*

$$m_x(t) = \left(\frac{1}{1-2t}\right)^{k/2}$$

Demostración.

Como

$$m_x(t) = E[\exp(tx)] = \int e^{tx} \left(\frac{1}{\Gamma(k/2)}\right) \left(\frac{1}{2}\right)^{k/2} x^{k/2-1} e^{-x/2} dx$$

$$m_x(t) = \left(\frac{1/2}{1/2-t}\right)^{k/2} \int \frac{(1/2-t)^{k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-(1/2-t)x} dx$$

al efectuar un cambio de variable de manera tal que

$$u = (1/2 - t)\, x$$

entonces

$$x = \frac{u}{1/2 - t} \quad y \quad dx = \frac{du}{1/2 - t}$$

por lo tanto,

$$m_x(t) = \left(\frac{1/2}{1/2 - t}\right)^{k/2} \left(\frac{1}{\Gamma(k/2)}\right) \int u^{k/2-1} e^{-u} du$$

y como $\Gamma(\alpha) = \int u^{\alpha-1} e^{-u} du$ entonces

$$\begin{aligned} m_x(t) &= \left(\tfrac{1/2}{1/2-t}\right)^{k/2} \\ &= \left(\tfrac{1}{1-2t}\right)^{k/2} \end{aligned}$$

Teorema 20 *Si $X_1, \ldots, X_k$ son variables aleatorias distribuidas normalmente e independientes con media μ_i y varianza σ_i^2 para $i = 1, \ldots, k$ entonces*

$$u = \sum_{i=1}^{k} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$$

tiene una distribución Ji-Cuadrada con k grados de libertad.

Demostración.

Si $Z_i = \dfrac{x_i - \mu_i}{\sigma_i}$ entonces Z_i tiene una distribución normal standard.

Con lo cual,

$$
\begin{aligned}
m_u(t) &= E\left[e^{tu}\right] \\
&= E\left[e^{t\sum_{i=1}^{k} Z_i^2}\right] \\
&= E\left[\prod_{i=1}^{k} e^{tZ_i^2}\right] \quad \textit{y por ser independientes las variables} \\
&= \prod_{i=1}^{k} E\left[e^{tZ_i^2}\right]
\end{aligned}
$$

por otro lado se tiene que

$$
\begin{aligned}
E\left[e^{tZ^2}\right] &= \int e^{tZ^2}\left(\frac{1}{\sqrt{2\pi}}\right)e^{-\frac{1}{2}Z^2}dz \\
&= \int\left(\frac{1}{\sqrt{2\pi}}\right)e^{-\frac{1}{2}(1-2t)Z^2}dz \\
&= \frac{1}{\sqrt{1-2t}}\int\frac{\sqrt{1-2t}}{\sqrt{2\pi}}e^{-\frac{1}{2}(1-2t)Z^2}dz
\end{aligned}
$$

como $\mu = 0$ y $\sigma^2 = \frac{1}{1-2t}$ entonces $\frac{1}{\sigma} = \sqrt{1-2t}$ y de esto se obtiene que

$$
\begin{aligned}
E\left[e^{tZ^2}\right] &= \frac{1}{\sqrt{1-2t}}\int\left(\frac{1}{\sqrt{2\pi}\sigma}\right)e^{-Z^2/2\sigma^2} \\
&= \frac{1}{\sqrt{1-2t}} \quad para\ t < \tfrac{1}{2}
\end{aligned}
$$

entonces

$$
\begin{aligned}
\prod_{i=1}^{k} E\left[e^{tZ^2}\right] &= \prod_{i=1}^{k}\left(\frac{1}{\sqrt{1-2t}}\right)^k \\
&= \prod_{i=1}^{k}\left(\frac{1}{1-2t}\right)^{k/2}
\end{aligned}
$$

y por el teorema anterior, observamos que la variable aleatoria U tiene la misma función generadora de momentos que la distribución Ji-Cuadrada con k grados de libertad. Por lo tanto U se distribuye como una Ji-Cuadrada.

Definición 5 Distribución F. *Si X es una variable aleatoria con función de densidad de probabilidad*

$$f_x(x) = \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)}\left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{x^{(m-2)/2}}{[1+(m/n)x]^{(m+n)/2}} I_{(0,\infty)}(x)$$

entonces X tiene una distribución F con m y n grados de libertad.

Teorema 21 *Sean U y V dos variables aleatorias con función de distribución Ji-Cuadrada con m grados de libertad y n grados de libertad respectivamente y además U y V son independientes entonces la variable aleatoria*

$$X = \frac{U/m}{V/n}$$

está distribuida como una función de distribución F con m y n grados de libertad.

Demostración.

Como U y V son independientes entonces su función de densidad conjunta es

$$\begin{aligned} f_{u,v}(u,v) &= f_u(u) f_v(v) \\ &= \frac{1}{\Gamma(m/2)\Gamma(n/2)2^{(m+n)/2}} u^{(m-2)/2} v^{(n-2)/2} e^{-\frac{1}{2}(u+v)} I_{(0,\infty)}(u) I_{(0,\infty)}(v) \end{aligned}$$

y para encontrar la función de distribución de X, haremos la transformación

$$X = \frac{U/m}{V/n} \quad y \quad Y = V$$

con lo cual se obtiene que

$$u = \left(\frac{m}{n}y\right)x$$

por lo tanto, el Jacobiano de la transformación es

$$\frac{du}{dx} = \left(\frac{m}{n}y\right)$$

y como

$$f_u(u) = \frac{1}{\Gamma(m/2)}\left(\frac{1}{2}\right)^{m/2} u^{m/2-1}e^{-u/2}I_{(0,\infty)}(u)$$

entonces

$$f_{x|y}(x \mid y) = \left(\frac{m}{n}y\right)\left[\frac{1}{\Gamma(m/2)}\left(\frac{1}{2}\right)^{m/2}\left(\frac{m}{n}xy\right)^{\left(\frac{m-2}{2}\right)}e^{-\frac{1}{2}\left(\frac{m}{n}xy\right)}\right]$$

Con lo cual la función de densidad conjunta de X y Y es

$$\begin{aligned} f_{x,y}(x,y) &= f_{x|y}(x \mid y)\, f_y(y) \\ &= \left(\tfrac{m}{n}y\right)\frac{1}{\Gamma(m/2)\Gamma(n/2)2^{(m+n)/2}}\left(\tfrac{m}{n}xy\right)^{(m-2)/2} y^{(n-2)/2}e^{-\frac{1}{2}\left[\left(\frac{m}{n}xy\right)+y\right]} \end{aligned}$$

y la función de densidad marginal de X es

$$
\begin{aligned}
f_x(x) &= \int f_{x,y}(x,y)\,dy \\
&= \frac{1}{\Gamma(m/2)\Gamma(n/2)2^{(m+n)/2}} \left(\frac{m}{n}\right)^{m/2} x^{(m-2)/2} \int y^{(m+n-2)/2} e^{-\frac{1}{2}\left[\left(\frac{m}{n}\right)x+1\right]y} dy
\end{aligned}
$$

Al efectuar el cambio de variable $u = \frac{1}{2}\left[\left(\frac{m}{n}\right)x + 1\right] y$ entonces $y = \frac{2u}{\left[\left(\frac{m}{n}\right)x+1\right]}$ y

$$
dy = \frac{2du}{\left[\left(\frac{m}{n}\right)x + 1\right]}
$$

obteniendo lo siguiente

$$
\begin{aligned}
f_x(x) &= \frac{1}{\Gamma(m/2)\Gamma(n/2)2^{(m+n)/2}} \left(\frac{m}{n}\right)^{m/2} x^{(m-2)/2} \int \left(\frac{2u}{\left(\frac{m}{n}\right)x+1}\right)^{(m+n-2)/2} e^{-u} \left(\frac{2}{\left(\frac{m}{n}\right)x+1}\right) du \\
&= \frac{1}{\Gamma(m/2)\Gamma(n/2)2^{(m+n)/2}} \left(\frac{m}{n}\right)^{m/2} \frac{x^{(m-2)/2}}{\left[(m/n)x+1\right]} \int u^{(m+n)/2-1} e^{-u} du
\end{aligned}
$$

Como $\Gamma(\alpha) = \int u^{\alpha-1} e^{-u} du$ entonces

$$
f_x(x) = \frac{\Gamma\left[(m+n)-2\right]}{\Gamma(m/2)\,\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} \frac{x^{(m-2)/2}}{\left[(m/n)\,x+1\right]^{}} I_{(0,\infty)}(x)
$$

Definición 6 Distribución t Student. *Si X es una variable aleatoria con función de densidad de probabilidad*

$$
f_x(x) = \frac{\Gamma\left[(k+1)/2\right]}{\Gamma(k/2)} \frac{1}{\sqrt{k\pi}} \frac{1}{(1+x^2/k)^{(k+1)/2}}
$$

entonces X tiene una distribución t Student.

Teorema 22 *Sean Z y U dos variables aleatorias que se distribuyen como una normal standard y como una Ji-Cuadrada con k grados de libertad, donde Z y U son independientes entonces la variable aleatoria $X = \frac{Z}{\sqrt{U/k}}$ tiene una distribución t Student con k grados de libertad.*

Demostración.

Puesto que Z y U son independientes entonces su función de densidad conjunta es

$$\begin{aligned} f_{z,u}(z,u) &= f_z(z) f_u(u) \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(k/2)} \left(\frac{1}{2}\right)^{k/2} u^{k/2-1} e^{-u/2} e^{-z^2/2} I_{(0,\infty)}(u) \end{aligned}$$

Para encontrar la función de distribución de X se hará la siguiente transformación

$$X = \frac{Z}{\sqrt{U/k}} \; y \; Y = U$$

obteniendo lo siguiente

$$Z = X\sqrt{Y/k}$$

por lo tanto, el Jacobiano de la transformación es

$$\frac{dz}{dx} = \sqrt{Y/k}$$

y como

$$f_z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

entonces

$$f_{x|y}(x \mid y) = \sqrt{y/k}\left(\frac{1}{\sqrt{2\pi}} e^{-x^2 y/k}\right)$$

Con lo cual, la función de densidad conjunta de X y Y es

$$\begin{aligned} f_{x,y}(x,y) &= f_{x|y}(x \mid y) f_y(y) \\ &= \sqrt{y/k}\frac{1}{\sqrt{2\pi}}\frac{1}{\Gamma(k/2)}\left(\frac{1}{2}\right)^{k/2} y^{k/2-1} e^{-y/2} e^{-x^2 y/k} \\ &= \frac{1}{\sqrt{k\pi}}\frac{1}{\Gamma(k/2)}\left(\frac{1}{2}\right)^{(k+1)/2} y^{(k+1)/2-1} e^{-\frac{1}{2}(1+x^2/k)y} \end{aligned}$$

obteniendo así la función de densidad marginal de X como

$$\begin{aligned} f_x(x) &= \int f_{x,y}(x,y)\, dy \\ &= \frac{1}{\sqrt{k\pi}}\frac{1}{\Gamma(k/2)}\left(\frac{1}{2}\right)^{(k+1)/2} \int y^{(k+1)/2-1} e^{-\frac{1}{2}(1+x^2/k)y} dy \end{aligned}$$

Al efectuar el cambio de variable

$$w = \frac{1}{2}\left(1 + x^2/k\right) y$$

entonces

$$y = \frac{2w}{1 + x^2/k} \quad y \quad dy = \frac{2}{1 + x^2/k} dw$$

Con lo cual obtenemos lo siguiente

$$\begin{aligned} f_x(x) &= \frac{1}{\sqrt{k\pi}}\frac{1}{\Gamma(k/2)}\left(\frac{1}{2}\right)^{(k+1)/2}\int\left(\frac{2w}{1+x^2/k}\right)^{(k+1)/2}e^{-w}\left(\frac{2}{1+x^2/k}\right)dw \\ &= \frac{1}{\sqrt{k\pi}}\frac{1}{\Gamma(k/2)}\frac{1}{(1+x^2/k)}\int w^{(k+1)/2-1}e^{-w}dw \\ &= \frac{\Gamma[(k+1)/2]}{\Gamma(k/2)}\frac{1}{\sqrt{k\pi}} \end{aligned}$$

2.3 Estimación Puntual

El problema de la estimación, consiste en obtener una muestra $X_1,...,X_n$ de variables aleatorias, con función de probabilidad $f_x(\cdot\,;\theta) = f(\cdot\,;\theta)$ conocida, pero el parámetro θ es desconocido.

Se supone además que los valores $x_1,...,x_n$ de la muestra aleatoria $X_1,...,X_n$ de la función de densidad de probabilidad $f(\cdot;\theta)$ pueden ser observados y en base a esos valores se estima el valor del parámetro desconocido θ, o el valor de alguna función de θ, es decir $\tau(\theta)$.

La estimación puntual presenta dos problemas que son los siguientes

i) Como obtener una estadística para usarla como estimador.

ii) Seleccionar un criterio o técnica para encontrar el mejor estimador entre todos.

Por lo tanto la estimación puntual consiste en dar un valor a alguna estadística $t(X_1, ..., X_n)$, la cual estima, el valor desconocida de $\tau(\theta)$.

La segunda, llamada estimación por intervalo y consiste en dar valor a dos estadísticas $t_1(X_1, ..., X_n)$ y $t_2(X_1, ..., X_n)$ donde $t_1(X_1, ..., X_n) < t_2(X_1, ..., X_n)$, con lo cual $(t_1(X_1, ..., X_n), t_2(X_1, ..., X_n))$ consiste en un intervalo, donde se determina la probabilidad de contener el valor desconocido de $\tau(\theta)$.

Definición 7 Estadística. *Una estadística es una función de los valores de las variables aleatorias observadas $x_1, ..., x_n$, la cual no contiene ningún parámetro desconocido.*

Definición 8 Estimador. *Un estimador es cualquier estadística cuyos valores son usados para estimar $\tau(\theta)$.*

Definición 9 Estimador Insesgado. *Un estimador $(T_1, ..., T_r)$, donde $T_j = t_j(X_1, ..., X_n)$ para $j = 1, ..., n$, es estimador insesgado de $(\tau_1(\theta), ..., \tau_r(\theta))$ si y sólo si $E[T_j] = \tau_j(\theta)$ para $j = 1, ..., n$ para todo $\theta \in \Theta$.*

Definición 10 Función de verosimilitud. *La función de verosimilitud de n variables aleatorias $X_1, ..., X_n$ es la función de densidad conjunta de las n variables, es decir $f_{x_1,...,x_n}(x_1,...,x_n;\theta)$, la cual es considerada como función de θ.*

Ejemplo 1 *Si $X_1, ..., X_n$ es una muestra aleatoria con función de densidad $f(x;\theta)$ entonces la función de verosimilitud es $L(\theta) = f(x_1;\theta)....f(x_n;\theta)$.*

Definición 11 Estimador Máximo Verosímil. *Sea* $L(\theta) = L(\theta; x_1, ..., x_n)$ *la función de verosimilitud para las variables aleatorias* $X_1, ..., X_n$. *Si* $\hat{\theta}$ *es el valor de* $\theta \in \Theta$, *el cual maximiza* $L(\theta)$ *entonces* $\hat{\Theta} = \hat{\vartheta}(X_1, ..., X_n)$ *es el estimador máximo verosímil de* θ *para la muestra* $x_1, ..., x_n$.

Definición 12 . Eficiente. *Sea* $T_1^*, ..., T_{n1}^*, ...$ *una secuencia de estimadores de* $\tau(\theta)$, *donde* $T_n^* = t_n(X_1, ..., X_n)$. *La secuencia de estimadores* $\left\{T_n^*\right\}$ *es eficiente si cumple con las siguientes condiciones:*

i) La distribución de $\sqrt{n}\left[T_n^* - \tau(\theta)\right]$ *es asintóticamente normal con media cero y varianza* $\sigma^{*2}(\theta)$.

ii) Si $\{T_n\}$ *es cualquier otra secuencia de estimadores para los cuales* $\sqrt{n}[T_n - \tau(\theta)]$ *es asintóticamente normal con media cero y varianza* σ^2.

entonces

iii) $\sigma^{*2}(\theta) < \sigma^2(\theta)$ *para todo* θ *en cualquier intervalo abierto.*

Definición 13 Consistente. *Sea* $T_1, ..., T_n, ...$ *una secuencia de estimadores de* $\tau(\theta)$, *donde* $T_n = t_n(X_1, ..., X_n)$. *La secuencia de estimadores* $\{T_n\}$ *es consistente. Si para cada* $\epsilon > 0$ *entonces*

$$\lim_{n \to \infty} P[\tau(\theta) - \epsilon < T_n < \tau(\theta) + \epsilon] = 1 \quad para \quad \theta \in \Theta.$$

Definición 14 El Mejor Estimador asintóticamente Normal. *Una secuencia de estimadores* $T_1^*, ..., T_n^*, ...$ *de* $\tau(\theta)$ *son los mejores asintóticamente si y sólo si cumplen las siguientes condiciones:*

i) La distribución de $\sqrt{n}\left[T_n^* - \tau(\theta)\right]$ es aproximadamente una distribución normal con media cero y varianza $\sigma^{*2}(\theta)$ cuando n tiende a infinito.

ii) Para todo $\epsilon > 0$

$$\lim_{n\to\infty} P\left[\left|T_n^* - \tau(\theta)\right| > \epsilon\right] = 0 \text{ para todo } \theta \in \Theta.$$

iii) Sea $\{T_n\}$ cualquier otra secuencia de estimadores consistentes para los cuales la distribución de $\sqrt{n}[T_n - \tau(\theta)]$ es aproximadamente una distribución normal con media cero y varianza $\sigma^2(\theta)$.

iv) $\sigma^2(\theta) > \sigma^{*2}(\theta)$ para todo θ en cualquier intervalo abierto.

Teorema 23 *Si la función de densidad* $f(x;\theta)$ *satisface las condiciones de regularidad y si* $\hat{\Theta}_n = \hat{\vartheta}(X_1, ..., X_n)$ *es el estimador máximo verosímil de* θ *para una muestra de tamaño n entonces se cumplen las siguientes condiciones:*

i) $\hat{\Theta}$ es distribuida asintóticamente normal con media θ y varianza

$$\frac{1}{nE\left[\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right]^2\right]}$$

ii) La secuencia de estimadores máximo verosímiles $\hat{\Theta}_1, ..., \hat{\Theta}_n$ son los mejores asintóticamente

normales.

Definición 15 Estadística Suficiente. *Sean $X_1, ..., X_n$ una muestra aleatoria de una función de densidad de probabilidad $f(\cdot;\theta)$. Cualquier estadística $S = s(X_1, ..., X_n)$ es una estadística suficiente si y sólo si la distribución condicional de cualquier estadística T dado S no depende de θ.*

Definición 16 Estadística suficiente conjuntamente. *Sean $X_1, ..., X_n$ una muestra aleatoria de una función de densidad de probabilidad $f(\cdot;\theta)$. Las estadísticas $S_1, ..., S_r$ son estadísticas suficientemente conjuntas si y sólo si la distribución condicional de $X_1, ..., X_n$ dado $S_1 = s_1, ..., S_r = s_r$ no depende de θ.*

Teorema 24 Factorización. *Sea $X_1, ..., X_n$ una muestra aleatoria con función de densidad de probabilidad $f_x(\ ;\theta)$, donde el parámetro puede ser un vector. Un conjunto de estadísticas $S_1 = s_1(X_1, ..., X_n), ..., S_r = s_r(X_1, ..., X_n)$ son conjuntamente suficientes si y sólo si la función de densidad de probabilidad conjunta de $X_1, ..., X_n$ puede ser factorizada como*

$$\begin{aligned} f_{x_1,...,x_n}(x_1,...,x_n;\theta) &= g(s_1(X_1,...,X_n),...,s_r(X_1,...,X_n);\theta)\,h(x_1,...,x_n) \\ &= g(s_1,...,s_r;\theta)\,h(x_1,...,x_n) \end{aligned}$$

donde la función $h(x_1, ..., x_n) > 0$ y $g(s_1, ..., s_r;\theta) > 0$ depende de $x_1,...,x_n$ solo a través de las funciones $s_1(x_1, ..., x_n), ..., s_r(x_1, ..., x_n)$.

Definición 17 Estadística Suficiente Mínimal. *Un conjunto de estadísticas suficientes es suficiente mínimal si y sólo si es función de cualquier otra estadística suficiente.*

Definición 18 Familia Exponencial con k parámetros. *Si una familia de funciones de densidad de probabilidad $f(\cdot;\theta)$ pueden expresarse como*

$$f(x;\theta_1,...,\theta_k) = a(\theta_1,...,\theta_k)\, b(x) \exp\left(\sum_{j=1}^{k} c_j(\theta_1,...,\theta_k)\, d_j(x)\right)$$

para una apropiada selección de funciones $a(\cdot,...,\cdot)$, $b(\cdot)$, $c(\cdot,...,\cdot)$ y $d_j(\cdot)$ para $j = 1,..,k$ entonces pertenece a la familia exponencial.

Definición 19 Estimador Insesgado de Varianza Mínima Uniforme. *Sea $X_1,...,X_n$ una muestra aleatoria con función de densidad de probabilidad $f(\cdot;\theta)$. Un estimador $T^* = t^*(X_1,...,X_n)$ de $\tau(\theta)$ es un estimador insesgado de varianza mínima uniforme de $\tau(\theta)$ si y sólo si*

i) $E[T^*] = \tau(\theta)$, es decir, T^* es un estimador insesgado.

ii) $Var[T^*] \leq Var[T]$ para cualquier estimador $T = t(X_1,...,X_n)$ de $\tau(\theta)$ el cual satisface que $E[T] = \tau(\theta)$.

Definición 20 Concentración Elipsoidal. *Sea $(T_1,...,T_r)$ estimadores insesgados de $(\tau_1(\theta),...,\tau_r(\theta$ Sea $\sigma^{ij}(\theta)$ el ij-ésimo elemento de la matriz inversa de la covarianza de $(T_1,...,T_r)$, donde ij-ésimo elemento de la matriz de covarianza es $\sigma_{ij}(\theta) = Cov[T_i,T_j]$. La concentración elipsoidal de $(T_1,...,T_r)$ es el interior y la elipsoide acotada por*

$$\sum_{i=1}^{r}\sum_{j=1}^{r} \sigma^{ij}(\theta)[T_i - T_i(\theta)][T_j - T_j(\theta)] = r + 2.$$

Definición 21 Familia de Funciones de Densidad de Probabilidad Conjuntamente Completas. *Sea $X_1, ..., X_n$ una muestra aleatoria con función de densidad de probabilidad $f(x;\theta_1,..,\theta_k)$ y sea $(T_1, ..., T_m)$ un conjunto de estadísticas. $T_1, ..., T_m$ son estadísticas completas si y sólo si $E[z(T_1, ..., T_m)] \equiv 0$ para todo $\theta \in \Theta$ implica que $p[z(T_1, ..., T_m) = 0] \equiv 1$ para todo $\theta \in \Theta$, donde $z(T_1, ..., T_m)$ es una estadística.*

Teorema 25 *Sea $X_1, ..., X_n$ una muestra aleatoria con función de densidad de probabilidad*

$$f(x;\theta_1,..,\theta_k) = a(\theta_1, ..., \theta_k)\, b(x) \exp\left(\sum_{j=1}^{k} c_j(\theta_1, ..., \theta_k)\, d_j(x)\right)$$

es decir, $f(x;\theta_1,..,\theta_k)$ es miembro de la familia exponencial, entonces

$$\left(\sum_{i=1}^{n} d_1(x_1), ..., \sum_{i=1}^{n} d_k(x_k)\right)$$

es un conjunto de estadísticas suficientes minimales y completas conjuntamente.

Teorema 26 *Sea $X_1, ..., X_n$ una muestra aleatoria con función de densidad de probabilidad $f(x;\theta_1,..,\theta_k)$ y sea $S_1 = s_1(X_1, ..., X_n), ..., S_m = s_m(X_1, ..., X_n)$ un conjunto de estadísticas suficientes conjuntamente.*

Sea $(T_1, ..., T_r)$ un estimador insesgado de $(\tau_1(\theta), ..., \tau_r(\theta))$ donde $\theta = (\theta_1, .., \theta_k)$.

Definamos $T_j' = E[T_j \mid S_1, ..., S_m]$ para $j = 1, .., k$ entonces

i) $\left(T_1', ..., T_r'\right)$ es un estimador insesgado de $(\tau_1(\theta), ..., \tau_r(\theta))$ y $T_j' = t_j'(S_1, .., S_m)$ es decir T_j' es una función de la estadísticas suficientes $S_1, ..., S_m$, para $j = 1, .., r$.

ii) $Var\left[T_j'\right] \leq Var[T_j]$ para cada $\theta \in \Theta$, para $j = 1, ..., r$.

iii) La concentración elipsoidal de $\left(T_1^*, ..., T_r^*\right)$ esta contenida en la concentración elipsoidal de $(T_1, ..., T_r)$ para cada $\theta \in \Theta$.

Teorema 27 Lehmann-Scheffe.

Sea $X_1, ..., X_n$ una muestra aleatoria con función de densidad de probabilidad $f(x; \theta_1, .., \theta_k)$.

Si $S_1 = s_1(X_1, ..., X_n), ..., S_m = s_m(X_1, ..., X_n)$ son estadísticas suficientes minimal y completas conjuntamente y si existe un estimador insesgado de $(\tau_1(\theta), ..., \tau_r(\theta))$ donde $\theta = (\theta_1, .., \theta_k)$ entonces existe un estimador único de $(\tau_1(\theta), ..., \tau_r(\theta))$, es decir, si $T_1^* = t_1^*(S_1, .., S_m), ..., T_r^* = t_r^*(S_1, .., S_m)$, donde cada t_j^* es función de $S_1, ..., S_m$ satisface que:

i) $Var\left[T_j^*\right] \leq Var[T_j]$ para cada $\theta \in \Theta$, donde $j = 1, .., r$, para cualquier estimador $(T_1, .., T_r)$ de $(\tau_1(\theta), ..., \tau_r(\theta))$.

ii) La concentración elipsoidal de $(T_1^*, ..., T_r^*)$ está contenida en la concentración elipsoidal de $(T_1, ..., T_r)$, donde $(T_1, ..., T_r)$ es cualquier estimador insesgado de $(\tau_1(\theta), ..., \tau_r(\theta))$.

2.4 Estimación por Intervalos

Intervalo de Confianza.

Sea $X_1, ..., X_n$ una muestra de una función de densidad de probabilidad $f(\cdot; \theta)$ y sea $T_1 = t_1(X_1, .., X_n)$ y $T_2 = t_2(X_1, .., X_n)$ dos estadísticas tales que $T_1 \leq T_2$ para las cuales $P[T_1 < \tau(\theta) < T_2] = \gamma$, donde $\gamma \in (0, 1)$ entonces el intervalo (T_1, T_2) es llamado

intervalo de confianza del 100γ ; γ es llamado coeficiente de confianza; T_1 y T_2 son llamados limite inferior y superior de confianza respectivamente para $\tau(\theta)$.

Al valor (t_1, t_2) del intervalo aleatorio (T_1, T_2) es llamado también intervalo de confianza del 100γ por ciento para $\tau(\theta)$.

2.5 Prueba de Hipótesis

Definición 22 Hipótesis estadísticas. *Una hipótesis estadística es una afirmación o conjetura acerca de una distribución de una o más variables.*

Definición 23 Prueba de Hipótesis Estadísticas. *Una prueba de hipótesis estadística H es una regla o procedimiento para decidir si se rechaza H.*

Definición 24 Región Crítica. *Sea Υ una prueba de una hipótesis estadística H. Una región crítica es el subconjunto para el cual la hipótesis H sería rechazada.*

Definición 25 Error Tipo I. *Rechazar H_0 dado que H_0 es verdadera, se define como error del tipo I.*

Definición 26 Error Tipo II. *Aceptar H_0 dado que H_0 es falsa, se define como error del tipo II.*

Definición 27 Probabilidad del Error Tipo I. *La probabilidad de rechazar H_0 dado que H_0 es cierta, se define como la probabilidad (o tamaño) del error tipo I y se denota por* α, $0 \leq \alpha \leq 1$.

Definición 28 Probabilidad del Error Tipo II. *La probabilidad de aceptar H_0 dado que H_0 es falsa, se define como la probabilidad (o tamaño) del error tipo II y se denota por* β, $0 \leq \beta \leq 1$.

Definición 29 Función Potencia. *Sea Υ una prueba de la hipótesis nula H_0. La función potencia de al prueba Υ, denotada por $\pi_\Upsilon(\theta)$ es la probabilidad de rechazar H_0.*

Capítulo 3

Modelo de Regresión Lineal Múltiple

Introducción.

El propósito de este capítulo es de proporcionar la metodología y los conceptos necesarios para extraer de grandes cantidades de datos las características principales de una relación.

De manera específica, se examinaran técnicas que permitan ajustar una ecuación de algún tipo al conjunto de datos dados, con el propósito de obtener una ecuación de predicción razonablemente precisa y que proporcione o justifique un modelo teórico.

Se supondrá que la existencia de un conjunto de n mediciones $Y_1, ..., Y_n$ de la variable

dependiente **Y**, las cuales se han observado bajo un conjunto de condiciones experimentales $x_{i1}, ...x_{ip}$ que representan los valores de p variables explicativas.

El interés recae en determinar una función matemática sencilla, que describa de manera razonable, el comportamiento de la variable dependiente, dados los valores de la variables explicativas.

3.1 El Modelo de Regresión Lineal Múltiple

El modelo de regresión lineal múltiple que relaciona una respuesta aleatoria **Y** con un conjunto de variables explicativas $X_1, ..., X_p$ es

$$\mathbf{Y} = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + e_i$$

en término de los datos observados es $x_1, ..., x_p$ es

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + e_i \quad para \quad i = 1, ..., n \qquad (3.1)$$

donde:

$\beta_1, ..., \beta_p$ *son los parámetros desconocidos*

e *es el error aleatorio*

i *es el índice asociado a observación*

n *es el tamaño de la muestra*

La ecuación (3.1) es una expresión abreviada del siguiente conjunto de ecuaciones simultáneas:

$$
\begin{aligned}
Y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + e_1 \\
Y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + e_2 \\
&\vdots \\
Y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + e_n
\end{aligned}
$$

En forma matricial

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}
$$

$\mathbf{Y}$ = vector columna de $n \times 1$ que contiene las observaciones de las variables dependientes,

$\mathbf{X}$ = matriz de $n \times p$ que contiene las observaciones de las p variables explicativas,

β = vector columna de $p \times 1$ que contiene los parámetros desconocidos $\beta_1, ..., \beta_p$,

$\mathbf{e}$ = vector columna de $n \times 1$ de los errores aleatorios $e_1, ..., e_n$,

Con lo cual el modelo. de regresión lineal con p variables, se puede escribir en forma matricial:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

Nuestro objetivo consistirá en estimar los parámetros desconocidos de este modelo y hacer inferencia acerca de ellos a partir de los datos que se tengan a mano.

Definición 30 *El modelo* $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ *será llamado modelo de regresión lineal múltiple de rango completo si sólo si el rango de* $\mathbf{X}$ *es igual a* p *donde* $p \leq n$.

3.2 El significado del Error e_i

El término e_i sustituye todas aquellas variables que han sido excluidas del modelo pero que afectan conjuntamente a $\mathbf{Y}$. La pregunta obvia es ¿por qué no se introducen explícitamente en el modelo todas estas variables? O, dicho de otro modo, ¿por qué no desarrollar un modelo de regresión múltiple con tantas variables como sea posible? Son varias las respuestas a esta pregunta, a saber:

1. La teoría si existe alguna, que determina el comportamiento de $\mathbf{Y}$, suele ser incompleta. Por ejemplo se puede estar seguro de que el ingreso semanal X afecta los gastos de consumo $\mathbf{Y}$, pero puede ocurrir que estemos inseguros o que desconozcamos otras variables que afectan a $\mathbf{Y}$. Por lo tanto, e_i puede utilizarse como un sustituto de todas las variables excluidas u omitidas del modelo.

2. Aun en el caso de que se conocieran algunas de las variables excluidas y que se procediera entonces a plantear un modelo de regresión múltiple, en lugar de regresión simple, es posible que no existan datos sobre dichas variables. Es muy común en los análisis empíricos, que la información que idealmente quisiéramos tener no se encuentre a nuestra disposición. Por ejemplo, podemos incluir la riqueza familiar como variable explicativa adicional al ingreso, para describir el consumo familiar. Desafortunadamente, a menudo ocurre que no existe información sobre esta variable, lo cual nos obliga a excluirla del modelo a pesar de su relevancia teórica para explicar los gastos de consumo.

3. Supongamos que además del ingreso X_1, los gastos de consumo también se ven afectados por el número de hijos de cada familia X_2, el sexo X_3, la religión X_4, la educación X_5 y la región geográfica X_6. Es muy posible que la influencia conjunta de todas o algunas de estas variables sea insignificante, y que del punto de vista práctico y por razones de costo no se justifique su introducción explícita en el modelo. Con optimismo, esperamos que el efecto combinado de todas estas variables se pueda tratar como una variable aleatoria e_i.

3.3 Supuestos del Modelo Clásico de Regresión Lineal Múltiple

El modelo general de regresión lineal en el cual se sustenta la mayor parte de la teoría econometrica, plantea los siguientes supuestos.

i) $E[\mathbf{e}] = 0$ este supuesto plantea que el valor promedio de e_i dado cualquier valor a $x_{1i}, \ldots, x_{ip}$ es igual a cero. Todo lo que se plantea es que aquellos factores que no están

incluidos explícitamente en el modelo, incorporados por lo tanto , en e_i, no afectan sistemáticamente el valor promedio de **Y**; es decir, los valores positivos de e_i se cancelan con los valores negativos de manera que su efecto promedio sobre **Y** es cero.

ii) $Cov[\mathbf{e}] = E[\mathbf{ee}'] = \sigma^2\mathbf{I}$ este supuesto plantea que no existe autocorrelación entre las e_i y además tienen igual varianza las e_i.

Este supuesto plantea que las perturbaciones asociadas a alguna observación no están influenciado por el término de perturbación asociado a cualquier otra observación; es decir las perturbaciones e_i y las e_j si $i \neq j$ no están corrrelacionadas.

Y también postula que la varianza de e_i para cada $x_{1i}, ..., x_{ip}$ es un número positivo constante igual a σ^2. Esto se conoce como homocedasticidad o igual dispersión o igual varianza. Esto significa que poblaciones de **Y** que corresponden a diferentes valores $x_{1i}, ..., x_{ip}$ tienen la misma varianza.

iii) la matriz **X** de orden $n \times p$ es no estocástica lo que implica que está formada por números fijos. Esto es obvio ya que el análisis de regresión los valores de **Y** dependen de los valores fijos de **X**.

iv) El rango de la matriz **X** es p y p es el número de columnas de la matriz. Lo que significa que las columnas de la matriz son linealmente independientes.

v) e esta distribuida como $N(\mathbf{0}, \sigma^2\mathbf{I})$ donde σ^2 es desconocido.

Como e_i representa la influencia combinada sobre la variable dependiente de un gran número de variables independientes que no se introducen explícitamente en el modelo de regresión lineal. Además se espera que la influencia de estas variables omitidas o no

tenidas en cuenta fuera pequeña y, en el mejor de los casos aleatoria: Y por el teorema del límite central , se demuestra que si existe un número grande de variables aleatorias independientes e identicamente distribuidas, entonces la distribución de su suma tenderá a seguir una distribución normal a medida que el número de estas variables aumenta indefinidamente. Es precisamente este teorema del límite central el que proporciona una justificación teórica para el supuesto de normalidad ϵ_i.

Una variante del teorema del límite central afirma que aunque el número de variables no sea muy grande o si estas variables no son estrictamente independientes, su suma puede seguir teniendo una distribución normal. En las siguientes secciones analizaremos los siguientes dos casos:

Caso A : $\mathbf{e}$ está distribuida como $\mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$ donde σ^2 es desconocido.

Caso B : $\mathbf{e}$ es una variable aleatoria tal que :

i) $E[\mathbf{e}] = 0$

ii) $Cov[\mathbf{e}] = E[\mathbf{e}\mathbf{e}^t] = \sigma^2\mathbf{I}$

3.4 Estimación Máximo Verosímil

Analizaremos primero el caso A, es decir bajo el supuesto de que $\mathbf{e}$ está distribuida como $\mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$ con varianza desconocida. Para este caso utilizaremos el método de estimación máximo verosímil, para estimar $\beta_1, ..., \beta_p$ y σ^2.

Puesto que $\mathbf{Y}$ esta distribuida como $\mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$ la función de verosimilitud es

$$L\left(\beta, \sigma^2\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{(\mathbf{Y}-\mathbf{X}\beta)^t(\mathbf{Y}-\mathbf{X}\beta)}{2\sigma^2}}$$

usando logaritmos obtenemos

$$\log L\left(\beta, \sigma^2\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)^t(\mathbf{Y} - \mathbf{X}\beta)$$

donde el espacio parametral de Ω es

$$\Omega = \left\{\left(\beta, \sigma^2\right) : \sigma^2 > 0, -\infty < \beta_i < \infty \ \forall\, i = 1, .., p\right\}$$

con lo cual, para encontrar los valores de β y σ^2 en Ω, tal que la función de verosimilitud tome su valor máximo, debemos derivar parcialmente con respecto a β y σ^2 e igualar las ecuaciones a cero, es decir,

$$\frac{\partial}{\partial\beta}\log L\left(\beta, \sigma^2\right) = \frac{2}{2\sigma^2}\left(\mathbf{X}^t\mathbf{Y} - \mathbf{X}^t\mathbf{X}\beta\right) = 0$$

$$\frac{\partial}{\partial\sigma^2}\log L\left(\beta, \sigma^2\right) = -\frac{n}{2\sigma^2} + \frac{(\mathbf{Y} - \mathbf{X}\beta)^t(\mathbf{Y} - \mathbf{X}\beta)}{2\sigma^4} = 0$$

Si $\hat{\beta}$ y $\hat{\sigma}^2$ son soluciones de las anteriores ecuaciones, tenemos que

$$\mathbf{X}^t\mathbf{X}\hat{\beta} = \mathbf{X}^t\mathbf{Y}$$

$$\hat{\sigma}^2 = \frac{\left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)^t \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)}{n}$$

Las ecuaciones

$$\mathbf{X}^t\mathbf{X}\hat{\beta} = \mathbf{X}^t\mathbf{Y}$$

son llamadas *ecuaciones normales.*

Como $\mathbf{X}$ es de rango completo entonces $\mathbf{X}^t\mathbf{X}$ es de rango p. Con lo cual existe la matriz inversa de $\mathbf{X}^t\mathbf{X}$.

Por lo tanto

$$\hat{\beta} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

y

$$\hat{\sigma}^2 = \frac{\left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)^t \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)}{n}$$

son las estimaciones máximo verosímil de β y σ^2 respectivamente.

Teorema 28 *Sea* $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ *el modelo de regresión lineal múltiple de rango completo. Si* $\mathbf{e}$ *está distribuido como* $\mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$ *entonces los estimadores*

$$\hat{\beta} = \mathbf{S}^{-1}\mathbf{X}^t\mathbf{Y} \quad \textit{donde } \mathbf{S} = \mathbf{X}^t\mathbf{X}$$

y

$$\hat{\sigma}^2 = \frac{\mathbf{Y}^t(\mathbf{I} - \mathbf{X}S^{-1}\mathbf{X}^t)\mathbf{Y}}{n - p}$$

cumplen las siguientes propiedades

i) Consistentes.

ii) Eficientes.

iii) Insesgados.

iv) Suficientes.

v) Completas.

vi) β *esta distribuido como una* $\mathbf{N}(\beta, \sigma^2\mathbf{S}^{-1})$.

vii) Insesgados de varianza mínima uniforme.

viii) $\frac{\hat{\sigma}^2(n-p)}{\sigma^2}$ *está distribuido como una distribución* χ^2_{n-p}.

ix) β *y* σ^2 *son independientes.*

Demostración.

i) y *ii)* Como $\hat{\beta}$ y $\hat{\sigma}^2$ son estimadores máximo verosímiles y por el teorema (23) entonces $\hat{\beta}$ y $\hat{\sigma}^2$ son estimadores consistentes y eficientes.

iii) ***Insesgados***

$$\begin{aligned} E\left[\hat{\beta}\right] &= E\left[\mathbf{S}^{-1}\mathbf{X}^t\mathbf{Y}\right] \\ &= \mathbf{S}^{-1}\mathbf{X}^t E\left[\mathbf{Y}\right] \\ &= \mathbf{S}^{-1}\mathbf{X}^t E\left[\mathbf{X}\beta + \mathbf{e}\right] \\ &= \mathbf{S}^{-1}\mathbf{X}^t\mathbf{X}\beta + \mathbf{S}^{-1}\mathbf{X}^t E\left[\mathbf{e}\right] \text{ como por hipótesis } E\left[\mathbf{e}\right] = 0 \text{ entonces} \\ &= \beta \end{aligned}$$

y

$$\begin{aligned} E\left[\hat{\sigma}^2\right] &= E\left[\frac{\mathbf{Y}^t(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t)\mathbf{Y}}{n-p}\right] \\ &= \frac{1}{n-p}E\left[(\mathbf{X}\beta+\mathbf{e})^t\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right)(\mathbf{X}\beta+\mathbf{e})\right] \text{ por el teorema (6)} \\ &= \frac{1}{n-p}E\left[(\beta^t\mathbf{X}^t+\mathbf{e}^t)\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right)(\mathbf{X}\beta+\mathbf{e})\right] \\ &= \frac{1}{n-p}E\left[\mathbf{e}^t\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right)\mathbf{e}\right] \end{aligned}$$

Y como

$$\mathbf{X}^t\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right) = 0 \; y \; \left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right)\mathbf{X} = 0$$

Por lo tanto por el teorema (14)

$$E\left[\hat{\sigma}^2\right] = \frac{\sigma^2}{n-p} tr\left(\mathbf{I} - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right)$$

Y como

$$tr\left(\mathbf{I} - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right) = tr\left(\mathbf{I}\right) - tr\left(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right)$$

donde $\mathbf{I}$ es la matriz identidad de $n \times n$, con lo cual

$tr\left(\mathbf{I}\right) = n$ y por el teorema (11)

$$tr\left(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right) = tr\left(\mathbf{X}^t\mathbf{X}\mathbf{S}^{-1}\right) = tr\left(\mathbf{I}\right) = p$$

Puesto que $\mathbf{X}^t\mathbf{X}\mathbf{S}^{-1} = \mathbf{I}$ es la matriz identidad de $p \times p$.

Por lo tanto,

$$E\left[\hat{\sigma}^2\right] = \frac{\sigma^2}{n-p}(n-p) = \sigma^2$$

iv) ***Suficiencia.***

La función de densidad conjunta de $\mathbf{Y}$ es

$$f\left(\beta, \sigma^2\right) = \left(\frac{1}{(2\pi\sigma^2)^{n/2}}\right) \exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)^t(\mathbf{Y} - \mathbf{X}\beta)\right]$$

Y como

$$
\begin{aligned}
(\mathbf{Y}-\mathbf{X}\beta)^{t}(\mathbf{Y}-\mathbf{X}\beta) &= \left[\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)-\mathbf{X}\left(\beta-\hat{\beta}\right)\right]^{t}\left[\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)-\mathbf{X}\left(\beta-\hat{\beta}\right)\right] \\
&= \left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)^{t}\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)-\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)^{t}\mathbf{X}\left(\beta-\hat{\beta}\right) \\
&\quad -\left(\beta-\hat{\beta}\right)^{t}\mathbf{X}^{t}\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)+\left(\beta-\hat{\beta}\right)^{t}\mathbf{X}^{t}\mathbf{X}\left(\beta-\hat{\beta}\right) \\
&= \left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)^{t}\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)+\left(\beta-\hat{\beta}\right)^{t}\mathbf{X}^{t}\mathbf{X}\left(\beta-\hat{\beta}\right) \\
&= (n-p)\hat{\sigma}^{2}+\left(\beta-\hat{\beta}\right)^{t}\mathbf{X}^{t}\mathbf{X}\left(\beta-\hat{\beta}\right)
\end{aligned}
$$

Puesto que

$$
\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)^{t}\mathbf{X}=\mathbf{Y}^{t}\mathbf{X}-\hat{\beta}^{t}\mathbf{X}^{t}\mathbf{X}=\mathbf{Y}^{t}\mathbf{X}-\mathbf{Y}^{t}\mathbf{X}\left(\mathbf{X}^{t}\mathbf{X}\right)^{-1}\mathbf{X}^{t}\mathbf{X}=0.
$$

y

$$
\mathbf{X}^{t}\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)=\mathbf{X}^{t}\mathbf{Y}-\mathbf{X}^{t}\mathbf{X}\hat{\beta}=\mathbf{X}^{t}\mathbf{Y}-\mathbf{X}^{t}\mathbf{X}\left(\mathbf{X}^{t}\mathbf{X}\right)^{-1}\mathbf{X}^{t}\mathbf{Y}=0.
$$

Con lo cual

$$
\begin{aligned}
f(\beta,\sigma^{2}) &= \left(\tfrac{1}{(2\pi\sigma^{2})^{n/2}}\right)\exp\left[-\tfrac{1}{2\sigma^{2}}\left[(n-p)\hat{\sigma}^{2}+\left(\beta-\hat{\beta}\right)^{t}\mathbf{X}^{t}\mathbf{X}\left(\beta-\hat{\beta}\right)\right]\right] \\
&= g\left(\hat{\beta},\hat{\sigma}^{2},\beta,\sigma^{2}\right)h(\mathbf{Y})
\end{aligned}
$$

donde $h(\mathbf{Y})=1$.

Por lo tanto, por el teorema (24)

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

$$\hat{\sigma}^{2}=\frac{\mathbf{Y}^{t}\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)\mathbf{Y}}{n-p}$$

son estadísticas suficientes.

v) Completas.

Como $\mathbf{Y}$ se distribuye como una $\mathbf{N}(\mathbf{X}\beta,\sigma^{2}\mathbf{I})$ y está pertenece a la familia exponencial entonces por el teorema (25)

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

$$\hat{\sigma}^{2}=\frac{\mathbf{Y}^{t}\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)\mathbf{Y}}{n-p}$$

son estadísticas suficientes y completas.

vi) β esta distribuido como una $\mathbf{N}(\beta,\sigma^{2}\mathbf{S}^{-1})$

Como $\hat{\beta}=\mathbf{S}^{-1}\mathbf{X}^{t}\mathbf{Y}$ entonces por el teorema (15) tenemos que $\hat{\beta}$ está distribuido como una normal y su media es $E\left[\hat{\beta}\right]=\beta$ esto fue calculado en el inciso *iii)* y la matriz de covarianza de $\hat{\beta}$ es

$$
\begin{aligned}
Cov\left[\hat{\beta}\right] &= E\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{t}\right] \\
&= E\left[\left(\mathbf{S}^{-1}\mathbf{X}^{t}\mathbf{Y}-\beta\right)\left(\mathbf{S}^{-1}\mathbf{X}^{t}\mathbf{Y}-\beta\right)^{t}\right]
\end{aligned}
$$

Si sustituimos $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ obtenemos que

$$
\begin{aligned}
Cov\left[\hat{\beta}\right] &= E\left\{\left[\mathbf{S}^{-1}\mathbf{X}^{t}\left(\mathbf{X}\beta+\mathbf{e}\right)-\beta\right]\left[\mathbf{S}^{-1}\mathbf{X}^{t}\left(\mathbf{X}\beta+\mathbf{e}\right)-\beta\right]^{t}\right\} \\
&= E\left[\left(\mathbf{S}^{-1}\mathbf{X}^{t}\mathbf{e}\right)\left(\mathbf{S}^{-1}\mathbf{X}^{t}\mathbf{e}\right)^{t}\right] \\
&= E\left[\mathbf{S}^{-1}\mathbf{X}^{t}\mathbf{e}\mathbf{e}^{t}\mathbf{X}\mathbf{S}^{-1}\right] \\
&= \mathbf{S}^{-1}\mathbf{X}^{t}E\left[\mathbf{e}\mathbf{e}^{t}\right]\mathbf{X}\mathbf{S}^{-1} \\
&= \sigma^{2}\mathbf{S}^{-1}\mathbf{X}^{t}\mathbf{X}\mathbf{S}^{-1} \\
&= \sigma^{2}\mathbf{S}^{-1}.
\end{aligned}
$$

Con lo cual

β esta distribuido como una $\mathbf{N}\left(\beta, \sigma^{2}\mathbf{S}^{-1}\right)$.

viii) $\frac{\hat{\sigma}^{2}(n-p)}{\sigma^{2}}$ está distribuido como una distribución $\chi^{2}_{(n-p)}$.

Puesto que $(n-p)\,\hat{\sigma}^{2} = \mathbf{Y}^{t}\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)\mathbf{Y}$ y $\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}$ es una matriz idempotente de rango $n-p$ por lo tanto por el teorema (16)

$\frac{\hat{\sigma}^{2}(n-p)}{\sigma^{2}}$ se distribuye como $\chi^{2}_{(n-p)}$ donde $\lambda = \frac{1}{2\sigma^{2}}\beta^{t}\mathbf{X}^{t}\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)\mathbf{X}\beta = 0$.

Con lo cual

$\frac{\hat{\sigma}^{2}(n-p)}{\sigma^{2}}$ se distribuye como una $\chi^{2}_{(n-p)}$

ix) β y σ^{2} son independientes.

Como

$$\hat{\beta} = \left(\mathbf{S}^{-1}\mathbf{X}^{t}\right)\mathbf{Y}$$

$$\hat{\sigma}^2 = \frac{1}{n-p}\left[\mathbf{Y}^{t}\left(\mathbf{I} - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)\mathbf{Y}\right]$$

Y por el teorema (17) sabemos que para que $\hat{\beta}$ y $\hat{\sigma}^2$ sean independientes el producto de la matriz lineal y la matriz de la forma cuadratica debe ser igual a cero.

Es decir, $\hat{\beta}$ y $\hat{\sigma}^2$ son independientes si y sólo si $\mathbf{S}^{-1}\mathbf{X}^{t}\left(\mathbf{I} - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right) = 0$ y esto es claro, ya que $\mathbf{X}^{t}\left(\mathbf{I} - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right) = 0$.

3.5 Estimación por Mínimos Cuadrados

Ahora analizaremos el caso B, es decir bajo los supuestos de que e es una variable aleatoria tal que:

i) $E[\mathbf{e}] = 0$

ii) $Cov[\mathbf{e}] = E[\mathbf{e}\mathbf{e}^{t}] = \sigma^2\mathbf{I}$

donde σ^2 es desconocido. Para este caso utilizaremos el método de mínimos cuadrados ordinarios, que consiste en hacer que la suma de los residuos al cuadrado sea tan pequeña como sea posible, es decir, minimizar $\sum e_i^2$.

Por lo tanto debemos de minimizar

$$\sum e_i^2 = \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{e}^t\mathbf{e}$$

Como el modelo es de la forma $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ tenemos que $\mathbf{e} = \mathbf{Y} - \mathbf{X}\beta$ con lo cual

$$\sum e_i^2 = \mathbf{e}^t\mathbf{e} = (\mathbf{Y} - \mathbf{X}\beta)^t (\mathbf{Y} - \mathbf{X}\beta)$$

Derivando con respecto a β obtenemos

$$\frac{\partial(\mathbf{e}^t\mathbf{e})}{\partial\beta} = 2\mathbf{X}^t\mathbf{Y} - 2\mathbf{X}^t\mathbf{X}\beta = 0$$

$$\mathbf{X}^t\mathbf{X}\beta = \mathbf{X}^t\mathbf{Y}$$

Como $\mathbf{X}$ es de rango completo entonces $\mathbf{X}^t\mathbf{X}$ es de rango p. Con lo cual existe la matriz inversa de $\mathbf{X}^t\mathbf{X}$.

Por lo tanto

$$\hat{\beta} = \mathbf{S}^{-1}\mathbf{X}^t\mathbf{Y} \text{ } \textit{donde} \text{ } \mathbf{S} = \mathbf{X}^t\mathbf{X}$$

Con lo cual hemos obtenido la solución

$$\hat{\beta} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

para la estimación de los parámetros desconocidos.

3.6 Teorema de Gauss Markoff

Teorema 29 *Gauss-Markoff*

El modelo general de regresión lineal múltiple de rango completo, $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$, que cumple las siguientes condiciones:

i) $E[\mathbf{e}] = 0$

ii) $Cov[\mathbf{e}] = E[\mathbf{e}\mathbf{e}^t] = \sigma^2\mathbf{I}$

entonces el mejor lineal insesgado de β es el dado por el mínimos cuadrados; es decir, $\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$ es el mejor estimador lineal insesgado de β.

Demostración.

Sea $\mathbf{A}$ cualquier matriz de $p \times n$ de valores conocidos y sea β^* cualquier otro estimador lineal insesgado de β tal que $\beta^* = \mathbf{AY}$. como debemos especificar los elementos de $\mathbf{A}$, tal que β^* sea el mejor estimador insesgado de β.

Sea $\mathbf{A} = \mathbf{S}^{-1}\mathbf{X}^t + \mathbf{B}$ y como $\mathbf{S}^{-1}\mathbf{X}^t$ es conocido debemos encontrar B para poder especi-

ficar **A**.

$$
\begin{aligned}
E[\beta^*] &= E(\mathbf{AY}) = E[(\mathbf{S}^{-1}\mathbf{X}^t+\mathbf{B})\mathbf{Y}] \\
&= (\mathbf{S}^{-1}\mathbf{X}^t+\mathbf{B})E[\mathbf{Y}] = (\mathbf{S}^{-1}\mathbf{X}^t+\mathbf{B})E[\mathbf{X}\beta+\mathbf{e}] \\
&= (\mathbf{S}^{-1}\mathbf{X}^t+\mathbf{B})[E[\mathbf{X}\beta]+E[\mathbf{e}]] \\
&= (\mathbf{S}^{-1}\mathbf{X}^t+\mathbf{B})(\mathbf{X}\beta+0) \qquad \text{por hipótesis } E[\mathbf{e}]=0 \\
&= \beta+\mathbf{BX}\beta
\end{aligned}
$$

Como β^* es insesgado entonces $\mathbf{BX}\beta = 0$ para toda β, por lo tanto $\mathbf{BX} = 0$.

Para demostrar que es el mejor estimador lineal insesgado, debemos encontrar la matriz **B** tal que $Var(\beta^*)$ sea mínimo para $i = 1, \ldots, p$ sujeto a la restricción $\mathbf{BX} = 0$.

$$
\begin{aligned}
Cov(\beta^*) &= E\left[(\beta^*-\beta)(\beta^*-\beta)^t\right] \\
&= E\left\{[(\mathbf{S}^{-1}\mathbf{X}^t+\mathbf{B})\mathbf{Y}-\beta][(\mathbf{S}^{-1}\mathbf{X}^t+\mathbf{B})\mathbf{Y}-\beta]^t\right\} \\
&= E\left\{[(\mathbf{S}^{-1}\mathbf{X}^t+\mathbf{B})(\mathbf{X}\beta+\mathbf{e})-\beta][(\mathbf{S}^{-1}\mathbf{X}^t+\mathbf{B})(\mathbf{X}\beta+\mathbf{e})-\beta]^t\right\} \\
&= E\left\{[\beta+\mathbf{BX}\beta+\mathbf{S}^{-1}\mathbf{X}^t\mathbf{e}+\mathbf{Be}-\beta][\beta+\mathbf{BX}\beta+\mathbf{S}^{-1}\mathbf{X}^t\mathbf{e}+\mathbf{Be}-\beta]^t\right\} \\
&= E\left\{[\mathbf{BX}\beta+\mathbf{S}^{-1}\mathbf{X}^t\mathbf{e}+\mathbf{Be}][\mathbf{BX}\beta+\mathbf{S}^{-1}\mathbf{X}^t\mathbf{e}+\mathbf{Be}]^t\right\} \\
&= E\left\{[\mathbf{S}^{-1}\mathbf{X}^t\mathbf{e}+\mathbf{Be}][\mathbf{S}^{-1}\mathbf{X}^t\mathbf{e}+\mathbf{Be}]^t\right\} \\
&= E\left[\mathbf{S}^{-1}\mathbf{X}^t\mathbf{e}\mathbf{e}^t\mathbf{X}\mathbf{S}^{-1}+\mathbf{B}\mathbf{e}\mathbf{e}^t\mathbf{B}^t+\mathbf{S}^{-1}\mathbf{X}^t\mathbf{e}\mathbf{e}^t\mathbf{B}^t+\mathbf{B}\mathbf{e}\mathbf{e}^t\mathbf{X}\mathbf{S}^{-1}\right] \\
&= \mathbf{S}^{-1}\mathbf{X}^t E[\mathbf{e}\mathbf{e}^t]\mathbf{X}\mathbf{S}^{-1}+\mathbf{B}E[\mathbf{e}\mathbf{e}^t]\mathbf{B}^t+\mathbf{S}^{-1}\mathbf{X}^t E[\mathbf{e}\mathbf{e}^t]\mathbf{B}^t+\mathbf{B}E[\mathbf{e}\mathbf{e}^t]\mathbf{X}\mathbf{S}^{-1} \\
&= \sigma^2\mathbf{S}^{-1}+\sigma^2\mathbf{B}\mathbf{B}^t+\sigma^2\mathbf{S}^{-1}\mathbf{X}^t\mathbf{B}^t+\sigma^2\mathbf{B}\mathbf{X}\mathbf{S}^{-1} \\
&= \sigma^2\mathbf{S}^{-1}+\sigma^2\mathbf{B}\mathbf{B}^t
\end{aligned}
$$

Sea

$$\mathbf{B}\mathbf{B}^t = \mathbf{G} = g_{ij}$$

Como $Cov(\beta^*) = \sigma^2\mathbf{S}^{-1} + \sigma^2\mathbf{BB}'$ la diagonal de los elementos de la $Cov(\beta^*)$ deben ser iguales a la varianza de β_i^*.

Para minimizar cada $Var(\beta_i^*)$ debemos, por lo tanto minimizar los elementos de la diagonal $Cov(\beta^*)$.

Como $\sigma^2\mathbf{S}^{-1}$ son constantes, debemos encontrar una matriz g tal que los elementos de la diagonal sean mínimos, pero como $\mathbf{G} = \mathbf{BB}'$ es semidefinida positiva tenemos que $g_{ij} \geq 0$ con lo cual los elementos de la diagonal serán mínimos cuando $g_{ij} = 0$ para $i = 1..., p$ pero como $B = b_{ij}$ entonces $g_{ij} = 0 = \sum b_{ij}^2$ y por lo tanto, $b_{ij} = 0$ para toda i y j, con lo cual $\mathbf{B} = 0$.

Con lo cual

$$\mathbf{A} = \mathbf{S}^{-1}\mathbf{X}'$$

y

$$\beta^* = \beta.$$

3.7 Estimación por Intervalo

3.7.1 Intervalo de Confianza de σ^2

Puesto que $\frac{\hat{\sigma}^2(n-p)}{\sigma^2}$ se distribuye como una $\chi^2_{(n-p)}$, un intervalo de confianza para σ^2 es el siguiente:

Si a_0 y a_1 son dos constantes tales que

$$P\left[a_0 \leq \frac{\hat{\sigma}^2(n-p)}{\sigma^2} \leq a_1\right] = 1 - \alpha$$

Por lo tanto,

$$P\left[\frac{\hat{\sigma}^2(n-p)}{a_1} \leq \sigma^2 \leq \frac{\hat{\sigma}^2(n-p)}{a_0}\right] = 1 - \alpha$$

El ancho del intervalo de confianza es

$$\frac{\hat{\sigma}^2(n-p)}{a_0} - \frac{\hat{\sigma}^2(n-p)}{a_1} = \hat{\sigma}^2(n-p)\left(\frac{1}{a_0} - \frac{1}{a_1}\right).$$

3.7.2 Intervalo de Confianza de β_i

Puesto que $\hat{\beta}_i$ se distribuye como $N(\beta_i, c_{ii}\sigma^2)$, donde c_{ij} es el $ij-ésimo$ elemento de $S^{-1} = C$.

Por lo tanto $\frac{(\hat{\beta}_i-\beta_i)}{\sigma\sqrt{c_{ii}}}$ se distribuye como $N(0,1)$ y es independiente de $\frac{\hat{\sigma}^2(n-p)}{\sigma^2}$, la cual se distribuye como $\chi^2_{(n-p)}$.

Por el teorema (22), tenemos que $u = \frac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{c_{ii}}}\sqrt{\frac{\sigma^2}{\hat{\sigma}^2}} = \frac{\hat{\beta}_i-\beta_i}{\sqrt{\hat{\sigma}^2 c_{ii}}}$ se distribuye como $t_{(n-p)}$.

Con lo cual,

$$P\left[-t_{\alpha/2} \leq \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 c_{ii}}} \leq t_{\alpha/2}\right] = 1-\alpha$$

Por lo tanto.

$$P\left[\hat{\beta}_i - t_{\alpha/2}\sqrt{\hat{\sigma}^2 c_{ii}} \leq \beta_i \leq \hat{\beta}_i + t_{\alpha/2}\sqrt{\hat{\sigma}^2 c_{ii}}\right] = 1-\alpha$$

Con lo cual, el ancho del intervalo de confianza es:

$$\hat{\beta}_i + t_{\alpha/2}\sqrt{\hat{\sigma}^2 c_{ii}} - \left(\hat{\beta}_i - t_{\alpha/2}\sqrt{\hat{\sigma}^2 c_{ii}}\right) = 2t_{\alpha/2}\sqrt{\hat{\sigma}^2 c_{ii}}$$

3.7.3 Intervalo de Confianza para una Función Lineal de β_i

Sea $\mathbf{r}$ un vector conocido de $p \times 1$ constantes entonces el intervalo de confianza para $\mathbf{r}^t\beta$ será el siguiente:

Puesto que $\mathbf{r}'\hat{\beta}$ se distribuye como $N(\mathbf{r}'\beta, \sigma^2\mathbf{r}'\mathbf{S}^{-1}\mathbf{r})$ entonces por el teorema (20)

$$\frac{\mathbf{r}'\hat{\beta}-\mathbf{r}'\beta}{\sigma\sqrt{\mathbf{r}'\mathbf{S}^{-1}\mathbf{r}}}$$

se distribuye como $N(0,1)$.

Con lo cual, la variable

$$u = \frac{\mathbf{r}'\hat{\beta}-\mathbf{r}'\beta}{\sigma\sqrt{\mathbf{r}'\mathbf{S}^{-1}\mathbf{r}}}\sqrt{\frac{\sigma^2}{\hat{\sigma}^2}} = \frac{\mathbf{r}'\hat{\beta}-\mathbf{r}'\beta}{\sqrt{\hat{\sigma}^2\mathbf{r}'\mathbf{S}^{-1}\mathbf{r}}}$$

se distribuye como $t_{(n-p)}$

Con lo cual,

$$P\left[-t_{\alpha/2} \leq \frac{\mathbf{r}'\hat{\beta}-\mathbf{r}'\beta}{\sqrt{\hat{\sigma}^2\mathbf{r}'\mathbf{S}^{-1}\mathbf{r}}} \leq t_{\alpha/2}\right] = 1-\alpha$$

Por lo tanto,

$$P\left[\mathbf{r}'\hat{\beta} - t_{\alpha/2}\sqrt{\hat{\sigma}^2\mathbf{r}'\mathbf{S}^{-1}\mathbf{r}} \leq \mathbf{r}'\beta \leq \mathbf{r}'\hat{\beta} + t_{\alpha/2}\sqrt{\hat{\sigma}^2\mathbf{r}'\mathbf{S}^{-1}\mathbf{r}}\right] = 1-\alpha$$

Con lo cual, la longitud el intervalo es

$$\mathbf{r}'\hat{\beta} + t_{\alpha/2}\sqrt{\hat{\sigma}^2\mathbf{r}'\mathbf{S}^{-1}\mathbf{r}} - \left(\mathbf{r}'\hat{\beta} - t_{\alpha/2}\sqrt{\hat{\sigma}^2\mathbf{r}'\mathbf{S}^{-1}\mathbf{r}}\right) = 2t_{\alpha/2}\sqrt{\hat{\sigma}^2\mathbf{r}'\mathbf{S}^{-1}\mathbf{r}}$$

3.8 Prueba de Hipótesis

La prueba de hipótesis $H_0 : \beta = \beta^*$ donde β^* es un vector conocido, es equivalente a la prueba de hipótesis de cada coeficiente $\beta_i = \beta_i^*$

En realidad para evaluar la función potencia de la distribución, también será importante conocer cuando la hipótesis alternativa $H_1 : \beta \neq \beta^*$ es verdadera.

Para esto usaremos la razón de verosimilitud L.

La función de verosimilitud es

$$f\left(\beta, \sigma^2\right) = \left(\frac{1}{(2\pi\sigma^2)^{n/2}}\right) \exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)^t(\mathbf{Y} - \mathbf{X}\beta)\right]$$

El criterio de prueba es $L = \frac{L(\hat{\omega})}{L(\hat{\Omega})}$, que a continuación explicaremos lo que significa $L(\hat{\omega})$ y $L\left(\hat{\Omega}\right)$.

La función de verosimilitud es una función de $p+1$ parámetros de Ω es de dimensión $p+1$, sujeto a las siguientes restricciones $0 < \sigma^2 < \infty$ y $-\infty < \beta_i < \infty$ para $i = 1, ..., p$.

El espacio $\hat{\Omega}$ serán los valores de los parámetros $\beta_1, ..., \beta_p$ y σ^2 en Ω tal que la función de verosimilitud sea máximo y $L\left(\hat{\Omega}\right)$ será el valor máximo.

El espacio ω, estará sujeto a las siguientes restricciones $0 < \sigma^2 < \infty$ y $\beta_1 = \beta_1^*, ..., \beta_p = \beta_p^*$, es decir ω es de una dimensión

Para encontrar $L(\hat{\omega})$ y $L\left(\hat{\Omega}\right)$ usaremos logaritmos en la función de verosimilitud.

Para encontrar $L(\hat{\omega})$ procedamos como sigue:

$$\log f\left(\beta,\sigma^2\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\beta)^t(\mathbf{Y}-\mathbf{X}\beta)$$

Como β_i^* son conocidos, la función solamente tienen un parámetro, que es σ^2 y por lo tanto el valor que maximiza la función de verosimilitud es

$$\frac{d}{d\sigma^2}\left(\log f\left(\beta,\sigma^2\right)\right) = \frac{1}{2\hat{\sigma}^4}(\mathbf{Y}-\mathbf{X}\beta^*)^t(\mathbf{Y}-\mathbf{X}\beta^*) - \frac{n}{2\hat{\sigma}^2} = 0$$

La solución es

$$\hat{\sigma}^2 = \frac{(\mathbf{Y}-\mathbf{X}\beta^*)^t(\mathbf{Y}-\mathbf{X}\beta^*)}{n}$$

entonces

$$L(\hat{\omega}) = \frac{n^{n/2}e^{-n/2}}{(2\pi)^{n/2}\left[(\mathbf{Y}-\mathbf{X}\beta^*)^t(\mathbf{Y}-\mathbf{X}\beta^*)\right]^{n/2}}$$

De manera similar obtengamos $L\left(\hat{\Omega}\right)$, para obtener el valor máximo $L\left(\hat{\Omega}\right)$ de la función de verosimilitud resolveremos las siguientes ecuaciones

$$\frac{d}{d\beta}\left(\log f\left(\beta,\sigma^2\right)\right) = \frac{\mathbf{X}^t\mathbf{Y}}{\hat{\sigma}^2} - \frac{\mathbf{X}^t\mathbf{X}\beta}{\hat{\sigma}^2} = 0$$

$$\frac{d}{d\sigma^2}\left(\log f\left(\beta,\sigma^2\right)\right) = \frac{\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)^t\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)}{2\hat{\sigma}^4} - \frac{n}{2\hat{\sigma}^2} = 0$$

Con lo cual,

$$L\left(\hat{\Omega}\right) = \frac{n^{n/2}e^{-n/2}}{(2\pi)^{n/2}\left[\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)^{t}\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)\right]^{n/2}}$$

Por lo tanto la razón de verosimilitud

$$L = \frac{L(\hat{\omega})}{L\left(\hat{\Omega}\right)} = \left[\frac{\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)^{t}\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)}{(\mathbf{Y}-\mathbf{X}\beta^{*})^{t}(\mathbf{Y}-\mathbf{X}\beta^{*})}\right]^{n/2}$$

Si $g(L,\beta^{*})$ es la distribución de L bajo $H_0 : \beta = \beta^{*}$, la región crítica es $0 \leq L \leq A$ donde A es tal que

$$\int g(L,\beta^{*})dL = \alpha$$

donde α es la probabilidad del error tipo I.

Para determinar A, debemos encontrar $g(L,\beta^{*})$ y la distribución de L bajo $H_0 : \beta = \beta^{*}$ entonces L sería determinada de la elección de los datos y H_0 sería rechazada si $L \leq A$.

Para determinar L, debemos estudiar las cantidades que están envueltas en L, es decir,

$$(\mathbf{Y}-\mathbf{X}\beta^{*})^{t}(\mathbf{Y}-\mathbf{X}\beta^{*})\, y\, \left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)^{t}\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)$$

Con lo cual,

$$
\begin{aligned}
(\mathbf{Y}-\mathbf{X}\beta^{*})'(\mathbf{Y}-\mathbf{X}\beta^{*}) &= \left[\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)-\mathbf{X}\left(\beta^{*}-\hat{\beta}\right)\right]'\left[\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)-\mathbf{X}\left(\beta^{*}-\hat{\beta}\right)\right] \\
&= \left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)'\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)+\left(\beta^{*}-\hat{\beta}\right)'\mathbf{X}'\mathbf{X}\left(\beta^{*}-\hat{\beta}\right) \\
&= \left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)'\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)+\left(\hat{\beta}-\beta^{*}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\beta}-\beta^{*}\right)
\end{aligned}
$$

Puesto que

$$
\mathbf{X}'\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)=\mathbf{X}'\mathbf{Y}-\mathbf{X}'\mathbf{X}\hat{\beta}=\mathbf{X}'\mathbf{Y}-\mathbf{X}'\mathbf{X}\left(\mathbf{S}^{-1}\mathbf{X}'\mathbf{Y}\right)=\mathbf{X}'\mathbf{Y}-\mathbf{X}'\mathbf{Y}=0
$$

De igual manera

$$
\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)\mathbf{X}=0
$$

Ahora si sustituimos $\hat{\beta}=\mathbf{S}^{-1}\mathbf{X}'\mathbf{Y}$ **en**

$$
(\mathbf{Y}-\mathbf{X}\beta^{*})'(\mathbf{Y}-\mathbf{X}\beta^{*})=\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)'\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)+\left(\hat{\beta}-\beta^{*}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\beta}-\beta^{*}\right)
$$

Obtenemos lo siguiente

$$
\begin{aligned}
(\mathbf{Y}-\mathbf{X}\beta^{*})^{t}(\mathbf{Y}-\mathbf{X}\beta^{*}) &= \left(\mathbf{Y}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\mathbf{Y}\right)^{t}\left(\mathbf{Y}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\mathbf{Y}\right) \\
&\quad +(\mathbf{S}^{-1}\mathbf{X}^{t}\mathbf{Y}-\beta^{*})^{t}\mathbf{X}^{t}\mathbf{X}(\mathbf{S}^{-1}\mathbf{X}^{t}\mathbf{Y}-\beta^{*}) \\
&= \mathbf{Y}^{t}\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)\mathbf{Y}+\mathbf{Y}^{t}\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\mathbf{Y}-\mathbf{Y}^{t}\mathbf{X}\beta^{*} \\
&\quad -(\beta^{*})^{t}\mathbf{X}^{t}\mathbf{Y}+(\beta^{*})^{t}\mathbf{X}^{t}\mathbf{X}\beta^{*} \\
&= \mathbf{Y}^{t}\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)\mathbf{Y}+\mathbf{Y}^{t}\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\mathbf{Y}-\mathbf{Y}^{t}\mathbf{X}\beta^{*} \\
&\quad -(\mathbf{X}\beta^{*})^{t}\mathbf{Y}+(\mathbf{X}\beta^{*})^{t}\mathbf{X}\beta^{*}
\end{aligned}
$$

Como

$$(\mathbf{X}\beta^{*})^{t}\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)(-\mathbf{X}\beta^{*})=0$$

$$\mathbf{Y}^{t}\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)(-\mathbf{X}\beta^{*})=0$$

$$(-\mathbf{X}\beta^{*})^{t}\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)\mathbf{Y}=0$$

Por lo tanto

$$\mathbf{Y}^{t}\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)\mathbf{Y}=(\mathbf{Y}-\mathbf{X}\beta^{*})^{t}\left(\mathbf{I}-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)(\mathbf{Y}-\mathbf{X}\beta^{*})$$

También como

$$\mathbf{Y}^{t}\left(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)(\mathbf{X}\beta^{*})=\mathbf{Y}^{t}\mathbf{X}\beta^{*}$$

$$(\mathbf{X}\beta^{*})^{t}\left(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)\mathbf{Y}=(\mathbf{X}\beta^{*})^{t}\mathbf{Y}$$

$$(\mathbf{X}\beta^{*})^{t}\left(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{t}\right)(\mathbf{X}\beta^{*})=(\mathbf{X}\beta^{*})^{t}\mathbf{X}\beta^{*}$$

Tenemos que

$$\mathbf{Y}^t\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\mathbf{Y} - \mathbf{Y}^t\mathbf{X}\beta^* - (\mathbf{X}\beta^*)^t\mathbf{Y} + (\mathbf{X}\beta^*)^t\mathbf{X}\beta^* =$$

$$= \mathbf{Y}^t\left(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right)\mathbf{Y} - \mathbf{Y}^t\left(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right)(\mathbf{X}\beta^*) - (\mathbf{X}\beta^*)^t\left(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right)\mathbf{Y} + (\mathbf{X}\beta^*)^t\left(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right)(\mathbf{X}\beta^*)$$

$$= (\mathbf{Y} - \mathbf{X}\beta^*)^t\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t(\mathbf{Y} - \mathbf{X}\beta^*)$$

Ahora si $\mathbf{Z} = (\mathbf{Y} - \mathbf{X}\beta^*)$ se obtiene que

$$\mathbf{Z}^t\mathbf{Z} = (\mathbf{Y} - \mathbf{X}\beta^*)^t(\mathbf{Y} - \mathbf{X}\beta^*) = \mathbf{Z}^t\mathbf{A}_1\mathbf{Z} + \mathbf{Z}^t\mathbf{A}_2\mathbf{Z}$$

donde $\mathbf{A}_1 = \mathbf{I} - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t$ y $\mathbf{A}_2 = \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t$ son matrices idempotentes.

Como $\mathbf{Y}$ está distribuido como $N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ entonces Z está distribuido como $N(\mathbf{X}\beta - \mathbf{X}\beta^*, \sigma^2\mathbf{I})$.

Como A_1 y A_2 son matrices idempotentes aplicando el teorema (18) tenemos que

i) $\frac{\mathbf{Z}^t\mathbf{A}_1\mathbf{Z}}{\sigma^2}$ es distribuido como $\chi^2_{(n-p)}$.

ii) $\frac{\mathbf{Z}^t\mathbf{A}_2\mathbf{Z}}{\sigma^2}$ es distribuido como $\chi'^2_{(p,\lambda)}$.

donde $\lambda = \frac{(\beta-\beta^*)^t X^t A_2 X(\beta-\beta^*)}{2\sigma^2}$.

iii) $\frac{\mathbf{Z}^t\mathbf{A}_1\mathbf{Z}}{\sigma^2}$ y $\frac{\mathbf{Z}^t\mathbf{A}_2\mathbf{Z}}{\sigma^2}$ son independientes.

Nosotros vemos que

$\frac{\mathbf{Z}^t\mathbf{A}_1\mathbf{Z}}{\sigma^2}$ está distribuida como $\chi'^2_{(k_1,\lambda_1)}$ donde $rango(\mathbf{A}_1) = k_1$ y

$$\lambda_1 = \frac{(\beta - \beta^*)^t \mathbf{X}^t \mathbf{A}_1 \mathbf{X} (\beta - \beta^*)}{2\sigma^2}$$

Puesto que A_1 es una matriz idempotente entonces

$$\begin{aligned} rango(\mathbf{A}_1) &= traza(\mathbf{A}_1) \\ &= tr(\mathbf{I}) - tr\left(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right) = n - p. \end{aligned}$$

y $\lambda_1 = 0$, ya que

$$\mathbf{A}_1\mathbf{X} = \left(\mathbf{I} - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right)\mathbf{X} = 0$$

Así que $\chi'^2_{(n-p,0)} = \chi^2_{(n-p)}$.

$\frac{\mathbf{Z}^t\mathbf{A}_2\mathbf{Z}}{\sigma^2}$ está distribuida como $\chi'^2_{(k_2,\lambda_2)}$ donde $rango(\mathbf{A}_2) = k_2$ y

$$\lambda_2 = \frac{(\beta - \beta^*)^t \mathbf{X}^t \left(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^t\right) \mathbf{X} (\beta - \beta^*)}{2\sigma^2}$$

De la misma manera como A_2 es una matriz idempotente de $rango(A_2) = tr(A_2) = p$ y

$$\lambda_2 = \frac{(\beta - \beta^*)^t \mathbf{X}^t\mathbf{X} (\beta - \beta^*)}{2\sigma^2}$$

Como $\mathbf{X}^t\mathbf{X}$ es semidefinida positiva entonces $\frac{\mathbf{Z}^t\mathbf{A}_2\mathbf{Z}}{\sigma^2}$ se distribuye como Ji-Cuadrada Cen-

tral si solo si $(\beta - \beta^*) = 0$, es decir si solo si $H_0 : \beta = \beta^*$ es verdadero.

Como $\frac{\mathbf{Z}'\mathbf{A}_1\mathbf{Z}}{\sigma^2}$ está distribuido como $\chi^2_{(n-p)}$

y $\frac{\mathbf{Z}'\mathbf{A}_2\mathbf{Z}}{\sigma^2}$ está distribuida como $\chi^2_{(n-p)}$

se sigue que

$$u = \frac{\mathbf{Z}'\mathbf{A}_2\mathbf{Z}}{\mathbf{Z}'\mathbf{A}_1\mathbf{Z}}\left(\frac{n-p}{p}\right)$$

se distribuye como $F'_{(p,n-p,\lambda)}$ si solo si H_0 es verdadera.

Y también se sigue que

$v = \frac{pu}{(n-p)+pu} = \frac{\mathbf{Z}'\mathbf{A}_2\mathbf{Z}}{\mathbf{Z}'\mathbf{Z}}$ se distribuye como una $E^2_{(p,n-p,\lambda)}$.

Ejemplo. Considere el modelo de regresión lineal simple.

$$Y_i = \beta_1 + \beta_2 x_i + e_i$$

donde e_i está distribuido como $N(0, \sigma^2)$ donde $n > 2$ y

$$\Omega = \left\{\left(\beta_1, \beta_2, \sigma^2\right) : \beta_1, \beta_2 \in R \text{ y } \sigma^2 > 0\right\}$$

Por lo tanto,

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \qquad e = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

$$S = X'X = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}, \qquad X'Y = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

$$S^{-1} = X'X = \frac{1}{n\sum\left(X_i - \overline{X}\right)^2} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix}$$

Como $\hat{\beta} = \mathbf{S^{-1}X'Y}$ entonces

$$\hat{\beta} = \frac{1}{n\sum\left(X_i - \overline{X}\right)^2} \begin{bmatrix} \sum Y_i \sum X_i^2 - \sum X_i \sum X_i Y_i \\ n\sum Y_i X_i - \sum X_i \sum Y_i \end{bmatrix}$$

Con lo cual

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_2 = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum\left(X_i - \overline{X}\right)^2}$$

3.9 Coeficiente de determinación R^2

Consideremos la bondad de ajuste de la línea de regresión ajustada a un conjunto de datos, por lo cual se intentará encontrar en que medida ajusta la línea de regresión muestral a los datos. Si todas las observaciones coincidieran con la línea de regresión, obtendríamos un ajuste "perfecto", lo que raras veces ocurre. Generalmente tienden a presentarse e_i positivos y negativos con la esperanza de que los residuos localizados alrededor de la línea de regresión sean lo más pequeños posibles.

En este sentido el coeficiente de determinación R^2 es una medida resumen que nos dice que también la línea de regresión muestral se ajusta a los datos. Utilizando el análisis de varianza obtendremos R^2. El análisis de varianza divide la variación total de las observaciones en sus partes componentes de acuerdo al modelo propuesto. Para esto analizaremos la desviación de la observación Y_i de la media de las observaciones $\overline{Y}$. Para esto consideraremos los siguientes dos casos:

a) Si suponemos que todas las observaciones Y_i son iguales entre sí, así que las $\beta_j = 0$, para $j = 1, ..., p$, $e_i = 0$ y $Y_i = \overline{Y}$ para toda i.

b) Por otro lado, si la magnitud de la desviación $Y_i - \overline{Y}$ es diferente de cero, ésta deberá atribuirse a las componentes del modelo.

Para la magnitud de la desviación

$$Y_i - \overline{Y}$$

sumemos y restemos el estimador $\widehat{Y}_i$ entonces

$$Y_i - \overline{Y} = \left(\widehat{Y}_i - \overline{Y}\right) + \left(Y_i - \widehat{Y}_i\right)$$

De aquí, la desviación total de la observación Y_i con respecto a la media $\overline{Y}$ es la suma de al desviación de $\widehat{Y}_i$ estimada de la media $\overline{Y}$ y la desviación de Y_i con respecto a $\widehat{Y}_i$.

Las desviaciones $Y_i - \widehat{Y}_i$ representan la contribución a la componente del error a la variación total. Si la magnitud de la desviación de $\widehat{Y}_i - \overline{Y}$ es grande, entonces se tiene un efecto lineal.

Siguiendo el análisis de varianza elevamos al cuadrado ambos miembros de la identidad

$$Y_i - \overline{Y} = \left(\widehat{Y}_i - \overline{Y}\right) + \left(Y_i - \widehat{Y}_i\right)$$

y se sumaran para todas las observaciones. Entonces se tiene

$$\sum \left(Y_i - \overline{Y}\right)^2 = \sum \left(\widehat{Y}_i - \overline{Y}\right)^2 + \sum \left(Y_i - \widehat{Y}_i\right)^2 + 2\sum \left(\widehat{Y}_i - \overline{Y}\right)\left(Y_i - \widehat{Y}_i\right)$$

Puesto que

$$\sum \left(\widehat{Y}_i - \overline{Y}\right)\left(Y_i - \widehat{Y}_i\right) = \sum \widehat{Y}_i \left(Y_i - \widehat{Y}_i\right) - \sum \overline{Y} \left(Y_i - \widehat{Y}_i\right)$$

y como $\sum \widehat{Y}_i \left(Y_i - \widehat{Y}_i\right) = \sum (\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})\, e_i = 0$

dado que $\beta_j \sum x_{ij} e_i = 0$ puesto que los residuos no están correlacionados con x_{ij}.

Por lo tanto, la ecuación del análisis de regresión es

$$\sum \left(Y_i - \overline{Y}\right)^2 = \sum \left(\widehat{Y}_i - \overline{Y}\right)^2 + \sum \left(Y_i - \widehat{Y}_i\right)^2$$

El término $\sum \left(Y_i - \overline{Y}\right)^2$ representa la variación total de las observaciones con respecto a la media (STC).

El término $\sum \left(\widehat{Y}_i - \overline{Y}\right)^2$ representa la variación total de los valores estimados de Y con respecto a la media (SCR).

El término $\sum \left(Y_i - \widehat{Y}_i\right)^2$ representa la variación residual o de las observaciones con respecto a los valores estimados (SEC).

Con lo cual, $STC = SEC + SCR$ donde nos muestra que la variación de los valores observados de Y_i alrededor de su media pueden dividirse en dos componentes, el primero en la línea de regresión y el segundo en los errores aleatorios. Dividiendo ambos miembros por STC, obtenemos que

$$1 = \frac{SEC}{STC} + \frac{SCR}{STC}$$

Tenemos que

$$\begin{aligned} SEC = \sum \left(Y_i - \widehat{Y}_i\right)^2 = \sum e_i^2 &= \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)' \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\beta} - \left(\mathbf{X}\hat{\beta}\right)' Y + \left(\mathbf{X}\hat{\beta}\right)' \mathbf{X}\hat{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\beta} - \hat{\beta}'\mathbf{X}'\mathbf{Y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y} \end{aligned}$$

Puesto que

$$\begin{aligned}\hat{\beta}^t\mathbf{X}^t\mathbf{X}\hat{\beta} &= (\mathbf{S}^{-1}\mathbf{X}^t\mathbf{Y})^t\,\mathbf{X}^t\mathbf{X}\hat{\beta} \\ &= \mathbf{Y}^t\mathbf{X}\hat{\beta}\end{aligned}$$

$$STC = \sum\left(Y_i - \overline{Y}\right)^2 = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n} = \mathbf{Y}^t\mathbf{Y} - n\overline{\mathbf{Y}}^2$$

Y como

$$\begin{aligned}SCR &= STC - SEC \\ &= \mathbf{Y}^t\mathbf{Y} - n\overline{\mathbf{Y}}^2 - \left(\mathbf{Y}^t\mathbf{Y} - \hat{\beta}^t\mathbf{X}^t\mathbf{Y}\right) \\ &= \hat{\beta}^t\mathbf{X}^t\mathbf{Y} - n\overline{\mathbf{Y}}^2\end{aligned}$$

Definición 31 *El coeficiente de determinación queda definido como*

$$R^2 = \frac{SCR}{STC} = \frac{STC - SEC}{STC} = 1 - \frac{SEC}{STC}$$

3.10 Análisis de Residuos

Siempre que se realiza un análisis estadístico, es importante comprobar que los datos observados cumplen las hipótesis en que se basa el análisis. En el modelo de regresión lineal múltiple una manera eficaz de descubrir posibles deficiencias en el modelo o violaciones de las suposiciones radica en llevar a cabo un análisis de residuos. Con el análisis de

residuos se puede detectar lo siguiente:

1. Si los residuos se distribuyen normalmente con una varianza constante.

2. Observar que una variable explicativa que ejerce influencia importante puede no estar incluida en el modelo.

3. Si se ha definido la ecuación de una manera correcta y no existe ninguna deficiencia.

4. Si hay observaciones aberrantes.

5. Si hay variables omitidas.

6. Si los residuos están correlacionados en vez de ser independientes como se supuso.

3.10.1 Gráfica de Probabilidad Normal

Un método simple para checar la suposición de normalidad es graficar los residuos en papel de probabilidad normal. Una gráfica de este tipo es la representación de la distribución acumulada de los residuos sobre papel de probabilidad normal. Si la distribución de los errores es normal, esta gráfica parecerá una línea recta. Al visualizar dicha línea hay que poner más énfasis en los valores centrales que en los extremos de la gráfica. Aunque pequeñas desviaciones de normalidad no afectan el modelo. Si llegamos al problema de que los residuos no siguen una distribución normal entonces los estimadores por mínimos cuadrados siguen siendo estimadores insesgados óptimos, pero todas las pruebas de significancia que apliquemos carecen de valor, ya que las estadísticas t o F y los intervalos de confianza y predicción dependen de la suposición de normalidad.

3.10.2 Gráfica de Residuos Contra Valores Estimados $\widehat{Y}_i$

Está gráfica sirve para detectar si la varianza del error es o no constante. Si el modelo es correcto y las suposiciones se satisfacen entonces la gráfica de los residuos contra los valores estimados $\widehat{Y}_i$ tenderán a encontrarse dentro de una banda horizontal centrada alrededor del valor cero, sin ninguna tendencia sistemática a ser positivos o negativos. Cualquier desviación con respecto a este comportamiento indicará la existencia de un problema. Un defecto que en ocasiones revela la gráfica es el de una varianza variable. Esto da origen a lo que se conoce como heterocedasticidad que analizaré en el capítulo cinco.

Capítulo 4

Multicolinealidad

Introducción.

En nuestro análisis del modelo de regresión lineal múltiple, nosotros asumimos que la matriz de variables explicativas, $\mathbf{X}$, tiene rango completo, es decir, no existe relación lineal entre las variables explicativas. En realidad para $(\mathbf{X}^t\mathbf{X})^{-1}$ exista, $\mathbf{X}$ debe ser de rango completo y también observamos que la interpretación del modelo, depende explícitamente o implícitamente de las estimaciones de los coeficientes. Las inferencias que generalmente se hacen son las siguientes :

i) Identificar los efectos relativos a las variables explicativas.

ii) Predecir y estimar.

iii) Seleccionar un conjunto apropiado de variables para el modelo.

Si no existe relación entre las variables explicativas, se dice que son ortogonales y bajo está condición, se pueden realizar las anteriores inferencias relativamente fácil. Algunas veces la falta de ortogonalidad es no seria. Sin embargo, en algunas situaciones las variables explicativas son casi linealmente dependientes y en tales casos las inferencias basadas en el modelo de regresión pueden ser erróneas o engañosas. En este capítulo analizaremos que pasa, si las variables explicativas, X, son linealmente dependientes o casi existe una relación de dependencia lineal.

4.1 El Significado de la Multicolinealidad

Una definición precisa de multicolinealidad no ha sido firmemente establecida en la literatura. Literalmente se dice que p variables explicativas tienen multicolinealidad si el vector que representan está situado en el subespacio de dimensión menor que p, es decir, si al menos uno de los vectores es una combinación lineal de otros. En la práctica, tal multicolinealidad raramente ocurre, por lo que el término generalmente se utiliza cuando las variables explicativas son casi linealmente dependientes.

De acuerdo con la precedente discusión la multicolinealidad tiene que ver con las características especificas de la matriz de datos $\mathbf{X}$ y no con los aspectos estadísticos de el modelo de regresión $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$, es decir , la multicolinealidad es un problema de los datos.

4.2 Efectos de la Multicolinealidad

La presencia de la multicolinealidad tiene un número potencialmente serio de efectos en la estimación por mínimos cuadrados de los coeficientes de regresión.

Sea el modelo

$$Y = \beta_1 X_1 + \beta_2 X_2 + e$$

las ecuaciones normales son:

$$\left(\mathbf{X}'\mathbf{X}\right)\hat{\beta} = \mathbf{X}'\mathbf{Y}$$

$$\begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

donde r_{12} es la correlación simple entre X_1 y X_2 y r_{iy} es la correlación simple entre X_i y Y para $i = 1, 2$.

Con lo cual,

$$S^{-1} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \begin{bmatrix} \frac{1}{(1-r_{12}^2)} & \frac{-r_{12}}{(1-r_{12}^2)} \\ \frac{-r_{12}}{(1-r_{12}^2)} & \frac{1}{(1-r_{12}^2)} \end{bmatrix}$$

y la estimación de los coeficientes son:

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{(1 - r_{12}^2)}, \qquad \hat{\beta}_2 = \frac{r_{1y} - r_{12}r_{1y}}{(1 - r_{12}^2)}$$

Si la multicolinealidad es bastante fuerte entre X_1 y X_2 entonces el coeficiente de correlación r_{12} es muy grande. Esto es debido a que cuando $|r_{12}| \to 1$ entonces $Var\left(\hat{\beta}_i\right) = \sigma_i^2 S_{ii} \to \infty$ y $Cov\left(\hat{\beta}_1, \hat{\beta}_2\right) = \sigma^2 S_{12} \to \pm\infty$.

Por lo tanto, cuando la multicolinealidad es severa entre X_1 y X_2 entonces los estimadores de los coeficientes tienen varianza y covarianza muy grande.

4.3 Como Detectar la Multicolinealidad

1. Examinar la matriz de correlación.

Una medida simple de multicolinealidad es por inspección de los elementos fuera de la diagonal r_{ij} de la matriz de correlación. Si las variables explicativas X_i y X_j tienen un coeficiente de correlación $|r_{ij}|$ cercano a la unidad entonces la multicolinealidad existe. Pero el problema de este criterio es que cuando hay más de dos variables explicativas aunque los coeficientes de correlación r_{ij} están cercanos a la unidad sugieren la existencia de multicolinealidad, no es necesario que dichos coeficientes estén cercanos a la unidad para que exista multicolinealidad en un determinado caso específico. Es decir, los coeficientes de correlación cercanos a la unidad, cuando hay mas de dos variables explicativas es una condición suficiente pero no necesaria para la existencia de multicolinealidad, debido a que ésta puede existir, a pesar de que los coeficientes de correlación sean bajos.

2. *La estadística F es significativa pero las estadísticas individuales de t no son significativas entonces la multicolinealidad existe.*

Antes de analizar este criterio, primero recordemos lo que quiere decir que una estadística es significativa y cuando no es significativa. Una estadística es significativa si el valor de la estadística de prueba se encuentra en la región crítica y por lo tanto se rechaza la hipótesis nula y cuando no es significativa el valor de la prueba no se encuentra en la región crítica y en este caso no se rechaza la hipótesis nula.

Con la estadística t evaluamos la hipótesis para la prueba de significancia de cualquier coeficiente individual del modelo de regresión Por lo tanto para β_j, evaluamos la hipótesis nula

$$H_0 : \beta_j = 0$$

contra la hipótesis alternativa

$$H_1 : \beta_j \neq 0$$

Si no rechazamos la hipótesis nula $H_0 : \beta_j = 0$ entonces esto indica que Y está relacionada con la variable explicativa X_j.

Con la estadística F evaluamos la hipótesis para la prueba de significancia, para determinar si hay una relación lineal entre la variable respuesta **Y** y las variables explicativas $X_1, X_2, \ldots, X_p$. Por lo tanto, evaluamos la hipótesis nula

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

contra la hipótesis alternativa

$$H_1 : \beta_j \neq 0 \qquad \textit{para al menos una } j.$$

Si se rechaza la hipótesis nula $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p$ entonces esto indica que al menos una de las variables explicativas $X_1, X_2, \ldots, X_p$ contribuye significativamente al modelo.

Por lo tanto, con este criterio no rechazamos individualmente los coeficientes de las variables explicativas sean cero. Sin embargo, cuando evaluamos conjuntamente los coeficientes de las variables explicativas están íntimamente relacionadas que es imposible aislar el impacto individual de las variables explicativas X_j.

3. *El factor de inflación de la varianza*

La covarianza de $\hat{\beta}$ es $cov(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ donde si $C = (\mathbf{X}'\mathbf{X})^{-1}$, tenemos que los elementos de la diagonal son la varianza de los coeficientes del modelo de regresión lineal, es decir,

$$V(\beta_j) = \sigma^2 C_{jj} = \sigma^2 \frac{1}{1 - R_j^2}$$

donde R_j^2 es el coeficiente de determinación obtenido cuando hacemos una regresión de X_j con el resto de las $p - 1$ variables explicativas. Si X_j es casi ortogonal el resto de las variables explicativas, R_j^2 es pequeño y por lo tanto C_{jj} es cercano a uno, mientras si X_j es casi linealmente dependiente, R_j^2 es cercano a uno y por lo tanto C_{jj} tiende a infinito. Como la varianza del coeficiente $j - ésimo$ es $V(\hat{\beta}_j) = \sigma^2 C_{jj}$ entonces C_{jj} es un factor

para el cual la varianza de $\hat{\beta}_j$ se incrementa debido a una casi dependencia lineal entre las variables explicativas.

Así, que

$$VIF_j = C_{jj} = \frac{1}{1 - R_j^2}$$

será llamado factor de inflación de la varianza . Diremos que existe la multicolinealidad si VIF_j es muy grande.

4. Examinar los eigenvalores de $X'X$ (eigensistema)

Este método utiliza los eigenvalores $\lambda_1, \ldots, \lambda_p$ de la matriz $\mathbf{X'X}$ como medida de multicolinealidad. Este método fue propuesto por Kendall (1957) y Silvey (1969) y dice que si existe casi dependencia lineal entre las columnas de X entonces existen eigenvalores muy pequeños. Este criterio tiene el defecto que no nos dice que significa que un eigenvalor es muy pequeño.

5. Indice de condición

Este criterio se basa en la descomposición de eigenvalores de $\mathbf{X}$ y consiste en descomponer la matriz X como

$$\mathbf{X} = \mathbf{UDV}'$$

donde $\mathbf{X}$ es una matriz de $n \times p$, $\mathbf{U'U} = \mathbf{V'V} = \mathbf{I}_p$ y $\mathbf{D}$ es una matriz diagonal que contiene los eigenvalores $\mu_1, \ldots, \mu_p$ de $\mathbf{X}$. La descomposición de eigenvalores está es-

trechamente relacionada con el anterior criterio, pero existen algunas diferencias que a continuación se discutira.

Tenemos que $\mathbf{X}^t\mathbf{X} = \left(\mathbf{UDV}^t\right)^t \mathbf{UDV}^t = \mathbf{VD}^2\mathbf{V}^t$ así que los eigenvalores al cuadrado de X son los eigenvalores de $\mathbf{X}^t\mathbf{X}$ y $\mathbf{V}$ es la matriz de eigenvectores de $\mathbf{X}^t\mathbf{X}$. Con lo cual la descomposición de eigenvalores de la matriz $\mathbf{X}$, provee información que abarca el dado por el eigensistema $\mathbf{X}^t\mathbf{X}$. Sin embargo, las razones para preferir la descomposición de eigenvalores son las siguientes:

a) La descomposición de eigenvalores se aplica directamente a la matriz $\mathbf{X}$ y el eigensistema a la matriz $\mathbf{X}^t\mathbf{X}$.

b) La descomposición de eigenvalores es más estable numéricamente que el eigensistema.

Si la matriz X tiene columnas linealmente dependientes entonces $rango(\mathbf{X}) = r$ y $r < p$. Ya que en la descomposición de eigenvalores de $\mathbf{X}$, $\mathbf{U}$ y $\mathbf{V}$ son ortogonales entonces $rango(\mathbf{X}) = rango(\mathbf{D})$. Por lo tanto existen tantos eigenvalores iguales a cero como columnas linealmente dependientes de X.

Esto es debido a que si particionamos la matriz X como

$$\mathbf{X} = \mathbf{UDV}^t = \mathbf{U}\begin{bmatrix} \mathbf{D}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{V}^t$$

donde $\mathbf{D}_{11}$ es una matriz diagonal no singular de tamaño $r \times r$. Si multiplicamos por $\mathbf{V}$ y además particionamos, tenemos que

$$X\begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{D}_{11} & 0 \\ 0 & 0 \end{bmatrix}$$

donde $\mathbf{V}_1$ es una matriz de $p \times r$, $\mathbf{U}_1$de $n \times r$, $\mathbf{V}_2$ de $p \times (p-r)$ y $\mathbf{U}_2$ es de $n \times (p-r)$

Por lo tanto

$$\mathbf{XV}_1 = \mathbf{U}_1\mathbf{D}_{11}$$

$$\mathbf{XV}_2 = \mathbf{0}$$

De aquí notamos que $\mathbf{XV}_2 = \mathbf{0}$ provee de un espacio nulo asociado con las columnas de X.

Con lo cual, si $\mathbf{X}$ posee $p-r$ columnas linealmente dependientes entonces existen $p-r$ eigenvalores iguales a cero en $\mathbf{D}$. Este resultado fue en el que se basaron Kendall y Silvey para determinar que existe multicolinealidad si existen eigenvalores pequeños.

El índice de condición tiene como objetivo saber cuando una matriz este bien condicionada. Se dice que una matriz esta bien condicionada si cambios relativamente pequeños en los coeficientes del sistema de ecuaciones lineales de la forma $\mathbf{AX} = b$, provocará cambios relativamente pequeños en la solución; lo contrario se llama mal condicionada. Con lo cual, debemos saber cuando una matriz está bien condicionada, para esto definiremos la norma espectral de una matriz A como sigue

$$\|A\| = \sup_{\|z\|=1} \|Az\|$$

que cumple las siguientes relaciones

a) $\|AX\| \leq \|AX\|$

b) $\|AB\| \leq \|A\| \|B\|$

Sea A una matriz no singular, $b \neq 0$ y $AX = b$ donde los cambios relativos de A, X y b se denotarán por δA, δX y δb respectivamente. Si A es fija y existen cambios en X y b. Tenemos por lo tanto, que

$$A(X + \delta X) = b + \delta b$$

con lo cual,

$$A\delta X = \delta b$$

implica que $\delta X = A^{-1}\delta b$

Con esto,

$$\|b\| = \|AX\| \leq \|A\| \|X\|$$

$$\|\delta X\| = \|A^{-1}\delta b\| \leq \|A^{-1}\| \|\delta b\|$$

entonces multiplicando las dos anteriores desigualdades obtenemos

$$\frac{\|\delta X\|}{\|X\|} = \|A\| \left\|A^{-1}\right\| \frac{\|\delta b\|}{\|b\|}$$

donde $k(A) = \|A\| \|A^{-1}\|$ nos da una cota de que cambios relativos en X resultan en cambios relativos b.

De la misma manera, se puede comprobar que

$$\frac{\|\delta X\|}{\|X + \delta X\|} \leq \|A\| \left\|A^{-1}\right\| \frac{\|\delta A\|}{\|A\|}$$

otra ves $k(A) = \|A\| \|A^{-1}\|$ nos provee de una cota para cambios relativos en A resulta en cambios relativos en b. Por lo tanto un índice de condición grande para la matriz A implica que A está mal condicionada.

Como nosotros sabemos que la norma espectral de una matriz cuadrada A no singular es

$$\|A\| = \sup_{\|z\|=1} \|Az\|$$

entonces $\|A\| = \mu_{\max}$ y $\|A^{-1}\| = \frac{1}{\mu_{\min}}$ donde $\mu_{\max}$ y $\mu_{\min}$ corresponden a los eigenvalores máximo y mínimo de la matriz A.

Por lo tanto, para medir el índice de condición de el modelo de regresión lineal $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, donde la matriz X es de tamaño $n \times p$ se define como

$$k(X) = \|X\| \left\|X^{-1}\right\| = \frac{\mu_{\max}}{\mu_{\min}}$$

Con lo cual, para medir el grado en que una matriz está mal condicionada depende por

lo tanto de que tan pequeño sea el eigenvalor mínimo relativo al eigenvalor máximo.

También podemos definir

$$\eta_k = \frac{\mu_{\max}}{\mu_k} \qquad k = 1, \ldots, p$$

como el índice de condición de la matriz X de tamaño $n \times p$.

De acuerdo a esto podemos concluir que hay tantas columnas casi linealmente dependientes como índice de condición grandes.

4.4 Soluciones a la Multicolinealidad

Hasta ahora nos hemos ocupado de cómo detectar la multicolinealidad. Cuando enfretamos el problema de la multicolinealidad, algunas de las soluciones son:

Obtención de datos adicionales

Eliminación de variables

Obtención de datos adicionales

Regresión Ridge

Una solución al problema de la multicolinealidad frecuentemente sugerido es obtener más datos. Esta recomendación tiene cierta importancia , pero no se debe tomar al

pie de la letra. Esta técnica es de la mas frecuentemente usadas y lo relevante no es el número de observaciones, sino el contenido informático. Por ejemplo, si al principio sólo tenemos datos anuales, podemos tratar de obtener datos trimestrales o mensuales, o bien una combinación de datos de series de tiempo y de sección cruzada. Pero esta combinación no resuelve necesariamente nuestro problema de datos confiables. A veces podemos estar añadiendo otras fuentes de variación , por ejemplo, si pasamos de datos anuales a datos mensuales, se introduce una estacionalidad, y los datos de sección cruzada introducen una variación de sección cruzada. También habría que tener en cuenta que los datos desagregados no sean simple interpolaciones. Si, por ejemplo, la serie mensual es una interpolación de la serie trimestral, acudiendo entonces a los datos mensuales obtendremos el triple de observaciones, pero esto es simplemente un incremento.

Eliminación de variables.

Está técnica es de las más frecuentemente utilizadas y consiste en eliminar del modelo una o mas de las variables colineales. Sin embargo, al eliminar una variable del modelo podemos cometer un sesgo de especificación . El sesgo de especificación surge debido a la especificación incorrecta del modelo utilizado en el análisis. Por lo tanto, la multicolinealidad provoca estimaciones no precisas de los parámetros del modelo, eliminar una variable puede llevar a equivocarse seriamente con respecto a los verdaderos valores de los parámetros.

4.5 Regresión Rigde

Hemos visto que cuando existe multicolinealidad las estimaciones de los coeficientes de regresión por mínimos cuadrados son muy grandes e inestables, es decir, sus magnitudes

y signos pueden cambiar considerablemente dadas diferentes muestras. Con el método de mínimos cuadrados el teorema de Gauss-Markov nos asegura que este estimador tiene varianza mínima dentro de la clase de los estimadores lineales insesgados, pero esto no nos garantiza que su varianza sea pequeña.

Una manera de resolver este problema es eliminar el requerimiento de que el estimador de $\hat{\beta}$ sea insesgado y suponer que se puede encontrar un estimador $\hat{\beta}_R$ sesgado de β, tal que tenga una varianza más pequeña, que el estimador $\hat{\beta}$ insesgado de β. Es decir, el objetivo es obtener un mejor estimador de β, que el obtenido por mínimos cuadrados para el modelo de regresión lineal múltiple cuando existe multicolinealidad.

Para mejorar un estimador se consideran generalmente los siguientes factores:

a) El estimador que se desea mejorar.

b) El parámetro.

c) La función f, que nos lleva a mejorar el estimador.

d) El criterio de comparación.

e) La región A donde se logra mejorar el estimador.

Siguiendo estos factores Hoerl y Kennard (1970) buscaron una solución mejorada al estimador $\hat{\beta}$, en el modelo de regresión lineal múltiple, cuando existe multicolinealidad. La técnica que de esto se desprende se conoce como *Regresión Ridge*.

4.5.1 Caracterización del Estimador Ridge

Otra manera, de observar que la multicolinealidad produce estimaciones muy grandes en valor absoluto es considerar la distancia de $\hat{\beta}$ a β al cuadrado, es decir,

$$d^2 = \left(\hat{\beta}-\beta\right)'\left(\hat{\beta}-\beta\right)$$

el valor esperado de está distancia, es

$$\begin{aligned} E\left[d^2\right] &= E\left[\left(\hat{\beta}-\beta\right)'\left(\hat{\beta}-\beta\right)\right] \\ &= \sum_{j=1}^{p} E\left[\left(\hat{\beta}_j-\beta_j\right)^2\right] \\ &= \sum_{j=1}^{p} V\left(\hat{\beta}\right) \\ &= \sigma^2 tr\left(\mathbf{X}'\mathbf{X}\right)^{-1} \end{aligned} \tag{4.1}$$

y como la traza de una matriz es igual a la suma de los elementos de la diagonal entonces también es igual a la suma de sus eigenvalores, es decir,

$$E\left[d^2\right] = E\left[\left(\hat{\beta}-\beta\right)'\left(\hat{\beta}-\beta\right)\right] = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$$

donde $\lambda_j > 0$ para $j = i, ..., p$ son eigenvalores de $\mathbf{X}'\mathbf{X}$. Por lo tanto, si existe multicolinealidad al menos uno de los eigenvalores de $\mathbf{X}'\mathbf{X}$ debe ser pequeño y esto implica que la distancia de $\hat{\beta}$ a β al cuadrado será muy grande.

En resumen, con el método de mínimos cuadrados el teorema de Gauss-Markov nos asegura que este estimador tiene varianza mínima dentro de la clase de los estimadores lineales insesgados, pero esto no nos garantiza que su varianza sea pequeña, cuando existe multicolinealidad. De tal manera, Hoerl y Kennard buscaron otro estimador $\hat{\beta}_R$

de β con varianza más pequeña, pero como dentro de los estimadores insesgados no se puede mejorar está varianza, entonces buscaron entre los sesgados.

Entonces el problema se reduce a mejorar el estimador $\hat{\beta}$. Tenemos el estimador que queremos mejorar, el parámetro en cuestión, nos falta encontrar una función f y una región A tal que exista un estimador $\hat{\beta}_R = f(\hat{\beta}, a)$, (donde $a \in A$) que mejore el error cuadrático medio, la cual es la función de riesgo que utilizaron como criterio de mejoría Hoerl y Kennard. Por lo anterior, sea $\hat{\beta}_R$ un estimador sesgado cualquiera y sea $\hat{\beta}$ el estimador obtenido por el método de mínimos cuadrados y la función de riesgo el error cuadrático medio (ECM).

El error cuadrático medio para el estimador de mínimos cuadrados es:

$$\begin{aligned} ECM(\hat{\beta}) &= E\left[\left(\hat{\beta}-\beta\right)'\left(\hat{\beta}-\beta\right)\right] \\ &= E\left[\hat{\beta}'\hat{\beta} - \hat{\beta}'\beta - \beta'\hat{\beta} + \beta'\beta\right] \\ &= E\left[\hat{\beta}'\hat{\beta} - 2\beta'\hat{\beta} + \beta'\beta\right] \\ &= E(\hat{\beta}'\hat{\beta}) - 2\beta' E(\hat{\beta}) + \beta'\beta \\ &= E(\hat{\beta}'\hat{\beta}) - 2\beta'\beta + \beta'\beta \\ &= E(\hat{\beta}'\hat{\beta}) - \beta'\beta \end{aligned}$$

y para el estimador $\hat{\beta}_R$ sesgado encontramos que

$$\begin{aligned} ECM(\hat{\beta}_R) &= E\left[\left(\hat{\beta}_R-\beta\right)'\left(\hat{\beta}_R-\beta\right)\right] \\ &= E\left[\hat{\beta}_R'\hat{\beta}_R - \hat{\beta}_R'\beta - \beta'\hat{\beta}_R + \beta'\beta\right] \\ &= E\left[\hat{\beta}_R'\hat{\beta}_R - 2\beta'\hat{\beta}_R + \beta'\beta\right] \\ &= E(\hat{\beta}_R'\hat{\beta}_R) - 2\beta' E(\hat{\beta}_R) + \beta'\beta \\ &= E(\hat{\beta}'\hat{\beta}) - 2\beta'(\beta+s) + \beta'\beta \quad \textit{donde } s=\textit{sesgo} \\ &= E(\hat{\beta}_R'\hat{\beta}_R) - 2\beta' s - \beta'\beta \end{aligned}$$

y como el objetivo es encontrar un estimador $\hat{\beta}_R$ sesgado que cumpla lo siguiente

$$ECM(\hat{\beta}_R) < ECM(\hat{\beta})$$

entonces se tiene que

$$E(\hat{\beta}_R'\hat{\beta}_R) - 2\beta's < E(\hat{\beta}'\hat{\beta})$$

y como también la expresión general para el error cuadrático es

$$ECM() = Var() + (sesgo)^2$$

entonces se tiene de igual manera que

$$Var(\hat{\beta}_R) + (sesgo)^2 < Var(\hat{\beta})$$

Como

$$\begin{aligned} ECM(\hat{\beta}) &= E\left[\left(\hat{\beta}-\beta\right)'\left(\hat{\beta}-\beta\right)\right] \\ &= E(\hat{\beta}'\hat{\beta}) - \beta'\beta \\ &= \sigma^2 tr\,(\mathbf{X}'\mathbf{X})^{-1} \ \textit{esto por (4.1)} \end{aligned}$$

entonces

$$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 tr(\mathbf{X}'\mathbf{X})^{-1}$$

es decir, el método de mínimos cuadrados produce estimaciones de los coeficientes de regresión que son mayores en valor absoluto que los verdaderos valores. En vista de esto, para acercarme a β en valor promedio (reducir el ECM) se debe comenzar por acortar la longitud del vector estimador $\hat{\beta}_R$ y se debe encontrar la región A donde

$$ECM(\hat{\beta}_R) < ECM(\hat{\beta})$$

y como la suma de los errores al cuadrado para el estimador $\hat{\beta}_R$ es

$$\begin{aligned}\Phi &= \left(\mathbf{Y}-\mathbf{X}\hat{\beta}_R\right)'\left(\mathbf{Y}-\mathbf{X}\hat{\beta}_R\right)\\ &= \left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)'\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)+\left(\hat{\beta}_R-\hat{\beta}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\beta}_R-\hat{\beta}\right)\\ &= \Phi_{\min}+\Phi\left(\hat{\beta}_R\right)\end{aligned}$$

donde $\Phi_{\min}$ es la que se obtiene con el estimador de mínimos cuadrados y $\Phi\left(\hat{\beta}_R\right)$ es el valor de la forma cuadrática $\left(\hat{\beta}_R-\hat{\beta}\right)$. Resulta que existe un continuo de valores de $\hat{\beta}_{R_0}$ que satisface

$$\Phi=\Phi_{\min}+\Phi_0$$

donde $\Phi_0 > 0$ es un incremento fijo. Por lo tanto, si nos movemos en la superficie de sumas de cuadrados se debe hacer en la dirección en la que se acorte la longitud del vector $\hat{\beta}_R$. Formalmente, para Φ_0 fijo , buscamos el estimador $\hat{\beta}_R$ con longitud mínima. Es decir, debemos minimizar

$$\hat{\beta}_R'\hat{\beta}_R \quad \textit{sujeto a} \quad \left(\hat{\beta}_R-\hat{\beta}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\beta}_R-\hat{\beta}\right)=\Phi_0$$

como este es un problema de Lagrange se tiene que minimizar

$$F=\hat{\beta}_R'\hat{\beta}_R+\frac{1}{k}\left(\hat{\beta}_R-\hat{\beta}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\beta}_R-\hat{\beta}\right)-\Phi_0$$

donde $\frac{1}{k}$ es el multiplicador. Entonces

$$\frac{\partial F}{\partial\hat{\beta}_R}=2\hat{\beta}_R+\frac{1}{k}\left[2\left(\mathbf{X}'\mathbf{X}\right)\hat{\beta}_R-2\left(\mathbf{X}'\mathbf{X}\right)\hat{\beta}\right]=0$$

$$2\hat{\beta}_R + \frac{1}{k}\left[2\left(\mathbf{X}'\mathbf{X}\right)\hat{\beta}_R - 2\left(\mathbf{X}'\mathbf{X}\right)\hat{\beta}\right] = 0$$

$$2\left[\mathbf{I}+\frac{1}{k}\left(\mathbf{X}'\mathbf{X}\right)\right]\hat{\beta}_R - \frac{2}{k}\left(\mathbf{X}'\mathbf{X}\right)\hat{\beta} = 0$$

$$\left[\mathbf{I}+\frac{1}{k}\left(\mathbf{X}'\mathbf{X}\right)\right]\hat{\beta}_R = \frac{1}{k}\left(\mathbf{X}'\mathbf{X}\right)\hat{\beta}$$

por lo tanto

$$\begin{aligned}
\hat{\beta}_R &= \left[\mathbf{I}+\tfrac{1}{k}\left(\mathbf{X}'\mathbf{X}\right)\right]^{-1}\tfrac{1}{k}\left(\mathbf{X}'\mathbf{X}\right)\hat{\beta} \\
&= \left[\tfrac{1}{k}\left(k\mathbf{I}+\left(\mathbf{X}'\mathbf{X}\right)\right)\right]^{-1}\tfrac{1}{k}\left(\mathbf{X}'\mathbf{X}\right)\hat{\beta} \\
&= \left[\left(\mathbf{X}'\mathbf{X}\right)+k\mathbf{I}\right]^{-1}\left(\mathbf{X}'\mathbf{X}\right)\hat{\beta} \\
&= \left[\left(\mathbf{X}'\mathbf{X}\right)+k\mathbf{I}\right]^{-1}\mathbf{X}'\mathbf{Y}
\end{aligned}$$

El estimador $\hat{\beta}_R$ que tiene longitud mínima, manteniendo la suma de los cuadrados de los errores en un contorno determinado es el estimador Ridge. Con esto se ha resuelto parte del problema. Se tiene un estimador lo único que falta analizar si existen valores de k para los cuales

$$ECM(\hat{\beta}_R) < ECM(\hat{\beta})$$

esto se analiza en la siguiente subsección. En la práctica se evalúa n los coeficientes de regresión para algunos valores de k y se hace una gráfica simultánea de los coeficientes contra k, a lo que se llama traza de Ridge y cuando los coeficientes se estabilizan, tienen varianza muy pequeña.

4.5.2 Propiedades de la Estimación Ridge

Como el estimador Ridge es

$$\left[\left(\mathbf{X}'\mathbf{X}\right) + k\mathbf{I}\right]\hat{\beta}_R = \mathbf{X}'\mathbf{Y}$$

o

$$\hat{\beta}_R = \left[\left(\mathbf{X}'\mathbf{X}\right) + k\mathbf{I}\right]^{-1}\mathbf{X}'\mathbf{Y}$$

donde $k \geq 0$ es una constante elegida por el estadístico

Las propiedades del estimador Ridge son las siguientes:

1. El estimador Ridge es una transformación lineal del estimador por mínimos cuadrados

$$\begin{aligned}\hat{\beta}_R &= [(\mathbf{X}'\mathbf{X}) + k\mathbf{I}]^{-1}\mathbf{X}'\mathbf{Y} \\ &= [(\mathbf{X}'\mathbf{X}) + k\mathbf{I}]^{-1}\mathbf{X}'\mathbf{X}\hat{\beta} \\ &= \mathbf{Z}\hat{\beta} \quad \textit{donde} \quad \mathbf{Z} = [(\mathbf{X}'\mathbf{X}) + k\mathbf{I}]^{-1}\mathbf{X}'\mathbf{X}\end{aligned}$$

2. Si $k = 0$ los estimadores coinciden

$$\hat{\beta}_R = \left[\left(\mathbf{X}'\mathbf{X}\right) + k\mathbf{I}\right]^{-1}\mathbf{X}'\mathbf{X}\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\hat{\beta} = \hat{\beta}$$

3. El estimador Ridge es sesgado

$$E\left[\hat{\beta}_R\right] = E\left[\mathbf{Z}\hat{\beta}\right] = \mathbf{Z}E\left[\hat{\beta}\right] = \mathbf{Z}\beta$$

4. La $Var\left(\hat{\beta}_R\right) = \sigma^2 tr\left(((\mathbf{X}'\mathbf{X}) + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}((\mathbf{X}'\mathbf{X}) + k\mathbf{I})^{-1}\right)$

$$\begin{aligned} Var\left(\hat{\beta}_R\right) &= Var\left(\mathbf{Z}\hat{\beta}\right) = Var\left(\mathbf{Z}\hat{\beta}\right) \\ &= \sigma^2 tr\left(\mathbf{Z}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{Z}'\right) \\ &= \sigma^2 tr\left[(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\right] \\ &= \sigma^2 tr\left[(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\right] \end{aligned}$$

5. Si $\mathbf{Z} = [(\mathbf{X}'\mathbf{X}) + k\mathbf{I}]^{-1}\mathbf{X}'\mathbf{X}$ *y* $\mathbf{W} = [(\mathbf{X}'\mathbf{X}) + k\mathbf{I}]^{-1}$ *entonces* $\mathbf{Z} = \mathbf{I} - k\mathbf{W}$

$$\mathbf{Z} + k\mathbf{W} = \mathbf{W}\mathbf{X}'\mathbf{X} + k\mathbf{W} = \mathbf{W}(\mathbf{X}'\mathbf{X}+k\mathbf{I}) = [(\mathbf{X}'\mathbf{X}) + k\mathbf{I}]^{-1}(\mathbf{X}'\mathbf{X}+k\mathbf{I}) = \mathbf{I}$$

entonces

$$\mathbf{Z} = \mathbf{I} - k\mathbf{W}$$

6. El $ECM\left(\hat{\beta}_R\right) = \sigma^2 tr\left[(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\right] + k^2\beta'[(\mathbf{X}'\mathbf{X}) + k\mathbf{I}]^{-2}\beta$

$$
\begin{aligned}
ECM\left(\hat{\beta}_R\right) &= E\left[\left(\hat{\beta}_R-\beta\right)^2\right] \\
&= E\left[\left(\hat{\beta}_R-E\left(\hat{\beta}_R\right)\right)^2\right]+\left[E\left(\hat{\beta}_R\right)-\beta\right]^2 \\
&= E\left[\left(\mathbf{Z}\hat{\beta}-\mathbf{Z}\beta\right)^2\right]+\left[(\mathbf{Z}\beta-\beta)^2\right] \\
&= Var\left(\mathbf{Z}\hat{\beta}\right)+(\mathbf{Z}\beta-\beta)^t(\mathbf{Z}\beta-\beta) \\
&= \sigma^2 tr\left[(\mathbf{X}^t\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X}+k\mathbf{I})^{-1}\right]+\beta^t(\mathbf{Z}-\mathbf{I})^t(\mathbf{Z}-\mathbf{I})\beta \\
&= \sigma^2 tr\left[(\mathbf{X}^t\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X}+k\mathbf{I})^{-1}\right]+\beta^t(-k\mathbf{W})^t(-k\mathbf{W})\beta \\
&= \sigma^2 tr\left[(\mathbf{X}^t\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X}+k\mathbf{I})^{-1}\right]+k^2\beta^t\left[(\mathbf{X}^t\mathbf{X})+k\mathbf{I}\right]^{-2}\beta \\
&= \gamma_1(k)+\gamma_2(k)
\end{aligned}
$$

A continuación demostraremos que existe al menos un $k>0$ tal que

$$ECM(\hat{\beta}_R) < ECM(\hat{\beta})$$

Teorema 30 *La varianza total $\gamma_1(k)$ es una función monótona decreciente de k y continua.*

Demostración

$$
\begin{aligned}
\gamma_1(k) &= \sigma^2 tr\left[(\mathbf{X}^t\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X}+k\mathbf{I})^{-1}\right] \\
&= \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2} \\
&= \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{\left[\lambda_i\left(1+\frac{k}{\lambda_i}\right)\right]^2} \\
&= \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i\left(1+\frac{k}{\lambda_i}\right)^2}
\end{aligned}
$$

Puesto que $\lambda_i > 0$ para cada i y la función $\frac{k}{\lambda_i}$ es monótona creciente cuando k crece

entonces cada sumando es monótonamente decreciente, por lo tanto, $\gamma_1(k)$ es monótona decreciente. Y $\gamma_1(k)$ es continua por ser suma de funciones continuas.

Corolario 1 *La primera derivada de la varianza total* $\gamma_1(k)$ *se aproxima a* $-\infty$ *cuando* $k \to 0^+$ *y* $\lambda_p \to 0$

Demostración

$$\gamma_1'(k) = -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3} < 0$$

por lo tanto,

$$\lim_{\substack{k\to 0^+ \\ \lambda_p \to 0}} \left(-2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3}\right) \to -\infty$$

Teorema 31 *El sesgo al cuadrado* $\gamma_2(k)$ *es una función monótona creciente y continua de* k.

Demostración

$$\gamma_2(k) = k^2\beta^t\left(\mathbf{X}^t\mathbf{X} + k\mathbf{I}\right)^{-2}\beta$$

Si $\mathbf{\Lambda}$ es la matriz de eigenvalores de $\mathbf{X}^t\mathbf{X}$ y $\mathbf{P}$ es una matriz ortogonal tal que

$$\mathbf{X}^t\mathbf{X} = \mathbf{P}^t\mathbf{\Lambda}\mathbf{P}$$

$$\begin{aligned}\gamma_2(k) &= k^2\beta^t\left[\mathbf{P}^t\mathbf{\Lambda}\mathbf{P}+k\mathbf{P}^t\mathbf{I}\mathbf{P}\right]^{-2}\beta \\ &= k^2\beta^t\left[\mathbf{P}^t\left(\mathbf{\Lambda}+k\mathbf{I}\right)\mathbf{P}\right]^{-2}\beta\end{aligned}$$

Si definimos $\alpha = \mathbf{P}\beta$ entonces $\sum_{i=1}^{p}\alpha_i^2 = \alpha^t\alpha = \beta^t\mathbf{P}^t\mathbf{P}\beta = \beta^t\beta$ obtenemos que

$$\gamma_2(k) = k^2\sum_{i=1}^{p}\frac{\alpha_i^2}{(\lambda_i+k)^2}$$

Puesto que $\lambda_i > 0$ para cada i y $k \geq 0$ entonces cada elemento $\lambda_i + k > 0$ y $\gamma_2(0) = 0$, por lo tanto, no existe indeterminación en ningún sumando. Además, $\gamma_2(k)$ se puede escribir como

$$\begin{aligned}\gamma_2(k) &= k^2\sum_{i=1}^{p}\frac{\alpha_i^2}{\left[k\left(1+\frac{\lambda_i}{k}\right)\right]^2} \\ &= \sum_{i=1}^{p}\frac{\alpha_i^2}{\left(1+\frac{\lambda_i}{k}\right)^2}\end{aligned}$$

donde cada $\frac{\lambda_i}{k}$ es monótona decreciente cuando k crece entonces cada sumando es monótona creciente y $\gamma_2(k)$ es continua por ser suma de funciones continuas.

Corolario 2 *El límite superior del sesgo al cuadrado $\gamma_2(k)$ se aproxima $\beta^t\beta$*

Demostración

Como

$$\gamma_2(k) = \sum_{i=1}^{p}\frac{\alpha_i^2}{\left(1+\frac{\lambda_i}{k}\right)^2}$$

entonces

$$\lim_{k\to\infty}\gamma_2(k)=\lim_{k\to\infty}\sum_{i=1}^{p}\frac{\alpha_i^2}{\left(1+\frac{\lambda_i}{k}\right)^2}=\sum_{i=1}^{p}\alpha_i^2=\alpha^t\alpha=\beta^t\mathbf{P}^t\mathbf{P}\beta=\beta^t\beta$$

Corolario 3 *La derivada del sesgo al cuadrado $\gamma_2(k)$ se aprozima a cero cuando $k\to 0^+$*

Demostración

$$\begin{aligned}\gamma_2'(k) &= 2k\sum_{i=1}^{p}\frac{\alpha_i^2}{(\lambda_i+k)^2}-2k^2\sum_{i=1}^{p}\frac{\alpha_i^2}{(\lambda_i+k)^3}\\ &= 2k\sum_{i=1}^{p}\frac{\alpha_i^2(\lambda_i+k)}{(\lambda_i+k)^3}-2k^2\sum_{i=1}^{p}\frac{\alpha_i^2}{(\lambda_i+k)^3}\\ &= 2k\sum_{i=1}^{p}\frac{\lambda_i\alpha_i^2}{(\lambda_i+k)^3}\end{aligned}$$

por lo tanto

$$\lim_{k\to 0^+}\gamma_2'(k)=\lim_{k\to 0^+}2k\sum_{i=1}^{p}\frac{\lambda_i\alpha_i^2}{(\lambda_i+k)^3}\to 0$$

Teorema 32 *Siempre existe $k>0$ tal que*

$$ECM\left(\hat{\beta}_R\right)<ECM\left(\hat{\beta}\right)$$

Demostración

Como

$$\gamma_1(k) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} \quad y \quad \gamma_1(0) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} = \sigma^2 tr\left(\mathbf{X}'\mathbf{X}\right)$$

y

$$\gamma_2(k) = k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i+k)^2} \quad y \quad \gamma_2(0) = 0$$

además

$\gamma_1(k)$ es monótona decreciente y $\gamma_1'(k)$ es no positiva.

$\gamma_2(k)$ es monótona creciente y $\gamma_2'(k)$ es no negativa.

por lo tanto, para demostrar el teorema sólo falta encontrar un $k > 0$ tal que la derivada del

$$ECM\left(\hat{\beta}_R\right) < 0$$

como

$$ECM\left(\hat{\beta}_R\right) = \gamma_1(k) + \gamma_2(k)$$

entonces

$$\begin{aligned}\gamma_1'(k)+\gamma_2'(k) &= -2\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3}+2k\sum_{i=1}^{p}\frac{\lambda_i\alpha_i^2}{(\lambda_i+k)^3}<0\\ &= 2k\alpha'\alpha\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3}-2\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3}<0\\ &= (k\alpha'\alpha-\sigma^2)\left(2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3}\right)<0\end{aligned}$$

y como

$$2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3}>0 \quad entonces \quad k\alpha'\alpha-\sigma^2<0$$

por lo tanto,

$$k<\frac{\sigma^2}{\alpha'\alpha}\leq\frac{\sigma^2}{\alpha_{max}^2}$$

Por lo tanto existe $k>0$ para el cual se cumple que

$$ECM\left(\hat{\beta}_R\right)<ECM\left(\hat{\beta}\right)$$

Capítulo 5

Heterocedasticidad

Introducción

Uno de los supuestos del modelo de regresión lineal múltiple es que los residuos tienen varianza constante σ^2. Si no se cumple este supuesto, se presenta el problema de heterocedasticidad. En este capítulo analizarenios las consecuencias de existir heterocedasticidad en el modelo de regresión lineal múltiple.

5.1 Como Detectar la Heterocedasticidad

Para detectar la heterocedasticidad, sucede lo mismo que en el caso de la multicolinealidad, no existen reglas fijas y seguras para detectarla, sino solamente algunas reglas muy

generales. A continuación examinaremos algunos de los métodos para detectar heterocedasticidad.

1. Método gráfico

Como se hizo notar en el capítulo tres, el análisis de los residuos puede descubrir las violaciones de las suposiciones o las deficiencias del modelo. Por lo tanto para detectar si la varianza es o no constante, se grafican los residuos estandarizados contra los correspondientes valores estimados de la respuesta. Si la varianza de el error es constante entonces los residuos tenderán a encontrarse dentro de una banda horizontal centrada alrededor del cero, sin ninguna tendencia sistemática a ser positivos o negativos. Cualquier desviación significativa con respecto a este comportamiento indicará la existencia de la heterocedasticidad.

2. Prueba de Goldfeld-Quandt

Esta prueba esta basada en la prueba de hipótesis

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$$

contra

$$H_1 : \sigma_1^2 \leq \sigma_2^2 \leq \ldots \leq \sigma_n^2$$

Esta consiste en que la muestra está dividida en dos subconjuntos de observaciones donde es la varianza del error es diferente en cada subconjuto, pero es constante en cada uno.

Como la prueba de hipótesis alternativa para la prueba Goldfeld-Quandt es

$$H_1 : \sigma_1^2 \leq \sigma_2^2 \leq \ldots \leq \sigma_n^2$$

es decir, bajo H_1, las observaciones pueden ser ordenadas de acuerdo con un incremento en la varianza donde hay solamente dos varianzas bajo la hipótesis alternativa.

Para evaluar lo anterior explícitamente, Goldfeld-Quandt sugirieron los siguientes pasos:

a) Asumamos que la prueba de hipótesis H_1 es verdadera ordenando las observaciones de acuerdo al incremento de la varianzas del error.

b) Omitir r observaciones centrales, dividir las restantes observaciones $(n - r)$ en dos grupos, una usando los primeros $(n - r)/2$ y el otro usando las últimas $(n - r)/2$.

c) Ajustar, regresiones separadas de mínimos cuadrados ordinarios a las primeras $(n - r)/2$ y a las últimas $(n - r)/2$ observaciones y obtener la suma de los errores al cuadrado SEC_1 y SEC_2 respectivamente, donde SEC_1, representa la SEC de la regresión correspondiente a los valores más pequeños de X_i (el grupo de varianzas más pequeño) y SEC_2 de los valores más grandes de X_i (el grupo de varianza más grande). Estas SEC tienen individualmente

$$\frac{(n - r - 2k)}{2}$$

grados de libertad

d) Si se supone que e se distribuye como $N(0, \sigma^2 \mathbf{I})$ y si la prueba de hipótesis H_0 es verdadera entonces

$$\lambda = \frac{SEC_2}{SEC_1}$$

tiene una distribución F con

$$\frac{(n - r - 2k)}{2}$$

Si en una aplicación el valor calculado λ es mayor que el valor crítico de F al nivel de significancia que se haya escogido, podemos proceder a rechazar la hipótesis de homocedásticidad, pudiendo entonces decir que existe una gran probabilidad de que haya heterocedasticidad.

Antes de ilustrar la prueba mencionemos algo sobre la omisión de las observaciones centrales r. Estas observaciones se omiten buscando acentuar las diferencia existente entre el grupo de varianza pequeña (es decir, SEC_1) y el grupo de varianza grande (es decir, SEC_2). Sin embargo, la habilidad de la prueba de Goldfeld-Quandt para llevar a cabo lo anterior en forma exitosa depende de la manera como se escoja r. En el caso de que exista más de una variable X en el modelo, el ordenamiento de las observaciones, primer paso en la prueba, se puede hacer de acuerdo a cualquiera de ellas.

5.2 Medidas Remediales para la Heterocedasticidad

Una suposición en la estimación por mínimos cuadrados es que la varianza del error es constante. Si los residuos tienden a aumentar o disminuir conforme se incrementan los valores estimados de la respuesta, la varianza no puede considerarse como constante. El

remedio apropiado para esta situación es aplicar mínimos cuadrados generalizado.

5.3 El Método de Mínimos Cuadrados Generalizados

Consideremos el modelo de regresión lineal de rango completo $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ que cumple las siguientes condiciones:

i) $E(\mathbf{e}) = 0$

ii) $E(\mathbf{e}\mathbf{e}^t) = \sigma^2\mathbf{V}$

donde $\mathbf{V}$ es definida positiva. Por el teorema (13) tenemos que si V es una matriz definida positiva entonces existe una matriz simétrica K de tamaño $n \times n$ y $rango(\mathbf{K}) = n$ tal que $\mathbf{V} = \mathbf{K}^t\mathbf{K}$.

Si definimos

$$\mathbf{Z} = \left(\mathbf{K}^t\right)^{-1}\mathbf{Y}$$

$$\mathbf{A} = \left(\mathbf{K}^t\right)^{-1}\mathbf{X}$$

$$\eta = \left(\mathbf{K}^t\right)^{-1} \mathbf{e}$$

entonces el modelo $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ se transforma en el modelo

$$\mathbf{Z} = \mathbf{A}\beta + \eta$$

que cumple las siguientes condiciones:

$$i)\ E(\eta) = \left(\mathbf{K}^t\right)^{-1} E(\mathbf{e}) = 0$$

$$\begin{aligned} ii)\ E(\eta\eta^t) &= (\mathbf{K}^t)^{-1} E(\mathbf{e}\mathbf{e}^t) (\mathbf{K}^t)^{-1} \\ &= (\mathbf{K}^t)^{-1} \sigma^2 \mathbf{V} (\mathbf{K}^t)^{-1} \\ &= \sigma^2 (\mathbf{K}^t)^{-1} (\mathbf{K}^t\mathbf{K}) \mathbf{K}^{-1} \\ &= \sigma^2 \mathbf{I} \end{aligned}$$

Como el modelo de regresión lineal

$$\mathbf{Z} = \mathbf{A}\beta + \eta$$

cumple con las condiciones de mínimos cuadrados ordinarios entonces cumple todas las propiedades deducidas en la sección (3.5).

Con lo cual las ecuaciones normales son:

$$\left(\mathbf{A}'\mathbf{A}\right)\hat{\beta} = \mathbf{A}'\mathbf{Z}$$

y la solución es

$$\begin{aligned}\hat{\beta} &= (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Z} \\ &= (\mathbf{X}'\mathbf{K}^{-1}(\mathbf{K}')^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{K}^{-1}(\mathbf{K}')^{-1}\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y})\end{aligned}$$

donde $\hat{\beta}$ es el estimador de mínimos cuadrados generalizado de β.

Ahora consideremos el modelo de regresión lineal de rango completo $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$, donde $\mathbf{e}$ está distribuido como $N(0, \sigma^2\mathbf{V})$ y donde $\mathbf{V}$ es una matriz definida positiva.

Y definamos de la misma manera

$$\mathbf{Z} = (\mathbf{K}')^{-1}\mathbf{Y}, \quad \mathbf{A} = (\mathbf{K}')^{-1}\mathbf{X} \text{ y } \mathbf{A} = (\mathbf{K}')^{-1}\mathbf{X}$$

entonces el modelo $\mathbf{Y} = \mathbf{X}\beta + e$ se transforma en el modelo

$$\mathbf{Z} = \mathbf{A}\beta + \eta$$

Como $\mathbf{Z} = (\mathbf{K}')^{-1}\mathbf{Y}$ entonces por el teorema (13) $\mathbf{Z}$ se distribuye como $N(\mathbf{A}\beta, \sigma^2\mathbf{I})$ y como $\eta = \mathbf{Z} - \mathbf{A}\beta$ entonces η se distribuye como $N(0, \sigma^2\mathbf{I})$.

Como el modelo de regresión lineal

$$\mathbf{Z} = \mathbf{A}\beta + \eta$$

cumple con las condiciones de estimación máximo verosímil entonces cumple todas las propiedades deducidas en la sección (3.4).

Con lo cual los estimadores son:

$$\hat{\beta} = \left(\mathbf{A}^t\mathbf{A}\right)^{-1}\mathbf{A}^t\mathbf{Z} = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{Y}$$

$$\hat{\sigma}^2 = \frac{\mathbf{Z}^t\left(\mathbf{I} - (\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\right)\mathbf{Z}}{n-p}$$

$$\hat{\sigma}^2 = \frac{\mathbf{Z}^t\mathbf{Z} - \hat{\beta}^t\mathbf{A}^t\mathbf{Z}}{n-p}$$

$$\hat{\sigma}^2 = \frac{\mathbf{Y}^t\left(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}\right)\mathbf{Y}}{n-p}$$

$$Cov\left(\hat{\beta}\right) = \sigma^2\left(\mathbf{A}^t\mathbf{A}\right)^{-1} = \sigma^2\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$$

Capítulo 6

Conclusiones

Hemos visto que si se cumplen los supuestos del modelo clásico de regresión lineal múltiple, se obtienen estimaciones de los coeficientes de regresión $\hat{\beta}$, si $\mathbf{e}$ se distribuye como $N(0, \sigma^2\mathbf{I})$ se cumplen las siguientes propiedades: son consistentes, eficientes, insesgados,suficientes, completos, insesgados de varianza mínima uniforme, también obtenemos intervalos de confianza y pruebas de hipótesis. Si $\mathbf{e}$ es una variable aleatoria tal que: $E[\mathbf{e}] = 0$ y $Cov[\mathbf{e}] = \sigma^2$, son los mejores estimadores lineales insesgados.

Sin embargo, cuando no se cumple el supuesto de multicolinealidad, obtenemos estimaciones que son mayores en valor absoluto que los verdaderos valores, así como una gran inestabilidad, ya que con muestras diferentes se obtienen valores de los coeficientes considerablemente distintos.

La multicolinealidad es un problema de grado y no de tipo , es decir, el problema

no es la presencia o ausencia de la multicolinealidad sino entre sus diferentes grados y magnitudes. Por lo tanto, un buen método para detectar la multicolinealidad es aquel que mide la severidad de la multicolinealidad.

Cuando existe multicolinealidad severa, el método a escoger va a depender de la naturaleza de los datos. Una alternativa es utilizar el método de Regresión Ridge, ya que las ventajas que ofrece son:

a) La traza Ridge permite ver la sensibilidad del modelo. Al graficar los coeficientes contra k, nos damos cuenta de cuales coeficientes son más inestables que otros.

b) Su poder predictivo es bueno.

c) Se obtienen estimaciones de los coeficientes de regresión más estables.

Cuando no se cumple el supuesto de homocedasticidad. La heterocedasticidad no destruye el insesgamiento, ni la propiedad de consistencia, pero, los estimadores no poseen varianza mínima, no siendo, por lo tanto, eficientes. Por lo tanto, es importante detectar la existencia de la heterocedasticidad, ya que de lo contrario obtendremos intervalos de confianza amplios y pruebas de significancia pobres cuando $\mathbf{e}$ se distribuye como $N(\mathbf{0}, \sigma^2\mathbf{I})$. Sin embargo cuando se tiene solamente un valor $\mathbf{Y}$ correspondiente a un valor de $\mathbf{X}$ dado, no es fácil de detectarla, imposibilitando averiguar σ_i^2 con base en esa única observación, en este sentido existen algunos métodos informales de aproximación para detectarla. Algunas veces, cuando no se conoce σ_i^2, generalmente se plantean algunos supuestos sobre la naturaleza de los σ_i^2, transformandolos para hacer que $\mathbf{e}$ sean homocedásticos.

Bibliografía

[1] **Belsley A. David; Kuh Edwin; Welsh E. Roy**

Regression Diagnostics Identying influential data and sources of collinearity

John Wiley&Sons,Inc; New York 1980.

[2] **Hoerl, A. E. y R. W. Kennard** [1970] Ridge regression: Biased estimation for nonorthogonal problems, *tecnometrics*

[3] **Graybill A. Franklin**

Theory and application of the linear model

Duxbury Press, 1976.

[4] **Gujarati Damodar N.**

Econometría

Mc Graw-Hill, 1990.

[5] **Maddala G. S.**

Econometría

Mc Graw-Hill, 1985.