UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE AGRONOMIA
COMISIÓN DE ESTUDIOS DE POSTGRADO
DOCTORADO EN CIENCIAS AGRICOLAS

UNA CONTRIBUCIÓN AL ESTUDIO DE LA MULTICOLINEALIDAD EN MODELOS DE REGRESIÓN LINEAL MÚLTIPLE USANDO DISTRIBUCIONES DE CONTORNO ELÍPTICO

DANNY A. VILLEGAS R.

Maracay, Febrero de 2017

UNA CONTRIBUCIÓN AL ESTUDIO DE LA MULTICOLINEALIDAD EN MODELOS DE REGRESIÓN LINEAL MÚLTIPLE USANDO DISTRIBUCIONES DE CONTORNO ELÍPTICO

Tesis Doctoral para optar al grado de
Doctor en Ciencias Agrícolas

Autor: Danny A. Villegas R.
C.I: V-13.041.626
Tutor: Dr. Manuel E. Milla P.

Comité Consejero: Dra. Margarita Cobo.

Dr. Franklin Chacín.

UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE AGRONOMÍA
COMISIÓN DE ESTUDIOS DE POSTGRADO
DOCTORADO EN CIENCIAS AGRÍCOLAS

Doctorado en Ciencias Agrícolas

Comisión de Estudios de Postgrado

VEREDICTO

Quienes suscriben, miembros del jurado designado por el Consejo de la Facultad de Agronomía y el Consejo de Estudios de Postgrado de la Universidad Central de Venezuela, para examinar la **Tesis Doctoral** presentada por: **DANNY ALBERTO VILLEGAS RIVAS, Cédula de identidad Nº 13.041.626** bajo el título **"UNA CONTRIBUCIÓN AL ESTUDIO DE LA MULTICOLINEALIDAD EN MODELOS DE REGRESIÓN LINEAL MÚLTIPLE USANDO DISTRIBUCIONES DE CONTORNO ELÍPTICO"**, a fin de cumplir con el requisito legal para optar al grado académico de **DOCTOR EN CIENCIAS AGRÍCOLAS**, dejan constancia de lo siguiente:

1.- Leído como fue dicho trabajo por cada uno de los miembros del jurado, se fijó el día 08 de Marzo de 2018 a las 2:00 PM., para que el autor lo defendiera en forma pública, lo que éste hizo en la Sala de Computación 2, Coordinación de Tecnología de Información y Comunicaciones (CTIC) del Centro Operacional de Computación (CODEC) adscrito a la Facultad de Agronomía de la Universidad Central de Venezuela, ubicado en la calle Francisco Fernández Yépez, Sector. Niño Jesús, El Limón, Maracay, estado Aragua, mediante un resumen oral de su contenido, luego de lo cual **respondió satisfactoriamente** a las preguntas que le fueron formuladas por el jurado, todo ello conforme con lo dispuesto en el Reglamento de Estudios de Postgrado.

2.- Finalizada la defensa del trabajo, el jurado decidió **aprobarlo**, por considerar, sin hacerse solidario con las ideas expuestas por el autor, que se ajusta a lo dispuesto y exigido en el Reglamento de Estudios de Postgrado.

Para dar este veredicto, el jurado estimó que el trabajo examinado constituye un valioso aporte, ya que las metodologías propuestas basadas en distribuciones de contorno elíptico

se muestran como robustas herramientas para el estudio de la multicolinealidad, si se comparan con las comúnmente utilizadas, ya que no solo permiten verificar la presencia de multicolinealidad, sino también el origen de la misma. Se crean estadísticos de prueba apropiados e innovadores y se demuestran las ventajas del uso de transformaciones logarítmicas como alternativa de tratamiento de la multicolinealidad.

3.- El jurado por unanimidad decidió otorgar la calificación de **EXCELENTE** al presente trabajo por considerarlo de excepcional calidad, presentando un manuscrito bien estructurado con la rigurosidad del método científico y además fue expuesto con amplia solvencia desde el punto de vista académico.

En fe de lo cual se levanta la presente ACTA, a los 08 días del mes de Marzo de 2018. Conforme a lo dispuesto en el Reglamento de Estudios de Postgrado, actuó como Coordinador del Jurado el **Dr. Manuel Emilio Milla.**

TEODORO R. GARCÍA LEÓN
C.I. No. 7.182.927
UNIVERSIDAD DE CARABOBO
Jurado designado por el Consejo de la Facultad

CATALINA MARÍA RAMIS JAUME
C.I. No. 5.614.038
FACULTAD DE AGRONOMÍA-UCV
Jurado designado por el Consejo de la Facultad

TONNY GARCIA RUJANO
C.I. No. V- 11.370.396
UNIVERSIDAD CENTRO OCCIDENTAL LISANDRO ALVARADO
Jurado designado por el Consejo de Estudios de Postgrado

IVIS JOSEFINA QUIROZ
C.I. No. V- 9.539.882
UNIVERSIDAD NACIONAL EXPERIMENTAL DE LOS LLANOS "EZEQUIEL ZAMORA"
Jurado designado por el Consejo de Estudios de Postgrado

MANUEL EMILIO MILLA
C.I. No. V-7.557.851
Tutor-Coordinador del Jurado
UNIVERSIDAD NACIONAL "TORIBIO RODRÍGUEZ DE MENDOZA" DE AMAZONAS, CHACHAPOYAS, PERÚ.

HB/MM.-
08-03-2018

DEDICATORIA

Dedico esta obra a mi entorno elíptico, Daniel Ernesto, mi amado hijo, Yary, mi esposa y Lirio, mi madre. Y a todos los familiares y amigos que de alguna u otra manera forman parte de ese entorno elíptico…

AGRADECIMIENTOS

En primera instancia quiero agradecer al Profesor Manuel Milla, amigo, guía y Tutor, quien junto a la Profesora Margarita Cobo han sido los pilares fundamentales en todo mi proceso de formación en la ciencia estadística, y en particular en el seguimiento de este trabajo. Su valiosa colaboración en la revisión y orientación de esta Tesis fue sin lugar a dudas una contribución importante para poder alcanzar los objetivos planteados. A ellos dos mi agradecimiento, afecto y respeto infinito. Así mismo, quiero agradecer al Profesor Jesús Tapia, quien en una conversación que sostuvimos hace algunos años me dio a conocer en detalle algunos aspectos teóricos de las distribuciones de contornos elípticos, lo que fue de gran utilidad en la elaboración del proyecto inicial. Al mismo tiempo, quiero agradecer a la Dra. Haydeé Bolívar, Coordinadora del Doctorado en Ciencias Agrícolas por su apoyo en las gestiones para la presentación y defensa pública del mismo, así como a la Comisión de Estudios de Postgrado de FAGRO-UCV. Finalmente, y de manera muy especial quiero agradecer al Profesor e Ictiólogo Otto Castillo, amigo entrañable y compañero de trabajo en la UNELLEZ, a quien agradezco haberme incluido hace más de una década en su equipo de colaboradores, con lo cual pude iniciar formalmente el estudio de la multicolinealidad con base en el análisis morfométrico de peces de ríos llaneros que realiza el Profesor Otto con estudiantes del Programa de Ciencias del Agro y del Mar en nuestra casa de estudios.

INDICE DE CONTENIDO

Contenido Página

INDICE DE CUADROS

UNA CONTRIBUCIÓN AL ESTUDIO DE LA MULTICOLINEALIDAD EN MODELOS DE REGRESIÓN LINEAL MÚLTIPLE USANDO DISTRIBUCIONES DE CONTORNO ELÍPTICO

Autor: Danny A. Villegas R.
Tutor: Dr. Manuel E. Milla P.

RESUMEN

En el presente trabajo se propone una metodología para el estudio de la multicolinealidad en modelos lineales, la cual está constituida por dos estadísticos basados en distribuciones de contornos elípticos; el primero T_s, el cual sigue una distribución *t* de Student generalizada no centrada, y Z_s, cuya distribución se aproxima a la normal *N*(0,1). Así mismo, se plantea e uso del estadístico $W^*_{ecm(p)}$, basado en el error cuadrático medio del estimador de mínimos cuadrados ordinarios del modelo, el cual sigue una distribución Chi-cuadrado y permite identificar el origen de la multicolinealidad. Las metodologías se validaron mediante un estudio de simulación, a través de un modelo lineal con tres variables independientes (X_{1i}, X_{2i} *y* X_{3i}), donde $X_{3i} = k_2 X_{2i}$, k_2 es un número real conocido. Se consideraron cinco tamaños de muestra (*n=7, n=10, n=20, n=30, n=50 y n=100*) y cinco distribuciones teóricas (uniforme, exponencial, normal, log-normal y gamma), así como errores heterocedasticos y homocedasticos. Las frecuencias relativas de los estadísticos de prueba se confrontaron con los cuantiles de la distribución. Las metodologías fueron aplicadas a un conjunto de datos reales provenientes de dos ensayos agrícolas, el primero de un experimento con el cultivo de maíz y el segundo a un experimento con el pasto *Brachiaria brizantha* cv. Toledo. Los resultados evidenciaron que ambas metodologías están condicionadas por la distribución de las variables independientes y por el tamaño de muestra, en particular, la distribución *t* generalizada no centrada y la normal. La distribución del estadístico T_s se aproximó bien para tamaños de muestra $n \leq 20$ y distribuciones exponencial, gamma y uniforme, así como para $n < 10$ y una distribución normal y log-normal, mientras que el estadístico Z_s se aproximó bien para tamaños de muestra $n > 20$ y una distribución normal y log-normal. Para la distribución uniforme este estadístico se aproximó bien para tamaños de muestra $n \geq 30$, mientras que para la distribución exponencial y gamma se requieren tamaños de muestra $n \geq 50$. Se verificó la propiedad que tienen las distribuciones *t* generalizada no centrada y normal de ser invariantes bajo leyes elípticas. El estadístico de prueba $W^*_{ecm(p)}$ verificó la propiedad de consistencia cuando las variables independientes X^*_{ij} siguen una distribución log-normal, y se demostró que este procedimiento de prueba se hace más robusto conforme se incrementan tanto el tamaño de muestra como el grado de multicolinealidad en el modelo. Finalmente, las metodologías propuestas en este trabajo se muestran como poderosas herramientas para el estudio de la multicolinealidad, si se comparan con las comúnmente utilizadas, ya que no solo permiten verificar la presencia de multicolinealidad, sino también el origen de la misma. Así mismo, se evidenció

las ventajas del uso de transformaciones logarítmicas como alternativa de tratamiento de la multicolinealidad.

Palabras clave: Variables colineales, elípticas, error cuadrático medio, transformaciones logarítmicas.

A CONTRIBUTION TO THE STUDY OF MULTICOLLINEARITY IN MULTIPLE LINEAR REGRESSION MODELS USING ELLIPTICAL CONTOUR DISTRIBUTIONS

ABSTRAC

In this paper we propose a methodology for the study of multicollinearity in linear models, which is constituted by the statistics based on distributions of elliptic contours; The first T_s, the first, which remains a t student generalized non-centered distribution, and Z_s, whose distribution approaches the normal *N* (0,1). In the same way, the statistical $W^*_{ecm(p)}$, based on the mean squared error of the ordinary least squares estimator of the model, is proposed and used, which remains a Chi-square distribution and allows the identification of the origin of multicollinearity . The methodologies are validated through a simulation study and a linear model with three independent variables ($X_{1i}, X_{2i}\, y\, X_{3i}$), where $X_{3i} = k_2 X_{2i}$, k_2 is a known real number. Five sample sizes (*n=7, n=10, n=20, n=30, n=50* and *n=100*) and five theoretical distributions (uniform, exponential, normal, lognormal and gamma) as well as homocedastic and heterocedastic errors. The relative frequencies of the test statistic are compared with the quantiles of the distribution. The methodologies were applied to a real data set of two agricultural trials, the first of an experiment with the maize crop and the second to an experiment with the grass *Brachiaria brizantha* cv. Toledo. The results showed that both methodologies are conditioned by the distribution of the independent variables and by the sample size, in particular the t student generalized non-centered and normal distribution. The distribution of the T_s statistic approximated well for sample sizes n $\leq$ 20 and an exponential, gamma and uniform distributions, as well as for n<10 and a normal and log-normal distributions, whereas the Z_s statistic approached well for sample sizes n>20 and a normal and log-normal distribution. For the uniform distribution this statistic was approached well for sample sizes n $\geq$ 30, while for an exponential and gamma distribution a sample size n $\geq$ 50 it is required. The property that has the *t* generalized distributions non central and normal to be invariant under elliptical laws was verified. The test statistic $W^*_{ecm(p)}$ verifies the consistency property when the independent variables X^*_{ij} follows a log-normal distribution, and it proves that this test procedure becomes more robust as both the sample size as the degree of multicollinearity in the model. Finally, the methodologies proposed in this work are powerful tools for the study of multicollinearity, when compared to those commonly used, since there is not only a verification of the presence of multicollinearity, but also the origin of the multicollinearity. Likewise, the

advantages of the use of logarithmic transformations as an alternative for the treatment of multicollinearity were evidenced.

Key words: Collinear, elliptic, mean square error, logarithmic transformations.

INTRODUCCIÓN

Multicolinealidad es el término generalmente utilizado para referirse a la existencia de relaciones lineales o cuasilineales entre las variables regresoras en un modelo de regresión múltiple, lo que indica que parte sustancial de la información en una o más de estas variables es redundante. En tal sentido, Mandell (1982) señala que una de las principales dificultades en el uso de estimaciones mínimo cuadráticas es la presencia de este problema que, aun cuando no afecta la capacidad predictiva del modelo, representa un problema grave si su propósito fundamental es evaluar la contribución individual de las variables explicativas. Esto es debido a que en presencia de multicolinealidad, los coeficientes *bj* tienden a ser inestables, es decir, sus errores estándar presentan magnitudes indebidamente grandes. Esta falta de precisión afecta los contrastes parciales diseñados para evaluar la contribución individual de cada variable explicativa, corriéndose un alto riesgo de no encontrar significación en variables que realmente la tengan. Por su parte, Jackson (1991) subraya además que, bajo condiciones de colinealidad resulta imposible distinguir los efectos individuales de cada variable predictora, debido a que la fuerza de la correlación entre ellas produce relaciones lineales de similar magnitud entre los coeficientes. Una de las hipótesis del modelo de regresión lineal múltiple establece que no existe relación lineal exacta entre los regresores, o, en otras palabras, establece que no existe multicolinealidad en el modelo. Esta hipótesis es necesaria para el cálculo del vector de estimadores de mínimos cuadrados ya que en caso contrario la matriz $\boldsymbol{X'X}$ será no singular. La multicolinealidad perfecta no se suele presentar en la práctica, salvo que se diseñe mal el modelo. En cambio, sí es frecuente que entre los regresores exista una relación aproximadamente lineal, en cuyo caso los estimadores que se obtengan serán en general poco precisos, aunque siguen conservando la propiedad de lineales, insesgados y de varianza mínima. La relación entre regresores hace que sea difícil cuantificar con precisión el efecto que cada regresor ejerce sobre la respuesta, lo que determina que las varianzas de los estimadores sean elevadas. Es por ello que, cuando se presenta una relación

aproximadamente lineal entre los regresores, se dice que hay presencia de multicolinealidad, aunque esta no es total. Chacín y Meneses (1984) señalan cuatro (4) causas posibles de multicolinealidad, las cuales serían: (1) El método de recolección de datos, (2) restricción en el modelo o en la población, (3) definición del modelo, (4) modelos sobre-definidos. En tal sentido, la multicolinealidad hace referencia, en concreto, a la existencia de relaciones aproximadamente lineales entre los regresores del modelo, cuando los estimadores obtenidos y la precisión de éstos se ven seriamente afectados. Así mismo, las inferencias que se hacen sobre la población, dependen de la exactitud de la estimación del valor del parámetro, es por ello que, elevados errores estándar (bajo pruebas de hipótesis) y coeficientes inestables (con signos distintos a los esperados o verdaderos) proporcionan una señal de alerta, es decir, las estimaciones no son confiables, esto implica también un incremento en la probabilidad de cometer error tipo II y una disminución de la potencia de las pruebas de hipótesis estadística. De igual forma, este problema está presente en diferentes grados y no tiene fácil solución, ya que en definitiva se trata de pedirle a la muestra de datos más información de la que posee, por lo que se hace necesario un análisis a priori del mismo al proponer un modelo de regresión lineal múltiple, y en la medida de lo posible corregir el problema de multicolinealidad. Para corregir este problema sólo cabe actuar en alguno de los siguientes sentidos: (1) Eliminar variables regresoras, con lo que se reduce el número de parámetros a estimar, (2) Incluir información externa a los datos originales. Si se opta por el primero de ellos, se trata de suprimir, o bien ciertas variables que se encuentran altamente correlacionadas, o bien algunas combinaciones lineales mediante el análisis de componentes principales aplicado a la regresión. La segunda alternativa conduce a trabajar con estimadores contraídos o bayesianos. En ambas opciones se sustituyen los estimadores mínimo cuadráticos de los coeficientes de regresión del modelo por estimadores sesgados. Estos procedimientos forman parte de la regresión sesgada, no lineal, pero que sigue cumpliendo el supuesto de los mínimos cuadrados. Además, estos estimadores, a pesar de ser sesgados, tienen un error cuadrático medio mucho menor, que es lo que se pretende para corregir la multicolinealidad.

De esta manera, dado que la multicolinealidad es un problema muestral, y está asociado a la configuración concreta de la matriz ***X***, son muy escasos los contrastes estadísticos, propiamente dichos, que sean aplicables para su detección, salvo el contraste propuesto por Farrar y Glauber (1967) para identificar la posible existencia de multicolinealidad entre las variables de un modelo lineal general. No obstante, se han desarrollado numerosas reglas prácticas que tratan de determinar en qué medida la multicolinealidad afecta gravemente a la estimación y contraste de un modelo. Estas reglas no son siempre fiables, siendo en algunos casos muy discutibles, entre las que destacan los indicadores considerados por Neter *et al.*, (1990), entre ellos el factor de inflación de la varianza (*VIF*) y la Tolerancia (*T*), calculados sobre cada columna de ***X***. No obstante**,** la utilización de los coeficientes *T* y *VIF* para detectar la presencia de la multicolinealidad ha recibido múltiples críticas, porque la conclusión obtenida con estos valores no siempre recoge adecuadamente la información y problema de los datos, así como es de resaltar el hecho de que las varianzas de los estimadores depende del *VIF*, σ^2 y $\sum(X_{ji} - \overline{X}_j)^2$, razón por la cual un *VIF* alto, no es condición suficiente ni necesaria para que dichas varianzas sean elevadas, ya que es posible que σ^2 sea pequeño o $\sum(X_{ji} - \overline{X}_j)^2$ grande y se compensen. De igual forma, otros procedimientos para detectar multicolinealidad reportados en la literatura incluyen la inspección de los autovalores y el análisis del "eigensistema", entre los cuales se encuentra el número de condición ***k(X)*** y la descomposición de las proporciones de varianza de los estimadores del modelo (π), siendo estos dos últimos considerados como las mejores alternativas para evaluar las dependencias lineales entre los regresores del modelo. Estos métodos, planteados inicialmente por Kendall (1957) y desarrollados posteriormente por Belsley *et al.,* (1980), y Belsley (1982) se enfocan en el abordaje de la multicolinealidad con base en los autovalores de la matriz ***X'X***.

Por otro lado, además de los procedimientos analíticos mencionados anteriormente, existen métodos gráficos de detección de la multicolinealidad,

como es el caso del *h-plot* de la inversa de la matriz de correlaciones introducido por Cornsten y Gabriel (1976) y considerado por Ramírez *et al.,* (2005), para obtener representaciones gráficas de la información contenida en una matriz de varianzas y covarianzas S_{pxp} de rango *r*, sobre espacios reducidos de baja dimensión. Además de los procedimientos antes señalados, existen otras aproximaciones informales tales como; un R^2 elevado, altas correlaciones entre parejas de regresores, examen de las correlaciones parciales y regresiones auxiliares. Así pues, como se ha observado, existen numerosos procedimientos desarrollados, en su mayoría, reglas prácticas para detectar multicolinealidad, en algunos casos, muchas de estas poco confiables, los cuales pueden ser enfocados desde dos puntos de vista, aquellos métodos basados en la correlación entre variables regresoras y los basados en la estructura de la matriz ***X'X,*** por lo que este trabajo persigue como fin proponer en primera instancia un procedimiento de prueba de hipótesis, cuyo aporte principal implica una alternativa metodológica con un enfoque del supuesto de multicolinealidad distinto a los antes señalados, el cual, además de considerar el potencial efecto de la dependencia lineal entre los predictores del modelo, en términos de la magnitud del error típico con relación al tamaño del coeficiente, partiendo de las ideas de Belsley *et al.,* (1980) y de los resultados obtenidos por Villegas *et al*., (2013), consiste en un procedimiento de prueba de hipótesis estructural de esta suposición del modelo de regresión lineal múltiple (multicolinealidad), incorporando un estadístico de prueba *Ts*, cuya distribución muestral depende del tamaño de muestra y principalmente de la ley elíptica de la cual fue obtenida la muestra, y además se constituye en una distribución de contornos elípticos, o simplemente, distribución elíptica, que permite detectar multicolinealidad de grado en modelos lineales. Así mismo, en esta investigación se plantea desarrollar un procedimiento de prueba con base en el error cuadrático medio (*ECM*) para identificar variables colineales en modelos de regresión lineal múltiple que permitan un análisis adicional en función del objetivo para el cual se plantea la construcción de un modelo lineal.

OBJETIVO GENERAL

Proponer una metodología para el diagnóstico de multicolinealidad basada en distribuciones de contorno elíptico y un procedimiento basado en el error cuadrático medio del estimador de mínimos cuadrados ordinarios para identificar variables colineales en modelos de regresión lineal múltiple.

OBJETIVOS ESPECÍFICOS

1. Estudiar metodologías existentes para la detección de multicolinealidad.
2. Analizar las debilidades y fortalezas de las metodologías existentes para la detección de multicolinealidad.
3. Generar una alternativa metodológica para la detección de multicolinealidad de grado basada en distribuciones elípticas en modelos de regresión lineal múltiple.
4. Desarrollar un procedimiento basado en el error cuadrático medio del estimador de mínimos cuadrados ordinarios para identificar variables colineales.
5. Comparar la metodología generada con las metodologías existentes en términos de sus debilidades y fortalezas.
6. Validar las alternativas metodológicas señaladas anteriormente a través de un estudio de simulación.
7. Ilustrar las metodologías generadas a través de su aplicación a un conjunto de datos reales provenientes de ensayos en ciencias agrícolas.

CAPÍTULO 1

Multicolinealidad, causas y consecuencias

1.1. Revisión de literatura sobre la multicolinealidad, causas y consecuencias sobre el modelo de regresión.

El análisis de regresión fue considerado por primera vez en el siglo XVIII en temas concernientes a la navegación con base en la astronomía. Por su parte, Gauss afirma haber desarrollado el método y en 1809 demostró que los mínimos cuadrados era la solución óptima cuando los errores están normalmente distribuidos. Inicialmente esta metodología era utilizada casi exclusivamente en las ciencias físicas hasta buena parte del siglo XIX. Francis Galton acuñó el término "regresión" en 1875 haciendo referencia a la ecuación de regresión simple. Galton usó esa ecuación para explicar el fenómeno de que los hijos de padres altos tienden a ser altos, pero no tan altos como sus padres, mientras que hijos de padres pequeños tienden a ser pequeños, pero no tan pequeños como ellos. Este efecto es llamado "efecto de regresión".

La multicolinealidad es un problema del análisis de regresión, que surge cuando las variables regresoras del modelo lineal general muestran dependencia lineal entre ellas o algunas de estas. Este es un problema complejo, porque en cualquier regresión las variables explicativas o independientes van a presentar algún grado de correlación. Matemáticamente, existe multicolinealidad cuando se presentan problemas a la hora de invertir la matriz ***X'X***. De hecho, si el determinante de esta matriz es cercano a cero (0) se dice que hay multicolinealidad no estricta o de grado. Más aun, si el determinante antes mencionado es igual a cero (0), se dice que hay multicolinealidad estricta o perfecta. En este último caso al menos una variable es combinación lineal de otra(s), la matriz ***X*** no es de rango completo, en consecuencia, la matriz ***X'X*** no tiene inversa única y el sistema de ecuaciones normales $(X'X)\hat{\beta}_i = X'Y$ tiene infinitas soluciones. La multicolinealidad estricta o perfecta, es un problema de identificación en el siguiente sentido. Si dada la

especificación del modelo hay un problema de multicolinealidad perfecta, distintos valores de los parámetros, generan el mismo valor medio de la variable respuesta, $E(Y) = X\beta$. Por tanto, dada una muestra (*X*,*Y*), no se pueden identificar aquellos valores de los parámetros que la han generado porque la función criterio que se minimiza $[Y - E(Y)]'[Y - E(Y)]$ no discrimina entre distintos valores de β. Por el contrario, si el problema no es de multicolinealidad perfecta, sino de un alto grado de colinealidad entre variables regresoras, los parámetros del modelo de regresión lineal múltiple se pueden estimar en forma única por mínimos cuadrados ordinarios, y los estimadores serán lineales, insesgados y de varianza mínima.

La literatura sobre el tema de la multicolinealidad ha sido ampliamente detallada por diferentes autores, entre los que destacan Neter y Wasserman (1974), quienes señalan que la multicolinealidad existe cuando las variables regresoras están correlacionadas entre ellas mismas, pero algunas veces este término es aplicado solamente en los casos donde la correlación entre las variables es muy alta, o casi perfecta. De igual forma, Chacín y Meneses (1984), refieren que la interpretación y el uso de los modelos de regresión múltiple dependen implícita o explícitamente de las estimaciones de los coeficientes de regresión individuales. Cuando no existen relaciones lineales entre las variables regresoras se dice que hay ortogonalidad. Lamentablemente en muchas aplicaciones del análisis de regresión, las variables regresoras no son ortogonales y aunque algunas veces los problemas de no ortogonalidad no son graves, en muchos casos las variables regresoras se encuentran relacionadas linealmente en forma estrecha y en tal situación, inferencias basadas en los modelos de regresión pueden estar completamente erradas. La multicolinealidad se refiere específicamente a la interdependencia que existe entre las variables regresoras y que tiene efecto directo sobre las estimaciones y varianzas de los parámetros. En tal sentido, la multicolinealidad es un problema que no tiene fácil solución, ya que en definitiva se trata de pedirle a la muestra de datos más información de la que posee (Peña, 1987). Por su parte,

Velilla (1987) señala que la multicolinealidad es una fuente clásica de distorsión en el problema de mínimos cuadrados lineales. Así mismo, López (1998) afirma que la multicolinealidad es un problema del análisis de regresión que consiste en que las variables regresoras del modelo están relacionadas, constituyendo entre sí una combinación lineal. Este hecho tiene consecuencias en el modelo de regresión, es decir, la influencia de cada una de las variables explicativas sobre la respuesta no puede distinguirse al quedar solapadas unas con otras, no se consigue una explicación del fenómeno en cuestión, en consecuencia los pronósticos no son nada confiables, puesto que otra combinación de variables en el modelo cambiando el orden, produciría predicciones de la respuesta contradictorias.

A su vez, Callaghan y Chen (2008) indican que frecuentemente en la investigación científica se establecen modelos con el fin de interpretar las estimaciones de los coeficientes de tales modelos como medidas de las características verdaderas de la población. Sin embargo, cuando la multicolinealidad está presente, el valor de los coeficientes estimados a partir de la muestra puede diferir notablemente del valor real de la población. Por desgracia, para los científicos, la multicolinealidad es un estado normal en la naturaleza; a menudo, las variables independientes están relacionadas linealmente entre ellas. Además, estos autores señalan que la multicolinealidad se presenta en diferentes grados.

Habshah *et al*., (2009) señalan que la multicolinealidad o no ortogonalidad es una dependencia casi lineal entre dos o más variables explicativas. Su presencia provoca dificultades para hacer inferencias o estimaciones, así como la selección de un conjunto adecuado de variables. Desafortunadamente, la mayoría de las variables explicativas en un modelo de regresión no son ortogonales, en tales casos, las inferencias basadas en las estimaciones de los parámetros del modelo serían inválidas.

Con relación a las causas de multicolinealidad Montgomery *et al*., (2001) señalan que la multicolinealidad puede ser causada por el método de colección de datos empleado, restricciones en el modelo o en la población que se muestrea. Por su parte, Kamruzzaman e Imon (2002) identificaron una nueva posible causa de multicolinealidad, la cual consiste en puntos influyentes, observaciones desviadas no solo de la línea de regresión misma como de los demás datos, sino también alejadas de la mayoría de las variables explicativas en el conjunto de datos. De la misma manera, Moller *et al*., (2005) y Hocking y Pendelton (1983) estudiaron esas nuevas fuentes de multicolinealidad partiendo de conjuntos de datos reales y simulados. De esta manera, utilizando la matriz de correlación, estos autores demostraron que la presencia de múltiples puntos influyentes iguales o distintos causa multicolinealidad severa. Así mismo, comprobaron que puntos influyentes considerablemente distintos incrementan los problemas de multicolinealidad en comparación a los casos donde estos puntos son similares.

En ese sentido, Hadi (1988) señala que esas nuevas fuentes de multicolinealidad puede ser colinealidad influenciada por las observaciones. De igual forma, el mismo autor señala que la multicolinealidad influenciada por las observaciones está generalmente asociada a puntos con una alta influencia, sin embargo, no todos estos son causantes de multicolinealidad. Por su parte, Sengupta y Bhimasankaram (1997) señalan que la debilidad de esta medida es la falta de simetría, la cual se debe al cambio aditivo del índice de condición de la matriz $\boldsymbol{X}$. Estos mismos autores afirman que se ha prestado poca atención al papel que juegan los casos individuales en multicolinealidad.

Los efectos adversos de la multicolinealidad han sido ampliamente estudiados, en tal sentido, Hoerl y Kennard (1970) demostraron que esta puede afectar el cuadrado de la distancia entre el estimador de mínimos cuadrados $\hat{\beta}_i$ y el parámetro estimado β_i. Por su parte, Webster *et al.,* (1974) afirman que si hay multicolinealidad fuerte entre una variable explicativa X_j y alguna de las demás

variables del modelo, el coeficiente de determinación múltiple de la regresión de X_j sobre el resto de las variables explicativas tendrá un valor cercano a uno (1), provocando que la varianza de $\hat{\beta}_i$ sea muy elevada. En ese mismo orden de ideas, de acuerdo a Cruz y Ragazzi (1994), la multicolinealidad también puede afectar el coeficiente de determinación múltiple del modelo (R^2). Por otro lado, Yu (1998) desarrolla un programa multimedia para ilustrar visualmente la forma como un modelo de regresión puede colapsar cuando las variables predictoras están intercorrelacionadas. Chacín (1998) señala que la presencia de multicolinealidad tiene efecto potencial sobre los estimadores mínimos cuadrados de los coeficientes de regresión, además que la estrecha multicolinealidad resulta en elevadas varianzas y covarianzas de estos estimadores.

1.2. Revisión de literatura sobre la detección de multicolinealidad.

Numerosos métodos han sido desarrollados con el objeto de detectar la posible existencia de multicolinealidad y sus efectos anómalos sobre un modelo de regresión. Una primera aproximación al problema plantea analizar la matriz de correlaciones R, procedimiento útil pero que no capta el fenómeno en toda su intensidad, puesto que estudia las relaciones entre las variables dos a dos, obviando las relaciones de estas con las otras variables predictoras. Otras propuestas alternativas se basan en el coeficiente de determinación múltiple de cada variable *Xj* con las restantes, y en los coeficientes de correlación parcial de las variables *Xj* y X_k, controlado por los efectos lineales de las restantes.

Por su parte, Gleason y Staelin (1975) proponen un índice basado en los autovalores de la matriz de correlaciones que toma el valor 0 cuando las variables son independientes (R = I) y el valor 1 cuando las variables están perfectamente correlacionadas (R = J). Raveh (1985) discute la importancia de ciertos elementos fuera de la diagonal principal de la inversa de la matriz de correlaciones para detectar predictores importantes en un análisis de regresión y como criterio para evaluar los supuestos requeridos para aplicar un análisis de factores. Whitakker

(1990) resalta la utilidad de la inversa de la matriz de correlaciones para establecer relaciones de dependencia entre variables y propone su representación gráfica mediante los denominados grafos de independencia condicional.

Marquardt (1970) propuso el factor de inflación de varianza (*VIF*) como una herramienta para la detección de multicolinealidad en modelos de regresión lineal múltiple. En ese sentido, Mandell (1982) demuestra que el error estándar del *j-ésimo* coeficiente de regresión puede expresarse como el producto del error estándar residual de la regresión por el factor de inflación de varianza (*VIF*), ampliamente utilizado para detectar multicolinealidad; en particular demuestra que el *VIF* se ve severamente afectado por los autovalores más pequeños de la matriz *R*. Este coeficiente mide el incremento que se produce en la varianza de b_j respecto del valor mínimo que se alcanzaría en ausencia total de colinealidad de la correspondiente variable X_j respecto de las restantes variables predictoras (Glantz y Slinker 2001). Adicionalmente, la expresión *VIF(j) - 1* coincide, excepto por los grados de libertad, con el estadístico que contrasta la bondad del ajuste de la regresión de X_j como función de las restantes variables explicativas. Belsley (1991) refiere las ventajas y debilidades de los *VIF* como medida para el diagnóstico de la colinealidad. Por su parte, Wetherill (1986) afirma que los *VIFs* no indican directamente el número de variables colineales ni cuáles de ellas están asociadas entre sí. Sin embargo, Neter *et al.,* (1983) señala que un *VIF* mayor que 10 es a menudo tratado como un indicativo de multicolinealidad. En ese mismo orden de ideas, Neter, Wasserman & Kutner (1990) consideran una serie de indicadores para analizar el grado de multicolinealidad entre los regresores de un modelo, como por ejemplo los llamados Tolerancia (*TOL*) y Factor de Inflación de la Varianza (*VIF*).

Villegas y García (2009) reportan la debilidad del *VIF* frente a otras alternativas (número de condición y descomposición de las proporciones de varianza) para detectar multicolinealidad en modelos de regresión lineal múltiple.

Por su parte, Farrar y Glauber (1967) propusieron un estadístico (*FG*) para identificar la posible existencia de multicolinealidad entre las variables explicativas en un modelo lineal general. Este estadístico considera el coeficiente de determinación (R_i^2) sobre la regresión lineal de la *i-ésima* variable regresora (X_i) en relación a las otras variables explicativas, el número de parámetros estimados (k) en el modelo y el tamaño de muestra (n). El estadístico (*FG*) sigue una distribución $F_{k-1,n-1}$.

En particular, Kendall (1957) enfoca el abordaje práctico de la multicolinealidad a través de un procedimiento diseñado en función de los autovalores y autovectores de la matriz de correlaciones.

Mason *et al.,* (1975) proponen el índice conocido como número de condición (*K),* definido como el cociente entre el mayor y el menor autovalor de la matriz *R*.

Por otra parte, un método práctico y útil para detectar multicolinealidad es el llamado número de condición (*CN*) de la matriz ***X***, el cual puede ser obtenido del análisis de la estructura de autovalores de la matriz de varianzas y covarianzas de ***X***. Por lo tanto, la matriz ***X'X*** puede ser factorizada en p autovalores y autovectores ordenados. El primer autovector es una combinación lineal de las variables independientes que explican la máxima varianza posible. Este autovector está asociado con el mayor autovalor. Subsecuentemente, los autovectores maximizan la varianza residual y están asociados a los autovalores menores. Así pues, un autovalor igual a cero (0) indica multicolinealidad perfecta. En ese sentido, Belsley *et al*., (1980) propusieron una aproximación similar para detectar multicolinealidad, con lo cual sugirieron dos medidas; la descomposición del valor singular (*SVD*) de la matriz de datos ***X*** con su índice de condición (*CN*) asociado y la descomposición de las varianzas de los coeficientes estimados usando la *SVD*, conocida como las proporciones de la descomposición de las varianzas.

Así mismo, Belsley (1991) realizó algunos estudios para comprobar si algunos métodos para detección de multicolinealidad podían identificar o no la existencia de este fenómeno en modelos de regresión lineal múltiple, así como las variables colineales. Su objetivo era proporcionar información sobre que tan alto debe ser el número de condición para sugerir un problema de multicolinealidad en el modelo.

En ese mismo orden de ideas, según afirman Judge *et al.,* (1985), el número de condición $k(\boldsymbol{X})$ y la descomposición de las proporciones de varianza de los estimadores del modelo son los procedimientos más adecuado entre los actualmente disponibles para detectar multicolinealidad.

A través de estudios de simulación Belsley *et al*., (2004) derivaron los valores límites del número de condición (*k*). Estos autores sugieren que esos valores están entre 10 y 30, lo que indica un grado de multicolinealidad moderado y severo, respectivamente.

Villegas *et al.,* (2011) en un estudio de simulación considerando tres (3) distribuciones teóricas continuas (uniforme, normal y exponencial) en las variables explicativas y cuatro (4) tamaños de muestra (*n=10, n=20, n=30 y n=50*) en un modelo con tres variables regresoras (X_1, X_2 y X_3), estableciendo a X_1 como una combinación lineal de X_2 y X_3, sugieren un potencial efecto de la distribución teórica de las variables regresoras y del tamaño de muestra (*n*) sobre el índice de condición (*K*).

A diferencia de los métodos antes mencionados, Cornsten y Gabriel (1976) introdujeron el uso de los h-plots como un procedimiento para obtener representaciones gráficas de la información contenida en una matriz de varianzas y covarianzas *Spxp* de rango *r*, sobre espacios reducidos de baja dimensión. Las ideas expuestas por estos autores son consideradas por Ramírez *et al.,* (2005), quienes proponen un dispositivo gráfico para obtener una representación aproximada de las relaciones de dependencia lineal que se producen entre un

conjunto de variables. Específicamente en dicha representación se visualizan los *VIF* para cada variable y los coeficientes de correlación parcial. En ese trabajo, los autores indicaron un aumento notable en el porcentaje de la correcta clasificación al eliminar del conjunto original las variables colineales.

1.3. Revisión de literatura sobre la distribución *t* de Student.

En probabilidad y estadística, la distribución *t* de Student centrada o simplemente distribución *t*, es una distribución de probabilidad continua que surge al estimar la media de una población distribuida normalmente en situaciones donde la muestra es pequeña y la desviación estándar poblacional es desconocida. Desempeña un papel en una serie de análisis estadísticos frecuentemente utilizados, incluidos la prueba de *t* de Student para hacer inferencias sobre la media de una y dos poblaciones independientes y normalmente distribuidas, intervalos de confianza para los casos antes mencionados y en el contexto del análisis de regresión. Esta distribución también es utilizada en el análisis bayesiano de datos provenientes de una población con distribución normal. La distribución *t* de Student al principio fue derivada como una distribución posterior por Helmert (1875, 1876a y 1876b). Una derivación de la distribución *t* de Student fue publicada por Gosset (1908), mientras trabajaba en la destilería Guinness en Dublín. La prueba de *t* y la teoría asociada a la misma se dio a conocer a través del trabajo de Fisher (1925), quien la denominó distribución *t* de Student. Se ha demostrado que una serie de estadísticos tiene una distribución *t* de Student para muestras de moderado tamaño bajo hipótesis nulas que son de interés. Por ejemplo, la distribución del coeficiente de correlación por rangos de Spearman se aproxima bastante bien a la distribución *t* de Student para muestras por encima de veinte (20).

Una generalización de la conocida distribución *t* de Student, es la distribución *t* de Student no centrada, usando un parámetro de no centralidad. Al igual que la distribución *t* de Student centrada, la distribución *t* de Student no centrada es

utilizada principalmente en inferencia estadística, aunque también puede ser usada en el modelaje de datos robustos, así como también en el análisis de robustez.

1.4. Revisión de literatura sobre las distribuciones elípticas.

La distribución normal ha servido durante mucho tiempo como modelo estándar para observaciones provenientes de diferentes áreas, siendo ésta el objetivo central del análisis paramétrico clásico. Sin embargo, a través de los años los estadísticos han tratado de extender la teoría de este análisis a casos más generales y, por lo tanto, con mayor cobertura. Últimamente se ha planteado una clase de distribuciones cuyos contornos de sus densidades tienen la misma forma elíptica de la Normal, pero además contienen distribuciones de colas más y menos pesadas que las de ésta. Esta clase de distribuciones simétricas se denomina distribuciones Contornos Elípticos o simplemente distribuciones Elípticas. La utilización de distribuciones Elípticas en problemas que solo han sido resueltos con la distribución Normal, permite generalizar la teoría hasta ahora planteada. Algunos de estos problemas ya han sido resueltos para las distribuciones Elípticas no singulares, aunque no así para el caso singular. Específicamente, cuando se muestrea desde una población Elíptica, el problema de las distribuciones de muestreo tradicionalmente usadas, esto es, las distribuciones chi-cuadrado, *t* y *F*, que bajo distribuciones Elípticas se denominan distribuciones chi-cuadrado, *t* y *F*, ha sido parcialmente resuelto.

Bien se sabe que las medidas de dispersión se utilizan para evaluar la variabilidad de un conjunto de datos, siendo la desviación estándar la más utilizada; sin embargo, ésta entrega poca información del conjunto si es interpretada en forma aislada, sólo cuando se la relaciona con la media su interpretación tiene mayor sentido. Por esta razón el coeficiente de variación, que relaciona a ambas medidas, es sumamente útil. Este coeficiente mide la dispersión u homogeneidad de un conjunto de datos asociados a una variable aleatoria y es una medida de variabilidad relativa, es decir, es adimensional, pues representa a la desviación

estándar por una unidad de media y resulta de particular interés cuando se desea comparar la variabilidad entre grupos cuyas medias y varianzas difieren.

Existen en la actualidad excelentes libros y artículos que recogen las investigaciones hechas hasta la fecha acerca de la distribuciones Elípticas, como lo son los libros, entre los que destacan los publicados por Anderson y Fang (1990), Fang y Zhang (1990), Fang *et al.*, (1990), Gupta y Varga (1993) y Rao (1993), y diversas publicaciones, entre las que se encuentran los trabajos de Kelker (1970), Chu (1973), Dawid (1977), Kariya y Eaton (1977), Chmielewski (1980), Cambanis *et al.*, (1981) y Tyler (1982), y los trabajos recientes de Arellano (1994) y Leiva (1999), entre otros. Aunque a partir de 1970 estas distribuciones comenzaron su auge, se registran estudios anteriores, como lo son Schoenberg (1938) y Lord (1954). Asimismo, las distribuciones de formas cuadráticas y distribuciones asociadas al muestreo, obtenidas a partir de distribuciones Elípticas, fueron estudiadas y publicadas en Cacoullos y Koutras (1984), Fang y Wu (1984), Anderson y Fang (1987) y Smith (1989), entre otros.

1.5. Revisión de literatura sobre la distribución de formas cuadráticas.

La distribución de formas cuadráticas y lineales, y expresiones relacionas, obtenidas a partir de distribuciones Elípticas, fueron estudiadas y publicadas en Cacoullos y Koutras (1984), Fang y Wu (1984) y Anderson y Fang (1987), entre otros. Estas distribuciones corresponden a las distribuciones χ^2, t y F obtenidas bajo distribuciones Elípticas.

Las distribuciones t y F obtenidas a partir de distribuciones Elípticas se denominan distribuciones t y F generalizadas. Sin embargo, si el parámetro de posición de la ley Elíptica es igual a cero (0), entonces las distribuciones t y F generalizadas coinciden con las distribuciones t y F obtenidas bajo normalidad, esto es, estas distribuciones son invariantes bajo leyes Elípticas cuando $v = 0$ (Fang y Zhang, 1990). Por otro lado, las distribuciones t y F generalizadas no

centradas dependen de la ley Elíptica particular bajo la cual fueron obtenidas. Asimismo, las distribuciones *t* y *F* generalizadas doble no centradas, análogas a las del caso Normal, dependen también de la ley Elíptica asociada.

CAPÍTULO 2

Diagnóstico de multicolinealidad

La estimación de parámetros en modelos de regresión lineal múltiple usando métodos que involucran matrices de correlación o de varianza-covarianza, puede verse afectada por los problemas en la estructura de las matrices causados por los efectos adversos de la multicolinealidad entre las variables del modelo. La presencia de multicolinealidad en la matriz de regresión puede ser diagnosticada mediante métodos que suministran información acerca de su intensidad (grados de multicolinealidad), así como también de las variables regresoras involucradas.

A continuación se presentan algunos procedimientos para el diagnóstico de multicolinealidad en modelos de regresión lineal múltiple:

2.1. Método del determinante de la matriz de regresión.

Se tiene la matriz de correlación, en consecuencia, su determinante varía de cero (0) a uno (1). El determinante de esta matriz toma el valor de uno (1) si las variables regresoras son ortogonales, y tomará el valor de cero (0) si éstas se encuentran en una completa combinación lineal entre sí. En la medida en que el determinante de la matriz se aproxima a cero (0) la multicolinealidad será severa. Sin embargo, aun cuando éste método es bastante sencillo, ya que involucra operaciones con matrices ampliamente conocidas y de fácil computo, no permite la identificación de las variables regresoras causantes de la multicolinealidad.

2.2. Análisis de la matriz de correlación.

Este procedimiento involucra el análisis de los elementos (r_{ij}) fuera de la diagonal principal de la matriz de correlación. Si las variables consideradas, X_i y X_j son aproximadamente linealmente dependientes, entonces r_{ij} será cercano a uno (1),

en valor absoluto. En consecuencia, un coeficiente de correlación alto indicaría multicolinealidad. No obstante, cuando el número total de variables independientes es mayor que dos (2), esto sería una condición suficiente, mas no necesaria, y la ausencia de altas correlaciones entre dos variables no indicaría ausencia de multicolinealidad.

2.3. Factor de inflación de varianza (*VIF*).

Marquardt (1970) propuso el factor de inflación de varianza (*VIF*) de los elementos de la diagonal principal de la inversa de la matriz *X'X* como alternativa para el diagnóstico de multicolinealidad, el cual puede definirse de la siguiente manera:

$$VIF_j = \frac{1}{1-R_j^2} \quad j = 1,\text{K}\ ,k$$

donde R_j^2 es el coeficiente de determinación múltiple cuando se hace la regresión de cada variable independiente sobre el resto de las otras variables regresoras mediante la estimación de mínimos cuadrados ordinarios. En este sentido, la varianza del j-ésimo estimador de mínimos cuadrados ordinarios ($\hat{\beta}_j$) es dada por:

$$V(\hat{\beta}_j) = V(\hat{b}_j)\frac{1}{1-R_j^2}$$

donde $V(\hat{b}_i)$ es la varianza de la estimación de la pendiente en la regresión simple de *Y* sobre X_j.

Así pues, el factor de inflación de varianza (*VIF*) se define como el cociente de la varianza de la estimación de una pendiente en regresión lineal múltiple y la varianza de la misma pendiente en regresión simple

$$VIF_j = \frac{V(\hat{\beta}_j)}{V(\hat{b}_j)} = \frac{1}{1-R_j^2}$$

De acuerdo a Neter *et al*., (1983), cualquier valor de un *VIF* entre 0 y 5 inclusive ambos indica que existe multicolinealidad leve en el conjunto de datos, mientras que un valor de cualquier *VIF* entre 5 y 10 inclusive sugiere multicolinealidad moderada, y un valor de VIF superior a 10 sugiere que es posible que algunos de los estimadores de mínimos cuadrados ordinarios del modelo de regresión ($\hat{\beta}_j$) asociados con tales valores estén severamente afectados por la multicolinealidad.

Aunque el *VIF* es un método de detección de multicolinealidad frecuentemente utilizado, tiene la desventaja de no suministrar información sobre cuáles variables estarían asociadas linealmente, sobre todo cuando existen relaciones complejas de dependencia lineal.

2.4. Test de Farrar-Glauber.

Definiendo *X* como una muestra de n observaciones sobre *k* variables independientes, cada una convertida en unidades estandarizadas (por el tamaño de muestra y su desviación estándar), se tiene que (*X'X*) es una matriz de correlación de primer orden. Es fácil observar que si existen pares ortogonales para todas las columnas de *X* entonces (*X'X*) se reduce a la matriz identidad y $|X'X| = 1$. Por el contrario, cuando *X* tiende a la singularidad, $|X'X|$ se aproxima a cero (0). Así, $0 \leq |X'X| \leq 1$ proporciona una medida general del grado en que *X* es interdependiente.

Esta prueba propuesta por Farrar y Glauber (1967) contrasta las hipótesis:

Ho: Las variables X_j son ortogonales

Ha: Existe multicolinealidad

Este método se basa en el estadístico

$$G = -\left[(n-1) - (2k+5)/6\right]\ln\left[\det(X'X)\right]$$

La distribución de *G* ha sido derivada por Bartlett y Rajalakshman (1953), quienes indican que la distribución del mismo, si Ho es cierta, $n \to \infty$, *X* es normal multivariada y que la matriz de correlación poblacional es igual a la matriz identidad, es χ^2_v, siendo *v=(k+1)k/2*, por lo que si $G > \chi^2_\alpha$ se acepta la existencia de multicolinealidad al nivel de significación α. Sin embargo, la procedencia de esta prueba es discutible, pues la multicolinealidad es generalmente un problema de la muestra concreta disponible y no de la población que la ha generado. Lo importante también es el grado de colinealidad en los datos.

2.5. Análisis de autovalores y autovectores en la matriz de correlación.

Cuando hay una o más relaciones de dependencia lineal entre las variables regresoras, uno o más autovalores $(\lambda_1, \mathrm{K}\,, \lambda_p)$ de la matriz de correlación serán pequeños (Belsley *et al*., 1980, y Silvey, 1969). Montgomery y Peck (1981) proponen el valor del número de condición (*CN*) de la matriz de correlación, dada su simetría, definiendo *CN* como la relación entre el mayor y el menor autovalor. Los autores señalan que, si *CN* < 100, la multicolinealidad es leve, si 100 < *CN* < 1000 la multicolinealidad es de moderada a fuerte, y si *CN* > 1000 la multicolinealidad es severa. El análisis de los autovalores puede identificar la naturaleza aproximada de las relaciones de dependencia lineal entre las variables. Para este análisis, $R = V\Lambda V'$, donde Λ es una matriz diagonal con dimensiones *p*p*, (*p* es el número de variables usadas para obtener la matriz de correlación *R*), cuyos elementos son los autovalores λ_j $(j = 1, \mathrm{K}\,, p)$ de *R*, y *V* una matriz ortogonal con dimensión *p*p*, cuyas columnas $(v_1, \mathrm{K}\,, v_p)$ son los autovectores normalizados de *R*. Un autovalor (λ_j) cercano a cero (0) indica dependencia

lineal entre las observaciones. Los elementos del autovector (v_j) asociado con este autovalor describe la naturaleza de esta dependencia.

2.6. Diagnóstico BKW.

Belsley *et al.*, (1980) propusieron las medidas de diagnóstico de multicolinealidad en modelos de regresión lineal múltiple conocidas como: la primera, descomposición en valor singular (*SDV*) y su índice de condición asociado, y la segunda, la descomposición de las varianzas de los estimadores de mínimos cuadrados ordinarios usando *SDV*, conocida como descomposición de las proporciones de varianza. Estas medidas son básicamente desarrolladas para examinar de qué manera las varianzas de los parámetros estimados ($\hat{\beta}_j$) se ven afectadas por los predictores colineales.

2.6.1. Descomposición en valor singular.

Para el modelo de regresión lineal múltiple $Y = X\beta + \varepsilon$, la matriz de datos X de n observaciones sobre p variables regresoras puede ser descompuesta de acuerdo a *SDV* como:

$$X = UDV'$$

donde $U_{(nxp)}$ (matriz cuyas columnas son los autovectores asociados con los p autovalores no nulos de $X'X$), $V_{(pxp)}$ (matriz de autovectores de $X'X$) son matrices ortogonales, $U'U = I$, $V'V = I$, y $D_{(pxp)}$ es una matriz diagonal con elementos de la diagonal no nulos μ_j, $j = 1,\text{K}, p$, llamados valores singulares. Entonces, se tiene

$$X'X = VD^2V'$$

En la *SDV* de X, U y V son matrices ortogonales de rango completo, lo cual implica que *Rango(X)=Rango (V)*. Por tanto, si X presenta una perfecta dependencia entre sus columnas, el *Rango(X)*= $r < p$, entonces exactamente $(p- r)$

valores singulares de *D* serán cero. Esto implica que un valor singular pequeño indicará presencia de una posible dependencia entre las variables regresoras.

El k-ésimo índice de condición de *X* está definido como:

$$\eta_j = \frac{\mu_{\max}}{\mu_j} = \sqrt{k_j}, \quad j = 1,\text{K}\ , p\ ,$$

donde $\eta_1,\text{K}\ ,\eta_p$ son los valores singulares de la matriz *X*. Es de resaltar que el mayor valor de k_j y η_j pueden ser definidos como el número de condición de la matriz *X'X* y de la matriz *X*, respectivamente.

Así, un valor singular pequeño en relación a $\mu_{\max}$ tendrá un alto índice de condición asociado. Belsley (1991) recomendó un índice de condición de la matriz *X* entre 10 y 30 será un indicativo de multicolinealidad moderada, mientras que para valores mayores que 30 resultará en multicolinealidad severa. La regla sugerida por este autor ha sido aceptada como un estándar de aplicación. Sin embargo, existen varias limitaciones en experimentos en los cuales el número de réplicas es pequeño. Muchos estudios se han orientado a este tema, entre los que destacan Mason y Perreault (1991), Rosen (1999), Schindler (1986) y Stinnett (1993).

Para hacer que los índices de condición sean comparables unos con otros en un conjunto de datos, en primer lugar, las variables regresoras deben estar a escala, dividiendo cada una de éstas entre su desviación estándar, para tener la misma longitud. Este escalamiento evitará que el análisis de los autovalores dependa de la unidad de medición de las variables. Posteriormente, también pueden ser centradas mediante la corrección de *X* por su promedio. Sin embargo, la elección de centrar es en cierto modo arbitraria, ya que algunos autores sostienen que centrar elimina cualquier colinealidad que involucre al intercepto. Al centrar, el intercepto será removido del modelo de regresión y en consecuencia se eliminará cualquier colinealidad que pueda existir entre el intercepto y las otras variables regresoras.

2.6.2. Descomposición de proporciones de varianza.

La matriz de varianza-covarianza de los estimadores $\hat{\beta} = (X'X)^{-1}X'Y$ de los parámetros de regresión es $\sigma^2(X'X)^{-1}$. Utilizando la *SDV*, esta puede ser escrita como:

$$V(\hat{\beta}_k) = \sigma^2 VD^{-2}V' \Rightarrow V(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^{p} \frac{v_{kj}^2}{\mu_j^2}$$

Por lo tanto, la varianza estimada de cada coeficiente de regresión se puede descomponer en una suma de términos cada uno asociado a un valor singular. Las proporciones $V(\hat{\beta}_k)$ pueden ser calculadas para cada valor singular y el término correspondiente a un valor singular pequeño será grande en relación a los demás términos de la descomposición. Suponga,

$$\varphi_{kj} = \frac{v_{kj}^2}{\mu_j^2},\ \varphi_k = \sum_{j=1}^{p} \varphi_{kj};\ k = 1,\text{K}\ , p$$

Entonces, la proporción de varianza del k-ésimo coeficiente de regresión asociado con el j-ésimo componente de su descomposición se define como:

$$\pi_{jk} = \frac{\varphi_{kj}}{\varphi_k},\ k, j = 1,\text{K}\ , p$$

Valores altos de la proporción de descomposición de varianza (0,50) para dos o más varianzas de coeficientes de regresión estimadas asociadas a valores singulares pequeños y elevados índices de condición identificarán a las variables regresoras involucradas en la dependencia lineal.

2.7. Método h-plot.

Es un procedimiento introducido por Cornsten y Gabriel (1976) para obtener representaciones gráficas de la información contenida en una matriz de varianzas

y covarianzas S_{pxp} de rango *r*, sobre espacios reducidos de baja dimensión. La selección adecuada de los vectores, denominados marcadores por sus autores, garantiza que en su representación sobre el primer plano h-plot se cumple que:

(a) El producto escalar entre dos marcadores aproxima la covarianza entre las variables correspondientes.

(b) La longitud de los marcadores aproxima la desviación estándar de las variables.

(c) El coseno del ángulo entre dos marcadores aproxima la correlación entre las variables correspondientes.

(d) El plano proporciona la mejor representación bidimensional aproximada, desde el punto de vista de los mínimos cuadrados, de las relaciones entre las variables en términos de varianzas y correlaciones.

La representación en referencia es aplicable a cualquier matriz simétrica A_{pxp} de rango *r*, y se efectúa eligiendo vectores marcadores $(h_1, h_2, \text{K}, h_p)$ para sus columnas, tales que los elementos de la matriz se obtienen a partir de operaciones de producto interno entre los marcadores:

$$a_{jj} = h_j' h_j \quad y \quad a_{jk} = h_j' h_k$$

El procedimiento para la selección de los marcadores se basa en la descomposición espectral de la matriz *A*:

$$A = V_{(r)} D_{(r)} V_{(r)}'$$

de manera que la matriz cuyas columnas contienen los marcadores se define de la siguiente manera:

$$H_{(r)}' = D_{(r)}^{1/2} V_{(r)}' = (h_1, h_2, \text{K}, h_p)$$

siendo $V_{(r)}$ una matriz cuyas columnas son los autovectores de *A* asociados con sus *r* autovalores no nulos, y $D_{(r)}$ una matriz diagonal que contiene tales autovalores. Además, Gabriel (1971) ha propuesto como medida de la bondad de la aproximación sobre el primer plano h-plot, al cociente:

$$\left(\frac{\lambda_1^2 + \lambda_2^2}{\sum \lambda_\alpha^2} \right)$$

siendo λ_1 *y* λ_2 los dos mayores autovalores de la matriz *A*.

Partiendo de las ideas expuestas anteriormente por los dos autores, Ramírez *et al.,* (2005) proponen un dispositivo gráfico cuyo propósito es obtener una representación aproximada de las relaciones de dependencia lineal que se producen entre un conjunto de variables. Específicamente, en dicha representación se visualizan los *VIF* para cada variable y los coeficientes de correlación parcial.

En este caso la descomposición espectral de rango *r* de la matriz R^{-1} toma la forma:

$$R^{-1} = r_{(ij)} = V_{(r)} D_{(r)}^{-1} V'_{(r)}$$

siendo $V_{(r)}$ la matriz cuyas columnas son los autovectores de R^{-1} asociados con sus *r* autovalores no nulos, organizados sobre la matriz diagonal $D_{(r)}^{-1}$.

Al definir la matriz de marcadores como $H'_{(r)} = D_{(r)}^{1/2} V'_{(r)} = (h_1, h_2, \mathrm{K}, h_p)$, su representación gráfica sobre el primer plano h-plot garantiza que:

1.- El coeficiente de inflación de varianza de la variable X_j es aproximado por el cuadrado de la longitud del marcador correspondiente:

$$h'_{j(2)} h_{j(2)} \approx VIF_{(j)} \quad \forall j = 1, 2, \mathrm{K}, p$$

2.- La correlación parcial entre las variables X_j y X_k es aproximada, excepto por el signo, a través del coseno del ángulo entre sus marcadores:

$$\frac{h'_j h_k}{\sqrt{h'_j h_j}\sqrt{h'_k h_k}} = -r_{jk.X(-j,-k)}$$

La información captada por la representación R^{-1}h-plot es la siguiente:

1.- Imprecisión global. Este indicador se define como *traza*(*R-1*) y se interpreta como una medida de la imprecisión global de los coeficientes de regresión debido a la multicolinealidad, ya que:

$$traza(R^{-1}) = \sum VIF_{(j)} = \sum \lambda_\alpha$$

siendo λ_α el α-ésimo autovalor de R^{-1}.

2.- Imprecisión captada por un eje. Este indicador se define como el cociente $\lambda_\alpha / traza(R^{-1})$ y se interpreta como la proporción de la imprecisión global que es captada por el α-ésimo eje h-plot.

3.- Contribución de cada variable a la imprecisión captada por un eje. Dado que:

$$\sum h_{j\alpha}^2 = \sum \left(\sqrt{\lambda_\alpha} v_{j\alpha}\right)^2 = \lambda_\alpha$$

se define como contribución de la variable X_j a la imprecisión captada por el eje α al cociente:

$$CV_j F_\alpha = \frac{h_{j\alpha}^2}{\lambda_\alpha}$$

4.- Contribución de cada eje a la imprecisión del coeficiente asociado a una variable. Dado que:

$$\sum h_{j\alpha}^2 = h_j' h_j = VIF_{(j)}$$

se define como contribución del eje α a la imprecisión del coeficiente de regresión b_j al cociente:

$$CF_\alpha V_j = \frac{h_{j\alpha}^2}{VIF_{(j)}}$$

CAPÍTULO 3

Diagnóstico de multicolinealidad basado en distribuciones de contorno elíptico

3.1. Distribuciones de contornos elípticos.

Una clase de distribuciones cuyos contornos de sus densidades tienen la misma forma elíptica de la distribución Normal, pero además contienen distribuciones de colas más y menos pesadas que las de ésta, dicha clase de distribuciones se denomina de Contornos Elípticos o simplemente distribuciones Elípticas. La teoría para esta clase de distribuciones será descrita en este capítulo, la cual ha sido ampliamente revisada y detallada por Leiva y Díaz (2001) como sigue a continuación.

3.2. Distribución elíptica singular.

Sea X un vector aleatorio *n*-dimensional, si $X \in IR^n$ es un vector aleatorio con parámetro de posición $\mu \in IR^n$, matriz de escala $\Sigma \in IR^{n\times n}$ y con $rk\Sigma = r \leq n$, entonces la distribución de X se dice singular o no singular, dependiendo si $r<n$ o si $r=n$, respectivamente.

Sea $X \in IR^n$, entonces, X pertenece a la familia de distribuciones elípticas de parámetros μ, Σ y ϕ si y solo si su función característica es

$$\psi_X(t) = e^{(it^T\Sigma t)}$$

Y se denota por $X \sim EC_n(\mu, \Sigma; \phi)$, con $\mu \in IR^n$, $\Sigma \in IR^{n\times n}$ y $\phi: IR^n \rightarrow IR$.

Observe que la función característica existe aun cuando Σ sea semidefinida positiva, es decir, cuando $rk\Sigma = r < n$, en cuyo caso la distribución elíptica obtenida es llamada distribución elíptica singular.

Sea $X \sim EC_n(\mu, \Sigma; \phi)$, con $\mu \in IR^n$, $\Sigma \in IR^{n\times n}$ y $\Sigma > 0$. Entonces la densidad de X es

$$g_X(X) = |\Sigma|^{-\frac{1}{2}} f[(X-\mu)^T \Sigma^{-1}(X-\mu)],$$

Siendo *f(u)*, con *u≥0*, una función real y que se denomina generadora de densidades. En este caso, se usa la notación $X \sim EC_n(\mu, \Sigma; f)$.

Así, la condición necesaria para que la densidad de X exista con respecto a la media en IR^n es que $rk\Sigma = r = n$, esto es, que $\Sigma > 0$, con lo cual la distribución de X es no singular.

Sea $X \sim EC_n(\mu, \Sigma; f)$, con $\mu \in IR^n$, $\Sigma \in IR^{n\times n}$ y $\Sigma > 0$, entonces X puede representarse de la forma

$$X = \mu + A^T Y,$$

de modo que

$$Y = (A^T)^{-1}(X-\mu),$$

con $\Sigma = A^T A, A \in IR^{n\times n}$.

3.3. Momentos de las distribuciones elípticas.

Si $X \sim N_n(\mu, \Sigma)$, con $\mu \in IR^n$, $\Sigma \in IR^{n\times n}$, se sabe que $IE(X) = \mu \; y \; Var(X) = \Sigma$. En el caso de las distribuciones elípticas no siempre existen estos momentos (por ejemplo la distribución de Cauchy).

Sea $X = RU^{(n)} \sim EC_n(\mu, \Sigma; f)$, con $\mu \in IR^n$, $\Sigma \in IR^{n\times n}$ y $\Sigma > 0$ y R independiente de $U^{(n)}$. Entonces, si $IE(R) < \infty \; y \; IE(R^2) < \infty$, el primer y segundo momento de X existen.

Sea $X = RU^{(n)} \sim EC_n(\mu, \Sigma; f)$, con $\mu \in IR^n$, $\Sigma \in IR^{n\times n}$ y $\Sigma > 0$ y R independiente de $U^{(n)}, IE(R) y \; IE(R^2) < \infty$. Entonces,

$$IE(X) = \mu \; y \; Var(X) = \frac{IE(R^2)}{rk\Sigma}\Sigma.$$

Se sabe que, en una distribución elíptica la matriz de varianzas-covarianzas, Σ_o, es proporcional al parámetro Σ de su distribución y en general no es igual éste.

Sean Sea $X \sim EC_n(\mu, \Sigma; \phi)$, con $\mu \in IR^n$, $\Sigma \in IR^{n\times n}$ y $\Sigma > 0$, una distribución no degenerada y M_k el *k*-ésimo momento de X. Entonces,

1. $M_1(X) = \mu$.
2. $M_2(X) = \mu\mu^T - 2\phi'(0)\Sigma \ y \ Var(X) = -2\phi'(0)\Sigma$.
3. $M_3(X) = \mu\mu^T\mu - 2\phi'(0)[\mu\Sigma + \mu\Sigma + vec(\Sigma)\mu^T]$,

si ellos existen.

Cuando la matriz de varianzas-covarianzas Σ_o de $X \sim EC_n(\mu, \Sigma; \phi)$ existe, al considerar Σ como Σ_o sólo se tiene un caso particular; la igualdad se cumple si se elige ϕ de manera que $-2\phi(0) = 1$, esto es, cuando $X \sim N_n(\mu, \Sigma)$.

3.4. Distribución normal.

La distribución Normal pertenece a la clase de distribuciones Elípticas. A continuación se presentan ciertas propiedades de esta distribución, las que no pueden extenderse a otras distribuciones Elípticas, caracterizando así a la distribución Normal.

Sea $X \sim EC_n(\mu, \Sigma; f)$, con $\mu \in IR^n$, $\Sigma \in IR^{n\times n}$ y $\Sigma > 0$. Entonces, cualquier distribución marginal es normal si y sólo si, X se distribuye normalmente.

Sea $X \sim EC_n(\mu, \Sigma; f)$, con $\mu \in IR^n$, $\Sigma \in IR^{n\times n}$ y $\Sigma = diag(\sigma_{ii}),\ i = 1, \dots, n$. Entonces, las siguientes proposiciones son equivalentes:

1. X se distribuye normalmente.
2. Las componentes de X son independientes.
3. $X_i \ y \ X_j (1 \le i < j \le n)$ son independientes.

3.5. Distribución elíptica singular.

Como se sabe, si $X \sim EC_n(\mu, \Sigma; \phi)$, con $\mu \in IR^n$, $\Sigma \in IR^{n\times n}$ y $\Sigma \geq 0$, es decir, $rk\Sigma = r < n$, entonces X tiene una distribución elíptica singular.

Sea $X \sim EC_n(\mu, \Sigma; \phi)$, con $\mu \in IR^n$, $\Sigma \in IR^{n\times n}$ y $\Sigma \geq 0$ y $rk\Sigma = r < n$. Entonces, la función de densidad de X viene dada por

$$\left(\prod_{i=1}^{r} \lambda_i^{-\frac{1}{2}}\right) f[(X-\mu)^T \Sigma^{-}(X-\mu)]$$

y

$$N^T X = N^T \mu \text{ , con probabilidad de 1,}$$

lo cual se denotará por $X \sim EC_n^r(\mu, \Sigma; f)$, donde Σ^{-} es un inverso generalizado simétrico de Σ, los λ_i son los valores propios no nulos de Σ y $N \epsilon V_{n,(n-r)} \equiv \{N \epsilon IR^{n\times(n-r)} / N^T N = I_{(n-r)}\}$, tal que $N^T \Sigma = 0$.

3.6. Distribución del estadístico *t* generalizado.

A continuación la distribución del estadístico *t* generalizado, basado en una muestra obtenida desde una población elíptica, que involucra a la distribución elíptica singular.

$X \sim EC_n(v, \sigma^2 I_n; f)$, con $v = \mu 1_n, v \in IR^n, \mu \in IR, 1_n \in IR^n$, $\sigma_0^2 = c_0\sigma^2, c_0 = -2\emptyset'(0)$. Entonces,

$$T = \sqrt{n}\frac{\bar{X}}{S} \sim Gt(n-1, \delta; f),$$

donde $\bar{X} = \sum_{i=1}^{n} {X_i}/{n}$ y $S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$ y $\delta = \mu/\sigma$ es el parámetro de no centralidad de la distribución t generalizada no centrada con (*n – 1*) grados de libertad.

3.7. Inferencias para el coeficiente de variación bajo una ley elíptica.

Las medidas de dispersión se emplean para evaluar la variabilidad de un conjunto de datos, siendo la desviación estándar la más utilizada. Sin embargo, ésta entrega poca información de la variabilidad del conjunto si es interpretada en forma aislada, sólo cuando se la relaciona con la media su interpretación tiene mayor sentido. Es por ello que, el coeficiente de variación (CV), el cual relaciona ambas medidas de dispersión es empleado habitualmente.

Considere una población con media μ y desviación estándar σ. Entonces, el coeficiente de variación se define como el cociente entre la desviación estándar y la media de esta población, es decir, $\gamma = \sigma/\mu$, con $\mu \neq 0$.

3.8. Modelo elíptico con estructura dependiente.

Se denomina modelo elíptico con estructura dependiente (Modelo D) a un vector aleatorio conjuntamente dependiente con distribución elíptica multivariada. Específicamente, sea $X = (X_1, X_2, \dots, X_n)^T$ una muestra de tamaño n de una población elíptica bajo el Modelo D, esto es, $X \sim EC_n(v, \Sigma; f)$. Entonces X tendrá:

1. Parámetro de posición $v = \mu 1_n, con\ v \in IR^n, \mu \in IR, y\ 1_n \in IR^n$, donde $1_n = (1,1, \dots, 1)^T$. El parámetro de posición coincidirá con la media cuando exista el primer momento del modelo.
2. Parámetro de escala $\Sigma = \sigma^2 I_n, con\ \Sigma \in IR^{n \times n}\ y\ \sigma^2 > 0$.
3. Matriz de varianza-covarianzas (si existe) Σ_0, proporcional al parámetro de escala, definida por $\Sigma_0 = c_0\Sigma = c_0\sigma^2 I_n,\ con\ c_0 = -2\emptyset'(0)$.

3.9. Modelo elíptico con estructura independiente.

Se denomina modelo elíptico con estructura independiente (Modelo I) a un vector aleatorio compuesto por variables aleatorias independientes e idénticamente distribuidas de forma elíptica univariada. Específicamente, sea

$X = (X_1, X_2, \dots, X_n)^T$ una muestra aleatoria de tamaño *n* de una población elíptica bajo el Modelo I, esto es, $X_i \sim EC_1(\mu, \sigma^2; f), con\ i = 1,2, \dots, n$. Entonces X_i tendrá:

1. Parámetro de posición $\mu \in IR$ y coincidirá con la media cuando exista el primer momento del modelo.
2. Parámetro de escala $\sigma^2 > 0$.
3. Varianza (si existe) σ_0^2, proporcional a σ^2 y definida por $\sigma_0^2 = c_0\sigma^2$.

Para los dos modelos planteados (D e I), la población tendrá coeficiente de variación generalizado definido por $\gamma = \sigma/\mu$, esto es, la raíz cuadrada del parámetro de escala sobre el parámetro de posición. De este modo, siempre es posible construir una medida adimensional de variabilidad relativa, existan o no los dos primeros momentos del modelo subyacente.

3.10. Estimación del coeficiente de variación de una población elíptica bajo el modelo con estructura dependiente.

Para obtener el estimador de máxima verosimilitud del coeficiente de variación con base en una muestra de una población elíptica, considerando el caso de la estimación puntual, se procede de manera análoga al caso normal, obteniendo una expresión general que depende de las funciones *f* y ϕ asociadas a toda ley elíptica, que al especificarse permiten encontrar una expresión explícita para el estimador del CV.

Sea $X = (X_1, X_2, \dots, X_n)^T \sim EC_n(\mu 1_n, \sigma^2 I_n; f)$ una muestra de tamaño *n* y γ el coeficiente de variación. Entonces el estimador de máxima verosimilitud de γ bajo el modelo con estructura dependiente es

$$\widehat{\gamma_D} = \sqrt{n\lambda(f)}\, S/\bar{X},$$

donde $\lambda(f)$ es el máximo de la función $f^*(\lambda) = \lambda^{-n/2} f(1/\lambda)$ y que corresponde a una constante que depende de cada ley elíptica.

3.11. Estimación del coeficiente de variación de una población elíptica bajo el modelo con estructura independiente.

Para obtener el estimador de máxima verosimilitud del coeficiente de variación con base en una muestra de una población elíptica bajo el modelo con estructura independiente, se procede de la misma que en el caso del modelo con estructura dependiente, sólo que en ese caso la expresión general que se obtiene debe resolverse numéricamente.

Sea $X = (X_1, X_2, \dots, X_n)^T$ una muestra aleatoria de tamaño *n*, con $X_i \sim EC_1(\mu, \sigma^2; f), (i = 1,2, \dots, n)$, y γ el coeficiente de variación. Entonces el estimador de máxima verosimilitud de γ bajo el modelo con estructura independiente es

$$\widehat{\gamma_I} = \hat{\sigma}/\hat{\mu},$$

donde $\hat{\mu}\ y\ \hat{\sigma}$ son los estimadores de máxima verosimilitud de $\mu\ y\ \sigma$, respectivamente, dados por

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} w_i (X_i - \hat{\mu})^2 \ \ y \ \ \hat{\mu} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i},$$

con

$$w_i = -2W_f(U_i), W_f(U_i) = f'(U_i)/f(U_i)\ y\ U_i = \left((X_i - \hat{\mu})/\hat{\sigma}\right)^2, \ i = 1,2, \dots, n$$

máximo de la función $f^*(\lambda) = \lambda^{-n/2} f(1/\lambda)$ y que corresponde a una constante que depende de cada ley elíptica.

3.12. Distribución del estimador de máxima verosimilitud del coeficiente de variación bajo el modelo con estructura dependiente.

La distribución exacta del inverso del estimador de máxima verosimilitud del coeficiente de variación de una población elíptica bajo el modelo con estructura dependiente está relacionada con la distribución *t* generalizada no centrada.

Sea $X = (X_1, X_2, \dots, X_n)^T \sim EC_n(\mu 1_n, \sigma^2 I_n; f)$ una muestra aleatoria de tamaño *n*, $\eta = 1/\gamma$ y $\widehat{\eta_D}$ el estimador de máxima verosimilitud de η bajo el modelo con estructura dependiente. Entonces

$$T = \sqrt{n}(\bar{X}/S) = n\sqrt{\lambda(f)}\widehat{\eta_D} \sim Gt(n-1, \delta; f),$$

con $\delta = \sqrt{n}\eta$ es el parámetro de no centralidad de la distribución *t* generalizada (*Gt*) con *n – 1* grados de libertad.

3.13. Distribución asintótica del estimador de máxima verosimilitud del coeficiente de variación bajo el modelo con estructura independiente.

En este caso se requiere la existencia de los cuatro primeros momentos de la distribución, pues la ley elíptica incide sobre la distribución asintótica a través del parámetro de curtósis (*k*). Además la distribución de este estimador está relacionado con el problema de muestras grandes.

Sea $X = (X_1, X_2, \dots, X_n)^T$ una muestra aleatoria de tamaño *n*, con $X_i \sim EC_1(\mu, \sigma^2; f)$ $(i = 1,2, \dots, n)$, γ el coeficiente de variación y $\widehat{\gamma_I}$ el estimador de máxima verosimilitud de γ bajo el modelo con estructura independiente. Entonces si $n \to \infty$

$$Z = \frac{\sqrt{n}(\widehat{\gamma_I} - \gamma)}{\gamma^2\left(\frac{\gamma^2}{4a(f)} + \frac{1}{4b(f)-1}\right)} \to N(0,1),$$

donde $a(f) = IE\left(UW_f^2(U)\right)$ y $b(f) = IE\left(U^2W_f^2(U)\right), U^{1/2} \sim EC_1(0,1; f)$.

3.14. Intervalo de confianza para el coeficiente de variación bajo el modelo con estructura dependiente.

Con base en la distribución del inverso del estimador de máxima verosimilitud del coeficiente de variación de una población elíptica bajo el modelo con estructura dependiente y el estadístico *t*, se presenta a continuación un intervalo de confianza (IC) del 100(1–α)% para el coeficiente de variación.

Sea $X = (X_1, X_2, \dots, X_n)^T \sim EC_n(\mu 1_n, \sigma^2 I_n; f)$ una muestra aleatoria de tamaño *n* y $\widehat{\gamma_D}$ el estimador de máxima verosimilitud de γ bajo el modelo con estructura dependiente. Entonces un IC aproximado del 100(1–α)% para γ es

$$\left[\left((\widehat{\gamma_D})^{-1} + \frac{t_e}{n\sqrt{\lambda(f)}}\right)^{-1}; \left((\widehat{\gamma_D})^{-1} + \frac{t_e}{n\sqrt{\lambda(f)}}\right)^{-1}\right],$$

con t_e el percentil (1-α/2) de la distribución t con (*n-1*) grados de libertad.

3.15. Intervalo de confianza para el coeficiente de variación bajo el modelo con estructura independiente.

Con base en la distribución asintótica del estimador de máxima verosimilitud y del estimador de momentos del coeficiente de variación de una población elíptica bajo el modelo con estructura independiente, y usando el método de la cantidad pivotal y el estadístico *t*, se presenta a continuación un intervalo de confianza asintótico del 100(1–α)% para el coeficiente de variación.

Sea $X = (X_1, X_2, \dots, X_n)^T$ una muestra aleatoria de tamaño *n*, con $X_i \sim EC_1(\mu, \sigma^2; f)$ $(i = 1,2, \dots, n)$, y $\widehat{\gamma_I}$ el estimador de momento de γ bajo el modelo con estructura independiente. Entonces un IC asintótico del 100(1–α)% para γ es

$$\left[\widehat{\gamma_I} \pm Z_e\sqrt{\frac{\widehat{\gamma_I}^2}{n}\left(\frac{\widehat{\gamma_I}^2}{4a(f)} + \frac{1}{4b(f)-1}\right)}\right],$$

con Z_e el percentil (1-α/2) de la distribución N(0,1).

3.16. Prueba de hipótesis para el coeficiente de variación bajo el modelo con estructura dependiente.

Sea $X = (X_1, X_2, \dots, X_n)^T \sim EC_n(\mu 1_n, \sigma^2 I_n; f)$ una muestra aleatoria de tamaño *n* y $\widehat{\gamma_D}$ el estimador de máxima verosimilitud de γ bajo el modelo con estructura dependiente. Entonces a través del estadístico

$$t = n\lambda(f)(\widehat{\gamma_D}^{-1} - \gamma^{-1}) \sim t(n-1),$$

se puede contrastar la hipótesis

$$H_0: \gamma = \gamma_0 \ \ vs. \ H_a: \gamma \neq \gamma_0,$$

(o una alternativa bilateral) para un γ_0 dado, mediante la siguiente regla de decisión: Rechace H_0 si $|t| > t_e$.

3.17. Prueba de hipótesis para el coeficiente de variación bajo el modelo con estructura independiente.

Sea $X = (X_1, X_2, \dots, X_n)^T$ una muestra aleatoria de tamaño *n*, con $X_i \sim EC_1(\mu, \sigma^2; f)$ $(i = 1,2, \dots, n)$, y $\widehat{\gamma_I}$ el estimador de momentos de γ bajo el modelo con estructura independiente. Entonces a través del estadístico

$$Z = \frac{\sqrt{n}(\widetilde{\gamma}_I - \gamma)}{c_0\gamma^2\left(\frac{4c_0\gamma^2 + 3k + 2}{4}\right)} \rightarrow N(0,1),$$

con $c_0 = 1, k = 0, k_o = 1,$ se puede contrastar la hipótesis

$$H_0: \gamma = \gamma_0 \ \ vs. \ H_a: \gamma \neq \gamma_0,$$

para un γ_0 dado, mediante la siguiente regla de decisión: Rechace H_0 si $|Z| > Z_e$.

3.18. Estadístico de prueba propuesto para el diagnóstico de multicolinealidad basado en distribuciones elípticas.

Una de las consecuencias marcadas de la multicolinealidad sobre los estimadores mínimos cuadráticos ordinarios $\hat{\beta}_i$ del modelo de regresión, es la inestabilidad que produce en los mismos, resultando esto en elevadas varianzas de dichos estimadores.

De las ideas de Belsley *et al.*, (1980) en relación al efecto del grado de multicolinealidad sobre la varianza del estimador de mínimos cuadrados ordinarios del modelo de regresión lineal y con base en lo señalado por Leiva y Díaz (2001) en cuanto a las medidas de dispersión empleadas para evaluar la variabilidad de un conjunto de datos, siendo la desviación estándar la más utilizada, y en especial el coeficiente de variación que involucra a dicha medida, cuya distribución ha sido desarrollada, se tiene el siguiente estadístico

$$T_s = {\hat{\sigma}_{\hat{\beta}_i}}\Big/{\hat{\beta}_i}$$

T_s permite tener un indicio del potencial efecto de la multicolinealidad sobre el error típico de los estimadores $\hat{\beta}_i$, es decir, sobre la estabilidad de los mismos, donde:

$\hat{\beta}_i$: representa el estimador mínimo cuadrático ordinario.

$\hat{\sigma}_{\hat{\beta}_i}$: representa el error típico de la estimación.

Se quiere probar la existencia de relaciones de dependencia lineal entre los predictores de un modelo de regresión lineal múltiple (Ho: ausencia de multicolinealidad de grado vs. Ha: presencia de multicolinealidad de grado), tomando como referencia el estadístico T_s para la construcción de un procedimiento de prueba de hipótesis estadística de ausencia de multicolinealidad de grado en un modelo de regresión lineal múltiple.

3.19. Estimador de máxima verosimilitud de γ_D bajo el modelo con estructura dependiente.

La distribución exacta del inverso del estimador de máxima verosimilitud del coeficiente de variación de una población elíptica bajo el modelo con estructura dependiente está relacionada con la distribución *t* generalizada no centrada. En tal sentido, estas ideas son aplicadas para derivar el estadístico $\widehat{\gamma_D} = \sqrt{(n-p)}\,\hat{\sigma}_{\hat{\beta}_i}/\hat{\beta}_i$ y su distribución, el cual permite medir el efecto de la

multicolinealidad sobre el estimador de mínimos cuadrados ordinarios del modelo lineal múltiple.

Para obtener el estimador EVM de γ_D de una población elíptica bajo el modelo D se procede de manera análoga al caso normal, solo que en este caso se obtiene una expresión general que depende de las funciones $f\ y\ \emptyset$ asociadas a toda ley elíptica, que al especificarse permiten encontrar una expresión explícita para el estimador de σ_{β_i}/β_i.

Sea $X = (X_1, X_2, \dots, X_n)^T \sim EC_n(\mu, \sigma^2 I_n; f)$ una muestra aleatoria de tamaño *n*, *Sea* $\hat{\beta}_i \sim EC_p(\beta_p, \Sigma; \emptyset) con\ \beta_i \in R^p, \Sigma \in R^{p\times p}, \Sigma \geq 0\ y\ rk\Sigma = r = p < n$, $\gamma = \sigma_{\beta_i}/\beta_i$, $\eta = 1/\gamma$ y $\widehat{\gamma_D} = \hat{\sigma}_{\hat{\beta}_i}/\hat{\beta}_i$. La función de densidad de $\hat{\beta}_i$ viene dada por

$$\prod_{i=1}^{n-p} \lambda_i^{-1/2} f\left[\left(\hat{\beta}_i - \beta_i\right)^T \Sigma^- (\hat{\beta}_i - \beta_i)\right]$$

lo cual se denotará por $\hat{\beta}_i \sim EC_p^{n-p}\ (\beta_i, \Sigma; f_i)$ donde Σ^- una inversa generalizada simétrica de Σ, los λ_i son los valores propios no nulos de Σ.

Entonces, el estimador de máxima verosimilitud de γ bajo el modelo con estructura dependiente es

$$\widehat{\gamma_D} = \sqrt{(n-p)\lambda(f)}\ \hat{\sigma}_{\hat{\beta}_i}/\hat{\beta}_i$$

donde $\lambda(f)$ es el máximo de la función $f^*(\lambda) = \lambda^{-\frac{n-p}{2}} f(1/\lambda)$ y que corresponde a una constante de cada ley elíptica.

La función de verosimilitud de una muestra proveniente de una población elíptica bajo el modelo D puede expresarse

$$L(\beta, \gamma, X) = \left(\frac{1}{\beta^2\gamma^2}\right)^{(n-p)/2} f\left(\frac{(n-p)}{\beta^2\gamma^2}\left(\hat{\sigma}_{\beta_i}^2 + (\hat{\beta}_i - \beta_i)^2\right)\right)$$

Al ser *f* una función monótona entonces $L(\beta, \gamma, X)$ como función de β alcanza su máximo en $\beta_i = \hat{\beta}_i$ con lo cual la verosimilitud se reduce a

$$L\left(\hat{\beta}_i, \gamma, X\right) = \left(\frac{1}{(n-p)\hat{\sigma}_{\beta_i}^2}\right)^{(n-p)/2} \left(\frac{(n-p)\hat{\sigma}_{\beta_i}^2}{\hat{\beta}_i{}^2\gamma^2}\right)^{(n-p)/2} f\left(\frac{(n-p)\hat{\sigma}_{\beta_i}^2}{\hat{\beta}_i{}^2\gamma^2}\right)$$

$$= \left(\frac{1}{(n-p)\hat{\beta}_i{}^2\gamma^2}\right)^{(n-p)/2} h\left(\frac{(n-p)\hat{\sigma}^2_{\beta_i}}{\hat{\beta}_i{}^2\gamma^2}\right)$$

Donde $h(x) = \lambda^{-\frac{(n-p)}{2}} f(x)$ con $x = \left(\frac{(n-p)\hat{\sigma}^2_{\beta_i}}{(\hat{\beta}_i\gamma)^2}\right)$ y $x \geq 0$

Como f es no creciente y continua, entonces el EVM de γ viene dado por la solución de $f'(x) + f(x)((n-p)/2x) = 0$ y esta es

$$\widehat{\gamma_D} = \sqrt{(n-p)\lambda(f)}\,\hat{\sigma}_{\hat{\beta}_i}/\hat{\beta}_i$$

Donde $\lambda(f)$ es el máximo de $f^*(\lambda) = \lambda^{-\frac{(n-p)}{2}} f\left(\frac{1}{\lambda}\right)$.

Como un estimador alternativo al de verosimilitud máxima $\widehat{\gamma_D}$ se puede proponer uno de tipo momentos dado por

$$\widetilde{\gamma_D} = \sqrt{(n-p)}\,\hat{\sigma}_{\hat{\beta}_i}/\hat{\beta}_i$$

3.20. Caso Modelo con estructura independiente (Modelo I).

Similar al modelo D, un estimador alternativo al de verosimilitud máxima bajo el modelo I es

$$\widetilde{\gamma_I} = \sqrt{(n-p)}\,\hat{\sigma}_{\hat{\beta}_i}/\hat{\beta}_i$$

3.21. Distribución del estimador de máxima verosimilitud $\widehat{\gamma_D}$ bajo el modelo con estructura dependiente.

Dado que el EVM de γ bajo el modelo D es $\widehat{\gamma_D} = \sqrt{(n-p)\lambda(f)}\,\hat{\sigma}_{\hat{\beta}_i}/\hat{\beta}_i$, $\hat{\eta}_D = 1/\widehat{\gamma_D}$

Entonces

$$T = \sqrt{(n-p)\lambda(f)}\hat{\eta}_D = \sqrt{(n-p)}\,\frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}} \sim Gt(n-p, \delta; f)$$

Con $\lambda(f)$ el máximo de $f^*(\lambda) = \lambda^{-\frac{n-p}{2}} f\left(\frac{1}{\lambda}\right)$.

Se sabe que $\lambda(f)$ corresponde a una constante que depende de cada ley elíptica. En el caso normal N(0,1) $\lambda(f)$=$\sigma^2 = 1$ y $\eta = \frac{1}{\gamma}$ es el parámetro de no centralidad de la distribución t generalizada (*Gt*) no centrada con (*n* - *p*) grados de libertad.

La singularidad de la ley elíptica afecta los grados de libertad de la distribución de T, de manera similar al caso de la distribución $G\chi^2$.

3.22. Distribución asintótica del estimador de máxima verosimilitud $\widehat{\gamma_I}$ bajo el modelo con estructura independiente.

En este caso se requiere la existencia de los cuatro primeros momentos de la distribución, pues la ley elíptica incide sobre la distribución asintótica a través del parámetro de curtósis (*k*). Además la distribución de este estimador está relacionada con el problema de muestras grandes.

Sea $X = (X_1, X_2, \dots, X_n)^T$ una muestra aleatoria de tamaño *n*, con $\hat{\beta}_i \sim EC_p^{n-p}(\beta_i, \Sigma; f_i)$ con $\beta_i \in R^p, \Sigma \in R^{p\times p}, \Sigma \geq 0$ y $rk\Sigma = r = p < n$, $\gamma = \sigma_{\beta_i}/\beta_i$, $\eta = 1/\gamma$y $\hat{\gamma}_I = \sqrt{(n-p)}\,\hat{\sigma}_{\hat{\beta}_i}/\hat{\beta}_i$el estimador de máxima verosimilitud de γ bajo el modelo con estructura independiente, el cual es función de $\hat{\beta}_i$ y $\hat{\sigma}_{\hat{\beta}_i}$.

Entonces

$$\sqrt{n-p}\big(T(\hat{\theta}) - T(\theta)\big) = \sqrt{n-p}\big(T\big(\hat{\beta}_i, \hat{\sigma}^2_{\beta_i}\big) - T(\beta_i, \sigma^2_{\beta_i})\big)$$

$$\sqrt{n-p}\big(\hat{\sigma}_{\hat{\beta}_i}/\hat{\beta}_i - \sigma_{\beta_i}/\beta_i\big) = \sqrt{n-p}(\hat{\gamma} - \gamma)$$

De Anderson (1993), referente a la distribución asintótica de una matriz de varianza-covarianza, y como se sabe

$$\sqrt{n-p}(\hat{\gamma}_I - \gamma) \to N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_0^2 & 0\\ 0 & (3k+2)\sigma_0^4\end{bmatrix}\right) \equiv N(0, V)$$

Note que $\gamma = T(\theta) = T\big(\beta_i, \sigma^2_{\beta_i}\big) = \sigma_{\beta_i}/\beta_i$ es función de β_i *y* $\sigma^2_{\beta_i}$. Análogamente, el EVM de γ, $\hat{\gamma}_I$ es función de $\hat{\beta}_i$ *y* $\hat{\sigma}^2_{\beta_i}$.

Entonces

$$\sqrt{n-p}\big(T(\hat{\theta}) - T(\theta)\big) = \sqrt{n-p}\big(T\big(\hat{\beta}_i, \hat{\sigma}^2_{\beta_i}\big) - T(\beta_i, \sigma^2_{\beta_i})\big)$$

$$\sqrt{n-p}\big(\hat{\sigma}_{\hat{\beta}_i}/\hat{\beta}_i - \sigma_{\beta_i}/\beta_i\big) = \sqrt{n-p}(\hat{\gamma}_I - \gamma)$$

Aplicando el método delta (Rao, 1965), se tiene que

$$\sqrt{n-p}(\hat{\gamma}_I - \gamma) \xrightarrow{d} N\left(0, \left(\frac{\partial T(\theta)}{\partial \theta}\right) V \left(\frac{\partial T(\theta)}{\partial \theta}\right)^T\right)$$

Donde

$$= \left(\frac{\partial T\left(\beta_i, \sigma^2_{\beta_i}\right)}{\partial \beta_i} \quad \frac{\partial T(\beta_i, \sigma^2_{\beta_i})}{\partial \sigma^2{}_{\beta_i}}\right) V \left(\frac{\partial T\left(\beta_i, \sigma^2_{\beta_i}\right)}{\partial \beta_i} \quad \frac{\partial T(\beta_i, \sigma^2_{\beta_i})}{\partial \sigma^2{}_{\beta_i}}\right)^T$$

$$= c_0 \gamma^2 \left(\frac{4c_0\gamma^2 + 3k + 2}{4}\right)$$

$$Z = \frac{\sqrt{n-p}(\tilde{\gamma}_I - \gamma)}{\sqrt{c_0\gamma^2\left(\frac{4c_0\gamma^2 + 3k + 2}{4}\right)}} \rightarrow N(0,1),$$

Para demostrar (Anderson, 1993).

3.23. Distribución asintótica de la media muestral y de la matriz de varianza-covarianza (Anderson, 1993).

Se define la matriz de varianza-covarianza muestral como $S = \frac{1}{n}A$ donde $n = N - 1$ el número de grados de libertad. Entonces, la media y la matriz de varianza-covarianza muestral se son estimadores insesgados de la media y de la matriz de varianza-covarianza del modelo:

$$E(\bar{x}) = \mu,\ E(S) = \Sigma$$

Por la ley de los grandes números son estimadores consistentes dado que $N \rightarrow \infty$, $\bar{x} \xrightarrow{p} \mu,\ S \xrightarrow{p} \Sigma$.

Las covarianzas de $\bar{x}\ y\ S$ son: $Cov(\bar{x}) = \frac{1}{N}\Sigma$, $E\left(S_{ij} - \sigma_{ij}\right)(\bar{x} - \mu) = 0$

$$Cov\left(S_{ij}, S_{kl}\right) = \frac{k}{N}\left(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}\right) + \frac{1}{n}\left(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}\right)$$

Entonces como $N \rightarrow \infty$

$$Cov\left(S_{ij}, S_{kl}\right) = \frac{k}{N}\left(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}\right) + \frac{1}{n}\left(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}\right)$$

$$nCov\left(S_{ij}, S_{kl}\right) \rightarrow (1 + k)\left(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}\right) + k\sigma_{ij}\sigma_{kl}$$

Aquí es conveniente usar algebra matricial.

3.24. Distribución asintótica de la media muestral y de la matriz de varianza-covarianza (Anderson, 1993).

Se define $vecB$, $B \otimes C$ (el producto Kronecker) y K_{mn} (la matriz conmutador) por

$$vecB = vec(b_1, \dots, b_n) = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$B \otimes C = \begin{bmatrix} b_{11}C & \cdots & b_{1n}C \\ \vdots & \ddots & \vdots \\ b_{m1}C & \cdots & b_{mn}C \end{bmatrix}$$

$$K_{mn} = vecB = vecB'$$

Se puede escribir $nCov(S_{ij}, S_{kl})$ como

$$nCov(vecS) = E(vecS - vec\Sigma)(vecS - vec\Sigma)'$$

$$\rightarrow (k+1)\left(I_{p^2} + k_{pp}\right)(\Sigma \otimes \Sigma) + kvec\Sigma(\text{vec}\Sigma)'$$

Entonces

$$\sqrt{n}\begin{bmatrix} (\bar{x} - \mu)' \\ vecS - vec\Sigma \end{bmatrix} \xrightarrow{d} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma & 0 \\ 0 & (k+1)\left(I_{p^2} + k_{pp}\right)(\Sigma \otimes \Sigma) + kvec\Sigma(\text{vec}\Sigma)' \end{pmatrix}\right]$$

por el Teorema Central de Límite y para vectores aleatorios independientes e idénticamente distribuidos (con los cuartos momentos finitos) esta afirmación constituye la base para la inferencia de muestras grandes.

Se puede escribir $nCov(S_{ij}, S_{kl})$ como

$$nCov(vecS) = E(vecS - vec\Sigma)(vecS - vec\Sigma)'$$

$$\rightarrow (k+1)\left(I_{p^2} + k_{pp}\right)(\Sigma \otimes \Sigma) + kvec\Sigma(\text{vec}\Sigma)'$$

$$(k+1)\left(I_{p^2} + k_{pp}\right)(\Sigma \otimes \Sigma) + kvec\Sigma(\text{vec}\Sigma)'$$

$$=(k+1)(2)\sigma_0^4 + k\sigma_0^4$$

$$=(3k+2)\sigma_0^4$$

3.25. Prueba de hipótesis para η bajo el modelo con estructura dependiente.

Sea $X = (X_1, X_2, \dots, X_n)^T$ una muestra aleatoria de tamaño *n*, $\hat{\beta}_i \sim EC_p^{n-p}(\beta_i, \Sigma; f_i)$, $\gamma = \sigma_{\beta_i}/\beta_i$, $\eta = 1/\gamma$ y $\widehat{\eta_D} = \sqrt{(n-p)}\,\hat{\sigma}_{\hat{\beta}_i}/\hat{\beta}_i$ el estimador de máxima verosimilitud de η bajo el modelo con estructura dependiente.

Entonces a través del estadístico

$$T = \frac{\hat{\beta}_i}{\sqrt{(n-p)}\,\hat{\sigma}_{\hat{\beta}_i}} \sim Gt(n-p, \delta; f)$$

con $\delta = \eta$ es el parámetro de no centralidad de la distribución *t* generalizada (*Gt*) no centrada con *n* – *p* grados de libertad, se puede contrastar la hipótesis

$H_0: \gamma = 1 \ \ vs. \ \ H_a: \gamma \neq 1,$

(o una alternativa bilateral), para $\gamma = 1$, mediante la siguiente regla de decisión:

Rechace H_0 si $|t| > t_e$.

3.26. Prueba de hipótesis para γ bajo el modelo con estructura independiente.

Sea $X = (X_1, X_2, \dots, X_n)^T$ una muestra aleatoria de tamaño *n*, $\hat{\beta}_i \sim EC_p^{n-p}(\beta_i, \Sigma; f_i)$, $\gamma = \sigma_{\beta_i}/\beta_i$ y $\widehat{\gamma_I} = \sqrt{(n-p)}\,\hat{\sigma}_{\hat{\beta}_i}/\hat{\beta}_i$el estimador de máxima verosimilitud de γ bajo el modelo con estructura independiente. Entonces si $n \to \infty$, a través del estadístico

$$Z = \frac{\sqrt{n-p}(\widetilde{\gamma}_I - \gamma)}{\sqrt{c_0\gamma^2\left(\frac{4c_0\gamma^2+3k+2}{4}\right)}} \to N(0,1),$$

con $c_0 = 1$, $k = 0$, $k_o = 1$, se puede contrastar la hipótesis

$H_0: \gamma = 1 \ \ vs. \ \ H_a: \gamma \neq 1,$

para un $\gamma = 1$, mediante la siguiente regla de decisión:

Rechace H_0 si $|Z| > Z_e$.

Resumen del procedimiento de prueba.

Para probar la existencia de relaciones de dependencia lineal entre los predictores de un modelo de regresión lineal múltiple, esto es:

Ho: ausencia de multicolinealidad de grado

Ha: presencia de multicolinealidad de grado.

Lo que equivale a probar las hipótesis anteriormente señaladas

$H_0: \gamma = 1 \ \ vs. \ \ H_a: \gamma \neq 1.$

Se propone un procedimiento que plantea el uso del estadístico

$T_s = \sqrt{(n-p)}\, \hat{\sigma}_{\hat{\beta}_i} / \hat{\beta}_i \, ; i = s = 1,2, \cdots, k$ para probar la hipótesis (*Ho:* Ausencia de multicolinealidad de grado), en modelos de regresión lineal múltiple, donde *n* representa el tamaño de la muestra, *k* el número de variables del modelo, $\hat{\beta}_i$representa el estimador de mínimos cuadrados ordinarios y $\hat{\sigma}_{\hat{\beta}_i}$el error típico de la estimación.

El procedimiento sugiere el cálculo de *k* valores de T_s conforme al número de variables del modelo de regresión lineal múltiple.

Seguidamente se selecciona el máximo valor T_s , el cual en el caso de muestras pequeñas, se sabe que el recíproco de este estadístico T_s^{-1}, sigue una distribución *t* generalizada (*Gt*) con *n–p* grados de libertad y parámetro de no centralidad $\delta = \sigma_{\beta_i} / \beta_i$, la cual pertenece a la familia de distribuciones de contornos elípticos.

Este valor se contrasta con el valor $T_{s_{\alpha;(n-p)}} = \frac{1}{t_{\alpha/2;\delta;(n-p)}}$.

Decisión: Si $max|Ts| \geq T_{s_{\alpha;(n-p)}}$ se rechaza *Ho* en favor de *Ha*, lo que sugiere la presencia de multicolinealidad de grado en el modelo.

Para muestras grandes, estos es, $n \to \infty$, la distribución de *Ts* es *N*(*0,1*), con

$$Z_s = \frac{\sqrt{n-p}\left(\hat{\sigma}_{\hat{\beta}_i} / \hat{\beta}_i - \sigma_{\beta_i} / \beta_i \right)}{\sqrt{c_0 \gamma^2 \left(\frac{4c_0\gamma^2 + 3k + 2}{4}\right),}} \to N(0,1)$$

Decisión: Si $max|Z_s| \geq Z_e$, donde Z_e es el percentil 1-α de la distribución *N*(0,1), se rechaza *Ho* en favor de *Ha*, lo que sugiere la presencia de multicolinealidad de grado en el modelo.

3.27. Comparación de la metodología generada con las empleadas frecuentemente.

La alternativa metodológica propuesta se confrontó con algunas de las metodologías descritas, principalmente el *VIF* y el diagnóstico BKW, las cuales, además de ser empleadas frecuentemente en el diagnóstico de multicolinealidad en modelos de regresión lineal múltiple, se constituyen como criterios o reglas prácticas.

3.28. Validación de la alternativa metodológica generada.

La validación del procedimiento metodológico propuesto se realizó a través de un estudio de simulación, para lo cual se consideró un modelo lineal múltiple como sigue:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

donde $X_{3i} = k_2 X_{2i}$, k_2 es un número real conocido, con lo cual se establecen las condiciones necesarias para que el modelo presente multicolinealidad.

Se consideraron cinco tamaños de muestra (n=7, *n=10, n=20, n=30, n=50 y n=100*) y cinco (5) distribuciones teóricas (uniforme, exponencial, normal, lognormal y gamma) para los predictores del modelo.

La simulación antes descrita se realizó bajo dos modelos (errores heterocedasticos y homocedasticos), para lo cual se utilizó el entorno de programación del softwareR 3.3.1, considerando mil (1000) simulaciones en cada uno de los casos.

Los resultados producto de la simulación permitieron calcular las frecuencias relativas de los estadísticos de prueba y confrontarlos con los cuantiles de la distribución.

CAPÍTULO 4

Identificación de variables colineales mediante el estimador del error cuadrático medio

4.1. Estadístico de prueba propuesto para identificar variables colineales basado en el error cuadrático medio del estimador de mínimos cuadrados ordinarios.

Se sabe que una de las consecuencias marcadas del fenómeno de la multicolinealidad sobre los estimadores mínimos cuadráticos ordinarios $\hat{\beta}_i$ del modelo de regresión, es la inestabilidad que produce en los mismos, resultando esto en elevadas varianzas de los estimadores de los coeficientes del modelo.

Se quiere probar que la variable regresora X_i es una variable colineal (Ho: X_i es colineal vs. Ha: X_i no es colineal), tomando como referencia el error cuadrático medio (*ECM*) del estimador de mínimos cuadrados ordinarios del modelo de regresión lineal múltiple para la construcción de un procedimiento de prueba de hipótesis estadística que permita identificar variables colineales en un modelo de regresión lineal múltiple. En ese sentido, la idea de usar el error cuadrático medio (*ECM*) tiene como fundamento los resultados presentados por Villegas *et al.*, (2013), quienes mediante un estudio de simulación reportan que el *ECM* del estimador de mínimos cuadrados ordinarios muestra una tendencia a cero (0) en el caso de las variables colineales.

Definición 1.

Sea Z una variable aleatoria normal estándar y X una variable aleatoria Chi-cuadrado con v grados de libertad. Si Z y X son independientes, entonces la

variable aleatoria $T = \dfrac{Z}{\sqrt{X/v}}$ tiene una distribución *t* de Student con *v* grados de libertad y una función de densidad de probabilidad dada por

$$f(t;v) = \frac{\Gamma[(v+1)/2]}{\sqrt{\pi v}\,\Gamma(v/2)}\left[1+(t^2/2)\right]^{-(v+1)/2}\ , -\infty < t < \infty,\ v > 0$$
$$E(t) = 0;\ V(t) = \frac{n}{n-2} \qquad (4.1.1)$$

Se sabe que el estimador $\hat{\beta}_i$ tiene una distribución normal multivariada de orden k + 1

$$\hat{\beta}_i \sim \mathrm{N}_{(k+1)}\left(\hat{\beta}_i;\sigma^2 q_{ii}\right) i = 0,1,2,...,k \qquad (4.1.2)$$

donde q_{ii} representa el *i-ésimo* elemento de la matriz $(X'X)^{-1}$.

Al mismo tiempo, se sabe que la distribución de $\hat{\sigma}^2_{\hat{\beta}_i}$ es una Chi-cuadrado con *n –(k+ 1)* grados de libertad.

Prueba.

Utilizando la hipótesis de normalidad se obtiene la siguiente relación que permite conocer la distribución de $\hat{\sigma}^2_{\hat{\beta}_i}$

Sea

$$W = \frac{(n-(k+1))\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{\mathrm{n\text{-}(k+1)}} \qquad (4.1.3)$$

De la distribución de $\hat{\beta}_i$, dada en (4.1.2), se deduce lo siguiente:

$$\hat{\beta}_i \sim \mathrm{N}_{(k+1)}\left(\hat{\beta}_i;\sigma^2 q_{ii}\right) \Rightarrow \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{q_{ii}}} \sim \mathrm{N}(0;1)\ \mathrm{i} = 0,1,...,\mathrm{k} \qquad (4.1.4)$$

Como σ^2 se desconoce, se sustituye por un estimador $\hat{\sigma}^2$, lo que permite obtener el siguiente estadístico:

De la definición dada en (4.1.3) se sabe que $\hat{\sigma}^2_{\hat{\beta}_i}$ se distribuye como una Chi-cuadrado con *n-(k+1)* grados de libertad, entonces, se tiene:

$$W = \frac{(n-(k+1))\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{\text{n-(k+1)}}. \qquad (4.1.5)$$

De igual forma, de la definición dada en (4.1.4), se tiene que la distribución de $\hat{\beta}_i$ es una normal como sigue a continuación

$$\hat{\beta}_i \sim \text{N}_{(k+1)}\left(\hat{\beta}_i ; \sigma^2 q_{ii}\right) \Rightarrow \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{q_{ii}}} \sim \text{N}(0;1)\ \text{i} = 0,1,...,\text{k} \qquad (4.1.6)$$

$$\hat{\beta}_i \sim \text{N}_{(k+1)}\left(\hat{\beta}_i ; \sigma^2 q_{ii}\right) \Rightarrow Z = \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{q_{ii}}} \sim \text{N}(0;1)\ \text{i} = 0,1,...,\text{k} \qquad (4.1.7)$$

Se tiene que, el método de mínimos cuadrados ordinarios (MCO) plantea que los parámetros del modelo pueden ser estimados minimizando la suma de los errores al cuadrado ($S_E(\hat{\beta})$), la que en términos matriciales equivale a: $S_E(\hat{\beta}) = \sum_{i=1}^{n} \hat{u}_i^2 = \hat{u}'\hat{u}$, donde $\hat{u} = Y - X\hat{\beta}$. En ese caso, el problema de minimizar la suma de los errores al cuadrado se expresa de la manera siguiente:

$$\begin{aligned} \min_{\hat{\beta}} S_E(\hat{\beta}) &= \min_{\hat{\beta}} \left[(Y - X\hat{\beta})'(Y - X\hat{\beta})\right] \\ &= \min_{\hat{\beta}} \left[Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}\right] \\ &\frac{\partial S_E(\hat{\beta})}{\partial \hat{\beta}'} = -2X'Y + 2X'X\hat{\beta} = 0 \\ &\Rightarrow \hat{\beta} = (X'X)^{-1}X'Y \end{aligned} \qquad (4.1.8)$$

De (4.1.8) se tiene:

$$X'(Y - X\hat{\beta}) = 0 \Rightarrow X'\hat{u} = 0 \qquad (4.1.9)$$

De allí que (4.1.9) es la condición de ortogonalidad.

De esta manera, el vector de parámetros estimados $\hat{\beta}$ se obtiene al resolver el siguiente sistema de ecuaciones normales:

$$X'X\hat{\beta} = X'Y$$

Se sabe que, el vector $\hat{\beta}$ es un vector aleatorio, ya que depende del vector de errores:

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + u) = \beta + (X'X)^{-1}X'u$$

(1.10)

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

$$= \beta + (X'X)^{-1}X'E(u)$$

$$\Rightarrow E(\hat{\beta}) = \beta \qquad (4.1.10)$$

De (4.1.10) queda demostrado que, el estimador MCO es insesgado.

De (4.1.10) se puede definir el error de estimación o sesgo como:

$$\hat{\beta} - \beta = (X'X)^{-1}X'u$$

Ahora, se puede calcular la varianza de $\hat{\beta}$:

$$var(\hat{\beta}) = E\left[\left(\hat{\beta} - E(\hat{\beta})\right) \cdot \left(\hat{\beta} - E(\hat{\beta})\right)'\right]$$

$$= E[(\hat{\beta} - \beta) \cdot (\hat{\beta} - \beta)']$$

$$= E[(X'X)^{-1}X'uu'X(X'X)^{-1}]$$

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

$$= (X'X)^{-1}X'(\sigma^2 I_n)X(X'X)^{-1}$$

$$= \sigma^2(X'X)^{-1} \qquad (4.1.11)$$

Para poder estimar la varianza de $\hat{\beta}$ es necesario reemplazar σ^2 en (4.1.11) por su estimador insesgado.

El error cuadrático medio (*ECM*) es una propiedad de los estimadores que mezcla los conceptos de eficiencia e insesgamiento. El *ECM* de $\hat{\beta}$ se define como:

$$ECM(\hat{\beta}) = E[(\hat{\beta} - \beta)^2]$$

Lo que se puede expresar de forma equivalente como sigue:

$$ECM(\hat{\beta}) = Var(\hat{\beta}) + [Sesgo(\hat{\beta})]^2 \qquad (4.1.12)$$

$ECM(\hat{\beta}_p^*)$ de $\hat{\beta}$ es igual a la varianza del estimador $\hat{\beta}$.

De lo antes mencionado, se tiene que, para estimadores insesgados de regresión, como es el caso de los estimadores de mínimos cuadrados ordinarios $\hat{\beta}$:

$$ECM(\hat{\beta}) = Var(\hat{\beta}) \qquad (4.1.13)$$

Nota: se usará la notación $ECM(\hat{\beta}_p^*)$ en lugar de $ECM(\hat{\beta}_p)$, dado que el procedimiento plantea el uso de variables estandarizadas Y_i^* y X_{ij}^*, como se detalla más adelante.

Partiendo de (4.1.13) y (4.1.5) se propone el siguiente estadístico:

$$W_{ecm(p)}^* = [n - (k+1)]ECM(\hat{\beta}_p^*) = [n - (k+1)]Var(\hat{\beta}_p^*) \sim \chi^2_{\alpha;\, n-(k+1)} \qquad (4.1.14)$$

el cual, de lo planteado en (4.1.14) sigue una distribución Chi-cuadrado con *n-(k+1)* grados de libertad.

4.2. Procedimiento de prueba para identificar variables colineales.

Se sabe que, las unidades con que se expresan las variables independientes (regresoras) y la variable dependiente (respuesta) en un modelo lineal influyen en la interpretación de los coeficientes de regresión, en términos del valor del coeficiente y su error estándar. Sin embargo, debe tenerse en cuenta que, la transformación de una escala de las variables del modelo a otra escala no afecta las propiedades de los estimadores de mínimos cuadrados ordinarios (Gujarati y Porter, 2010). En tal sentido, para el desarrollo del presente método, se considera en primera instancia emplear transformaciones logarítmicas sobre las variables del

modelo; log(Y_i) y log(X_{ij}), dado los resultados obtenidos que se derivan del estudio de simulación de Monte Carlo como se detalla más adelante, y luego transformarlas en variables estandarizadas, para evitar principalmente los efectos adversos que puedan reflejarse en las varianzas de los estimadores como resultado de las unidades en las que se expresan las variables originales del modelo.

Así, en la regresión de Y_i y $X_{i1}, X_{i2}\ldots, X_{ip}$, si se redefinen como:

$$Y_i^* = \frac{\log(Y_i) - \log(\overline{Y})}{\log(S_Y)};\ X_{ij}^* = \frac{\log(X_{ij}) - \log(\overline{X})}{\log(S_X)};\ i = 1, \ldots, n; j = 1, \ldots, p$$

Donde $\overline{Y}$ es la media muestral de Y_i , S_Y es la desviación estándar muestral de Y_i, $\overline{X}$ es la media muestral de X_{ij} y S_X la desviación estándar muestral de X_{ij}; las variables Y_i^* y X_{ij}^* se denominan variables estandarizadas. Una propiedad interesante de una variable estandarizada es que el valor de su media siempre es cero y el de su desviación estándar siempre es uno.

Como resultado, no importa en qué unidades se expresen las variables del modelo (Y_i y X_{ij}). En consecuencia, en lugar de llevar a cabo la regresión múltiple:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

se puede realizar la regresión sobre las variables estandarizadas de la manera siguiente:

$$Y_i^* = \beta_1^* + \beta_2^* X_{i2}^* + \beta_3^* X_{i3}^* + \cdots + \beta_p^* X_{ip}^* + \varepsilon_i^*$$

$$= \beta_2^* X_{i2}^* + \beta_3^* X_{i3}^* + \cdots + \beta_p^* X_{ip}^* + \varepsilon_i^*,$$

Pues resulta sencillo demostrar que, en la regresión que involucra a la variable respuesta estandarizada Y_i^* y a las variables regresoras estandarizadas X_{ij}^*, el término del intercepto siempre es cero. Los coeficientes de regresión de las variables regresoras estandarizadas, denotados por β_p^* se conocen en la bibliografía como los coeficientes beta.

Para probar que la variable regresora X_i es una variable colineal (Ho: X_i es colineal vs. Ha: X_i no es colineal), se propone un procedimiento que plantea el uso del error cuadrático medio $\left(ECM\left(\hat{\beta}_p^*\right)\right)$ del estimador de mínimos cuadrados ordinarios del modelo de regresión lineal múltiple estandarizado, como un estadístico denominado $W_{ecm(p)}^*$, el cual, de lo planteado en (1.33) y por lo señalado en (1.34) sigue una distribución Chi-cuadrado con *n-(k+1)* grados de libertad, donde *n* representa el tamaño de la muestra, *k* el número de parámetros del modelo.

El procedimiento sugiere el cálculo del estadístico $W_{ecm(p)}^*$, el cual representa el error cuadrático medio $\left(ECM\left(\hat{\beta}_p^*\right)\right)$ del estimador de parámetros del modelo de regresión estandarizado, que en el caso de estimadores insesgados como lo son los estimadores de mínimos cuadrados ordinarios es igual a la varianza del estimador de los *k* parámetros estimados del modelo.

$$W_{ecm(p)}^* = [n-(k+1)]ECM\left(\hat{\beta}_p^*\right) = [n-(k+1)]Var\left(\hat{\beta}_p^*\right) \sim \chi^2_{\alpha;\, n-(k+1)}$$

Seguidamente se contrastan los *k* valores estimados del estadístico de prueba $W_{ecm(p)}^*$con el valor $\chi^2_{\alpha;(n-(k+1))}$.

Decisión: Si $W_{ecm(p)}^* \leq \chi^2_{\alpha;(n-(k+1))},\ i = 1, \cdots, k$ se rechaza *Ho* en favor de *Ha*, lo que sugiere que la variable X_i es colineal.

Note que se calcula el estadístico solo para los coeficientes de las variables regresoras estandarizadas $\left(\hat{\beta}_1^*, \ldots, \hat{\beta}_p^*\right)$.

4.3. Validación de la alternativa metodológica propuesta para identificar variables colineales.

La validación del procedimiento metodológico propuesto se realizó a través de un estudio de simulación de Monte Carlo, para lo cual se considerará un modelo lineal múltiple como sigue:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

donde $X_{3i} = k_2 X_{2i}$, k_2 es un número real conocido, con lo cual se establecen las condiciones necesarias para que el modelo presente multicolinealidad. Así mismo, se consideraron cuatro (4) tamaños de muestra (*n=10, n=20, n=30, n=50 y n=100*) y cinco (5) distribuciones teóricas (uniforme, exponencial, normal, lognormal y gamma) para los regresores del modelo. La simulación antes descrita se realizó mediante el uso de algoritmos bajo el entorno de programación R 3.3.1, considerando mil simulaciones en cada uno de los casos. En tal sentido, los resultados producto de la simulación permitieron calcular las frecuencias relativas del estadístico de prueba.

4.4. Aplicación de la metodología propuesta para identificar variables colineales.

El procedimiento propuesto se ilustró mediante la aplicación del mismo a un conjunto de datos reales proveniente de ensayos en ciencias agrícolas.

CAPÍTULO 5

Diagnóstico de Multicolinealidad en un modelo de regresión lineal múltiple.

5.1. Los Datos.

Para validar los procedimientos metodológicos propuestos para el diagnóstico de multicolinealidad en modelos lineales, se realizó un estudio de simulación con un modelo lineal múltiple como sigue:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

Donde $X_{3i} = k_2 X_{2i}$, k_2 es un número real conocido, con lo cual se establecen las condiciones necesarias para que el modelo presente multicolinealidad. Así mismo, se consideraron cinco (5) tamaños de muestra (n=7, *n=10, n=20, n=30, n=50 y n=100*) y cinco (5) distribuciones teóricas de X_{1i} y X_{2i} (uniforme, exponencial, normal, log-normal y gamma) para los predictores del modelo. De igual forma, la simulación antes descrita se realizó considerando dos modelos (errores heterocedasticos y homocedasticos), para lo cual se utilizó el entorno de programación del software R 3.3.1, considerando mil (1000) simulaciones en cada uno de los casos. Los resultados producto de la simulación permitieron calcular las frecuencias relativas de los estadísticos de prueba y confrontarlos con los cuantiles de la distribución. De igual manera, se utilizaron dos grupos de datos reales; el primero correspondiente a un experimento realizado con el cultivo de maíz donde X_1 (Distancia entre hileras), X_2 (Número de mazorcas por m^2), X_3 (Número de granos por mazorca) y Y (Rendimiento granos en Ton/ha), donde *n*=20 (Chacín, 2000). El segundo grupo de datos corresponde a un experimento realizado con el pasto *Brachiaria brizantha* cv. Toledo bajo fertilización nitrogenada (30 kg nitrógeno/ha) y 21 días de corte. Las variables fueron X_1 (largo de la hoja), X_2 (ancho de la hoja), X_3 (altura de la planta) e Y (área foliar), donde *n*=50

5.2. Resultados del diagnóstico de multicolinealidad mediante un estadístico basado en distribuciones de contorno elíptico con base en un estudio de simulación con un modelo lineal.

En el Cuadro 1 se muestran las frecuencias relativas de los estadísticos de prueba T_s y Z_s basados en distribuciones de contorno elíptico para diagnosticar multicolinealidad en un modelo lineal con errores heterocedasticos, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i}$, $k_2 = 1/4$. Las frecuencias relativas de los estadísticos fueron comparados con los correspondientes cuantiles de la distribución 90%, 95% y 99% y los tamaños de muestra simulados fueron *n*=7, 20, 30, 50 y 100. Así mismo, fueron utilizadas cinco distribuciones teóricas para generar X_{1i} y X_{2i}, uniforme, normal, log-normal, exponencial y gamma.

El estudio de simulación mostró que la distribución del estadístico T_s se aproxima bien para tamaños de muestra $n \leq 20$ cuando se consideran las distribuciones exponencial y gamma y para tamaños de muestra $n \leq 10$ en el caso de la distribución uniforme, mientras que la prueba basada en el estadístico Z_s conduce a decisiones más liberales. Por otro lado, cuando se consideran la distribución normal y log-normal la distribución del estadístico Z_s se aproxima bien para tamaños de muestra $n \geq 20$. Así mismo, para la distribución uniforme este estadístico se aproxima bien para tamaños de muestra $n \geq 30$, mientras que para la distribución exponencial se requieren tamaños de muestra *n*=100 y para la distribución gamma $n \geq 50$, mientras que la prueba basada en el estadístico T_s conduce a decisiones más liberales. En líneas generales se puede decir que los resultados de esta prueba están condicionados por la distribución X_{1i} y X_{2i} y por el tamaño de muestra. Estos resultados verifican lo señalado por Fan (1984), quien demuestra que la distribución t generalizada no centrada y la normal dependen de la ley elíptica particular bajo la cual fueron obtenidas y con lo planteado por Leiva y Díaz (2001) quienes sugieren el efecto del tamaño de muestra sobre la distribución del estimador del coeficiente de variación.

Cuadro 1. Cuantiles aproximados de pruebas basadas en distribuciones de contornos elípticos para el diagnóstico de multicolinealidad en un modelo lineal con errores heterocedasticos, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i},\ k_2 = 1/4$.

n	Cuantil de distribución	Uniforme		Normal		Log-Normal		Exponencial		Gamma	
		T_s	Z_s	T_s	Z_s	T_s	Z_s	T_s	Z_s	T_s	Z_s
7	0,99	0,994	0,082	0,828	0,001	0,847	0,001	0,991	0,159	0,992	0,124
	0,95	0,935	0,088	0,417	0,000	0,443	0,001	0,938	0,176	0,955	0,144
	0,90	0,876	0,100	0,244	0,005	0,276	0,005	0,893	0,222	0,910	0,148
10	0,99	0,948	0,050	0,113	0,000	0,120	0,000	0,962	0,165	0,968	0,121
	0,95	0,842	0,058	0,017	0,000	0,032	0,000	0,906	0,228	0,900	0,134
	0,90	0,739	0,047	0,010	0,000	0,019	0,000	0,882	0,250	0,851	0,172
20	0,99	0,623	0,004	0,000	0,000	0,000	0,000	0,951	0,171	0,895	0,048
	0,95	0,338	0,152	0,000	0,999	0,000	0,999	0,884	0,206	0,746	0,096
	0,90	0,258	0,615	0,000	0,999	0,000	0,999	0,849	0,309	0,585	0,311
30	0,99	0,234	0,346	0,000	0,999	0,000	0,999	0,936	0,123	0,720	0,073
	0,95	0,079	0,930	0,000	0,999	0,000	0,999	0,824	0,321	0,455	0,589
	0,90	0,035	0,974	0,000	0,999	0,000	0,999	0,762	0,486	0,355	0,737
50	0,99	0,012	0,999	0,000	0,999	0,000	0,999	0,848	0,465	0,289	0,880
	0,95	0,003	0,999	0,000	0,999	0,000	0,999	0,653	0,715	0,110	0,966
	0,90	0,000	0,999	0,000	0,999	0,000	0,999	0,518	0,779	0,087	0,975
100	0,99	0,000	0,999	0,000	0,999	0,000	0,999	0,438	0,949	0,005	0,999
	0,95	0,000	0,999	0,000	0,999	0,000	0,999	0,238	0,957	0,003	0,999
	0,90	0,000	0,999	0,000	0,999	0,000	0,999	0,171	0,968	0,003	0,999

Análogamente, en el Cuadro 2 se muestran las frecuencias relativas de los estadísticos de prueba T_s y Z_s para diagnosticar multicolinealidad en un modelo lineal con errores homocedasticos, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i},\ k_2 = 1/4$. La simulación mostró que la distribución del estadístico T_s se aproxima bien para tamaños de muestra $n \leq 10$ cuando se considera la distribución uniforme y gamma y para tamaños de muestra $n \leq 20$ en el caso de una distribución exponencial, mientras que la prueba basada en el

estadístico Z_s conduce a decisiones más liberales. Por otro lado, cuando se consideran la distribución normal y log-normal la distribución del estadístico Z_s se aproxima bien para tamaños de muestra $n \geq 20$. Así mismo, para la distribución uniforme este estadístico se aproxima bien para tamaños de muestra $n \geq 50$, mientras que para la distribución exponencial se requieren tamaños de muestra $n \geq 50$ y para la distribución gamma $n = 50$, mientras que la prueba basada en el estadístico T_s conduce a decisiones más liberales. Al igual que en el caso del modelo lineal con errores heteroscedasticos, los resultados de esta prueba están condicionados por la distribución X_{1i} y X_{2i} y por el tamaño de muestra, es decir, por la ley elíptica de la cual son obtenidas.

Cuadro 2. Cuantiles aproximados de pruebas basadas en distribuciones de contornos elípticos para el diagnóstico de multicolinealidad en un modelo lineal con errores homocedasticos, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i},\ k_2 = 1/4$.

n	Cuantil de distribución	Uniforme		Normal		Log-Normal		Exponencial		Gamma	
		T_s	Z_s	T_s	Z_s	T_s	Z_s	T_s	Z_s	T_s	Z_s
7	0,99	0,994	0,102	0,841	0,001	0,820	0,001	0,996	0,142	0,991	0,099
	0,95	0,957	0,125	0,433	0,002	0,400	0,001	0,940	0,177	0,953	0,133
	0,90	0,914	0,135	0,241	0,006	0,208	0,000	0,907	0,186	0,882	0,113
10	0,99	0,968	0,077	0,133	0,000	0,109	0,000	0,968	0,171	0,969	0,086
	0,95	0,896	0,105	0,028	0,001	0,020	0,000	0,920	0,191	0,855	0,102
	0,90	0,820	0,124	0,022	0,001	0,011	0,000	0,885	0,199	0,808	0,109
20	0,99	0,842	0,018	0,000	0,000	0,000	0,000	0,947	0,111	0,753	0,015
	0,95	0,618	0,090	0,000	0,999	0,000	0,999	0,846	0,156	0,521	0,131
	0,90	0,495	0,351	0,000	0,999	0,000	0,999	0,777	0,267	0,435	0,433
30	0,99	0,587	0,083	0,000	0,999	0,000	0,999	0,877	0,074	0,487	0,185
	0,95	0,364	0,707	0,000	0,999	0,000	0,999	0,714	0,382	0,276	0,777
	0,90	0,256	0,850	0,000	0,999	0,000	0,999	0,616	0,544	0,177	0,894
50	0,99	0,224	0,922	0,000	0,999	0,000	0,999	0,640	0,658	0,091	0,973
	0,95	0,087	0,973	0,000	0,999	0,000	0,999	0,399	0,839	0,017	0,998
	0,90	0,046	0,984	0,000	0,999	0,000	0,999	0,301	0,891	0,014	0,996
100	0,99	0,004	0,999	0,000	0,999	0,000	0,999	0,121	0,996	0,000	0,999
	0,95	0,001	0,999	0,000	0,999	0,000	0,999	0,046	0,997	0,000	0,999
	0,90	0,000	0,999	0,000	0,999	0,000	0,999	0,023	0,996	0,000	0,999

Por otro lado, en el Cuadro 3 se muestran las frecuencias relativas de los estadísticos de prueba T_s y Z_s para diagnosticar multicolinealidad en un modelo lineal con errores heterocedasticos, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i},\ k_2 = 2$. En este caso, se incrementó el grado de multicolinealidad representado en la constante $k_2 = 2$. La simulación mostró que la distribución del estadístico T_s se aproxima bien para tamaños de muestra $n \leq 10$ cuando se considera la distribución uniforme y gamma y para tamaños de

muestra $n \leq 30$ en el caso de una distribución exponencial, mientras que la prueba basada en el estadístico Z_s conduce a decisiones más liberales. Por otro lado, cuando se consideran la distribución normal y log-normal la distribución del estadístico Z_s se aproxima bien para tamaños de muestra $n \geq 30$. Así mismo, para la distribución uniforme este estadístico se aproxima bien para tamaños de muestra $n \geq 50$, mientras que para la distribución exponencial se requieren tamaños de muestra $n = 100$ y para la distribución gamma tamaños de muestra $n \geq 50$, mientras que la prueba basada en el estadístico T_s conduce a decisiones más liberales. Al igual que en el caso del modelo lineal con errores heteroscedasticos y una combinación lineal $X_{3i} = k_2 X_{2i}$, $k_2 = 1/4$, los resultados de esta prueba están condicionados por la distribución X_{1i} y X_{2i} y por el tamaño de muestra, es decir, por la ley elíptica de la cual son obtenidas.

Cuadro 3. Cuantiles aproximados de pruebas basadas en distribuciones de contornos elípticos para el diagnóstico de multicolinealidad en un modelo lineal con errores heterocedasticos, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i},\ k_2 = 2.$

n	Cuantil de distribución	Uniforme		Normal		Log-Normal		Exponencial		Gamma	
		T_s	Z_s	T_s	Z_s	T_s	Z_s	T_s	Z_s	T_s	Z_s
7	0,99	0,998	0,089	0,960	0,009	0,975	0,017	0,990	0,157	0,992	0,121
	0,95	0,952	0,102	0,799	0,016	0,798	0,012	0,936	0,179	0,935	0,122
	0,90	0,892	0,112	0,619	0,019	0,620	0,019	0,891	0,196	0,924	0,169
10	0,99	0,956	0,056	0,644	0,003	0,597	0,004	0,962	0,186	0,966	0,113
	0,95	0,854	0,074	0,338	0,005	0,318	0,005	0,932	0,218	0,895	0,148
	0,90	0,787	0,075	0,209	0,010	0,191	0,003	0,878	0,217	0,854	0,185
20	0,99	0,719	0,008	0,045	0,000	0,030	0,000	0,960	0,174	0,895	0,049
	0,95	0,454	0,114	0,013	0,867	0,010	0,890	0,887	0,220	0,751	0,082
	0,90	0,317	0,459	0,004	0,987	0,003	0,991	0,832	0,285	0,629	0,301
30	0,99	0,328	0,215	0,002	0,981	0,001	0,987	0,941	0,139	0,718	0,075
	0,95	0,149	0,879	0,000	0,999	0,001	0,999	0,838	0,311	0,501	0,534
	0,90	0,079	0,955	0,000	0,999	0,000	0,999	0,745	0,490	0,369	0,743
50	0,99	0,036	0,994	0,000	0,999	0,000	0,999	0,834	0,396	0,338	0,867
	0,95	0,010	0,999	0,000	0,999	0,000	0,999	0,646	0,668	0,177	0,949
	0,90	0,001	0,998	0,000	0,999	0,000	0,999	0,559	0,752	0,101	0,964
100	0,99	0,000	0,999	0,000	0,999	0,000	0,999	0,450	0,925	0,023	0,998
	0,95	0,000	0,999	0,000	0,999	0,000	0,999	0,247	0,953	0,004	0,999
	0,90	0,000	0,999	0,000	0,999	0,000	0,999	0,171	0,970	0,002	0,999

Análogamente, en el Cuadro 4 se muestran las frecuencias relativas de los estadísticos de prueba T_s y Z_s para diagnosticar multicolinealidad en un modelo lineal con errores homocedasticos, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i},\ k_2 = 2.$ La simulación mostró que la distribución del estadístico T_s se aproxima bien para tamaños de muestra $n \leq 10$ cuando se considera la distribución uniforme y gamma y para tamaños de muestra $n \leq 20$ en el caso de una distribución exponencial, mientras que la prueba basada en el

estadístico Z_s conduce a decisiones más liberales. Por otro lado, cuando se consideran la distribución normal y log-normal la distribución del estadístico Z_s se aproxima bien para tamaños de muestra $n \geq 20$. Así mismo, para la distribución uniforme este estadístico se aproxima bien para tamaños de muestra $n \geq 50$, mientras que para la distribución exponencial se requieren tamaños de muestra $n = 100$ y para la distribución gamma tamaños de muestra $n \geq 50$, mientras que la prueba basada en el estadístico T_s conduce a decisiones más liberales. Al igual que en el caso del modelo lineal con errores heteroscedasticos y una combinación lineal $X_{3i} = k_2 X_{2i},\ k_2 = 2$, los resultados de esta prueba están condicionados por la distribución X_{1i} y X_{2i} y por el tamaño de muestra, es decir, por la ley elíptica de la cual son obtenidas.

Cuadro 4. Cuantiles aproximados de pruebas basadas en distribuciones de contornos elípticos para el diagnóstico de multicolinealidad en un modelo lineal con errores homocedasticos, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i},\ k_2 = 2$.

n	Cuantil de distribución	Uniforme		Normal		Log-Normal		Exponencial		Gamma	
		T_s	Z_s	T_s	Z_s	T_s	Z_s	T_s	Z_s	T_s	Z_s
7	0,99	0,993	0,116	0,956	0,010	0,953	0,010	0,986	0,128	0,993	0,099
	0,95	0,961	0,142	0,795	0,023	0,767	0,013	0,946	0,185	0,956	0,143
	0,90	0,899	0,147	0,622	0,024	0,603	0,016	0,906	0,189	0,911	0,133
10	0,99	0,965	0,086	0,635	0,003	0,611	0,003	0,964	0,174	0,978	0,090
	0,95	0,902	0,104	0,348	0,009	0,290	0,005	0,925	0,179	0,879	0,107
	0,90	0,822	0,123	0,248	0,010	0,193	0,006	0,884	0,229	0,802	0,131
20	0,99	0,836	0,026	0,053	0,000	0,042	0,000	0,957	0,111	0,837	0,021
	0,95	0,628	0,091	0,013	0,879	0,007	0,876	0,877	0,178	0,630	0,089
	0,90	0,563	0,323	0,002	0,987	0,005	0,986	0,797	0,274	0,492	0,379
30	0,99	0,646	0,080	0,001	0,976	0,000	0,975	0,888	0,081	0,527	0,144
	0,95	0,441	0,630	0,000	0,999	0,001	0,999	0,739	0,389	0,331	0,735
	0,90	0,312	0,802	0,000	0,999	0,001	0,999	0,632	0,545	0,221	0,872
50	0,99	0,302	0,883	0,000	0,999	0,000	0,999	0,639	0,647	0,111	0,972
	0,95	0,131	0,966	0,000	0,999	0,000	0,999	0,447	0,816	0,044	0,996
	0,90	0,063	0,983	0,000	0,999	0,000	0,999	0,286	0,876	0,029	0,998
100	0,99	0,013	0,999	0,000	0,999	0,000	0,999	0,126	0,991	0,002	0,999
	0,95	0,003	0,999	0,000	0,999	0,000	0,999	0,047	0,997	0,000	0,999
	0,90	0,001	0,999	0,000	0,999	0,000	0,999	0,010	0,994	0,000	0,999

5.3. Resultados del diagnóstico de multicolinealidad mediante un estadístico basado en el error cuadrático medio del estimador de mínimos cuadrados ordinarios para identificar variables colineales con base en un estudio de simulación.

En los Cuadros 5 y 6 se muestran las frecuencias relativas del estadístico de prueba basado en el error cuadrático medio del estimador de mínimos cuadrados ordinarios $W^*_{ecm(p)}$ para identificar variables colineales en un modelo lineal con errores heterocedasticos y homocedasticos, respectivamente, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i}$, $k_2 = 1/4$. Las frecuencias relativas del estadístico fueron comparados con los correspondientes cuantiles de la distribución 90%, 95% y 99% y los tamaños de muestra simulados fueron *n*=7, 20, 30, 50 y 100. Así mismo, fueron utilizadas cinco distribuciones teóricas para generar X_{1i} y X_{2i}, uniforme, normal, log-normal, exponencial y gamma.

El estudio de simulación mostró que el estadístico de prueba $W^*_{ecm(p)}$ está condicionado por la distribución teórica de los regresores, quedando en evidencia que, la forma y la simetría de la distribución producen resultados distintos que se reflejan en la frecuencia relativa de este estadístico. De allí que, cuando se considera una distribución simétrica y rectangular, como es el caso de la distribución uniforme U(a,b), en los dos modelos (errores homocedasticos y heterocedasticos), en la medida en que se incrementa el tamaño de muestra, la prueba basada en el estadístico $W^*_{ecm(p)}$ conduce a decisiones más liberales, específicamente para valores de $n > 20$. En el caso de la distribución normal, que al igual que la distribución uniforme, es también simétrica pero campaniforme o de contorno, la prueba se muestra más liberal, es decir, no se aproxima a la distribución en todo el dominio muestral ($7 \leq n \leq 100$) en ambos escenarios de los errores (homocedasticos y heterocedasticos). De igual manera, para el caso de distribuciones asimétricas, como lo son la exponencial y gamma, considerando errores heterocedasticos y homocedasticos, la distribución del estadístico de

prueba se aproxima bien y mejora conforme se incrementa el tamaño de muestra en el intervalo $7 \leq n \leq 20$, sin embargo, esta prueba muestra una tendencia a ser más liberal conforme n→∞. No obstante, las frecuencias relativas del estadístico de prueba en el caso de la distribución exponencial superan a las estimadas en la distribución gamma. Esto se explica dada la relación que existe entre ambas distribuciones, es decir, se sabe que la distribución exponencial es un caso especial de la distribución gamma, dado que cuando en la distribución gamma, el parámetro k=1, se tiene la distribución exponencial. Note que en la distribución gamma, el parámetro $k \in N$. Finalmente, en el caso del modelo con errores heterocedasticos, la distribución log-normal, que al igual que las distribuciones exponencial y gamma, esta también es asimétrica, se observa un comportamiento distinto al resto de las distribuciones consideradas en este estudio, incluyendo al de la normal. Esto se evidencia en el hecho de que, las frecuencias relativas del estadístico de prueba se incrementan conforme n→∞ en comparación con las otras distribuciones (uniforme, normal, exponencial y gamma), es decir, converge asintóticamente a la distribución y el estadístico se aproxima bien cuando $n \geq 50$.

Estos resultados observados en el caso de la distribución log-normal están relacionados directamente con las propiedades de esta distribución, especialmente cuando se sabe que si $X \sim N(\mu, \sigma^2) \rightarrow e^X \sim Log - N(\mu, \sigma^2)$, de ahí que, el hecho de usar distribuciones basadas en logaritmos, como es el caso de la log-normal, esto hace que con frecuencia se estabilicen las varianzas de los estimadores y se mantenga la propiedad de consistencia de los mismos. Así pues, es importante señalar que, la heterocedasticidad no afecta las propiedades de insesgamiento y de consistencia de los estimadores de MCO, sin embargo, dejan de tener varianza mínima, es decir, pierden la propiedad de eficiencia (Gujarati y Porter, 2010). En ese sentido, el estadístico de prueba $W^*_{ecm(p)}$ propuesto en este trabajo verifica la propiedad de consistencia cuando las variables X^*_{ij} de un modelo lineal con errores heterocedasticos y una combinación lineal $X_{3i} = k_2 X_{2i}$, $k_2 = 1/4$ siguen una distribución lognormal, dado que este converge asintóticamente a la distribución cuando n→∞.

Cuadro 5. Cuantiles aproximados de pruebas basadas en el error cuadrático medio del estimador de mínimos cuadrados ordinarios $W^*_{ecm(p)}$ para identificar variables colineales en un modelo lineal con errores heterocedasticos, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i},\ k_2 = 1/4$.

n	Cuantil de distribución	Uniforme	Normal	Log-Normal	Exponencial	Gamma
7	0,99	0,416	0,003	0,000	0,617	0,507
	0,95	0,646	0,004	0,000	0,784	0,736
	0,90	0,733	0,016	0,004	0,820	0,779
10	0,99	0,747	0,002	0,000	0,911	0,862
	0,95	0,745	0,003	0,000	0,899	0,831
	0,90	0,597	0,010	0,003	0,829	0,722
20	0,99	0,704	0,001	0,000	0,936	0,808
	0,95	0,373	0,002	0,002	0,768	0,543
	0,90	0,178	0,000	0,127	0,598	0,371
30	0,99	0,437	0,000	0,000	0,839	0,641
	0,95	0,108	0,000	0,106	0,574	0,285
	0,90	0,030	0,000	0,748	0,404	0,145
50	0,99	0,059	0,000	0,022	0,643	0,270
	0,95	0,007	0,000	0,914	0,319	0,053
	0,90	0,000	0,000	0,862	0,182	0,022
100	0,99	0,000	0,000	0,986	0,187	0,010
	0,95	0,000	0,000	0,850	0,038	0,000
	0,90	0,000	0,000	0,552	0,015	0,000

Cuadro 6. Cuantiles aproximados de pruebas basadas en el error cuadrático medio del estimador de mínimos cuadrados ordinarios $W^*_{ecm(p)}$ para identificar variables colineales en un modelo lineal con errores homocedasticos, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i},\ k_2 = 1/4$.

n	Cuantil de distribución	Uniforme	Normal	Log-Normal	Exponencial	Gamma
7	0,99	0,449	0,005	0,001	0,580	0,432
	0,95	0,695	0,011	0,000	0,789	0,677
	0,90	0,726	0,015	0,000	0,823	0,728
10	0,99	0,813	0,002	0,000	0,903	0,816
	0,95	0,761	0,002	0,000	0,878	0,829
	0,90	0,588	0,008	0,001	0,801	0,755
20	0,99	0,663	0,000	0,000	0,938	0,825
	0,95	0,320	0,001	0,000	0,747	0,569
	0,90	0,132	0,000	0,046	0,588	0,391
30	0,99	0,372	0,000	0,000	0,837	0,668
	0,95	0,062	0,000	0,029	0,589	0,333
	0,90	0,020	0,000	0,175	0,398	0,176
50	0,99	0,051	0,000	0,005	0,640	0,308
	0,95	0,000	0,000	0,134	0,296	0,086
	0,90	0,002	0,000	0,050	0,172	0,025
100	0,99	0,000	0,000	0,080	0,193	0,013
	0,95	0,000	0,000	0,017	0,040	0,001
	0,90	0,000	0,000	0,001	0,016	0,000

Análogamente, en los Cuadros 7 y 8 se muestran las frecuencias relativas del estadístico de prueba basado en el error cuadrático medio del estimador de mínimos cuadrados ordinarios $W^*_{ecm(p)}$ para identificar variables colineales en un modelo lineal con errores heterocedasticos y homocedasticos, respectivamente, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i},\ k_2 = 2$. En este caso, se incrementó el grado de multicolinealidad representado en la constante $k_2 = 2$.

Al igual que en el caso de los modelos con errores heterocedasticos y homocedasticos con la constante $k_2 = 1/4$, el estudio de simulación mostró que el estadístico de prueba $W^*_{ecm(p)}$ está condicionado por la distribución teórica de los regresores. No obstante, en ambos escenarios de los errores (heterocedasticos y homocedasticos), cuando se utilizó la distribución log-normal, se observa un incremento considerable de las frecuencias relativas del estadístico de prueba cuando n→∞ en comparación con las observadas cuando se consideró un valor de la constante $k_2 = 1/4$, en ese caso, el estadístico converge asintóticamente a la distribución y se aproxima bien en el modelo con errores heterocedasticos cuando $n \geq 10$ y en el modelo con errores homocedasticos en el intervalo $10 \leq n \leq 50$.

Estos resultados observados verifican el efecto del grado de multicolinealidad sobre las frecuencias del estadístico $W^*_{ecm(p)}$, lo que demuestra como este procedimiento de prueba se hace más robusto conforme se incrementan tanto el tamaño de muestra como el grado de multicolinealidad en el modelo, lo que se constituye en una ventaja del mismo, ya que según lo planteado por Gujarati y Porter (2010), la multicolinealidad es una cuestión de grado y no de clase, y en cuyo caso, la distinción importante no es entre presencia o ausencia de multicolinealidad, sino entre sus diferentes grados.

Cuadro 7. Cuantiles aproximados de pruebas basadas en el error cuadrático medio del estimador de mínimos cuadrados ordinarios $W^{*}_{ecm(p)}$ para identificar variables colineales en un modelo lineal con errores heterocedasticos, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i},\ k_2 = 2.$

n	Cuantil de distribución	Uniforme	Normal	Log-Normal	Exponencial	Gamma
7	0,99	0,401	0,185	0,223	0,599	0,474
	0,95	0,640	0,441	0,498	0,761	0,726
	0,90	0,769	0,640	0,742	0,841	0,791
10	0,99	0,799	0,570	0,655	0,916	0,862
	0,95	0,894	0,853	0,906	0,905	0,891
	0,90	0,833	0,851	0,961	0,849	0,793
20	0,99	0,963	0,958	0,998	0,957	0,905
	0,95	0,732	0,677	0,999	0,806	0,721
	0,90	0,456	0,317	0,999	0,650	0,497
30	0,99	0,839	0,794	0,999	0,895	0,791
	0,95	0,398	0,208	0,999	0,650	0,462
	0,90	0,144	0,031	0,998	0,461	0,269
50	0,99	0,432	0,220	0,999	0,684	0,483
	0,95	0,061	0,002	0,999	0,368	0,137
	0,90	0,005	0,000	0,999	0,178	0,038
100	0,99	0,003	0,000	0,999	0,231	0,046
	0,95	0,000	0,000	0,999	0,052	0,002
	0,90	0,000	0,000	0,999	0,011	0,000

Cuadro 8. Cuantiles aproximados de pruebas basadas en el error cuadrático medio del estimador de mínimos cuadrados ordinarios $W^{*}_{ecm(p)}$ para identificar variables colineales en un modelo lineal con errores homocedasticos, tres variables regresoras y una combinación lineal $X_{3i} = k_2 X_{2i},\ k_2 = 2.$

n	Cuantil de distribución	Uniforme	Normal	Log-Normal	Exponencial	Gamma
7	0,99	0,419	0,174	0,207	0,537	0,418
	0,95	0,660	0,429	0,484	0,767	0,682
	0,90	0,749	0,598	0,654	0,843	0,795
10	0,99	0,836	0,537	0,584	0,894	0,831
	0,95	0,728	0,764	0,855	0,924	0,927
	0,90	0,702	0,706	0,894	0,872	0,877
20	0,99	0,857	0,847	0,967	0,970	0,974
	0,95	0,502	0,397	0,903	0,860	0,838
	0,90	0,269	0,112	0,829	0,702	0,622
30	0,99	0,594	0,503	0,945	0,927	0,922
	0,95	0,181	0,054	0,870	0,747	0,659
	0,90	0,066	0,000	0,774	0,535	0,414
50	0,99	0,156	0,027	0,915	0,805	0,737
	0,95	0,012	0,000	0,773	0,483	0,286
	0,90	0,000	0,000	0,596	0,261	0,105
100	0,99	0,000	0,000	0,781	0,348	0,176
	0,95	0,000	0,000	0,477	0,092	0,010
	0,90	0,000	0,000	0,243	0,021	0,001

5.4. Resultados del diagnóstico de multicolinealidad con base en ensayos agrícolas (datos reales).

En el Cuadro 9 se muestran los resultados del análisis de regresión lineal múltiple sobre datos reales correspondientes a un experimento realizado con el cultivo de maíz donde X_1 (Distancia entre hileras), X_2 (Número de mazorcas por m^2), X_3 (Número de granos por mazorca) y Y (Rendimiento granos en Ton/ha), donde n=20. Allí se observa que el modelo presenta un coeficiente de ajuste alto (R^2= 0,9544), además de un valor del estadístico *F* del análisis de varianza del modelo significativo (p<0,05). Así mismo, las pruebas de *t* de student para los coeficientes β_1 y β_2 significativas (p<0,05); sin embargo la prueba de *t* de student para el coeficiente β_3 resultó no significativa, aun cuando en el análisis de correlación lineal mostrado en el Cuadro 10 se evidencia que la asociación lineal entre X_3 (Número de granos por mazorca) y Y (Rendimiento granos en Ton/ha) es significativa (p<0,05).

Del mismo modo, se observan signos negativos de los coeficientes de correlación, lo que indica tendencias opuestas entre las variables. En ese sentido, Chacín (2000) señala que el hecho de que la correlación entre X_1 (Distancia entre hileras) y X_3 (Número de granos por mazorca) sea de signo positivo sugiere que si bien cuando se aumenta la distancia entre las plantas disminuye el número de mazorca por metro cuadrado (X_2), el tamaño o peso de mazorcas obtenidas a baja densidad de siembra (distancia de separación grande) debe ser mayor, ya que el número de granos por mazorca aumenta cuando la separación entre las plantas es mayor. Lo antes señalado puede estar relacionado con algunas de las consecuencias de la multicolinealidad explicadas por Gujarati y Porter (2010), entre ellas razones *t* "no significativas", ya que en casos de alta colinealidad los errores estándar estimados aumentan drásticamente, lo que disminuye los valores *t*, y por consiguiente, en tales casos se acepta cada vez con mayor facilidad la hipótesis nula de que el verdadero valor poblacional relevante es cero, así como cambios en el signo de los

coeficientes de correlación que no explica la verdadera naturaleza de la relación lineal entre las variables.

Cuadro 9. Análisis de regresión lineal en un modelo para estimar el rendimiento correspondiente a un experimento realizado con el cultivo de maíz y un tamaño de muestra *n*=20.

Coeficientes	Estimado $(\hat{\beta}_i)$	Error estándar $(\hat{\sigma}_{\hat{\beta}_i})$	*t*	*P* valor
β_0	7,35290	2,977355	2,470	0,02520
β_1	-0,08291	0,010645	-7,789	0,00000
β_2	0,25039	0,115596	2,166	0,04580
β_3	0,00474	0,003728	1,270	0,22220
$F_{(\alpha,3;16)}$		111,5		0,00000
R^2		0,95440		

Cuadro 10. Análisis de correlación lineal correspondiente a un experimento realizado con el cultivo de maíz y un tamaño de muestra *n*=20.

Variables	Coeficiente de Pearson (r)	*T*	*P* valor
(Y, X_1)	-0,9641842	-15,2430	0,0000000
(Y, X_2)	0,8684349	7,4313	0,0000006
(Y, X_3)	-0,8001346	-5,6595	0,0000228
(X_1, X_2)	-0,8156203	-5,9807	0,0000172
(X_1, X_3)	0,7578859	4,9287	0,0001084
(X_2, X_3)	-0,9685284	-16,509	0,0000000

En el Cuadro 11 se muestran los resultados del diagnóstico de multicolinealidad en un modelo lineal para estimar el rendimiento correspondiente a un experimento realizado con el cultivo de maíz. En ese sentido, se observan los estadísticos de las dos metodologías propuestas en esta investigación, entre ellos los basados en distribuciones de contornos elípticos (T_s *y* Z_s) para diagnosticar multicolinealidad de grado en el modelo y el basado en el error cuadrático medio del estimador MCO ($W^*_{ecm(p)}$), que permite identificar variables involucradas en una combinación lineal en el modelo. Así mismo, se presentan otras metodologías utilizadas con frecuencia, como el valor de inflación de varianza (*VIF*), además del diagnóstico BKW; autovalores (λ_i), índice de condición (*k*) de la matriz $X´X$ y

proporciones de las varianzas. Los resultados presentados verifican en primera instancia el efecto del tamaño de muestra sobre los estadísticos de prueba T_s y Z_s, toda vez que los datos se constituyen en una muestra *n*=20, y en consecuencia el estudio de simulación anterior mostró que el estadístico T_s se aproxima bien para tamaños de muestra $n \leq 20$, mientras que Z_s se aproxima bien para tamaños de muestra *n*>20. En ese caso, se observa como el estadístico T_s es significativo (p<0,05), mientras que Z_s resultó no significativo (p>0,05), por lo que se verificó el supuesto de multicolinealidad únicamente con base en el estadístico de prueba T_s dado el tamaño de muestra del experimento.

De igual manera se muestran los valores infladores de varianza (*VIF*), los cuales dan indicios de multicolinealidad severa, ya que existen algunos *VIF* mayores a 10 (*VIF* X_2=21,59 y *VIF* X_3=16,98*)*, y según lo señalado por Neter *et al*., (1983), cualquier valor de un *VIF* superior a 10 sugiere que es posible que algunos de los estimadores de mínimos cuadrados ordinarios del modelo de regresión ($\hat{\beta}_i$) asociados con tales valores estén severamente afectados por la multicolinealidad. Así mismo, allí se observa un autovalor cercano a cero ($\lambda_3 = 0{,}0055$), lo cual es evidencia de multicolinealidad, ya que según Belsley *et al*., (1980) y Silvey (1969), cuando hay una o más relaciones de dependencia lineal entre las variables regresoras, uno o más autovalores de la matriz ($X'X$) serán pequeños. En ese orden, se presenta un índice de condición elevado (*K*=9976,118), que según lo señalado por Belsley (1991) es un indicativo de multicolinealidad severa, ya que el mismo autor sugiere que un índice de condición entre 10 y 30 será un indicativo de multicolinealidad moderada, mientras que para valores mayores que 30 resultará en multicolinealidad severa. Seguidamente se muestran los resultados de las proporciones asociadas a cada raíz característica en cada uno de los regresores, esto con el objeto de identificar el origen de la multicolinealidad. Allí se observa que las tres variables regresoras están involucradas en relaciones de dependencia lineal, ya que muestran proporciones elevadas asociadas a cada raíz característica (en las tres Componentes principales), y según lo señalado por Belsley (1991), valores altos de la proporción de descomposición de varianza

(mayores a 0,50) para dos o más varianzas de coeficientes de regresión estimadas asociadas a valores singulares pequeños y elevados índices de condición identificarán a las variables regresoras involucradas en la dependencia lineal.

Por último, se muestran los resultados del estadístico de prueba basado en el error cuadrático medio del estimador MCO ($W^{*}_{ecm(p)}$) que permite identificar variables involucradas en una combinación lineal en el modelo. En tal sentido, este estadístico arroja valores significativos ($p<0,05$) para cada una de las variables regresoras; X_1 (Distancia entre hileras), X_2 (Número de mazorcas por m^2) y X_3 (Número de granos por mazorca), lo que permite señalar que las tres variables están involucradas en una o más relaciones de dependencia lineal entre sí. Así pues, al comparar las metodologías propuestas en este trabajo (T_s, Z_s y $W^{*}_{ecm(p)}$) con las comúnmente utilizadas (*VIF*, λ_i e índice de condición *K*) para el diagnóstico de multicolinealidad, se puede decir que las mismas se muestran como poderosas herramientas para el estudio de la multicolinealidad, toda vez que no solo permiten verificar la presencia de multicolinealidad, sino más bien por el hecho de que una de ellas, como es el caso del estadístico $W^{*}_{ecm(p)}$ es capaz de identificar si una variable es colineal, lo que se constituye en un aporte sustancial en el estudio del supuesto de multicolinealidad en el modelo.

Cuadro 11. Diagnóstico de multicolinealidad en un modelo lineal para estimar el rendimiento correspondiente a un experimento realizado con el cultivo de maíz y un tamaño de muestra *n*=20.

Estadístico de prueba	Valor	*P* valor
T_s(máx)	0,7874072	0,0097047
Z_s (máx)	0,5489122	0,7084671
$W^*_{ecm(X_1)}$	3,569041	0,0005311
$W^*_{ecm(X_2)}$	3,819344	0,0008194
$W^*_{ecm(X_3)}$	0,167796	0,0000000
VIF (X_1)	3,143002	
VIF (X_2)	21,592070	
VIF (X_3)	16,983284	
λ_0	5494267	
λ_1	6101,368	
λ_2	1373,367	
λ_3	0,005520605	
Índice de condición (*K*)	9976,11803	
Componente 1	0,114*X_1+0,993* X_3	
Componente 2	0,992*X_1-0,116* X_3	
Componente 3	0,998*X_2	

En el Cuadro 12 se muestran los resultados del análisis de regresión lineal múltiple sobre datos reales correspondientes a un experimento realizado con el pasto *Brachiaria brizantha* cv. Toledo bajo fertilización nitrogenada (30 kg nitrógeno/ha) y 21 días de corte. Las variables fueron X_1 (largo de la hoja), X_2 (ancho de la hoja), X_3 (altura de la planta) e Y (área foliar), donde *n*=50. Allí se observa que el modelo presenta un coeficiente de ajuste alto (R^2= 0,9549), además de un valor del estadístico *F* del análisis de varianza del modelo significativo (p<0,05). Así mismo, las pruebas de *t* de student para los coeficientes β_1 y β_2 significativas (p<0,05); sin embargo la prueba de *t* de student para el coeficiente β_3 resultó no significativa, lo que coincide con el análisis de correlación lineal

mostrado en el Cuadro 13, en donde se evidencia que la asociación lineal entre X_3 (Altura de la planta) y Y (Área foliar) es no significativa (p>0,05).

Del mismo modo, se observan signos positivos en cada uno de los coeficientes de correlación, lo que indica tendencias similares entre las variables. En ese sentido, la literatura señala algunas alternativas de relaciones empíricas entre el área foliar y otras variables, entre ellas; longitud y ancho de folíolos, altura de la planta, que permiten estimar el área foliar mediante modelos de regresión lineal Jesús *et al*., (2001).

Cuadro 12. Análisis de regresión lineal en un modelo para estimar el área foliar correspondiente a un experimento realizado con el pasto *Brachiaria brizantha* cv. Toledo y un tamaño de muestra *n*=50.

Coeficientes	Estimado $(\hat{\beta}_i)$	Error estándar $(\hat{\sigma}_{\hat{\beta}_i})$	*t*	*P* valor
β_0	-15,59802	3,33853	-4,672	0,000002
β_1	0,91558	0,07147	12,811	0,000000
β_2	16,52434	2,28570	7,229	0,000000
β_3	-0,02448	0,05271	-0,464	0,645000
$F_{(\alpha,3;46)}$		325		0,00000
R^2		0,9549		

Cuadro 13. Análisis de correlación lineal correspondiente a un experimento realizado con el pasto *Brachiaria brizantha* cv. Toledo y un tamaño de muestra *n*=50.

Variables	Coeficiente de Pearson (r)	*T*	*P* valor
(Y, X_1)	0,9490105	20,857	0,0000000
(Y, X_2)	0,8856328	13,213	0,0000000
(Y, X_3)	0,1539402	1,0794	0,2858000
(X_1, X_2)	0,7797587	8,6288	0,0000000
(X_1, X_3)	0,2196596	1,5599	0,1253000
(X_2, X_3)	0,06209574	0,4310	0,6684000

En el Cuadro 14 se muestran los resultados del diagnóstico de multicolinealidad en un modelo lineal para estimar el área foliar en un experimento realizado con el

pasto *Brachiaria brizantha* cv. Toledo bajo fertilización nitrogenada (30 kg nitrógeno/ha) y 21 días de corte, donde *n*=50. En ese sentido, se observan los estadísticos de las dos metodologías propuestas en esta investigación, además de otras metodologías utilizadas con frecuencia. Los resultados presentados verifican en primera instancia el efecto del tamaño de muestra sobre los estadísticos de prueba T_s *y* Z_s, toda vez que los datos se constituyen en una muestra *n*=50, y en consecuencia el estudio de simulación anterior mostró que el estadístico T_s se aproxima bien para tamaños de muestra $n \leq 20$, mientras que Z_s se aproxima bien para tamaños de muestra *n*>20. En ese caso, se observa como el estadístico T_s es no significativo (p>0,05), mientras que Z_s resultó significativo (p<0,05), por lo que se verificó el supuesto de multicolinealidad únicamente con base en el estadístico de prueba Z_s dado el tamaño de muestra del experimento.

De igual manera se muestran los valores infladores de varianza (*VIF*), los cuales no dan indicios de multicolinealidad severa, ya que los *VIF* asociados a las tres variables independientes (VIFX_1=2,758 VIF*X2* =2,635 y VIFX_3=1,201*)* no superan el umbral (*VIF*=10) señalado por Neter *et al*., (1983). En ese sentido, estos resultados muestran las debilidades del *VIF* frente a otras metodologías de diagnóstico de multicolinealidad (Villegas y García, (2009); Villegas *et al*., (2011) y Villegas *et al*., (2013)). Por otro lado, no se observan autovalores menores que uno, lo cual no evidencia un grado de multicolinealidad severa presente en el modelo (Belsley *et al*., (1980) y Silvey (1969)). No obstante, se presenta un índice de condición (*K*=397,71) mayor que 30, que según lo señalado por Belsley (1991) es un indicativo de multicolinealidad severa. Seguidamente se muestran los resultados de las proporciones asociadas a cada raíz característica en cada uno de los regresores, esto con el objeto de identificar el origen de la multicolinealidad. Allí se observa que las tres variables regresoras están involucradas en relaciones de dependencia lineal, ya que muestran proporciones elevadas asociadas a cada raíz característica (en las tres Componentes principales), y según lo señalado por Belsley (1991), valores altos de la proporción de descomposición de varianza (mayores a 0,50) para dos o más

varianzas de coeficientes de regresión estimadas asociadas a valores singulares pequeños y elevados índices de condición identificarán a las variables regresoras involucradas en la dependencia lineal.

Por último, se muestran los resultados del estadístico de prueba basado en el error cuadrático medio del estimador MCO ($W^{*}_{ecm(p)}$) que permite identificar variables involucradas en una combinación lineal en el modelo. En tal sentido, este estadístico arroja valores significativos (p<0,05) para cada una de las variables regresoras; X_1 (largo de la hoja), X_2 (ancho de la hoja), X_3 (altura de la planta), lo que permite señalar que las tres variables están involucradas en una o más relaciones de dependencia lineal entre sí. Así pues, al comparar las metodologías propuestas en este trabajo (T_s, Z_s y $W^{*}_{ecm(p)}$) con las comúnmente utilizadas (*VIF*, λ_i e índice de condición *K*) para el diagnóstico de multicolinealidad, se puede decir que las mismas se muestran como poderosas herramientas para el estudio de la multicolinealidad, tanto en muestras grandes como pequeñas, toda vez que no solo permiten verificar la presencia de multicolinealidad, sino más bien por el hecho de que una de ellas, como es el caso del estadístico $W^{*}_{ecm(p)}$ es capaz de identificar si una variable es colineal, lo que se constituye en un aporte sustancial en el estudio del supuesto de multicolinealidad en el modelo.

Cuadro 14. Diagnóstico de multicolinealidad en un modelo lineal para estimar el área foliar correspondiente a un experimento realizado con el pasto *Brachiaria brizantha* cv. Toledo y un tamaño de muestra *n*=50.

Estadístico de prueba	Valor	*P* valor
T_s(máx)	0,1383231	0,5002991
Z_s (máx)	3,853536	0,0000058
$W^*_{ecm(X_1)}$	0,02657586	0,0000000
$W^*_{ecm(X_2)}$	0,02612616	0,0000000
$W^*_{ecm(X_3)}$	0,00921667	0,0000000
VIF (X_1)	2,758319	
VIF (X_2)	2,635390	
VIF (X_3)	1,085381	
λ_0	190066,7	
λ_1	6715,632	
λ_2	5,006530	
λ_3	1,201633	
Índice de condición (*K*)	397,710607	
Componente 1	0,893*X_1+0,450* X_3	
Componente 2	0,449*X_1-0,893* X_3	
Componente 3	-X_2	

CONCLUSIONES

En líneas generales se puede decir que los resultados del procedimiento de prueba propuesto basado en distribuciones de contorno elíptico para diagnosticar multicolinealidad en un modelo lineal están condicionados por la distribución de las variables regresoras y por el tamaño de muestra, ya que la distribución *t* generalizada no centrada y la normal dependen de la ley elíptica particular bajo la cual fueron obtenidas.

En presencia de errores heterocedasticos y homocedasticos, además de una combinación lineal $X_{3i} = k_2 X_{2i}$, $k_2 \in R$, la distribución del estadístico T_s se aproxima bien para tamaños de muestra $n \leq 20$ cuando se consideran las distribuciones exponencial, gamma y uniforme, así como también para tamaños de muestra $n < 10$ cuando se consideran la distribución normal y log-normal. Por otro lado, el estadístico Z_s se aproxima bien para tamaños de muestra $n > 20$ cuando se consideran la distribución normal y log-normal. No obstante, para la distribución uniforme este estadístico se aproxima bien para tamaños de muestra $n \geq 30$, mientras que para la distribución exponencial y gamma se requieren tamaños de muestra $n \geq 50$.

Los resultados de esta investigación verificaron la propiedad que tienen las distribuciones *t* generalizada no centrada y normal de ser invariantes bajo leyes elípticas, evidenciada en las frecuencias relativas obtenidas de los estadísticos de prueba T_s y Z_s cuando se incrementó el grado de multicolinealidad expresado en la constante k_2 de la combinación lineal.

En relación al procedimiento de prueba basado en el error cuadrático medio del estimador de mínimos cuadrados ordinarios en modelos con errores heterocedasticos y homocedasticos se observó que el estadístico de prueba $W^*_{ecm(p)}$ está condicionado por la distribución teórica de los regresores, quedando

en evidencia que, la forma y la simetría de la distribución producen resultados distintos que se reflejan en la frecuencia relativa de este estadístico.

En el caso de la distribución uniforme U(a,b), en la medida en que se incrementó el tamaño de muestra, la prueba basada en el estadístico $W^*_{ecm(p)}$ conduce a decisiones más liberales, específicamente para valores de $n \geq 20$. Con la distribución normal, que al igual que la distribución uniforme, es también simétrica pero campaniforme o de contorno, la prueba se mostró más liberal en todo el dominio muestral ($7 \leq n \leq 100$).

Para el caso de distribuciones asimétricas como la exponencial y gamma, la distribución del estadístico de prueba se aproximó bien y mejoró conforme se incrementó el tamaño de muestra en el intervalo $7 \leq n \leq 20$, sin embargo, esta prueba mostró una tendencia a ser más liberal conforme $n \to \infty$. No obstante, las frecuencias relativas del estadístico de prueba en el caso de la distribución exponencial superan a las estimadas en la distribución gamma, con lo que se verifica la relación que existe entre ambas distribuciones, es decir, ya que la distribución exponencial es un caso especial de la distribución gamma.

Finalmente, en el caso del modelo con errores heterocedasticos y una distribución log-normal de los regresores, se observó un comportamiento distinto al resto de las distribuciones consideradas en este estudio, incluyendo al de la normal en cuanto a las frecuencias relativas del estadístico de prueba, toda vez que dichas frecuencias se incrementaron conforme $n \to \infty$ en comparación con las otras distribuciones, lo que permitió inferir que este estadístico, converge asintóticamente a la distribución y se aproxima bien cuando $n \geq 50$.

Los resultados observados en el caso de la distribución log-normal están relacionados directamente con las propiedades de esta distribución, especialmente cuando se sabe que si $X \sim N(\mu, \sigma^2) \rightarrow e^X \sim Log - N(\mu, \sigma^2)$, de ahí que, el hecho de usar distribuciones basadas en logaritmos, como es el caso de la log-normal,

esto hace que con frecuencia se estabilicen las varianzas de los estimadores y se mantenga la propiedad de consistencia de los mismos. Así pues, es importante señalar que la heterocedasticidad no afecta las propiedades de insesgamiento y de consistencia de los estimadores de MCO, sin embargo, dejan de tener varianza mínima, es decir, pierden la propiedad de eficiencia. En ese sentido, el estadístico de prueba $W^*_{ecm(p)}$ propuesto en este trabajo verifica la propiedad de consistencia cuando las variables X^*_{ij} de un modelo lineal con errores heterocedasticos siguen una distribución log-normal.

Se verificó el efecto del grado de multicolinealidad sobre las frecuencias del estadístico $W^*_{ecm(p)}$, lo que demuestra como este procedimiento de prueba se hace más robusto conforme se incrementan tanto el tamaño de muestra como el grado de multicolinealidad en el modelo, lo que se constituye en una ventaja del mismo, ya que la multicolinealidad es una cuestión de grado y no de clase, y en cuyo caso, la distinción importante no es entre presencia o ausencia de multicolinealidad, sino entre sus diferentes grados.

Partiendo de los resultados antes señalados, para el desarrollo del presente método, se considera en primera instancia emplear transformaciones logarítmicas sobre las variables del modelo, esto es, $\log(Y_i)$ y $\log(X_{ij})$ y luego transformarlas en variables estandarizadas, para evitar principalmente los efectos adversos que puedan reflejarse en las varianzas de los estimadores como resultado de las unidades en las que se expresan las variables originales del modelo.

Finalmente, las metodologías propuestas en este trabajo (T_s, Z_s y $W^*_{ecm(p)}$) se muestran como poderosas herramientas para el estudio de la multicolinealidad si se comparan con las comúnmente utilizadas (*VIF*, λ_i, índice de condición *K* y raíces características), ya que no solo permiten verificar la presencia de multicolinealidad, sino más bien por el hecho de que una de ellas, el estadístico

$W^*_{ecm(p)}$ es capaz de identificar el origen de la multicolinealidad, lo que constituye un aporte sustancial en el estudio de este supuesto del modelo lineal múltiple.

RECOMENDACIONES

En virtud de los resultados obtenidos en esta investigación, se recomienda el uso de las metodologías propuestas (T_s, Z_s y $W^*_{ecm(p)}$), las cuales permiten no solo diagnosticar la presencia de multicolinealidad en el modelo, sino también identificar el origen de la misma, lo que resulta en una valiosa contribución al estudio de este supuesto del modelo de regresión lineal múltiple.

Así mismo, se recomienda evaluar el estadístico $W^*_{ecm(p)}$ basado en el error cuadrático medio considerando estimadores de máxima verosimilitud, con el fin de estudiar el efecto del uso de estimadores sesgados sobre este estadístico de prueba.

En ese orden, se sugiere incrementar el número de variables regresoras (X_{ij}) del modelo, así como, considerar un rango de valores (k_2=*1/2,1,3,4,*...) de la constante utilizada en la combinación lineal, esto con el fin de estudiar de forma más exhaustiva las consecuencias del grado de multicolinealidad sobre las metodologías propuestas.

Finalmente, los resultados obtenidos en relación a la distribución logarítmica dan indicios de considerar el uso de transformaciones logarítmicas como alternativa para el tratamiento de la multicolinealidad dada las propiedades de dicha distribución, especialmente aquellas relacionadas con el sesgo de la distribución. No obstante, se recomienda el uso de otras transformaciones, como por ejemplo la familia de transformadas de Box-Cox por sus efectos sobre el sesgo y las varianzas.

REFERENCIAS BIBLIOGRÁFICAS

Anderson, T.; K. Fang. 1987. Cochran's theorem for elliptically contoured distributions. Sankhya, ser a. 49: 305-315.

Anderson, T.; K. Fang. 1990. Statistical inference in elliptically contoured and related distributions. Allerton press, inc.

Arellano, R. 1994. Distribucoes elipticas: Propriedades, inferencia aplicacoes a modelos de regressao. Tesis doctoral. Universidad de Sao Paulo. Brasil.

Bartlett, M.; D. Rajalakshman. 1953. Goodness of fit tests for simultaneous autoregressive series. J. Roy. Statist. Soc. Ser. B. 15: 107-124.

Belsley, D. 1991. Conditioning diagnostics, collinearity and weak data in regression. Wiley. New York.

Belsley, D. 1982. Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise. Journal of econometrics. Elsevier. 20(2): 211-253.

Belsley, D.; E. Kuh.; R. Welch. 1980. Regression diagnostics: identifying data and source of collinearity. John wiley & sons. New York. Pp. 292.

Cacoullos, T.; M. Koutras. 1984. Quadratic form in spherical random variables generalized noncentral $\hat{a}2$-distributions. Naval res. Logist. Quart. 31: 447-461.

Callaghan, K.; J. Chen. 2008. Revisiting the collinear data problem: An assessment of estimator 'ill-conditioning' in linear regression. Practical assessment research & evaluation, 13(5): 6p.

Cambanis, S.; S. Huang.; G. Simons. 1981. On the theory of elliptically contoured distributions. J. Of multivariate analysis. 11: 365-385.

Chacín, F. 1998. Análisis de regresión y superficies de respuesta. Facultad de agronomía. Universidad central de venezuela. Venezuela. 274p.

Chacín, F.; P. Meneses. 1984. La regresión ridge y componentes principales como alternativa para fijar modelos de regresión lineal múltiple con datos con problemas de multicolinealidad. Mimeografiado. Facultad de agronomía. Universidad central de venezuela. Postgrado en estadística. Venezuela. 40p.

Chmielewski, M. 1980. Invariant scale matriz hypothesis tests under elliptical symmetry. J. Of mult. Analysis. 10: 343-350.

Chu, K. 1973. Estimation and decision for linear systems with elliptical random process. Ieee transaction on automatic control. 18: 499-505.

Cornsten, L.; K. Gabriel. 1976. 'Graphical exploration in comparing variance matrices'. Biometrics. 32: 851-863.

Cruz, C.; A. Ragazzi.1994. Modelos biométricos aplicados en mejoramiento genético. Universidad Federal de Vicosa. Mg. pp. 390.

Dawid, A. 1977. Spherical matrix distribtutions and a multivariate model. J. Of the Royal Stat. Soc. Ser. B. 39: 254-261.

De Jesús, W.; F. Do Vale.; R. Coelho.; L. Costa. 2001. Comparison of two methods for estimating leaf area index on common bean. Agronomy Journal 93: 989-991.

Fang, K.; S. Kotz.; K. Ng. 1990. Symetric, multivariate and related distribution. Monographs on Statistics and Applied Probability. 36: 329-384. Chapman and Hall. London.

Fang, K.; Y. Wu. 1984. Distribution of cuadratic forms and cochran's theorem. Mathematics in economics. 1: 29-48.

Fang, K.; Y. Zhang. 1990. Generalized multivariate analysis. Sciences press. Beijing. Springer-verlag, Berlin.

Farrar, D.; R. Glauber. 1967. 'Multicollinearity in regression analysis: The problem revisited'. Review of economic statistics. 49: 92–107.

Fisher, R. 1925. Applications of "student's" distribution. *Metron* 5: 90-104.

Gabriel, K. 1971. The biplot graphic display of matrices whit application to principal components analysis. *Biometrika*. 58: 453-467.

Glantz, S.; B. slinker. 2001. Primer of applied regression and analysis of variance. Mcgraw-Hill. New York.

Gleason, T.; R. Staelin. 1975. 'A proposal for handling missing data. Psychometrika. 40: 229–252.

Gosset, W. 1908. The probable error of a mean. *Biometrika*. 6(1): 1-25.

Gupta, A.; T. Varga. 1993. Elliptically contoured models in statistics. Kluwer academic publishers. Boston.

Habshah, M.; M. Norazan,; A. Imon. 2009. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. Journal of Applied Statistics 36(5): 507-520.

Hadi, A. 1988. Diagnosing collineariy-influential observations. Comput. Statist. Data anal. 7: 143-159.

Helmert, F. 1875. "Über die bestimmung des wahrscheinlichen fehlers aus einer endlichen anzahl wahrer beobachtungsfehler". *Z. Math. Phys.* 20: 300–3.

Helmert, F. 1876a. "Über die wahrscheinlichkeit der potenzsummen der beobachtungsfehler und uber einige damit in zusammenhang stehende fragen". *Z. Math. Phys.* 21: 192–218.

Helmert, F. 1876b. "Die genauigkeit der formel von peters zur berechnung des wahrscheinlichen beobachtungsfehlers directer beobachtungen gleicher genauigkeit". *Astron. Nachr.* 88: 113–32.

Hocking, R.; O. Pendelton. 1983. The regression dilemma. Comm. Stat. Theory meth. 12: 497-527.

Hoerl, A.; R. Kennard. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics.* 12(1): 55-67.

Jackson, J. 1991. A user's guide to principal components. Wiley. New York.

Judge, G.; W. Griffiths.; R. Hill.; H. Lutkepohl.; T. Lee. 1985. The theory and practice of econometrics (2nd ed.). New York. Wiley.

Kamruzzaman, M.; A. Imon. 2002. High leverage point: another source of multicollinearity. Pak. J. Statist. 18: 435-448.

Kariya, T.; M. Eaton. 1977. Robust tests for spherical symmetry. The Annals of Statistics. 5: 206-215.

Kelker, D. 1970. Distribution theory os spherical distributions and a location scale parameter generalization. Sankhya, Ser. A. 32: 419-430.

Kendall, M. 1957. A course in multivariate analysis. Griffin. London.

Leiva, V. 1999. "Inferencias sobre el coeficiente de variación bajo poblaciones elípticas". Tesis doctoral. Universidad de Granada. España.

Leiva, V.; J. Díaz. 2001. Distribuciones elípticas multivariadas singulares y no singulares: Teoría y aplicaciones. Universidad Autónoma Agraria Antonio Narro. Coahuila, México. 141 p.

López, E. 1998. Tratamiento de la colinealidad en regresión múltiple. Psicothema. 10(2): 491-507.

Lord, R. 1954. The use of hankel transform in statistics. I. General theory and examples. Biometrika. 41: 44-45.

Mandell, J. 1982. ‘Use of the singular value decomposition in regression analysis’. The American Statistician. 36(1): 15–24.

Marquardt, D. 1970. Generalized inverses, ridge regression and biased linear estimation. *Technometric.,* 12: 591-612.

Mason, C.; W. Perreault. 1991. Collinearity, power and interpretation of multiple regression analysis. Journal of Marketing Research. 28: 268-280.

Mason, R.; R. Gunst.; J. Webster. 1975. Regression analysis and the problem of multicollinearity. Communications in Statistics. 4: 277–292.

Moller, S.; J. Frese.; R. Bro. 2005. Robust methods for multivariate data analysis. J. Chemometr. 19: 549-563.

Montgomery, D.; E. Peck. 1981. "The multicollinearity problem in regression". Invited tutorial session at the southeast institute for decision sciences meeting. Orlando. Florida.

Montgomery, D.; E. Peck.; G. Viving. 2001. Introduction to linear regression analysis. 3th ed. New York. John Wiley and Sons.

Neter, J.; W. Wasserman. 1974. Applied linear statistical models. Homewood. I. Irwin.

Neter, J.; W. Wasserman.; M. Kutner. 1983. Applied linear regression models. Richard d. Irwin, inc. Homewood. Illinois.

Neter, J.; W. Wasserman.; M. Kutner. 1990. Applied linear statistical models. 3 edn. M.A. Irwin.

Peña, D. 1987. Estadística. Modelos y métodos. Modelos lineales y series temporales. Madrid. Alianza.

Ramírez, G.; M. Vasquez.; A. Camardiel.; B. Pérez.; P. Galindo. 2005. Detección gráfica de la multicolinealidad mediante el h-plot de la inversa de la matriz de correlaciones. Rev. Colombiana de Estadística. 28(2): 207-219.

Rao, C. 1993. Multivariate analysis: Future directions. Elsevier Science Publishers b. V.

Raveh, A. 1985. 'On the use of the inverse of the correlation matrix in multivariate data analysis'. The American Statistician 39(1): 39–42.

Rosen, D. 1999. The diagnosis of collinearity: A monte carlo simulation study. Department of epidemiology. Phd Thesis School of Emory University.

Schindler, J. 1986. Regression diagnostics: Mechanical and structural aspects of collinearity. Phd Thesis Department of Biostatistics. University of North Carolina at Chapel Hill.

Schoenberg, J. 1938. Metric spaces and completely monotone functions variates. Australian j. Stat. 7: 110-114.

Sengupta, D.; P. Bhimasankaram. 1997. On the roles of observations in collineariy in the linear model. J. Am. Stat. Assoc. 92: 1024-1032.

Silvey, S. 1969. Multicollinearity and imprecise estimation. J. Of the Royal Stat. Soc. Ser. B. 31(3): 539-552.

Smith, M. 1989. On the expectation of a ratio of cuadratic forms in normal variables. J. Mult. Analysis. 31: 244-257.

Stinnett, S. 1993. Collinearity and mixed models. Phd Thesis Department of Biostatistics. University of North Carolina at Chapel Hill.

Tyler, D. 1982. Radial estimates and the test for sphericity. Biometrika. 69: 429-436.

Velilla, S. 1987. Contribuciones al análisis de los problemas de influencia y multicolinealidad en regresión lineal. Tesis Doctoral. Universidad Complutense de Madrid.

Villegas, D.; T. García. 2009. Diagnóstico y tratamiento de la multicolinealidad en regresión lineal múltiple. Memorias VI encuentro Colombia-Venezuela de estadística. UC. Valencia. Venezuela.

Villegas, D.; M. Milla.; B. Cobo. 2011. Multicolinealidad en regresión basada en rangos. Memorias IX Congreso Latinoamericano de Sociedades de Estadística. Viña del mar. Chile.

Villegas, D.; M. Ascanio.; B. Cobo. 2013. Evaluación de la multicolinealidad en modelos de regresión lineal múltiple con presencia de valores atípicos. Rev. Fac. Agron. (UCV). 39(3): 134-143.

Webster, J.; R. Gunst.; R. Mason. 1974. Latent root regression analysis. Technometrics, 16: 513-522.

Wetherill, G. 1986. Regression analysis with applications. Chapman and Hall. London.

Whitakker, J. 1990. Graphical models in applied multivariate analysis. Wiley. New York.

Yu, C. 1998. Multicollinearity variance inflation and orthogonalization in regression.*http://seamonkey.ed.asu.edu/alex/computer/sas/collinear.html.

ANEXOS

Anexo 1. Algoritmo en el entorno de programación del software R para el cálculo del estadístico Z_s.

```
Y<-c()
X1<-c()
X2<-c()
X3<-c()
g<-lm(Y~X1+X2+X3)
X<-as.matrix(cbind(1,X1,X2,X3)[,-7])
B<-solve(t(X)%*% X,t(X)%*% Y)
sqrt(vcov(g)[2,2])
sqrt(vcov(g)[3,3])
sqrt(vcov(g)[4,4])
Ts1<-sqrt(vcov(g)[2,2])/B[2,1]
Ts1
Ts2<-sqrt(vcov(g)[3,3])/B[3,1]
Ts2
Ts3<-sqrt(vcov(g)[4,4])/B[4,1]
Ts3
TS<-c(Ts1,Ts2,Ts3)
TS
r<-max(TS)
no=length(Y)
k=ko
tz<-sqrt(no-k-1)*(r-1)/1.5
Ze<-sqrt(tz*tz)
Zo<-qnorm(1-a/2,0,1)
W<-Ze>=Zo
```

Anexo 2. Algoritmo en el entorno de programación del software R para el cálculo del estadístico T_s.

```
Y<-c()
X1<-c()
X2<-c()
X3<-c()
g<-lm(Y~X1+X2+X3)
X<-as.matrix(cbind(1,X1,X2,X3)[,-7])
B<-solve(t(X)%*% X,t(X)%*% Y)
sqrt(vcov(g)[2,2])
sqrt(vcov(g)[3,3])
sqrt(vcov(g)[4,4])
Ts1<-sqrt(vcov(g)[2,2])/B[2,1]
Ts2<-sqrt(vcov(g)[3,3])/B[3,1]
Ts3<-sqrt(vcov(g)[4,4])/B[4,1]
TS<-c(Ts1,Ts2,Ts3)
n<-length(Y)
Ta<-1/qt(1-a/2,n-5,max(TS))
max(TS)>=sqrt(Ta*Ta)->r
```

Anexo 3. Algoritmo en el entorno de programación del software R para el cálculo del estadístico $W^{*}_{ecm(p)}$.

```
Y<-c()
X1<-c()
X2<-c()
X3<-c()
LX3<-log(X3)
Z1<-(LX1-mean(LX1))/sqrt(var(LX1))
Z2<-(LX2-mean(LX2))/sqrt(var(LX2))
Z3<-(LX3-mean(LX3))/sqrt(var(LX3))
Zy<-(LY-mean(LY))/sqrt(var(LY))
g1<-lm(Zy~Z1+Z2+Z3)
summary(g1)
vcov(g1)
v1<-vcov(g1)[2,2]
v2<-vcov(g1)[3,3]
v3<-vcov(g1)[4,4]
no<-length(Y)
a<-0.01
k<-4
v1*(no-k)<=qchisq(a,no-k)->w1
w1
v2*(no-k)<=qchisq(a,no-k)->w2
w2
v3*(no-k)<=qchisq(a,no-k)->w3
```

Anexo 4. Algoritmo en el entorno de programación del software R para el cálculo de VIF, análisis de correlación lineal de Pearson y diagnóstico BKW.

```
#Correlación de Pearson en R
cor.test(Xi,Xj,alternative="two.sided",method="pearson")

#Multicolinealidad en Datos reales(VIF y BKW)
Y<-c()
X1<-c()
X2<-c()
X3<-c()
g<-lm(Y~X1+X2+X3)
summary(g)
#Valores VIF
library(faraway)
vif(g)
X<-as.matrix(cbind(1,X1,X2,X3))
X
#Matrix X´X
x.x<-t(X)%*%X
x.x
#Valores propio de X´X
lambda<-eigen(x.x)$values;lambda
#Número de condición
kappa<-max(lambda)/min(lambda);kappa
#Índices de condición
kappa.j<-sqrt(max(lambda))/sqrt(lambda);kappa.j
#Proporciones de la varianza
pr<-princomp(~X1+X2+X3)
pr$loadings
```