Pecvnia Monográfico 2012, pp. 85-110

DISTRIBUCIÓN DE LAS TRANSFORMACIONES LINEALES DE LOS RESIDUOS MÍNIMOS CUADRADOS *STUDENTIZADOS* INTERNAMENTE / *DISTRIBUTION OF LINEAR TRANSFORMATIONS OF INTERNALLY STUDENTIZED LEAST SQUARES RESIDUALS*

Seppo Pynnönen[1]
sjp@uwasa.fi

University of Vaasa (Finland)

Resumen

Los residuos de regresión por mínimos cuadrados ordinarios tienen una distribución que depende de un parámetro escalar. El término "*Studentización*" se utiliza comúnmente para describir una cantidad U dependiente de un parámetro de escala dividida por una estimación de escala S, de forma que el ratio resultante, U/S, sigue una distribución que no tiene el inconveniente del parámetro de escala desconocido. La *Studentización* externa hace referencia a un ratio en que el numerador y el denominador son independientes, mientras que la *Studentización* interna se refiere al ratio en que ambos son dependientes. La ventaja de la *Studentización* interna es que puede utilizarse cualquier estimador de escala común, mientras que en la *Studentización* externa, cada residuo es obtenido por un estimador de escala diferente, con el fin de alcanzar la independencia. Con errores de regresión normales, la distribución conjunta de un conjunto arbitrario (linealmente independiente) de residuos *Studentizados* internamente está bien documentada. Sin embargo, en algunas aplicaciones una combinación lineal de residuos internamente *Studentizados* puede resultar útil. Sus limitaciones han sido bien documentadas, pero la distribución no parece haberse derivado en la literatura. Este trabajo contribuye a la literatura existente, en el sentido de obtener la distribución conjunta de una transformación arbitraria lineal de residuos de regresión por mínimos cuadrados ordinarios internamente *Studentizados* con distribución esférica de error. Todas las principales versiones de los residuos de regresión internamente *Studentizados* que se han utilizado comúnmente en la literatura son casos especiales de la transformación lineal.

Palabras clave: Transformación de Borel de residuos *Studentizados*; Residuos normados; Distribución esférica; Distribución elíptica.

[1] Department of Mathematics and Statistics, University of Vaasa, P.O.Box 700, FI-65101, Vaasa, Finland.

Abstract

Ordinary least squares regression residuals have a distribution that is dependent on a scale parameter. The term 'Studentization' is commonly used to describe a scale parameter dependent quantity U divided by a scale estimate S such that the resulting ratio, U/S, has a distribution that is free of from the nuisance unknown scale parameter. *External* Studentization refers to a ratio in which the nominator and denominator are independent, while *internal* Studentization refers to a ratio in which these are dependent. The advantage of the internal Studentization is that typically one can use a single common scale estimator, while in the external Studentization every single residual is scaled by different scale estimator to gain the independence. With normal regression errors the joint distribution of an arbitrary (linearly independent) subset of internally Studentized residuals is well documented. However, in some applications a linear combination of internally Studentized residuals may be useful. The boundedness of them is well documented, but the distribution seems not be derived in the literature. This paper contributes to the existing literature by deriving the joint distribution of an arbitrary linear transformation of internally Studentized residuals from ordinary least squares regression with spherical error distribution. All major versions of commonly utilized internally Studentized regression residuals in literature are obtained as special cases of the linear transformation.

Keywords: Borel transformation of Studentized residuals; Normed residuals; Spherical distribution; Elliptical distribution.

1. INTRODUCCIÓN

Las transformaciones de residuos juegan un papel clave en los diagnósticos de regresión. Por consiguiente, las propiedades de las distribuciones de los residuos subyacentes son inherentes a las inferencias estadísticas eficientes sobre la calidad del modelo. Los residuos *Studentizados* se consideran útiles en el análisis de datos atípicos, en particular (por ejemplo, Chatterjee and Hadi, 1988, Sec. 4.2.1). Como consecuencia, hay un continuo interés en estudiar las propiedades estadísticas de diferentes formas de residuos *Studentizados*; entre otros podemos citar como ejemplos Abrahamse y Koerts (1971), Beckman y Trussel (1974), Chatterjee y Hadi (1988), Díaz-García y Gutiérrez-Jáimez (2006, 2007), Pynnönen (2012).

1. INTRODUCTION

Transformations of residual play a key role in regression diagnostics. Therefore, the distributional properties of the underlying residuals are eminent to efficient statistical inferences about the quality of the model. Studentized residual are considered useful in analysis of outliers, in particular (e.g., Chatterjee and Hadi, 1988, Sec. 4.2.1). As a consequence there is a continuous interest to study the statistical properties of different forms of Studentized residuals; examples are among others Abrahamse and Koerts (1971), Beckman and Trussel (1974), Chatterjee and Hadi (1988), Díaz-García and Gutiérrez-Jáimez (2006, 2007), Pynnönen (2012).

Este trabajo continúa esta línea de investigación y deduce, explícitamente, la distribución de una transformación lineal arbitraria de residuos de una regresión múltiple con errores elípticos que han sido interna y externamente *Studentizados*. En principio, la mayoría de resultados presentados en este trabajo pueden obtenerse como casos especiales del de regresión multivariante abordado en Pynnönen (2012). Sin embargo, muchos de esos resultados no son tan obvios y por ello se incluyen dentro de la estructura de regresión múltiple.

Para este objetivo, consideremos una modelo de regresión con n observaciones:

$$y = X\beta + u \qquad (1)$$

donde X es una matriz no estocástica de orden $n \times p'$ con rango $p \leq p' < n$, y es un vector n-dimensional de respuestas observables, β es un vector p'-dimensional de parámetros de pendiente, y u es un vector n-dimensional de errores homoscedásticos no observables que siguen una distribución de contorno esférico con matriz de varianzas-covarianzas proporcional a $\sigma_u^2 I_n$, donde $\sigma_u^2 > 0$ es el parámetro de escala de la varianza e I_n es la matriz identidad de orden $n \times n$.

Los residuos por mínimos cuadrados ordinarios (MCO) vienen dados por:

$$\begin{aligned} \hat{u} &= Q_y \\ &= Q_u \end{aligned} \qquad (2)$$

donde

$$Q = I_n - X(X'X)^{-}X' \qquad (3)$$

This paper continues this research and explicitly derives the distribution of an arbitrary linear transformation of internally and externally Studentized residuals from a multiple regression with elliptical errors. In principle most of the results presented in this paper can be obtained as special cases from multivariate regression dealt with in Pynnönen (2012). However, many of the results are not that obvious and motivates us to reconstruct them within the multiple regression framework.

For the purpose, consider a regression model with n observations

$$y = X\beta + u \qquad (1)$$

where X is an $n \times p'$ nonstochastic matrix with rank $p \leq p' < n$, y is an n-vector of observable responses, β is a p'-vector of slope parameters, and u is an n-vector of unobservable homoscedastic errors that follow some spherical contoured distribution with variance-covariance matrix proportional to $\sigma_u^2 I_n$, where $\sigma_u^2 > 0$ is the scalar variance parameter and I_n is the $n \times n$ identity matrix.

The ordinary least squares (OLS) residuals are given by

$$\begin{aligned} \hat{u} &= Q_y \\ &= Q_u \end{aligned} \qquad (2)$$

where

$$Q = I_n - X(X'X)^{-}X' \qquad (3)$$

es una matriz simétrica idempotente de orden $n \times n$ con rango $n - p$ en la que la prima denota la matriz traspuesta y $(X'X)^-$ es la inversa generalizada de $X'X$.

La *Studentización* es un término común utilizado para describir la división de un estadístico dependiente de un parámetro de escala, U, por una estimación de escala S de forma que la distribución del ratio resultante U/S ya no depende de los parámetros de escala (véase, por ejemplo, Margolin 1977). Por lo general, U y S se obtienen de los mismos datos, en cuyo caso el ratio U/S se denomina internamente *Studentizado* si U y S son dependientes y externamente *Studentizado* si son independientes (véase Cook y Weisberg, 1982: 18).

En la regresión por mínimos cuadrados los residuos internamente *Studentizados* se definen como:

$$\tilde{r}_i = \frac{\hat{u}_i}{s\sqrt{q_{ii}}} \qquad (4)$$

donde $\hat{u}_i$ es el componente *i*-ésimo del vector $\hat{u}$, $s = \sqrt{\hat{u}'\hat{u}/(n-p)}$, y q_{ii} es el elemento *i*-ésimo de la diagonal principal de la matriz Q. Incluso con errores distribuidos normalmente, si bien $\tilde{r}_i$ es el conocido ratio de una variable aleatoria normalmente distribuida y la raíz cuadrada de una variable aleatoria *chi*-cuadrado, el resultado final no es una variable aleatoria distribuida según una *t*. La razón es que el numerador y el denominador no son independientes debido al hecho de que $\hat{u}_i^2 < \hat{u}'\hat{u}$ para todo $i = 1,\ldots, n$. Además, puesto que $q_{ii} \geq 1 + 1/n$ (p.e., Cook y Weisberg, 1982: 12), a diferencia de una variable aleatoria que sigue una distribución t y que supone todos los valores reales, se verifica que $\tilde{r}_i^2 \leq n - p$.

is an $n \times n$ symmetric idempotent matrix with rank $n - p$ in which the prime denotes the matrix transposition and $(X'X)^-$ is a generalized inverse of $X'X$.

Studentization is a common term used to describe division of a scale parameter dependent statistic, say U, by a scale estimate S such that the distribution if the resulting ratio U/S is free from the nuisance scale parameters (see e.g. Margolin, 1977). Typically U and S are derived from the same data in which case the ratio U/S is called internally Studentized if U and S are dependent and externally Studentized if they are independent (see Cook and Weisberg, 1982: 18).

In the least squares regression the internally Studentized residuals are defined as

$$\tilde{r}_i = \frac{\hat{u}_i}{s\sqrt{q_{ii}}} \qquad (4)$$

where $\hat{u}_i$ is the *i*th component of the vector $\hat{u}$, $s = \sqrt{\hat{u}'\hat{u}/(n-p)}$, and q_{ii} is the *i*th diagonal element of the matrix Q. Even with normally distributed errors, although $\tilde{r}_i$ is the familiar ratio of a normally distributed random variable and a square root of a scaled *chi*-square random variable, the end result is not a *t*-distributed random variable. The reason is that the nominator and the denominator are not independent due to the fact that $\hat{u}_i^2 < \hat{u}'\hat{u}$ for all $i = 1,\ldots, n$. Furthermore, because $q_{ii} \geq 1 + 1/n$ (e.g., Cook and Weisberg, 1982: 12), it follows that unlike a *t*-distributed random variable that assumes all real values, $\tilde{r}_i^2 \leq n - p$.

Stefansky (1972), Ellenberg (1973) y Díaz-García y Gutiérrez-Jáimez (2007) obtuvieron la distribución conjunta de un conjunto arbitrario (no singular) de los residuos internamente *Studentizados* definidos en (4). Beckman and Trussell (1974) obtuvieron la distribución del estadístico t para un valor arbitrario $\tilde{r}_i$. Pynnönen (2012) obtuvo la distribución de los resultados referidos para una transformación lineal arbitraria de residuos internamente *Studentizados* de una regresión multivariante con errores elípticos. El presente trabajo obtiene las distribuciones conjuntas de una transformación arbitraria lineal no singular de residuos interna y externamente *Studentizados*, de las que son casos especiales las distribuciones conjuntas de un conjunto de residuos interna y externamente *Studentizados*. El trabajo también muestra que las inferencias relativas a las transformaciones lineales de residuos (*Studentizados*) son, de hecho, un problema de variable omitida en regresión. Como hemos señalado, la mayoría de estos resultados son casos especiales de los presentados en Pynnönen (2012). Sin embargo, debido a que la regresión univariante es un procedimiento de modelización estadística mucho más popular en análisis empíricos aplicados y porque los resultados de la regresión multivariante no siempre pueden trasladarse de forma sencilla al caso univariante, creemos que la discusión de los resultados en el contexto de la regresión múltiple está suficientemente justificada.

2. PRINCIPALES RESULTADOS

Partimos de la siguiente definición para la familia de las distribuciones de contorno esféricos (p.e. Kariya y Eaton, 1977),

Definición 1 Un vector aleatorio u de orden $n \times 1$ sigue una distribución de contorno esférico si

$$Hu \overset{d}{=} u \tag{5}$$

Stefansky (1972), Ellenberg (1973), and Díaz-García and Gutiérrez-Jáimez (2007) derived the joint distribution of an arbitrary (nonsingular) subset of the internally Studentized residuals defined in (4). Beckman and Trussell (1974) derive the distribution for a t-statistic for an arbitrary single $\tilde{r}_i$. Pynnönen (2012) derived the distribution of related results of an arbitrary linear transformation of internally Studentized residuals of multivariate regression with elliptical errors. The present paper derives the joint distributions of an arbitrary non-singular linear transformation of internally and externally Studentized residuals of which the joint distributions of a subset of internally and externally Studentized residuals are special cases. The paper also shows that inference regarding linear transformations of (Studentized) residuals is factually an omitted variable problem in regression. As noted above, most of these results are factually special cases of those given in Pynnönen (2012). However, because univariate regression is vastly more popular statistical modeling devise in applied empirical analyses and because the results from the multivariate regression may not always be straightforward to translate to the univariate counterpart, we think that discussion of the result in the context of the multiple regression is well warranted.

2. MAIN RESULTS

We use the following definition for the family of spherical contoured distributions (e.g. Kariya and Eaton, 1977).

Definition 1 A $n \times 1$ random vector u is spherical contoured distributed if

$$Hu \overset{d}{=} u \tag{5}$$

donde $H \in O(n)$ siendo

$$O(n) = \{H(n \times n) matrix : H'H = I\}$$

el grupo de matrices ortogonales de orden $n \times n$ y "$\overset{d}{=}$" significa que los dos vectores aleatorios siguen la misma distribución. Denotamos la familia de distribuciones esféricas por $S(n)$, y denotamos $u \in S(n)$ para indicar que la distribución de la variable aleatoria u pertenece a la familia de distribuciones esféricas.

Una propiedad importante de las distribuciones esféricas (y más general, de las distribuciones elípticas) es que todas las distribuciones marginales son esféricas (elípticas) (p.e. Muirhead, 1982: 34).

Para el operador "$\overset{d}{=}$" utilizamos el siguiente resultado. Si las variables aleatorias x e y siguen la misma distribución, es decir, $x \overset{d}{=} y$, entonces para cualquier función de Borel f, se verifica que $f(x) \overset{d}{=} f(y)$. Para su comprobación, veáse Anderson y Fang (1990).

Sea

$$C_n = \{x \in \Re^n : \| x \| = 1\} \qquad (6)$$

la esfera unitaria en el espacio euclídeo n-dimensional $\Re^n$, donde $\| x \| = \sqrt{x'x}$ es la norma euclídea. Entonces, la distribución uniforme es la única distribución en C_n que es invariante bajo $O(n)$ (p.e., Kariya y Eaton, 1977).

where $H \in O(n)$ with

$$O(n) = \{H(n \times n) matrix : H'H = I\}$$

the group of $n \times n$ orthogonal matrices, and "$\overset{d}{=}$" means that the two random vectors have the same distributions. We denote the family of spherical distributions by $S(n)$, and denote $u \in S(n)$ to mean that the distribution of the random variable u belongs to the family of spherical distributions.

An important property of spherical distributions (and more generally of elliptical distributions) is that all marginal distributions are spherical (elliptical) (e.g., Muirhead, 1982, p. 34).

For the operator "$\overset{d}{=}$" we utilize the following important result. If random variables x and y have the same distribution, i.e., $x \overset{d}{=} y$ then for any Borel function f, $f(x) \overset{d}{=} f(y)$. For a proof, see Anderson and Fang (1990).

Let

$$C_n = \{x \in \Re^n : \| x \| = 1\} \qquad (6)$$

be the unit sphere in the n-dimensional Euclidian space $\Re^n$, where $\| x \| = \sqrt{x'x}$ is the Euclidian norm. Then the uniform distribution is the unique distribution on C_n that is invariant under $O(n)$ (e.g., Kariya and Eaton, 1977).

Lema 1 Para cualquier $u \in S(n)$ con $P(u=0) = 0$, la variable aleatoria normada

$$\frac{u}{\|u\|} \overset{d}{=} u \tag{7}$$

donde U es un vector aleatorio que sigue una distribución uniforme en la esfera C_n.

Comprobación. Véase Kariya y Eaton (1977, teorema 2.1).

Proposición 1 Una importancia particular del lema 1 es que, debido a que $N(0, I_n)$, la distribución conjunta de n variables aleatorias normales estándar independientes, es un miembro de $S(n)$, la unicidad de la distribución uniforme en la esfera C_n implica que para cualquier $u \in S(n)$, se verifica:

$$\frac{u}{\|u\|} \overset{d}{=} \frac{z}{\|z\|} \tag{8}$$

donde $z \sim N(0, I_n)$. Esto es, podemos estudiar las propiedades de la distribución de varias tranformaciones de los vectores aleatorios normados $u/\|u\|$ con $u \in S(n)$ en términos del ratio $z/\|z\|$, donde $z \sim N(0, I_n)$.

Definición 2 Los residuos normalizados

$$r = \frac{\hat{u}}{\|\hat{u}\|} \tag{9}$$

se denominan residuos normalizados internamente (*INRs: internally normalized residuals*), donde $\hat{u}$, definido en la ecuación (2), es un vector de residuos por mínimos cuadrados de una regresión con errores que siguen una distribución esférica.

Lemma 1 For any $u \in S(n)$ with $P(u = 0) = 0$, the normed random variable

$$\frac{u}{\|u\|} \overset{d}{=} u \tag{7}$$

where U is a uniformly distributed random vector on the sphere C_n.

Proof. See Kariya and Eaton (1977, Theorem 2.1).

Remark 1 A particular importance of Lemma 1 is that because $N(0, I_n)$, the joint distribution of n independent standard normal random variables, is a member of $S(n)$, the uniqueness of the uniform distribution on the sphere C_n implies that for any $u \in S(n)$

$$\frac{u}{\|u\|} \overset{d}{=} \frac{z}{\|z\|} \tag{8}$$

where $z \sim N(0, I_n)$. That is, we can study the distributional properties of various transformations of the normed random vector $u/\|u\|$ with $u \in S(n)$ in terms of the ratio $z/\|z\|$, where $z \sim N(0, I_n)$.

Definition 2 The normalized residuals

$$r = \frac{\hat{u}}{\|\hat{u}\|} \tag{9}$$

are called internally normalized residuals (*INRs*), where $\hat{u}$ defined in equation (2) is a vector of least squares residuals from a regression with spherical distributed errors.

Los siguientes lemas son versiones univariantes de Pynnönen (2012).

Lema 2 (Pynnönen 2012, Lema 1) Bajo el supuesto de que los errores $u \in S(n)$ con $P(u = 0) = 0$ en la regresión $y = X\beta + u$ definida en la ecuación (1)

$$r = \frac{\hat{u}}{\| \hat{u} \|} \overset{d}{=} \frac{v}{\| v \|} \tag{10}$$

donde

$$v = Qz \tag{11}$$

con $z \sim N(0, I_n)$ y $Q = I_n - X(X'X)^{-}X'$ está definido en (3).

Demostración: Utilizando la ecuación (2) y considerando que para una cantidad fija Q la transformación $Q / \| Q \|$ es continua (a.e.) en $\Re^n$ y, por tanto, una función de Borel, tenemos

The following Lemmas are univariate versions of Pynnönen (2012).

Lemma 2 (Pynnönen (2012), Lemma 1) Under the assumption that the errors $u \in S(n)$ with $P(u = 0) = 0$ in the regression $y = X\beta + u$ defined in equation (1),

$$r = \frac{\hat{u}}{\| \hat{u} \|} \overset{d}{=} \frac{v}{\| v \|} \tag{10}$$

where

$$v = Qz \tag{11}$$

with $z \sim N(0, I_n)$ and $Q = I_n - X(X'X)^{-}X'$ is defined in (3).

Proof: Using equation (2) and noting that for fixed Q the transformation $Q / \| Q \|$ continuous (a.e.) on $\Re^n$ and hence a Borel function, we have

$$\frac{\hat{u}}{\| \hat{u} \|} = \frac{Qu}{\| Qu \|} = \frac{Qu / \| u \|}{\| (Qu / \| u \|) \|} \overset{d}{=} \frac{Qz / \| z \|}{\| (Qz / \| z \|) \|} = \frac{Qz}{\| Qz \|} = \frac{v}{\| v \|} \tag{12}$$

donde $z \sim N(0, I_n)$ y $v = Qz$. Esto completa la comprobación del lema.

Proposición 2 El lema 2 muestra explícitamente que *r* definido en la ecuación (9) no depende del parámetro de escala σ_u^2 y, por consiguiente, es una transformación *Studentizada* de los residuos.

Proposición 3 El resultado en el lema 2 es fundamental en el sentido de que todos los problemas inferenciales en la regresión $y = X\beta + u$ que podrían resolverse en términos de funciones (de Borel) de residuos internamente normalizados, aunque no dependiendo de la normalidad de $u \in S(n)$, comparten exactamente las

where $z \sim N(0, I_n)$ and $v = Qz$. This completes the proof of the lemma.

Remark 2 Lemma 2 shows explicitly that *r* defined in (9) does not depend on the scale parameter σ_u^2, and thus indeed is a Studentizing transformation of the residuals.

Remark 3 The result in Lemma 2 is fundamental in the sense that all inference problems in regression $y = X\beta + u$ that can be worked out in terms of (Borel) functions of the internally normalized residuals, although not depending on the normality of $u \in S(n)$,

mismas propiedades estadísticas que si u estuviera normalmente distribuido. Ejemplos son las distribuciones nulas de los estadísticos estándar t y F en regresión (véase también Chmielewski, 1981). Esto se puede ver fácilmente considerando la hipótesis lineal general de la que pueden obtenerse como casos especiales los tests para un coeficiente simple relacionados con el estadístico t. Supongamos, por simplicidad, que X es de rango completo. Expresando la hipótesis lineal general como

$$H_0 : R\beta = q \tag{13}$$

donde R es una matriz conocida de orden $k \times p$ de rango k y q es un vector conocido de orden $k \times 1$. El estadístico F es de la forma (p.e. Johnston y DiNardo, 1997: 97)

$$F = \frac{(\hat{u}_R'\hat{u}_R - \hat{u}'\hat{u})/k}{\hat{u}'\hat{u}/(n-p)} \tag{14}$$

donde $\hat{u}_R$ son los residuos de mínimos cuadrados bajo las restricciones de la hipótesis fijada en (13). Bajo estas restricciones, el estadístico F puede escribirse como

$$F = \frac{u'Q_R u/k}{u'Qu/(n-p)} \tag{15}$$

donde $u \in S(n)$, Q_R es una matriz simétrica idempotente de rango k, Q es la matriz idempotente simétrica de rango $(n-p)$ definida en la ecuación (3) y $Q_R Q = 0$. Por tanto, utilizando el mismo método que en la ecuación (12), obtenemos

$$F \overset{d}{=} \frac{z'Q_R z/k}{z'Qz/(n-p)} \tag{16}$$

share exactly the same statistical properties as if u were normally distributed. Examples are the null distributions of the standard t and F-statistics in regression (see also Chmielewski, 1981). This is easily seen by considering the general linear hypothesis from which single coefficient tests with related t-statistics can be obtained as special cases. Assume for simplicity that X is of full rank. Write the general linear hypothesis as

$$H_0 : R\beta = q \tag{13}$$

where R is a $k \times p$ known matrix of rank k and q is a $k \times 1$ known vector. The F-statistic is of the form (e.g. Johnston and DiNardo, 1997: 97)

$$F = \frac{(\hat{u}_R'\hat{u}_R - \hat{u}'\hat{u})/k}{\hat{u}'\hat{u}/(n-p)} \tag{14}$$

where $\hat{u}_R$ are the least squares residuals under the restrictions of the hypothesis in (13). Under these restrictions the F-statistic can be written in the form

$$F = \frac{u'Q_R u/k}{u'Qu/(n-p)} \tag{15}$$

where $u \in S(n)$, Q_R is a symmetric idempotent matrix of rank k, Q is the symmetric rank $(n-p)$ idempotent matrix given in equation (3), and $Q_R Q = 0$. Thus, using the same method as in equation (12), we obtain

$$F \overset{d}{=} \frac{z'Q_R z/k}{z'Qz/(n-p)} \tag{16}$$

donde $z \sim N(0, I_n)$, que implica que el lado derecho de la ecuación (16) sigue una distribución F con k y $n - p$ grados de libertad. Por tanto, con los errores siguiendo una distribución esférica general, la distribución nula del estadístico F en (14) es $F(k; n - p)$.

El interés de este trabajo se centra en las transformaciones lineales de los residuos internamente normados provenientes de regresiones con errores que siguen una distribución esférica general. Por el lema 2, si M es una matriz $m \times n$, entonces

$$Mr \overset{d}{=} Mv / \| v \| \tag{17}$$

donde $v = Qz$ con $z \sim N(0, I_n)$. Esto es, de nuevo los resultados de la distribución de Mr pueden derivarse en términos de variables aleatorias independientes distribuidas según una normal de media cero y varianza unitaria.

Lema 3 (Pynnönen 2012, Lema 2) Sea M una matriz de orden $m \times n$. Entonces

$$Mr \overset{d}{=} \tilde{M}\tilde{z} / \| \tilde{z} \| \tag{18}$$

donde $\tilde{z} \sim N(0, I_{n-p})$ y

$$\tilde{M} = MH_{n-p} \tag{19}$$

es una matriz de orden $m \times (n - p)$ en la que H_{n-p} es una matriz $n \times (n - p)$ que contiene los vectores propios de los autovalores unitarios de la matriz simétrica idempotente Q, tal que

$$\tilde{M}\tilde{M}' = MQM' \tag{20}$$

Demostración: La comprobación es análoga a la de Pynnönen (2012), Lema 2.

where $z \sim N(0, I_n)$, which implies that the right hand side of (16) has the F-distribution with k and $n - p$ degrees of freedom. Thus, with the general spherical distributed errors, the null distribution of the F-statistic in (14) is $F(k; n - p)$.

The interest of this paper is in the linear transformations of the internally normed residuals stemming from regressions with general spherical distributed errors. By Lemma 2 if M is an $m \times n$ matrix, then

$$Mr \overset{d}{=} Mv / \| v \| \tag{17}$$

where $v = Qz$ with $z \sim N(0, I_n)$. That is, again the distribution results of Mr can be derived in terms of independent zero mean and unit variance normal distributed random variables.

Lemma 3 (Pynnönen (2012), Lemma 2) Let M be an $m \times n$ matrix. Then

$$Mr \overset{d}{=} \tilde{M}\tilde{z} / \| \tilde{z} \| \tag{18}$$

where $\tilde{z} \sim N(0, I_{n-p})$ and

$$\tilde{M} = MH_{n-p} \tag{19}$$

is an $m \times (n - p)$ matrix in which H_{n-p} is an $n \times (n - p)$ matrix containing the eigenvectors of the unit eigenvalues of the symmetric idempotent matrix Q, such that

$$\tilde{M}\tilde{M}' = MQM' \tag{20}$$

Proof: Proof is analogous to that of Pynnönen (2012), Lemma 2.

Podríamos proceder con los resultados de este lema para obtener todos los resultados que siguen. Sin embargo, elegimos utilizar los resultados de este lema más adelante y seguir en términos de la matriz v definida en la ecuación (11). El motivo de esta elección es que algunos de los resultados intermedios que siguen pueden ser de interés en el análisis de los residuos brutos (es decir, no normalizados) de las regresiones con errores normales.

Lema 4 (Pynnönen 2012, Lemma 4) Como en la ecuación (11), sea $v = Qz$, donde $z \sim N(0, I_n)$ y sea M una matriz (no estocástica) de orden $m \times n$ con $r = rank(M) < n - p$, entonces

$$V_M = v'v - v'M'(MQM')^{-}Mv \qquad (21)$$

y

$$U_M = v'M'(MQM')^{-}Mv \qquad (22)$$

se distribuyen independientemente como $\chi^2(n-p-r)$ y $\chi^2(r)$, respectivamente.

Demostración: La comprobación es análoga a la de Pynnönen (2012), Lema 4.

Corolario 1

$$\frac{v'M'(MQM')^{-}Mv}{v'v} \leq 1 \qquad (23)$$

Demostración: Como $V_M \geq 0$ en (21), resulta la desigualdad (23), que completa la comprobación.

Corolario 2 Sea M una matriz de orden $m \times n$, entonces

$$r'M'(MQM')^{-}Mr \leq 1 \qquad (24)$$

We could proceed with the results of this lemma to obtain all the results what follow. However, we choose to utilize the results of this lemma only later and proceed in terms of v defined in equation (11). The motivation of this choice is that some of the intermediate distributional results what follow may be of interest in analysis of raw residuals (i.e., non-normalized) from regressions with normal errors.

Lemma 4 (Pynnönen (2012), Lemma 4) As in equation (11), let $v = Qz$, where $z \sim N(0, I_n)$ and let M be an $m \times n$ (nonstochastic) matrix with $r = rank(M) < n - p$, then

$$V_M = v'v - v'M'(MQM')^{-}Mv \qquad (21)$$

and

$$U_M = v'M'(MQM')^{-}Mv \qquad (22)$$

are independently distributed as $\chi^2(n-p-r)$ and $\chi^2(r)$, respectively.

Proof: Proof is parallel to Pynnönen (2012), Lemma 4.

Corollary 1

$$\frac{v'M'(MQM')^{-}Mv}{v'v} \leq 1 \qquad (23)$$

Proof: Because $V_M \geq 0$ in (21), the inequality (23) follows, which completes the proof.

Corollary 2 Let M be an $m \times n$ matrix, then

$$r'M'(MQM')^{-}Mr \leq 1 \qquad (24)$$

Demostración: Por el lema 2 $V_M \geq 0$, lo que implica que $I_n - M'(MQM')^{-}M)$ es semidefinida positiva. Por tanto,

$$r'(I_n - M'(MQM')^{-}M)r \geq 0$$

Además, por la definición de *r* en (9), *r'r* = 1, lo que implica el resultado del corolario en (24).

Lema 5 Bajo los supuestos del lema 4, V_M y Mv son independientes.

Demostración: Considerando $Mv = MQz$, la ecuación (21) sería $V_M = (\tilde{Q}z)'(\tilde{Q}z)$, donde $\tilde{Q} = Q - QM'(MQM')^{-}MQ$ es una matriz simétrica idempotente. Entonces, la multiplicación directa y las propiedades de las inversas generalizadas implican que $\tilde{Q}QM' = 0$. Esto significa que la variable aleatoria $\tilde{Q}z$ de V_M y MQz están incorreladas, que junto con la normalidad de *z* implica que V_M y Mz son independientes, lo que completa la comprobación del lema.

Con los resultados de los lemas anteriores, podemos obtener los principales resultados de esta sección respecto a las propiedades de la distribución de una transformación lineal de residuos internamente *Studentizados*.

Teorema 1 Suponiendo el modelo de regresión lineal en (1) con errores distribuidos esféricamente, $u \in S(n)$, consideremos una transformación lineal arbitraria

$$r_M = Mr \qquad (25)$$

de residuos internamente normalizados *r* definidos en la ecuación (9), donde M es una matriz de orden $m \times n$ tal que MQM' es definida positiva. Entonces, para

Proof: By Lemma 2 $V_M \geq 0$, which implies that $I_n - M'(MQM')^{-}M)$ is positive semi-definite. Thus,

$$r'(I_n - M'(MQM')^{-}M)r \geq 0$$

and by the definition of *r* in (9), *r'r* = 1, which imply the result in (24) of the corollary.

Lemma 5 Under the assumptions of Lemma 4, V_M and Mv are independent.

Proof. Write $Mv = MQz$ and in (21) $V_M = (\tilde{Q}z)'(\tilde{Q}z)$, where $\tilde{Q} = Q - QM'(MQM')^{-}MQ$ is a symmetric idempotent matrix. Then direct multiplication and the properties of generalized inverses imply $\tilde{Q}QM' = 0$. This implies that the defining random variable $\tilde{Q}z$ of V_M and MQz are uncorrelated, which together with the normality of *z* imply that V_M and Mz are independent, completing the proof of the lemma.

With the results of the above Lemmas we can derive the main results of this section regarding the distributional properties of a linear transformation of internally Studentized residuals.

Theorem 1 Assuming the linear regression model in (1) with spherically distributed errors, $u \in S(n)$, consider an arbitrary linear transformation

$$r_M = Mr \qquad (25)$$

of the internally normalized residuals *r* defined in equation (9), where M is an $m \times n$ matrix such that MQM' is positive definite. Then for $m < n - p$ the joint

$m < n - p$ la distribución conjunta del vector aleatorio r_M de orden $m \times 1$ es

distribution of the $m \times 1$ random vector r_M is

$$f_{r_M}(x) = c_{n-p,m} \mid MQM' \mid^{-1/2} \left(1 - x'(MQM')^{-1}x\right)^{\frac{1}{2}(n-p-m)-1} \tag{26}$$

para $x'(MQM')^{-1}x \leq 1$, donde

for $x'(MQM')^{-1}x \leq 1$, where

$$c_{n-p,m} = \frac{\Gamma[(n-p)/2]}{\pi^{m/2}\Gamma[(n-p-m)/2]} \tag{27}$$

$$c_{n-p,m} = \frac{\Gamma[(n-p)/2]}{\pi^{m/2}\Gamma[(n-p-m)/2]} \tag{27}$$

y $\Gamma(\cdot)$ es la función Gamma.

and $\Gamma(\cdot)$ is the Gamma function.

Demostración: Bajo el supuesto de no singularidad, MQM' es definida positiva, existe la inversa $(MQM')^{-1}$ y $r = m$, es decir, el rango de la matriz es m. El límite, $x'(MQM')^{-1}x \leq 1$, sigue del colorario 2. Por el lema 2, $r = v/\|v\|$, donde v es un vector aleatorio normal. Esto implica que $Mr/\|r\| \overset{d}{=} Mv/\|v\|$ Por tanto, encontrando la distribución de

Proof. Under the nonsingularity assumption MQM' is positive definite, the inverse $(MQM')^{-1}$ exists, and $r = m$, i.e., the rank of the matrix is m. The bounds, $x'(MQM')^{-1}x \leq 1$, follow from Corollary 2. By Lemma 2, $r = v/\|v\|$, in which v is a normal random vector. This implies $Mr/\|r\| \overset{d}{=} Mv/\|v\|$. Thus, finding the distribution of

$$r_M^v = \frac{Mv}{\|v\|} \tag{28}$$

$$r_M^v = \frac{Mv}{\|v\|} \tag{28}$$

obtenemos la distribución de r_M. Con estos resultados, el resto puede seguirse de forma análoga a Ellenberg (1973). Esto es, debido al resultado de la distribución χ^2 de V_M en el lema 4, la normalidad de $v_M = Mv$ y la independencia de v_M y V_M por el lema 5, su función de densidad conjunta es el producto de sus funciones de densidad, que resulta en

gives the distribution of r_M. With these results, the rest can be proceeded parallel to Ellenberg (1973). That is, due to the χ^2-distribution result of V_M in Lemma 4, normality of $v_M = Mv$, and independence of v_M and V_M by Lemma 5, their joint density is the product of their densities, resulting to

$$f_{v_M,V_M}(u,v) = \frac{1}{\pi^{\frac{1}{2}m}2^{\frac{1}{2}(n-p)} \mid MQM' \mid^{\frac{1}{2}} \Gamma[(n-p-m)/2]} v^{\frac{1}{2}(n-p-m)-1} \times \exp\left\{-\frac{1}{2}[u'(MQM')^{-1}u + v]\right\} \tag{29}$$

Definiendo las siguientes transformaciones

$$x = u/\sqrt{y/(n-p)}$$
$$y = u'(MQM')^{-1}u + v \quad (30\text{-}31)$$

el Jacobiano de la transformación es

$$y^{\frac{1}{2}m} \quad (32)$$

Utilizando los resultados anteriores, la función de densidad conjunta de r_M^v y $s = v'v$ resulta

Define next transformations

$$x = u/\sqrt{y/(n-p)}$$
$$y = u'(MQM')^{-1}u + v \quad (30\text{-}31)$$

The Jacobian of the transformation is

$$y^{\frac{1}{2}m} \quad (32)$$

Using these, the joint density of r_M^v and $s = v'v$ becomes

$$\begin{aligned} f_{r_M^v,s}(x,y) &= y^{\frac{1}{2}m} f_{v_M,V_M}\left(x\sqrt{y/(n-p)}, y - x'(MQM')^{-1}x\right) \\ &= \frac{|MQM'|^{-\frac{1}{2}}}{\pi^{\frac{1}{2}m}\Gamma[(n-p-m)/2]} \\ &\times\left(1-x'(MQM')^{-1}x\right)^{\frac{1}{2}(n-p-m)-1} \\ &\times\frac{1}{2^{(n-p)/2}} y^{\frac{1}{2}(n-p)-1} e^{-\frac{1}{2}y} \end{aligned}$$

Integrando respecto a y se obtiene finalmente la función de densidad marginal de r_M^v, que es de la forma definida en la ecuación (26). Esto completa la comprobación del teorema.

El mencionado teorema determina la función de densidad conjunta en el caso de $m < n - p$. En el límite con $m = n - p = \text{rank}(Q)$ se demuestra que la distribución es uniforme (con respecto a una medida de volumen adecuada). Esto resulta porque por el lema 3

$$r_M = Mr \overset{d}{=} \tilde{M}\tilde{z}/\|\tilde{z}\| \quad (33)$$

donde la matriz $\tilde{M}$ es una matriz cuadrada e invertible. Por tanto, puesto que (por el lema 1) $\tilde{z}/\|\tilde{z}\|$ sigue una distribución uniforme en la esfera C_{n-p} y r_M sigue la misma distribución que la transformación lineal unívoca de $\tilde{z}/\|\tilde{z}\|$, tenemos:

Integrating with respect to y yields finally the marginal density of r_M^v, which is of the form in (26). This completes the proof of the theorem.

The above theorem gives the joint density in the case of $m < n - p$. On the borderline with $m = n - p = \text{rank}(Q)$, the distribution proves to be uniform (w.r.t suitable volume measure). This is because by Lemma 3

$$r_M = Mr \overset{d}{=} \tilde{M}\tilde{z}/\|\tilde{z}\| \quad (33)$$

where the matrix $\tilde{M}$ is a square matrix and invertible. Thus, because (by Lemma 1) $\tilde{z}/\|\tilde{z}\|$ is uniformly distributed on the sphere C_{n-p} and r_M has the same distribution as the one-to-one linear transformation of $\tilde{z}/\|\tilde{z}\|$, we have

Teorema 2 Bajo los supuestos del teorema 1, cuando $m = n - p$, Mr se distribuye uniformemente en la esfera $\{x \in \Re^{n-p} : x'(MQM')^{-1}x = 1\}$.

Con estos resultados, también obtenemos los siguientes resultados secundarios relativos a la distribución uniforme en la esfera C_n.

Corolario 3 Sea U un vector aleatorio n-dimensional que sigue una distribución uniforme en la esfera C_n. Entonces, la distribución conjunta de una transformación lineal arbitraria

$$u_M = MU \tag{34}$$

donde M es una matriz $m \times n$ con rango $m < n$ tal que la función de densidad es:

Theorem 2 Under the assumptions of Theorem 1, when $m = n - p$, Mr is uniformly distributed on the sphere $\{x \in \Re^{n-p} : x'(MQM')^{-1}x = 1\}$.

With these results we obtain also the following side results regarding uniform distribution on the sphere C_n.

Corollary 3 Let U be a random n-vector that is uniformly distributed on the sphere C_n. Then the joint distribution of an arbitrary linear transformation

$$u_M = MU \tag{34}$$

where M is an $m \times n$ matrix with rank $m < n$ is such that the density is

$$f_{u_M}(x) = c_{n,m} \mid MM' \mid^{-1/2} \left(1 - x'(MM')^{-1}x\right)^{\frac{1}{2}(n-m)-1} \tag{35}$$

para $x'(MM')^{-1}x \le 1$, $m < n$ y $c_{n,m}$ es el definido en la ecuación (27). Para $m = n$ la distribución es uniforme en la esfera $\{x \in \Re^n : x'(MM')^{-1}x = 1\}$.

Casos particulares son la distribución conjunta de los márgenes m-variables $(m < n)$ U_m de U que se obtienen seleccionando, por ejemplo $M = (I_m : 0_{n-m})$. Esto resulta en

for $x'(MM')^{-1}x \le 1$, $m < n$, and $c_{n,m}$ is defined equation (27). For $m = n$ the distribution is uniform on the sphere $\{x \in \Re^n : x'(MM')^{-1}x = 1\}$.

Particular special cases are the joint distribution of m-variate $(m < n)$ margins U_m of U that are obtained by selecting for example $M = (I_m : 0_{n-m})$. This results to

$$f_{u_m}(x) = \frac{\Gamma[n/2]}{\pi^{m/2}\Gamma[(n-m)/2]}(1 - x'x)^{\frac{1}{2}(n-m)-1} \quad ,, \quad x'x \le 1 \tag{36}$$

obtenido en Eaton (1981: 392). En el siguiente epígrafe discutiremos aplicaciones más concretas relativas a diversos aspectos inferenciales de la regresión basada en MCO. Sin embargo, antes de eso, abordamos algunos resultados relativos a residuos externamente *Studentizados.*

derived for example in Eaton (1981: 392). In the next section we will discuss more concrete applications related to various aspects OLS based regression inference. Before that we, however, deal with some results related to *externally* Studentized residuals by noting first:

Proposición 4. El lema 5 junto con el lema 2 y el resultado en (8) implican que

$$\frac{M\hat{u}}{\sqrt{\hat{V}_M}} \overset{d}{=} \frac{Mv}{\sqrt{V_M}} \tag{37}$$

donde $\hat{V}_M = \hat{u}'\hat{u} - \hat{u}M'(MQM')^{-}M\hat{u}$. El numerador y el denominador en el lado derecho del ratio, $Mv = V_M$, son independientes. Además, de forma análoga a la comprobación del lema 2, es fácil observar que la distribución de $M\hat{u}/\sqrt{\hat{V}_M}$ es de nuevo independiente del parámetro de escala σ_u^2. Así, una definición natural para los residuos externamente *Studentizados* de una transformación lineal $M\hat{u}$ de residuos sería

$$e_M = M\hat{u}/\sqrt{\hat{V}_M} \tag{38}$$

Puesto que en la ecuación (37), el numerador se distribuye normalmente y es independiente de la raíz cuadrada de la variable *Chi*-cuadrado en el denominador, el ratio sigue una distribución que es una constante múltiple de la distribución t multivariante con una matriz de covarianzas proporcional a MQM' y número de grados de libertad, por el lema 4, igual a $n - p - r$ con $r = \text{rank}(M) = n - p$.

3. APLICACIONES

3.1. Distribución conjunta de las clases de residuos *Studentizados*

En primer lugar, notemos que todas las principales clases de residuos definidos en la literatura pueden obtenerse como casos especiales de la ecuación (25). En particular, todos los resultados implícitos no son dependientes de la normalidad de los errores, $u \in S(n)$, del modelo de regresión definido en (1).

Remark 4 Lemma 5 together with Lemma 2 and the result in (8) imply that

$$\frac{M\hat{u}}{\sqrt{\hat{V}_M}} \overset{d}{=} \frac{Mv}{\sqrt{V_M}} \tag{37}$$

where $\hat{V}_M = \hat{u}'\hat{u} - \hat{u}M'(MQM')^{-}M\hat{u}$. The nominator and the denominator in the right hand side ratio, $Mv = V_M$, are independent. Furthermore, paralleling the proof of Lemma 2, it is easy to see that the distribution of $M\hat{u}/\sqrt{\hat{V}_M}$ is again independent of the scale parameter σ_u^2. Thus, a natural definition for externally Studentized residuals of a linear transformation $M\hat{u}$ of the residuals would be

$$e_M = M\hat{u}/\sqrt{\hat{V}_M} \tag{38}$$

Because in (37) the nominator is normally distributed and independent of the square root of the *Chi*-squared variable in the denominator, the ratio has the distribution that is a constant multiple of multivariate *t*-distribution with a covariance matrix proportional to MQM' and degrees of freedom by Lemma 4 equal to $n - p - r$ with $r = \text{rank}(M) = n - p$.

3. APPLICATIONS

3.1. Joint distribution of classes of Studentized residuals

We note first that all the major classes of residuals defined in literature can be obtained as special cases of (25). In particular all the implied results are not dependent on normality of the errors, $u \in S(n)$, of the regression model in (1).

Por ejemplo, consideremos además de los residuos internamente *Studentizados* definidos en la ecuación (4), otras formas de residuos discutidos p.e. en Chatterjee y Hadi (1988) y Lloynes (1979):

Residuos normalizados:

$$\hat{u}_i / \sqrt{\hat{u}'\hat{u}} \qquad (39)$$

Residuos estandarizados:

$$\hat{u}_i / s \qquad (40)$$

donde $s = \sqrt{u'u/(n-p)}$.

Residuos de Abrahamse-Koerts (normalizados):

$$B'\hat{u} / \sqrt{\hat{u}'\hat{u}} \qquad (41)$$

donde B es una matriz de orden $n \times n$ definida en Abrahamse and Koerts (1971), cumpliendo $\hat{u}'BB'\hat{u} = \hat{u}'\hat{u}$.

Consideremos un subconjunto arbitrario linealmente independiente $I_m = \{i_1, i_2, \ldots, i_m\} \subset \{1, \ldots, n\}$, $m \le n$ de las clases anteriores de residuos. Puede observarse fácilmente que cada uno de ellos es un caso especial de la transformación lineal definida en la ecuación (25). Para nuestro objetivo, definamos M_I como una matriz de orden $m \times n$ en la que cada fila $j = 1, \ldots, m$ es un vector de orden $1 \times n$ con elemento $i_j = 1$ y ceros en el resto, $i_j \in I_m$. Además, denotemos $D^{-1/2}$ a una matriz diagonal de orden $n \times n$ con elementos $(q_{11})^{-1/2}, \ldots, (q_{nn})^{-1/2}$. Entonces, un conjunto, I_m, de residuos internamente *Studentizados*, definidos en la ecuación (4), se obtiene definiendo en la ecuación (25), $M = (n-p)^{1/2} D^{-1/2} M_I$; un conjunto de residuos normalizados, definidos en la ecuación (39), se obtiene definiendo $M = M_I$; un conjunto de residuos

For example, consider in addition to the internally Studentized residuals defined in equation (4), other forms of residuals discussed e.g. in Chatterjee and Hadi (1988) and Lloynes (1979):

Normalized residuals:

$$\hat{u}_i / \sqrt{\hat{u}'\hat{u}} \qquad (39)$$

Standardized residuals:

$$\hat{u}_i / s \qquad (40)$$

where $s = \sqrt{u'u/(n-p)}$.

Abrahamse-Koerts residuals (normalized):

$$B'\hat{u} / \sqrt{\hat{u}'\hat{u}} \qquad (41)$$

where B is an $n \times n$ matrix defined in Abrahamse and Koerts (1971), satisfying $\hat{u}'BB'\hat{u} = \hat{u}'\hat{u}$.

Consider an arbitrary linearly independent subset $I_m = \{i_1, i_2, \ldots, i_m\} \subset \{1, \ldots, n\}$, $m \le n$ of the above classes of residuals. It is easily seen that each of these is a special case of the linear transformation defined in equation (25). For the purpose, define M_I as an $m \times n$ matrix in which each row $j = 1, \ldots, m$ is a $1 \times n$ vector with element $i_j = 1$ and zeros elsewhere, $i_j \in I_m$. Furthermore, let $D^{-1/2}$ denote an $n \times n$ diagonal matrix with elements $(q_{11})^{-1/2}, \ldots, (q_{nn})^{-1/2}$. Then a set, I_m, of internally Studentized residuals, defined in equation (4), is obtained by defining in equation (25), $M = (n-p)^{1/2} D^{-1/2} M_I$, a set of normalized residuals, defined in equation (39), is obtained by defining $M = M_I$, a set of standardized residuals, defined in equation (40), is obtained by defining $M = (n-p)^{1/2} M_I$, and a set of

estandarizados, definidos en la ecuación (40), se obtiene definiendo $M = (n-p)^{1/2} M_I$; y un conjunto de residuos de Abrahamse-Koerts, definidos en la ecuación (41), se obtiene definiendo $M = M_I B'$.

3.2. Inferencia estadística

3.2.1. Variables omitidas

En particular, seleccionando $m = 1$ de forma que M se convierte en un vector fila, el teorema 1 implica que la función de densidad de la distribución de una combinación lineal (no singular) arbitraria $r_m = m'\hat{u} / \| \hat{u} \|$, donde m es un vector de orden $n \times 1$ de números reales que verifican $m'Qm > 0$, es

$$f_{r_m}(x) = \frac{\Gamma[(n-p)/2](m'Qm)^{-1/2}}{\Gamma[(n-p-1)/2]\Gamma[1/2]} \left(1 - \frac{x^2}{mQ'm}\right)^{\frac{1}{2}(n-p-1)-1} \tag{42}$$

para $|x| \leq \sqrt{m'Qm}$. Por consiguiente, $r_m^2 / m'Qm$ sigue una distribución Beta con parámetros 1/2 and $(n - p - 1)/2$. Fijando el componente i-ésimo en m igual a 1 y el resto de elementos igual a cero, lleva a la función de densidad para un simple residuo internamente *Studentizado* $\tilde{r}_i = \hat{u}_i / s\sqrt{q_{ii}}$ definido en la ecuación (4) donde $s = \sqrt{\hat{u}\hat{u}/(n-p)}$,

$$f_{r_i}(r) = \frac{\Gamma[(n-p)/2]}{\Gamma[(n-p-1)/2]\Gamma[1/2]\sqrt{n-p}} \left(1 - \frac{r^2}{n-p}\right)^{\frac{1}{2}(n-p-1)-1} \tag{43}$$

$|r| \leq \sqrt{(n-p)}$.

De nuevo, $r_i^2/(n-p)$ sigue una distribución Beta con parámetros 1/2 and $(n - p - 1)/2$. Por tanto, las distribuciones de residuos simples internamente *Studentizados* y sus

Abrahamse-Koerts residuals are obtained, defined in equation (41), by defining $M = M_I B'$.

3.2. Statistical inference

3.2.1. Omitted variables

In particular, selecting $m = 1$ such that M becomes a row vector, Theorem 1 implies that the density function of the distribution of an arbitrary (nonsingular) linear combination $r_m = m'\hat{u} / \| \hat{u} \|$, where m is an $n \times 1$ vector of real numbers satisfying $m'Qm > 0$, is

for $|x| \leq \sqrt{m'Qm}$. Thus, $r_m^2 / m'Qm$ follows a Beta distribution with parameters 1/2 and $(n - p - 1)/2$. Setting the ith component in m equal to 1 and all others equal to zero gives the density function for a single internally Studentized residual $\tilde{r}_i = \hat{u}_i / s\sqrt{q_{ii}}$ defined in equation (4), where $s = \sqrt{\hat{u}\hat{u}/(n-p)}$,

Again, $r_i^2/(n-p)$ follows a Beta distribution with parameters 1/2 and $(n - p - 1)/2$. Thus, the distributions of single internally Studentized residuals and their arbitrary

combinaciones lineales arbitrarias pertenecen a la misma familia de distribuciones de forma que a través de simples transformaciones están idénticamente distribuidos como una distribución Beta con parámetros 1/2 and $(n-p-1)/2$. Como es obvio, esto facilita las inferencias basadas en residuos individuales o en sus combinaciones lineales. Además, como se muestra más adelante, la situación puede resultar más fácil introduciendo una transformación que lleva a una ditribución *t* común.

De hecho, si *studentizamos* $m'\hat{u}$ de la misma forma que $\tilde{r}_i$ en la ecuación (4) definiendo $\tilde{r}_m = m'\hat{u}/s\sqrt{q_m}$, donde $s=\sqrt{\hat{u}'\hat{u}/(n-p)}$ y $q_m = m'Qm$, podemos escribir

linear combinations belong to the same family of distributions such that through simple transformations they are identically distributed as the Beta distribution with parameters 1/2 and $(n-p-1)/2$. This obviously facilitates inference based on individual residuals or their linear combinations. Moreover, as is shown below, the situation can be further facilitated by introducing a transformation that leads to a common *t*-distribution.

In fact, if we studentize $m'\hat{u}$ in a manner of $\tilde{r}_i$ in equation (4) by defining $\tilde{r}_m = m'\hat{u}/s\sqrt{q_m}$, where $s=\sqrt{\hat{u}'\hat{u}/(n-p)}$ and $q_m = m'Qm$, we can write

$$t_m = \tilde{r}_m\left(\frac{n-p-1}{n-p-\tilde{r}_m^2}\right)^{\frac{1}{2}} = \frac{m'\hat{u}/\sqrt{m'Qm}}{\sqrt{\hat{V}_m/(n-p-1)}} \overset{d}{=} \frac{m'v/\sqrt{m'Qm}}{\sqrt{V_m/(n-p-1)}} \qquad (44)$$

donde, $\hat{V}_m = \hat{u}'\hat{u}-(m'\hat{u})^2/(m'Qm)$, $V_m = v'v-(m'v)^2/(m'Qm)$ y $v = Qz$ con $z \sim N(0, I_n)$. Así, t_m es una variable aleatoria que sigue una distribución *t* con $n-p-1$ grados de libertad en virtud de los lemas 4 y 5 y que $m'v/\sqrt{m'Qm} \sim N(0,1)$. De nuevo, un caso especial es un residuo simple, en cuyo caso la segunda relación en la ecuación (44) puede escribirse como

where $\hat{V}_m = \hat{u}'\hat{u}-(m'\hat{u})^2/(m'Qm)$, $V_m = v'v-(m'v)^2/(m'Qm)$ and $v = Qz$ with $z \sim N(0, I_n)$. Thus, t_m is a *t*-distributed random variable with $n-p-1$ degrees of freedom by virtue of Lemma 4, Lemma 5, and that $m'v/\sqrt{m'Qm} \sim N(0,1)$. Again, a special case of this is a single residual, in which case the second relation in the middle of (44) can be written as

$$t_i = \frac{\hat{u}_i}{s_{(i)}\sqrt{q_{ii}}} \qquad (45)$$

donde

$$t_i = \frac{\hat{u}_i}{s_{(i)}\sqrt{q_{ii}}} \qquad (45)$$

where

$$s_{(i)}^2 = \frac{(n-p)s^2-\hat{u}_i^2/q_{ii}}{n-p-1} \qquad (46)$$

es la media cuadrática residual de una muestra de la que se ha eliminado la observación *i*-ésima de la regresión (p.e.,

$$s_{(i)}^2 = \frac{(n-p)s^2-\hat{u}_i^2/q_{ii}}{n-p-1} \qquad (46)$$

is the residual means square from a sample with the *i*th observation removed from the regression (e.g., Beckman and Trussell,

Beckman y Trussell, 1974). Por la ecuación (44), la relación entre el residuo externamente *Studentizado*, t_i, y el residuo internamente *Studentizado*, $\tilde{r}_i$, definido en la ecuación (4), es

$$t_i = \tilde{r}_i \left(\frac{n-p-1}{n-p-\tilde{r}_i^2} \right)^{\frac{1}{2}} \qquad (47)$$

(c.f.r. Cook y Weisberg 1982: 20). Los residuos individuales se utilizan normalmente como estadísticos de diagnóstico para la comprobación del modelo, buscando datos atípicos y observaciones influyentes (véase Cook y Weisberg, 1982). El resultado final es que una función lineal (no singular) arbitraria de residuos internamente *Studentizados* puede transformarse en un estadístico t con $n - p - 1$ grados de libertad. Por tanto, utilizando la sencilla transformación recogida en (44), más que confiando en la distribución Beta, puede utilizarse la conocida distribución t en las correspondientes inferencias estadísticas.

Si m en la definición de $\widetilde{r}_m$ en el estadístico definido en (44) es un vector de observaciones de la variable omitida en la regresión $y = X\beta + u$, t_m es un estadístico t para contrastar la significación de la omisión. Esto puede generalizarse inmediatamente a matrices de forma que las transformaciones lineales de los residuos normalizados pueden utilizarse para contrastar variables omitidas en una regresión. De forma más precisa, denotemos como Z m variables adicionales de la regresión en (1) tal que

$$y = X\beta + Z\gamma + e \qquad (48)$$

donde γ es un vector m-dimensional con los coeficientes de pendiente adicionales y $e \in S(n)$.

La hipótesis nula a contrastar es:

$$H_0 : \gamma = 0 \qquad (49)$$

1974). By equation (44), the relationship between the externally Studentized residual, t_i, and the internally Studentized residual, $\tilde{r}_i$, defined in equation (4) is

$$t_i = \tilde{r}_i \left(\frac{n-p-1}{n-p-\tilde{r}_i^2} \right)^{\frac{1}{2}} \qquad (47)$$

(c.f. Cook and Weisberg (1982), p. 20). Individual residuals are typically used as diagnostic tools for the model checking, testing for outliers and influential observations (see, Cook and Weisberg, 1982). The end result is that an arbitrary (nonsingular) linear function of internally Studentized residuals can be transformed to a t-statistic with $n - p - 1$ degrees of freedom. Thus, utilizing the simple transformation in (44), rather than relying on the Beta distribution the familiar t-distribution can be used instead in the related statistical inference.

If m in the definition of $\widetilde{r}_m$ in statistic (44) is an observation vector of an omitted variable from the regression $y = X\beta + u$, t_m is a t-statistic for testing the significance of the omission. This generalizes immediately to matrices such that linear transformations of the normalized residuals can be utilized as such for testing omitted variables from a regression. More precisely, let Z denote m additional variables of the regression in (1) such that

$$y = X\beta + Z\gamma + e \qquad (48)$$

where γ is an m-vector of the additional slope coefficients and $e \in S(n)$.

The null hypothesis to be tested is

$$H_0 : \gamma = 0 \qquad (49)$$

Dados los residuos $\tilde{u}$ procedentes de la regresión $y = X\beta + u$, el contraste de la hipótesis (49) puede basarse en la transformación lineal definida en (25). Eso es así porque, en el caso general, por el lema 2

Given residuals $\tilde{u}$ from the regression $y = X\beta + u$, the testing for hypothesis (49) can be based on linear transformation in (25). This is because, in the general case, by Lemma 2

$$t_M^2 = \frac{r'M'(MQM')^{-1}Mr/m}{(r'r - r'M'(MQM')^{-1}Mr)/(n-p-m)} \overset{d}{=} \frac{v'M'(MQM')^{-1}Mv/m}{v'(I - M'(MQM')^{-1}M)v/(n-p-m)} \tag{50}$$

donde las normas $\|v\|$ se han anulado. Por el lema 4 el último ratio es un cociente entre dos variables aleatorias independientes que siguen una distribución *chi*-cuadrado con m y $n-p-m$ grados de libertad, respectivamente, implicando que t_M^2 sigue una distribución F con m y $n-p-m$ grados de libertad. Esto es,

where the norms $\|v\|$ have been canceled out. By Lemma 4 the last ratio is a ratio of two independent *chi*-squared random variables with degrees of freedom m and $n-p-m$, respectively, implying that t_M^2 is F-distributed with m and $n-p-m$ degrees of freedom. That is,

$$t_M^2 \sim F(m, n-p-m) \tag{51}$$

Por tanto, fijando $M = Z'$, y notando que las normas $\|\hat{u}\|$ vuelven a anularse, obtenemos

Thus, setting $M = Z'$, and noting that the norms $\|\hat{u}\|$, again cancel out, we obtain

$$t_Z^2 = \frac{\hat{u}'Z(Z'QZ)^{-1}Z'\hat{u}/m}{\hat{u}'(I - Z(Z'QZ)^{-1}Z')\hat{u}/(n-p-m)} \tag{52}$$

que, por la ecuación (50), sigue una distribución F con m y $n-p-m$ grados de libertad bajo la hipótesis nula. Utilizando álgebra de matrices resulta que:

which by (50) is F-distributed with m and $n-p-m$ degrees of freedom under the null hypothesis. Using little matrix algebra shows that

$$t_Z^2 = \frac{(\hat{u}'\hat{u} - \hat{e}'\hat{e})/m}{\hat{e}'\hat{e}/(n-p-m)} \tag{53}$$

es decir, el habitual estadístico F para contrastar variables omitidas, donde $\hat{e}$ es el vector de residuos por MCO de la regresión en (48).

i.e., the usual F-statistic for testing omitted variables, where $\hat{e}$ is the vector of OLS residuals from regression (48).

Es de notar que todas las hipótesis lineales del tipo $R\beta = q$ definido en la ecuación (13) pueden convertirse de nuevo en el problema de variables omitidas con el estadístico F definido en (52). Así, los resultados desarrollados con anterioridad muestran desde otro ángulo la robustez de los habituales estadísticos *t* y F en el análisis de regresión, en el sentido de que sus distribuciones nulas son independientes de la normalidad de $u \in S(n)$.

3.2.2. Medidas de sensibilidad

Díaz-García *et al*. (2007) proponen varias generalizaciones de las distancias de Cook multivariantes para detectar una o más observaciones influyentes. A continuación demostramos las posibilidades de utilizar los resultados obtenidos en este trabajo en dicha dirección.

Sea $y = X\beta + u$ la regresión inicial como está definida en la ecuación (1) y consideremos la regresión

$$y = X\beta + m\gamma + e \qquad (54)$$

donde m es el vector de observaciones de una simple variable explicativa adicional. Entonces, podría ser de interés mediar la influencia de la omisión de la variable m en la estimación de β. Una medida apropiada es la distancia de Cook.

En lo que sigue podemos suponer sin pérdidad de generalidad que en la regresión definida en (1) $\text{rank}(X) = p$ e $y \sim N(0, \sigma_u^2 I_n)$. Utilizando entonces los resultados del epígrafe 3.2.1 y siguiendo las deducciones de Díaz-García *et al*. (2007) hasta la ecuación (10), obtenemos la distancia de Cook

$$D_m = \frac{1}{qs^2}(\hat{\beta} - \hat{\beta}_{(m)})'(X'X)(\hat{\beta} - \hat{\beta}_{(m)}) \quad (55)$$

It is notable that all linear hypotheses of the form $R\beta = q$ defined in equation (13) can be returned to the above omitted variables problem with F-statistic of the form (52). Thus the above results show from another angle the robustness of the usual t and F-statistics in regression analysis in the sense that their null-distributions are independent of the normality of $u \in S(n)$.

3.2.2. Sensitivity measures

Díaz-García *et al*. (2007) propose several generalizations to multivariate Cook's distances to detect one or more influential observations. Below we demonstrate possibilities to utilize results derived in this paper in that direction.

Let the initial regression be $y = X\beta + u$ as defined in equation (1) and consider the regression

$$y = X\beta + m\gamma + e \qquad (54)$$

where m is the observation vector of a single additional explanatory variable. Then it might be of interest to measure the influence of the omission of variable m on the estimate of β. An appropriate measure is the Cook distance.

In what follows we can assume without loss of generality that in regression (1) $\text{rank}(X) = p$ and $y \sim N(0, \sigma_u^2 I_n)$. Utilizing then results in Section 3.2.1 and following the derivations up to equation (10) in Díaz-García *et al*. (2007), we get the Cook distance

$$D_m = \frac{1}{qs^2}(\hat{\beta} - \hat{\beta}_{(m)})'(X'X)(\hat{\beta} - \hat{\beta}_{(m)}) \quad (55)$$

donde $\hat{\beta}$ es el estimador MCO de β en la regresión inicial con m omitido; $\hat{\beta}_{(m)}$ es el estimador MCO de β cuando m está incluido, y $s^2 = \hat{u}'\hat{u}/(n-p) = \| \hat{u} \|^2 /(n-p)$ es la varianza residual de la regresión incial con m eliminado. Por tanto, D_m mide la influencia o sensibilidad de la omisión en las estimaciones de β. Con el resultado para matrices particionadas, obtenemos una expresión equivalente (Díaz-García *et al*. 2007, ecuación (9)) a la diferencia explícita de las estimaciones debida a la omisión como

$$\hat{\beta} - \hat{\beta}_{(m)} = (X'X)^{-1} X'mm'\hat{u} / q_m \qquad (56)$$

con $q_m = m'Qm = (I_n - X(X'X)^{-1}X')m$. Sustituyendo (56) en (55) y operando resulta:

where $\hat{\beta}$ is the OLS estimator of β from the initial regression with m omitted, $\hat{\beta}_{(m)}$ is the OLS estimator of β when m is included, and $s^2 = \hat{u}'\hat{u}/(n-p) = \| \hat{u} \|^2 /(n-p)$ is the residual variance from the initial regression with m omitted. Thus, D_m measures the influence or sensitivity of the omission on the estimates of β. With result for partitioned matrices, we obtain an analog to (Díaz-García *et al*. (2007), equation (9)) the explicit difference of the estimates due to the omission as

$$\hat{\beta} - \hat{\beta}_{(m)} = (X'X)^{-1} X'mm'\hat{u} / q_m \qquad (56)$$

with $q_m = m'Qm = (I_n - X(X'X)^{-1}X')m$. Using (56) in (55) and arranging terms yields

$$D_m = \frac{(n-p)}{qq_m^2} m' \frac{\hat{u}}{\| \hat{u} \|} \frac{\hat{u}'}{\| \hat{u} \|} mm'X(X'X)^{-1}X'm \qquad (57)$$

$$= \frac{(n-p)(m'm - q_m}{qq_m^2} q_m^2 \qquad (58)$$

donde $r_m = m'\hat{u} / \| \hat{u} \|$. Se observa inmediatamente que si m es ortogonal a las columnas de X, $q_m = m'm$ y $D_m = 0$. Utilizando la expresión (42) y su relación con la distribución Beta presentada en la página 102, tenemos:

where $r_m = m'\hat{u} / \| \hat{u} \|$. It is immediately observed that if m is orthogonal to columns of X, $q_m = m'm$ and $D_m = 0$. Utilizing (42) and its relation to the Beta distribution discussed on page 102, we get

$$D_m \sim a_m Beta\left(\frac{1}{2}, \frac{n-p-1}{2}\right) \qquad (59)$$

donde

$$a_m = \frac{(n-q)(m'm - q_m)}{qq_m} \qquad (60)$$

Si m es un vector de coordenadas con valor uno en la posición *i*-ésima y cero en el resto, obtenemos la distancia de Cook, una versión multivariante de la propuesta y debatida en Díaz-García *et al*. (2007).

$$D_m \sim a_m Beta\left(\frac{1}{2}, \frac{n-p-1}{2}\right) \qquad (59)$$

where

$$a_m = \frac{(n-q)(m'm - q_m)}{qq_m} \qquad (60)$$

If m is a coordinate vector with one in the *i*th position and zeros elsewhere we obtain the Cook distance, a multivariate version of which is proposed and discussed in Díaz-García *et al*. (2007).

3.3. Algunas aplicaciones especiales

Los resultados de este trabajo respecto a las distribuciones pueden aportar mayor comprensión en el campo de estudios de eventos en economía financiera (una excelente revisión puede consultarse en Campbell, Lo, and MacKinlay, 1997, Cap. 4). El tradicional enfoque paramétrico está basado en residuos *Studentizados* externos. Sin embargo, un enfoque no paramétrico basado en sumas de rango del tipo Wilcoxon, sugerido por Corrado (1989) para contrastar las rentabilidades de un único periodo, y ampliado en Campbell y Wasley (1993) para contrastar las sumas de rango multi-día acumuladas, está incrementando su popularidad. Los estadísticos de contraste utilizados son del tipo de residuos internamente *Studentizados*, donde el numerador y el denominador no son independientes. Por consiguiente, la teoría sobre las distribuciones desarrollada en este trabajo puede utilizarse fácilmente en el examen de las propiedades de las distribuciones de los estadísticos de contraste respectivos. Kolari y Pynnönen (2011) desarollan un estudio de evento para contrastar la estrategia de rentabilidades anormales acumuladas con contrastes de rango, donde la teoría de la distribución asintótica se obtiene utilizando los resultados de variables aleatorias internamente normalizadas. Luoma y Pynnönen (2010) y Luoma (2011) son otros ejemplos, donde los resultados anteriores se utilizan en la obtención de distribuciones asintóticas de las versiones de contrastes de rango y signo discutidos en sus trabajos.

4. CONCLUSIONES

Este trabajo obtiene la distribución conjunta de una transformación lineal general de residuos internamente

3.3. Some special applications

The distribution results of this paper can give additional insight in the field of event studies in financial economics (an excellent review is in Campbell, Lo, and MacKinlay, 1997, Ch. 4). The traditional parametric approach is based on external Studentized residuals. However, a non-parametric approach based on Wilcoxon-type rank sums, suggested by Corrado (1989) for testing single period returns, and extended in Campbell and Wasley (1993) for testing cumulative multi-day rank sums, is gaining increasingly popularity. The used test statistics are obviously of the type of internally Studentized residuals, where the nominator and denominator are not independent. Thus, the distribution theory developed in this paper can be readily utilized in examination of the distributional properties of the related test statistics. Kolari and Pynnönen (2011) develop an event study testing strategy of cumulative abnormal returns with rank tests, where the asymptotic distribution theory is derived by utilizing the results of internally normalized random variables. Luoma and Pynnönen (2010) and Luoma (2011) are other examples, where the above results are utilized in deriving the asymptotic distributions of the versions of rank and sign tests discussed in their papers.

4. CONCLUSIONS

This paper derives the joint distribution of a general linear transformation of internally Studentized residual from a general linear regression. Other types of residuals, commonly used in practical applications, can be easily obtained as special cases by

Studentizados procedentes de una regresión lineal general. Otros tipos de residuos, utilizados comúnmente en aplicaciones prácticas, pueden obtenerse fácilmente como casos especiales definiendo de forma apropiada la transformación lineal.

Las distribuciones de conjuntos arbitrarios además de las distribuciones marginales de residuos simples se obtienen como casos especiales de la distribución general mediante la definición de transformaciones lineales de un modo adecuado. Este trabajo también discute algunas aplicaciones potenciales en las que los resultados pueden utilizarse de forma inmediata.

defining the linear transformation appropriately.

The distributions of arbitrary subsets as well as marginal distributions of single residuals are obtained as special cases from the general distribution by defining linear transformation in a suitable manner. The paper discusses also some potential applications in which the results can be readily applied.

BIBLIOGRAFÍA/REFERENCES

Abrahamse, A.P.J. y Koerts, J. (1971). New estimators in regression analysis. *Journal of the American Statistical Association, 66*, 71-74.

Anderson, T.W. y Kai-Tai Fang (1990). On the theory of multivariate elliptically contoured distributions and their applications. In T.W. Anderson & Kai-Tai Fang (Eds.), *Statistical inference in elliptically contoured and related distributions* (pp. 1-23). New York: Allerton Press Inc.

Beckman, R.J. y Trussell, H.J. (1974). The distribution of an arbitrary studentized residual and the effects of updating in multiple regression. *Journal of the American Statistical Association, 69*, 199-201.

Campbell, C.J. y Wasley, C.E. (1993). Measuring security price performance using NASDAQ returns. *Journal of Financial Economics, 33*, 73-92.

Campbell, J.Y., Lo, A.W. y Craig MacKinlay, A. (1997). *The econometrics of financial markets*. Princeton, NJ: Princeton University Press.

Chatterjee, S. y Hadi, A.S. (1988). *Sensitivity analysis in linear regression*. New York: Wiley.

Chmielewski, A.K. (1981). Elliptically symmetric distributions: A review and bibliography. *International Statistical Review, 49*, 67-74.

Corrado, C.J. (1989). A nonparametric test for abnormal security price performance in event studies. *Journal of Financial Economics, 23*, 385-395.

Cook, D.R. y Weisberg, S. (1982). *Residuals and influence in regression*. Chapman and Hall.

Díaz-García, J.A. y Gutiérrez-Jáimez, R. (2006). The distribution of the residual from a general elliptical multivariate linear model. *Journal of Multivariate Analysis, 97*, 1829-1841.

Díaz-García, J.A. y Gutiérrez-Jáimez, R. (2007). The distribution of residuals from a general elliptical linear model. *Journal of Statistical Planning and Inference, 137*, 2347-2354.

Díaz-García, J.A., Gutiérrez-Jáimez, R. y Alvarado-Castro, V.M. (2007). Exact distributions for sensitivity analysis in linear regression. *Applied Mathematical Sciences, 1*, 1083-1100.

Eaton, M.L. (1981). On the projections of isotropic distributions. *Annals of Statistics, 9*, 391-400.

Ellenberg, J.K. (1973). The joint distribution of the standardized least squares residuals from a general linear regression. *Journal of the American Statistical Association, 68*, 941-943.

Johnston, J. y DiNardo, J. (1997). *Econometric methods* (4th ed.). Singapore: McGraw-Hill.

Kariya, T. y Eaton, M.L. (1977). Robust tests for spherical symmetry. *Annals of Statistics, 5*, 206-215.

Kolari, J. y Pynnönen, S. (2011). Nonparametric rank tests for event studies. *Journal of Empirical Finance, 18*, 953-971.

Luoma, T. (2011). *Nonparametric event study tests for testing cumulative abnormal returns*. Acta Wasaensia.

Luoma, T. y Pynnönen, S. (2010). Testing for cumulative abnormal returns in event studies with the rank tests, *Working Paper*, University of Vaasa (Submitted).

Lloynes, R.M. (1979). A note on Prescott's upper bound for normed residuals. *Biometrika, 66*, 387-389.

Margolin, B.H. (1977). The distribution of internally studentized statistics via Laplace transform inversion. *Biometrika, 64*, 573-582.

Muirhead, R.J. (1982). *Aspects of multivariate analysis*. New York: Wiley.

Pynnönen, S. (2012). Distribution of an arbitrary linear transformation of internally Studentized residuals of multivariate regression with elliptical errors. *Journal of Multivariate Analysis, 107*, 40-52.

Stefansky, W. (1972). Rejecting outliers in fractional designs. *Technometrics*, *14*, 469-479.