Revista Electrónica de Comunicaciones y Trabajos de ASEPUMA. **Rect@**
Volumen 17 (2016), páginas 167 a 177.

EL PROBLEMA DE UN TAMAÑO MUESTRAL PEQUEÑO EN LA REGRESIÓN LINEAL: MICRONUMEROSIDAD

ROMÁN SALMERÓN

romansg@ugr.es
Departamento de Métodos Cuantitativos para la Economía y la Empresa
Universidad de Granada.

VÍCTOR BLANCO

vblanco@ugr.es
Departamento de Métodos Cuantitativos para la Economía y la Empresa
Universidad de Granada

Recibido (05/05/2016)
Revisado (12/11/2016)
Aceptado (22/12/2016)

RESUMEN: El económetra Arthur Goldberger introdujo (sarcásticamente) el término micronumerosidad argumentando que los textos sobre Econometría dedican varias páginas al problema de multicolinealidad pero no comentan nada sobre el problema análogo que surge al estimar con un tamaño muestral pequeño/reducido. Por tanto, podríamos referirnos a la micronumerosidad como la multicolinealidad debida a mustras pequeñas. Puesto que su origen es muy concreto su tratamiento también debe serlo, por lo que en este trabajo se obvian las soluciones tradicionales a la multicolinealidad proponiendo otra basada en la peculiaridad del problema tratado.

Palabras Clave: Micronumerosidad, multicolinealidad, muestras pequeñas, regresión múltiple.

ABSTRACT: The econometrician Arthur Goldberg introduced the notion of micronumerosity motivating that classical Econometrics textbooks used to explain the problem of multicollinearity but nothing is explain about the analogous problem of estimating using an small size sample. Then, micronumerosity refers to multicollinearity because of small samples. Since its origins are very particular, its treatment should also be specific. In this paper we obviate standard multicollinearity solutions and we propose a new scheme based on the specific charasteristics of the problem.

Keywords: Micronumerosity, multicollinearity, small samples, multiple regression.

1. Introducción

El modelo de regresión múltiple es una de las técnicas más usadas cuando se desea establecer relaciones lineales entre variables. Así, dado el siguiente modelo de regresión lineal con n observaciones y p variables exógenas:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{1}$$

donde $\mathbf{u}$ representa a la perturbación aleatoria (que se presupone centrada y esférica), el objetivo es estimar los coeficientes de las variables exógenas, $\boldsymbol{\beta}$, para a partir de los valores obtenidos establecer el sentido de las relaciones (con el signo) y cuantificar las mismas (con el número).

Para obtener las estimaciones comentadas, la estimación por el método de Mínimos Cuadrados Ordinarios (MCO) nos proporciona la expresión del estimador: $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$, por lo que es necesario que exista la inversa de la matriz $\mathbf{X}^t\mathbf{X}$, es decir, se asume la independencia lineal entre las variables exógenas presentes en $\mathbf{X}$. Cuando esta condición no se verifica se dice que en el modelo hay multicolinealidad. Por tanto, la multicolinealidad describe la situación de ausencia de ortogonalidad entre las variables independientes del modelo de regresión.

Si la multicolinealidad es exacta, lo cual ocurre cuando una de las variables exógenas es combinación lineal exacta de algunas o todas las demás, no es posible obtener la inversa de la matriz $\mathbf{X}^t\mathbf{X}$ (puesto que $det(\mathbf{X}^t\mathbf{X}) = 0$) y, en tal caso, el objetivo marcado sería inalcanzable ya que no existiría una estimación única para $\boldsymbol{\beta}$. Si la multicolinealidad es aproximada, lo cual ocurre cuando una de las variables exógenas es aproximadamente igual a una combinación lineal de las restantes, sí es posible calcular dicha inversa. Sin embargo, la estimación realizada será inestable ya que se pueden obtener coeficientes estimados sensibles a pequeños cambios en los datos, con la obtención de signos y magnitudes erróneas. Por tanto, este caso es problemático, pues pone en entredicho las conclusiones obtenidas en el análisis realizado, y puesto que los estimadores pueden obtenerse con la expresión de MCO, no es sencillo detectar la existencia del problema.

Entre las causas que pueden provocar multicolinealidad en un modelo de regresión lineal se tiene el mal condicionamiento de los datos usados constituyendo un problema puramente numérico. Así, por ejemplo, Spanos y McGuirk (2002) etiquetan esta situación como multicolinealidad errática, dentro de la cual podría incluirse el término de micronumerosidad, distinguiendo entre micronumerosidad exacta (caso en el que $n = 0$) y micronumerosidad aproximada (si la condición $n > 0$ es apenas satisfecha, en adelante, micronumerosidad).

Nótese que la varianza de los coeficientes estimados responde a la siguiente expresión:

$$\widehat{\text{var}}\left(\widehat{\beta}_i\right) = \frac{\widehat{\sigma}^2}{n \cdot \text{var}(\mathbf{X}_i)} \cdot \frac{1}{1 - R_i^2}, \tag{2}$$

donde $\widehat{\beta}_i$ y $\text{var}(\mathbf{X}_i)$ son la estimación por Mínimos Cuadrados Ordinarios (MCO) del coeficiente de la variable exógena i-ésima y su correspondiente varianza, y $\widehat{\sigma}^2$ es la estimación de la varianza de la perturbación aleatoria. Diversos factores pueden provocar un valor alto de $\widehat{\text{var}}\left(\widehat{\beta}_i\right)$. Entre estos factores, es evidente que una muestra pequeña también puede provocar varianzas grandes para los coeficientes estimados (ver O'Brien, 2007). Es decir, la micronumerosidad aproximada puede suponer que $\widehat{\text{var}}\left(\widehat{\beta}_i\right)$ sea grande y, por tanto, exista tendencia a no rechazar la hipótesis nula en los contrastes de significación individual (uno de los síntomas de presencia de multicolinealidad grave en el modelo de regresión junto a un coeficiente de determinación alto y un modelo significativo conjuntamente).

Entre las diversas soluciones que se suelen barajar se tienen aquellas relacionadas con los datos que se disponen (como, por ejemplo, mejora del diseño muestral, incorporar más observaciones a la muestra o el uso de información a priori), la aplicación de técnicas de estimación alternativas a MCO (como, por ejemplo, la regresión cresta, regresión alzada, regresión de componentes

principales, regresión con variables ortogonales, regresión LASSO o máxima entropía) o incluso la eliminación del modelo de las variables que se consideran provocan el problema. En este trabajo profundizamos en la propuesta realizada en Salmerón y Blanco (2016), que consiste en identificar aquellas observaciones que pudieran estar provocando el problema de micronumerosidad y valorar si es factible su eliminación del análisis. Aunque la micronumerosidad ha sido referenciada en manuales básicos de Econometría como son los de Gujarati (2004) y Wooldrigde (2006), dentro de nuestro conocimiento, no ha sido tratada de forma específica hasta el momento.

El trabajo se estructura como sigue: en la sección 2 se presenta la metodología para mitigar el problema de micronumerosidad, estableciendo los pasos a seguir y posibles criterios para decidir qué conjunto de observaciones eliminar, en el caso de que exista más de uno. Nos gustaría destacar que en el trabajo de Salmerón y Blanco (2016) se plantea la situación en la que se elimina una sóla observación tras comprobar que el Factor de Inflación de la Varianza queda por debajo de los umbrales establecidos, mientras que en el presente trabajo se aumenta el número de observaciones a eliminar y se plantean cuatro criterios más para determinar el conjunto de observaciones idóneo a eliminar. En la sección 3 se plantea la ampliación y automatización de esta metodología, expresándola como un problema de programación no lineal y entera. Sin duda, esta se trata de una de las principales aportaciones del presente artículo, distinguiéndolo del trabajo de Salmerón y Blanco (2016), y el lugar por donde deberían de discurrir las futuras investigaciones sobre el tratamiento de la micronumerosidad siguiendo el planteamiento de prescindir de observaciones. En la sección 4 se ilustran los resultados anteriores a partir de dos ejemplos, el segundo de ellos especialmente enfocado a mostrar los problemas que se pueden encontrar con los criterios anteriormente comentados. Finalmente, en la sección 5 se destacan las principales conclusiones obtenidas en el trabajo.

2. Metodología para mitigar la micronumerosidad

En Salmerón y Blanco (2016) se establecen los pasos necesarios para identificar las observaciones que puedan estar provocando el problema de micronumerosidad, la cual parte del Factor de Inflación de la Varianza (FIV).

El FIV es una de las medidas más usadas para detectar si el grado de multicolinealidad presente en un modelo de regresión lineal es preocupante. Para cada una de las variables exógenas del modelo (1) se obtiene a partir de la expresión:

$$\text{FIV}(i) = \frac{\text{var}\left(\widehat{\beta}_i\right)}{\text{var}\left(\widehat{\beta}_i^o\right)} = \frac{1}{1-R_i^2}, \quad i = 2, \ldots, p,$$

siendo $\widehat{\boldsymbol{\beta}}$ el estimador por MCO del modelo (1), $\widehat{\boldsymbol{\beta}}^o$ el estimador por MCO del modelo (1) suponiendo que las variables exógenas son ortogonales, y R_i^2 el coeficiente de determinación de la regresión auxiliar:

$$\mathbf{X}_i = \mathbf{X}_{-i}\boldsymbol{\delta} + \mathbf{w},$$

donde $\mathbf{X}_{-i}$ es igual a la matriz $\mathbf{X}$ tras eliminar la variable $\mathbf{X}_i$, $i = 2, \ldots, p$.

Puesto que el FIV se obtiene como el cociente entre la varianza observada y la varianza que se hubiera obtenido en el caso de que $\mathbf{X}_i$ estuviese incorrelada con el resto de variables exógenas del modelo, muestra en qué medida se agranda la varianza del estimador como consecuencia de la relación lineal existente entre los regresores. Es comúnmente aceptado que valores del FIV superiores a 10 indicarían que el grado de multicolinealidad presente en el modelo es preocupante. Es decir, una vez calculados los FIVs asociados a cada uno de los regresores, se diría que la multicolinealidad no es grave si todos son inferiores a 10.

Analizando esta medida de detección, se tiene que cuanto mayor sea R_i^2 mayor será FIV(i). Es decir, aquellos valores que hagan que R_i^2 sea alto pueden considerarse como los responsables de la multicolinealidad. Por tanto, un procedimiento para mitigar la micronumerosidad puede ser el siguiente:

(i) Estimar por MCO la regresión auxiliar (2) y obtener sus residuos. Puesto que puede existir más de una regresión auxiliar posible, elegir aquella con un mayor FIV asociado.
(ii) Ordenar (en valor absoluto) de menor a mayor dichos residuos. Aquellos más pequeños deben ser los causantes de un R_i^2 alto.
(iii) Seleccionar las m observaciones con menor residuo y formar grupos de k elementos ($1 \leq k \leq m$), donde m se elige en base al tamaño de la muestra (por ejemplo, se podría establecer que m es $0.1 \cdot n$ o $0.05 \cdot n$).
(iv) Eliminar los grupos de observaciones formados y ajustar por MCO el modelo inicial (1) con las observaciones restantes, comprobando si en cada uno de los modelos resultantes se ha mitigado el problema de micronumerosidad.

En relación a los pasos anteriores hay que tener en cuenta que:

- Es posible que no exista un conjunto de observaciones cuya eliminación de la muestra suponga que se mitigue el problema de micronumerosidad.
- En el caso de que exista, se debe valorar si es factible su eliminación de la muestra ya que se pueden tratar de observaciones relevantes dentro de la misma o que el número de observaciones a eliminar sea elevado en relación con el tamaño de la muestra. Tampoco se ha de obviar que al eliminar observaciones, en teoría, se está ahondando en el problema de micronumerosidad.
- Se considerará que el problema de micronumerosidad se ha mitigado si todos los FIVs son menores que 10. Sin embargo, tal y como se muestra en el segundo ejemplo de la sección 4, este es uno de los puntos que han de mejorarse en futuras investigaciones.
- En el caso de que exista más de un conjunto de observaciones que mitigue el problema, no es claro cuál de ellos elegir. Posibles criterios pueden ser los siguientes: seleccionar aquel conjunto de observaciones que tras ser eliminadas conduzcan a un modelo inicial (1) con mayor coeficiente de determinación, menor valor para la estimación de la varianza de la perturbación aleatoria (para hacer disminuir la varianza de los coeficientes estimados, ver expresión (2)), menor suma de cuadrados de los residuos (buscando un mejor ajuste) o menor FIV. También se puede optar por eliminar el menor número de observaciones posible.

Con respecto al último punto, se ha de tener en cuenta que el objetivo perseguido es el de mitigar el problema de micronumerosidad y obtener el mejor ajuste posible del modelo inicial. Por tanto, es posible que sea necesario considerar una combinación de los criterios anteriores y no limitarse simplemente a obtener un FIV por debajo de los umbrales establecidos tal y como se hace en Salmerón y Blanco (2016). En la subsección 4.2 se ponen de manifiesto las limitaciones de este último planteamiento.

3. El problema de selección de observaciones

Es claro que el procedimiento descrito en la sección anterior, nos permite seleccionar m observaciones que probablemente estén provocando un problema de multicolinealidad en el modelo, pues se están considerando aquellas que precisamente se ajustan mejor según el modelo original. Sin embargo, esto no es más que una aproximación a un problema más general que consiste en la selección de las m observaciones tales que al ajustar el modelo eliminando éstas, el ajuste es el

mejor posible. Es obvio, que aunque queremos mitigar, si existe, el problema de multicolinealidad para poder explotar convenientemente el modelo, lo deseable finalmente es estimar de la mejor forma posible (con mínimos residuos) los datos finales con los que trabajemos.

Una primera opción consiste en enumerar todas las posibles opciones de selección de m observaciones sobre el total, n, estimar los modelos y seleccionar aquel conjunto de observaciones con las que obtengamos un mejor ajuste (R^2). Incluso en el caso de muestras pequeñas, la enumeración supone estimar un número exponencial de modelos, lo que lo hace inviable en la práctica. Es por esto, por lo que planteamos un modelo de programación matemática que permite obtener m observaciones a descartar, en base a un mejor ajuste robusto. El modelo está basado en el uso de operadores, llamados m-centro, sobre los residuos en MCO:

$$\mathrm{F}(e_1, \ldots, e_n) = \sum_{i=m+1}^{n} e_{(i)}^2,$$

donde $e_{(i)} \in \{e_1, \ldots, e_n\}$ tal que $e_{(1)} \leq \cdots \leq e_{(n)}$, $e_{(1)}, \ldots, e_{(n)}$ representa la secuencia, ordenada de menor a mayor, de los residuos del modelo. En la función F se tienen en cuenta los $n-m$ mayores residuos, obviando los m más pequeños. Este operador se encuentra enmarcado dentro de los operadores OWA (Ordered Weighted Averaging) y ha sido utilizado en distintas disciplinas dentro de la toma de decisiones estratégicas (Nickel y Puerto (2005), Blanco et. al (2016), entre otros), y permite obtener soluciones (en este caso, estimaciones) robustas, menos sensibles a variaciones en los datos (véase Rousseauw y Leroy (2003)).

Es por tanto que si planteamos el problema de programación matemática:

$$\begin{aligned} \min \quad & \mathrm{F}(e_1, \ldots, e_n) \\ \text{sujeto a} \quad & e = y - \mathbf{X}\boldsymbol{\beta}, \\ & \boldsymbol{\beta} \in \mathbb{R}^{p+1}, \\ & e_1, \ldots, e_n \in \mathbb{R}, \end{aligned} \tag{3}$$

se pueden estimar los coeficientes del modelo $\boldsymbol{\beta}$, minimizando la suma de los cuadrados de los $n-m$ mayores residuos en un modelo lineal, obviando los m residuos menores (aunque de alguna forma teniendo en cuenta que estos m residuos son menores que los $n-m$ restantes). Obsérvese que el problema anterior no puede resolverse por técnicas estándard de optimización convexa, pues al no conocer a priori cuál es el orden de los residuos (pues este orden dependerá de los coeficientes $\boldsymbol{\beta}$), para ser resuelto de forma óptima necesita describir este orden en las restricciones. Aunque existen diferentes metodologías para reformular el modelo anterior utilizando formulaciones de programación entera mixta sobre el cono de segundo (MISOCO), en Blanco et. al (2014) se describe una formulación convexa para el problema, que permite resolverlo de forma eficiente con los solvers comerciales disponibles en el mercado. A diferencia de los métodos de detección de outliers/eliminación de observaciones, este planteamiento considera todas las observaciones en el proceso de estimación, pues aunque los residuos cuadráticos a optimizar son únicamente los m mayores, el orden establecido mantiene el resto de residuos por debajo de éstos.

La obtención de estimaciones robustas utilizando este tipo de operadores y técnicas de optimización moderna, permitiría mitigar el problema de micronumerosidad, sin tener que obviar observaciones en el estudio, aunque dándole un peso menor a algunas de ellas. Está por ver cómo, una vez estimados los parámetros del modelo, podemos medir el nivel de multicolinealidad en el problema, para una rigurosa interpretación de los resultados. En particular, observese que la Suma de los Cuadrados de los Residuos (SCR), debiera representar el valor de $\mathrm{F}(e_1, \ldots, e_n)$, pues esta es la cantidad que realmente representa el objetivo buscado, y por tanto, la Suma de Cuadrados Totales (SCT) o Suma de Cuadrados Explicada (SCE), y en tal caso las varianzas correspondientes, habría que calcularlas de forma truncada para ser consistentes con el método de estimación.

Table 1. Residuos de la regresión auxiliar de L sobre X y S

Observación	1	2	3	4	5	6	7	8	9
Residuo	6.23	-46.348	8.337	4.755	-14.324	-21.511	-12.418	6.161	-7.826
Observación	10	11	12	13	14	15	16	17	
Residuo	-40.076	-23.957	59.835	7.2754212	-36.317	0.895	155.503	-46.213	

Finalmente, destacar que una la formulación anterior podría completarse incorporando otro tipo de medidas de bondad sobre los modelos a obtener. En particular, podrían incorporse restricciones del tipo $R^2_J \geq R^2$, $SCR_J \leq SCR$, $\widehat{\sigma}^2_J \leq \widehat{\sigma}^2$, ó $\max \text{FIV}_J \leq \max \text{FIV}$ (aquí R^2_J, SCR_J, $\widehat{\sigma}^2_J$ y $\max FIV_J$ representan al coeficiente de determinación, suma de cuadrados de los residuos, estimación de la varianza de la perturbación aleatoria y máximo FIV tras eliminar el conjunto de observaciones que forman el conjunto J).

4. Ejemplos

En el presente apartado se aplica la metodología presentada a dos conjuntos de datos ampliamente usados para ilustrar la detección y tratamiento de la multicolinealidad. En el primero de ellos se ilustran los pasos mostrados en la sección 2 y 3, mientras que el segundo está enfocado a mostrar que el trabajo planteado en la sección 3 no es inmediato, constituyendo una interesante línea de trabajo. Puesto que se disponen de muestras pequeñas, se podría hablar de micronumerosidad.

4.1. *Datos sobre hospitales navales*

Consideraremos el siguiente modelo para predecir el número de horas mensuales por hombre, H, con fines de dotación en 17 hospitales navales de Estados Unidos (Myers, 1990):

$$H_i = \beta_0 + \beta_1 L_i + \beta_2 X_i + \beta_3 S_i + u_i, \quad i = 1, \ldots, 17, \tag{4}$$

donde las variables exógenas son la carga media de pacientes diaria, L, la exposición mensual a rayos equis, X, y la duración media (en días) de permanencia de los pacientes, S.

La estimación del modelo (4) por MCO conduce a los siguientes resultados:

$$\begin{array}{rcccccc} \widehat{H}_i & = & 1475.024 & + 29.731 \cdot L_i & + 0.0534 \cdot X_i & - 318.14 \cdot S_i & R^2 = 0.9894 \quad \widehat{\sigma} = 635 \\ & & (811.002) & (3.3169) & (0.0208) & (158.4065) & F_{3,13} = 404.6 \quad \text{SCR} = 5241925 \end{array}$$

Teniendo en cuenta que la siguiente matriz de correlaciones de las variables independientes:

$$R = \begin{pmatrix} 1 & 0.9073 & 0.6711 \\ 0.9073 & 1 & 0.4466 \\ 0.6711 & 0.4466 & 1 \end{pmatrix},$$

tiene un determinante próximo a cero, 0.0707, y que uno de los FIVs asociados a cada una de las variables, $FIV(1) = 11.32138, FIV(2) = 7.771393, FIV(3) = 2.498503$, es superior a 10, se tiene que la estimación anterior podría estar afectada por el grado de micronumerosidad presente en el modelo.

Para poner en práctica la metodología anterior, puesto que hay tres regresiones auxiliares posibles, se trabajará con aquella que ha conducido a un mayor FIV (regresión de L sobre X y S). Los residuos de esta regresión se muestran en la Tabla 1. Se puede observar que las observaciones con menor residuo (de menor a mayor) son 15, 4, 8, 1, 13.

A partir de estas 5 observaciones es posible hacer 31 subconjuntos distintos (5 de un elemento, 10 de dos y tres, 5 de cuatro y 1 de cinco). De estas 31 opciones, en 18 de ellas se obtienen regresiones, realizadas tras eliminar las correspondientes observaciones, con FIVs menores que 10. En la Tabla 2 se tienen los resultados obtenidos en cada una de ellas. Se puede observar que:

- El coeficiente de la variable L, que es inicialmente significativamente distinto de cero, lo sigue siendo en todas las regresiones.
- El coeficiente de la variable X, que es inicialmente significativamente distinto de cero, lo sigue siendo en todas las regresiones exceptuando la 9, 15 y 16.
- Las estimaciones de las variables L y X son muy parecidas en todos los casos, donde se producen grandes fluctuaciones es en la estimación de la constante y del coeficiente de la variable S.
- La regresión con mayor coeficiente de determinación corresponde al modelo 17, donde se han eliminado las observaciones (4, 8, 13, 15).
- La regresión con menores FIVs corresponde al modelo 18, donde se han eliminado las observaciones (1, 4, 8, 13, 15). Este modelo también corresponde al de menor suma de cuadrados de los residuos.
- La regresión con menor estimación de la varianza de la perturbación aleatoria corresponde al modelo 3, donde se han eliminado las observaciones (4, 15).
- La regresión en la que se han eliminado un menor número de observaciones corresponde al modelo 1, donde se ha eliminado únicamente la observación 15.

Se puede observar que en todos los modelos destacados aparece la observación 15, este dato no debe sorprendernos si se tiene en cuenta que es la observación con un residuo muy pequeño si se compara con todos los demás.

Si se usan éstas cuatro regresiones y la original para realizar predicción a partir de las medias de las variables exógenas, $\overline{L} = 148.27,\ \overline{X} = 18163.24,\ \overline{S} = 5.89$, se obtienen los siguientes valores:

$$5010.876,\ 4988.478,\ 5016.281,\ 5039.014,\ 18253.58.$$

Se puede observar que los valores obtenidos en las regresiones en las que se han eliminado observaciones ofrecen predicciones muy parecidas entre sí y, a su vez, muy distintas de la que se obtendría con la regresión original. Esta situación se vuelve a repetir si se usan los valores de los hospitales 1, 8 y 17 (ver Tabla 3).

Finalmente, en lo que se refiere a la apliación de técnicas de optimización avanzada para obtener estimaciones robustas, hemos implementado la metodología mencionada en la sección 3, de forma que para $m = 5$ (esto es, las 5 observaciones con menor residuo no computarían en el cálculo de la SCR) se tiene el siguiente modelo estimado:

$$\widehat{H}_i = 1452.71 + 29.947L_i + 0.05329X_i - 325.622S_i,$$

siendo los residuos obtenidos los siguientes:

Obs.	1	2	3	4	5	6	7	8	9
Res.	**-34.7091**	**70.0017**	**152.605**	512.949	**171.184**	-300.694	**262.006**	291.488	-412.35
Obs.	10	11	12	13	14	15	16	17	
Res.	-851.526	449.48	-631.92	-273.192	1667.89	-314.155	282.049	-421.907	

donde se marcaron en negrita los 5 residuos cuyo cuadrado es menor de los 17. Las estimaciones (y los residuos) corresponden, por tanto, con las mejores de entre todas las posibles combinaciones de 5 observaciones obviadas, cuando en el ajuste se mide en base a los residuos a las 12 restantes. Si realizamos la predicción utilizando esta estimación, obtenemos los datos de la siguiente tabla:

	$\widehat{H}$
Media	4943.046
Hospital 1	601.2202
Hospital 8	1869.103
Hospital 17	19275.91

Table 2. Estimaciones del modelo de hospitales norteamericanos

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6	Modelo 7	Modelo 8	Modelo 9
Constante	751.13	740.26	58.84	694.05	792.99	1190.37	-105.16	664.98	796.54
	(1131.33)	(1235.12)	(1243.9)	(1181.19)	(1171.01)	(988.73)	(1393.92)	(1299.5)	(1285.9)
L	29.757	29.753	29.65	29.71	29.82	29.594	29.608	29.69	29.827
	(3.335)	(3.486)	(3.269)	(3.46)	(3.44)	(3.65)	(3.41)	(3.63)	(3.61)
X	0.0509	0.0509	0.0503	0.0515	0.0505	0.054	0.0506	0.0516	0.0505
	(0.021)	(0.022)	(0.0206)	(0.021)	(0.021)	(0.023)	(0.021)	(0.023)	(0.022)
S	-177.97	-176.28	-59.82	-171.69	-181.88	-279.64	-33.64	-167.23	-182.42
	(220.05)	(236.51)	(236.14)	(228.6)	(227.33)	(183.32)	(259.41)	(247.22)	(245.5)
R^2	0.989	0.9886	0.9901	0.989	0.9893	0.9891	0.9898	0.9886	0.9888
F_{exp} (g.l.)	360.7 (3, 12)	318 (3, 11)	368.4 (3, 11)	331 (3, 11)	338.2 (3, 11)	302.1 (3, 10)	323.8 (3, 10)	288.4 (3, 10)	295.4 (3, 10)
$\widehat{\sigma}$	638.6	667	**625.6**	662.1	659.4	699.1	652.8	694.2	691.6
	9.34	8.96	8.98	9.204	9.36	9.96	8.54	8.78	8.97
FIVs	7.57	7.34	7.45	7.47	7.58	7.17	7.18	7.21	7.35
	1.75	1.68	1.67	1.73	1.75	2.22	1.58	1.66	1.68
SCR	4893720	4893779	4305129	48221141	4782892	4887408	4261478	4819136	4783106
Obs. eliminadas	15	(1, 15)	(4, 15)	(8, 15)	(13, 15)	(1, 4, 8)	(1, 4, 15)	(1, 8, 15)	(1, 13, 15)

	Modelo 10	Modelo 11	Modelo 12	Modelo 13	Modelo 14	Modelo 15	Modelo 16	Modelo 17	Modelo 18
Constante	-63.15	113.8	738.54	1253.7	-288.8	-33.99	724.91	-7.628	-219.406
	(1301.1)	(1300.5)	(1229.8)	(1051.8)	(1472.4)	(1471.8)	(1363.08)	(1371.39)	(1571.8)
L	29.582	29.71	29.77	29.66	29.515	29.66	29.77	29.638	29.57
	(3.37)	(3.403)	(3.59)	(3.83)	(3.531)	(3.57)	(3.79)	(3.538)	(3.73)
X	0.051	0.05	0.051	0.054	0.0516	0.0503	0.051	0.0508	0.051
	(0.022)	(0.024)	(0.022)	(0.024)	(0.022)	(0.022)	(0.024)	(0.022)	(0.023)
S	-43.8	-66.6	-175.9	-287.1	-8.15	-43.15	-173.8	-50.698	-17.47
	(245.2)	(246.14)	(237.5)	(193.2)	(271.4)	(272.74)	(258.41)	(257.55)	(288.15)
R^2	0.9903	0.9903	0.9893	0.9892	0.99	0.9899	0.9888	**0.9904**	0.9901
F_{exp} (g.l.)	340.3 (3, 10)	340.2 (3, 10)	307 (3, 10)	273.5 (3, 9)	296.7 (3, 9)	294.9 (3, 9)	264.3 (3, 9)	309.9 (3, 9)	265.4 (3, 8)
$\widehat{\sigma}$	645.7	650.7	687.2	732.2	674.6	683.1	724.3	676	711.8
	8.804	8.99	9.21	9.92	8.328	8.55	8.79	8.81	**8.326**
FIVs	7.32	7.45	7.47	7.16	7.03	7.18	7.21	7.32	**7.024**
	1.64	1.67	1.73	2.21	1.55	1.58	1.66	1.64	**1.55**
Obs. eliminadas	(4, 8, 15)	(4, 13, 15)	(8, 13, 15)	(1, 4, 8, 13)	(1, 4, 8, 15)	(1, 4, 13, 15)	(1, 8, 13, 15)	(4, 8, 13, 15)	(1, 4, 8, 13, 15)
SCR	4169285	4234105	4722438	4825052	4095766	4199630	4721494	4112784	**4053274**

Table 3. Predicción para los hospitales 1, 8 y 17

						Predicción ($\widehat{H}$) según el modelo			
Individuo	H	L	X	S	Inicial	17	18	3	1
Hospital 1	566.52	15.57	2463	4.45	13928.95	353.35	288.87	378.18	547.84
Hospital 8	2160.55	59.28	5969	5.15	15193.02	1791.443	1747.952	1808.66	1902.402
Hospital 17	18854.45	510.22	86533	6.35	32520.26	19188.22	19170.05	19159.62	19208.17

Se observa la cercanía de las estimaciones obtenidas con respecto a las que se obtenían con la aproximación descrita en la sección 2, tanto para los valores medios, como para los valores reales de la variable H, L, X y S en los hospitales 1, 8 y 17.

4.2. *Datos de consumo e ingresos agrícolas*

En este caso, se analizará el modelo usado por Klein y Goldberger (1964) sobre el consumo e ingresos salariales en los Estados Unidos para los años 1936 a 1952 (los datos de 1942 a 1944 no están disponibles por estar en guerra). En este caso, la estimación del modelo por MCO conduce a los siguientes resultados:

$$\begin{array}{lcccccc} \widehat{C}_t & = & 18.7021 & + 0.3803 \cdot I_t & + 1.4186 \cdot InA_t & 0.5331 \cdot IA_t & R^2 = 0.9187 \quad \widehat{\sigma} = 6.06 \\ & & (6.8454) & (0.3121) & (0.7204) & (1.3998) & F_{3,10} = 37.68 \end{array}$$

donde C es el consumo, I son los ingresos salariales, InA los ingresos no agrícolas e IA los ingresos agrícolas. Se observa que ninguna de las variables tienen coeficientes significativamente distintos de cero, el coeficiente de determinación es alto y el modelo es significativo conjuntamente. Es decir, todos los síntomas de la existencia de micronumerosidad grave en el modelo. Esta sospecha se confirma con los FIVs asociados 12.2965, 9.23007 y 2.9766.

Si se realizan todas las regresiones posibles tras eliminar dos observaciones, 91, se tiene que en sólo 10 se obtienen que todos los FIVs son menores que 10 (según lo establecido, la micronumerosidad se habría mitigado) y en ninguno de estos casos cambian los síntomas anteriores (significación individual y conjunta). Sin embargo, eliminando los pares de observaciones (1, 6) ó (1, 7) sí que se obtienen coeficientes significativamente distintos de cero aunque existan FIVs superiores a 10 (ver, respectivamente, modelos 1 y 2 de la Tabla 4).

Por otro lado, si se realizan todas las regresiones posibles tras eliminar tres observaciones, 364, se tiene que en 61 se obtienen que todos los FIVs son menores que 10 y, al igual que antes, ninguno de los síntomas de micronumerosidad cambian. Sin embargo, eliminando los tríos de observaciones (1, 2, 6) ó (1, 2, 7) sí que se obtienen coeficientes significativamente distintos de cero aunque existan FIVs superiores a 10 (ver, respectivamente, modelos 3 y 4 de la Tabla 4).

Por tanto, este ejemplo muestra que la restricción establecida en la sección 2 de que el FIV sea menor que 10, usada en Salmerón y Blanco (2016), no tiene por qué ser la mejor opción y que establecer criterios que presupongan que la micronumerosidad ha sido mitigada no es trivial. Esta cuestión no es sorprendente, ya que tal y como se muestra en la expresión (2) valores altos para el FIV pueden ser compensados, por ejemplo, con estimaciones de la varianza de la perturbación aleatoria pequeños (como en este caso en los modelos 1, 2 y 4) y que se mitiguen así los síntomas de la micronumerosidad sobre significación individual.

Al mismo tiempo, adviértase que al quitar observaciones se tiene que la varianza estimada de los coeficientes aumenta a no ser que mejore la estimación de σ^2 o que el FIV dismuya lo suficiente (ver expresión (2)). Esta situación hace replantearse (una vez más) las restricciones a las que deben estar sujetas el problema de programación matemática mostrado en (**??**). Quizás podría considerarse una situación más general en la se imponga que $\widehat{var}\left(\widehat{\beta}_i\right)_J \leq \widehat{var}\left(\widehat{\beta}_i\right)$, $i = 2, \ldots, p$,

Table 4. Estimaciones del modelo de consumo e ingresos agrícolas

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Constante	17.9288	17.3314	18.9787	18.1844
	(7.5969)	(4.23)	(8.9985)	(4.9701)
Ingresos salariales	0.05403	0.8553	0.04502	0.8553
	(0.3599)	(0.194)	(0.3843)	(0.2051)
Ingresos no agrícolas	2.1814	0.3235	2.1658	0.2937
	(0.8373)	(0.4528)	(0.8926)	(0.4847)
Ingresos agrícolas	0.9677	0.5643	0.9937	0.5793
	(1.4044)	(0.7725)	(1.4971)	(0.8178)
R^2	0.9264	0.9781	0.9108	0.9744
F_{exp} (g.l.)	33.59 (3, 8)	119.3 (3, 8)	23.81 (3, 7)	88.73 (3, 7)
$\widehat{\sigma}$	5.844	3.254	6.217	3.441
	14.4818	12.9238	12.025	10.9759
FIVs	11.1156	10.382	9.1708	8.795
	2.6193	2.4073	2.43	2.2618

donde $\widehat{var}\left(\widehat{\beta}_i\right)_J$ denota la estimación de la varianza de los coeficientes obtenida tras eliminar el subconjunto de observaciones que hay en J.

5. Conclusiones

En el presente trabajo se aborda el caso en el que el escaso número de observaciones puede provocar la existencia de multicolinealidad grave en un modelo de regresión. En este contexto, Goldberger introduce el concepto de micronumerosidad como sinónimo de multicolinealidad. Como alternativa a las soluciones tradicionales de este problema, en Salmerón y Blanco (2016) se propone una metodología para la detección de las posibles observaciones que pudieran estar provocándolo.

En el presente trabajo se profundiza en esta metodología ampliando tanto el número de observaciones a eliminar como el número de criterios a considerar para determinar el conjunto de observaciones idóneo que se ha de eliminar (si es que existe). Ahora bien, es importante destacar que una reducción del tamaño muestral puede conducir a mayores varianzas estimadas de los coeficientes y entonces es posible que se llegue a concluir que la micronumeosidad ha sido mitigada (FIVs por debajo de los umbrales) y que, sin embargo, sus síntomas (varianzas estimadas grandes) perduren o incluso hayan empeorado.

Sin embargo, en la generalización y automatización propuesta al plantear el problema de optimización matemática no se eliminan observaciones sino que son consideradas en el proceso de estimación. Es decir, el tamaño muestral permanece intacto evitando las consecuencias nocivas comentadas. De esta forma, se selecciona el subconjunto de observaciones que conduce a una mejor estimación robusta y con menor valor en las restricciones establecidas. En cualquier caso, no queda claro cuáles han de ser las restricciones a las que debe estar sujeto dicho planteamiento, por lo que constituye una línea de investigación futura más que interesante.

Referencias bibliográficas

1. A. Spanos, A. McGuirk (2002): The problem of near-multicollinearity revisited: erratic vs systematic volatily, *Journal of Econometrics.* **108** (2), 365–393.

2. A. Goldberger (1991): *A course in Econometrics* (Cambridge, MA: Harvard University Press).
3. R.M. O'Brien (2007): A caution regarding rules of thumb for variance inflation factors, *Quality and Quantity,* **41**, 673–690.
4. R. Salmerón, V. Blanco (2016): Micronumerosidad aproximada y regresión lineal múltiple, *XXIV Jornadas de ASEPUMA y XII Encuentro Internacional*, Granada, España, 159–168.
5. D. Gujarati (2004): *Basic Econometrics* (McGraw-Hill, 4a Edición).
6. J.M. Wooldridge (2006): *Introducción a la Econometría: Un Enfoque Moderno* (Thomson, 2a Edición).
7. S. Nickel, J. Puerto (2005): *Location Theory: A Unified Approach.* Berlin, Springer.
8. V. Blanco, J. Puerto, R. Salmerón, R. (2016): *A general framework for multiple linear regression.* Submitted. Preprint available at `https://arxiv.org/abs/1505.03451`.
9. P. Rousseeuw, A. Leroy, A. (2003): *Robust Regression and Outlier Detection.* New York: Wiley.
10. V. Blanco, J. Puerto, S. El-Haj Ben-Ali (2014): *Revisiting several problems and algorithms in continuous location with ℓ_τ norms.* Comput. Optim. Appl. **58** (3), 563–595.
11. R.H. Myers (1990): *Classical and modern regression with apllications"* (2a Edición, PWS-Kent).
12. L.R. Klein, A.S. Goldberger (1964): *An economic model of the United States, 1929-1952* (Amsterdan: North Holland Publishing Company).