CONSECUENCIAS DE ALTA MULTICOLINEALIDAD EN UN MODELO DE REGRESIÓN LINEAL

Dr. (C) Coa Clemente, Ramiro[1]

✉ *clementecoa@gmail.com*

RESUMEN

En este artículo se revisan e ilustran algunas consecuencias de la alta multicolinealidad entre covariables presentes en la parte sistemática de un modelo de regresión lineal. Con este propósito, se comparan dos modelos. En el primero no existe el problema de multicolinealidad, es decir las covariables son linealmente independientes. En el segundo modelo se tiene el problema de alta multicolinealidad, es decir las covariables están muy asociadas linealmente. Se analizan cuatro tipos de consecuencias: (i) sobre la magnitud de los coeficientes de regresión, (ii) sobre las sumas de cuadrados adicionales, (iii) sobre la magnitud de los errores estándar para los estimadores de coeficientes y (iv) sobre pruebas estadísticas de los coeficientes. En presencia de alta multicolinealidad entre las covariables del modelo, estas consecuencias podrían conducir a inferencias estadísticas erróneas y consecuentemente a conclusiones incorrectas.

PALABRAS CLAVE

Multicolinealidad, multicolinealidad perfecta, alta multicolinealidad o multicolinealidad imperfecta.

ABSTRACT

In this article, some consequences of the high multicollinearity among covariates present in the systematic part of a linear regression model are reviewed and illustrated. For this purpose, two models are compared. In the first there is no problem of multicollinearity, that is, the covariates are linearly independent. In the second model there is the problem of high multicollinearity, that is, the covariates are very linearly associated. Analyze four types of consequences: (i) on the magnitude of the regression coefficients, (ii) on the sums of additional squares, (iii) on the magnitude of the standard errors for the coefficient estimators and (iv) on statistical tests of the coefficients. In the presence of high multicollinearity among the covariates of the model, these consequences can lead to erroneous statistical inferences and consequently to incorrect conclusions.

KEYWORDS

Multicollinearity, perfect multicollinearity, high multicollinearity or imperfect multicollinearity.

1. EL PROBLEMA DE MULTICOLINEALIDAD

Consideremos el modelo de regresión lineal múltiple $Y=X\beta+\varepsilon$, donde Y es el vector $n*1$ de variables respuesta, X es la matriz $n*p$ de observaciones, β es el vector $p*1$ de coeficientes y ε es un vector de errores aleatorios con los supuestos clásicos, es decir $\varepsilon \sim N(0,\sigma^2 I_n)$.

Formalmente, la multicolinealidad se define en términos de la dependencia lineal entre las columnas de la matriz X. Recordemos que los vectores $X_1,X_2,...,X_K$ son linealmente dependientes si hay un conjunto de constantes $a_1,a_2,...,a_K$, no todos ceros, tal que $\sum_{k=1}^{K} a_k x_k = 0$. Si esta ecuación se cumple para algún subconjunto de las columnas de X, entonces el rango de la matriz

[1] Ex-Director de Investigación de la Unidad de Analisis y Politica Social (UDAPSO)

$X'X$ es inferior a p, y por tanto no existe una solución única para el vector de coeficientes. En esta situación se tiene un problema denominado multicolinealidad perfecta. El problema más frecuente, sin embargo, es el denominado multicolinealidad imperfecta o alta multicolinealidad. Este se presenta cuando todas o algunas covariables del modelo están altamente correlacionadas, es decir cuando se tiene una dependencia "casi lineal" entre las columnas de X. Por este hecho, la multicolinealidad es un problema principalmente de grado o de nivel. Se podría decir, entonces, que cada conjunto de datos a ser analizado con un modelo de regresión sufre del problema en alguna medida, a no ser que las columnas de X sean ortogonales, en cuyo caso no existe problema de multicolinealidad, ni perfecta ni imperfecta, lo que se da generalmente en un experimento diseñado apropiadamente.

2. CONSECUENCIAS DE LA MULTICOLINEALIDAD IMPERFECTA

En este artículo se examinan cuatro efectos de la presencia de multicolinealidad imperfecta: (i) efectos sobre la magnitud de los coeficientes de regresión, (ii) sobre la suma de cuadrados adicional, (iii) sobre el error estándar de los coeficientes estimados y (iv) sobre las decisiones en pruebas de hipótesis.

Para visualizar apropiadamente estos efectos, la información a ser analizada previamente es transformada mediante la denominada *transformación correlación*. Dos de las interesantes características de esta transformación son: (i) la nueva matriz simétrica $X'X$ representa la matriz de correlaciones entre las columnas de la nueva matriz X y que (ii) el nuevo vector $X'Y$ consiste de las correlaciones entre cada nueva covariable y la nueva variable respuesta. Consecuentemente, el nuevo vector de coeficientes puede ser estimado a partir de la matriz y el vector de correlaciones.

2.1 EFECTOS SOBRE LA MAGNITUD DE LOS COEFICIENTES DE REGRESIÓN

Para ilustrar el efecto de la multicolinealidad imperfecta sobre la magnitud de los coeficientes consideremos dos modelos de regresión, cada uno con dos covariables. En el primer modelo no existe asociación lineal entre las dos covariables, es decir la correlación lineal es 0; en el segundo modelo ambas covariables están altamente asociadas, una correlación lineal de 0,92.

Cuando ambas covariables no están correlacionadas, los estimadores de los coeficientes permanecen invariables. El coeficiente para X_1 es el mismo (0,742) cuando X_1 es la única covariable en el modelo o cuando ambas covariables están en el modelo. Lo mismo ocurre para el coeficiente de X_2. Sin embargo, cuando ambos regresores están fuertemente asociados, los valores de los coeficientes tienen grandes cambios. Cuando X_1 es la única covariable en el modelo, su coeficiente es 0,915; mientras cuando ambas covariables están presentes en el modelo, su coeficiente es 0,294, una reducción de 67,0%. De manera similar, aunque menos acentuada, la reducción para el coeficiente de X_2 es 28,8% (Cuadro Nº 1).

Cuando las covariables están fuertemente asociadas, las magnitudes de sus coeficientes cambian significativamente. Consecuentemente, esos coeficientes no reflejan los efectos reales de sus correspondientes covariables sobre la respuesta, sólo reflejan un efecto parcial o marginal, que podría conducir a conclusiones erradas.

Cuadro Nº 1
Cambios en las magnitudes de los coeficientes

Variables en el Modelo	Modelo 1 (Correlación 0)		Modelo 2 (Correlación 0,92)	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
X_1	0,742		0,915	
X_2		0,638		0,945
X_1 X_2	0,742	0,638	0,294	0,673
Cambio Relativo (%)	0	0	67,9%	28,8%

Fuente: Elaboración Propia

2.2 EFECTOS SOBRE LAS SUMAS DE CUADRADOS ADICIONALES

Consideremos los dos modelos de regresión anteriores. Recordemos que el primero es un modelo en el que no existe asociación lineal entre los dos regresores (correlación lineal 0), mientras en el segundo ambos regresores se encuentran altamente asociados (correlación lineal de 0,92).

Cuando ambos regresores no están asociados linealmente, la contribución adicional de X_1 a la suma de cuadrados de la regresión (SCReg) es la misma que cuando X_1 es la única covariable presente en el modelo. Esto es, la $SCReg(X_1)$=0,550 $=SCReg(X_1 \mid X_2)$. De manera similar, la $SCReg(X_2)$=0,408=$SCReg(X_2 \mid X_1)$. En cambio, cuando ambas covariables están altamente relacionadas, la contribución marginal de cualquiera de ellas a la suma de cuadrados de la regresión es diferente de la contribución a la suma de cuadrados cuando la misma variable es la única en el modelo; es decir, la $SCReg(X_1)$=0,838≠0,013=$SCReg(X_1 \mid X_2)$ y la $SCReg(X_2)$=0,892≠0,067=$SCReg(X_2 \mid X_1)$ (Cuadro Nº 2).

Cuadro Nº 2
Cambios en las sumas de cuadrados

	Modelo 1 (Correlación 0)	Modelo 2 (Correlación 0,92)
$SCReg(X_1)$	0,550	0,838
$SCReg(X_1 \mid X_2)$	0,550	0,013
$SCReg(X_2)$	0,408	0,892
$SCReg(X_2 \mid X_1)$	0,408	0,067

Fuente: Elaboración Propia

La razón por la que la $SCReg(X_1 \mid X_2)$=0,013 es muy pequeña comparada con la SCReg(X_1)=0,838 es que X_1 y X_2 están altamente correlacionadas. Cuando X_2 se encuentra en el modelo, la contribución marginal de X_1 a la suma de cuadrados de la regresión es comparativamente pequeña debido a que X_2 contiene mucha de la misma información contenida en X_1. Dicho en otros términos, cuando las covariables del modelo están muy asociadas linealmente, no hay una única suma de cuadrados que pueda ser adscrita a cualquiera de estas covariables, por lo que no es posible determinar el efecto

neto de esas covariables en la explicación de la variabilidad de la respuesta.

2.3 EFECTOS SOBRE LA MAGNITUD DE LOS ERRORES ESTÁNDAR DE LOS ESTIMADORES DE COEFICIENTES

Nuevamente consideremos los dos modelos de regresión anteriores. Ahora se analiza el efecto de una fuerte asociación lineal entre las covariables del modelo sobre los errores estándar de sus correspondientes coeficientes estimados.

En ausencia de asociación lineal entre las covariables del modelo, los errores estándar de los coeficientes estimados se reducen. Por ejemplo, cuando el modelo contiene sólo X_1, el error estándar de su coeficiente asociado β_1 es 0,274, valor que se reduce a 0,092 cuando ambas covariables se encuentran en el modelo. Similar comportamiento se observa para X_2 (Cuadro Nº 3). Este hecho es deseable puesto que los estimadores de los coeficientes tienen mayor precisión.

Cuadro Nº 3
Cambios en los errores estándar de los coeficientes estimados

Variables en el Modelo	Modelo 1 (Correlación 0)		Modelo 2 (Correlación 0,92)	
	$ee(\hat{\beta}_1)$	$ee(\hat{\beta}_2)$	$ee(\hat{\beta}_1)$	$ee(\hat{\beta}_2)$
X_1	0,274		0,142	
X_2		0,314		0,945
X_1, X_2	0,092	0,092	0,303	0,303
Cambio	Reducción	Reducción	Incremento	Incremento

Fuente: Elaboración Propia

En cambio, cuando las covariables están fuertemente asociadas, los errores estándar de sus correspondientes coeficientes estimados se incrementan considerablemente. Por ejemplo, el error estándar de $\hat{\beta}_1$se incrementa de 0,142 a 0,303, un incremento de un poco más del doble. El incremento para el error estándar de $\hat{\beta}_2$ es de casi el triple, pasando de 0,116 a 0,303. En consecuencia, el alto grado de multicolinealidad entre las covariables produce un incremento considerable en los errores estándar de los estimadores, lo que conduce a inferencias menos imprecisas e incluso a falsas conclusiones.

2.4 EFECTOS SOBRE PRUEBAS ESTADÍSTICAS DE LOS COEFICIENTES

Un abuso frecuente en el análisis de modelos de regresión lineal es examinar para cada coeficiente de regresión la estadística $t = \hat{\beta}_k / e.e(\hat{\beta}_k)$ que resulta de dividir el estimador del coeficiente por su error estándar, esto con el propósito de decidir si la hipótesis nula H_0: $\beta_k = 0$ es o no rechazada, para un determinado nivel de significancia α. En presencia de alta multicolinealidad, sin embargo, las conclusiones derivadas de dicho examen podrían ser incorrectas. Esta situación es analizada a continuación.

En ausencia del problema de multicolinealidad, el test t para probar individualmente cada una de la dos hipótesis nulas H_0:β_1=0 y H_0:β_2=0 conduce a la misma decisión a la que se llega con el test F, un test adecuado para probar la hipótesis de que simultáneamente ambos coeficientes son nulos, es decir H_0:β_1=β_2=0. En efecto, los valores de las estadísticas t para $\hat{\beta}_1$(8,10) y

para $\hat{\beta}_2$ (6,97) son superiores al valor crítico 2,97, por lo que individualmente se rechazan ambas hipótesis, H_0:β_1=0 y H_0:β_2=0 (Cuadro Nº 4). Estas decisiones son coherentes con la decisión tomada con base en el test *F*. Este test también conduce a rechazar la hipótesis de que simultáneamente ambos coeficiente son nulos H_0:β_1=β_2=0, pues el valor de la estadística *F*(57,06) supera el percentil 95 de la distribución *F*(5,79). En consecuencia, cuando las covariables no están asociados linealmente, ambos tests, *t* y *F*, son coherentes, conducen a las mismas decisiones y, por ende, a las mismas conclusiones.

Cuadro Nº 4
Efectos sobre pruebas estadísticas de los coeficientes

Test	Modelo 1 (Correlación 0)		Modelo 2 (Correlación 0,92)	
	Hipotesis	**Hipotesis**	**Hipotesis**	**Hipotesis**
	H_0: β_1=0 H_0: β_2=0	H_0: β_1= β_2=0	H_0: β_1=0 H_0: β_2=0	H_0: β_1= β_2=0
t	$t_{tabla}(0{,}9875; 6) = 2{,}97$ $t_{cal,\hat{\beta}_1} = 8{,}10$ $t_{cal,\hat{\beta}_2} = 6{,}97$ Como: $t_{cal,\hat{\beta}_1} > t_{tabla}$ $t_{cal,\hat{\beta}_2} > t_{tabla}$ Decisión: **Rechazar** ambas H_0		$t_{tabla}(0{,}9875; 8) = 2{,}75$ $t_{cal,\hat{\beta}_1} = 0{,}97$ $t_{cal,\hat{\beta}_2} = 2{,}22$ Como: $t_{cal,\hat{\beta}_1} < t_{tabla}$ $t_{cal,\hat{\beta}_2} < t_{tabla}$ Decisión: **No Rechazar** ambas H_0	
F		$F_{tabla}(0{,}95; 2{,}5) = 5{,}79$ $F_{cal} = 57{,}06$ Como: $F_{cal} > F_{tabla}$ Decisión: **Rechazar** H_0		$F_{tabla}(0{,}95; 2{,}7) = 4{,}74$ $F_{cal} = 33{,}36$ Como: $F_{cal} > F_{tabla}$ Decisión: **Rechazar** H_0

Fuente: Elaboración Propia

En cambio, en presencia de alta multicolinealidad entre ambas covariables, los dos tests, *t* y *F*, conducen a decisiones contradictorias. La prueba *t* conduce a no rechazar individualmente cada una de las dos hipótesis H_0:β_1=0 y H_0:β_2=0; mientras la prueba F, una prueba más apropiada para este problema, conduce a rechazar la hipótesis de que simultáneamente ambos coeficientes son nulos H_0:β_1=β_2=0.

La razón para este resultado contradictorio es que cada una de las dos pruebas t es una prueba marginal. Esto es, un valor pequeño de la $SCReg(X_1 \mid X_2)$ (Cuadro Nº 2) indica que X_1 no proporciona mucha información adicional sobre la que proporciona la covariable X_2, por lo que se arriba a la

conclusión de que $\beta_1=0$. De manera similar se llega a la conclusión de que $\beta_2=0$ porque la *SCReg* $(X_2 \mid X_1)$ es pequeña, reflejando que X_2 no proporciona información adicional substancial cuando X_1 se encuentra en el modelo. En consecuencia, los dos tests, *t* y *F*, conducen a decisiones y consecuentemente a conclusiones contradictorias.

3. CONCLUSIÓN

La presencia de alta multicolinealidad en la parte sistemática del modelo de regresión tiene varias consecuencias. Una primera es la disminución en la magnitud de los coeficientes correspondientes a covariables fuertemente asociadas, razón por lo que esos coeficientes sólo reflejan un efecto parcial o marginal sobre la respuesta, no reflejan los efectos reales de sus correspondientes covariables. Una segunda consecuencia tiene que ver con el hecho de que no hay una única suma de cuadrados que pueda ser adscrita a cualquiera de las covariables altamente asociadas, por lo que no es posible determinar el efecto neto de esas covariables. La tercera consecuencia es el incremento en los errores estándar de los coeficientes estimados, lo que conduce a inferencias menos precisas. Por último, con relación a las pruebas de hipótesis, la presencia de alta multicolinealidad lleva a decisiones contradictorias entre las pruebas *t* y *F*.

BIBLIOGRAFÍA

Greene, W. (1997), "*Econometric Analysis*", 3ra Ed. MacMillan.

Seber, G.A.F. and Lee, A.J, (2003), "*Linear Regression Analysis*", 2da Ed. Wiley.

Sen, A. and Srivastava, M. (1990) "*Regression Analysis: Theory, Methods, and Applications*". Springer-Verlag.