

Máster Universitario en Técnicas Cuantitativas en Gestión Empresarial



ugr | Universidad
de **Granada**

El problema de la multicolinealidad en modelos de
regresión con interacciones: análisis de las tasas de
suicidio en España.

Tutor/a:

Román Salmerón Gómez

Alumno/a:

Pilar Rodríguez Puente

Granada, 2026

DECLARACIÓN EXPLÍCITA DE ORIGINALIDAD

- La alumna D^a Pilar Rodríguez Puente declara que la presente memoria se corresponde con un trabajo original, en el sentido de que no se ha utilizado fuentes sin citarlas debidamente.

Fdo. D^a Pilar Rodríguez Puente

- D. Román Salmerón Gómez tutor de la alumna D^a Pilar Rodríguez Puente da el visto bueno para la presentación y defensa de la presente memoria.

Fdo. D. Román Salmerón Gómez

Índice

1. Introducción	1
2. Preliminares	2
2.1. Formulación matemática del problema de multicolinealidad	3
2.2. Herramientas para la detección de multicolinealidad	4
2.3. Posibles soluciones al problema de la multicolinealidad	5
3. Metodología: regresión alzada	5
3.1. Estimación y principales características	6
3.2. Multicolinealidad: FIV, número de condición y variabilidad	8
3.3. Selección de la variable a alzar y del factor de alzado	9
3.4. Relación con la residualización	9
4. Aplicación de la metodología del alzado a modelos de regresión con interacciones	10
4.1. Estimación por Mínimos Cuadrados Ordinarios	12
4.2. Estimación mediante la regresión alzada	15
4.2.1. Alzando la interacción	15
4.2.2. Alzando la variable cuantitativa sin tener en cuenta la interacción . .	16
4.2.3. Alzando la variable cuantitativa teniendo en cuenta la interacción . .	19
4.3. Estimación mediante el centrado de variables	21
5. Discusión de resultados	23
6. Conclusiones	26
A. Relación entre el centrado de variables y la residualización	27

1. Introducción

En el análisis econométrico, los modelos de regresión lineal múltiple constituyen una herramienta fundamental para entender las relaciones entre variables económicas. Sin embargo, la utilidad de estos modelos puede verse comprometida cuando ciertas condiciones teóricas no se cumplen. Una de las problemáticas más comunes y potencialmente conflictiva es la multicolinealidad, es decir, la presencia de una alta relación lineal entre al menos dos de las variables explicativas incluidas en el modelo.

Tal y como señalan Stock y Watson [31], la multicolinealidad no implica un sesgo en los estimadores por mínimos cuadrados ordinarios (MCO), pero sí puede provocar una gran imprecisión en las estimaciones, lo cual dificulta la interpretación individual de los coeficientes y reduce la potencia estadística de las pruebas de hipótesis. En contextos aplicados, esto se traduce en intervalos de confianza más amplios, valores experimentales elevados y, en definitiva, en una menor capacidad para identificar relaciones estadísticamente significativas entre las variables del modelo.

Según Brambor et al. [2], este problema puede aumentar cuando el modelo incluye términos de interacción, los cuales se emplean para capturar efectos no aditivos entre variables explicativas. Si bien estos términos permiten una mayor flexibilidad y riqueza en la especificación del modelo —posibilitando, por ejemplo, que el efecto de una variable sobre la variable dependiente dependa del nivel de otra—, su inclusión puede generar (según estos autores, recalquemos de nuevo) un aumento notable en la multicolinealidad entre los regresores.

Esta cuestión llama la atención, ya que aunque los términos de interacción son construcciones algebraicas que combinan variables ya presentes en el modelo, lo que puede inducir una relación lineal fuerte entre los regresores originales y sus interacciones, realmente se están añadiendo relaciones que no son lineales, por lo que no se debería ver aumentado el grado de multicolinealidad existente de forma relevante.

Brambor et al. [2] también resaltan que la afirmación de que el centrado de variables (relevantes) puede mitigar cualquier problema de multicolinealidad que pudiera existir simplemente no es cierta, recomendando a los académicos que dejen de justificar el uso de variables centradas en los modelos de interacción alegando que esto reduce el grado de multicolinealidad existente. Esta recomendación no es baladí ya que, tal y como indican Hainmueller et al. [10], el trabajo de Brambor, Clark y Golder se considera como la mejor práctica en Ciencias Políticas en cuanto a la aplicación de modelos de interacciones.

En cambio Iacobucci et al. [13] indican que el centrado de variables puede aliviar la multicolinealidad de tipo micro (relaciones dos a dos medidas con el coeficiente de correlación) aunque no la de tipo macro (relaciones globales medidas mediante la matriz de correlaciones).

Debido a las contradicciones anteriores y dada la utilización de modelos con interacciones en investigación aplicada, resulta esencial estudiar en profundidad cómo y en qué medida la multicolinealidad afecta a estos modelos, así como explorar posibles soluciones o estrategias de diagnóstico y mitigación. Este trabajo se centra precisamente en ese objetivo: analizar el problema de la multicolinealidad en presencia de interacciones en modelos de regresión.

Para ilustrar el análisis, se emplea un caso aplicado al ámbito de la salud pública: el estudio de la tasa de suicidio en España por comunidades autónomas. Este análisis resulta de gran interés por su relevancia social y por presentar una fuerte heterogeneidad territorial. Se emplean variables que recogen dimensiones socioeconómicas y demográficas que nos van a permitir analizar cómo el efecto de cada una puede variar según el nivel de las demás.

Diversos trabajos han analizado las tasas de suicidio en España desde enfoques agregados, poniendo de manifiesto la influencia de factores socioeconómicos, demográficos y ambientales, así como la existencia de una marcada heterogeneidad territorial (De Irala et al. [5]). En particular, algunos estudios evidencian una relación entre el desempleo y la mortalidad por suicidio en determinados periodos económicos (crisis) y contextos macroeconómicos (Iglesias et al. [14]), mientras que otros destacan el papel conjunto de determinantes sociodemográficos, económicos y del entorno en la explicación de este fenómeno desde enfoques territoriales y geoespaciales (Irala et al. [5]). No obstante, la literatura existente es relativamente escasa y presenta resultados dispares.

La elección de este caso de estudio responde también a una motivación personal, fruto del interés y la curiosidad por comprender qué factores del contexto económico y social pueden influir en que personas de distintas edades y territorios lleguen a tomar una decisión tan extrema, y cómo estas relaciones pueden variar según el entorno en el que se desarrollan.

Más concretamente, el trabajo se estructura como sigue: en la sección 2 se revisa brevemente la literatura dando una definición formal de multicolinealidad desde el punto de vista matemático, así como presentando las diferentes herramientas para su detección y las posibles soluciones al problema. Dentro de las soluciones comentadas en la sección anterior nos vamos a centrar en estudiar la regresión alzada presentada en la sección 3. En la sección 4 se aplica ésta metodología al caso de estudio y se comentan los resultados obtenidos. En la sección 5 se discuten dichos resultados obtenidos y en la 6 se presentan las conclusiones obtenidas en el trabajo y las posibles líneas de investigación futuras.

2. Preliminares

En esta sección se presenta una exposición formal del problema de multicolinealidad en modelos de regresión lineal, así como las principales herramientas para su detección y algunas estrategias de mitigación.

2.1. Formulación matemática del problema de multicolinealidad

Consideremos el modelo de regresión lineal múltiple con n observaciones y p variables explicativas expresado en forma matricial:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad (1)$$

donde $\mathbf{y} \in \mathbb{R}^n$ es el vector de la variable dependiente, $\mathbf{X} \in \mathbb{R}^{n \times p}$ es la matriz de diseño que contiene una columna de unos en primer lugar (término constante) y las variables explicativas, $\boldsymbol{\beta} \in \mathbb{R}^p$ es el vector de parámetros a estimar que recoge la relación entre las variables independientes y la dependiente, y $\mathbf{u} \in \mathbb{R}^n$ es el vector de términos de perturbación, que se supone esférico (es decir, centrado, homocedástico e incorrelado).

La matriz $\mathbf{X}^\top \mathbf{X}$ desempeña un papel clave en la estimación de los parámetros por mínimos cuadrados ordinarios (MCO), cuya expresión es:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \quad (2)$$

dado que la multicolinealidad ocurre cuando las columnas de $\mathbf{X}$ están altamente relacionadas linealmente, lo que implica que $\mathbf{X}^\top \mathbf{X}$ sea casi singular (tiene un rango deficiente). En este caso, la matriz inversa $(\mathbf{X}^\top \mathbf{X})^{-1}$ es inestable, lo que se traduce en una alta varianza de los estimadores ya que ésta se expresa como sigue:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}, \quad (3)$$

siendo σ^2 la varianza del término de perturbación.

Como se ha comentado anteriormente, se entiende por multicolinealidad la relación lineal alta entre al menos dos variables explicativas del modelo de regresión lineal. Sin embargo, es posible profundizar más en este concepto.

Así, Marquardt y Snee [16], Marquardt [17] y Snee y Marquardt [29] distinguieron entre multicolinealidad esencial y no esencial:

- **Multicolinealidad esencial:** existencia de una relación lineal entre las variables explicativas del modelo excluyendo la constante.
- **Multicolinealidad no esencial:** existencia de una relación lineal entre la constante y al menos una de las variables explicativas del modelo.

Esta distinción es relevante, ya que como se mencionará a continuación, no todas las medidas de detección son capaces de detectar ambos tipos de multicolinealidad.

2.2. Herramientas para la detección de multicolinealidad

Existen diversas herramientas estadísticas para detectar la presencia de multicolinealidad en un modelo. Los principales son los siguientes:

- **Matriz de correlaciones:** permite detectar relaciones lineales entre pares de variables. Según Salmerón et al. [22, 26] para que una correlación lineal indique una relación preocupante ha de ser mayor que $\sqrt{0.9}$.
- **Regresiones auxiliares** y su coeficiente de determinación, R_i^2 , utilizado para calcular el Factor de Inflación de la Varianza (FIV):

$$FIV(i) = \frac{\text{var}(\hat{\beta}_i)}{\text{var}(\hat{\beta}_i^o)} = \frac{1}{1 - R_i^2}, \quad i = 2, \dots, p, \quad (4)$$

donde R_i^2 representa el coeficiente de determinación de $\mathbf{X}_i$ sobre el resto de variables independientes $\mathbf{X}_{-i}$.

En la práctica, cuando el valor del FIV supera 10, se supone que el grado de multicolinealidad presente en el modelo es preocupante (Marquardt [15]).

Es importante señalar que el FIV no capta la posible relación entre las variables $\mathbf{X}_2, \dots, \mathbf{X}_p$ y la constante, por lo que no detecta la llamada multicolinealidad no esencial (ver, por ejemplo, Salmerón et al. [24]).

- **Número de condición:** herramienta basada en la descomposición en valores singulares de la matriz $\mathbf{X}$, que permite medir la sensibilidad de la estimación ante perturbaciones numéricas:

$$K(\tilde{\mathbf{X}}) = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}},$$

donde $\lambda_{\max}$ y $\lambda_{\min}$ son, respectivamente, los autovalores máximo y mínimo de $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$, donde:

$$\tilde{\mathbf{X}} = [\tilde{\mathbf{1}} \ \tilde{\mathbf{X}}_2 \ \dots \ \tilde{\mathbf{X}}_p], \quad \tilde{\mathbf{1}} = \frac{1}{\sqrt{n}}, \quad \tilde{\mathbf{X}}_i = \frac{\mathbf{X}_i}{\sqrt{\sum_{j=1}^n X_{ji}^2}}, \quad i = 2, \dots, p.$$

Es decir, $\tilde{\mathbf{X}}$ representa la transformación de $\mathbf{X}$ en longitud unidad.

Si esta medida es superior a 20 se supone que el grado de multicolinealidad presente en el modelo es moderado y si es superior a 30 preocupante (Belsley [1]).

- **Coeficiente de variación:** en Salmerón et al. [24] se muestra que una variable explicativa tenga un coeficiente de variación inferior a 0.1002506 es indicativo de que dicha variable es responsable de un problema de multicolinealidad aproximada del tipo no esencial. Además, según Montgomery et al. [18] este problema se resuelve centrando (restando la media) la variable que provoca el problema.

2.3. Posibles soluciones al problema de la multicolinealidad

Cuando se detecta multicolinealidad severa, existen diversas estrategias para mitigar sus efectos:

- **Eliminar o combinar variables:** si dos variables son altamente colineales, puede considerarse eliminar una o reemplazarlas por un índice compuesto.
- **Recoger más datos:** aumentar el tamaño muestral puede ayudar a reducir la varianza de los estimadores.
- **Transformaciones:** técnicas como la estandarización o centrado de las variables pueden aliviar la colinealidad. Sobre todo en el caso de la multicolinealidad del tipo no esencial (ver, por ejemplo, Montgomery et al. [18] o Salmerón et al. [24]).
- Técnicas alternativas de estimación a los Mínimos Cuadrados Ordinarios:
 - **Regresión Ridge:** introduce un término de penalización en la función de mínimos cuadrados, que regulariza la estimación haciéndola converger hacia cero y reduce la varianza a costa de un pequeño sesgo (ver Hoerl y Kennard [11, 12] para más detalles).
 - **Regresión LASSO:** también introduce un término de penalización en la función de mínimos cuadrados, la cual en este caso permite la selección de variables (ver Tibshirani [33] para más detalles).
 - **Regresión elastic-net:** combinación de las dos anteriores (ver Zou y Hastie [35] para más detalles).
 - **Regresiónalzada y residualizada:** ver siguiente sección.

Estas soluciones no están exentas de limitaciones, por lo que su aplicación debe analizarse cuidadosamente en función del contexto del modelo y los objetivos del análisis, especialmente la primera.

3. Metodología: regresión alzada

Entre las alternativas planteadas a la estimación por mínimos cuadrados ordinarios (MCO), nos centraremos en la regresión alzada (García et al. [6] y Salmerón et al. [21, 28]), la cual constituye una herramienta útil para mitigar los inconvenientes derivados de la multicolinealidad. Su principal ventaja es que lo hace sin sacrificar la capacidad de inferencia ni el coeficiente de determinación o significación global: no solo se conserva el mismo valor obtenido mediante MCO, sino también su interpretación estadística.

En este apartado se presenta el marco teórico y metodológico que fundamenta el análisis empírico desarrollado en este trabajo. El objetivo es exponer los fundamentos conceptuales y las aportaciones más relevantes de la literatura sobre la regresión alzada.

La regresión alzada fue introducida formalmente en García et. al. [21] como un procedimiento alternativo para estimar parámetros en presencia de colinealidad. Posteriormente, Salmerón et. al. [28] desarrollan de manera rigurosa su justificación geométrica y analítica, sus propiedades estadísticas y su aplicación práctica, constituyendo el trabajo central sobre el que se apoya este trabajo.

3.1. Estimación y principales características

En presencia de multicolinealidad, la matriz $\mathbf{X}^T \mathbf{X}$ se vuelve casi singular, provocando que su inversa contenga valores muy grandes, en consecuencia, se inflan las varianzas de los estimadores de MCO.

La regresión alzada modifica una de las columnas, digamos $\mathbf{X}_i$, mediante:

$$\tilde{\mathbf{X}}_i = \mathbf{X}_i + \lambda \mathbf{e}_i, \quad \lambda \geq 0, \quad i = 2, \dots, p,$$

donde $\mathbf{e}_i$ son los residuos de la regresión auxiliar de $\mathbf{X}_i$ sobre $\mathbf{X}_{-i}$. En tal caso, la nueva matriz de diseño es $\tilde{\mathbf{X}} = [\mathbf{X}_{-i}, \tilde{\mathbf{X}}_i]$.

Esta transformación hace que la variable alzada se distancie de la combinación lineal formada por el resto de regresores, reduciendo así la colinealidad sin eliminar información del modelo. Entonces el estimador alzado siguiente se obtiene estimando por MCO el modelo alzado $\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{u}}$:

$$\hat{\boldsymbol{\beta}}(\lambda) = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{y},$$

donde la notación usada ilustra que el estimador alzado depende del parámetro de alzado λ .

Dicho estimador puede escribirse como:

$$\hat{\boldsymbol{\beta}}(\lambda) = \mathbf{M}_\lambda^{-1} \hat{\boldsymbol{\beta}},$$

donde $\mathbf{M}_\lambda$ es la matriz de transformación lineal que recoge el efecto del alzado sobre la matriz original $\mathbf{X}$:

$$M_\lambda = \begin{pmatrix} 1 & 0 & \cdots & 0 & -\lambda\hat{\alpha}_0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & -\lambda\hat{\alpha}_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & -\lambda\hat{\alpha}_{i-1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 + \lambda & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & -\lambda\hat{\alpha}_{i+1} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & -\lambda\hat{\alpha}_p & 0 & \cdots & 1 \end{pmatrix}.$$

Se observa que el estimador es sesgado a no ser que $\mathbf{M}_\lambda^{-1} = I$, lo cual sólo ocurre cuando $\lambda = 0$.

No obstante, a pesar del sesgo, la regresión alzada mantiene invariable las medidas de bondad de ajuste del modelo, puesto que los residuos del alzado coinciden exactamente con los del modelo inicial, $\mathbf{e}(\lambda) = \mathbf{e}$, donde $\mathbf{e}$ son los residuos MCO y $\mathbf{e}(\lambda)$ son los residuos del modelo alzado. En consecuencia, puesto que el modelo inicial y alzado tienen la misma variable dependiente:

- ambos modelos proporcionan la misma estimación de la varianza de la perturbación aleatoria: $\hat{\sigma}^2(\lambda) = \hat{\sigma}^2$,
- la suma de cuadrados de los residuos también coincide, por ende, el coeficiente de determinación de ambos modelos es el mismo: $R^2(\lambda) = R^2$, y
- el estadístico F de significación conjunta permanece inalterado.

Esto es lo que la distingue de la regresión ridge, que sí altera dichas medidas.

Puesto que se trata de un estimador sesgado, es interesante analizar su error cuadrático medio (ECM). El ECM del estimador alzado es:

$$\text{ECM}(\hat{\beta}(\lambda)) = \sigma^2 \text{tr} \left((\mathbf{X}_{-i}^\top \mathbf{X}_{-i})^{-1} \right) + \left(1 + \sum_{j=0, j \neq i}^p \hat{\alpha}_j^2 \right) \beta_i^2 \frac{\lambda^2 + h}{(1 + \lambda)^2},$$

donde $h = \frac{\sigma^2}{\beta_i^2 \mathbf{e}_i^\top \mathbf{e}_i}$, $i = 2, \dots, p$. El valor óptimo:

$$\lambda_{\min} = \frac{\hat{\sigma}^2}{\hat{\beta}_i^2 \mathbf{e}_i^\top \mathbf{e}_i}$$

minimiza el ECM y mejora siempre respecto al de MCO, es decir, hace que el estimador alzado tenga un ECM menor que el estimador de MCO.

Finalmente, en cuanto a la significación individual, para la variable alzada se tiene que:

$$t_{\text{exp}}(\beta(\lambda)_i) = \left| \frac{\mathbf{e}_i^\top \mathbf{y}}{\hat{\sigma} \sqrt{\mathbf{e}_i^\top \mathbf{e}_i}} \right| = t_{\text{exp}}(\beta_i),$$

es decir, el valor experimental coincide con el del modelo inicial (no cambia).

Por otro lado los estadísticos experimentales del resto de variables dependen de λ , por lo que cambian con respecto a los del modelo inicial, aunque tienden a estabilizarse al crecer este parámetro.

3.2. Multicolinealidad: FIV, número de condición y variabilidad

Comentamos a continuación cómo se adaptan las principales herramientas de detección de la multicolinealidad al ambiente de la regresión alzada.

- **Factor de Inflación de la Varianza:** el FIV en la regresión alzada viene dado por:

$$\text{FIV}(l, \lambda) = \frac{1}{1 - R_l^2(\lambda)}, \quad l = 2, \dots, p,$$

donde para calcular $R_l^2(\lambda)$ es necesario distinguir dos casos, dependiendo de cuál sea la variable dependiente de la regresión auxiliar: a) si es la variable alzada, $\tilde{\mathbf{X}}_i$ con $i = 2, \dots, p$; b) si no es la variable alzada, $\mathbf{X}_j$ con $j = 2, \dots, p$ siendo $j \neq i$.

En Salmerón et al. [25] se demuestra que el $\text{FIV}(l, \lambda)$ posee las siguientes propiedades deseables:

- es continuo $\text{FIV}(l, 0) = \text{FIV}(l)$ para $l = 2, \dots, p$,
 - es monótono decreciente en el parámetro de alzado $\text{FIV}(l, \lambda)$ disminuye al aumentar λ , y
 - es siempre mayor o igual que uno $\text{FIV}(l, \lambda) \geq 1$, para $\lambda \geq 0$ y $l = 2, \dots, p$.
- **Número de condición (NC):** el número de condición en la regresión alzada viene dado por la siguiente expresión:

$$K(\tilde{\mathbf{X}}, \lambda) = \sqrt{\frac{\tilde{\xi}_{\text{máx}}}{\tilde{\xi}_{\text{mín}}}},$$

donde $\tilde{\xi}_{\text{máx}}$ y $\tilde{\xi}_{\text{mín}}$ son, respectivamente, el mayor y el menor autovalor de la matriz $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$.

Del Hierro et al. [4] demuestra que $K(\tilde{\mathbf{X}}, \lambda)$ tiene las propiedades de ser continua, monótona en el parámetro de alzado de forma que $K(\tilde{\mathbf{X}}, \lambda)$ decrece si λ aumenta y es mayor que 1.

- **Coefficiente de variación:** en Salmerón et al. [28] se muestra que el coeficiente de variación de la variable alzada $\tilde{\mathbf{X}}_i$ es igual a:

$$CV(\tilde{\mathbf{X}}_i) = \frac{\sqrt{\text{var}(\mathbf{X}_i) + (\lambda^2 + 2\lambda) \cdot \text{var}(\mathbf{e}_i)}}{\bar{\mathbf{X}}_i},$$

de forma que es mayor o igual al coeficiente de variación de la variable $\mathbf{X}_i$ (es decir, se mitiga la multicolinealidad de tipo no esencial).

3.3. Selección de la variable a alzar y del factor de alzado

Teniendo en cuenta los resultados de los apartados anteriores, los diferentes criterios para seleccionar la variable a alzar son:

- Alzar la variable cuyo coeficiente asociado sea inicialmente significativamente distinto de cero.
- Alzar la variable con el FIV más alto.
- Alzar la variable con el CV más bajo.

En relación con el factor de alzado encontramos diferentes posibilidades para su estimación:

- Seleccionar el valor que minimice el ECM.
- Seleccionar, si existiese, el valor que proporcione valores del FIV o del NC más bajos.

3.4. Relación con la residualización

La residualización fue estudiada de forma exhaustiva por García et al. [9], este método de estimación consiste en estimar el modelo residualizado $\mathbf{y} = \mathbf{X}_O \boldsymbol{\gamma} + \mathbf{v}$ donde la nueva matriz de diseño es $\mathbf{X}_O = [\mathbf{X}_{-i}, \mathbf{e}_i]$, siendo $\mathbf{e}_i$ los residuos de la regresión auxiliar de $\mathbf{X}_i$ sobre el resto de variables independientes $\mathbf{X}_{-i}$ (es decir, la misma que para calcular el FIV).

Comparando la matriz de diseño de la regresión alzada y residualizada:

$$\tilde{\mathbf{X}} = [\mathbf{X}_{-i}, \mathbf{X}_i + \lambda \mathbf{e}_i], \quad \mathbf{X}_O = [\mathbf{X}_{-i}, \mathbf{e}_i],$$

parece claro que para un valor de λ suficientemente alto, la influencia de $\mathbf{X}_i$ en $\tilde{\mathbf{X}}_i$ será testimonial, de forma que $\tilde{\mathbf{X}}_i$ será proporcional a $\mathbf{e}_i$:

$$\tilde{\mathbf{X}} = [\mathbf{X}_{-i}, \mathbf{X}_i + \lambda \mathbf{e}_i] \xrightarrow[\lambda \rightarrow +\infty]{} [\mathbf{X}_{-i}, \lambda \mathbf{e}_i] \approx [\mathbf{X}_{-i}, \mathbf{e}_i] = \mathbf{X}_O.$$

Es decir, la regresión alzada converge hacia la residualizada cuando $\lambda \rightarrow +\infty$ (para más detalles sobre esta relación ver García et al. [9]).

4. Aplicación de la metodología del alzado a modelos de regresión con interacciones

Como se ha comentado en la introducción, el objetivo del presente trabajo es el de analizar la existencia de un problema de multicolinealidad preocupante cuando en el modelo de regresión lineal (1) se tienen interacciones. Sin pérdida de generalidad y con el objetivo de simplificar el modelo, se considerará el caso más sencillo posible dado por la expresión:

$$\mathbf{y} = \beta_1 + \beta_2 \mathbf{D} + \beta_3 \mathbf{X} + \beta_4 (\mathbf{D} \times \mathbf{X}) + \mathbf{u}, \quad (5)$$

donde $\mathbf{D}$ es una variable binaria y $\mathbf{X}$ es continua.

Concretamente, aplicaremos el método del alzado al siguiente modelo:

$$\mathbf{ts} = \beta_1 + \beta_2 \mathbf{tp} + \beta_3 \mathbf{m} + \beta_4 (\mathbf{tp} \times \mathbf{m}) + \mathbf{u}, \quad (6)$$

donde $\mathbf{ts}$ es la tasa de suicidio (medida en porcentaje) en cada una de las Comunidades y Ciudades Autónomas de España en el año 2022, $\mathbf{tp}$ es una variable binaria que toma el valor 1 si la tasa de paro de la correspondiente comunidad es superior a la media nacional y a 0 en caso contrario y $\mathbf{m}$ es el porcentaje de población femenina en cada comunidad. Los datos disponibles se muestran en la Tabla 1.

Mediante este modelo se puede analizar si la influencia del porcentaje de población femenina sobre la tasa de suicidio se ve moderada por la tasa de paro. Es decir, si difiere en aquellas comunidades donde la tasa de paro es superior a la media nacional de las que no.

En este sentido, teniendo presente el modelo (6) y siguiendo las indicaciones de Brambor et al. [2], se tiene que:

- β_2 captura el efecto de $\mathbf{tp}$ sobre $\mathbf{ts}$ cuando $\mathbf{m}$ es cero ya que:

$$E[\mathbf{ts}|\mathbf{tp} = 1] - E[\mathbf{ts}|\mathbf{tp} = 0] = \beta_2 + \beta_4 \mathbf{m}.$$

- β_3 captura el efecto de $\mathbf{m}$ sobre $\mathbf{ts}$ cuando $\mathbf{tp}$ es cero ya que:

$$\frac{\partial \mathbf{ts}}{\partial \mathbf{m}} = \beta_3 + \beta_4 \mathbf{tp}.$$

- $\beta_3 + \beta_4$ captura el efecto de $\mathbf{m}$ sobre $\mathbf{ts}$ cuando $\mathbf{tp}$ es uno.

Gráficamente, la interpretación anterior se puede resumir como se muestra en la Figura 1. Se observa que el coeficiente β_2 implica que ambas rectas tengan distinto origen y β_4 que haya distinta pendiente. En consecuencia, si $\beta_2 = 0$ ambas rectas parten del mismo origen y si $\beta_4 = 0$ la misma pendiente.

Finalmente, destacar que el código elaborado en R [20] para obtener los resultados que vienen a continuación está disponible en <https://github.com/pilarropu/Trabajo-Fin-de-Master>.

Tabla 1: Datos disponibles sobre la tasa de suicidio en España para el año 2022

Obs.	ts	tp	m	tp × m
Andalucía	9.71	1	50.74	50.74
Aragón	8.54	0	50.58	0
Asturias	12.53	0	52.31	0
Islas Baleares	9.10	1	50.22	50.22
Canarias	10.59	1	50.62	50.62
Cantabria	10.05	0	51.53	0
Castilla y León	10.11	0	50.78	0
Castilla La Mancha	8.88	1	49.93	49.93
Cataluña	7.84	0	50.80	0
Comunidad Valenciana	8.62	0	50.77	0
Extremadura	9.00	1	50.56	50.56
Galicia	12.17	0	51.88	0
Comunidad de Madrid	5.92	0	52.15	0
Región de Murcia	8.89	0	49.90	0
Navarra	8.53	0	50.52	0
País Vasco	7.87	0	51.38	0
Rioja	9.35	0	50.65	0
Ceuta	2.41	1	49.54	49.54
Melilla	1.17	1	49.54	49.54

$$\mathbf{ts} = \beta_1 + \beta_2 \mathbf{tp} + \beta_3 \mathbf{m} + \beta_4 (\mathbf{tp} \times \mathbf{m}) + \mathbf{u}$$

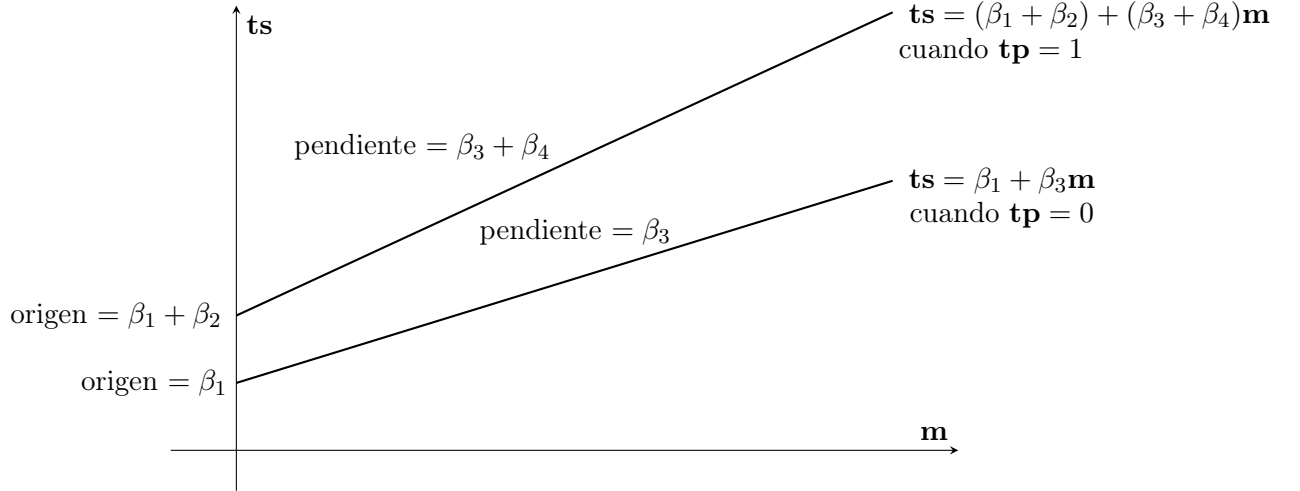


Figura 1: Ilustración gráfica del modelo de interacción (6)

4.1. Estimación por Mínimos Cuadrados Ordinarios

En primer lugar, se estima el modelo (6) por el método de MCO. A partir de los resultados mostrados en la segunda columna de la Tabla 2 se tiene que:

- El modelo es globalmente válido ya que el p-valor asociado es menor que el valor de significación del 5 % (todas las interpretaciones futuras se realizarán teniendo en cuenta este nivel de significación).
- Interpretación del coeficiente de **tp**: En las Comunidades y Ciudades Autónomas donde la tasa de paro es superior a la media nacional, la tasa de suicidio es menor que en aquellas donde es inferior cuando el porcentaje de población femenina de la correspondiente comunidad es cero (**m** = 0).
- Interpretación del coeficiente de **m**: El porcentaje de población femenina en cada Comunidad y Ciudad Autónoma no influye en su tasa de suicidio cuando la tasa de paro de la correspondiente comunidad es inferior a la media nacional (**tp** = 0).
- Sin embargo, interpretando el coeficiente de la interacción, el efecto del porcentaje de población femenina sí es inuyente en la tasa de suicidio en aquellas Comunidades y Ciudades Autónomas donde la tasa de paro es superior a la media nacional (**tp** = 1).
- Puesto que las estimaciones de los coeficientes β_2 y β_4 son significativamente distintas de cero, se tendría distinto origen y pendiente en las rectas dadas al distinguir entre las comunidades con tasa de paro superior e inferior a la media.

Si bien no se observan coeficientes estimados con signos conflictivos con las correlaciones simples entre la variable dependiente y las independientes continuas (análisis numérico del modelo) ni contradicciones entre los contrastes de significación individual y conjunta (análisis estadístico del modelo), sí se observa un valor estimado del coeficiente de **tp** anormalmente alto: en las Comunidades y Ciudades Autónomas donde la tasa de paro es superior a la media nacional y el porcentaje de población femenina es nula, la tasa de suicidio es un 298.472 % menor que en aquellas donde es inferior. Por tal motivo, a continuación se estudia el grado de multicolinealidad existente en el modelo.

Puesto que entre las variables independientes hay una variable binaria, no es aconsejable el uso del factor de inflación de la varianza ya que en el caso de dicha variable se estaría usando el coeficiente de determinación de una regresión donde la variable dependiente es binaria (ver, por ejemplo, Salmerón et al. [27]). Por tal motivo, en este trabajo se opta por usar el número de condición (NC).

Por tal motivo, partiendo de la matriz de diseño del modelo (1) dada por $\mathbf{X} = [cte \ \mathbf{tp} \ \mathbf{m} \ \mathbf{tp} \times \mathbf{m}]$, se calcula el NC en las siguientes situaciones (ver Tabla 2):

Tabla 2: Estimaciones del modelo de interacciones del modelo original, transformado mediante el alzado y centrado (entre paréntesis la desviación estándar en el caso de los coeficientes y el p-valor en el resto; *p<0.1; **p<0.05; ***p<0.01)

	MCO	Alzado ($\lambda = 16$)	Centrado
Constante	-26.97 (39.423)	7.074*** (2.382)	8.956*** (0.607)
tp	-298.472*** (86.04)	-19.378*** (5.14)	2.247 (1.302)
m / $\tilde{\mathbf{m}}$ / $\mathbf{m}_c$	0.708 (0.771)	0.042 (0.045)	0.708 (0.771)
interacción	5.925*** (1.708)	0.349*** (0.1)	5.925*** (1.708)
n	19	19	19
R^2	0.619	0.619	0.619
F_{exp} sig. conjunta	8.138*** (0.001883)	8.138*** (0.001883)	8.138*** (0.001883)
ECM	8960.539	32.10437	5.578386
corr(ts , m)	0.5603356** (0.01258)	0.4574185** (0.04894)	0.5603356** (0.01258)
corr(ts , tp $\times$ m)	-0.3351441 (0.1607)	-0.2010429 (0.4092)	0.7317676*** (0.0003693)
NC sin cte	238.9524	14.15295	3.371368
NC con cte	332.4112	19.70192	3.612554
NC sin cte y sin tp	1.999483	1.982361	1.974194
NC con cte y sin tp	179.9288	10.70572	2.706405
NC sin cte y sin m	214.3817	12.68922	2.883579
NC con cte y sin m	239.3712	14.19873	3.398603
NC sin cte y sin tp $\times$ m	1.999402	1.960152	1.653938
NC con cte y sin tp $\times$ m	180.6231	10.78243	2.736447
CV(tp)	1.309	1.309	1.309
CV(m)	0.015	0.2109	$5.18 \cdot 10^{14}$
CV(tp $\times$ m)	1.309	1.335	1.844

- NC sin cte, $\mathbf{X} = [\mathbf{tp}, \mathbf{m}, \mathbf{tp} \times \mathbf{m}]$, es igual a 238.9524, lo que indica que el grado de multicolinealidad existente del tipo esencial es preocupante.
- NC con cte, $\mathbf{X} = [cte, \mathbf{tp}, \mathbf{m}, \mathbf{tp} \times \mathbf{m}]$, es igual a 332.4112, lo que indica que la inclusión del término independiente aumenta las relaciones lineales del modelo.
- NC sin cte y sin $\mathbf{tp}$, $\mathbf{X} = [\mathbf{m}, \mathbf{tp} \times \mathbf{m}]$, es igual a 1.999483, lo que indica que la relación lineal entre $\mathbf{m}$ y la interacción es baja.
- NC con cte y sin $\mathbf{tp}$, $\mathbf{X} = [cte, \mathbf{m}, \mathbf{tp} \times \mathbf{m}]$, es igual a 179.9288, lo que indica de nuevo que la inclusión del término independiente aumenta (y mucho) las relaciones lineales del modelo.
- NC sin cte y sin $\mathbf{m}$, $\mathbf{X} = [\mathbf{tp}, \mathbf{tp} \times \mathbf{m}]$, es igual a 214.3817, lo que indica que la relación lineal entre $\mathbf{tp}$ y la interacción es alta.
- NC con cte y sin $\mathbf{m}$, $\mathbf{X} = [cte, \mathbf{tp}, \mathbf{tp} \times \mathbf{m}]$, es igual a 239.3712, lo que indica de nuevo que la inclusión del término independiente aumenta (levemente) las relaciones lineales del modelo.
- NC sin cte y sin $\mathbf{tp} \times \mathbf{m}$, $\mathbf{X} = [\mathbf{tp}, \mathbf{m}]$, es igual a 1.999402, lo que indica que la relación lineal entre $\mathbf{tp}$ y $\mathbf{m}$ es baja.
- NC con cte y sin $\mathbf{tp} \times \mathbf{m}$, $\mathbf{X} = [cte, \mathbf{tp}, \mathbf{m}]$, es igual a 180.6231, lo que indica de nuevo que la inclusión del término independiente aumenta (y mucho) las relaciones lineales del modelo (especialmente cuando está $\mathbf{m}$ en el modelo).

De los resultados anteriores se deduce que:

- La relación lineal entre $\mathbf{tp}$ y $\mathbf{m}$ es baja, por lo que el grado de multicolinealidad esencial es preocupante una vez que se incluye el término de interacción.
- El término independiente tiene un papel relevante en el grado de multicolinealidad existente, es decir, la multicolinealidad de tipo no esencial es preocupante.
Esta suposición se refrenda al existir un coeficiente de variación (CV), el de $\mathbf{m}$, inferior a 0.1002506 (ver, para más detalles, Salmerón et al. [24]).

En definitiva, la variable $\mathbf{m}$ es la causante de que el grado de multicolinealidad aproximada de tipo no esencial sea preocupante y la interacción de que lo sea la de tipo esencial.

4.2. Estimación mediante la regresión alzada

Dentro de los métodos de estimación alternativos a los MCO para tratar el problema de multicolinealidad en un modelo de regresión lineal múltiple, a continuación usaremos la regresión alzada, la cual ha sido presentada en la sección 3.

Como se ha descrito, en primer lugar se ha de decidir qué variable se ha de alzar. Si bien en Salmerón et al. [28] se describen distintas opciones (algunas de las cuales han sido recogidas en el apartado 3.3), éstas se hacen para un modelo de regresión en el que no hay interacciones. Por tal motivo, a continuación se van a explorar las siguientes posibilidades:

- Alzar el término de interacción a partir de la regresión auxiliar $(\mathbf{tp} \times \mathbf{m}) = \alpha_1 + \alpha_2 \mathbf{tp} + \alpha_3 \mathbf{m} + \mathbf{v}$.
- Alzar la variable cuantitativa, $\mathbf{m}$, a partir de las regresiones auxiliares:
 - $\mathbf{m} = \alpha_1 + \alpha_2 \mathbf{tp} + \mathbf{v}$ (sin la interacción) ó
 - $\mathbf{m} = \alpha_1 + \alpha_2 \mathbf{tp} + \alpha_3 (\mathbf{tp} \times \mathbf{m}) + \mathbf{v}$ (con la interacción).

Adviértase que se descarta la posibilidad de alzar la variable binaria $\mathbf{tp}$ ya que anteriormente se han identificado a $\mathbf{m}$ y la interacción como las causantes de los problemas de multicolinealidad preocupante.

4.2.1. Alzando la interacción

En este caso, el modelo alzado corresponde a la siguiente expresión:

$$\mathbf{ts} = \gamma_1 + \gamma_2 \mathbf{tp} + \gamma_3 \mathbf{m} + \gamma_4 \left(\widetilde{\mathbf{tp} \times \mathbf{m}} \right) + \varepsilon,$$

donde la variable alzada, $\widetilde{\mathbf{tp} \times \mathbf{m}} = (\mathbf{tp} \times \mathbf{m}) + \lambda \cdot \mathbf{e}$, se alza a partir de los residuos, $\mathbf{e}$, de la regresión auxiliar $(\mathbf{tp} \times \mathbf{m}) = \alpha_1 + \alpha_2 \mathbf{tp} + \alpha_3 \mathbf{m} + \mathbf{v}$.

Con el objetivo de encontrar el valor de λ que minimice el grado de multicolinealidad preocupante existente, se calcula el NC para distintos valores de λ . Los resultados mostrados en la Tabla 3 indican que no existe un valor de λ que consiga dejar al NC por debajo del umbral de 20. Concretamente, se vislumbra que la asíntota a la que se refiere Salmerón et al. [28] en su apartado 3.4 está en torno al 180.

En definitiva, alzar el término de interacción no permite mitigar el grado de multicolinealidad existente en el modelo inicial.

Tabla 3: Valores del NC cuando se alza la interacción para $\lambda \in \{0, 100, 200, \dots, 1500\}$

λ	Número de Condición
0	332.41
100	196.93
200	189.02
300	185.31
400	183.53
500	182.58
600	182.02
700	181.67
800	181.43
900	181.27
1000	181.15
1100	181.06
1200	180.99
1300	180.94
1400	180.89
1500	180.86

4.2.2. Alzando la variable cuantitativa sin tener en cuenta la interacción

En este caso, el modelo alzado corresponde a la siguiente expresión:

$$\mathbf{ts} = \gamma_1 + \gamma_2 \mathbf{tp} + \gamma_3 \tilde{\mathbf{m}} + \gamma_4 (\mathbf{tp} \times \tilde{\mathbf{m}}) + \varepsilon, \quad (7)$$

donde la variable alzada, $\tilde{\mathbf{m}} = \mathbf{m} + \lambda \cdot \mathbf{e}$, se alza a partir de los residuos, $\mathbf{e}$, de la regresión auxiliar $\mathbf{m} = \alpha_1 + \alpha_2 \mathbf{tp} + \mathbf{v}$.

Al calcular el número de condición para distintos valores de λ , se observa (ver Tabla 4) que cuando $\lambda > 15$ el NC es inferior a 20 (umbral establecido como preocupante). Por tal motivo se decide estimar por MCO el modelo alzado (7) considerando que $\lambda = 16$.

Adviértase que podría haberse planteado el modelo alzado:

$$\mathbf{ts} = \gamma_1 + \gamma_2 \mathbf{tp} + \gamma_3 \tilde{\mathbf{m}} + \gamma_4 (\mathbf{tp} \times \mathbf{m}) + \varepsilon, \quad (8)$$

sin embargo, esta opción no tiene mucho sentido ya que la relación lineal entre $\mathbf{tp}$ y $\mathbf{tp} \times \mathbf{m}$ no se estaría modificando. Por tanto, si era preocupante de inicio, seguirá siéndolo en el modelo alzado. Esta situación queda clara a partir de los valores obtenidos para el NC mostrados en la Tabla 4.

Estimado el modelo (7) para $\lambda = 16$, los resultados se muestran en la tercera columna de la Tabla 2. Se observa que:

Tabla 4: Valores del NC cuando se alza la variable cuantitativa (sin tener en cuenta la interacción) para $\lambda \in \{0, 5, 10, \dots, 100\}$

λ	NC para (7)	NC para (8)
0	332.41	332.41
5	55.45	304.70
10	30.31	303.97
15	20.92	303.37
20	16.01	302.65
25	13.01	301.81
30	10.99	300.86
35	9.55	299.81
40	8.46	298.70
45	7.62	297.53
50	6.95	296.34
55	6.41	295.15
60	5.96	293.95
65	5.59	292.78
70	5.27	291.63
75	5.00	290.52
80	4.76	289.44
85	4.55	288.41
90	4.37	287.43
95	4.21	286.49
100	4.07	285.60

- Como sabíamos, las características globales del modelo no se han modificado. Luego el modelo sigue siendo válido.
- El coeficiente de variación de la variable alzada (y de la interacción) ha aumentado superando el límite establecido como preocupante e indicando que se ha mitigado la multicolinealidad de tipo no esencial.
- Todo los valores del NC están por debajo del umbral, 20, establecido como preocupante. Es decir, también se ha mitigado la multicolinealidad de tipo esencial.
- Se obtiene un error cuadrático medio (ECM) inferior al de MCO. Esta cuestión es relevante ya que los estimadores obtenidos son sesgados.
- El término independiente es significativamente distinto de cero, de forma que la tasa de suicidio es del 7.074 % cuando la tasa de paro de la Comunidad o Ciudad Autónoma es inferior a la media nacional ($\mathbf{tp} = 0$) y el porcentaje de población femenina es cero ($\mathbf{m} = 0$).
- Si bien la estimación del coeficiente de $\mathbf{tp}$ se ha reducido en cuanto a magnitud se refiere (y sigue siendo significativamente distinto de cero), sigue proporcionando una interpretación que puede estar en entredicho al proporcionar una estimación negativa ($-12.304 = 7.074 - 19.378$) para $\mathbf{ts}$ cuando la tasa de paro es superior a la media nacional ($\mathbf{ts} = 0$) y el porcentaje de población femenina es inexistente ($\mathbf{m} = 0$). Probablemente, el problema de la interpretación resida en que el valor $\mathbf{m} = 0$ no tiene sentido teniendo en cuenta las observaciones de la Tabla 1, pero se observa que sí se ha mitigado la interpretación inicial ($-325.442 = -26.97 - 298.472$), que era aún menos defendible.
- La interpretación del coeficiente de $\mathbf{m}$, indica que el porcentaje de población femenina en cada Comunidad y Ciudad Autónoma no influye en su tasa de suicidio cuando la tasa de paro de la correspondiente comunidad es inferior a la media nacional ($\mathbf{tp} = 0$).
- Sin embargo, de la interpretación del coeficiente de la interacción se deduce que el efecto del porcentaje de población femenina sí es inuyente en la tasa de suicidio en aquellas Comunidades y Ciudades Autónomas donde la tasa de paro es superior a la media nacional ($\mathbf{tp} = 1$). En este caso un 0.349 % más.
- Puesto que las estimaciones de los coeficientes β_2 y β_4 son significativamente distintos de cero, se tendría distinto origen y pendiente en las rectas dadas al distinguir entre las comunidades con tasa de paro superior e inferior a la media (ver Figura 1).

Tabla 5: Valores del NC cuando se alza la variable cuantitativa (teniendo en cuenta la interacción) para $\lambda \in \{0, 1000, 2000, \dots, 10000\}$

λ	NC para (7)	NC para (8)
0	332.41	332.41
1000	239.76	239.76
2000	239.47	239.47
3000	239.42	239.42
4000	239.40	239.40
5000	239.39	239.39
6000	239.38	239.38
7000	239.38	239.38
8000	239.38	239.38
9000	239.38	239.38
10000	239.38	239.38

4.2.3. Alzando la variable cuantitativa teniendo en cuenta la interacción

En este caso, el modelo alzado corresponde a la expresión (7) donde la variable alzada, $\tilde{\mathbf{m}} = \mathbf{m} + \lambda \cdot \mathbf{e}$, se alza a partir de los residuos, $\mathbf{e}$, de la regresión auxiliar $\mathbf{m} = \alpha_1 + \alpha_2 \mathbf{tp} + \alpha_3 (\mathbf{tp} \times \tilde{\mathbf{m}}) + \mathbf{v}$.

Aunque no tenga sentido, tal y como se ha comentado antes, también se considerará el modelo alzado de la expresión (8) con propósitos comparativos.

En la Tabla 5 se muestra el número de condición para los modelos (7) y (8) alzados según se acaba de indicar. Se observa que al tener en cuenta la interacción en la regresión auxiliar, ambos modelos alzados conducen a los mismos valores del número de condición.

En consecuencia, la matriz de diseño del modelo (7), $\tilde{\mathbf{X}}_1 = [cte, \mathbf{tp}, \tilde{\mathbf{m}}, \mathbf{tp} \times \tilde{\mathbf{m}}]$, y del modelo (8), $\tilde{\mathbf{X}}_2 = [cte, \mathbf{tp}, \tilde{\mathbf{m}}, \mathbf{tp} \times \mathbf{m}]$, coinciden. Por lo que se ha de verificar que $\mathbf{tp} \times \tilde{\mathbf{m}} = \mathbf{tp} \times \mathbf{m}$. Este hecho se observa en las siguientes matrices que conducen a los valores del NC mostrados en la Tabla 5 para $\lambda = 10000$:

$$\tilde{\mathbf{X}}_1 = \begin{pmatrix} cte & \mathbf{tp} & \tilde{\mathbf{m}} & \mathbf{tp} \times \tilde{\mathbf{m}} \\ 1 & 1 & 50,740 & 50,74 \\ 1 & 0 & -5191,087 & 0 \\ 1 & 0 & 12110,643 & 0 \\ 1 & 1 & 50,220 & 50,22 \\ 1 & 1 & 50,620 & 50,62 \\ 1 & 0 & 4309,863 & 0 \\ 1 & 0 & -3190,887 & 0 \\ 1 & 1 & 49,930 & 49,93 \\ 1 & 0 & -2990,867 & 0 \\ 1 & 0 & -3290,897 & 0 \\ 1 & 1 & 50,560 & 50,56 \\ 1 & 0 & 7810,213 & 0 \\ 1 & 0 & 10510,483 & 0 \\ 1 & 0 & -11991,767 & 0 \\ 1 & 0 & -5791,147 & 0 \\ 1 & 0 & 2809,713 & 0 \\ 1 & 0 & -4491,017 & 0 \\ 1 & 1 & 49,540 & 49,54 \\ 1 & 1 & 49,540 & 49,54 \end{pmatrix}, \quad \tilde{\mathbf{X}}_2 = \begin{pmatrix} cte & \mathbf{tp} & \tilde{\mathbf{m}} & \mathbf{tp} \times \mathbf{m} \\ 1 & 1 & 50,740 & 50,74 \\ 1 & 0 & -5191,087 & 0 \\ 1 & 0 & 12110,643 & 0 \\ 1 & 1 & 50,220 & 50,22 \\ 1 & 1 & 50,620 & 50,62 \\ 1 & 0 & 4309,863 & 0 \\ 1 & 0 & -3190,887 & 0 \\ 1 & 1 & 49,930 & 49,93 \\ 1 & 0 & -2990,867 & 0 \\ 1 & 0 & -3290,897 & 0 \\ 1 & 1 & 50,560 & 50,56 \\ 1 & 0 & 7810,213 & 0 \\ 1 & 0 & 10510,483 & 0 \\ 1 & 0 & -11991,767 & 0 \\ 1 & 0 & -5791,147 & 0 \\ 1 & 0 & 2809,713 & 0 \\ 1 & 0 & -4491,017 & 0 \\ 1 & 1 & 49,540 & 49,54 \\ 1 & 1 & 49,540 & 49,54 \end{pmatrix}.$$

Teniendo presente la Tabla 1, se observa que $\tilde{\mathbf{m}} = \mathbf{m}$ cuando $\mathbf{tp} = 1$ y que $\tilde{\mathbf{m}} \neq \mathbf{m}$ cuando $\mathbf{tp} = 0$. En consecuencia $\mathbf{tp} \times \tilde{\mathbf{m}} = \mathbf{tp} \times \mathbf{m}$.

Es evidente que $\mathbf{tp} \times \tilde{\mathbf{m}} \neq \mathbf{tp} \times \mathbf{m}$ cuando en la regresión alzada usada para alzar la variable cuantitativa del modelo no se tiene en cuenta a la interacción ya que los valores del NC mostrados en la Tabla 4 son distintos entre sí para los distintos valores de λ considerados. Así, para $\lambda = 100$, las matrices de diseño anteriores son:

$$\tilde{\mathbf{X}}_1 = \begin{pmatrix} cte & \mathbf{tp} & \tilde{\mathbf{m}} & \mathbf{tp} \times \tilde{\mathbf{m}} \\ 1 & 1 & 108,311429 & 108,31143 \\ 1 & 0 & -1,836667 & 0 \\ 1 & 0 & 172,893333 & 0 \\ 1 & 1 & 55,791429 & 55,79143 \\ 1 & 1 & 96,191429 & 96,19143 \\ 1 & 0 & 94,113333 & 0 \\ 1 & 0 & 18,363333 & 0 \\ 1 & 1 & 26,501429 & 26,50143 \\ 1 & 0 & 20,383333 & 0 \\ 1 & 0 & 17,353333 & 0 \\ 1 & 1 & 90,131429 & 90,13143 \\ 1 & 0 & 129,463333 & 0 \\ 1 & 0 & 156,733333 & 0 \\ 1 & 0 & -70,516667 & 0 \\ 1 & 0 & -7,896667 & 0 \\ 1 & 0 & 78,963333 & 0 \\ 1 & 0 & 5,233333 & 0 \\ 1 & 1 & -12,888571 & -12,88857 \\ 1 & 1 & -12,888571 & -12,88857 \end{pmatrix}, \quad \tilde{\mathbf{X}}_2 = \begin{pmatrix} cte & \mathbf{tp} & \tilde{\mathbf{m}} & \mathbf{tp} \times \mathbf{m} \\ 1 & 1 & 108,311429 & 50,74 \\ 1 & 0 & -1,836667 & 0 \\ 1 & 0 & 172,893333 & 0 \\ 1 & 1 & 55,791429 & 50,22 \\ 1 & 1 & 96,191429 & 50,62 \\ 1 & 0 & 94,113333 & 0 \\ 1 & 0 & 18,363333 & 0 \\ 1 & 1 & 26,501429 & 49,93 \\ 1 & 0 & 20,383333 & 0 \\ 1 & 0 & 17,353333 & 0 \\ 1 & 1 & 90,131429 & 50,56 \\ 1 & 0 & 129,463333 & 0 \\ 1 & 0 & 156,733333 & 0 \\ 1 & 0 & -70,516667 & 0 \\ 1 & 0 & -7,896667 & 0 \\ 1 & 0 & 78,963333 & 0 \\ 1 & 0 & 5,233333 & 0 \\ 1 & 1 & -12,888571 & 49,54 \\ 1 & 1 & -12,888571 & 49,54 \end{pmatrix}.$$

En este caso se observa que $\tilde{\mathbf{m}} \neq \mathbf{m}$ tanto cuando $\mathbf{tp} = 1$ como cuando $\mathbf{tp} = 0$.

En consecuencia, puede ser interesante analizar teóricamente por qué $\mathbf{tp} \times \tilde{\mathbf{m}} = \mathbf{tp} \times \mathbf{m}$ cuando la variable $\mathbf{m}$ se alza a partir de los residuos de la regresión auxiliar $\mathbf{m} = \alpha_1 + \alpha_2 \mathbf{tp} + \alpha_3 (\mathbf{tp} \times \tilde{\mathbf{m}}) + \mathbf{v}$ y $\mathbf{tp} \times \tilde{\mathbf{m}} \neq \mathbf{tp} \times \mathbf{m}$ cuando la variable $\mathbf{m}$ se alza a partir de los residuos de la regresión auxiliar $\mathbf{m} = \alpha_1 + \alpha_2 \mathbf{tp} + \mathbf{v}$.

En cualquier caso, se confirma que no tiene sentido plantear el modelo alzado (8) en ningún caso y que tampoco se ha de incluir la interacción en la regresión auxiliar que proporciona los residuos usados para alzar la variable cuantitativa del modelo.

4.3. Estimación mediante el centrado de variables

Es bien conocido que para solventar el problema de multicolinealidad no esencial es suficiente con centrar (restar la media) la variable que provoca el problema (ver, por ejemplo, Stewart [30], Montgomery et al. [18] o Salmerón et al. [23]). Por tal motivo, a continuación se estima el siguiente modelo:

$$\mathbf{ts} = \beta_1 + \beta_2 \mathbf{tp} + \beta_3 \mathbf{m}_c + \beta_4 (\mathbf{tp} \times \mathbf{m}_c) + \mathbf{u}, \quad (9)$$

donde $\mathbf{m}_c = \mathbf{m} - \overline{\mathbf{m}}$. Los resultados se muestran en la última columna de la Tabla 2.

Se observa que:

- Las desviaciones típicas estimadas de los coeficientes estimados (entre paréntesis) han disminuido en aquellas variables no transformadas y se mantienen inalteradas con respecto al modelo inicial en aquellas donde se considera la variable $\mathbf{m}$ centrada. De forma similar, las estimaciones de las variables donde interviene la variable centrada tampoco han cambiado, al contrario de lo que ocurre en el resto de variables. Ambas cuestiones se deben a que el centrado de variables es un caso particular del método de residualización (ver Apéndice A y, especialmente, García et al. [9], para más detalles).
Es llamativa especialmente la disminución sufrida en la desviación típica estimada en el caso del término independiente, lo cual es lógico ya que juega un papel clave en la multicolinealidad de tipo no esencial solventada.
- Las características globales del modelo (coeficiente de determinación y valor experimental del contraste de significación conjunta) no cambian al igual que en el alzado. Esto se debe de nuevo a que el centrado de variables es un caso particular del método de residualización.
- El comportamiento destacado en los dos puntos anteriores es análogo al mostrado en el ejemplo de Salmerón et al. [23].
- En todas las combinaciones del cálculo del NC se obtienen valores muy próximos a su mínimo valor (el 1), lo cual indica que el grado de multicolinealidad existente no es preocupante (y casi inexistente).
- Igualmente, el coeficiente de variación de la variable centrada $\mathbf{m}$ es muy grande (debería ser infinito), lo cual indica que la relación de esta variable con el término constante se ha suprimido.

En definitiva, se observa que al solventar el problema de multicolinealidad no esencial en el modelo de interacción (6) también se ha mitigado el de tipo esencial. Además, las características globales del modelo inicial no se han visto modificadas y se obtiene el menor ECM de los tres modelos considerados.

Por último, interpretando las estimaciones obtenidas se concluye que:

- El término independiente es significativamente distinto de cero, de forma que la tasa de suicidio es del 8.956 % cuando la tasa de paro de la Comunidad o Ciudad Autónoma es inferior a la media nacional ($\mathbf{tp} = 0$) y el porcentaje de población femenina es igual a su media (que es 50.75 %) ($\mathbf{m}_c = 0$).

- La tasa de paro no influye en la tasa de suicidio cuando el porcentaje de población femenina de la correspondiente comunidad es igual a su media ($\mathbf{m}_c = 0$). Aunque el coeficiente no sea significativamente distinto de cero, adviértase que el centrar la variable cuantitativa facilita la interpretación de la variable cualitativa ya que el hecho de que la variable centrada sea igual a cero se interpreta directamente como que la variable original es igual a su media (en lugar de a cero).
- El porcentaje de población femenina no influye en la tasa de suicidio cuando la tasa de paro de la correspondiente comunidad es inferior a la media nacional ($\mathbf{tp} = 0$). Aunque no sea significativamente distinto de cero, adviértase que se ha corregido la magnitud anormalmente alta del coeficiente de la tasa de paro y que tiene un signo acorde con el del coeficiente de correlación calculado entre la variable dependiente y esta variable.
- El efecto del porcentaje de población femenina sí es influyente en la tasa de suicidio en aquellas Comunidades y Ciudades Autónomas donde la tasa de paro es superior a la media nacional ($\mathbf{tp} = 1$), concretamente, un 5.925 %. De nuevo, el coeficiente estimado de la interacción tiene el mismo signo que el de la correlación entre la variable dependiente y la interacción (siendo ambos significativamente distintos de cero).

Finalmente, puesto que las estimaciones del coeficiente β_2 no es significativamente distinta de cero y el de β_4 sí lo es, se tendría el mismo origen (desde un punto de vista estadístico) y distinta pendiente en las rectas dadas al distinguir entre las comunidades con tasa de paro superior e inferior a la media. Replicando la Figura 1, esta situación se ilustra en la Figura 2 donde se representan las estimaciones de la Tabla 2 para la estimación del modelo de interacción centrado (9).

5. Discusión de resultados

A partir del modelo centrado estimado, los resultados reflejan que los efectos de la tasa de paro y del porcentaje de población femenina sobre la tasa de suicidio no son independientes, sino que dependen del contexto económico en el que se analizan. Este comportamiento concuerda con la literatura previa, que subraya que el suicidio responde a una combinación de factores socioeconómicos, demográficos y territoriales, como muestran De Irala et al. [5] y Pérez Díez et al. [19].

Analizando la interpretación de los coeficientes observamos que:

- *La tasa de paro no influye en la tasa de suicidio cuando el porcentaje de población femenina es igual a su media.*

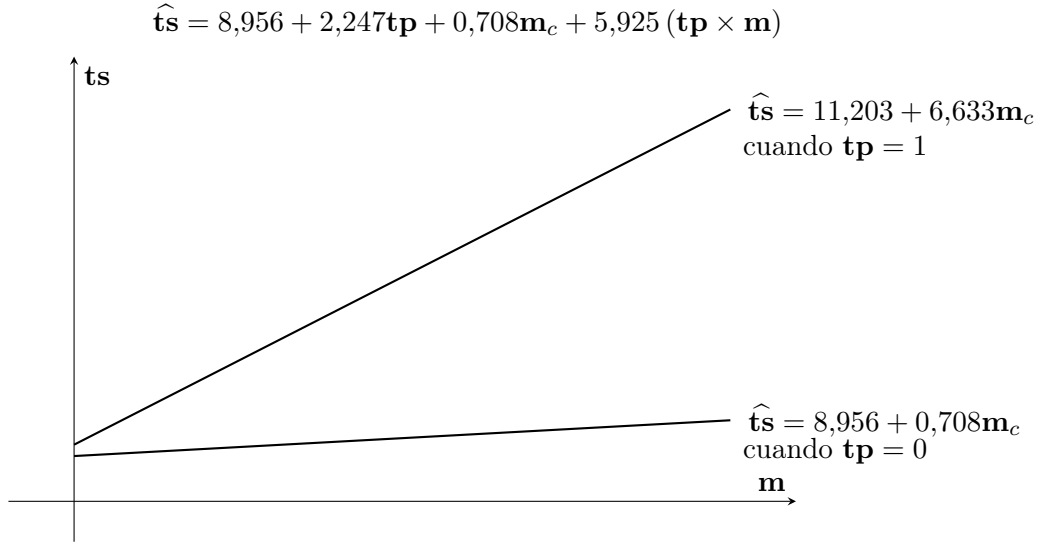


Figura 2: Ilustración gráfica del modelo de interacción centrado (9)

En términos económicos, este resultado sugiere que el desempleo considerado de forma aislada no es suficiente para explicar las diferencias territoriales en la tasa de suicidio. Esta interpretación coincide con la evidencia empírica existente, la cual señala que la relación entre desempleo y suicidio no es directa ni estable en el tiempo, sino que depende del contexto económico y social en el que se produce.

Concretamente, Iglesias et al. [14] muestran que la relación entre desempleo y suicidio en España varía según el periodo económico, lo que sugiere la existencia de mecanismos sociales o institucionales capaces de amortiguar el impacto del desempleo sobre la salud mental. En consonancia con esta evidencia, el modelo estimado muestra que en ausencia de tensiones demográficas adicionales el paro considerado de forma aislada no genera un efecto significativo sobre la tasa de suicidio.

- *La población femenina no influye en la tasa de suicidio cuando la tasa de paro es inferior a la media.*

Interpretándolo desde una óptica económica y social, indica que en entornos laborales relativamente favorables, las variaciones en la estructura demográfica de género no se traducen en cambios estadísticamente significativos en la tasa de suicidio. Estudios previos que analizan la mortalidad por suicidio a nivel territorial señalan que los factores socioeconómicos y demográficos tienden a tener un impacto más limitado cuando el contexto económico es menos adverso.

En concreto, De Irala et al. [5] muestran que la influencia de variables socioeconómicas y demográficas sobre el suicidio no es homogénea entre territorios y que su efecto tiende a intensificarse en contextos de mayor vulnerabilidad. En línea con este resultado, el modelo estimado sugiere que cuando el desempleo es bajo las posibles tensiones asociadas a la estructura de género pierden relevancia.

- *La población femenina influye positivamente en la tasa de suicidio cuando la tasa de paro es superior a la media.*

Este resultado indica que la influencia de la población femenina sobre la tasa de suicidio solo emerge en contextos de elevada vulnerabilidad económica, lo que apunta a un efecto de interacción entre factores económicos, sociales y de género. La literatura ha mostrado que los efectos del desempleo sobre la salud mental y el suicidio dependen en gran medida del nivel de desigualdad económica y del contexto institucional en el que se produce, tal y como señalan Stuckler et al. [32] y Wilkinson y Pickett [34]. Podemos analizarlo desde diferentes perspectivas:

- Desde la perspectiva territorial, estudios recientes destacan que las comunidades autónomas con mayores tasas de suicidio suelen presentar una combinación de factores económicos adversos, menor diversificación productiva y un acceso más limitado a recursos sanitarios y sociales, como muestran Pérez Díez et al. [19]. Estas características coinciden con muchas de las comunidades que presentan tasas de paro superiores a la media nacional de nuestro estudio.
- Desde una perspectiva de género, la literatura señala que aunque las mujeres presentan mayores tasas de conducta suicida, los hombres registran una mayor mortalidad por suicidio, este fenómeno se conoce como la “paradoja de género”, descrito por Canetto y Sakinofsky [3]. Este patrón se ha vinculado a normas sociales de masculinidad, mayor dificultad para expresar malestar emocional y mayor tendencia al aislamiento social, factores que se intensifican en contextos de desempleo elevado y estructuras sociales más tradicionales tal y como apuntan De Irala et al. [5] e Iglesias-García et al. [14].

En este sentido, un mayor porcentaje de población femenina en comunidades con alto desempleo puede estar asociado a una mayor asunción de responsabilidades familiares y de cuidados por parte de las mujeres, mientras que los hombres ante la pérdida del rol tradicional de proveedor experimentan una mayor presión económica y social. Este mecanismo ha sido señalado por la literatura como un factor de riesgo relevante en entornos con elevada desigualdad y menor cohesión social, como señalan Wilkinson y Pickett [34].

Relacionado con las políticas públicas, los resultados obtenidos sugieren adoptar estrategias de prevención del suicidio adaptadas a las características económicas y sociales de cada territorio. En particular, reforzando y evaluando de forma sistemática los planes

autonómicos de prevención del suicidio, analizando su grado de implementación, la dotación presupuestaria destinada a salud mental y la eficiencia en el acceso a los servicios sanitarios especializados, especialmente en comunidades con elevados niveles de desempleo estructural.

Así, es necesario mejorar la accesibilidad y la rapidez de atención por parte de los profesionales del sistema sanitario, dado que los tiempos de espera pueden constituir una barrera relevante para la detección temprana y el tratamiento de conductas de riesgo. Por último, sería fundamental profundizar hacia una mayor desagregación de la información disponible, en particular diferenciando por sexo y grupos de edad, para así poder diseñar intervenciones más específicas y eficaces.

6. Conclusiones

En este trabajo se aborda el problema de multicolinealidad aproximada en un modelo de regresión lineal con interacciones entre las variables explicativas. Este tipo de modelos son interesantes de usar cuando se considera que el efecto de una variable explicativa sobre la explicada puede estar moderada por una segunda variable explicativa. Es decir, el efecto cambia en función de los niveles considerados en esa segunda variable.

Así, en el caso de estudio planteado se considera que el efecto del porcentaje de población femenina sobre la tasa de suicidio difiere dependiendo del nivel de paro que hay en cada Comunidad y Ciudad Autónoma. Por tal motivo se ha de plantear un modelo de regresión lineal con una interacción entre el porcentaje de población femenina y la tasa de paro.

Concretamente, a partir del modelo centrado analizado, se concluye que la población femenina influye (positivamente) en la tasa de suicidio cuando la tasa de paro es superior a la media nacional y no influye cuando es inferior. Además, se deducen las siguientes cuestiones teóricas:

- Se tiene un problema de multicolinealidad preocupante de tipo no esencial entre la variable cuantitativa y el término independiente, y entre la variable binaria y la interacción del tipo esencial. También se observa que entre la variable binaria y la cuantitativa las relaciones lineales son bajas.
- En cuanto a la aplicación del método del alzado: a) alzar la interacción no mitiga el problema de multicolinealidad y b) alzar la variable cuantitativa mitiga el problema sólo cuando en la regresión auxiliar usada en el alzado no se tiene en cuenta el término de la interacción.
- Al eliminar la multicolinealidad no esencial centrando la variable cuantitativa, también se mitiga la multicolinealidad de tipo esencial. Esto sugiere que para que haya una relación lineal del tipo esencial preocupante (entre la variable binaria y

la interacción) tiene que haberla previamente del tipo no esencial (entre la variable cuantitativa y el término independiente).

- Esta afirmación tiene mucho sentido, ya que si en el término de interacción una de ellas es prácticamente constante (en este caso, $\mathbf{X} \approx cte$), entonces el término de interacción estaría añadiendo relaciones lineales al modelo (5) al ser de la forma $\mathbf{D} \times \mathbf{X} \approx cte \cdot \mathbf{D}$ ya que éste modelo se podría expresar como:

$$\begin{aligned}\mathbf{y} &= \beta_1 + \beta_2 \mathbf{D} + \beta_3 \mathbf{X} + \beta_4 (\mathbf{D} \times \mathbf{X}) + \mathbf{u} \\ &\approx \beta_1 + \beta_2 \mathbf{D} + \beta_3 \mathbf{X} + \beta_4 cte \cdot \mathbf{D} + \mathbf{u}.\end{aligned}$$

- Igualmente, esta afirmación justificaría que el tratamiento adecuado de la multicolinealidad en este tipo de modelos es el centrado de la variable que provoca el problema de multicolinealidad no esencial, siendo entonces totalmente innecesario la aplicación de técnicas más complejas como el ridge, LASSO, elastic-net o alzado.

En consecuencia, como futuras líneas de trabajo:

- Se desea comprobar las conclusiones anteriores desde un punto de vista teórico. Especialmente la tercera, ya que de confirmarse, contribuiría a clarificar la disparidad de opiniones comentada en la introducción de los trabajos de Brambor et al. [2] y Iacobucci et al. [13] sobre la necesidad de centrar o no las variables en un modelo de interacciones.
- También sería interesante analizar el caso en el que se considera la interacción entre dos variables cuantitativas.
- Como se ha indicado, sería conveniente analizar teóricamente por qué $\mathbf{tp} \times \tilde{\mathbf{m}} = \mathbf{tp} \times \mathbf{m}$ cuando la variable $\mathbf{m}$ se alza a partir de los residuos de la regresión auxiliar $\mathbf{m} = \alpha_1 + \alpha_2 \mathbf{tp} + \alpha_3 (\mathbf{tp} \times \tilde{\mathbf{m}}) + \mathbf{v}$ y $\mathbf{tp} \times \tilde{\mathbf{m}} \neq \mathbf{tp} \times \mathbf{m}$ cuando la variable $\mathbf{m}$ se alza a partir de los residuos de la regresión auxiliar $\mathbf{m} = \alpha_1 + \alpha_2 \mathbf{tp} + \mathbf{v}$.
- Puesto que la variable dependiente es una tasa (por tanto, acotada entre 0 y 1), podría considerarse que se modeliza según una distribución beta en lugar de una normal, por lo que se puede plantear un modelo lineal generalizado (GLM) para su estimación en lugar de los MCO.

A. Relación entre el centrado de variables y la residualización

Al aplicar la regresión alzada comentada en el apartado 3.4 al modelo de regresión lineal simple $\mathbf{y} = \beta_1 + \beta_2 \mathbf{X} + \mathbf{u}$, se tendría que estimar por MCO el modelo residualizado $\mathbf{y} = \gamma_1 + \gamma_2 \mathbf{e} + \mathbf{v}$ donde $\mathbf{e}$ son los residuos de la regresión auxiliar $\mathbf{X} = \alpha + \mathbf{w}$.

Puesto que en esta regresión auxiliar se verifica que $\hat{\alpha} = \bar{\mathbf{X}}$, es claro que $\mathbf{e} = \mathbf{X} - \bar{\mathbf{X}}$, es decir, el modelo residualizado se puede expresar como $\mathbf{y} = \gamma_1 + \gamma_2 (\mathbf{X} - \bar{\mathbf{X}}) + \mathbf{v}$.

En consecuencia, se tiene que el centrado de variables se puede considerar como un caso particular de la regresión residualizada. En tal caso, el apartado 3.4 conectaría el centrado de variables con la regresión alzada.

Referencias

- [1] Belsley, D.A. (1991). Conditioning Diagnostics: Collinearity and Weak Data in Regression. John Wiley and Sons, Hoboken.
- [2] Brambor, T. Clark, W.R. y Golder, M. (2006). Understanding interaction models: improving empirical analyses. *Political Analysis*, 14 (1), 63-82, doi: <https://www.jstor.org/stable/25791835>.
- [3] Canetto, S. S. y Sakinofsky, I. (1998). The gender paradox in suicide. *Suicide and Life-Threatening Behavior*, 28(1), 1-23, doi: <https://doi.org/10.1111/j.1943-278X.1998.tb00622.x>.
- [4] Del Hierro, A.R.L., García, C.B. y Salmerón, R. (2020). Analysis of the condition number in the raise regression. *Communications in Statistics-Theory and Methods*, 1-16, doi: <https://doi.org/10.1080/03610926.2020.1740737>.
- [5] De Irala Estévez, J., Serrano del Castillo, A., Pelayo Orozco, A., Navas Martínez, J. L., y Fernández-Crehuet Navajas, R. (1997). Mortalidad por suicidio en España: factores socioeconómicos y ambientales asociados. *Atención Primaria*, 19(5), 250-256, url: <https://www.elsevier.es/es-revista-atencion-primaria-27-articulo-mortalidad-por-suicidio-espana-factores-14499>.
- [6] García, J., García, C.B. y Liria, J. (2011). The Raise Method: An Alternative Procedure to Estimate the Parameters in Presence of Collinearity, *Quality and Quantity*, 45, 403-423, doi: <https://doi.org/10.1007/s11135-009-9305-0>.
- [7] García, C., Salmerón, R., García, C. y García Pérez, J. (2016). Raise regression mitigating collinearity in moderated regression. In *Proceedings of the 5th Advanced Research in Scientific Areas Conference (ARSA 2016)*, 23-29, url: <https://www.researchgate.net/publication/311810153>.
- [8] García, C., García, J. y Salmerón, R. (2018). Variance inflation factor and condition number in multiple linear regression. *Journal of Statistical Computation and Simulation*, 88(12), 2365-2384, doi: <https://www.tandfonline.com/doi/abs/10.1080/00949655.2018.1463376>.

- [9] García, C.B., Salmerón, R., García, C. y García, J. (2020). Residualization: justification, properties and application. *Journal of Applied Statistics*, 47(11), 1990-2010, doi: <https://doi.org/10.1080/02664763.2019.1701638>.
- [10] Hainmueller, J., Mummolo, J. y Xu, Y. (2019). How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. *Political Analysis*, 27(2), 163–192, url: https://jmmumolo.scholar.princeton.edu/sites/g/files/toruqf3341/files/hainmueller_mummolo_xu_2019.pdf.
- [11] Hoerl, A.E. y Kennard, R.W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12 (1), 55–67, doi: <https://doi.org/10.2307/1271436>.
- [12] Hoerl, A.E. y Kennard, R.W. (1970b). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12 (1), 69–82, doi: <https://doi.org/10.2307/1267352>.
- [13] Iacobucci, D., Schneider, M.J., Popovich, D.L. y Bakamitsos, G.A. (2015). Mean centering helps alleviate “micro” but not “macro” multicollinearity. *Behavior Research Methods*, 48, 1308–1317, doi: <https://doi.org/10.3758/s13428-015-0624-x>.
- [14] Iglesias-García, C., Sáiz, P. A., Burón, P., Sánchez-Lasheras, F., Jiménez-Treviño, L., Fernández-Artamendi, S., Al-Halabí, S., Corcoran, P., García-Portilla, M. P. y Bobes, J. (2017). Suicidio, desempleo y recesión económica en España. *Revista de Psiquiatría y Salud Mental*, 10(2), 70–77, doi: <https://doi.org/10.1016/j.rpsm.2016.04.005>.
- [15] Marquardt, D.W. (1970). Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. *Technometrics*, 12(3), 591–612, doi: <https://doi.org/10.1080/00401706.1970.10488699>.
- [16] Marquardt, D.W. y Snee, R.D. (1975). Ridge Regression in Practice. *The American Statistician*, 29(1), 3—20, doi: <https://www.jstor.org/stable/2683673>.
- [17] Marquardt, D.W. (1980). A Critique of Some Ridge Regression Methods: Comment. *Journal of the American Statistical Association*, 75(369), 87—91, doi: <https://doi.org/10.2307/2287388>.
- [18] Montgomery, D.C., Peck, E.A. y Vining, G.G. (2012). *Introduction to linear regression analysis*. John Wiley and Sons, Hoboken.
- [19] Pérez Díez, I., La Calle Rodríguez-Peña, M., Corral Martínez, J. L. y de la Torre Luque, A. (2025). ¿Cómo influyen los determinantes sociodemográficos, económicos y ambientales en las tasas de suicidio en España? Un análisis geoespacial. *Plataforma Nacional para el Estudio y la Prevención del Suicidio*, url: <https://www.plataformanacionalsuicidio.es/noticias/cmo-influyen-los-determinantes-sociodemogrifi>

cos-econmicos-y-ambientales-en-las-tasas-de-suicidio-en-espaa-un-analisis-geoespacial.

- [20] R Core Team (2025). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (Austria), <https://www.R-project.org/>.
- [21] Salmerón, R., García, C.B., García, J. y López, M.M. (2017). The raise estimator estimation, inference, and properties. *Communications in Statistics-Theory and Methods*, 46 (13), 6446–6462, doi: <https://hdl.handle.net/10481/89676>.
- [22] Salmerón, R., García, C.B. y García, J. (2018). Variance Inflation Factor and Condition Number in multiple linear regression. *Journal of Statistical Computation and Simulation*, 88(12), 2365–2384, doi: <https://doi.org/10.1080/00949655.2018.1463376>.
- [23] Salmerón, R., García, C.B. y García, J. (2020). Comment on “A Note on Collinearity Diagnostics and Centering” by Velilla (2018). *American Statistician*, 74(1), 68–71, doi: <https://doi.org/10.1080/00031305.2019.1635527>.
- [24] Salmerón, R., Rodríguez, A. y García, C.B. (2020b). Diagnosis and quantification of the non-essential collinearity. *Computational Statistics*, 35(2), 647–666, doi: <https://doi.org/10.1007/s00180-019-00922-x>.
- [25] Salmerón, R., Rodríguez Sánchez, A., García, C.B. y García Pérez, J. (2020). The VIF and MSE in raise regression. *Mathematics* 8 (4), 605, doi: <https://www.mdpi.com/2227-7390/8/4/605>.
- [26] Salmerón, R., García, C.B. y García, J. (2021). A Guide to Using the R Package “multiColl” for Detecting Multicollinearity. *Computational Economics*, 57, 529–536, doi: <https://doi.org/10.1007/s10614-019-09967-y>.
- [27] Salmerón, R., García, C.B. y García, J. (2022). The multiColl Package Versus Other Existing Packages in R to Detect Multicollinearity. *Computational Economics*, 60(2), 439–450, doi: <https://doi.org/10.1007/s10614-021-10154-1>.
- [28] Salmerón, R., García, C.B. y García, J. (2025). The raise regression: justification, properties and application. *International Statistical Review*, 93 (1), 73–101, doi: <https://doi.org/10.1111/insr.12575>.
- [29] Snee, R.D. y Marquardt, D.W. (1984). Comment: Collinearity Diagnostics Depend on the Domain of Prediction, the Model, and the Data. *The American Statistician*, 38(2), 83–87, doi: <https://www.jstor.org/stable/2683239>.

- [30] Stewart, G.W. (1987). Collinearity and least squares regression. *Statistical Science*, 2, 68–84, doi: <https://www.jstor.org/stable/2245615>.
- [31] Stock, J.H. y Watson, M.W. (2020). *Introducción a la econometría* (4.^a ed.). Pearson Educación, url: https://gc.scalahed.com/recursos/files/r161r/w25589w/Introduccion_a_la_Econometria.pdf
- [32] Stuckler, D., Basu, S., Suhrcke, M., Coutts, A. y McKee, M. (2009). The public health effect of economic crises and alternative policy responses in Europe: an empirical analysis. *The Lancet*, 374(9686), 315–323, url: [https://doi.org/10.1016/S0140-6736\(09\)61124-7](https://doi.org/10.1016/S0140-6736(09)61124-7)
- [33] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58(1), 267–288, doi: <https://www.jstor.org/stable/2346178>.
- [34] Wilkinson, R. G. y Pickett, K. E. (2009). *The Spirit Level: Why More Equal Societies Almost Always Do Better*. Allen Lane, London, url: <https://www.equalitytrust.org.uk/resources/the-spirit-level>.
- [35] Zou, H. y Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2), 301–320, doi: <https://www.jstor.org/stable/3647580>.