Tarea 6: Modelos Estadísticos I.

Rojas Gutiérrez Rodolfo Emmanuel

28 de abril de 2021

Observación 1. *Para todos los modelos considerados aquí se supondrá que sus términos de error cumplen los supuestos usuales del modelo de regresión lineal, además para algunos resultados será de utilidad agregar el supuesto de normalidad de dichos términos de error.* △

Ejercicio 1 (Análisis de servicio hospitalario.)**:**

En la tabla 1 se muestran datos relacionados con las necesidades para el trabajo hospitalario provenientes de una muestra de 17 hospitales, de un sistema hospitalario determinado. La variable $V1$ representa la carga promedio diaria de pacientes, $V2$ denota el número mensual de rayos X, $V3$ denota el número de días cama de ocupación mensual, $V4$ denota la población elegible en el área (dividida por 1000), $V5$ denota la longitud promedio de permanencia de los pacientes, en días. La variable respuesta W es el número de horas de trabajo durante el mes. Hacer lo siguiente:

a) Grafique los diagramas de dispersión de la variable respuesta con cada una de las variables independientes y también evalué los coeficientes de correlación entre la variable respuesta y cada una de las cinco variables regresoras. Comente los resultados de este inciso.

b) ¿Que evidencia hay de la presencia de multicolinealidad en los datos de este problema? Haga los cálculos necesarios para la detección de la multicolinealidad.

c) Calcule los factores de inflación de la varianza VIF para las cinco variables regresoras y deles una interpretación.

d) Discuta la posibilidad de obtener modelos de regresión donde el impacto de la multicolinealidad sea menor que en el modelo completo original con las cinco variables regresoras.

e) Proponga un modelo de regresión en el que la multicolinealidad es menor que en el modelo original.

Solución. **a)** En la figura 1 se dejan las gráficas de dispersión solicitadas, en cada una de estás gráficas se puede ver un modelo de regresión lineal que ocupa en cada caso a una y solo una de las variables independientes $V1, \ldots, V2$ para explicar a la variable respuesta W, es importante destacar que tanto el modelo que tiene por variable explicativa a $V1$ como el que tiene por variable explicativa a $V3$, parecen ser bastante razonables para explicar a la variable respuesta W ya que presentan[1] una

[1] Al menos visualmente.

Cuadro 1: Datos de necesidades del trabajo en un sistema hospitalario determinado.

$V1$	$V2$	$V3$	$V4$	$V5$	W
15.57	2463	472.82	18	4.45	556.52
44.02	2048	1339.75	9.5	6.92	696.82
20.42	3940	620.25	12.8	4.28	1033.15
18.74	6505	568.33	36.7	3.9	1603.62
49.2	5723	1497.6	35.7	5.5	1611.37
44.92	11520	1365.83	24	4.6	1613.27
55.48	5779	1687	43.3	5.62	1854.17
59.28	5969	1639.92	46.7	5.15	2160.55
94.39	8461	2872.33	78.7	6.18	2305.58
128.02	20106	3655.08	180.5	6.15	3503.93
96	13313	2912	60.9	5.88	3571.89
131.42	10771	3921	103.7	4.88	3741.4
127.21	15543	3865.67	126.8	5.5	4026.52
252.9	36194	7684.1	157.7	7	10343.81
409.2	34703	12446.33	169.4	10.78	11732.17
463.7	39204	14098.4	331.4	7.05	15414.94
510.22	86533	15524	371.6	6.35	18854.45

variabilidad aceptable en sus bandas de confianza y predicción, aquí uno podría pensar erróneamente en usar ambas variables como predictores para 'mejorar' estos modelos, o incluso usar todas las variables independientes, pero habrá que analizar la existencia de multicolinealidad entre ellas para determinar cuales podrían realmente aportar más a un modelo que busque explicar a W, por otro lado las correlaciones solicitadas se encuentran a continuación, las mismas no agregan mucho más al análisis previamente mencionado.

$Cor(V1, W)$	$Cor(V2, W)$	$Cor(V3, W)$	$Cor(V4, W)$	$Cor(V5, W)$
0.986	0.945	0.986	0.940	0.579

Cuadro 2: Correlaciones de las variables independientes con la variable respuesta.

Como primer paso para detectar problemas de multicolinealidad entre las variables independientes, se pensó que resultaría de gran utilidad realizar gráficas de dispersión para las mismas, por ende, se puede observar en las figuras 2-5 dichas gráficas, además, se agregó nuevamente una recta de regresión lineal e intervalos de confianza y predicción al 95 % en cada caso a modo de indicador de la relación lineal existente entre las variables independientes. Se observa que en la mayoría de los casos pareciera existir colinealidad cuando menos moderada, sin embargo, la variable $V5$ aparenta ser la que tiene la colinealidad más débil entre las diversas variables independientes, esto porque en la mayoría de los gráficos de dispersión los intervalos de confianza y predicción parecen abrirse mucho más que en los otros casos, indicando que hay gran variabilidad que no es explicada por ninguna de las otras variables independientes por separado.

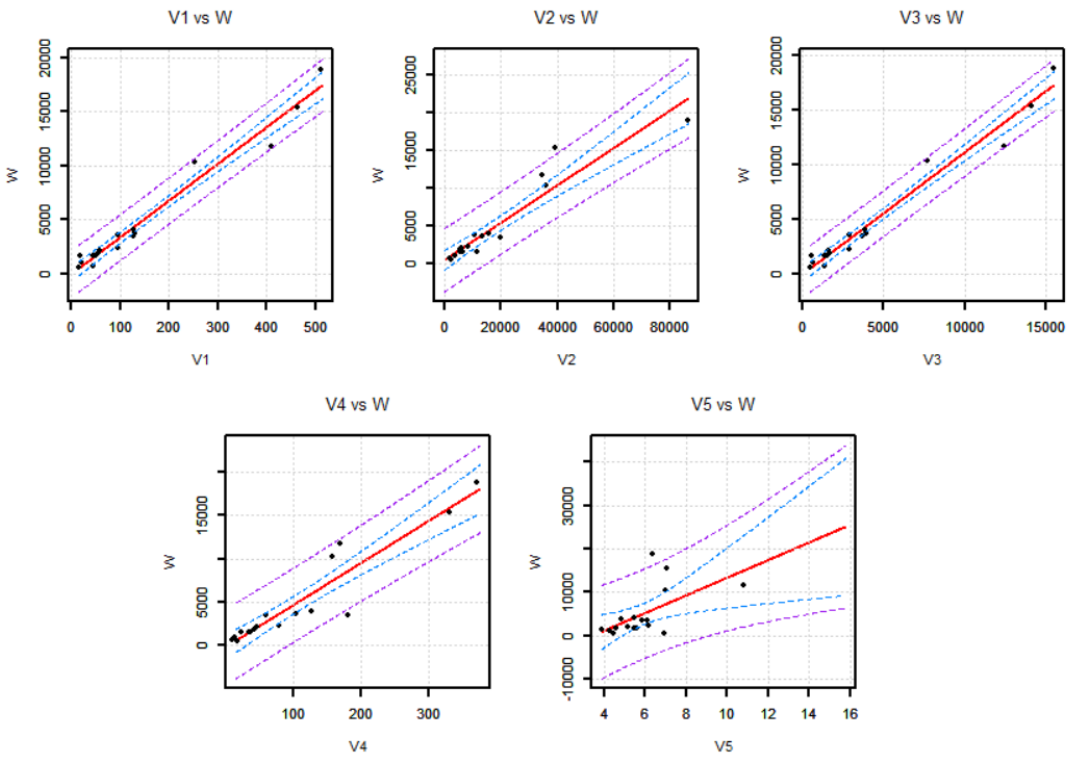

Figura 1: Gráficos de dispersión de cada una de las variables independientes $\{V1, \ldots, V5\}$ contra la variable respuesta W. Se deja recta de regresión lineal (recta roja) e intervalos de confianza al 95 % (rectas azules) y predicción al 95 % confianza (rectas purpuras).

b) y c) Para $i \in \{1, \ldots, 5\}$ se define el vector columna de datos X_i como aquel que resulta de centrar el vector columna Vi y posteriormente escalar el vector resultante dividiendo por su longitud.[2] Sea ahora $\mathbf{X} = \begin{pmatrix} X_1 & X_2 & \ldots & X_5 \end{pmatrix}$ se ajustará inicialmente el modelo[3]

$$E[W_i | X_{i1}, \ldots, X_{i5}] = \alpha + \delta_1 X_{1i} + \ldots + \delta_5 X_{i5}, \ i \in \{1, \ldots, 17\}. \tag{1}$$

esto con la finalidad de estudiar la existencia de multicolinealidad entre todas las variables independientes.[4] Bajo el modelo (1) se tiene que la estimación[5] para α es $\hat{\alpha} = \overline{W} = 4977.892$, por lo

[2] Entiéndase por longitud su norma euclideana.

[3] El centrar y escalar tiene dos propósitos en este trabajo, el evitar que los problemas generados por las diversas escalas afecten los diagnósticos de multicolinealidad, y el poder hacer comparables los modelos calculados en los ejercicios 1 y 2, con aquellos que se calcularan en los ejercicios 3 y 4, en los cuales será preferible el escalar $\mathbf{X}$ de modo que $\mathbf{X}'\mathbf{X}$ resulte una matriz de correlación.

[4] Note que las X_i son transformaciones afines de las Vi, por lo que de existir multicolinealidad entre las Vi esta se vera reflejada en las X_i y viceversa.

[5] Por mínimos cuadrados ordinarios.

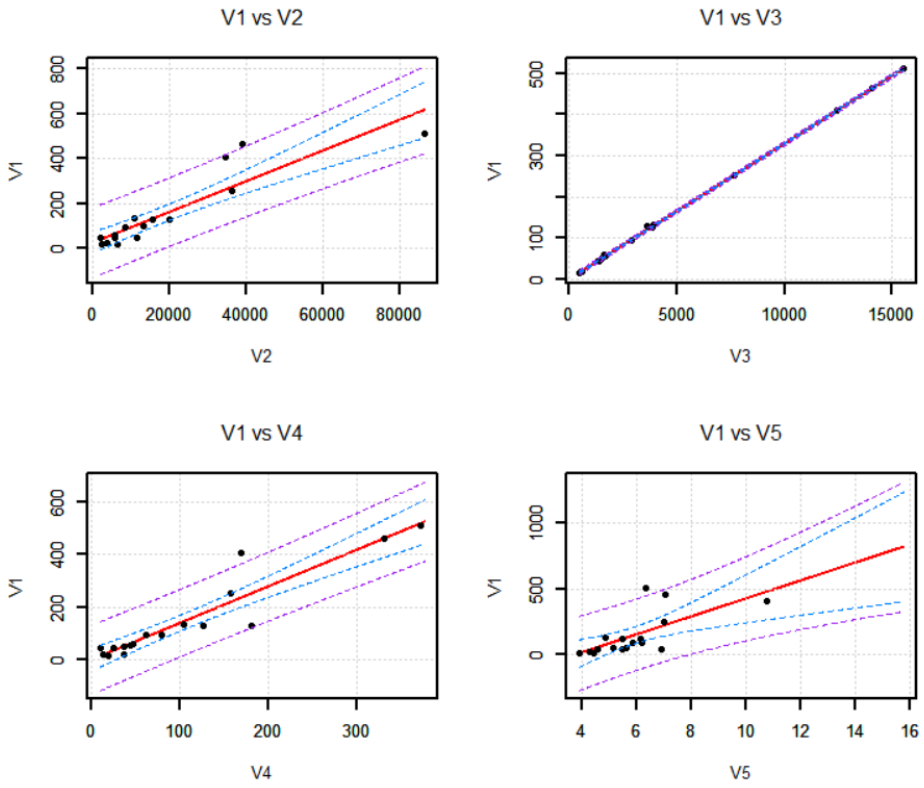

Figura 2: Gráficos de dispersión de las variables independientes $\{V2, V3, V4, V5\}$ contra la variable independiente $V1$. Se deja recta de regresión lineal (recta roja) e intervalos de confianza al 95 % (rectas azules) y predicción al 95 % confianza (rectas purpuras).

cual, es posible olvidarnos del intercepto y ajustar de manera equivalente el modelo

$$E[W_i - \overline{W}|X_{i1}, \ldots, X_{i5}] = \delta_1 X_{1i} + \ldots + \delta_5 X_{i5}, \ i \in \{1, \ldots, 17\}\,. \tag{2}$$

dicho ajuste se realizó en R mediante mínimos cuadrados para las estimaciones de los coeficientes, arrojando los resultados presentados en la tabla 3, algunos otros detalles sobre el ajuste de este modelo también se presentan en (3). Observe que en este primer paso se tiene un segundo indicador[6] de la posible existencia de multicolinealidad entre las variables independientes utilizadas para el modelo, ya que en (3) se observa que el coeficiente de determinación del modelo ajustado (3) es bastante alto, de hecho es casi uno, sin embargo solo el coeficiente δ_2 resulta significativamente distinto de cero bajo un nivel de significancia del 5 %, estos primeros indicios pese a indicarnos la existencia de multicolinealidad en el modelo no nos dicen nada sobre la gravedad de la misma, por lo que se analizarán otros indicadores importantes a continuación. Primeramente, se obtuvo la matriz de correlación de las columnas de $\mathbf{X}$ (4), en la misma es posible observar que X_1 y X_3 están casi

[6]Recuerde que el primero fueron las gráficas de dispersión del inciso anterior.

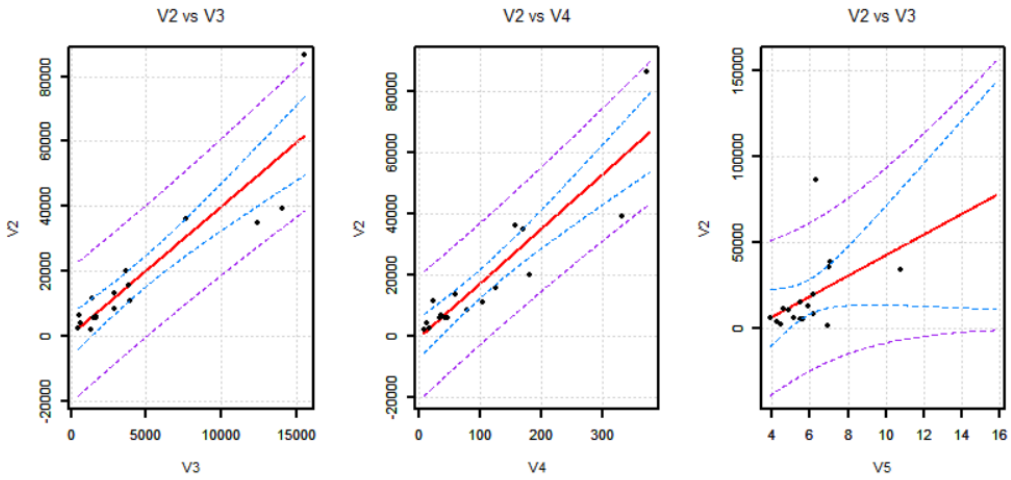

Figura 3: Gráficos de dispersión de las variables independientes $\{V3, V4, V5\}$ contra la variable independiente $V2$. Se deja recta de regresión lineal (recta roja) e intervalos de confianza al 95 % (rectas azules) y predicción al 95 % confianza (rectas purpuras).

perfectamente correlacionadas, lo cual ya es un indicador claro de que la colinealidad entre estas dos variables es fuerte, por otro lado, observe que hay varios coeficientes de correlación que superan al 0.9 los cuales corresponden en su totalidad a las correlaciones entre las variables independientes en el conjunto $\{X_1, \ldots, X_4\}$, de este modo, los coeficientes de correlación más pequeños son aquellos que involucran a X_5 lo cual era de esperarse después de lo comentado en lo gráficos de dispersión en el inciso anterior.

Coeficiente	Estimación	t-valor	p-valor
δ_1	-10175	-0.169	0.8687
δ_2	4762	2.748	0.0177*
δ_3	31159	0.536	0.6018
δ_4	-1818	-0.613	0.5516
δ_5	-2495	-1.961	0.0735

Cuadro 3: Resultados análisis de regresión: $W - \overline{W}$ variable respuesta $X_1, \ldots, X_5$ variables independientes.

$$R^2 = 0.9908, \quad AIC = 272.660. \tag{3}$$

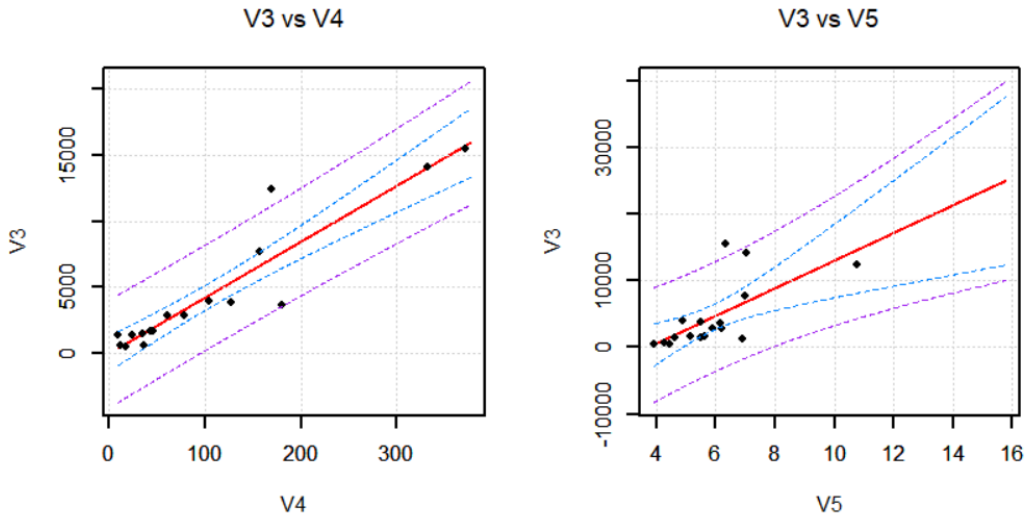

Figura 4: Gráficos de dispersión de las variables independientes $\{V4, V5\}$ contra la variable independiente $V3$. Se deja recta de regresión lineal (recta roja) e intervalos de confianza al 95 % (rectas azules) y predicción al 95 % confianza (rectas purpuras).

$$R_X = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{array} \begin{array}{c} \begin{array}{ccccc} X_1 & X_2 & X_3 & X_4 & X_5 \end{array} \\ \begin{pmatrix} 1.000 & 0.907 & 0.999 & 0.936 & 0.671 \\ 0.907 & 1.000 & 0.907 & 0.910 & 0.447 \\ 0.999 & 0.907 & 1.000 & 0.933 & 0.671 \\ 0.936 & 0.910 & 0.933 & 1.000 & 0.463 \\ 0.671 & 0.447 & 0.671 & 0.463 & 1.000 \end{pmatrix} \end{array} \tag{4}$$

Por otra parte, se presenta en la figura 6 un mapa de calor de la matriz de correlaciones (4) con un dendograma generado de manera automática por el comando *heatmap* de R, utilizando un algoritmo de machine learning conocido como hierarchical clustering, los clusters formados se basan en que tan cercanas son las distintas variables independientes utilizando como distancia a la correlación existente entre las mismas, mediante este agrupamiento es posible constatar que las variables X_1 y X_3 están fuertemente correlacionadas, y qué posiblemente haya una colinealidad de moderada a fuerte entre estas variables independientes y las variables independientes X_2 y X_4. Por otro lado, en la tabla 4 se pueden observar los factores de inflación de la varianza correspondientes a cada una de los coeficientes de nuestro modelo, note que en este caso dichos factores de inflación resultan particularmente preocupantes debido a que al menos 3 de ellos, los correspondientes a los coeficientes que multiplican a las variables independientes X_1, X_3 y X_4, superan con claridad al numero 10 y el asociado al coeficiente que multiplica a la variable independiente X_2 es mayor a 5, lo cual indica que estos coeficientes están siendo pobremente estimados debido a la multicolinealidad

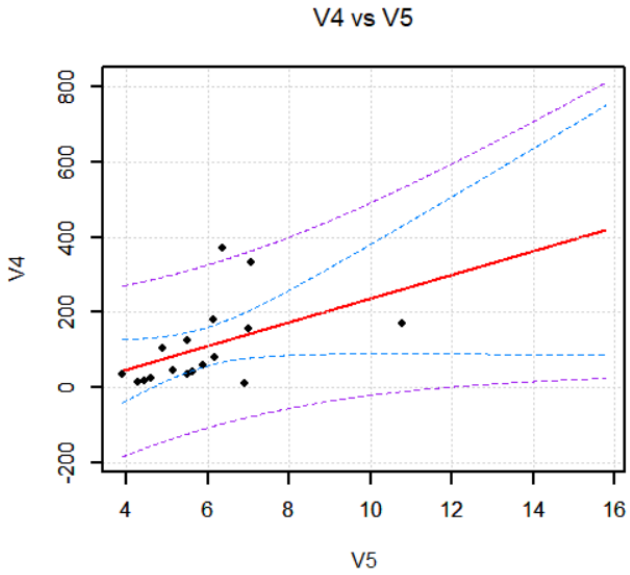

Figura 5: Gráfico de dispersión de la variable independiente $V5$ contra la variable independiente $V4$. Se deja recta de regresión lineal (recta roja) e intervalos de confianza al 95 % (rectas azules) y predicción al 95 % confianza (rectas purpuras).

VIF_1	VIF_2	VIF_3	VIF_4	VIF_5
9598.007	7.941	8933.566	23.293	4.280

Cuadro 4: Factores de inflación de la varianza, modelo 3.

De igual modo, se realizó un análisis utilizando los valores propios de la matriz $\mathbf{X}'\mathbf{X}$ con los cuales se obtuvo el número de condición de la matriz $\mathbf{X}'\mathbf{X}$ de la siguiente manera, sea $\lambda_{\text{máx}}$ el máximo valor propio de la matriz $\mathbf{X}'\mathbf{X}$ y $\lambda_{\text{mín}}$ el mínimo de ellos entonces el número de condición de la matriz $\mathbf{X}'\mathbf{X}$ esta dado por:

$$K(\mathbf{X}'\mathbf{X}) = \frac{\lambda_{\text{máx}}}{\lambda_{\text{mín}}} = 77773.50. \tag{5}$$

Dado que este número es mayor a 1000 se considera que existe multicolinealidad severa en las columnas de la matriz de diseño. Por último, se llevó a cabo de acuerdo a Belsey y Kuh un procedimiento conocido como descomposición de varianza, en el cual se descompone la varianza de los estimadores[7] de los coeficientes del modelo de acuerdo a los valores singulares de la matriz $\mathbf{X}$, que son simplemente las raíces cuadradadas positivas de lo valores propios de la matriz $\mathbf{X}'\mathbf{X}$, este procedimiento está fuertemente relacionado con los índices de condición (ó condicionamiento) de la matriz $\mathbf{X}'\mathbf{X}$, los cuales se obtienen de la siguiente manera, sean $\{\lambda_1, \ldots, \lambda_5\}$ los valores propios

[7] Entiéndase la parte constante que multiplica al valor desconocido de σ^2.

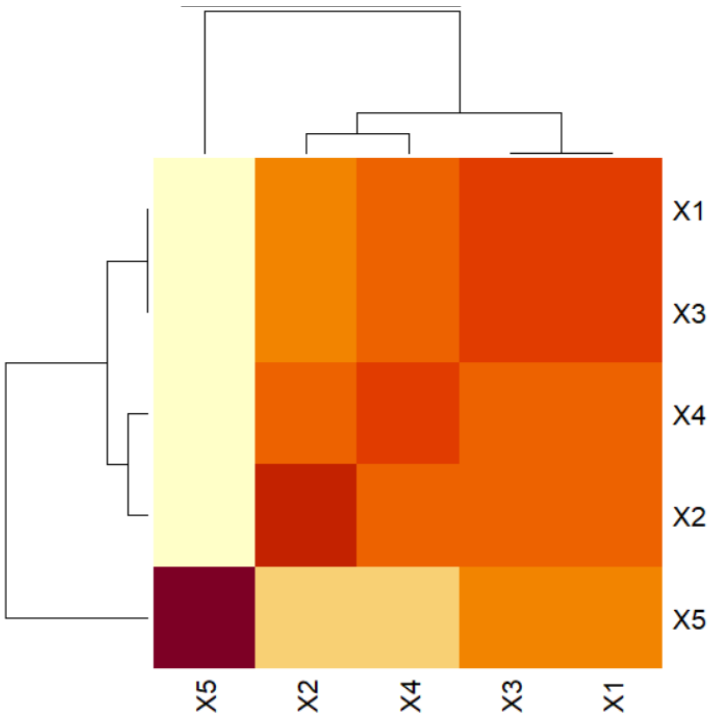

Figura 6: Mapa de calor de matriz de correlaciones de las columnas de $\mathbf{X}$.

ordenados de manera ascendente de la matriz $\mathbf{X}'\mathbf{X}$, entonces $\left\{\lambda_1^{1/2}, \ldots, \lambda_5^{1/2}\right\}$ es el conjunto de valores singulares de la matriz $\mathbf{X}$, y los índices de condición de la matriz $\mathbf{X}'\mathbf{X}$ se calculan como $i_j = \lambda_5/\lambda_j,\ j = 1, \ldots, 5$, observe que a cada valor singular le corresponde un índice de condición, así que para hacer referencia de la proporción de varianza asociada a un valor singular de la matriz $\mathbf{X}$ se hará referencia al porcentaje de varianza asociado a su índice de condición. Los índices de condición junto con la descomposición de varianzas mencionada y los valores singulares de $\mathbf{X}$ correspondientes a cada índice se encuentran expuestos en la tabla 5

Valor Singular	$V(\hat{\delta}_1)$	$V(\hat{\delta}_2)$	$V(\hat{\delta}_3)$	$V(\hat{\delta}_4)$	$V(\hat{\delta}_5)$	Índice de Cond.
2.05	$5.85 \cdot 10^{-6}$	$6.16 \cdot 10^{-3}$	$6.27 \cdot 10^{-6}$	$2.17 \cdot 10^{-3}$	$6.20 \cdot 10^{-3}$	1.00
0.817	$6.42 \cdot 10^{-10}$	0.0213	$1.22 \cdot 10^{-10}$	$6.21 \cdot 10^{-3}$	0.277	2.51
0.308	$3.04 \cdot 10^{-5}$	0.861	$2.80 \cdot 10^{-5}$	0.131	0.0328	6.66
0.202	$5.61 \cdot 10^{-4}$	0.109	$7.13 \cdot 10^{-4}$	0.424	0.485	10.15
$7.35 \cdot 10^{-3}$	$9.99 \cdot 10^{-1}$	$3.13 \cdot 10^{-3}$	$9.99 \cdot 10^{-1}$	0.437	0.199	278.88

Cuadro 5: Matriz de descomposición de varianzas.

La forma de obtener las proporciones de varianza asociadas a cada valor singular es un proce-

dimiento complejo que se puede consultar en el libro de Belsey y Kuh[8]. Una vez teniendo la tabla anterior, el procedimiento se basa en ver cuales índices de condición exceden[9] el 30, y fijarse en cuales coeficientes concentran los mayores porcentajes de varianza en estos índices, lo que es un indicativo de que las variables asociadas a dichos coeficientes presentan problemas de multicolinealidad. En este caso observe que únicamente el último índice de condicionamiento excede el 30, y se tiene que tanto el coeficiente de X_1 como el coeficiente de X_3 concentran casi el 100 % de su variabilidad en dicho índice, lo que no es de sorprender y coincide con lo ya comentado hasta este punto, se destaca además que el coeficiente de X_4 pareciera tener una buena proporción de variabilidad asignada a este índice de condición[10]. Las conclusiones de todo este análisis se presentan en el siguiente inciso.

d) Después de todo el análisis realizado en el inciso anterior, es claro que solo es posible quedarnos con la variable independiente X_1 o con la variable independiente X_3 debido a alta colinealidad que tienen, por lo que, parece ser una buena idea considerar un modelo reducido primeramente eliminando a una de estas variables, analizando detenidamente el mapa de calor de la matriz de correlación de las columnas de $\mathbf{X}$ y los factores de inflación de la varianza presentados en la tabla 5, se considera que la mejor opción es deshacernos de la variable X_1 por tener el mayor VIF y por presentar además problemas de multicolinealidad con las demás variables independientes, por otro lado, dado que la variable independiente X_4 apareció en el análisis de descomposición de la varianza como otra variable independiente que podría tener problemas de multicolinealidad con X_3, aunado al hecho de que la correlación entre X_3 y X_4 es igual a 0.933 y observando que la correlación entre X_3 y X_2 es de 0.907 y la de X_2 con X_4 es de 0.910; se piensa que estas variables pueden representar problemas de multicolinealidad[11] de incluirse mas de una ellas en el modelo, por lo que, se decidió construir un modelo que únicamente considerará a alguna de ellas y a X_5 que es la única variable independiente que no pareciera tener problemas de multicolinealidad fuerte con ninguna de las demás variables independientes, en la tabla 6 se puede ver el coeficiente de determinación R^2 y el AIC de cada uno de estos modelos de regresión lineal ajustados por mínimos cuadrados

Variables consideradas	X_2 y X_5	X_3 y X_5	X_4 y X_5
R^2	0.9239	0.9847	0.9104
AIC	302.619	275.302	305.394

Cuadro 6: Coeficientes de determinación y AIC de los ajustes por mínimos cuadrados para los modelos propuestos.

Dado que el modelo que considera únicamente a las variables independientes X_3 y X_5 es claramente superior a los otros dos modelos en todos los aspectos considerados en la tabla 6 se decidió presentar este en el siguiente inciso.

e) El modelo reducido a considerar es

$$E[W_i - \overline{W}|X_{i3}, X_{i5}] = \delta_1' X_{3i} + \delta_2' X_{5i},\ i \in \{1, \ldots, 17\}\,, \tag{6}$$

[8]Referencia 3, pp. 100 - 107.

[9]Este criterio es coherente en el sentido de que el máximo índice de condición es igual a la raíz cuadrada del número de condición, y si el número de condición excede a 1000 entonces el máximo índice de condición excede a 30.

[10]Al hacer referencia a un buen porcentaje de su variabilidad, se entenderá la regla de dedo siguiente 0.4 o más.

[11]Estas relaciones inclusive se ven marcadas en el mapa de calor de la matriz de correlaciones de las columnas de $\mathbf{X}$.

para el cual se obtuvieron con el uso de R las estimaciones por mínimos cuadrados y los diagnósticos presentados en la tabla 7 y en (7)

Coeficiente	Estimación	t-valor	p-valor
δ_1'	24189.2	25.282	$1.03 \cdot 10^{-13}$
δ_2'	-3362.5	-3.514	$3.13 \cdot 10^{-3}$

Cuadro 7: Resultados análisis de regresión: $W - \overline{W}$ variable respuesta, X_3, X_5 variables independientes.

$$R^2 = 0.9847, \quad AIC = 275.302. \tag{7}$$

Observe que para este modelo, todos los coeficientes resultan significativamente distintos de cero bajo un nivel de significancia del 5 %, más aún, se tiene un coeficiente de determinación R^2 aceptable y a pesar de que hubo un incremento en el AIC comparado con el modelo ajustado en el inciso b) (3), este incremento es muy pequeño considerando la gran cantidad de variables que se quito al modelo. Además, se analizaron nuevamente algunos de los distintos índices de multicolinealidad presentados en el inciso c) entre los cuales se destacan los siguientes. Sea $\mathbf{X}_1 = \begin{pmatrix} X_3 & X_5 \end{pmatrix}$ la matriz de diseño para el nuevo modelo se tiene que la matriz de correlación de las columnas de $\mathbf{X}_1$ esta dada por

$$R_{\mathbf{X}_1} = \begin{array}{c} \\ X_3 \\ X_5 \end{array} \begin{array}{c} \begin{array}{cc} X_3 & X_5 \end{array} \\ \begin{pmatrix} 1.000 & 0.671 \\ 0.671 & 0.447 \end{pmatrix} \end{array} \tag{8}$$

note que la correlación entre X_3 y X_5 pareciera no ser tan elevada, por lo menos no como algunas de las correlaciones presentadas en el inciso **c)** para el modelo completo. Por otro lado, el indice de condición de la matriz $\mathbf{X}_1'\mathbf{X}_1$ esta dado por

$$K(\mathbf{X}_1'\mathbf{X}_1) = 5.081. \tag{9}$$

El mismo es menor a 100, por lo que colinealidad que pudiera existir entre X_3 y X_5 no considerarse como un problema serio para el modelo. Por otra parte, dado que este número de condición coincide con el cuadrado del mayor índice de condicionamiento, se concluye que no será necesario realizar un procedimiento de descomposición de la varianza, debido a que la raíz cuadrada de (9) es igual a 2.254 y esta cantidad es claramente menor a 30, lo que implica que ningún índice de condicionamiento podrá ser mayor a 30. Por último, los factores de inflación de la varianza se presentan en el cuadro 8 en los mismos se observa que ninguno excede el 5, por lo cual no pareciera que la colinealidad entre X_3 y X_5, sea un problema para las estimaciones producidas por este modelo.

VIF_3	VIF_5
1.819	1.819

Cuadro 8: Factores de inflación de la varianza, modelo 3.

Adicional Si gusta regresar de este modelo a uno que considere las escalas originales para las variables independientes, la variable respuesta sin centrar y que se base en el que ya fue construido,

basta tomar los coeficientes estimados 2583.212 para el intercepto 1.232 para el coeficiente que multiplica a la variable independiente $V3$ y -530.676 para la el coeficiente restante. ■

Ejercicio 2 (Datos de Evolución Calorífica del Cemento)**:**

Los datos de la tabla 9 están relacionados con la evolución calorífica en calorías por gramo de cemento (y), como una función de las cantidades de cada uno de los siguientes cuatro ingredientes que participan en la mezcla. Aluminato tricálcico (x_1), silicato tricálcilo (x_2), ferrita de alumino tetracálcico (x_3), silicato dicálcico (x_4). Hacer lo siguiente

a) Obtenga la matriz de las correlaciones entre los regresores x_1, x_2, x_3 y x_4.

b) Calcule los factores de inflación de la varianza.

c) Calcule los eigenvalores de la matriz $X'X$ también calcule el número de condición K. Interprete los valores propios de la matriz $X'X$.

d) De acuerdo a los resultados obtenidos en los incisos anteriores ¿sospecha usted que la multicolinealidad está presente en los datos? ¿Qué puede decir de la fuente de la multicolinealidad?

e) Discuta la posibilidad de obtener modelos de regresión donde el impacto de la multicolinealidad sea menor que en el modelo completo original con las cuatro variables regresoras.

f) Proponga un modelo de regresión en el que la multicolinealidad es menor que en el modelo original.

Cuadro 9: Datos de Evolución Calorífica del Cemento.

obs	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

Solución. **a)** El objetivo a lo largo de este ejercicio será determinar un modelo de regresión lineal para la respuesta y, utilizando las variables independientes x_1, x_2, x_3 y x_4 de alguna manera, pero

primero se deberá determinar si existen problemas severos de multicolinealidad entre las variables independientes los cuales puedan afectar las estimaciones del modelo. Con esto en mente, en la figura 7 se pueden ver todos los elementos que se encuentran por encima de la diagonal de la matriz de correlación de x_1, x_2, x_3 y x_4, debido que los elementos de la diagonal de la misma son iguales a uno y dada la simetría de la matriz de correlación, esto es suficiente para determinar la matriz de correlación completa, observe que las correlaciones más altas se encuentran señaladas con un asterisco, dichas correlaciones son la correspondiente a x_1 y x_3 con un valor de -0.824 y la correspondiente a x_2 y x_4 con un valor de -0.973, note que esto se ve reflejado en las correspondientes gráficas de dispersión, sobre todo en los modelos de regresión lineal ajustados para estos datos, ya que los modelos asociados a las variables independientes mencionadas son los que parecen tener los intervalos de confianza con la menor variabilidad entre todas los modelos de regresión lineal presentados.

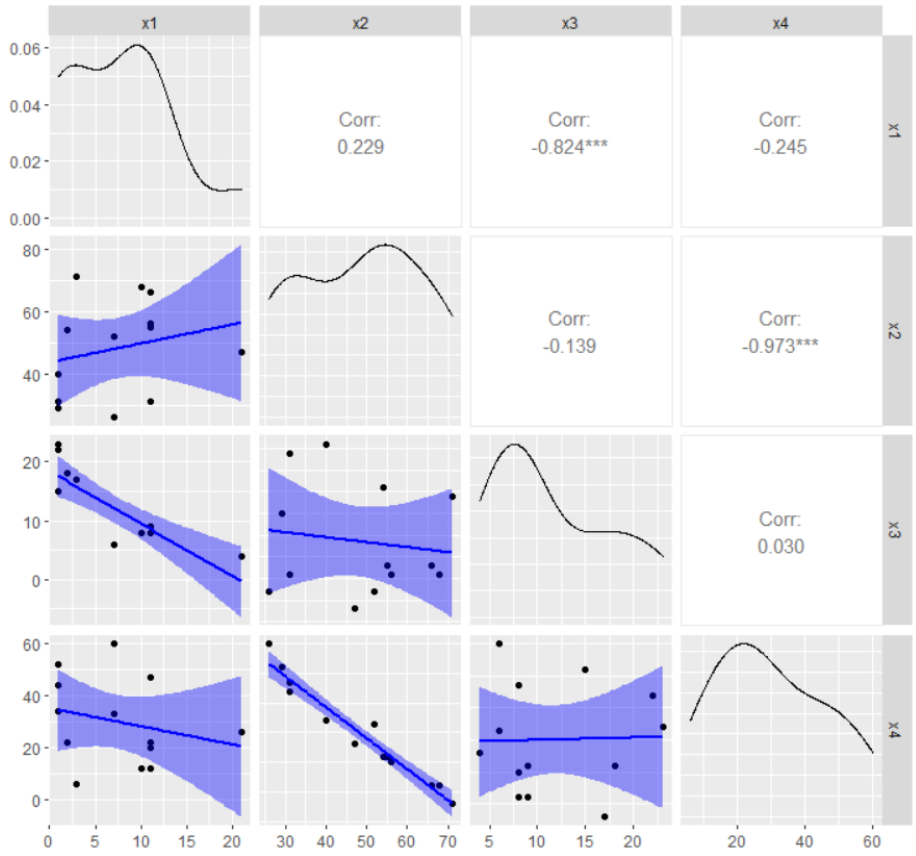

Figura 7: PairPlot con correlaciones de las variables independientes x_1, x_2, x_3 y x_4.

b) Para $i \in \{1, \ldots, 4\}$ se define el vector columna de datos X_i como aquel que resulta de centrar el vector columna x_i y posteriormente escalar el vector resultante dividiendo por su longitud. Sea

ahora $\mathbf{X} = \begin{pmatrix} X_1 & X_2 & X_3 & X_4 \end{pmatrix}$ se ajustará inicialmente el modelo[12]

$$E[y_i|X_{i1}, \dots X_{i4}] = \alpha + \alpha_1 X_{1i} + \dots \alpha_4 X_{i4},\ i \in \{1, \dots, 13\}. \tag{10}$$

esto con la finalidad de estudiar la existencia de multicolinealidad entre todas las variables independientes. Bajo el modelo (10) se tiene que la estimación[13] para α es $\hat{\alpha} = \overline{y} = 95.423$, por lo cual, es posible olvidarnos del intercepto y ajustar de manera equivalente el modelo

$$E[y_i - \overline{y}|X_{i1}, \dots X_{i4}] = \alpha_1 X_{1i} + \dots + \alpha_4 X_{i4},\ i \in \{1, \dots, 13\}. \tag{11}$$

dicho ajuste se realizó en R utilizando estimación por mínimos cuadrados para los coeficientes del mismo arrojando los resultados presentados en la tabla 10 y en (12)

Coeficiente	Estimación	t-valor	p-valor
α_1	31.607	2.209	0.0545
α_2	27.500	0.748	0.4738
α_3	2.261	0.143	0.8893
α_4	-8.353	-0.215	0.8342

Cuadro 10: Resultados análisis de regresión: $y - \overline{y}$ variable respuesta $X_1, \dots, X_4$ variables independientes.

$$R^2 = 0.9824, \quad AIC = 63.837. \tag{12}$$

Aquí podemos encontrar un segundo indicio de la existencia de multicolinealidad entre las variables independientes, esto porque ninguno de los coeficientes resulta ser significativamente distinto de cero bajo un nivel de significancia del 5 %, sin embargo, el coeficiente de determinación R^2 del modelo resulta bastante elevado. Por otro lado, se calcularan los factores de inflación de la varianza asociados a los coeficientes del modelo, calculando primeramente la matriz de correlación de las columnas de $\mathbf{X}$ la cual coincide con la matriz de correlación expuesta en el inciso **a)**, con el objetivo de no repetir información se presenta únicamente un mapa de calor de la misma, con un dendográma construido utilizando hierarchical clustering con el que se pueden apreciar las relaciones anteriormente[14] expuestas entre las variables independientes X_1 y X_3 y las variables independientes X_2 y X_4. Posteriormente, se utilizó la diagonal de la inversa de esta matriz de correlación con lo cual se obtuvieron los factores de inflación de la varianza mencionados, los cuales se presentan en la tabla 12, estos factores son un signo de alerta debido a que ninguno es menor en magnitud al número 10, por lo que se concluye que todos los coeficientes del modelo están siendo pobremente estimados debido a la multicolinealidad existente entre las variables independientes.

VIF_1	VIF_2	VIF_3	VIF_4
38.496	254.423	46.868	282.513

Cuadro 11: Factores de inflación de la varianza, modelo 3.

[12] Ya se explicó el porque de centrar y escalar.
[13] Por mínimos cuadrados ordinarios.
[14] Anteriormente en su versión no escalada ni centrada.

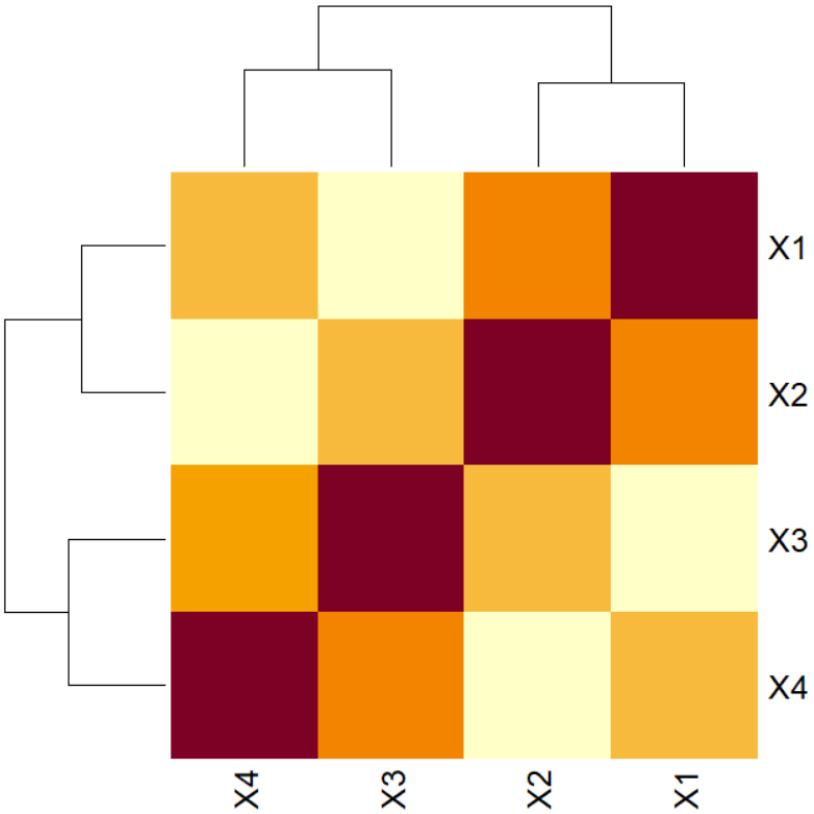

Figura 8: Mapa de calor de la matriz de correlaciones de las columnas de $\mathbf{X}$.

c) Los valores propios de la matriz $\mathbf{X}'\mathbf{X}$ ordenados en forma ascendente se encuentran dados en la tabla 12

λ_1	λ_2	λ_3	λ_4
$1.623 \cdot 10^{-3}$	0.187	1.576	2.236

Cuadro 12: Factores de inflación de la varianza, modelo 3.

Por si solos estos valores propios no son muy interpretables, pero serán de utilidad para realizar cálculos similares a los hechos anteriormente en el inciso **c)** del ejercicio 1, para determinar la gravedad de la multicolinealidad entre las variables independientes. En primera instancia observe que el número de condición de la matriz $\mathbf{X}'\mathbf{X}$ esta dado por

$$K(\mathbf{X}'\mathbf{X}) = \frac{\lambda_4}{\lambda_1} = 1376.881. \tag{13}$$

Debido a que el número de condición de la matriz $\mathbf{X}'\mathbf{X}$ es mayor a 1000 se puede decir que la

multicolinealidad existente entre las variables independientes del modelo es severa. Por lo cual, se decidió utilizar el procedimiento de descomposición de la varianza, el cual ya fue previamente explicado en el ejercicio anterior, los resultados de este procedimiento están expuestos en la tabla 13

Valores Singulares	$V(\hat{\alpha}_1)$	$V(\hat{\alpha}_2)$	$V(\hat{\alpha}_3)$	$V(\hat{\alpha}_4)$	Índice de Cond
1.50	$2.63 \cdot 10^{-3}$	$5.59 \cdot 10^{-4}$	$1.48 \cdot 10^{-3}$	$4.75 \cdot 10^{-4}$	1.000
1.26	$4.27 \cdot 10^{-3}$	$4.27 \cdot 10^{-4}$	$4.95 \cdot 10^{-3}$	$4.57 \cdot 10^{-4}$	1.191
0.432	0.0635	$2.08 \cdot 10^{-3}$	0.0465	$7.24 \cdot 10^{-4}$	3.461
0.0403	0.930	0.997	0.947	0.998	37.106

Cuadro 13: Matriz de descomposición de varianzas.

En la tabla 13 podemos observar que el único índice de condicionamiento que supera al 30 es el último, el cual coincide con la raíz cuadrada positiva del número de condición de la matriz $\mathbf{X}'\mathbf{X}$, y que todos los estimadores de los coeficientes tienen casi el 100 % de su varianza concentrada en este índice, por lo que todas las variables independientes parecen padecer de problemas por la multicolinealidad. Las conclusiones de los análisis hechos hasta el momento se presentan en el siguiente inciso.

d) Se observó que todas las variables independientes presentan en este modelo problemas de multicolinealidad los cuales afectan por ejemplo a las estimaciones de los coeficientes del modelo, como ya se mencionó cuando se realizaron los cálculos de los VIF, sin embargo, se notó en repetidas ocasiones que la multicolinealidad en el modelo parece ser una consecuencia de la relación existente entre las variables independientes X_1 y X_3, y las variables independientes X_2 y X_4, de lo que se concluye que lo mas sensato sería tomar una y solo una variable entre X_1 y X_3 para el modelo y lo mismo para las variables X_2 y X_4.

e) Para este inciso se ajustaron dos modelos, el que únicamente considera como regresoras a las variables independientes X_1 y X_2 y el que considera como regresoras únicamente a las variables independientes X_3 y X_4, se destaca que el primer modelo mencionado obtuvo un coeficiente de determinación R^2 igual a 0.979 y un AIC de 62.312, mientras que el segundo modelo obtuvo un coeficiente de determinación R^2 de 0.935 y un AIC de 76.745, de este modo se decidió que el modelo más adecuado para este problema era el que considera como regresoras únicamente a las variables independientes X_1 y X_2, todos los detalles sobre este modelo se expondrán en el siguiente inciso.

Observación 2. *Es importante destacar que se consideraron otros modelos, con otras combinaciones de las variables regresoras de acuerdo a las conclusiones obtenidas en el inciso* ***d)****, sin embargo, estos dos fueron los que obtuvieron los mejores resultados.* △

f) El modelo reducido a considerar es

$$E[y_i - \overline{y}|X_{i1}, X_{i2}] = \alpha_1' X_{1i} + \alpha_2' X_{2i},\ i \in \{1, \ldots, 13\}\,, \tag{14}$$

para el cual se obtuvieron con el uso de R las estimaciones y diagnósticos presentados en la tabla 14 y en (15)

Coeficiente	Estimación	t-valor	p-valor
α_1'	29.920	12.70	$6.51 \cdot 10^{-8}$
α_1'	35.698	15.15	$1.03 \cdot 10^{-8}$

Cuadro 14: Resultados análisis de regresión: $y - \overline{y}$ variable respuesta, X_1, X_2 variables independientes.

$$R^2 = 0.979, \quad AIC = 62.312. \tag{15}$$

Se destaca que todos los coeficientes resultan ser significativos bajo un nivel de significancia del 5 %, que el coeficiente R^2 es aceptable y que el AIC de este modelo mejoró un poco en comparación con el del modelo que considera a todas las variables independientes (12). Más aún, referente a a los temas de multicolinealidad, sea $\mathbf{X}_1 = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$ la matriz de diseño del modelo (14) se tiene que la matriz de correlación de sus columnas esta dada por

$$R_{\mathbf{X}_1} = \begin{array}{c} \\ X_1 \\ X_2 \end{array} \begin{array}{c} \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{pmatrix} 1.000 & 0.229 \\ 0.229 & 1.000 \end{pmatrix} \end{array}. \tag{16}$$

Claramente la correlación entre X_1 y X_2 es considerablemente pequeña, tomando en cuenta que en valor absoluto es más cercana a cero que a uno, y además los Factores de Inflación de la varianza para los estimadores de los coeficientes de este modelo, los cuales se encuentran dados en la tabla 15, resultan ambos menores que 5, por que la colinealidad entre X_1 y X_2 no resultará un problema en las estimaciones que se hagan de los coeficientes

VIF_1	VIF_2
1.055	1.055

Cuadro 15: Factores de inflación de la varianza, modelo 3.

Por último, el número de condición de la matriz $\mathbf{X}_1'\mathbf{X}_1$ esta dado por

$$K(\mathbf{X}_1'\mathbf{X}_1) = 1.26. \tag{17}$$

Dicho número de condición es mucho menor a 100 lo que constata que la colinealidad existente entre las variables independientes X_1 y X_2 es débil.

Adicional Si gusta regresar de este modelo a uno que considere las escalas originales y se base en el que ya fue construido, basta tomar los coeficientes estimados 52.577 para el intercepto 1.468 para el coeficiente que multiplica a la variable independiente $x1$ y 0.662 para la el coeficiente restante. ■

Observación 3. *A partir de este punto I_p representa a una matriz identidad de dimensión $p \times p$ con $p \in \mathbb{N}$, y 0_p representa un vector columna de dimensión $p \times 1$ con todas sus entradas iguales a cero.* △

Ejercicio 3 (Análisis de datos de servicio hospitalario (Ridge Regression))**:**

Hacer el análisis de los datos de la tabla 1 que correspondan al servicio hospitalario, utilizando el método de regresión Ridge. Hacer lo siguiente:

a) Determine el valor del parámetro de sesgo k siguiendo el procedimiento propuesto por Hoerl et al. (1975).

b) Determinar el valor del parámetro de sesgo k utilizando el método de validación cruzada

c) Estime para ambos valores de k obtenidos en los incisos (a y b) los coeficientes de regresión ridge $\beta_{(k)}$, considerando las ecuaciones normales (tipo ridge)

$$(X'X + kI_p)\hat{\beta}_{(k)} = X'Y.$$

d) Compare y comente las soluciones del inciso anterior.

e) Compare y comente la mejor solución del inciso (c) con el modelo propuesto en el inciso e) del problema.

Solución. **a)** En este ejercicio se buscará ajustar el modelo (1) utilizando regresión de Ridge, sin embargo, de acuerdo a Hastie y Tibsharani[15] considerando este modelo con las variables independientes centradas y escaladas se tiene que la estimación Ridge para el intercepto esta dada por $\hat{\alpha} = \overline{W} = 4977.892$, y por ende es posible considerar de manera equivalente el modelo (2) y obtener mediante regresión Ridge las estimaciones para los coeficientes de dicho modelo utilizando la matriz de diseño $\mathbf{X} = (X_1, \ldots, X_5)$ como fue definida en el inciso **b)** del ejercicio 1, de este modo además se cumplirá una de las condiciones estipuladas en el artículo de Hoerl (1975), es decir que la matriz de diseño $\mathbf{X}$ estará escalada de tal suerte que $\mathbf{X'X}$ sea una matriz de correlación. Ahora, de acuerdo al artículo de Hoerl (1975) una manera de estimar el valor del parámetro de sesgo k que resulte óptimo[16] es

$$k_h = \frac{p\hat{\sigma}^2}{\hat{\delta}'\hat{\delta}},$$

donde p es el número de parámetros en el modelo (2), es decir $p = 5$, $\hat{\sigma}^2$ es una estimación de la varianza de los términos de error en el modelo (2), la cual fue calculada usando la suma de cuadrados de los residuales del modelo (2) estimado por mínimos cuadrados y arrojó un valor de $\hat{\sigma}^2 = 378270$, y $\hat{\delta}$ es el estimador de mínimos cuadrados para los coeficientes del modelo (2) el cual puede encontrarse en la tabla 3. De este modo el valor del parámetro de sesgo obtenido por este método es

$$k_h = 1.709 \cdot 10^{-3}. \tag{18}$$

b) Primeramente, para determinar un rango posible en el que buscar el valor de k óptimo mediante validación cruzada, se realizó un gráfico de la traza de Ridge[17] el cual se presenta en la figura 9, en la misma se puede observar como el vector de parámetros Ridge estimados $\hat{\delta}_k$ comienza a degenerarse a cero para valores cercanos a 10, por lo que se esperaría que el valor óptimo para k

[15] Ver referencia 1, pp. 64.

[16] En el sentido de que este k sea tal que el error cuadrático promedio que comete el estimador Ridge sea el mínimo posible.

[17] En dicho gráfico se comete adrede un abuso de notación, ya que se omite la notación $\hat{\delta}$ para el estimador, con la finalidad de que los elementos en la gráfica no se vean apelmazados.

se encontrará dentro del rango de 0 a 10. Por otro lado, se estimó la cota superior para este valor de k óptimo, vista en ayudantías, utilizando las cantidades calculadas en el ejercicio anterior de la siguiente manera

$$cota = \frac{2\hat{\sigma}^2}{\hat{\delta}'\hat{\delta}} = 6.837 \cdot 10^{-4}.$$

Dado que las estimaciones utilizadas para este cálculo cuentan con una variabilidad alta,[18], y de acuerdo a lo observado en la traza de Ridge, se decidió buscar el k óptimo en un rango de 0 a 1.

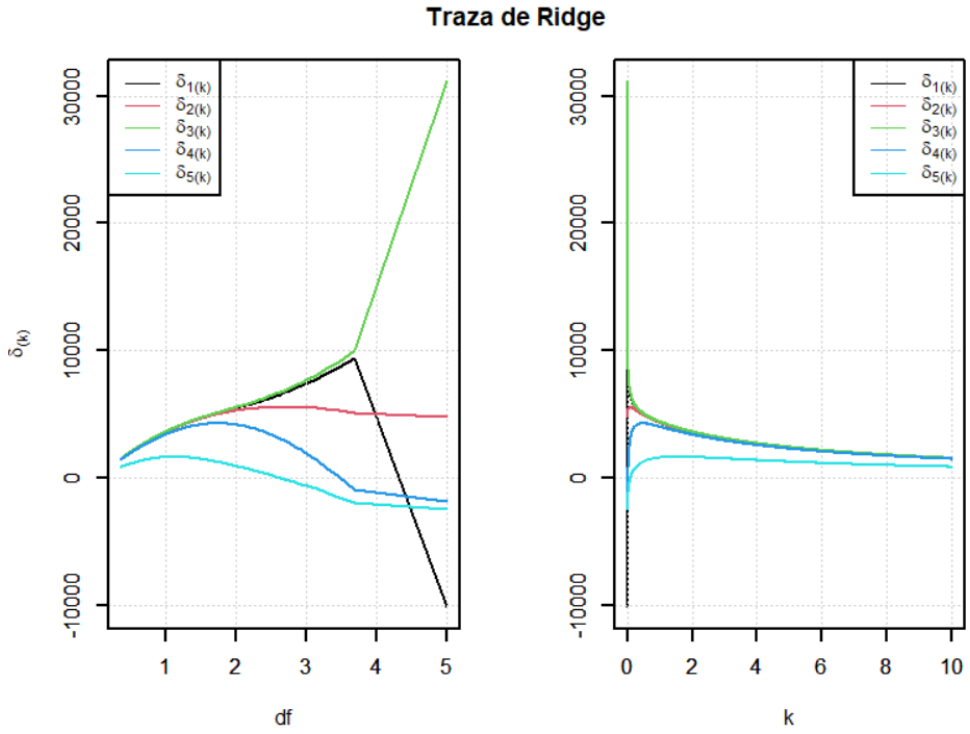

Figura 9: Gráficos de la traza de Ridge.

Para poder llevar a cabo la validación cruzada se programó en R el método conocido como K-fold cross validation y para determinar el valor adecuado para el parámetro de sesgo k se utilizaron $K = 17$ grupos realizando entonces una validación cruzada del tipo $Leave - One - Out$, este procedimiento se repitió 10 veces utilizando una rejilla de 1000 valores entre cero y uno, dado que la estimación para el k óptimo resultaba ser menor a 0.5, se decidió repitir este algoritmo pero con una rejilla de 1000 valores entre 0 y 0.5 para intentar mejorar la precisión de la estimación, el valor

[18]Debido a los problemas de multicolinealidad detectados en el modelo (2) en el ejercicio 1.

resultante para k por este método fue[19]

$$k_{vc} = 0.0846, \tag{19}$$

Por último, se deja en la figura 10 una gráfica de los diversos valores de k entre cero y cero punto cinco utilizados y el error promedio estimado que comete el modelo en sus predicciones con dichos valores de k.

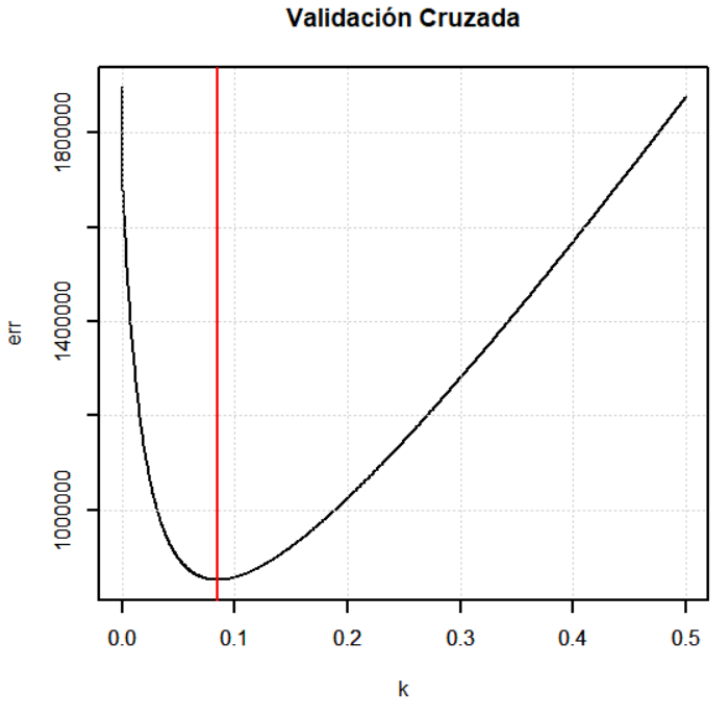

Figura 10: gráfica de los diversos valores de k entre cero y cero punto cinco utilizados y el error cuadrático promedio estimado que comete el modelo con dichos valores de k. En linea roja vertical se remarca el valor óptimo para k (19).

c) Denote por $\hat{\delta}_{(k_h)}$ al vector de estimaciones Ridge de los coeficientes del modelo obtenido usando el parámetro de sesgo k_h, calculado en el inciso **a)** de este ejercicio, entonces se tiene que

$$\hat{\delta}_{(k_h)} = (\mathbf{X}'\mathbf{X} + k_h I_p)^{-1}\mathbf{X}'Y_W = \begin{pmatrix} 9743.096 \\ 4860.627 \\ 11479.145 \\ -2156.766 \\ -2540.559 \end{pmatrix},$$

[19]Este valor fue comprobado con la función *RidgeCV* de Python para corroborar el resultado.

donde $Y_W = W - \overline{W} = \begin{pmatrix} W_1 - \overline{W} & \cdots & W_{17} - \overline{W} \end{pmatrix}'$. De este modo los valores ajustados por este modelo pueden escribirse como:

$$\widehat{E}[W_i - \overline{W} | X_{i1}, \ldots, X_{i5}] = 9743.096 X_{1i} + \ldots - 2540.559 X_{i5}, \ i \in \{1, \ldots, 17\}. \tag{20}$$

Por otra parte, denote por $\hat{\delta}_{(k_{vc})}$ al vector de estimaciones Ridge de los coeficientes del modelo obtenido usando el parámetro de sesgo k_{vc}, calculado en el inciso **b)** de este ejercicio, entonces se tiene que

$$\hat{\delta}_{(k_{vc})} = (\mathbf{X}'\mathbf{X} + k_h I_p)^{-1} \mathbf{X}' Y_W = \begin{pmatrix} 6778.872 \\ 5570.405 \\ 6968.747 \\ 2825.609 \\ -139.701 \end{pmatrix}.$$

De este modo los valores ajustados por este modelo pueden escribirse como:

$$\widehat{E}[W_i - \overline{W} | X_{i1}, \ldots, X_{i5}] = 6778.872 X_{1i} + \ldots - 139.701 X_{i5}, \ i \in \{1, \ldots, 17\}. \tag{21}$$

d) Para este inciso se ocuparon dos criterios, el primero de ellos serán los factores de inflación de la varianza de los modelos ajustados (20) y (21), dichos factores de inflación de la varianza se calcularon de acuerdo a las notas del doctor Rogelio Ramos Quiroga de la siguiente manera, primero se realizaron en R los siguientes productos matriciales

$$(\mathbf{X}'\mathbf{X} + kI_p)^{-1} (\mathbf{X}'\mathbf{X}) (\mathbf{X}'\mathbf{X} + kI_p)^{-1}, \tag{22}$$

con $k \in \{k_{vc}, k_h\}$. Así, los factores de inflación de la varianza para el modelo (20) se obtuvieron como la diagonal del producto (22) con $k = k_h$ y se presentan en la tabla 16

$VIF_1^{k_h}$	$VIF_2^{k_h}$	$VIF_3^{k_h}$	$VIF_4^{k_h}$	$VIF_5^{k_h}$
14.284	7.606	14.528	12.236	3.254

Cuadro 16: Factores de inflación de la varianza, modelo (20).

De manera análoga se obtuvieron los factores de inflación de la varianza del modelo ajustado (21) arrojando los resultados expuestos en la tabla 17

$VIF_1^{k_{vc}}$	$VIF_2^{k_{vc}}$	$VIF_3^{k_{vc}}$	$VIF_4^{k_{vc}}$	$VIF_5^{k_{vc}}$
0.708	2.177	0.800	2.055	1.217

Cuadro 17: Factores de inflación de la varianza, modelo ajustado (21).

Y como segundo criterio se obtuvieron los coeficientes de determinación R^2 para ambos modelos, los cuales se calcularon en Python arrojando un coeficiente de determinación

$$R^2 = 0.9907, \tag{23}$$

para el modelo ajustado (20). Y un coeficiente de determinación

$$R^2 = 0.9846, \tag{24}$$

para el modelo ajustado (21). Ahora, note que pese a que el modelo ajustado (20) posee un R^2 mayor al del modelo ajustado (21), hay tres factores de inflación de la varianza en este modelo que superan al 10 y por ende, al menos tres de las estimaciones de los coeficientes estarán siendo pobremente estimados debido a problemas de multicolinealidad. Por ende, dado que uno de los principales objetivos de utilizar regresión Ridge consiste en conservar todas las variables independientes disminuyendo los problemas de multicolinealidad, se prefiere el modelo ajustado (21) debido a que en el mismo ninguno de los factores de inflación de la varianza supera al 10, además de que la diferencia entre los coeficientes de determinación de ambos modelos no parece ser tan grande.

e) Observe que los factores de inflación de la varianza para el mejor modelo[20] ajustado en el inciso **c)** de este problema y los factores de inflación de la varianza del modelo (6)-(7) propuesto en el inciso **e)** del problema 1, los cuales fueron presentados en las tablas 17 y 8 respectivamente, no parecen presentar una diferencia significativa que nos haga decantarnos por uno u otro modelo, inclusive todos estos factores de inflación de la varianza resultan menores a 5 por lo que la multicolinealidad existente en ambos modelos no es un problema para las estimaciones de los coeficientes de los mismos. Sin embargo, observe que el coeficiente de determinación R^2 del modelo ajustado (21) es de 0.9846, mientras que el coeficiente de determinación del modelo propuesto en el inciso **e)**[21] del ejercicio 1 es de 0.9847, por lo que, se concluye que el modelo propuesto en el ejercicio 1 inciso **e)** pareciera ser mejor opción, ya que utilizando un menor número de variables explicativas consigue un coeficiente de determinación superior al del mejor modelo del inciso **c)**. ■

Ejercicio 4 (Análisis de datos de la tabla 9, correspondientes a la evolución calorífica del cemento, utilizando el método de regresión Ridge):

Hacer el análisis de los datos de la tabla 9 que correspondan a la evolución calorífica del cemento, utilizando el método de regresión Ridge. Hacer lo siguiente:

a) Determine el valor del parámetro de sesgo k siguiendo el procedimiento propuesto por Hoerl et al. (1975).

b) Determinar el valor del parámetro de sesgo k utilizando el método de validación cruzada

c) Estime para ambos valores de k obtenidos en los incisos (a y b) los coeficientes de regresión ridge $\beta_{(k)}$, considerando un procedimiento de mínimos cuadrados ordinarios, aumentando las matrices de la siguiente forma: Sean

$$X_A = \begin{pmatrix} X \\ \sqrt{k} I_p \end{pmatrix} \text{ y } Y_A = \begin{pmatrix} Y \\ 0_p \end{pmatrix},$$

Las estimaciones tipo ridge de los coeficientes de regresión tipo ridge son:

$$\hat{\beta}_{(k)} = (X_A' X_A)^{-1} X_A' Y_A.$$

[20] Modelo ajustado (21).
[21] Ver (7).

d) Compare y comente las soluciones del ejercicio anterior.

e) Compare y comente la mejor solución del inciso (c) con el modelo propuesto en el inciso (f) del problema 2.

Solución. **a)** El objetivo de los primeros incisos de este ejercicio será estimar los coeficientes del modelo (10) utilizando regresión de Ridge, sin embargo, recuerde que considerando este modelo con las variables independientes centradas y escaladas se tiene que la estimación Ridge para el intercepto esta dada por $\hat{\alpha} = \overline{y} = 95.423$, y por ende es posible considerar de manera equivalente el modelo (11) y obtener mediante regresión Ridge las estimaciones para los coeficientes de dicho modelo utilizando la matriz de diseño $\mathbf{X} = (X_1, X_2, X_3, X_4)$ como fue definida en el inciso **b)** del ejercicio 2, de esto modo además se cumplirá una de las condiciones estipuladas en el artículo de Hoerl (1975), es decir que la matriz de diseño $\mathbf{X}$ estará escalada de tal suerte que $\mathbf{X}'\mathbf{X}$ sea una matriz de correlación. Ahora, de acuerdo al artículo de Hoerl (1975) una manera de estimar el valor del parámetro de sesgo k que resulte óptimo es

$$k_h = \frac{p\hat{\sigma}^2}{\hat{\delta}'\hat{\delta}},$$

donde p es el número de parámetros en el modelo (11), es decir $p = 4$, $\hat{\sigma}^2$ es una estimación de la varianza de los términos de error en el modelo (11), la cual fue calculada usando la suma de cuadrados de los residuales del modelo (11) estimado por mínimos cuadrados y arrojó un valor de $\hat{\sigma}^2 = 5.318$, y $\hat{\alpha}$ es el vector de estimadores de mínimos cuadros el cual puede encontrarse en la tabla 10. De este modo

$$k_h = 0.0116.$$

b) Nuevamente para determinar un rango posible en el que buscar el valor de k óptimo mediante validación cruzada, se realizó un gráfico de la traza de Ridge[22] el cual se presenta en la figura 11, en la misma se puede observar como el vector de parámetros ridge estimados $\hat{\alpha}_k$ comienza a degenerarse a cero para valores cercanos a 35, por lo que se esperaría que el valor óptimo para k se encontrará dentro del rango de 0 a 35. También se estimó la cota superior para este valor de k óptimo, aprovechando las cantidades calculadas en el ejercicio anterior de la siguiente manera

$$cota = \frac{2\hat{\sigma}^2}{\hat{\delta}'\hat{\delta}} = 5.812 \cdot 10^{-3}.$$

Atendiendo a lo ocurrido en el inciso anterior, en el cual la estimación de esta cota resulto un poco pequeña para el valor del k óptimo, se decidió buscar en un rango de 0 a 10 tomando en cuenta los resultados obtenidos en la gráfica de la traza de Ridge. De este modo, utilizando nuevamente la función programada para realizar validación cruzada, se estimó el valor de k óptimo estableciendo $K = 13$ grupos, es decir, se realizó una validación cruzada del tipo $Leave - One - Out$ utilizando una rejilla de 1000 valores posibles para k entre 0 y 10, este se repitió 10 veces con lo que se obtuvo un valor menor a 0.5 para el k óptimo, de este modo se decidió volver a correr todo este algoritmo pero esta vez utilizando una rejilla de 1000 valores posibles para k entre 0 y 0.5 con la finalidad de

[22] En dicho gráfico se comete adrede un abuso de notación, ya que se omite la notación $\hat{\alpha}$ para el estimador, con la finalidad de que los elementos en la gráfica no se vean apelmazados.

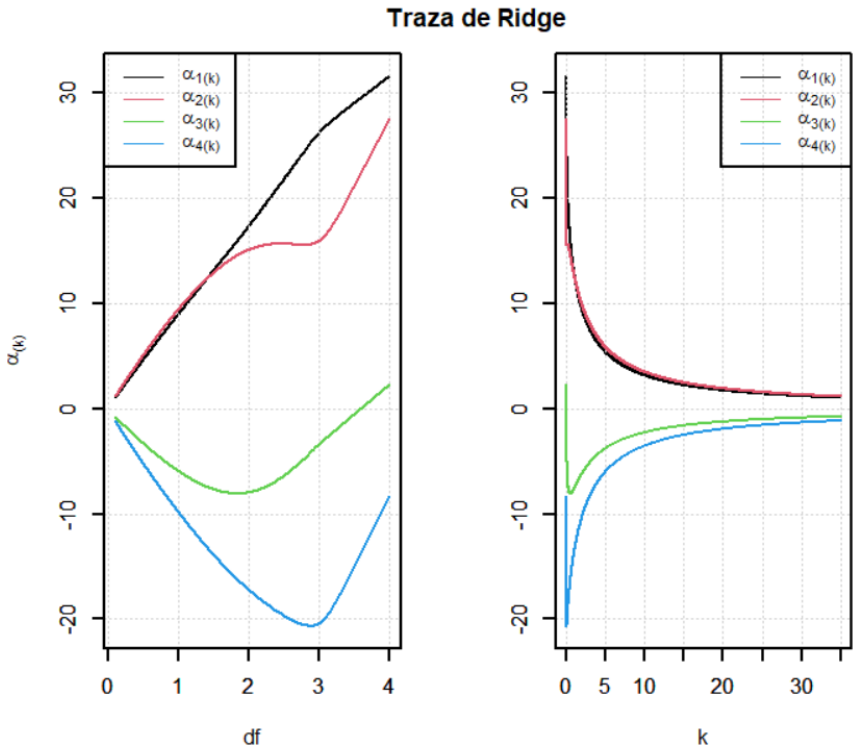

Figura 11: Gráficos de la traza de Ridge.

mejorar la precisión de la estimación dada, dando como resultado el siguiente valor del parámetro de sesgo

$$k_{vc} = 1.001 \cdot 10^{-2}.$$

Para finalizar este inciso, puede verse en la figura 12 una gráfica de algunos de los valores de k entre cero y cero punto cinco utilizados y el error promedio estimado que comete el modelo en sus predicciones con dichos valores de k.

c) Se define ahora para cada $k > 0$ la matriz aumentada $\mathbf{X}_{A,k}$ de la siguiente manera

$$\mathbf{X}_{A,k} = \begin{pmatrix} \mathbf{X} \\ \sqrt{k} I_p \end{pmatrix}.$$

Y la matriz columna aumentada Y_A de observaciones como

$$Y_A = \begin{pmatrix} Y_{\overline{y}} \\ 0_p \end{pmatrix},$$

donde $Y_{\overline{y}} = \begin{pmatrix} y_1 - \overline{y} & \cdots & y_{13} - \overline{y} \end{pmatrix}$. De esta manera, si se denota por $\hat{\alpha}_{(k_h)}$ al vector de estimaciones Ridge de los coeficientes del modelo (11) obtenido usando el parámetro de sesgo k_h, calculado en

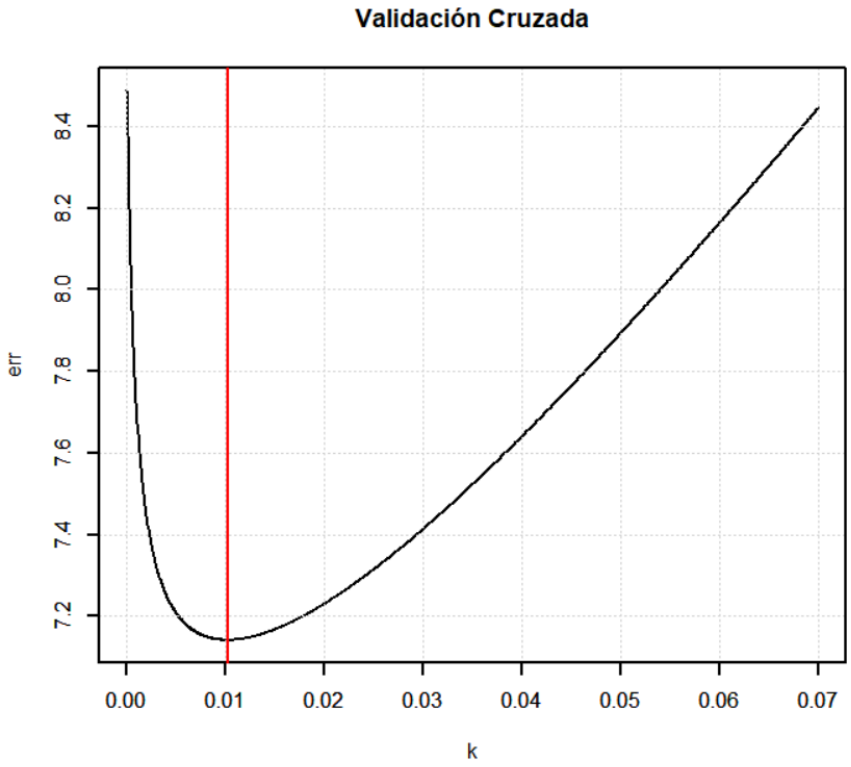

Figura 12: gráfica de los algunos valores de k entre cero y cero punto cinco utilizados y el error cuadrático promedio estimado que comete el modelo con dichos valores de k. En linea roja vertical se remarca el valor óptimo para k.

el inciso **a)** de este ejercicio, entonces se tiene que

$$\hat{\alpha}_{(k_h)} = (X'_{A,k_h} X_{A,k_h})^{-1} X'_{A,k_h} Y_A = \begin{pmatrix} 26.623 \\ 16.303 \\ -3.024 \\ -20.081 \end{pmatrix}.$$

Por lo cual, los valores ajustados por este modelo pueden escribirse como:

$$\widehat{E}[y_i - \overline{y} | X_{i1}, \ldots, X_{i4}] = 26.623 X_{1i} + \ldots - 20.081 X_{i4}, \ i \in \{1, \ldots, 13\}. \tag{25}$$

Por otro lado, si se denota por $\hat{\alpha}_{(k_{vc})}$ al vector de estimaciones Ridge de los coeficientes del modelo obtenido usando el parámetro de sesgo k_{vc}, calculado en el inciso **a)** de este ejercicio, entonces se tiene que

$$\hat{\alpha}_{(k_{vc})} = (X'_{A,k_{vc}} X_{A,k_{vc}})^{-1} X'_{A,k_{vc}} Y_A = \begin{pmatrix} 26.799 \\ 16.500 \\ -2.864 \\ -19.886 \end{pmatrix}.$$

De este modo, los valores ajustados por este modelo pueden escribirse como:

$$\widehat{E}[y_i - \overline{y}|X_{i1}, \ldots, X_{i4}] = 26.799X_{1i} + \ldots - 19.886X_{i4},\ i \in \{1, \ldots, 13\}\,. \tag{26}$$

d) Para comparar los modelos ajustados (25) y (26) se volverá a hacer uso de los factores de inflación de la varianza de ambos modelos, los cuales se encuentran dados en las tablas 18 y 19 respectivamente,[23] y de sus coeficientes de determinación R^2 los cuales se encuentran dados en (27) y (28) respectivamente

$VIF_1^{k_h}$	$VIF_2^{k_h}$	$VIF_3^{k_h}$	$VIF_4^{k_h}$
2.967	4.528	2.896	4.679

Cuadro 18: Factores de inflación de la varianza, modelo (25).

$VIF_1^{k_{vc}}$	$VIF_2^{k_{vc}}$	$VIF_3^{k_{vc}}$	$VIF_4^{k_{vc}}$
3.162	5.667	3.126	5.939

Cuadro 19: Factores de inflación de la varianza, modelo ajustado (26).

$$R^2 = 0.98210, \tag{27}$$

$$R^2 = 0.98213. \tag{28}$$

Observe que entre estos dos modelos las diferencias resultan muy sutiles, lo que era de esperarse dado que los dos parámetros de sesgo estimados k_h y k_{vc} son bastante similares, sin embargo, en el modelo ajustado que considera al parámetro k_{vc} se tiene que dos de sus Factores de la inflación de la varianza superan por poco al 5, que es el primer nivel de alerta sobre la posibilidad de que la multicolinealidad existente en el modelo pueda estar causando problemas en la estimación de los coeficientes realizada para ese ajuste, de este modo y dado que las diferencias en los coeficientes de determinación son prácticamente inexistentes se determino elegir el modelo ajustado (25).

e) Observe que los factores de inflación de la varianza para el mejor modelo[24] ajustado en el inciso **c)** de este problema y los factores de inflación de la varianza del modelo (14)-(14) propuesto en el inciso **f)** del problema 2, los cuales fueron presentados en las tablas 18 y 15 respectivamente, no parecen presentar una diferencia significativa que nos haga decantarnos por uno u otro modelo, inclusive todos estos factores de inflación de la varianza resultan menores a 5 por lo que la multicolinealidad existente en ambos modelos no es un problema para las estimaciones de los coeficientes de los mismos. Sin embargo, observe que el coeficiente de determinación R^2 del modelo ajustado (21) es de 0.9821, mientras que el coeficiente de determinación del modelo propuesto en el inciso **f)**[25] del ejercicio 2 es de 0.978, dado que la diferencia entre estos coeficientes de determinación no se

[23] Y fueron calculados en la manera que ya fue explicada anteriormente en el ejercicio anterior.
[24] Modelo ajustado (26).
[25] Ver (15).

considera excesivamente grande, y dado que el modelo propuesto en el inciso **f)** del problema 2 ocupa dos variables independientes menos que el mejor modelo ajustado en el inciso **c)** de este ejercicio me decantaría por este último, ya que con menos variables independientes logra un ajuste similar. ■

1. Anexo

Se dará un breve compendio de criterios vistos en clase y algunos complementos obtenidos en la bibliografía indicada. En clase se vio que un factor de la inflación de la varianza VIF mayor a 10 indica que el coeficiente asociado a este factor esta siendo pobremente estimado por efectos asociados con la multicolinealidad entre las variables independientes consideradas en el modelo, sin embargo, en clase también se recalco que desde que algún VIF excede el valor de 5 se debe de ser precavido al respecto de la multicolinealidad en el modelo y su posible efecto adverso en la estimación del coeficiente asociado a dicho VIF. Por otro lado, el número de condición es un indicador que nos dice la gravedad de la multicolinealidad existente en el modelo de la siguiente manera, si dicho número de condición es mayor a 1000 se considera que la multicolinealidad existente entre las variables independientes del modelo es severa, por otra parte si el número de condición se encuentra entre 100 y 1000 se dirá que la multicolinealidad entre las variables independientes en el modelo es moderada, por último, si el numero de condición K es menor a 100 se dirá que no existe un problema de multicolinealidad serio. Por otro lado, los índices de condicionamiento indican problemas de multicolinealidad cuando al menos uno de ellos excede el 30, dado que el mayor índice de condicionamiento coincide con la raiz cuadrada positiva del número de condición, si este no excede el 1000 entonces no hay necesidad alguna de hacer un análisis del modelo haciendo uso de estos índices.

2. Bibliografía y Referencias.

1. Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (2nd 2009, Corr. 9th Printing 2017 ed.). Springer.

2. Belsley, D. A., Kuh, E., Welsch, R. E. (2013). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity: 546. Wiley-Interscience.

3. Rawlings, J. O. (2001). Applied Regression Analysis: A Research Tool (Springer Texts in Statistics) (English Edition) (2nd ed.). Springer.

4. Ramos, R. Modelos Estadísticos I.

Por último, una prueba formal de lo comentado por Tibsharani y Hastie puede consultarse dando click aquí.