Modelos Estadísticos I (Fecha: 28/04/2021)

Tarea #6

Estudiante: Roberto Vásquez Martínez *NUA:* 424662

Problema 1

(a) En la siguiente gráfica además de ver la dispersión entre la variable respuesta W con cada V_i, $i = 1, 2, \ldots, 5$ vemos también la dispersión entre las mismas variables regresoras. Cabe decir que hemos centrado y escalado cada variable para tener media 0 y norma 1.

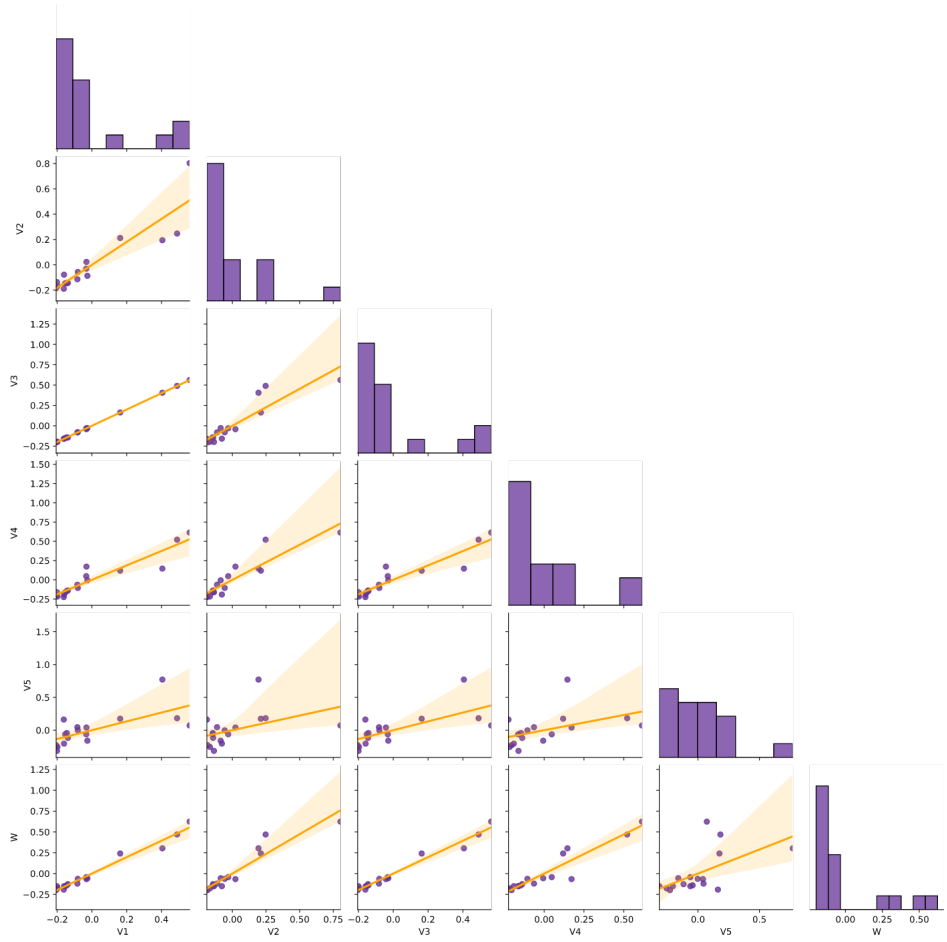

Figura 1: Diagrama de dispersión entre variables

Y los coeficientes de correlación son

```
Correlaciones con la variable respuesta

 Cor(W,V1): 0.985646

 Cor(W,V2): 0.945173

 Cor(W,V3): 0.985992

 Cor(W,V4): 0.940356

 Cor(W,V5): 0.578580
```

Podemos observar que la variable menos correlacionada con la respuesta es V_5, y podemos ver que las variables restantes están muy correlacionadas, pero hay que analizar las correlaciones entre estas variables para proponer un modelo de regresión adecuado.

(b)

A continuación veremos si hay evidencia de multicolinealidad. Visualizamos primero la matriz de correlaciones de las variables regresoras.

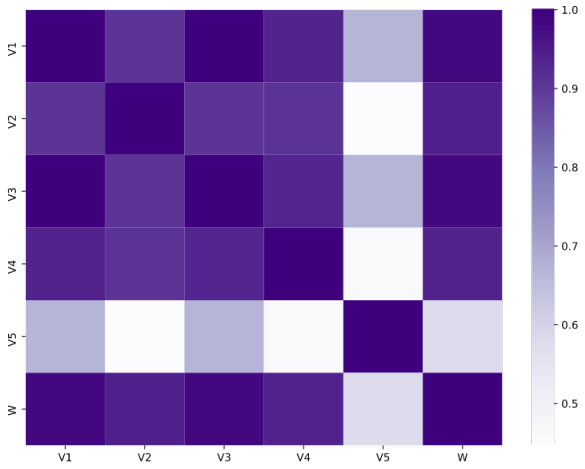

Figura 2: Matriz de correlaciones

y calculamos los valores explícitamente

```
Matriz de correlaciones:
[[1.      0.90738 0.9999  0.93569 0.6712 ]
 [0.90738 1.      0.90715 0.91047 0.44665]
 [0.9999  0.90715 1.      0.93317 0.67111]
 [0.93569 0.91047 0.93317 1.      0.46286]
 [0.6712  0.44665 0.67111 0.46286 1.     ]]
```

Podemos ver a partir de aquí que el conjunto de variables $\{V_1, \ldots, V_4\}$ están altamente correlacionadas entre sí, y V_5 es la menos correlacionada con cada una de estas variables.

Obtenemos el número de condición

```
Numero de condicion:
278.87212383682476
```

Vemos del número de condición que hay un problema de colinealidad moderada pues no se supera el umbral de 1000 para el número de condición.

(c)

Calculamos los factores de inflación de la varianza

```
Factores de inflación de la varianza
[9597.57076    7.94059   8933.0865    23.29386    4.27984]
```

Las variables que V_1, V_3 están en una relación de colinealidad alta con respecto a las restantes por su factor de inflación, pero esta relación se puede deber a que V_1 es linealmente dependiente de las demás y no necesariamente por la correlación que tiene con V_3 y viceversa. Además V_5 tiene un VIF aceptable, por lo que V_5 no esta inmersa en el problema de multicolinealidad. Del inciso anterior y los factores de inflación podemos ver que el problema de multicolinealidad esta sujeta a relaciones de dependencia lineal entre las variables $\{V_1, V_2, V_3, V_4\}$

(d) Del inciso anterior lo que haremos será elegir un subconjunto de las variables $\{V_1, \ldots, V_4\}$ e incluir V_5 en el modelo pues esta variable parece ser linealmente independiente. Probaremos las selecciones de variables $\{V_i, V_5\}$ para $i = 1, 2, 3, 4$ y el modelo $\{V_1, V_3, V_5\}$, pues V_1, V_3 tienen correlaciones altas con la variable respuesta W, y aunque sus VIF sean altos no necesariamente se tiene esto por la correlación alta que hay entre V_1, V_3.

Hacemos las regresiones anteriores, comparamos su R^2 ajustado y su AIC, el que tenga mejor desempeño será el seleccionado.

Modelo	Adj R^2	AIC
$[V_1, V_5]$	0.982	-66.2515
$[V_2, V_5]$	0.914	-39.7114
$[V_3, V_5]$	0.983	-67.0398
$[V_4, V_5]$	0.898	-36.9340
$[V_1, V_3, V_5]$	0.982	-65.3453

En el último modelo a pesar de tener buenos resultados tenemos un número de condición de 26789, por lo que hay un severo problema de multicolinealidad aquí a pesar de que V_1, V_3 esten muy correlacionadas con la respuesta, el primer análisis de multicolinealidad no nos garantizaba este resultado, aunque se podía intuir por la fuerte correlación entre V_1 y V_2.

(e)

El modelo propuesto será una regresión con las variables $\{V_3, V_5\}$, la regresiones anteriores las hicimos con los datos centrados y normalizados, también para la variable respuesta.

El análisis de multicolinealidad para este caso nos arroja lo siguiente

```
Matriz de correlacion:
[[1.      0.67111]
 [0.67111 1.     ]]

Factores de inflacion:
[1.81947 1.81947]

Valores propios:
[1.67111 0.32889]

Numero de condicion:
5.081
```

Por lo que en este modelo el problema de multicolinealidad es mucho menor con respecto al modelo original.

□

Problema 2

(a) La matriz de correlaciones de las variables regresoras X_1, X_2, X_3, X_4 es la siguiente

```
Matriz de correlaciones:
[[ 1.       0.22858 -0.82413 -0.24545]
 [ 0.22858  1.      -0.13924 -0.97295]
 [-0.82413 -0.13924  1.       0.02954]
 [-0.24545 -0.97295  0.02954  1.     ]]
```

Podemos observar que hay una correlación muy fuerte entre las parejas de variables (X_1, X_3) y (X_2, X_4).

(b) Los factores de inflación para cada una de las variables son los siguientes

```
Factores de inflacion:
[ 38.49621   254.42317   46.86839   282.51286]
```

De aquí podemos ver que todas las variables están inmersas en un problema de multicolinealidad con las restantes, en particular este problema es más fuerte en las variables X_2 y X_4, que viendo los resultados del inciso anterior se puede deber a la correlación mayor que hay entre ellas.

(c) Los valores propios de X^TX son

```
Los valores propios son:
[2.2357  1.57607 0.18661 0.00162]
```

Si $L^2 = (\hat{\beta} - \beta)^T(\hat{\beta} - \beta)$ es la distancia entre β el valor verdadero y la estimación, por lo visto en clase tenemos que

$$\mathbb{E}[L^2] = \sigma^2 \sum_{i=1}^{4} \frac{1}{\lambda_i},$$

donde λ_i son los valores propios de X^TX, al tener valores propios pequeños tenemos que la esperanza anterior es grande, y esto da evidencia de tener problemas de multicolinealidad en este modelo.

El número de condición en este caso es

```
Numero de condicion:
1376.8806213592732
```

por lo que tenemos un problema de multicolinealidad severa.

(d) Como hemos mencionado antes el número de condición nos dice que hay multicolinealidad severa, el tamaño de los VIF se puede explicar por las correlaciones altas que hay entre las parejas de variables (X_1, X_3) y (X_2, X_4), en particular la pareja (X_2, X_4) tiene una correlación más fuerte lo que podría explicar el VIF de cada una de estas variables respecto a la otra pareja.

Lo anterior da indicio a que la fuente de colinealidad esta en las correlaciones altar entre estas parejas de variables.

(e) La idea para proponer un modelo es elegir una variable de cada pareja (X_1, X_3) y (X_2, X_4), al igual que antes ajustaremos cada uno de los cuatro selecciones de variables que tenemos, nos fijaremos en su R^2 y AIC.

Los resultados los resumimos en la siguiente tabla

Modelo	Adj R^2	AIC
$[X_1, X_2]$	0.975	-42.4764
$[X_1, X_4]$	0.967	-39.1547
$[X_2, X_3]$	0.819	-16.8592
$[X_3, X_4]$	0.924	-28.0438

Nos quedamos así con el modelo de regresión respecto a las variables $\{X_1, X_2\}$.

(f) Elegimos en el inciso anterior el modelo de regresión con las variables $\{X_1, X_2\}$. Hacemos un análisis de multicolinealidad para ver si este modelo reduce la multicolinealidad severa del modelo original.

Los resultados con esta selección de variables son los siguientes

```
Matriz de correlacion:
[[1.      0.22858]
 [0.22858 1.     ]]

Factores de inflacion:
[1.05513 1.05513]

Valores propios:
[1.22858 0.77142]

Numero de condicion:
1.5926196192852162
```

y bajo estos resultados que en este problema no hay evidencia de un problema de multicolinealidad.

□

Problema 3

(a) La propuesta de Hoerl (1975) es la siguiente

$$k_H = \frac{p\hat{\sigma}^2}{\hat{\beta}^T\hat{\beta}}$$

donde $\hat{\sigma}$ y $\hat{\beta}$ las calculamos a partir de una regresión con todas las variables en el problema 1. El valor que obtenemos con los datos del problema 1 es

```
Lambda Hoerl:
0.001702
```

(b)

Ahora determinamos el valor del parámetro mediente validación cruzada, por el tamaño de la muestra lo hacemos mediante LOO, es decir usando todos los datos menos 1 para entrenamiento y el restante para validación.

```
Lambda_CV:
0.087036
```

Observamos que en este caso el valor del parámetro de sesgo es considerablemente más grande, como usamos en el parámetro de Hoerl el producto interior de $\hat{\beta}$, este se esperaría más pequeño, pues por el problema de multicolinealidad la distancia de $\hat{\beta}$ a el valor verdadero es grande, luego la norma de $\hat{\beta}$ es grande.

Variamos el valor de λ y hacemos la regresión Ridge también calculando los grados de libertad efectivos. Las gráficas de la traza de Ridge son las siguientes

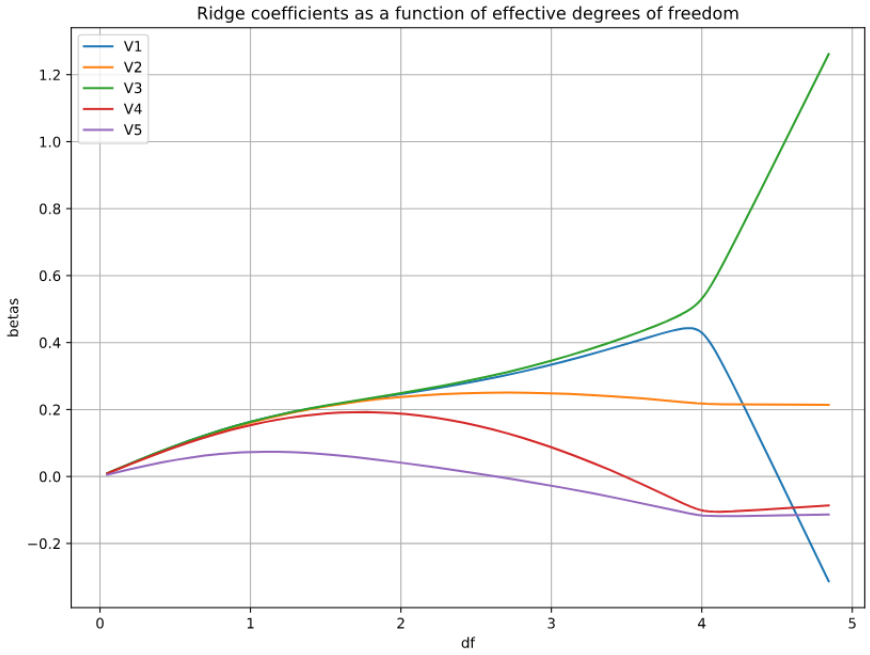

Figura 3: Traza de Ridge grados de libertad efectivos

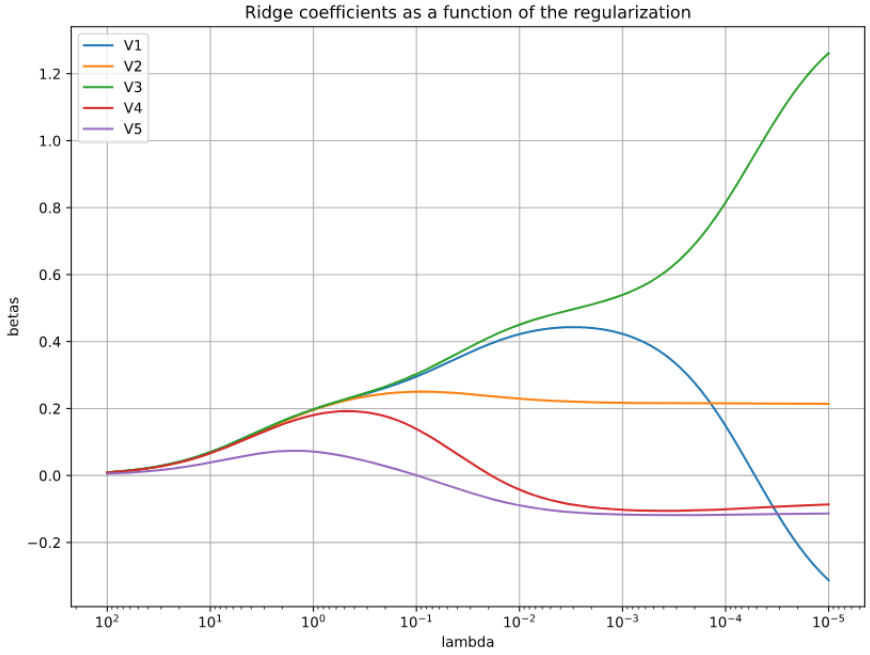

Figura 4: Traza de Ridge con λ

(c)

Ajustamos el modelo Ridge para cada λ ya sea por validación cruzada y con la propuesta de Hoerl. Los resultados con el λ de Hoerl fue el siguiente

```
Parametros Ridge Problema 1 (Hoerl):
[ 0.43805  0.21846  0.51643 -0.09722 -0.1144 ]

 Error cuadrático medio Problema 1 (Hoerl):
0.000545
```

Mientras que con validación cruzada LOO es

```
Parametros Ridge Problema 1 (CV):
[ 0.30317  0.25043  0.31154  0.1291  -0.00515]

 Error cuadrático medio Problema 1 (CV):
0.000913
```

(d)

Observamos que el error en en CV es mayor, pero esto se puede explicar debido a que la reestricción con el parámetro λ de Hoerl es considerablemente más pequeño y por ende la estimación es menos sesgada. A pesar de la diferencia el parámetro elegido por CV es mejor pues por ser la reestricción menor en el λ de Hoerl se hace la estimación de los parámetros en este caso en un espaio parametral más grande, mientras que en CV al variar el lambda podemos evitar un sobre ajuste.

(e) Para comparar los modelos, el Ridge con el parámetro λ el elegido por validación cruzada y el propuesto en el **Problema 1 e)** hacemos validación cruzada. Cabe aclarar que para ajustar ambos modelos con los datos completos consideramos las variables explicativas y la respuesta centradas y escaladas. Con esta observación cada que hagamos LOO con los todos los datos menos 1 consideramos el intercepto, pues cuando quitamos un dato ya no tenemos en el ajuste a los datos centrados y normalizados.

El resultado para el error cuadrático medio en ambos casos es el siguiente

```
Error cuadrático medio modelo Ridge_CV P1:
[0.00172]
Error cuadrático modelo [V3,V5]:
[0.0015]
```

En este caso los errores es mejor el modelo elegido en V_3, V_5, y por la diferencia de los parámetros de sesgo λ de Hoerl y el calculado por validación cruzada se puede deber a que en efecto es mejor quitar algunas variables. Aunque al ser la diferencia tan pequeña también podemos usar regresión Ridge con todas las variables a pesar de tener una estimación sesgada.

□

Problema 4

(a)

Al igual que en el problema anterior calculamos el parámetro λ de Hoerl. El resultado es el siguiente

```
Lambda Hoerl:
0.01162
```

(b)

Usamos validación cruzada como en el problema anterior. Al igual que en el caso anterior todas las estimaciones las hacemos con los datos centrados y escalados, considerando el intercepto cuando quitamos alguna observación.

El resultado para λ con validación cruzada LOO es

```
Lambda CV_Auto
0.01059
```

Graficamos las trazas de Ridge

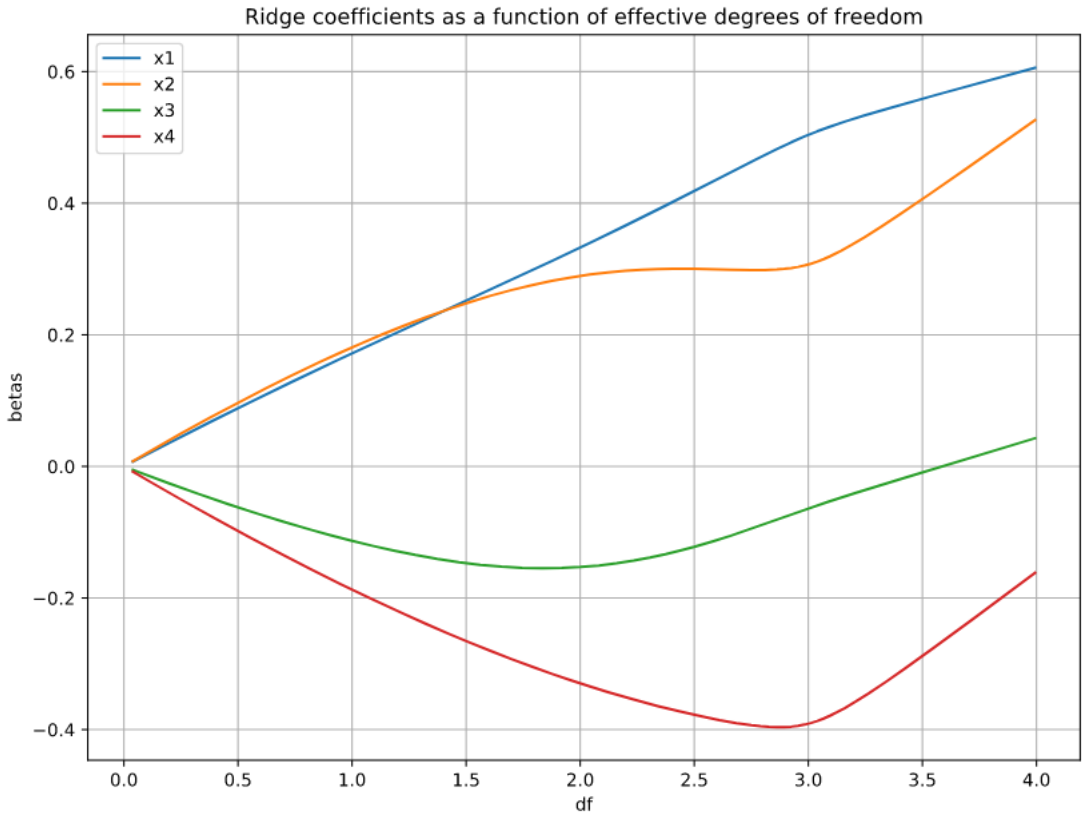

Figura 5: Traza de Ridge considerando grados de libertad efectivos

Y la traza variando el λ

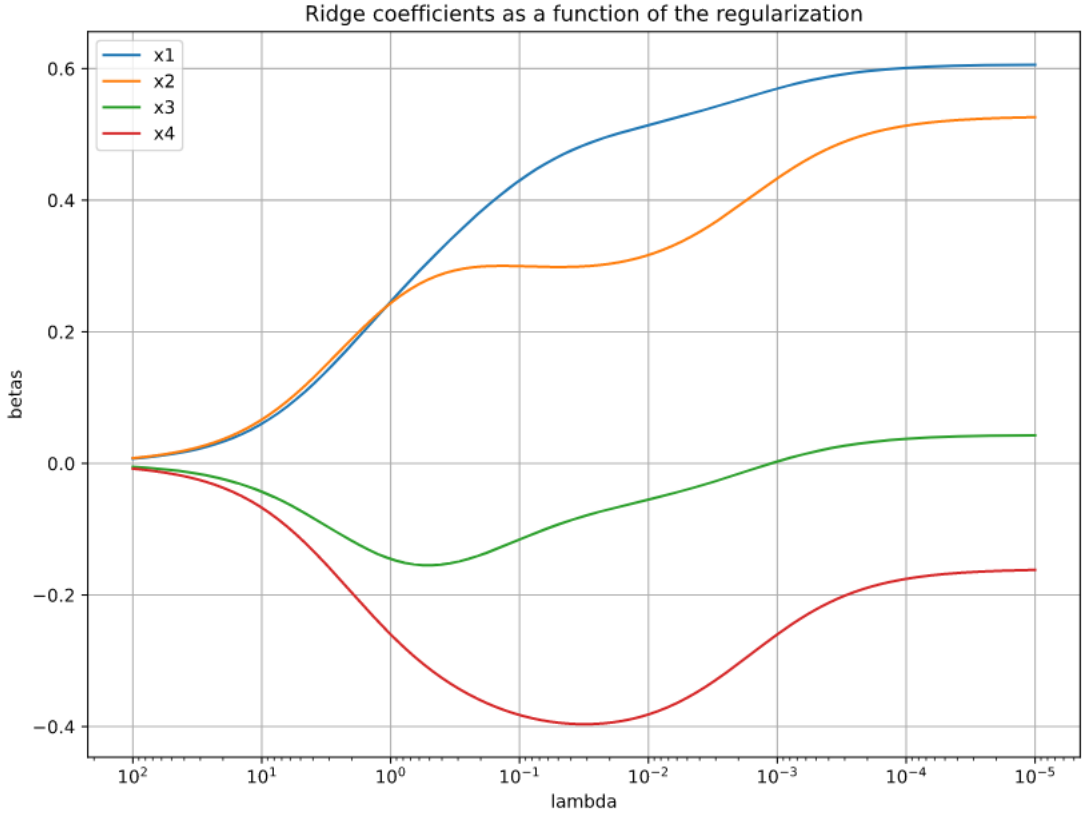

Figura 6: Traza de Ridge variando λ

En este caso los valores del parámetro de sesgo son bastante parecidos, aún así consideramos mejor el elegido por validación cruzada por tener la bondad de detectar un sobre ajuste.

(c)

Obtenemos las estimaciones de los coeficientes de los parámetros con la estimación sugerida. Las estimaciones para cada caso son las siguientes

```
Parámetros regresión Ridge (Hoerl):
[ 0.51086  0.31285 -0.05803 -0.38534]

 Parámetros regresión Ridge (CV):
[ 0.51297  0.31512 -0.05613 -0.38309]
```

(d)

Observamos que las soluciones son parecidas lo que era de esperarse dado que los parámetros de sesgo λ en cada caso eran muy parecidos. Consideramos utilizar la estimación del parámetro de sesgo el obtenido por validación cruzada debido a la bondad para evitar algún sobre ajuste.

(e)

Al igual que en el **Problema 3 e)** hacemos validación cruzada LOO para calcular el error cuadrático medio entre el modelo propuesto en **Problema 2 e)** y el modelo Ridge. Los resultados son los siguientes

```
Error cuadrático medio modelo Ridge_CV P2:
[0.00263]
Error cuadrático modelo [x1,x2]:
[0.00266]
```

En este caso tiene mejor desempeño el modelo Ridge con el parámetro de sesgo λ calculado por validación cruzada.

□