Tarea 6: Modelos Estadísticos I.

Rojas Gutiérrez Rodolfo Emmanuel

3 de mayo de 2021

Ejercicio 1 (Ejercicio 1. Datos consumidores)**:**

a) y b) Primeramente observe en la figura 1 En ella podemos observar en parte superior las

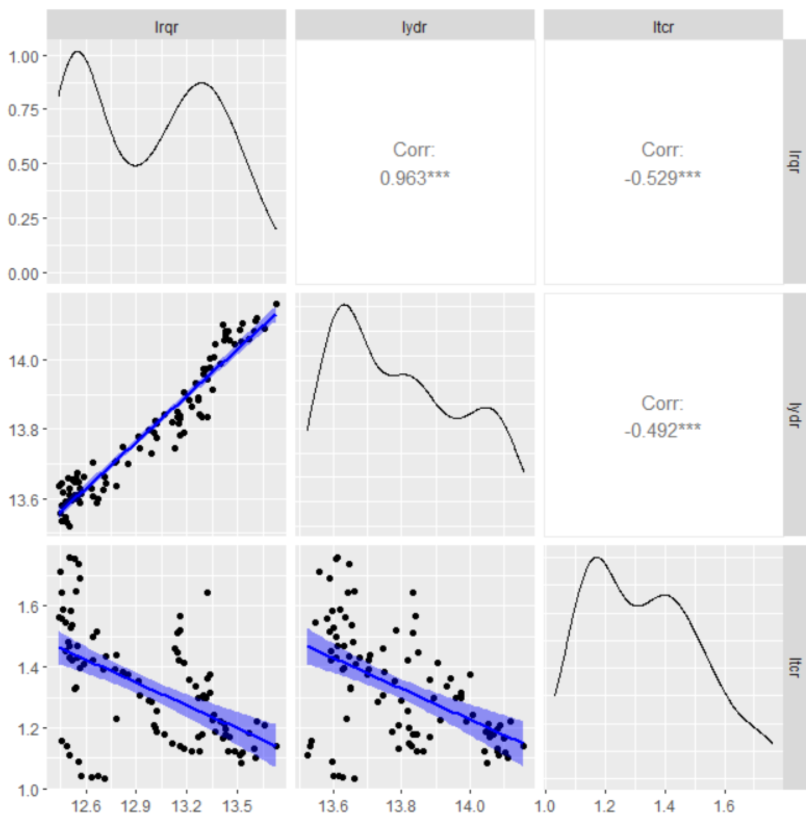

Figura 1: PairPlot con correlaciones de las variables independientes $lrqr, lydr$ y $ltcr$.

entradas por encima de la diagonal de la matriz de correlación entre las variables independientes, dado que las entradas en la diagonal son 1 y por la simetría de está matriz, esto es suficiente para determinar toda la matriz de correlación entre las variables independientes, note ahora que la correlación más alta es la existente entre las variables independientes $lrqr$ y $lydr$. Este hecho se ve reflejado en la parte inferior del gráfico donde es posible observar los gráficos de dispersión de las variables independientes con modelos de regresión lineal simple ajustados con sus respectivos intervalos de confianza. Note como el modelo para $lrqr$ y $lydr$ parece presentar una variabilidad casi inexistente en sus bandas de confianza, o que indica que una variable explica casi perfectamente a la otra. A partir de aquí se consideraron dos maneras de realizar los diagnósticos la primera es considerando la matriz $\mathbf{X_{CE}}$ que es la matriz cuyas columnas corresponden a los valores de las variables independientes centrados y escalados, y $\mathbf{X_E}$ que es la matriz cuyas columnas consideran un intercepto y a las variables independientes escaladas. Dado que la matriz de correlaciones de los datos no cambia si los centramos y escalamos, o solo los escalamos los $VIF's$ para ambos casos resultan iguales y se presentan en la figura 2: Dado que los VIFS asociados a los coeficientes que

```
  lrqr   lydr   ltcr
14.673 13.945  1.396
```

Figura 2: Vif's para los modelos de regresión.

pre-multiplican a las variables $lrqr$ y $lydr$ exceden el 10, se considera que estos coeficientes están siendo pobremente estimados debido a problemas de multicolinealidad. Por otro lado, se obtuvieron los números de condición de las matrices $\mathbf{X_{CE}}'\mathbf{X_{CE}}$ y $\mathbf{X_E}'\mathbf{X_E}$ los cuales se presentan en las figuras (1) y 2 respectivamente

$$K(\mathbf{X_{CE}}'\mathbf{X_{CE}}) = 486325.2. \tag{1}$$

$$K(\mathbf{X_E}'\mathbf{X_E}) = 65.699. \tag{2}$$

Llaman la atención dos cosas, en el modelo en que consideramos las columnas centradas y escaladas el número de condición es menor a 100 y por ende se considera que no existe un problema de multicolinealidad serio en este caso, sin embargo, el modelo que solo esta escalado tiene un número de condición por arriba de 1000 lo que indica un problema serio de multicolinealiadad. Esto no es de sorprender ya que al centrar y escalar las variables de respuesta recuerde que las variables independientes se vuelven ortogonales al intercepto, eliminando cualquier rastro de colinealiadad con este, por lo que se podría sospechar de la existencia de multicolinealidad con este. Dado que los VIFs en los modelos centrados y escalados arrojaron resultados alarmantes en ambos modelos, y debido a la fuerte relación lineal detectada entre las variables independientes $lydr$ y $lrqr$ se decidió que lo correcto sería quedarse con una y solo uno de estas variables independientes, de este modo para el modelo con variables independientes centradas y escaladas y variable respuesta centrada se obtuvieron los siguientes resúmenes eliminando $lydr$ y $lrqr$:

```
Call:
lm(formula = YCE ~ lrqr + ltcr - 1, data = dataC)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.120729 -0.035090  0.002992  0.037276  0.102336 

Coefficients:
     Estimate Std. Error t value Pr(>|t|)    
lrqr  1.81925    0.05508  33.027   <2e-16 ***
ltcr -0.02733    0.05508  -0.496    0.621    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04676 on 94 degrees of freedom
Multiple R-squared:  0.9424,	Adjusted R-squared:  0.9412 
F-statistic: 769.1 on 2 and 94 DF,  p-value: < 2.2e-16
```

Figura 3: Resumen modelo con variables independientes centradas y escaladas y variable respuesta centrada. Eliminando la variable independiente $lydr$.

```
Call:
lm(formula = YCE ~ lydr + ltcr - 1, data = dataC)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.074269 -0.020489 -0.001975  0.018901  0.082560 

Coefficients:
     Estimate Std. Error t value Pr(>|t|)    
lydr   1.8069     0.0402  44.951   <2e-16 ***
ltcr  -0.1005     0.0402  -2.499   0.0142 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.035 on 94 degrees of freedom
Multiple R-squared:  0.9677,	Adjusted R-squared:  0.967 
F-statistic:  1410 on 2 and 94 DF,  p-value: < 2.2e-16
```

Figura 4: Resumen modelo con variables independientes centradas y escaladas y variable respuesta centrada. Eliminando la variable independiente $lrqr$.

Observe que el modelo en el que se elimina a $lrqr$ obtiene mejor R^2, mejor R^2 ajustada y sus dos coeficientes resultan significativos, por lo que, nos decantamos por este modelo. Los coeficientes en la escala original para el modelo en la figura 4 se presentan en la figura 5. Podemos ver que en este caso el intercepto no resulta significativo, de este modo se exploró la posibilidad de removerlo, (debido a los resultados obtenidos en el análisis que consideraba al intercepto y las variables solamente escaladas), lo cual arrojo el resumen presentado en la 6. Sorprende lo bueno de que resulta este modelo ya que alcanza un R^2 y R^2 ajustado de 1 y todos sus coeficientes resultan significativamente distintos de cero, bajo un nivel de significacnia del 5 % Además la correlación entre $lydr$ y $lrqr$ es la más pequeña en valor absoluto entre todas las variables independientes, cosa que puede corroborarse

```
Call:
lm(formula = lcpr ~ lydr + ltcr, data = data)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.074269 -0.020489 -0.001975  0.018901  0.082560 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.31390    0.32972  -0.952   0.3436    
lydr         1.01813    0.02277  44.711   <2e-16 ***
ltcr        -0.05534    0.02226  -2.486   0.0147 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03519 on 93 degrees of freedom
Multiple R-squared:  0.9677,	Adjusted R-squared:  0.967 
F-statistic:  1395 on 2 and 93 DF,  p-value: < 2.2e-16
```

Figura 5: Resumen modelo con variables en escala original. Eliminando la variable independiente $lrqr$.

en la figura 1, con un valor -0.492, de este modo el los VIFS de este modelo, ambos son iguales a

$$1/(1-(-0.492)^2) \approx 1.319.$$

Por lo que ningún parámetro esta siendo pobremente estimado debido a la multicolinealidad.

c) Para $i \in \{lqrq, \ldots, ltcr\}$ se define el vector columna de datos X_i como aquel que resulta de centrar el vector columna i y posteriormente escalar el vector resultante dividiendo por su longitud.[1] Sea ahora $\mathbf{X_{CE}} = \begin{pmatrix} X_{lrqr} & X_{lydr} & X_{ltcr} \end{pmatrix}$ ajustará[2] inicialmente el modelo

$$E[lpcr_i|X_{ilqrq}, \ldots, X_{iltcr}] = \alpha + \delta_1 X_{ilqrq} + \ldots + \delta_3 X_{iltcr},\ i \in \{1, \ldots, 96\}\,. \tag{3}$$

Sin embargo, de acuerdo a Hastie Tibsharani[3] considerando este modelo con las variables independientes centradas y escaladas se tiene que la estimación Ridge para el intercepto esta dada por $\hat{\alpha} = \overline{lpcr} = 3.6568$, y por ende es posible considerar de manera equivalente el modelo (4) y obtener mediante regresión Ridge las estimaciones para los coeficientes de dicho modelo utilizando la matriz de diseño $\mathbf{X_{CE}}$

$$E[lpcr_i - \overline{lpcr}|X_{ilqrq}, \ldots, X_{iltcr}] = \delta_1 X_{ilqrq} + \ldots + \delta_3 X_{iltcr},\ i \in \{1, \ldots, 96\}\,. \tag{4}$$

De este modo además se cumplirá una de las condiciones estipuladas en el artículo de Hoerl (1975), es decir que la matriz de diseño $\mathbf{X_{CE}}$ estará escalada de tal suerte que $\mathbf{X_{CE}}'\mathbf{X_{CE}}$ sea una matriz de correlación. Ahora, de acuerdo al artículo de Hoerl (1975) una manera de estimar el valor del

[1] Entiéndase por longitud su norma euclideana.
[2] Note que $\mathbf{X_{CE}}$ es exactamente la misma matriz que se definió en el inciso anterior
[3] Ver referencia 1, pp. 64.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.08161 -0.02055 -0.00203  0.02253  0.08096

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
lydr  0.996522   0.001798 554.135  < 2e-16 ***
ltcr -0.067180   0.018459  -3.639 0.000447 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03517 on 94 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 7.239e+06 on 2 and 94 DF,  p-value: < 2.2e-16
```

Figura 6: Resumen modelo con variables en escala original sin intercepto. Eliminando la variable independiente $lrqr$.

parámetro de sesgo k que resulte óptimo[4] es

$$k_h = \frac{p\hat{\sigma}^2}{\hat{\delta}'\hat{\delta}} = 0.001522308,$$

donde p es el número de parámetros en el modelo (4), es decir $p = 3$, $\hat{\sigma}^2$ es una estimación de la varianza de los términos de error en el modelo (4), la cual fue calculada usando la suma de cuadrados de los residuales del modelo (4) estimado por mínimos cuadrados y arrojó un valor de $\hat{\sigma}^2 = 0.0009841351$, y $\hat{\delta}$ es el estimador de mínimos cuadrados para los coeficientes del modelo (4). 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

$$\hat{\delta}_{(k_h)} = (\mathbf{X_{CE}}'\mathbf{X_{CE}} + k_h I_p)^{-1}\mathbf{X_{CE}}'Y_{CE} = \begin{pmatrix} 0.60119040 \\ 1.24672385 \\ -0.05802807 \end{pmatrix},$$

donde $Y_{CE} = lpcr - \overline{lpcr} =$. De este modo los valores ajustados por este modelo pueden escribirse como:

$$E[lpcr_i - \overline{lpcr}|X_{ilqrq}, \ldots, X_{iltcr}] = 0.60119040X_{ilqrq} + \ldots + -0.05802807X_{iltcr}, \ i \in \{1, \ldots, 96\}.$$

Para esto modelo se obtuvo un R^2 de

$$R^2 = 1 - SS(RES)/SS(TOT) = 0.9743479.$$

Este coeficiente de determinación es menor al R^2 del modelo propuesto en el inciso anterior, y debido a que este modelo considera más parámetros se prefiere al modelo anterior. Además, el objetivo de la regresión Ridge no fue logrado, ya que en la figura 7 podemos ver los factores de inflación de la varianza bajo este modelo a penas cambiaron, comparados con los dados al inicio de este ejercicio, con esta elección de k y dos de ellos siguen resultando mayores a 5

```
      lrqr      lydr      ltcr 
13.523944 12.858682  1.386553 
```

Figura 7: VIFS Modelo Ridge.

```
Call:
lm(formula = CPR ~ ., data = dataOr)

Residuals:
   Min     1Q Median     3Q    Max 
-53767 -17351   -661  15691  59654 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.470e+05  3.714e+04   3.958 0.000149 ***
RQR          3.093e-01  6.091e-02   5.077 1.99e-06 ***
YPDR         6.106e-01  5.806e-02  10.517  < 2e-16 ***
TCR         -7.404e+03  4.185e+03  -1.769 0.080140 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25870 on 92 degrees of freedom
Multiple R-squared:  0.9785,	Adjusted R-squared:  0.9778 
F-statistic:  1395 on 3 and 92 DF,  p-value: < 2.2e-16
```

Figura 8: Resumen modelo con variables en escala original. Todas las variables

d) Por último, el resumen del modelo original sin ninguna escala esta dado en la figura 8 Se destaca que este modelo con todas las variables predictoras, no resulta mejor en R^2 ni en R^2 ajustada al mejor modelo presentado en los primeros dos incisos.

Ejercicio 2:

Solución. **a) y b)** Primeramente observe en la figura 9 En ella podemos observar en parte superior las entradas por encima de la diagonal de la matriz de correlación entre las variables independientes, dado que las entradas en la diagonal son 1 y por la simetría de está matriz, esto es suficiente para determinar toda la matriz de correlación entre las variables independientes, preocupan en especial la correlación existentes entre LWT y CWT con un valor de 0.901, la correlación existente entre LEA y $WTWAT$ con un valor de 0.861 y la correlación existente entre DEP y $WTWAT$ con un valor de -0.833. En un segundo plano también llaman la atención las correlaciones entre $LESL$ y DEP $LESL$ y LEA, por lo que, deberemos tener cuidado con estas variables. A partir de aquí se consideraron dos maneras de realizar los diagnósticos la primera es considerando la matriz $\mathbf{X_{CE}}$ que es la matriz cuyas columnas corresponden a los valores de las variables independientes centrados y escalados, y $\mathbf{X_E}$ que es la matriz cuyas columnas consideran un intercepto y a las variables independientes escaladas. Dado que la matriz de correlaciones de los datos no cambia si los centramos y escalamos, o solo los escalamos los $VIF's$ para ambos casos resultan iguales y se presentan en la figura 10:

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

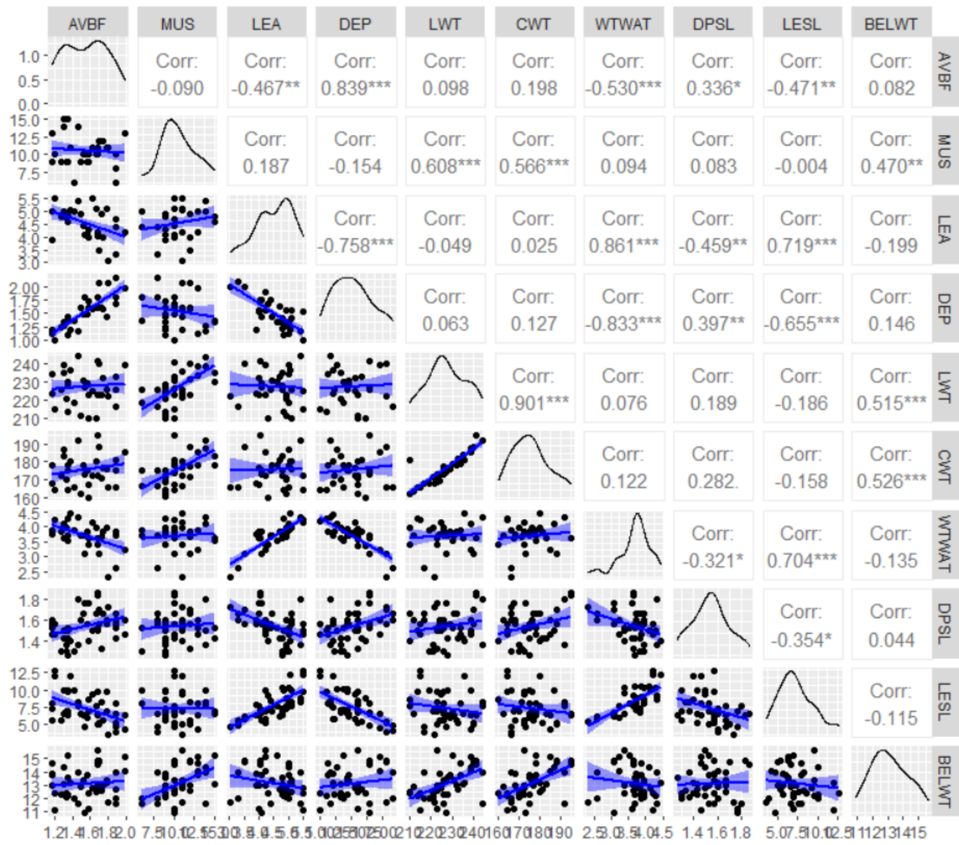

Figura 9: PairPlot con correlaciones de las 10 variables independientes.

Se observa que los VIFs asociados a los coeficientes que premultiplican a las variables $AVBF$ y LWT exceden al 5, lo que da una primera alerta de que estos coeficientes pueden estar siendo pobremente estimados por problemas de multicolinealidad en el modelo, por otro lado, los coeficientes asociados a las variables DEP, CWT y $WTWAT$ exceden con claridad al 10, por lo que estos coeficientes si estan siendo pobremente estimados debido a la multicolinealidad. Note que varias de estas variables fueron mencionadas en el análisis de la matriz de correlaciones de las variables independientes. Por otro lado, se obtuvieron los números de condición de las matrices $\mathbf{X_{CE}}'\mathbf{X_{CE}}$ y

AVBF	MUS	LEA	DEP	LWT	CWT	WTWAT	DPSL	LESL	BELWT
5.464	2.849	8.007	17.375	8.004	10.474	12.150	1.889	2.670	2.054

Figura 10: Vif's para los modelos de regresión.

posible.

$\mathbf{X_E}'\mathbf{X_E}$ los cuales se presentan en las figuras (5) y 6 respectivamente

$$K(\mathbf{X_{CE}}'\mathbf{X_{CE}}) = 103834.9. \tag{5}$$

$$K(\mathbf{X_E}'\mathbf{X_E}) = 130.386. \tag{6}$$

Llaman la atención dos cosas, en el modelo en que consideramos las columnas centradas y escaladas el número de condición se encuentra entre 100 y 1000 por ende se considera la multicolinealidad existente en el modelo es moderada, sin embargo, el modelo que solo esta escalado tiene un número de condición por arriba de 1000 lo que indica un problema serio de multicolinealiadad. Esto no es de sorprender ya que al centrar y escalar las variables de respuesta recuerde que las variables independientes se vuelven ortogonales al intercepto, eliminando cualquier rastro de colinealiadad con este, por lo que se podría sospechar de la existencia de multicolinealidad con dicho término de intercepto. De este modo, se tiene que las variables LWT y CWT estan fuertemente correlacionadas, por lo que se debería considerar quedarnos con una de ellas únicamente, dado que CWT posee el mayor factor de inflación de la varianza nos quedaremos con LWT, por otro lado, todo el conjunto de variables independientes en el conjunto $\{LEA, WTWAT, LESL, DEP\}$ parecen tener problemas de colinealidad, ya que las correlaciones de todas estas variables resultan ser bastante elevadas, sin embargo, observando la figura 10 se uno puede notar fácilmente que los factores de inflación de la varianza de DEP y $WTWAT$ resultan ser los mas grandes en todo el conjunto de datos, por lo que se decidió eliminar igualmente estas variables. [5]

c) Consideramos primeramente el modelo con las variables independientes centradas y escaladas y la variable respuesta centrada, en el que se removieron como ya se mencionó anteriormente las variables independientes en el conjunto $\{CWT, DEP, WTWAT\}$, las estimaciones así como un resumen para este modelo se consideran en la figura 11 Este mismo resumen pero con el modelo equivalente con los datos en las escalass originales se presenta en la figura 12 Luego por lo observado cuando se el análisis de la matriz de datos escalada en la que se considera la multicolinealidad con el intercepto, se decidió intentar ajustar el modelo anterior sin intercepto, a modo de analizar si la inclusión del mismo tenía efectos negativos por la posible multicolinealidad que este genera, lo que arrojó el análisis presentado en la figura 13. Puede verse que la significancia de varios coeficientes creció, además, en la última linea de la imagen presentada en la 13 se presentan todos los factores de inflación de la varianza de los coeficientes de este modelo, observe que ninguno rebasa el 5, por lo que la multicolinealidad no es un problema para las estimaciones de los coeficientes en este modelo.

Sin embargo, deseabamos encontrar un modelo con un mayor poder predictivo que el modelo original, y que además considerará el menor número de variables independientes en el, por lo cual, se decidió omitir las variables independientes $DSPL$ y LEA, las cuales no resultaban significativamente distintas de 0 bajo un nivel de significancia del 5 % en el modelo presentado en la figura 13, obteniendo así el modelo presentado en la figura 14, en el mismo todos los coeficientes resultan significativamente distintos de cero y se tiene una R^2 de 0.9972 y un R^2 ajustada de 0.9969. Por último, los factores de inflación de la varianza para este modelo se presentan al final de la imagen en la figura 14.

d) Para $i \in \{AVBF, \ldots, BELWT\}$ se define el vector columna de datos X_i como aquel que resulta de centrar el vector columna i y posteriormente escalar el vector resultante dividiendo por

[5] Si da tiempo, se presentará la tabla de descomposición de varianzas para este problema en el anexo.

```
Call:
lm(formula = YCE ~ . - WTWAT - CWT - DEP - 1, data = dataC)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2634 -1.8746 -0.0119  1.5919  4.5915

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
AVBF     4.900      3.265   1.501   0.1417
MUS     -7.874      3.927  -2.005   0.0521 .
LEA    -13.013      4.859  -2.678   0.0109 *
LWT      6.450      3.826   1.686   0.1000 .
DPSL    -1.505      3.316  -0.454   0.6525
LESL    -9.850      4.312  -2.284   0.0280 *
BELWT    8.641      3.646   2.370   0.0230 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.765 on 38 degrees of freedom
Multiple R-squared:  0.736,     Adjusted R-squared:  0.6873
F-statistic: 15.13 on 7 and 38 DF,  p-value: 2.943e-09
```

Figura 11: Resumen para modelo con las variables independientes centradas y escaladas y la variable respuesta centrada. Eliminado $\{CWT, DEP, WTWAT\}$.

su longitud.[6] Sea ahora $\mathbf{X_{CE}} = \begin{pmatrix} X_{AVBF} & \dots & X_{BELWT} \end{pmatrix}$ ajustará[7] inicialmente el modelo

$$E[FAT_i|X_{AVBFi}, \dots, X_{BELWTi}] = \alpha + \delta_1 X_{AVBF} + \dots + \delta_{10} X_{BELWTi},\ i \in \{1, \dots, 45\}\,. \quad (7)$$

Sin embargo, de acuerdo a Hastie Tibsharani[8] considerando este modelo con las variables independientes centradas y escaladas se tiene que la estimación Ridge para el intercepto esta dada por $\hat{\alpha} = \overline{FAT} = 55.082$, y por ende es posible considerar de manera equivalente el modelo (4) y obtener mediante regresión Ridge las estimaciones para los coeficientes de dicho modelo utilizando la matriz de diseño $\mathbf{X_{CE}}$

$$E[FAT_i - \overline{FAT}|X_{iAVBF}, \dots, X_{iBELWT}] = \delta_1 X_{iAVBF} + \dots + \delta_{10} X_{iBELWT},\ i \in \{1, \dots, 45\}\,. \quad (8)$$

De este modo además se cumplirá una de las condiciones estipuladas en el artículo de Hoerl (1975), es decir que la matriz de diseño $\mathbf{X_{CE}}$ estará escalada de tal suerte que $\mathbf{X_{CE}}'\mathbf{X_{CE}}$ 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[9] es

$$k_h = \frac{p\hat{\sigma}^2}{\hat{\delta}'\hat{\delta}} = 0.08565412,$$

donde p es el número de parámetros en el modelo (8), es decir $p = 10$, $\hat{\sigma}^2$ es una estimación de la varianza de los términos de error en el modelo (8), la cual fue calculada usando la suma de

[6] Entiéndase por longitud su norma euclideana.

[7] Note que $\mathbf{X_{CE}}$ es exactamente la misma matriz que se definió en el inciso anterior

[8] Ver referencia 1, pp. 64.

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

```
Call:
lm(formula = FAT ~ . - WTWAT - DEP - CWT, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.2634 -1.8746 -0.0119  1.5919  4.5915 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 39.53582   14.42668   2.740  0.00939 **
AVBF         3.04597    2.05709   1.481  0.14715   
MUS         -0.55789    0.28197  -1.979  0.05535 . 
LEA         -3.11137    1.17742  -2.643  0.01199 * 
LWT          0.10491    0.06305   1.664  0.10461   
DPSL        -1.53645    3.43093  -0.448  0.65689   
LESL        -0.69286    0.30743  -2.254  0.03023 * 
BELWT        1.09625    0.46880   2.338  0.02488 * 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.802 on 37 degrees of freedom
Multiple R-squared:  0.736,     Adjusted R-squared:  0.686 
F-statistic: 14.73 on 7 and 37 DF,  p-value: 5.394e-09
```

Figura 12: Resumen para modelo con las variables en la escala original. Eliminado $\{CWT, DEP, WTWAT\}$.

cuadrados de los residuales del modelo (8) estimado por mínimos cuadrados y arrojó un valor de $\hat{\sigma}^2 = 6.123102$, y $\hat{\delta}$ es el estimador de mínimos cuadrados para los coeficientes del modelo (4). 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

$$\hat{\delta}_{(k_h)} = (\mathbf{X_{CE}}'\mathbf{X_{CE}} + k_h I_p)^{-1}\mathbf{X_{CE}}'Y_{CE} = \begin{pmatrix} -3.049124 \\ -7.406572 \\ -5.581604 \\ 10.893713 \\ 3.202395 \\ 5.997023 \\ -6.552797 \\ -1.565149 \\ -6.561477 \\ 6.4300717 \end{pmatrix},$$

donde $Y_{CE} = FAT - \overline{FAT}$. De este modo los valores ajustados por este modelo pueden escribirse como:

$$E[FAT_i - \overline{FAT}|X_{iAVBF}, \ldots, X_{iBELWT}] = -3.049124X_{iAVBF} + \ldots + 6.4300717X_{iBELWT},\ i \in \{1, \ldots, 45\}\,.$$

Para esto modelo se obtuvo un R^2 de

$$R^2 = 1 - SS(RES)/SS(TOT) = 0.7972822.$$

```
Call:
lm(formula = FAT ~ . - WTWAT - DEP - CWT - 1, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.1084 -1.9828 -0.2033  2.3217  5.2895 

Coefficients:
      Estimate Std. Error t value Pr(>|t|)    
AVBF   4.06958    2.18933   1.859 0.070809 .  
MUS   -0.94613    0.26385  -3.586 0.000944 ***
LEA   -1.80011    1.16432  -1.546 0.130378    
LWT    0.22301    0.04981   4.477  6.7e-05 ***
DPSL   1.97462    3.44460   0.573 0.569852    
LESL  -0.66500    0.33254  -2.000 0.052712 .  
BELWT  1.37070    0.49566   2.765 0.008726 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.033 on 38 degrees of freedom
Multiple R-squared:  0.9975,	Adjusted R-squared:  0.997 
F-statistic:  2132 on 7 and 38 DF,  p-value: < 2.2e-16
    AVBF      MUS      LEA      LWT     DPSL     LESL    BELWT 
1.394475 2.017050 3.088487 1.914321 1.437845 2.432516 1.738977 
```

Figura 13: Resumen para modelo con las variables en la escala original. Eliminado $\{CWT, DEP, WTWAT\}$ y el término de intercepto.

Llama la atención ya que es menor que el R^2 del modelo completo con multicolinealidad, y es mucho menor al R^2 del modelo propuesto en el inciso anterior. Pese a ello, el objetivo de la regresión Ridge fue logrado, ya que en la figura 15 podemos ver los factores de inflación de la varianza para este modelo y todos ellos resultan menores a 5 ■

```
Call:
lm(formula = FAT ~ . - WTWAT - DEP - CWT - DPSL - LEA - 1, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.0875 -2.2142 -0.2983  2.4534  5.8113 

Coefficients:
      Estimate Std. Error t value Pr(>|t|)    
AVBF   4.75876    2.14680   2.217 0.032395 *  
MUS   -1.04002    0.26314  -3.952 0.000307 ***
LWT    0.20006    0.03381   5.917 6.18e-07 ***
LESL  -1.06979    0.22919  -4.668 3.39e-05 ***
BELWT  1.59964    0.47103   3.396 0.001556 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.087 on 40 degrees of freedom
Multiple R-squared:  0.9972,    Adjusted R-squared:  0.9969 
F-statistic:  2880 on 5 and 40 DF,  p-value: < 2.2e-16
    AVBF      MUS      LWT     LESL    BELWT 
1.317761 1.755067 1.864424 1.324354 1.448392 
```

Figura 14: Resumen para modelo con las variables en la escala original. Eliminado $\{CWT, DEP, WTWAT, DSPL, LEA\}$ y el término de intercepto.

```
    AVBF      MUS      LEA      DEP      LWT      CWT    WTWAT     DPSL     LESL    BELWT 
1.454868 1.292499 1.933313 1.796016 1.737246 1.752992 1.813882 1.081488 1.545317 1.219537 
```

Figura 15: VIFS Modelo Ridge.