Solución a la práctica 4.2 con R

En esta práctica vamos a utilizar los operadores cor, head, linearHypothesis, lm, str y summary. Para utilizar el operador linearHypothesis debemos instalar y cargar el paquete *car*.

Se piensa que el apoyo electoral al bipartidismo disminuye cuando aumenta el tamaño de la población de las localidades y cuando se consigue una mayor participación de los electores. También puede influir la situación socioeconómica de los votantes. Para comprobar la veracidad de estos supuestos, se han tomado los siguientes datos sobre los resultados de las elecciones al Congreso de los Diputados celebradas en 2015 para cada municipio de la Región de Murcia (fichero de datos: *Practica42.RData*): porcentaje de votantes que votó a uno cualquiera de los dos partidos mayoritarios hasta ese momento (*bipartidismo)*, total de miles de votantes (*votantes*) y total de miles de inscritos en el censo electoral (*censados*). Igualmente, de cada municipio se dispone de los datos de paro registrado en 2015 en porcentaje de su población total (*paro*). Para analizar las cuestiones descritas se plantea el siguiente modelo:

$$bipartidismo = \beta_0 + \beta_1 votantes + \beta_2 censados + \beta_3 paro + \varepsilon$$

Antes de responder a las preguntas examinamos el tipo de datos que contiene el fichero *Practica42.RData*. Para ello ejecutamos el código *str(Practica42)* y obtenemos la salida de la Figura 4.2.1:

Figura 4.2.1: str(Practica42)

```
Classes 'tbl_df', 'tbl' and 'data.frame':       45 obs. of  5 variables:
 $ BIPARTIDISMO: num  78.9 69.8 67.4 81.7 58.2 ...
 $ CENSADOS    : num  4.65 10.05 24.57 1.15 29.89 ...
 $ MUNICIPIO   : chr  "Abanilla" "Abaran" "Aguilas" "Albudeite" ...
 $ PARO        : num  8.64 5.53 8.73 13.07 12.25 ...
 $ VOTANTES    : num  3.62 7.29 16.67 0.94 21.46 ...
```

Podemos ver que todas las variables son numéricas, excepto la variable *municipio*, la cual contiene el nombre de los municipios de nuestra muestra. Además, con el código *head(practica42)* podemos ver las seis primeras observaciones de todas la variables.

Figura 4.2.2: head(practica42)

	BIPARTIDISMO	CENSADOS	MUNICIPIO	PARO	VOTANTES
1	78.88920	4.646	Abanilla	8.639909	3.619
2	69.83539	10.052	Abaran	5.534373	7.290
3	67.38870	24.566	Aguilas	8.734758	16.666
4	81.70213	1.151	Albudeite	13.073144	0.940
5	58.21452	29.887	Alcantarilla	12.253785	21.462
6	82.43045	0.807	Aledo	5.068598	0.683

1) Estime el modelo propuesto por MCO.

A continuación, estimamos por MCO el modelo propuesto y el resultado lo guardamos en el objeto *mod.1*. Para ello ejecutamos el siguiente código:

```
mod.1 <- lm(BIPARTIDISMO ~ VOTANTES + CENSADOS + PARO, data =
        practica42)
```

Este resultado de esta estimación se muestra en la Figura 4.2.3.

Figura 4.2.3: summary(mod.1)

```
Call:
lm(formula = BIPARTIDISMO ~ VOTANTES + CENSADOS + PARO, data = practica42)

Residuals:
    Min      1Q  Median      3Q     Max
-11.989  -3.606  -0.508   3.205  15.444

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.3499     4.7913  15.100   <2e-16 ***
VOTANTES      0.6425     0.6273   1.024    0.312
CENSADOS     -0.5287     0.4611  -1.146    0.258
PARO         -0.4656     0.4994  -0.932    0.357
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.367 on 41 degrees of freedom
Multiple R-squared:  0.2083,    Adjusted R-squared:  0.1504
F-statistic: 3.597 on 3 and 41 DF,  p-value: 0.02138
```

2) Examine los resultados de la anterior estimación, en particular fíjese en los resultados de los contrastes de significatividad individuales de los coeficientes de las variables explicativas y en el contraste de significatividad conjunta del modelo. ¿Nota algo que le parezca anómalo? ¿Es capaz de aventurar una explicación a dicha anomalía?

Los resultados de la estimación mostrados por RStudio incluyen los resultados de los contrastes de significatividad individual de los coeficientes de cada una de las variables explicativas y el contraste de significatividad conjunta del modelo. Llama la atención el hecho de que el p-valor de cada uno de los tres contrastes de significatividad individual (31.2%, 25.8% y 35.7%) indique no significatividad de la explicativa correspondiente y que, por el contrario, el bajo p-valor del contraste de significatividad conjunto (2.138%) tienda a señalar que, conjuntamente, el grupo de las tres variables sí resulta significativo a la hora de explicar el apoyo del electorado a los partidos mayoritarios. Aun siendo posible que esa contradicción en los resultados se deba al azar, también es factible que sea consecuencia del efecto engañoso provocado por la existencia de multicolinealidad fuerte entre algún grupo de variables explicativas.

3) ¿Qué método(s) sencillo(s) se le ocurre emplear para intentar confirmar si la explicación que ha aventurado para la anomalía es un hecho que realmente se da en los datos? Aplíquelo(s) y obtenga conclusiones.

Para confirmar o descartar la existencia de multicolinealidad fuerte entre algunas de las variables explicativas, se puede comenzar por calcular las diferentes correlaciones entre ellas. Para ello obtenemos la matriz de correlaciones con los datos de las columnas 2, 4 y 5 del data.frame *practica42*. El resultado aparece en la Figura 4.2.4 y se obtiene con el código *cor(practica42[,c(2,4,5)])*. *cor* es el operador que nos permite obtener dicha matriz, *[,c(2,4,5)]* nos permite seleccionar las variables *censados*, *paro* y *votantes* indicando la columna en la que están colocadas. Esto los sabemos de la Figura 4.2.2. Notar que *c(2,4,5)* esta precedido por una coma para indicar a RStudio que seleccione columnas y no filas.

Figura 4.2.4: cor(practica42[,c(2,4,5)])

```
             CENSADOS        PARO     VOTANTES
CENSADOS   1.00000000 -0.01707311  0.99915711
PARO      -0.01707311  1.00000000 -0.01639622
VOTANTES   0.99915711 -0.01639622  1.00000000
```

El altísimo valor absoluto del coeficiente de correlación de las variables *censados* y *votantes*, 0.99916, y el bajo valor absoluto del resto de coeficientes de correlación, 0.01707 y 0.016396, confirman la existencia de multicolinealidad muy fuerte entre *censados* y *votantes*.

Otro modo de examinar la multicolinealidad entre las variables explicativas del modelo es calcular los coeficientes de determinación de las tres regresiones auxiliares de cada una de las variables explicativas sobre las otras dos $(R_j^2, j = 1,2,3)$. Los resultados obtenidos son:

Figura 4.2.5: summary(lm(CENSADOS ~ VOTANTES + PARO, data = practica42))

```
Call:
lm(formula = CENSADOS ~ VOTANTES + PARO, data = practica42)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2655 -0.5766 -0.4239  0.4083 11.8468

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.577421   1.600754   0.361    0.720
VOTANTES     1.359138   0.008616 157.742   <2e-16 ***
PARO        -0.018224   0.167070  -0.109    0.914
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.13 on 42 degrees of freedom
Multiple R-squared:  0.9983,    Adjusted R-squared:  0.9982
F-statistic: 1.244e+04 on 2 and 42 DF,  p-value: < 2.2e-16
```

Figura 4.2.6: summary(lm(VOTANTES ~ CENSADOS + PARO, data = practica42))

```
Call:
lm(formula = VOTANTES ~ CENSADOS + PARO, data = practica42)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5544 -0.3055  0.2883  0.4050  4.2376

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.391701   1.177049  -0.333    0.741
CENSADOS     0.734521   0.004656 157.742   <2e-16 ***
PARO         0.012850   0.122821   0.105    0.917
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.566 on 42 degrees of freedom
Multiple R-squared:  0.9983,    Adjusted R-squared:  0.9982
F-statistic: 1.244e+04 on 2 and 42 DF,  p-value: < 2.2e-16
```

Figura 4.2.7: summary(lm(PARO ~ VOTANTES + CENSADOS, data = practica42))

```
Call:
lm(formula = PARO ~ VOTANTES + CENSADOS, data = practica42)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2923 -0.9566 -0.0678  0.9793  4.4592

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.35963    0.32578  28.730   <2e-16 ***
VOTANTES     0.02028    0.19380   0.105    0.917
CENSADOS    -0.01554    0.14247  -0.109    0.914
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.967 on 42 degrees of freedom
Multiple R-squared:  0.000552,  Adjusted R-squared:  -0.04704
F-statistic: 0.0116 on 2 and 42 DF,  p-value: 0.9885
```

En las figuras anteriores se observa que dos regresiones auxiliares tienen coeficientes de determinación muy altos, superiores al 99%. Lo cual indica que dichas regresiones explican muy bien el comportamiento de las correspondientes variables dependientes (*censados* y *votantes*). Todo lo contrario ocurre con la regresión auxiliar cuya variable dependiente es *para*, pues su R^2 es casi 0. Además, en las regresiones de alto R_j^2, *paro* es siempre una variable explicativa no significativa, siendo relevante únicamente su pareja correspondiente. Todo ello confirma que la multicolinealidad fuerte existe y se limita a las variables *censados* y *votantes.*

Por otra parte, nótese también que, en el modelo original, el contraste de significatividad conjunta de las variables *votantes* y *censados* tiene un p-valor de 1.12%. Este contraste lo realizamos con el operador *linearHypothesis* del paquete *car*. El resultado se muestra en la siguiente figura:

Figura 4.2.8: linearHypothesis(mod.1,c("VOTANTES=0","CENSADOS=0"), test="F")

```
Linear hypothesis test

Hypothesis:
VOTANTES = 0
CENSADOS = 0

Model 1: restricted model
Model 2: BIPARTIDISMO ~ VOTANTES + CENSADOS + PARO

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     43 2069.1
2     41 1662.0  2    407.05 5.0208 0.01121 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

El bajo p-valor del contraste de significatividad conjunta de las variables *votantes* y *censados* sería señal de que ambas variables son conjuntamente significativas. En contradicción con lo que señalaban sus contrastes individuales. Esta aparente contradicción seguramente será de nuevo un efecto engañoso provocado por la multicolinealidad fuerte detectada.

4) ¿Qué se puede aplicar a esta situación con el fin de conseguir hacer desaparecer el efecto anómalo detectado en el apartado 3? Hágalo. Interprete los resultados, señalando los pros y contras del modelo alternativo que haya planteado. ¿Servirá ese modelo para analizar las cuestiones que motivaron este análisis?

La multicolinealidad fuerte y sus consecuencias indeseadas desaparecen si se elimina de la ecuación una de las variables explicativas involucradas en la relación multicolineal. Empleando como guía los valores de los estadísticos t de significatividad individual, se llega a la conclusión de que la mejor candidata a ser eliminada es *votantes* pues el valor absoluto de su estadístico t es menor que el de la variable *censados* y solo ligeramente superior a 1. Todo esto lleva a plantear el modelo simplificado:

$$bipartidismo = \beta_0 + \beta_2 censados + \beta_3 paro + \varepsilon$$

Los resultados de su estimación son los siguientes:

Figura 4.2.9: summary(lm(BIPARTIDISMO ~ CENSADOS + PARO, data = practica42))

```
Call:
lm(formula = BIPARTIDISMO ~ CENSADOS + PARO, data = practica42)

Residuals:
     Min       1Q   Median       3Q      Max
-11.3663  -3.8873  -0.5818   3.1142  15.6482

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 72.09826    4.78782  15.059  < 2e-16 ***
CENSADOS    -0.05677    0.01894  -2.997  0.00456 **
PARO        -0.45735    0.49959  -0.915  0.36519
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.371 on 42 degrees of freedom
Multiple R-squared:  0.1881,    Adjusted R-squared:  0.1494
F-statistic: 4.865 on 2 and 42 DF,  p-value: 0.01258
```

Con este procedimiento se consigue una reducción en la varianza de los estimadores que hace desaparecer el engañoso resultado de la no significatividad individual en los regresores multicolineales no suprimidos. Obsérvese que ahora el estadístico *t* para el contraste de significatividad individual de la variable *censados* tiene un p-valor igual a 0.00456. Este método tiene la desventaja de introducir sesgo en las estimaciones de los coeficientes de esos mismos regresores por implicar la eliminación de una variable relevante, aunque ese efecto negativo se palía, en la medida de lo posible, al haber seguido las indicaciones acerca de cómo elegir la variable explicativa a eliminar según los valores de los estadísticos *t*.

En todo caso hay que tener en cuenta que la estimación de β_2 en el modelo simplificado es sesgada y, de hecho, recoge una combinación de los efectos *ceteris paribus* reales de *votantes* y *censados*.

Además, al no aparecer la explicativa *votantes* en la ecuación, resulta imposible estudiar su efecto *ceteris paribus*, que era uno de los objetivos pretendidos inicialmente.

En la estimación del modelo simplificado se aprecia que la variable *paro* no parece ser relevante pues el p-valor asociado a su test de significatividad individual toma un valor elevado igual a 0.36519. De modo que resulta conveniente simplificar el modelo eliminando esa variable explicativa. El resultado de esta estimación se muestra en la Figura 4.2.10.

Figura 4.2.10: summary(lm(BIPARTIDISMO ~ CENSADOS, data = practica42))

```
Call:
lm(formula = BIPARTIDISMO ~ CENSADOS, data = practica42)

Residuals:
     Min       1Q   Median       3Q      Max
-11.3443  -4.2274  -0.3689   2.9572  14.6558

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.82019    1.03912  65.267  < 2e-16 ***
CENSADOS    -0.05647    0.01890  -2.987  0.00463 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.359 on 43 degrees of freedom
Multiple R-squared:  0.1719,    Adjusted R-squared:  0.1526
F-statistic: 8.925 on 1 and 43 DF,  p-value: 0.004633
```

5) Ante las carencias de la solución dada en el apartado 4 a las anomalías detectadas, un investigador sugiere otra solución alternativa. Propone estimar el modelo

$$bipartidismo = \beta_0^* + \beta_1^* votantesp + \beta_2^* censados + \beta_3^* paro + \varepsilon$$

donde *votantesp* mide ahora la participación electoral mediante el porcentaje de votantes sobre el total del censo. Indique ventajas de esta solución respecto a la del apartado anterior. Compruebe que los problemas detectados han desaparecido. Estime el modelo e interprete los resultados que obtenga.

Lo que el investigador propone es medir la participación en términos relativos mediante el porcentaje de votantes sobre el total del censo, en lugar de en términos absolutos mediante el número total de votantes. A este respecto, lo que se haría es equiparar la forma de medir la participación, a la que se emplea con el paro, variable que desde el principio ha estado medida en términos porcentuales y no absolutos.

La variable *votantesp* = 100**votantes*/*censados* se puede generar e incluir en el data.frame *practica42* con el siguiente código:

```
practica42$VOTANTESP <- 100*practica42$VOTANTES/practica42$CENSADOS
```

Ahora, el data.frame *practica42* consta de 6 variables. Ver Figura 4.2.11.

Figura 4.2.11: head(practica42)

	BIPARTIDISMO	CENSADOS	MUNICIPIO	PARO	VOTANTES	VOTANTESP
1	78.88920	4.646	Abanilla	8.639909	3.619	77.89496
2	69.83539	10.052	Abaran	5.534373	7.290	72.52288
3	67.38870	24.566	Aguilas	8.734758	16.666	67.84173
4	81.70213	1.151	Albudeite	13.073144	0.940	81.66811
5	58.21452	29.887	Alcantarilla	12.253785	21.462	71.81049
6	82.43045	0.807	Aledo	5.068598	0.683	84.63445

Entre los regresores *votantesp*, *censados* y *paro* no hay multicolinealidad fuerte, como se puede comprobar en la matriz de correlaciones de la Figura 4.2.12.

Figura 4.2.12: cor(practica42[,c(2,4,6)])

	CENSADOS	PARO	VOTANTESP
CENSADOS	1.00000000	-0.01707311	-0.1538562
PARO	-0.01707311	1.00000000	0.1063433
VOTANTESP	-0.15385623	0.10634325	1.0000000

También se debe comprobar que los resultados que indican ausencia de multicolinealidad se mantienen con las regresiones auxiliares de cada una de esas tres variables sobre las otras dos. Además de no sufrir del problema de la multicolinealidad, el siguiente modelo alternativo:

$$bipartidismo = \beta_0^* + \beta_1^* votantesp + \beta_2^* censados + \beta_3^* paro + \varepsilon$$

es útil para analizar el efecto *ceteris paribus* de la participación electoral y el tamaño de las poblaciones sobre el apoyo a los dos partidos tradicionalmente mayoritarios. Estas deben ser las razones que seguramente han llevado al investigador a proponer este modelo alternativo. Su estimación por MCO con R da los siguientes resultados:

Figura 4.2.13: summary(lm(BIPARTIDISMO ~ VOTANTESP + CENSADOS + PARO, data = practica42))

```
Call:
lm(formula = BIPARTIDISMO ~ VOTANTESP + CENSADOS + PARO, data = practica42)

Residuals:
    Min      1Q  Median      3Q     Max
-12.174  -3.016  -1.089   4.246  11.820

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.90454   12.98418   2.226  0.03156 *
VOTANTESP    0.60200    0.17101   3.520  0.00107 **
CENSADOS    -0.04761    0.01700  -2.801  0.00774 **
PARO        -0.62201    0.44556  -1.396  0.17023
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.65 on 41 degrees of freedom
Multiple R-squared:  0.3765,    Adjusted R-squared:  0.3309
F-statistic: 8.253 on 3 and 41 DF,  p-value: 0.0002059
```

Nótese, primero, que los resultados contradictorios entre sí observados en el modelo inicial (Figura 4.2.3) al examinar los test de significatividad, desaparecen. En el nuevo modelo alternativo los contrastes de significatividad individual indican que tanto *votantesp* como *censados* son variables relevantes. Además, se aprecia que el porcentaje de paro registrado no resulta significativo a la hora de explicar el bipartidismo (p-valor de su contraste de significatividad individual igual a 17.023%), por lo que *paro* es una variable explicativa irrelevante. Ello sugiere excluirla del modelo para mejorar las estimaciones. La estimación de este nuevo modelo se muestra en la Figura 4.2.14.

Figura 4.2.14: summary(lm(BIPARTIDISMO ~ VOTANTESP + CENSADOS, data = practica42))

```
Call:
lm(formula = BIPARTIDISMO ~ VOTANTESP + CENSADOS, data = practica42)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1106  -2.9980  -0.7198   4.2395  10.7401

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.94856   12.81349   1.947  0.05824 .
VOTANTESP    0.57694    0.17198   3.355  0.00169 **
CENSADOS    -0.04760    0.01719  -2.769  0.00834 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.714 on 42 degrees of freedom
Multiple R-squared:  0.3469,    Adjusted R-squared:  0.3158
F-statistic: 11.15 on 2 and 42 DF,  p-value: 0.0001302
```

La interpretación de los coeficientes de las dos variables explicativas restantes es clara: en el apoyo al bipartidismo, el tamaño de la población tiene un efecto negativo si permanece constante la participación medida en porcentaje de votantes sobre el censo. En otras palabras, según estos resultados, los pueblos pequeños son más proclives al bipartidismo. Pero ocurre al revés con el porcentaje de participación, que al aumentar, hace también aumentar el apoyo al bipartidismo si no varía el tamaño del censo de las localidades. Estas conclusiones discrepan en parte de las ideas preconcebidas de las que se partía.

Debemos tener en cuenta que la muestra usada se reduce a municipios de la Región de Murcia. Sería interesante comprobar qué ocurre si la muestra se centra en otras comunidades o se hace extensiva al conjunto del país.