Solución a la práctica 6 con R

En esta práctica vamos a utilizar los operadores head, layout, lm, matrix, plot, str y summary. Para utilizar el operador *bptest* debemos instalar y cargar el paquete *lmtest*: install.packages("lmtest") y library("lmtest"). Para utilizar el operador *coeftest* debemos instalar y cargar el paquete *sandwich*: install.packages("sandwich") y library("sandwich"), respectivamente.

El siguiente modelo de regresión relaciona la nota media que obtienen los alumnos en matemáticas (*nota*) en un centro, con el número de profesores disponibles en el centro (*profesores*), el porcentaje de repetidores (*repetidores*) sobre el total de alumnado del centro y el porcentaje de recursos con los que se financia el centro (*recursos*) que proceden de fondos públicos

$$nota_i = \beta_0 + \beta_1 profesores_i + \beta_2 repetidores_i + \beta_3 recursos_i + \varepsilon_i \,.$$

Usando las bases de datos de la prueba PISA[1] para centros educativos (PISA 2013) se han obtenido datos de una muestra de 225 centros escolares españoles que se encuentran en el fichero *practica6.RData* Sabiendo que la puntuación PISA es una puntuación normalizada tal que la media de los países de la OCDE en PISA se establece en 500 y la desviación típica en 100, responda a las siguientes cuestiones:

Antes de resolver esta práctica examinamos el tipo de datos que contiene el fichero *Practica6.RData*. Para ello ejecutamos el código *str(Practica6)* y obtenemos la salida de la Figura 6.1:

Figura 6.1: str(Practica6)

```
Classes 'tbl_df', 'tbl' and 'data.frame':       225 obs. of  4 variables:
 $ NOTA       : num  490 522 465 431 490 ...
 $ PROFESORES : num  67 67 15 60 61 16 59 60 81 65 ...
 $ RECURSOS   : num  85 83 75 98 100 70 100 100 95 97 ...
 $ REPETIDORES: num  5 2 15 8 13 9 27 14 8 39 ...
```

Observamos que el data.frame *Practica6* contiene 4 variables numéricas con 225 observaciones. También observamos las primeras observaciones de cada variable. Otro manera de obtener las primeras (seis) observaciones de las variables es con el código *head(practica6)*, como podemos ver en la Figura 6.2.

Figura 6.2: head(practica6)

	NOTA	PROFESORES	RECURSOS	REPETIDORES
1	489.63	67	85	5
2	522.47	67	83	2
3	464.96	15	75	15
4	431.49	60	98	8
5	489.95	61	100	13
6	455.45	16	70	9

1) Estime el modelo e interprete los coeficientes

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

[1] Datos disponibles en el Instituto Nacional de Evaluación Educativa
http://www.mecd.gob.es/inee/Bases-de-datos.html

mod <- lm(NOTA ~ PROFESORES + REPETIDORES + RECURSOS, data = practica6)

Figura 6.3: summary(mod)

```
Call:
lm(formula = NOTA ~ PROFESORES + REPETIDORES + RECURSOS, data = practica6)

Residuals:
     Min       1Q   Median       3Q      Max
-112.424  -14.670    0.366   18.713   69.435

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 511.8863     8.9801  57.003  < 2e-16 ***
PROFESORES    0.4029     0.0773   5.212 4.29e-07 ***
REPETIDORES  -1.7703     0.3008  -5.884 1.47e-08 ***
RECURSOS     -0.3229     0.1036  -3.116  0.00208 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.76 on 221 degrees of freedom
Multiple R-squared:  0.3182,    Adjusted R-squared:  0.3089
F-statistic: 34.38 on 3 and 221 DF,  p-value: < 2.2e-16
```

La nota media en matemáticas se relaciona positivamente con el número de profesores de los centros y negativamente con el porcentaje de repetidores y con el porcentaje de recursos públicos con los que se financia el centro. En concreto, un colegio con un profesor más que otro colegio con las mismas características obtendría aproximadamente 0.4 puntos más de nota. Además, un punto adicional en el porcentaje de repetidores o en el porcentaje que representan los recursos públicos disminuye la nota media en aproximadamente 1.8 y 0.3 puntos porcentuales, respectivamente.

2) Realice los diagramas de dispersión entre el cuadrado de los residuos y cada una de las variables explicativas, y entre el cuadrado de los residuos y la variable dependiente estimada por el modelo. Interprete los resultados.

El cuadrado de los residuos constituye una aproximación simple (y la única posible con la información disponible) a la varianza de las perturbaciones. Por tanto, si hay homoscedasticidad, se espera que el comportamiento general del cuadrado de los residuos sea el mismo a lo largo de todo el recorrido de valores de la variable con la que se compara, sea esta una de las explicativas del modelo o las predicciones de la explicada.

Para conocer los componentes del objeto *mod* es recomendable utilizar el código: *names(mod)*. De este modo podemos saber que el componente *residuals* contiene al vector de residuos del modelo y que *fitted.values* contiene el valor ajustado de la variable dependiente. Por tanto, el vector del cuadrado de los residuos de la regresión se obtiene con el código: *mod$residuals^2*. Para facilitar la programación cuando utilicemos este vector es aconsejable guardarlo en un objeto. En este caso, guardamos el vector del cuadrado de los residuos de la regresión en el objeto *e2*. Para ello, ejecutamos el siguiente código:

e2 <- mod$residuals^2

La variable dependiente ajustada o estimada por el modelo, $\hat{y}$, se obtiene con el código: *mod$fitted.values*. Y la guardamos en el objeto *notae* con el código siguiente:

notae <- mod$fitted.values

Para obtiener los diagramas de dispersión ejecutamos los siguientes códigos:

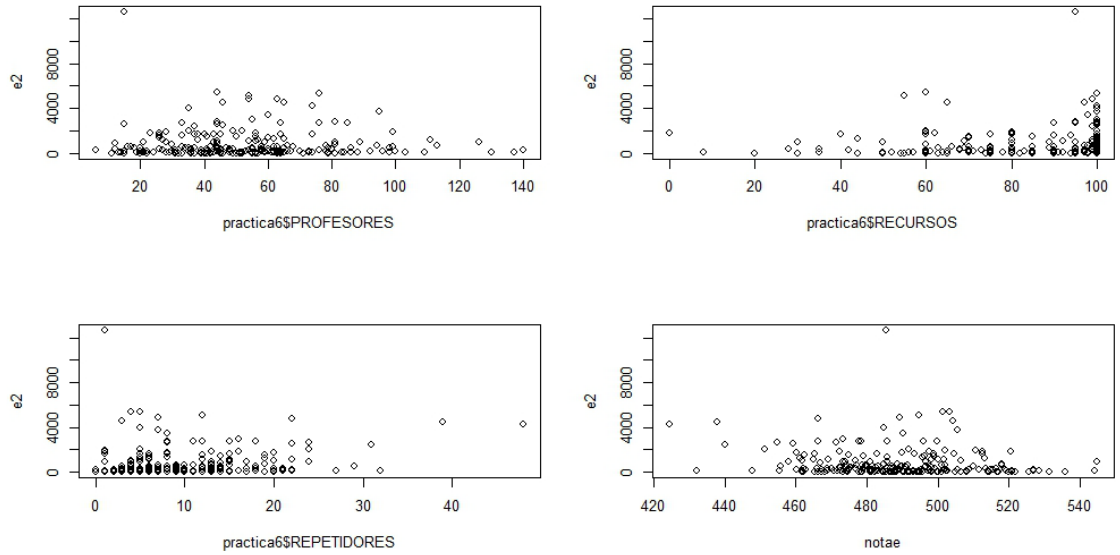
```
plot(practica6$PROFESORES, e2)
plot(practica6$RECURSOS, e2)
plot(practica6$REPETIDORES, e2)
plot(notae, e2)
```

Si queremos que los cuatro diagramas aparezcan en una misma figura ejecutamos los siguientes códigos:

```
M <- matrix(c(1,2,3,4),nrow = 2, byrow = T)
layout(M)
plot(practica6$PROFESORES, e2)
plot(practica6$RECURSOS, e2)
plot(practica6$REPETIDORES, e2)
plot(notae, e2)
layout(1)
```

Con las dos primeras líneas del código estamos diciendo a R que coloque los gráficos en una matriz con cuatro entradas. Como esta instrucción es permanente, los gráficos que se hagan en adelante se colocaran en una matriz con cuatro entradas. De modo que para volver al modo inicial en el que solo aparece un gráfico escribimos la última línea.

Figura 6.4: Gráficos de dispersión



En los diagramas de dispersión se intuye que podría existir una cierta relación entre el cuadrado de los residuos, que es una aproximación a la varianza de los errores, y la variable explicativa recursos (en menor medida con los profesores). Sin embargo, no se encuentra esa relación con la estimación de las notas, que recoge el efecto de las tres variables explicativas al mismo tiempo. Por tanto, podría existir un problema de heteroscedasticidad en este modelo. En cualquier caso, la decisión sobre la posible presencia de heteroscedasticidad tiene que apoyarse en el resultado de un contraste.

3) Utilice el contraste de White para detectar la presencia de heteroscedasticidad.

$$\left.\begin{array}{ll} H_0 & \text{homoscedasticidad} \\ H_1 & \text{heteroscedasticidad} \end{array}\right\}$$

Este contraste no precisa conocer la estructura de la heteroscedasticidad; ni siquiera requiere imponer qué variable(s) explicativa(s) genera(n) los problemas de heteroscedasticidad. Se aplica a través de los siguientes pasos.

Primero, se parte de una estimación de la varianza de los errores del modelo. Dicha estimación se hace con el cuadrado de los residuos, el cual hemos guardado en el objeto *e2*. A continuación, se plantea una regresión auxiliar que explique dicha varianza regresando *e2* sobre una constante, las explicativas del modelo inicial, sus cuadrados, y sus productos cruzados. Hay que estar atento para no se repetir ninguna explicativa en esta regresión auxiliar. De este modo, la aproximación que se hace de la varianza es muy flexible, por lo que debería ser capaz de capturar la posible presencia de heteroscedasticidad. El código para estimar la regresión auxiliar es el siguiente:

```
mod2 <- lm(e2 ~ PROFESORES + REPETIDORES + RECURSOS +
I(PROFESORES^2) + I(REPETIDORES^2) + I(RECURSOS^2) +
PROFESORES*REPETIDORES + PROFESORES*RECURSOS +
   REPETIDORES*RECURSOS, data = practica6)
```

Figura 6.5: summary(mod2)

```
Call:
lm(formula = e2 ~ PROFESORES + REPETIDORES + RECURSOS + I(PROFESORES^2) +
    I(REPETIDORES^2) + I(RECURSOS^2) + PROFESORES * REPETIDORES +
    PROFESORES * RECURSOS + REPETIDORES * RECURSOS, data = practica6)

Residuals:
    Min      1Q  Median      3Q     Max
-1505.9  -658.6  -400.0   199.2 10926.9

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)           1283.15070 1160.32042   1.106  0.27002
PROFESORES              -7.27817   21.37135  -0.341  0.73377
REPETIDORES            -83.36071  102.30023  -0.815  0.41605
RECURSOS                -9.11531   23.62340  -0.386  0.69998
I(PROFESORES^2)          0.01212    0.10182   0.119  0.90535
I(REPETIDORES^2)         3.26355    1.08054   3.020  0.00283 **
I(RECURSOS^2)            0.18065    0.19351   0.934  0.35157
PROFESORES:REPETIDORES   1.11134    0.66438   1.673  0.09583 .
PROFESORES:RECURSOS     -0.06885    0.21979  -0.313  0.75438
REPETIDORES:RECURSOS    -0.71036    1.17096  -0.607  0.54472
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1332 on 215 degrees of freedom
Multiple R-squared:  0.09607,   Adjusted R-squared:  0.05823
F-statistic: 2.539 on 9 and 215 DF,  p-value: 0.008677
```

Se obtiene que el R^2 de esta regresión vale $R_e^2 = 0.09607$.

El estadístico de este contraste es NR_e^2, que bajo la hipótesis nula sigue una χ_q^2, donde *q* es el número de explicativas que usamos en la regresión auxiliar. El valor crítico con el que hay que comparar el estadístico es $\chi_{q,\alpha}^2$. Por tanto, se rechaza la hipótesis nula de homoscedasticidad si $NR_e^2 > \chi_{q,\alpha}^2$.

En este caso, tenemos que $NR_e^2 = 225 \cdot 0.0961 = 21.61$ y q=9, por lo que se rechaza la hipótesis nula de homoscedasticidad si $21.61 > \chi_{9,\alpha}^2$. Para un tamaño del test del 5% tenemos que $\chi_{9,0.05}^2 = 16.92$, *qchisq(0.95,9)*, y para el 1% tenemos $\chi_{9,0.01}^2 = 21.67$, *qchisq(0.99,9)*. Por lo tanto, rechazamos la hipótesis nula para un tamaño del test del 5%, pero no para el 1%.

El contraste de White se puede hacer de *forma directa* con el operador *bptest*. Para utilizar este operador debemos instalar y cargar el paquete *lmtest*.[2] Una vez hecho esto, el código para hacer el contraste de White en esta práctica es:

*bptest(mod, ~ PROFESORES + REPETIDORES + RECURSOS + I(PROFESORES^2) + I(REPETIDORES^2) + I(RECURSOS^2) + PROFESORES*REPETIDORES + PROFESORES*RECURSOS + REPETIDORES*RECURSOS, data = practica6)*

Y el resultado es el que aparece en la siguiente figura:

Figura 6.6: Contraste de White

```
        studentized Breusch-Pagan test

data:  mod
BP = 21.615, df = 9, p-value = 0.01018
```

Como vemos, el valor del estadístico (21.615) coincide con el que hemos obtenido previamente. Además, R presenta el *p*-valor del contraste, que en este caso es 0.01018. Esto implica que rechazamos H_0 y aceptamos la existencia de heteroscedasticidad para un tamaño del test mayor que 0.01018, como ocurre si $\alpha = 0.05$.

Si ejecutamos el código *bptest(mod)*, R llevaría a cabo el test de heteroscedasticidad de Breusch-Pagan.

4) Lleve a cabo la estimación del modelo siguiendo el método propuesto por White. ¿Qué ventaja tiene aplicar este método?

Como se ha detectado la presencia de heteroscedasticidad en los datos, se puede estimar el modelo usando el "método de White". La ventaja de este método frente a MCO recae en que, aunque usa como estimación de los parámetros la que se obtienen por MCO (que no pierde su propiedad de consistencia ni de insesgadez en presencia de heteroscedasticidad), propone una estimación alternativa de varianzas de los estimadores de los coeficientes que permite hacer inferencia en caso de que exista heteroscedasticidad. En efecto, el estimador MCO habitual de estas varianzas es

[2] install.packages("lmtest") y library("lmtest"), respectivamente.

sesgado e inconsistente en presencia de heteroscedasticidad, mientras que la corrección de White nos proporciona un estimador consistente. Por tanto, la estimación de los coeficientes propuesta se obtiene de la estimación MCO que aparece en el primer apartado. Pero las varianzas se deben obtener mediante otra expresión distinta a la de MCO.

Para estimar con R las desviaciones típicas de los estimadores de los parámetros del modelo según el método propuesto por White necesitamos instalar y cargar el paquete *sandwich*.[3] Este paquete nos permite utilizar el operador *coeftest*. El resultado obtenido aparece en la Figura 6.7.

Figura 6.7: coeftest(mod, vcov. = vcovHC, type="HC1")

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 511.886261   8.386070 61.0401 < 2.2e-16 ***
PROFESORES    0.402887   0.072959  5.5221 9.363e-08 ***
REPETIDORES  -1.770264   0.444156 -3.9857 9.141e-05 ***
RECURSOS     -0.322902   0.105347 -3.0651  0.002447 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Se observa que las estimaciones de los parámetros son las mismas que las que se obtuvieron aplicando el método MCO al modelo original, Figura 6.3, pero que las estimaciones de las desviaciones típicas de los estimadores de los coeficientes han cambiado. El modelo estimado según el enfoque de White quedaría del siguiente modo:

$$\hat{notas}_i = \underset{(8.38)}{511.88} + \underset{(0.07)}{0.40}\, profesores_i - \underset{(0.44)}{1.77}\, repetidores_i - \underset{(0.11)}{0.32}\, recursos$$

Usando estos resultados podemos concluir que todas las variables son altamente significativas.

[3] install.packages("sandwich") y library("sandwich"), respectivamente.