Solución a la práctica 5.1 con R

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

Usando la base de datos de la prueba PISA (Programme for International Student Assessment) 2012 para Murcia, se han obtenido datos para 1374 estudiantes (edad media 15 años) sobre las siguientes variables:

NOTA DE LOS ALUMNOS (Unidades de medida: puntos. La puntuación PISA es una puntuación normalizada tal que la media de los países de la OCDE en PISA se establece por definición en 500, y la desviación típica en 100):

notamat = nota en matemáticas
notalec = nota en lectura
notacien = nota en ciencias

NIVEL EDUCATIVO DE LOS PADRES (variable ficticia):

educmadre = vale 1 si la madre tiene estudios superiores y 0 si no.
educpadre = vale 1 si el padre tiene estudios superiores y 0 si no.

VARIABLE DE GÉNERO (variable ficticia):
varon = vale 1 si el estudiante es varón y 0 si no.

VARIABLES DE ESTATUS:

convivienda = es un índice que refleja el entorno favorable al estudio en la casa del estudiante (habitación propia, ordenador propio, etc.) así como las condiciones generales de la vivienda.

estatus = índice PISA de estatus económico, social y cultural de la familia. Mide la riqueza familiar, las posesiones culturales de la familia y los recursos educativos disponibles en el hogar, entre otros.

Antes de resolver esta práctica examinamos el tipo de datos que contiene el fichero *Practica51.RData*. Para ello ejecutamos el código *str(Practica51)* y obtenemos la salida de la Figura 5.1.1:

Figura 5.1.1: str(practica51)

```
Classes 'tbl_df', 'tbl' and 'data.frame':       1374 obs. of  9 variables:
 $ CONVIVIENDA: num  -1.14 0.62 -0.5 0.62 -1.14 0.01 -1.14 0.2 0.62 -0.66 ...
 $ EDUCMADRE  : num  0 0 0 1 0 0 0 0 0 0 ...
 $ EDUCPADRE  : num  0 0 0 0 0 1 0 0 0 0 ...
 $ ESTATUS    : num  -1.04 -0.77 -1.42 1.43 -1.76 -0.35 -2.33 -1.17 -0.14 -0.41 ...
 $ MUJER      : num  1 1 0 0 1 0 0 0 1 1 ...
 $ NOTACIEN   : num  368 492 545 518 535 ...
 $ NOTALEC    : num  413 568 597 566 536 ...
 $ NOTAMAT    : num  392 434 541 534 578 ...
 $ VARON      : num  0 0 1 1 0 1 1 1 0 0 ...
```

Notar que todas las variables son numéricas. La Figura 5.1.1 también muestra las primeras observaciones de las variables. Otro manera de obtener las primeras (seis) observaciones es con el código *head(practica51)*, como podemos ver en la Figura 5.1.2.

Figura 5.1.2: head(practica51)

	CONVIVIENDA	EDUCMADRE	EDUCPADRE	ESTATUS	MUJER	NOTACIEN	NOTALEC	NOTAMAT	VARON
1	-1.14	0	0	-1.04	1	368.3700	413.2041	392.0470	0
2	0.62	0	0	-0.77	1	492.1112	568.3328	434.2655	0
3	-0.50	0	0	-1.42	0	545.2630	597.3506	541.2915	1
4	0.62	1	0	1.43	0	518.2208	566.0748	534.2810	1
5	-1.14	0	0	-1.76	1	535.2853	535.5278	578.2131	0
6	0.01	0	1	-0.35	0	495.9344	448.9908	506.0056	1

Responda a las siguientes cuestiones, utilizando los datos del fichero *Practica51.RData*:

1) **Para analizar el efecto del género sobre la nota en matemáticas, se proponen tres modelos alternativos diferentes:**

   a) $notamat = \beta_0 + \beta_1 varon + \beta_2 convivienda + \beta_3 estatus + \varepsilon$

   b) $notamat = \beta_0 + \beta_1 mujer + \beta_2 convivienda + \beta_3 estatus + \varepsilon$

   c) $notamat = \beta_0 + \beta_1 varon + \beta_2 mujer + \beta_3 convivienda + \beta_4 estatus + \varepsilon$

Estima cada uno de estos modelos, comenta los resultados e interpreta los coeficientes de las variables ficticias.

Primero, estimamos por MCO el modelo a) y guardamos el resultado en el objeto *modeloA*. Para ello ejecutamos el siguiente código:

modeloA <- lm(NOTAMAT~VARON+CONVIVIENDA+ESTATUS, data = practica51)

Figura 5.1.3: summary(modeloA)

```
Call:
lm(formula = NOTAMAT ~ VARON + CONVIVIENDA + ESTATUS, data = practica51)

Residuals:
    Min      1Q  Median      3Q     Max
-264.17  -52.63    2.88   53.95  283.03

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  471.918      3.374 139.862  < 2e-16 ***
VARON         13.424      4.404   3.048  0.00235 **
CONVIVIENDA   11.624      3.397   3.422  0.00064 ***
ESTATUS       24.466      2.786   8.781  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 81.39 on 1363 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.1403,    Adjusted R-squared:  0.1384
F-statistic: 74.12 on 3 and 1363 DF,  p-value: < 2.2e-16
```

Como podemos observar, las variables de estatus (*convivienda* y *estatus*) son altamente significativas, lo cual indica la importancia del nivel económico, social y cultural de la familia sobre la nota de los estudiantes. Por otro lado, vemos como la variable de género (*varon*) también es significativa. Como en este modelo la categoría de referencia para el género es la mujer, el resultado de la estimación refleja que, para dos estudiantes con el mismo nivel de estatus, en promedio el chico obtiene 13.424 puntos más que la chica.

En el caso de haber incluido la variable *mujer* para representar el género en vez de la variable *hombre* obtendríamos el modelo b). La estimación de este modelo se realiza con el siguiente código, el cual también nos permite guardar el resultado de la misma.

modeloB <- lm(NOTAMAT~MUJER+CONVIVIENDA+ESTATUS, data = practica51)

Figura 5.1.4: summary(modeloB)

```
Call:
lm(formula = NOTAMAT ~ MUJER + CONVIVIENDA + ESTATUS, data = practica51)

Residuals:
    Min      1Q  Median      3Q     Max
-264.17  -52.63    2.88   53.95  283.03

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  485.342      3.362 144.351  < 2e-16 ***
MUJER        -13.424      4.404  -3.048  0.00235 **
CONVIVIENDA   11.624      3.397   3.422  0.00064 ***
ESTATUS       24.466      2.786   8.781  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 81.39 on 1363 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.1403,    Adjusted R-squared:  0.1384
F-statistic: 74.12 on 3 and 1363 DF,  p-value: < 2.2e-16
```

Observamos como, en esta estimación, el resultado es exactamente el mismo que en la anterior, salvo por el signo del coeficiente de la variable ficticia *mujer*. Ahora, la categoría de referencia son los varones, por lo tanto el coeficiente de la ficticia refleja la diferencia ceteris paribus de las chicas respecto de los chicos. Como el signo es negativo, esto implica que en promedio las chicas obtienen 13.42443 puntos menos que los chicos en el examen de matemáticas.

Por último, consideramos el modelo c), el cual incluye el término constante y las dos variables ficticias de género. Este modelo tratamos de estimarlo con el siguiente código.

summary(lm(NOTAMAT~VARON+MUJER+CONVIVIENDA+ESTATUS, data = practica51))

El resultado que obtenemos es la estimación del modelo a) debido a que el modelo tiene un problema de multicolinealidad exacta o perfecta entre las variables *mujer*, *varon* y el término constante, ya que $varon + mujer = 1$ para todos los estudiantes de la muestra. Es decir, sea caído en la trampa de las ficticias. Para solventarlo R ha eliminado automáticamente la ficticia *mujer*, dado que la estimación del coeficiente que la multiplica no aparece (en su lugar aparece NA), como se puede ver en la Figura 5.1.4.

Figura 5.1.4: summary(lm(NOTAMAT~VARON+MUJER+CONVIVIENDA+ESTATUS, data = practica51))

```
Call:
lm(formula = NOTAMAT ~ VARON + MUJER + CONVIVIENDA + ESTATUS,
    data = practica51)

Residuals:
    Min      1Q  Median      3Q     Max
-264.17  -52.63    2.88   53.95  283.03

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  471.918      3.374 139.862  < 2e-16 ***
VARON         13.424      4.404   3.048  0.00235 **
MUJER             NA         NA      NA       NA
CONVIVIENDA   11.624      3.397   3.422  0.00064 ***
ESTATUS       24.466      2.786   8.781  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81.39 on 1363 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.1403,    Adjusted R-squared:  0.1384
F-statistic: 74.12 on 3 and 1363 DF,  p-value: < 2.2e-16
```

Si se quieren introducir las dos ficticias y evitar caer en la trampa de las ficticias, la solución es eliminar la constante del modelo. La estimación de este modelo, que guardamos en el objeto *modeloC*, aparece en la Figura 5.1.5 y se obtiene con el código:[1]

modeloC <- lm(NOTAMAT ~ -1+VARON+MUJER+CONVIVIENDA+ESTATUS, data = practica51)

Figura 5.1.5: summary(modeloC)

```
Call:
lm(formula = NOTAMAT ~ -1 + VARON + MUJER + CONVIVIENDA + ESTATUS,
    data = practica51)

Residuals:
    Min      1Q  Median      3Q     Max
-264.17  -52.63    2.88   53.95  283.03

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
VARON        485.342      3.362 144.351  < 2e-16 ***
MUJER        471.918      3.374 139.862  < 2e-16 ***
CONVIVIENDA   11.624      3.397   3.422  0.00064 ***
ESTATUS       24.466      2.786   8.781  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81.39 on 1363 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.9708,    Adjusted R-squared:  0.9707
F-statistic: 1.133e+04 on 4 and 1363 DF,  p-value: < 2.2e-16
```

Como no hemos dejado ninguna categoría de referencia, los coeficientes de las ficticias reflejan directamente el valor del término constante del modelo para cada una de las categorías. Así, el término constante para el modelo de los varones es 485.342 mientras que el del modelo de las mujeres es de 471.918. La diferencia entre ambos términos constantes es 13.424 (= 485.342 – 471.918), que coincide con el coeficiente de la

[1] Para eliminar el término constante hemos escrito *-1* en el lugar en el que se indican las variables explicativas.

variable *varon* en el modelo a) y, en valor absoluto, con el coeficiente de la variable *mujer* en el modelo b).

Una vez analizado este sencillo ejemplo, vamos a completar el análisis de la nota en los distintos exámenes con todas las variables de que disponemos, siguiendo un proceso ordenado para seleccionar el modelo que mejor se ajusta a los datos. Comenzaremos por la nota en ciencias, después analizaremos la nota en matemáticas y terminaremos con el análisis de la nota en lectura.

2) Para examinar los determinantes de la nota en ciencias, se propone el siguiente modelo de regresión:

$$notacien = \beta_0 + \beta_1 varon + \beta_2 convivienda + \beta_3 estatus + \beta_4 educmadre + \beta_5 educpadre + \\ + \beta_6 educmadre * estatus + \beta_7 educpadre * estatus + \varepsilon$$

a) Interprete los coeficientes de las variables ficticias.

$\beta_0 \equiv$ término constante del modelo para la categoría de referencia. En este caso, chicas con padres sin estudios superiores.

$\beta_1 \equiv$ diferencia de la nota en ciencias de los chicos respecto de las chicas, independiente del valor de las variables explicativas del modelo.

$\beta_4 \equiv$ diferencia de la nota en ciencias de los chicos/as cuya madre tiene estudios superiores respecto de aquellos cuya madre no tiene estudios superiores, independiente del valor de las variables explicativas del modelo.

$\beta_5 \equiv$ diferencia de la nota en ciencias de los chicos/as cuyo padre tiene estudios superiores respecto de aquellos cuyo padre no tiene estudios superiores, independiente del valor de las variables explicativas del modelo.

$\beta_6 \equiv$ diferencia de la nota en ciencias de los chicos/as cuya madre tiene estudios superiores respecto de aquellos cuya madre no tiene estudios superiores, que depende del estatus familiar.

$\beta_7 \equiv$ diferencia de la nota en ciencias de los chicos/as cuyo padre tiene estudios superiores respecto de aquellos cuyo padre no tiene estudios superiores, que depende del estatus familiar.

b) Estime el modelo por MCO.

Estimamos el modelo y guardamos el resultado en el objeto *modelo2* con el siguiente código:

```
modelo2 <- lm(NOTACIEN ~ VARON + CONVIVIENDA + ESTATUS +
    EDUCMADRE + EDUCPADRE + EDUCMADRE*ESTATUS +
    EDUCPADRE*ESTATUS, data = practica51)
```

El resultado de la estimación se muestra en la Figura 5.1.5.

Figura 5.1.5: summary(modelo2)

```
Call:
lm(formula = NOTACIEN ~ VARON + CONVIVIENDA + ESTATUS + EDUCMADRE +
    EDUCPADRE + EDUCMADRE * ESTATUS + EDUCPADRE * ESTATUS, data = practica51)

Residuals:
    Min      1Q  Median      3Q     Max
-453.69  -51.78    4.21   51.37  250.88

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        484.665      4.978  97.370  < 2e-16 ***
VARON                1.577      4.415   0.357 0.720992
CONVIVIENDA          9.307      3.553   2.620 0.008902 **
ESTATUS             14.912      4.075   3.659 0.000263 ***
EDUCMADRE           15.060      6.842   2.201 0.027892 *
EDUCPADRE            1.046      6.287   0.166 0.867923
ESTATUS:EDUCMADRE    3.801      8.000   0.475 0.634760
ESTATUS:EDUCPADRE   13.592      7.502   1.812 0.070259 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 79.75 on 1307 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.1274,    Adjusted R-squared:  0.1228
F-statistic: 27.27 on 7 and 1307 DF,  p-value: < 2.2e-16
```

b.1) Contraste la relevancia de la variable de género sobre la nota en ciencias. ¿Qué conclusiones obtiene? ¿Cómo se interpreta el resultado? Modifique el modelo de acuerdo con los resultados de su contraste.

Para contrastar la relevancia de la variable *varon* sobre la nota en ciencias miramos directamente el p-valor de su coeficiente en la Figura 5.1.5. Su valor es 0.720992, mayor que 0.10, por lo que concluimos que la variable *varon* no es relevante incluso a niveles altos de significación. Como consecuencia, podemos decir que el género no influye en la nota en ciencias, es decir, que no hay una diferencia significativa en la nota de ciencias entre los chicos y las chicas.

Dado que la variable *varon* no es relevante debemos eliminarla del modelo. El modelo resultante es el siguiente:

$$notacien = \beta_0 + \beta_1 convivienda + \beta_2 estatus + \beta_3 educmadre + \beta_4 educpadre + \\ + \beta_5 educmadre * estatus + \beta_6 educpadre * estatus + \varepsilon$$

Estimamos el modelo y guardamos el resultado en el objeto *modelo21* con el siguiente código:

```
modelo21 <- lm(NOTACIEN ~ CONVIVIENDA + ESTATUS + EDUCMADRE
    + EDUCPADRE+EDUCMADRE*ESTATUS+EDUCPADRE*ESTATUS,
    data = practica51)
```

El resultado de la estimación se muestra en la Figura 5.1.6.

Figura 5.1.6: summary(modelo21)

```
Call:
lm(formula = NOTACIEN ~ CONVIVIENDA + ESTATUS + EDUCMADRE + EDUCPADRE +
    EDUCMADRE * ESTATUS + EDUCPADRE * ESTATUS, data = practica51)

Residuals:
    Min      1Q  Median      3Q     Max
-452.89  -52.15    3.43   51.18  251.73

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        485.468      4.438 109.384  < 2e-16 ***
CONVIVIENDA          9.287      3.551   2.615  0.00902 **
ESTATUS             14.955      4.072   3.673  0.00025 ***
EDUCMADRE           14.973      6.835   2.191  0.02865 *
EDUCPADRE            1.177      6.274   0.188  0.85120
ESTATUS:EDUCMADRE    3.828      7.997   0.479  0.63230
ESTATUS:EDUCPADRE   13.438      7.487   1.795  0.07292 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 79.73 on 1308 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.1273,    Adjusted R-squared:  0.1233
F-statistic: 31.81 on 6 and 1308 DF,  p-value: < 2.2e-16
```

b.2) Contraste la significatividad del nivel educativo de la madre sobre la nota en ciencias en el modelo propuesto en el apartado anterior. Comente sus conclusiones y modifique el modelo si es necesario.

Para contrastar la significatividad del nivel educativo de la madre sobre la nota en ciencias hay que contrastar la significatividad conjunta de las variables ficticias aditiva y multiplicativa asociadas al nivel educativo de la madre. En concreto, el contraste es: H_0: $\beta_3 = \beta_5 = 0$, H_1: no H_0. Para hacerlo con R utilizamos el operador *linearHypothesis* y obtenemos los resultados de la Figura 5.1.7.[2]

Figura 5.1.7: linearHypothesis(modelo21,c("EDUCMADRE=0","ESTATUS:EDUCMADRE=0"),test = "F")

```
Linear hypothesis test

Hypothesis:
EDUCMADRE = 0
ESTATUS:EDUCMADRE = 0

Model 1: restricted model
Model 2: NOTACIEN ~ CONVIVIENDA + ESTATUS + EDUCMADRE + EDUCPADRE + EDUCMADRE *
    ESTATUS + EDUCPADRE * ESTATUS

  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1   1310 8357100
2   1308 8314202  2     42898 3.3744 0.03454 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El estadístico F para contrastar la significatividad conjunta de los parámetros β_3 y β_5 tiene un valor de 3.37 con un p-valor asociado de 0.03454. Por lo que a un tamaño del 5% rechazamos la hipótesis nula de no significación conjunta de los parámetros, pero no al 1%. Es decir, existe evidencia fuerte de que el nivel educativo de la madre es relevante para explicar la nota en ciencias. De modo

[2] Antes, debemos cargar el paquete *car* con el código *library("car")*, si lo hemos instalado. En caso contrario, debemos instalarlo con el código *install.packages("car")*.

que al menos una de las variables ficticias asociadas al nivel de educación de la madre es significativa individualmente, aunque este contraste no nos permite deducir cual. Para ello realizamos los contrastes de significatividad individual de los coeficientes β_3 y β_5, y obtenemos que los p-valores de estos contrastes son, respectivamente, 0.0286 y 0.6323, como podemos ver en la Figura 5.1.6. Por lo tanto, concluimos que mientras la variable ficticia aditiva *educmadre* es relevante, la variable ficticia multiplicativa *educmadre•estatus* no lo es. En consecuencia, eliminamos esta variable del modelo anterior y obtenemos el siguiente:

$$notacien = \beta_0 + \beta_1 convivienda + \beta_2 estatus + \beta_3 educmadre + \beta_4 educpadre + \\ + \beta_5 educpadre \cdot estatus + \varepsilon$$

Este modelo lo estimamos y el resultado, que aparece en la Figura 5.1.8, lo guardamos en el objeto *modelo22* con el siguiente código:

*modelo22 <- lm(NOTACIEN ~ CONVIVIENDA + ESTATUS + EDUCMADRE + EDUCPADRE+EDUCPADRE*ESTATUS, data = practica51)*

Figura 5.1.8: summary(modelo22)

```
Call:
lm(formula = NOTACIEN ~ CONVIVIENDA + ESTATUS + EDUCMADRE + EDUCPADRE +
    EDUCPADRE * ESTATUS, data = practica51)

Residuals:
    Min      1Q  Median      3Q     Max
-452.94  -51.96    2.98   51.65  252.03

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       485.7812     4.3886 110.692  < 2e-16 ***
CONVIVIENDA         9.3210     3.5495   2.626 0.008739 **
ESTATUS            15.3063     4.0040   3.823 0.000138 ***
EDUCMADRE          16.1932     6.3400   2.554 0.010758 *
EDUCPADRE           0.5972     6.1541   0.097 0.922713
ESTATUS:EDUCPADRE  15.2943     6.4018   2.389 0.017032 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79.7 on 1309 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.1272,    Adjusted R-squared:  0.1239
F-statistic: 38.15 on 5 and 1309 DF,  p-value: < 2.2e-16
```

Concluimos por tanto que el hecho de que la madre tenga educación superior influye positivamente en la nota de los alumnos con un impacto constante sea cual sea el valor de las explicativas. En concreto, la diferencia en la nota de ciencias es de 16.1932 puntos entre los estudiantes cuya madre tiene educación superior respecto a aquellos cuya madre no la tiene, ceteris paribus.

b.3) Haga lo mismo que en el apartado anterior para el nivel educativo del padre. ¿Influye la educación del padre en el efecto que tiene la variable *estatus* en la nota en ciencias?

De modo similar al subapartado anterior, para contrastar la significatividad del nivel educativo del padre sobre la nota en ciencias hay que contrastar la significatividad conjunta de las variables ficticias aditiva y multiplicativa

asociadas al nivel educativo del padre. En concreto, el contraste es: H_0: $\beta_4=\beta_5=0$, H_1: no H_0. Para hacerlo utilizamos el operador *linearHypothesis* y obtenemos los resultados de la Figura 5.1.9.

Figura 5.1.9: linearHypothesis(modelo22,c("EDUCPADRE=0","ESTATUS:EDUCPADRE=0"),test = "F")

El resultado se presenta en la siguiente tabla:

```
Linear hypothesis test

Hypothesis:
EDUCPADRE = 0
ESTATUS:EDUCPADRE = 0

Model 1: restricted model
Model 2: NOTACIEN ~ CONVIVIENDA + ESTATUS + EDUCMADRE + EDUCPADRE + EDUCPADRE *
    ESTATUS

  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1   1311 8354500
2   1309 8315658  2     38842 3.0571 0.04736 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

El estadístico F para contrastar la significatividad conjunta de los parámetros β_4 y β_5 tiene un valor de 3.05 con un p-valor asociado de 0.0474. Por lo que a un tamaño del 5% rechazamos la hipótesis nula de no significación conjunta de los parámetros, pero al 1% no la rechazamos. Es decir, existe evidencia fuerte de que el nivel educativo del padre es relevante para explicar la nota en ciencias. De modo que al menos una de las variables ficticias asociadas al nivel de educación del padre es significativa individualmente, aunque este contraste no nos permite deducir cual. Para ello realizamos los contrastes de significatividad individual de los coeficientes β_4 y β_5, y obtenemos que los p-valores de estos contrastes son, respectivamente, 0.9227 y 0.0170, como podemos ver en la Figura 5.1.8. Por lo tanto, concluimos que mientras la variable ficticia aditiva *educpadre* no es relevante, la variable ficticia multiplicativa *educpadre•estatus* si lo es. En consecuencia, eliminamos *educpadre* del modelo anterior y obtenemos el siguiente:

$$notacien = \beta_0 + \beta_1 convivienda + \beta_2 estatus + \beta_3 educmadre + \beta_4 educpadre \cdot estatus + \varepsilon$$

Este modelo lo estimamos y el resultado, que aparece en la Figura 5.1.10, lo guardamos en el objeto *modelo23* con el siguiente código:

```
modelo23 <- lm(NOTACIEN ~ CONVIVIENDA + ESTATUS + EDUCMADRE
        + I(EDUCPADRE*ESTATUS), data = practica51)
```

Figura 5.1.10: summary(modelo23)

```
Call:
lm(formula = NOTACIEN ~ CONVIVIENDA + ESTATUS + EDUCMADRE + I(EDUCPADRE *
    ESTATUS), data = practica51)

Residuals:
    Min      1Q  Median      3Q     Max
-453.05  -51.96    3.47   51.64  252.07

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            485.983      3.861 125.857  < 2e-16 ***
CONVIVIENDA              9.263      3.498   2.648  0.00818 **
ESTATUS                 15.441      3.756   4.111 4.18e-05 ***
EDUCMADRE               16.234      6.324   2.567  0.01037 *
I(EDUCPADRE * ESTATUS)  15.431      6.243   2.472  0.01357 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79.67 on 1310 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.1272,    Adjusted R-squared:  0.1245
F-statistic: 47.72 on 4 and 1310 DF,  p-value: < 2.2e-16
```

Concluimos por tanto que el hecho de que el padre tenga estudios superiores influye positivamente en la nota en ciencias de los estudiantes, alterando la manera en que el estatus familiar influye sobre esta nota. En concreto, el efecto de un incremento unitario de la variable *estatus* en la nota en ciencias se incrementa en 15.431 puntos para los estudiantes cuyo padre tienes estudios universitarios respecto a los que no.

3) Considere ahora el siguiente modelo para la nota en matemáticas:

$$notamat = \beta_0 + \beta_1 varon + \beta_2 convivienda + \beta_3 estatus + \beta_4 educmadre + \beta_5 educpadre \cdot estatus + \varepsilon$$

a) Estime el modelo, comente los resultados y mejore la especificación en caso de que sea necesario.

Estimamos el modelo y el resultado, que aparece en la Figura 5.1.11, lo guardamos en el objeto *modelo3* con el siguiente código:

*modelo3 <- lm(NOTAMAT ~ VARON + CONVIVIENDA + ESTATUS + EDUCMADRE + I(EDUCPADRE*ESTATUS), data = practica51)*

Figura 5.1.11: summary(modelo3)

```
Call:
lm(formula = NOTAMAT ~ VARON + CONVIVIENDA + ESTATUS + EDUCMADRE +
    I(EDUCPADRE * ESTATUS), data = practica51)

Residuals:
     Min       1Q   Median       3Q      Max
-265.095  -53.442    2.894   54.622  276.406

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            464.359      4.615 100.616  < 2e-16 ***
VARON                   15.177      4.485   3.384 0.000736 ***
CONVIVIENDA             12.032      3.564   3.376 0.000756 ***
ESTATUS                 17.403      3.831   4.543 6.06e-06 ***
EDUCMADRE               10.021      6.444   1.555 0.120195
I(EDUCPADRE * ESTATUS)  14.683      6.367   2.306 0.021256 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 81.15 on 1309 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.1411,    Adjusted R-squared:  0.1379
F-statistic: 43.02 on 5 and 1309 DF,  p-value: < 2.2e-16
```

Como podemos observar en la salida de estimación, Figura 5.1.11, todas las variables del modelo son significativas al 5%, excepto la variable *educmadre*, cuyo p-valor de su contraste de significatividad individual es 0.12. Debido a que su p-valor es mayor que los tamaños del contraste habitualmente usados (aquellos menores a 0.10), eliminamos *educmadre* del modelo. De modo que el modelo resultante es:

$$notamat = \beta_0 + \beta_1 varon + \beta_2 convivienda + \beta_3 estatus + \beta_4 educpadre \cdot estatus + \varepsilon$$

Estimamos el modelo y el resultado, que aparece en la Figura 5.1.12, lo guardamos en el objeto *modelo31* con el siguiente código:

```
modelo31 <- lm(NOTAMAT ~ VARON + CONVIVIENDA + ESTATUS +
      I(EDUCPADRE*ESTATUS), data = practica51)
```

Figura 5.1.12: summary(modelo31)

```
Call:
lm(formula = NOTAMAT ~ VARON + CONVIVIENDA + ESTATUS + I(EDUCPADRE *
    ESTATUS), data = practica51)

Residuals:
    Min      1Q  Median      3Q     Max
-265.72  -51.80    2.35   53.97  278.14

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           468.297      3.812 122.853  < 2e-16 ***
VARON                  14.619      4.468   3.272  0.00110 **
CONVIVIENDA            10.812      3.463   3.122  0.00184 **
ESTATUS                20.304      3.266   6.218 6.76e-10 ***
I(EDUCPADRE * ESTATUS)  15.529      6.339   2.450  0.01443 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81.08 on 1317 degrees of freedom
  (52 observations deleted due to missingness)
Multiple R-squared:  0.1399,    Adjusted R-squared:  0.1373
F-statistic: 53.55 on 4 and 1317 DF,  p-value: < 2.2e-16
```

En este modelo todas las variables son altamente significativos.

b) ¿Es $\hat{\beta}_1$ significativo? Interprete lo que este contraste indica sobre la influencia del género en la nota de matemáticas. ¿Podemos decir que los chicos sacan en promedio mejor nota en matemáticas que las chicas?

El parámetro β_1 es significativo casi para cualquier tamaño del test habitualmente usado, dado su bajo p-valor (0.0011). Esto implica que existe evidencia muy fuerte de que *varon* es una variable relevante en este modelo. Más concretamente, según la estimación, en promedio los chicos obtienen una nota en matemáticas superior en 14.619 puntos a la de las chicas, ceteris paribus.

4) Lleve a cabo el mismo análisis que en el apartado 3, utilizando esta vez la nota en lectura. ¿Qué conclusiones obtiene ahora?

En este caso, el modelo de partida es el siguiente:

$$notalec = \beta_0 + \beta_1 varon + \beta_2 convivienda + \beta_3 estatus + \beta_4 educmadre + \beta_5 educpadre \cdot estatus + \varepsilon$$

Su estimación por MCO, que aparece en la Figura 5.1.13 y guardamos en el objeto *modelo4*, la obtenemos con el siguiente código:

*modelo4 <- lm(NOTALEC ~ VARON + CONVIVIENDA + ESTATUS + EDUCMADRE + EDUCPADRE*ESTATUS, data = practica51)*

Figura 5.1.13: summary(modelo4)

```
Call:
lm(formula = NOTALEC ~ VARON + CONVIVIENDA + ESTATUS + EDUCMADRE +
    I(EDUCPADRE * ESTATUS), data = practica51)

Residuals:
     Min       1Q   Median       3Q      Max
-246.498  -50.768    5.376   56.378  272.839

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           486.720      4.787 101.666  < 2e-16 ***
VARON                 -31.395      4.652  -6.748 2.24e-11 ***
CONVIVIENDA             8.360      3.697   2.262   0.0239 *
ESTATUS                19.448      3.974   4.894 1.11e-06 ***
EDUCMADRE              11.530      6.685   1.725   0.0848 .
I(EDUCPADRE * ESTATUS) 15.037      6.604   2.277   0.0230 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84.18 on 1309 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.1545,    Adjusted R-squared:  0.1512
F-statistic: 47.83 on 5 and 1309 DF,  p-value: < 2.2e-16
```

Ahora, todos los parámetros del modelo son individualmente significativos al menos a un tamaño del test del 9%. Al contrario del modelo del apartado anterior, los chicos obtienen, en promedio, una nota en lectura inferior en 31.395 puntos respecto a la de las chicas, ceteris paribus. Además, esa diferencia no depende del valor del resto de variables explicativas.