Solución a la práctica 4.1 con R

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

Se pretende explicar la demanda de turismo nacional en las 52 provincias españolas durante el año 2014. Para ello se propone el siguiente modelo econométrico:

$$\log(turismo_i) = \beta_0 + \beta_1 \log(densidad_i) + \beta_2 \log(hoteles_i) + \beta_3 kmco_i + \beta_4 ipc_i + \beta_5 pibpc_i + \beta_6 paro_i + \varepsilon_i \quad \textbf{(1)}$$

donde *turismo* es el número de pernoctaciones hoteleras de turistas nacionales en cada una de las provincias españolas en 2014, *densidad* es la densidad de población de cada provincia (población según censo a 1 de enero de 2015), *hoteles* es el número de establecimientos hoteleros en la provincia de destino en 2014, *kmco* son los kilómetros de costa de cada provincia, *ipc* es el IPC provincial en 2014 (base 2011), *pibpc* es el índice del PIB per cápita provincial en 2013 (España=100), y *paro* es la tasa de paro (%) de cada provincia en 2014 según la EPA.[1]

Estime el modelo por MCO y responda a las siguientes preguntas (fichero de datos: *Practica41.RData*):

Como es habitual, antes de responder a las preguntas comprobamos el tipo de datos contenidos en el fichero *Practica41.RData*. Para ello ejecutamos el código *str(Practica41)* y obtenemos la salida de la Figura 4.1.1:

Figura 4.1.1: str(Practica41)

```
Classes 'tbl_df', 'tbl' and 'data.frame':       52 obs. of  7 variables:
 $ DENSIDAD: num  0.85 3.97 1.49 0.69 2.22 ...
 $ HOTELES : num  162 441 186.2 70.3 546.5 ...
 $ IPC     : num  103 103 103 103 103 ...
 $ KMCO    : num  0 244 249 0 401 ...
 $ PARO    : num  27.5 26 35.6 17.2 21.1 ...
 $ PIBPC   : num  79.3 77.2 76.6 150 88.1 ...
 $ TURISMO : num  532003 8169493 3378617 429773 2551975 ...
```

Podemos ver que todas las variables son numéricas y que contienen 52 observaciones. Además, con el código *head(Practica41)* podemos ver las seis primeras observaciones de nuestra muestra.

Figura 4.1.2: head(Practica41)

	DENSIDAD	HOTELES	IPC	KMCO	PARO	PIBPC	TURISMO
1	0.85	162.00000	103.374	0	27.52	79.3	532003
2	3.97	441.00000	103.265	244	26.03	77.2	8169493
3	1.49	186.25000	102.927	249	35.60	76.6	3378617
4	0.69	70.33333	103.468	0	17.20	150.0	429773
5	2.22	546.50000	103.333	401	21.13	88.1	2551975
6	0.36	104.33333	104.050	0	26.04	80.5	397614

Para obtener las estimaciones de los parámetros del modelo (1), Figura 4.1.3, ejecutamos el siguiente código:

lm(log(TURISMO)~log(DENSIDAD)+log(HOTELES)+KMCO+IPC+PIBPC+PARO, data = Practica41)

[1] Fuente de Datos: Instituto Nacional de Estadística (INE) e Instituto Geográfico Nacional.

Figura 4.1.3: lm(log(TURISMO)~log(DENSIDAD)+log(HOTELES)+KMCO+IPC+PIBPC+PARO, data = Practica41)

```
Call:
lm(formula = log(TURISMO) ~ log(DENSIDAD) + log(HOTELES) + KMCO +
    IPC + PIBPC + PARO, data = Practica41)

Coefficients:
  (Intercept)  log(DENSIDAD)   log(HOTELES)          KMCO           IPC         PIBPC          PARO
   32.2292589     0.5363583     0.5254474     0.0005464    -0.2131786     0.0050518     0.0159813
```

Para simplificar la programación llamamos a esta estimación *modelo1* con el siguiente código.

modelo1 <- lm(log(TURISMO)~log(DENSIDAD)+log(HOTELES)+KMCO+IPC+PIBPC+PARO, data = Practica41)

Para obtener más información de esta estimación usamos el operador *summary*. De este modo obtenemos la salida de la Figura 4.1.4.

Figura 4.1.4: summary(modelo1).

```
Call:
lm(formula = log(TURISMO) ~ log(DENSIDAD) + log(HOTELES) + KMCO +
    IPC + PIBPC + PARO, data = Practica41)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8319 -0.2282 -0.0661  0.2173  1.0252

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   32.2292589 12.1930711   2.643 0.011260 *
log(DENSIDAD)  0.5363583  0.1155664   4.641 3.01e-05 ***
log(HOTELES)   0.5254474  0.1309798   4.012 0.000225 ***
KMCO           0.0005464  0.0002320   2.355 0.022943 *
IPC           -0.2131786  0.1167413  -1.826 0.074478 .
PIBPC          0.0050518  0.0049179   1.027 0.309800
PARO           0.0159813  0.0170788   0.936 0.354402
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4161 on 45 degrees of freedom
Multiple R-squared:  0.8589,    Adjusted R-squared:  0.8401
F-statistic: 45.67 on 6 and 45 DF,  p-value: < 2.2e-16
```

1) Realice una estimación de la varianza de ε mediante MCO.

El estimador de la varianza de ε mediante MCO es $s^2 = \frac{\sum_{i=1}^{N} e_i^2}{N-K}$. Para obtener el numerador debemos tener en cuenta el vector de residuos con el código *modelo1$residuals*, y su cuadrado con el código *modelo1$residuals^2*, y, finalmente, sumar el cuadrado del residuo de cada observación con el código *sum(modelo1$residuals^2)*. Al ejecutarlo obtenemos que la suma del cuadrado de los residuos es 7.792953. El denominador vale 45 porque hay 52 observaciones y el modelo (1) contiene 7 parámetros. Por tanto, la varianza estimada es

$$s^2 = \frac{7.792953}{52\text{-}7} = 0.1731767$$

Este resultado también se puede obtener directamente con el código *summary(modelo1)$sigma^2*, dado que *summary(modelo1)$sigma* es la estimación de la desviación típica de ε mediante MCO, s. Notar que esta estimación aparece en el ítem *Residual standard error* de la Figura 4.1.4.

2) ¿Cuál es la estimación de la varianza de los estimadores MCO de los parámetros del modelo?

En la columna *Std. Error* de la salida de la estimación, Figura 4.1.4, tenemos la desviación típica estimada de los coeficientes MCO, $\sqrt{\hat{V}(\hat{\beta}_j)}$. Por tanto, para obtener la estimación de la varianza de los estimadores MCO tenemos que elevar al cuadrado los valores que aparecen en dicha columna. Por ejemplo, $\hat{V}(\hat{\beta}_1) = 0.1155664^2 = 0.01335559$.

3) Contraste la significatividad conjunta de las variables incluidas en el modelo.

Las hipótesis del contraste son:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$
$$H_1 : \text{no } H_0$$

Bajo la hipótesis alternativa al menos uno de los coeficientes de los regresores del modelo (1) difiere de 0. El estadístico de contraste es:

$$F = \frac{\frac{SCE_R - SCE_{NR}}{q}}{\frac{SCE_{NR}}{N-K}} \sim F_{q,N-K}$$

El modelo no restringido es el modelo (1), cuya estimación aparece en la Figura 4.1.5. Por tanto, K = 7 y N=52. El número de restricciones bajo la hipótesis nula es 6, de modo que q = 6. Del apartado 1) sabemos que SCE_{NR} = 7.792953. Solo nos queda por calcular SCE_R. Para ello estimamos el modelo restringido, el cual en este contraste es: $\log(turismo_i) = \beta_0 + \varepsilon_i$. Su estimación aparece en la Figura 4.1.5:

Figura 4.1.5: summary(lm(log(TURISMO)~1, data = Practica41))

```
Call:
lm(formula = log(TURISMO) ~ 1, data = Practica41)

Residuals:
     Min       1Q   Median       3Q      Max
-2.31657 -0.84724 -0.02633  0.86091  2.08191

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.0310     0.1443   97.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.041 on 51 degrees of freedom
```

Para obtener la suma del cuadrado de los residuos de este modelo restringido, SCE_R, ejecutamos:

sum(lm(log(TURISMO)~1, data = Practica41)$residuals^2)

Obtenemos que dicha suma es 55.24567. Por lo tanto, el estadístico de contraste se calcula como:

$$F = \frac{\frac{55.24567 - 7.792953}{6}}{\frac{7.792953}{52-7}} = 45.66887$$

((sum(lm(log(TURISMO)~1, data = Practica41)$residuals^2)-
sum(modelo1$residuals^2))/6)/(sum(modelo1$residuals^2)/(52-7))

La regla de rechazo es que si $F > F_{q,N-K;\alpha}$, se rechaza la hipótesis nula. Dado que el enunciado no especifica el tamaño del contraste, consideramos los tamaños del contraste habituales, los cuales son 0.01, 0.05 y 0.10. Para obtenerlos con R utilizamos el operador *qf* que hace referencia a los cuantíles de la F-Snedecor. El valor crítico al tamaño del 10% se obtiene con el código *qf(0.90,6,45)* y vale $F_{6,45;0.10}$ = 1.909351, el del 5% se obtiene con el código *qf(0.95,6,45)* y vale $F_{6,45;0.05}$ = 2.308273, y el del 1% se obtiene con el código *qf(0.99,6,45)* y vale $F_{6,45;0.01}$ = 3.232472.
Como el valor del estadístico *F*, 45.66887 es mayor que los valores críticos al 1%, 5% y 10%, rechazamos la hipótesis nula. Por lo tanto, concluimos que al menos una de las variables explicativas es relevante y que el modelo es estadísticamente significativo.

Si observamos la salida de la estimación del modelo (1), Figura 4.1.4, vemos que el valor de este estadístico *F* aparece en la parte inferior izquierda bajo el ítem *F-Statistic*. En concreto, aparece *F-statistic: 45.67 on 6 and 45 DF*. También aparece el p-valor asociado al contraste de significatividad conjunta, cuyo valor es *p-valor < 2.2e-16*. Dado que *e-16* significa 10^{-16}, este p-valor es prácticamente a cero. De modo que se rechaza la hipótesis nula para cualquier tamaño del test o nivel de significatividad.

4) Contraste la significatividad individual de las variables incluidas en el modelo ¿Está de acuerdo con el modelo propuesto? Si no es así, ¿qué modelo propone?

Para contrastar la relevancia estadística de cada una de las variables incluidas en el modelo, debemos realizar un contraste de significatividad individual de su coeficiente. Las hipótesis del contraste son:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0 \text{ , donde j=1, 2, 3, 4, 5, 6}$$

El estadístico de contraste es: $t = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} \underset{H_0}{\sim} t_{N-K}$

La regla de rechazo es que si $|t| > t_{52-7;\alpha/2}$, se rechaza la hipótesis nula. Dado que el enunciado no especifica el tamaño del contraste o nivel de significatividad, consideramos los tamaños del contraste habituales. Para obtenerlos con R utilizamos el operador *qt* que hace referencia a los cuantíles de la t-Student. El valor crítico al tamaño del 10% se obtiene con el código *qt(0.95,45)* y vale $t_{45;0.05}$ = 1.679427, el del 5% se obtiene con el código *qt(0.975,45)* y vale $t_{45;0.025}$ = 2.014103, y el del 1% se obtiene con el código *qt(0.995,45)* y vale $t_{45;0.005}$ = 2.689585. Con los datos de la estimación, Figura 4.1.4, se obtiene que:

- El estadístico *t* para β_1 es $t = \frac{0.5364}{0.1156} = 4.641$, por lo que su estimación es estadísticamente significativa. De tal manera que la densidad de población es relevante para explicar la demanda turística.

- El estadístico t para β_2 es $t = \frac{0.5254}{0.13098} = 4.012$, por lo que su estimación es estadísticamente significativa. De tal manera que el número de hoteles es relevante para explicar la demanda turística.

- El estadístico t para β_3 es $t = \frac{0.0005464}{0.0002320} = 2.355$, por lo que su estimación es estadísticamente significativa al 10% y 5%, pero no al 1%. De tal manera que la variable que recoge los kilómetros de costa de cada provincia es relevante para explicar la demanda turística al 5% y 10%, pero no al 1%.

- El estadístico t para β_4 es $t = \frac{-0.2132}{0.1167} = -1.826$, por lo que su estimación es estadísticamente significativa al 10% pero no al 5% y 1%. De tal manera que el IPC de cada provincia es relevante para explicar la demanda turística al 10%, pero no al 1% y 5%.

- El estadístico t para β_5 es $t = \frac{0.00505}{0.00492} = 1.027$, por lo que su estimación no es estadísticamente significativa a ningún nivel de significatividad. De modo que el PIB per cápita provincial no es una variable relevante para explicar el número de pernoctaciones hoteleras en la provincia.

- El estadístico t para β_6 es $t = \frac{0.01598}{0.01708} = 0.936$, por lo que su estimación no es estadísticamente significativa a ningún nivel de significatividad. De modo que la tasa de paro de la provincia no es relevante para explicar el número de pernoctaciones hoteleras en la provincia.

Si observamos la salida de la estimación del modelo (1), Figura 4.1.4, vemos que el valor del estadístico *t* del contraste de significatividad individual de cada parámetro aparece en la columna *t value*. Además, la salida de la estimación también contiene el p-valor asociado a estos contrastes en la columna *Pr(>|t|)*:

- log(*densidad*) tiene un p-valor asociado igual a 0.0000301, por tanto es relevante para cualquier nivel de significatividad.
- log(*hoteles)* tiene un p-valor asociado igual a 0.000225, por tanto es relevante para cualquier nivel de significatividad,
- *kmco* tiene un p-valor asociado igual a 0.022943, por tanto es relevante al 5% y 10% pero no al 1%.
- *ipc* tiene un p-valor asociado igual a 0.074478, por tanto es relevante al 10% pero no al 1% y 5%.
- *pibpc* tiene un p-valor asociado igual a 0.3098, por tanto no es relevante a ningún nivel de significatividad.
- *paro* tiene un p-valor asociado igual a 0.354402, por tanto no es relevante a ningún nivel de significatividad.

La salida de la estimación del modelo (1), Figura 4.1.4, también informa del grado de evidencia en contra de la hipótesis nula del contraste de significatividad individual de cada parámetro. De modo que *** indica que el p-valor es menor o igual a 0.001, y se interpreta como que existe evidencia extremadamente fuerte de que la hipótesis nula no es verdadera; ** indica que el p-valor es mayor que 0.001 y menor o igual que 0.01, y se interpreta como que existe evidencia muy fuerte de que la hipótesis nula no es verdadera; * indica que el p-valor es mayor que 0.01 y menor o igual que 0.05, y se interpreta como que existe evidencia fuerte de que la hipótesis nula no es verdadera; . indica que el p-valor es mayor que 0.05 y menor o igual que 0.10, y se interpreta como que existe evidencia de que la hipótesis nula no es verdadera; y finalmente, la ausencia de símbolo indica que el p-valor es mayor que 0.10, y se interpreta como que no existe evidencia en contra de la hipótesis nula.

El modelo propuesto no es correcto porque presenta inclusión de variables irrelevantes, lo que supone una pérdida de precisión en las estimaciones. Para obtener la especificación adecuada hay que eliminar una a una las variables irrelevantes, comenzando por la que tiene un mayor p-valor. En nuestro caso eliminamos en primer lugar la variable *paro*, y volvemos a estimar el modelo. El resultado que se obtiene es el siguiente:

Figura 4.1.6: summary(lm(log(TURISMO)~log(DENSIDAD)+log(HOTELES)+KMCO+IPC+PIBPC, data = Practica41))

```
Call:
lm(formula = log(TURISMO) ~ log(DENSIDAD) + log(HOTELES) + KMCO +
    IPC + PIBPC, data = Practica41)

Residuals:
     Min       1Q   Median       3Q      Max
-0.75267 -0.23693 -0.06767  0.20424  1.07146

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   37.8681988 10.5854726   3.577 0.000831 ***
log(DENSIDAD)  0.5862073  0.1024168   5.724  7.5e-07 ***
log(HOTELES)   0.4977050  0.1274076   3.906 0.000305 ***
KMCO           0.0005351  0.0002314   2.313 0.025272 *
IPC           -0.2592939  0.1056859  -2.453 0.017998 *
PIBPC          0.0015458  0.0031812   0.486 0.629324
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como la variable *pibpc* sigue sin ser relevante (p-valor=0.629324), la eliminamos y estimamos el siguiente modelo:

$$\log(turismo_i) = \beta_0 + \beta_1 \log(densidad_i) + \beta_2 \log(hoteles_i) + \beta_3 kmco_i + \beta_4 ipc_i + \varepsilon_i \qquad (2)$$

A la estimación de este modelo, que aparece en la Figura 4.1.7, la llamamos *modelo2* con el siguiente código:

modelo2 <- lm(log(TURISMO)~log(DENSIDAD)+log(HOTELES)+KMCO+IPC, data = Practica41)

Figura 4.1.7: summary(modelo2)

```
Call:
lm(formula = log(TURISMO) ~ log(DENSIDAD) + log(HOTELES) + KMCO +
    IPC, data = Practica41)

Residuals:
     Min       1Q   Median       3Q      Max
-0.76892 -0.23539 -0.05295  0.20401  1.03921

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   36.3929357 10.0580223   3.618 0.000723 ***
log(DENSIDAD)  0.5866364  0.1015774   5.775 5.88e-07 ***
log(HOTELES)   0.4992534  0.1263285   3.952 0.000259 ***
KMCO           0.0005305  0.0002293   2.314 0.025109 *
IPC           -0.2437204  0.0998876  -2.440 0.018516 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4122 on 47 degrees of freedom
Multiple R-squared:  0.8555,    Adjusted R-squared:  0.8432
F-statistic: 69.54 on 4 and 47 DF,  p-value: < 2.2e-16
```

De la estimación del modelo (2), Figura 4.1.7, observamos que todas las variables son relevantes al 5%, lo cual indica que existe evidencia fuerte en contra de que las variables explicativas no son relevantes. Por lo tanto, este es el modelo que proponemos para estudiar la demanda de noches de hotel.

5) ¿Cuál es la interpretación de β_2 y de su estimación?

El coeficiente β_2 es la elasticidad de la demanda de noches de hotel respecto al número de establecimientos hoteleros y mide aproximadamente el cambio en la variable *turismo* en % por cada incremento de un 1% en *hoteles*, manteniendo fijas el resto de variables del modelo. Según los resultados, en promedio, se estima que un incremento de un 1% en el número de hoteles de una provincia produce aproximadamente un incremento de un 0.499 % en las pernoctaciones hoteleras en dicha provincia.

6) Contraste la hipótesis nula de que un incremento de un 1% en el número de establecimientos hoteleros produce, en promedio, un incremento de la misma cuantía en las pernoctaciones frente a la alternativa de que dicho incremento sea distinto.

Las hipótesis del contraste son:

$$H_0 : \beta_2 = 1$$
$$H_1 : \beta_2 \neq 1$$

De modo que el estadístico de contraste es: $t = \dfrac{\hat{\beta}_2 - 1}{\sqrt{\hat{Var}(\hat{\beta}_2)}} \underset{H_0}{\sim} t_{N-K}$

La regla de rechazo es que si $|t| > t_{47;\alpha/2}$, se rechaza la hipótesis nula. Dado que el enunciado no especifica el tamaño del contraste o nivel de significatividad, consideramos los tamaños del contraste habituales que son el 1%, 5% y 10%. Por lo tanto, los valores críticos son: $t_{47;0.005}$ = 2.684556, *qt(0.995,47)*; $t_{47;0.025}$ = 2.011741, *qt(0.975,47)*; y, $t_{47;0.05}$ = 1.677927, *qt(0.95,47)*. Con los datos de la estimación, Figura 4.1.7, se obtiene que el estadístico t es -3.963845. Dado que se cumple la regla de rechazo, se rechaza la hipótesis nula y

concluimos que un incremento de un 1% en el número de establecimientos hoteleros produce, en promedio, un incremento de distinta cuantía en las pernoctaciones.

$$t = \frac{0.4992534 - 1}{0.1263285} = -3.963845$$

7) Contraste la hipótesis nula de que un incremento de un 1% en el número de establecimientos hoteleros produce, en promedio, un incremento de la misma cuantía en las pernoctaciones frente a la alternativa de que dicho incremento sea menor.

Las hipótesis del contraste son:

$$H_0 : \beta_2 = 1$$
$$H_1 : \beta_2 < 1$$

El estadístico de contraste es el mismo que en el apartado anterior. De modo que su valor es -3.963845. Este es un contraste unilateral por lo que la región crítica está en una sola cola de la distribución, en este caso en la cola de la izquierda. La regla de rechazo es que si $t < -t_{47;\alpha}$, se rechaza la hipótesis nula. Dado que el enunciado no especifica el tamaño del contraste o nivel de significatividad, consideramos los tamaños del contraste del 1%, 5% y 10%. Por lo tanto, los valores críticos son: $t_{47;0.01}$ = 2.408345, *qt(0.99,47)*; $t_{47;0.05}$ = 1.677927, *qt(0.95,47)*; y, $t_{47;0.10}$ = 1.299825, *qt(0.90,47)*. Dado que se cumple la regla de rechazo, se rechaza la hipótesis nula y concluimos que un incremento de un 1% en el número de establecimientos hoteleros produce, en promedio, un incremento de menor cuantía en la demanda hotelera.

8) Construya un intervalo de confianza para β_2 al 95% e interprételo.

A partir de la estimación del modelo (2), Figura 4.1.7, calculamos el intervalo de confianza para β_2 al 95%, el cual es:

$$IC\left(\hat{\beta}_2\right)_{95\%} = \left(\hat{\beta}_2 - t_{47;0.025}\sqrt{\hat{Var}(\hat{\beta}_2)}, \hat{\beta}_2 + t_{47;0.025}\sqrt{\hat{Var}(\hat{\beta}_2)}\right) =$$
$$= \left(0.4992534 - 2.011741 \times 0.1263285, 0.4992534 + 2.011741 \times 0.1263285\right) = \left(0.2451132,\ 0.7533936\right)$$

El intervalo de confianza se puede calcular utilizadon RStudio como calculador o directamente con el operador *confint*. De modo que si ejecutamos el código *confint(modelo2,level = 0.95)* obtenemos dicho intervalo, más los intervalos de confianza del resto de parámetros del modelo (2) como aparece en la Figura 4.18.

Figura 4.1.8: confint(modelo2,level = 0.95)

	2.5 %	97.5 %
(Intercept)	1.615880e+01	56.6270665532
log(DENSIDAD)	3.822891e-01	0.7909837731
log(HOTELES)	2.451132e-01	0.7533935134
KMCO	6.923314e-05	0.0009918569
IPC	-4.446683e-01	-0.0427725158

Interpretación: El verdadero valor de β_2 pertenece al intervalo (0.245, 0.753) con un nivel de confianza del 95%.

9) Contraste conjuntamente las hipótesis de que la elasticidad de la demanda respecto al número de hoteles es idéntica a la elasticidad de la demanda respecto a la densidad de población, y que por

cada punto que se incrementa el IPC provincial se espera una disminución del 50% de las pernoctaciones. Realice el contraste de forma manual y automática.

Como las variables *turismo*, *densidad* y *hoteles* están en logaritmos, β_1 y β_2 son elasticidades. Sin embargo, como el IPC no está en logaritmos, β_4 es una semielasticidad. Para interpretar el efecto de *ipc* sobre *turismo* en términos porcentuales hay que multiplicar β_4 por 100. Por tanto, las hipótesis del contraste en términos de los parámetros del modelo son:

$$H_0 : \beta_1 = \beta_2$$
$$\beta_4 = -0.5$$
$$H_1 : \text{ no } H_0$$

Bajo la alternativa, al menos una de las dos restricciones no se cumple. El estadístico de contraste es

$$F = \frac{\frac{SCE_R - SCE_{NR}}{q}}{\frac{SCE_{NR}}{N-K}} \underset{H_0}{\sim} F_{q,N-K}$$

La suma del cuadrado de los residuos en el modelo no restringido, modelo (2), es 7.985369 y se calcula con el código *sum(modelo2$residuals^2)*. El modelo restringido se obtiene incorporando las restricciones de la hipótesis nula en el modelo (2) y reordenando los términos.

$$\log(turismo_i) = \beta_0 + \beta_2 \log(densidad_i) + \beta_2 \log(hoteles_i) + \beta_3 kmco_i - 0.5 \cdot ipc_i + \varepsilon_i$$

Bajo la hipótesis nula no hay incertidumbre sobre el parámetro que multiplica a la variable *ipc*. De modo que ese término lo pasamos al lado izquierdo de la igualdad en el que esta la variable explicativa. También hacemos factor común en el parámetro β_2. Nos queda el siguiente modelo:

$$\log(turismo_i) + 0.5 \cdot ipc_i = \beta_0 + \beta_2 (\log(densidad_i) + \log(hoteles_i)) + \beta_3 kmco_i + \varepsilon_i$$

Estimamos el modelo (2) restringido con el código:

*lm(I(log(TURISMO)+0.5*IPC)~I(log(DENSIDAD)+log(HOTELES))+KMCO+PIBPC, data = Practica41)*

donde el código *I()* permite que los términos del modelo incluyan símbolos matemáticos normales. La Figura 4.1.9 muestra la estimación del modelo (2) restringido, el cual llamamos *modelo2_r*.

Figura 4.1.9: summary(modelo2_r)

```
Call:
lm(formula = I(log(TURISMO) + 0.5 * IPC) ~ I(log(DENSIDAD) + 
    log(HOTELES)) + KMCO, data = Practica41)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.86630 -0.25375  0.00013  0.28790  1.10971 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    6.250e+01  2.142e-01 291.796   <2e-16 ***
I(log(DENSIDAD) + log(HOTELES)) 5.821e-01  3.904e-02  14.911   <2e-16 ***
KMCO                           4.227e-04  2.355e-04   1.795   0.0788 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4326 on 49 degrees of freedom
Multiple R-squared:  0.8568,    Adjusted R-squared:  0.8509 
F-statistic: 146.6 on 2 and 49 DF,  p-value: < 2.2e-16
```

La suma del cuadrado de los residuos del modelo (2) restringido es 9.171892 y se calcula con el código *sum(modelo2_r$residuals^2)*. Dado que hay dos restricciones bajo la hipótesis nula (q = 2), N = 52, el modelo (2) contiene cinco parámetros (K = 5), SCE_R = 9.171892 y SCE_{NR} = 7.985369, obtenemos que el valor del estadístico F es:

$$F = \frac{(9.171892 - 7.985369)/2}{7.985369/(52-5)} = 3.491797$$

La regla de rechazo es que si $F > F_{2,47;\alpha}$, se rechaza la hipótesis nula. Dado que el enunciado no especifica el tamaño del contraste, consideramos los tamaños del contraste 0.01, 0.05 y 0.10. Los valores críticos son: $F_{2,47;0.01}$ = 5.087373, *qf(0.99,2,47)*; $F_{2,47;0.05}$ = 3.195056, *qf(0.95,2,47)*; y $F_{2,47;0.10}$ = 2.419168, *qf(0.90,2,47)*. Por lo tanto, la hipótesis nula se rechaza al 5% y 10%, pero no al 1%.

Este contraste también se puede realizar de una forma más directa con R o RStudio. Para ello, debemos instalar el paquete *car*, si no se ha instalado previamente, con el código *install.packages("car")*, y después, cargarlo con el código *library(car)*. La Figura 4.1.10 muestra el resultado del contraste con el operador *linearHypothesis*. Notar que el código hace referencia al modelo no restringido, modelo (2), a las restricciones y al tipo de contraste, F.

Figura 4.1.10: linearHypothesis(modelo2,c("log(DENSIDAD)=log(HOTELES)", "IPC=-0.5"),test="F")

```
Linear hypothesis test

Hypothesis:
log(DENSIDAD) - log(HOTELES) = 0
IPC = - 0.5

Model 1: restricted model
Model 2: log(TURISMO) ~ log(DENSIDAD) + log(HOTELES) + KMCO + IPC

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
1     49 9.1719                              
2     47 7.9854  2    1.1865 3.4918 0.03856 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

El resultado del contraste, Figura 4.1.10, informa del valor del estadístico F del contraste, 3.4918, y su p-valor 0.03856. Notar que el valor del estadístico F coincide con el obtenido anteriormente. Dado que el p-

valor es 0.03856, rechazamos la hipótesis nula al 5% y 10%, pero no al 1%. Este resultado coincide con el obtenido anteriormente. Este resultado significa que hay evidencia fuerte de que al menos una de las restricciones bajo la hipótesis nula no se cumple.

10) Considere ahora la hipótesis nula de que la elasticidad de la demanda de noches de hotel respecto al número de hoteles es idéntica a la elasticidad de la demanda respecto a la densidad de población. Estime un modelo que proporcione directamente el estadístico t para contrastar dicha hipótesis nula ¿A qué conclusión se llega?

$$\begin{matrix} H_0:\beta_1=\beta_2 \\ H_0:\beta_1\neq\beta_2 \end{matrix} \rightarrow \begin{matrix} H_0:\beta_1-\beta_2=0 \\ H_1:\beta_1-\beta_2\neq 0 \end{matrix} \rightarrow \begin{matrix} H_0:\delta=0 \\ H_1:\delta\neq 0 \end{matrix}$$

donde $\delta=\beta_1-\beta_2$. Despejando tenemos que $\beta_1=\delta+\beta_2$. Incorporando $\beta_1=\delta+\beta_2$ en el modelo (2) y reordenando los términos obtenemos el siguiente modelo:

$$\log(turismo_i)=\beta_0+(\delta+\beta_2)\log(densidad_i)+\beta_2\log(hoteles_i)+\beta_3 kmco_i+\beta_4 ipc_i+\varepsilon_i$$

Hacemos factor común en el parámetro β_2 y obtenemos el modelo reparametrizado a estimar.

$$\log(turismo_i)=\beta_0+\delta\log(densidad_i)+\beta_2\left(\log(densidad_i)+\log(hoteles_i)\right)+\beta_3 kmco_i+\beta_4 ipc_i+\varepsilon_i$$

Estimamos el modelo reparametrizado con el código:

lm(log(TURISMO)~log(DENSIDAD)+I(log(DENSIDAD)+log(HOTELES))+KMCO+IPC, data = Practica41)

donde el código *I()* permite que los términos del modelo incluyan símbolos matemáticos normales. La Figura 4.1.11 muestra la estimación del modelo reparametrizado, la cual llamamos *modelo10*.

Figura 4.1.11: summary(modelo10)

```
Call:
lm(formula = log(TURISMO) ~ log(DENSIDAD) + I(log(DENSIDAD) +
    log(HOTELES)) + KMCO + IPC, data = Practica41)

Residuals:
     Min       1Q   Median       3Q      Max
-0.76892 -0.23539 -0.05295  0.20401  1.03921

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 36.3929357 10.0580223   3.618 0.000723 ***
log(DENSIDAD)                0.0873831  0.2136712   0.409 0.684426
I(log(DENSIDAD) + log(HOTELES))  0.4992534  0.1263285   3.952 0.000259 ***
KMCO                         0.0005305  0.0002293   2.314 0.025109 *
IPC                         -0.2437204  0.0998876  -2.440 0.018516 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4122 on 47 degrees of freedom
Multiple R-squared:  0.8555,    Adjusted R-squared:  0.8432
F-statistic: 69.54 on 4 and 47 DF,  p-value: < 2.2e-16
```

La Figura 4.1.11 muestra que el valor del estadístico t para el contraste de significatividad individual de δ es 0.409 y su p-valor es 0.684426. Dado que el p-valor es mayor que 0.10, no rechazamos la hipótesis nula para todos los tamaños de test habitualmente usados.

11) Contraste la hipótesis nula del apartado 10) utilizando el estadístico F. Realice el contraste de forma manual y automática.

La expresión del estadístico de contraste es:

$$F = \frac{\left(SCE_R - SCE_{NR}\right)/q}{SCE_{NR}/\left(N-K\right)} \sim F_{q,N-K}$$

Anteriormente, hemos obtenido que la suma del cuadrado de los residuos en el modelo no restringido, modelo (2), es 7.985369. En este contraste, el modelo restringido es:

$$\log(turismo_i) = \beta_0 + \beta_2\left(\log(densidad_i) + \log(hoteles_i)\right) + \beta_3 kmco_i + \beta_4 ipc_i + \varepsilon_i$$

Con el siguiente código, estimamos el modelo restringido:

lm(log(TURISMO)~I(log(DENSIDAD)+log(HOTELES))+KMCO+IPC, data = Practica41)

La Figura 4.1.12 muestra el resultado de su estimación, la cual llamamos *modelo2_r_10*.

Figura 4.1.12: summary(modelo2_r_10)

```
Call:
lm(formula = log(TURISMO) ~ I(log(DENSIDAD) + log(HOTELES)) +
    KMCO + IPC, data = Practica41)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77265 -0.23058 -0.03407  0.18834  1.05909

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   37.7753477  9.3904462   4.023 0.000203 ***
I(log(DENSIDAD) + log(HOTELES))  0.5483432  0.0390313  14.049  < 2e-16 ***
KMCO                          0.0005182  0.0002253   2.300 0.025859 *
IPC                          -0.2595274  0.0913039  -2.842 0.006555 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4086 on 48 degrees of freedom
Multiple R-squared:  0.8549,    Adjusted R-squared:  0.8459
F-statistic:  94.3 on 3 and 48 DF,  p-value: < 2.2e-16
```

La suma del cuadrado de los residuos en este modelo restringido es 8.013785 y se calcula con el código *sum(modelo2_r_10$residuals^2)*. Dado que hay una restricción bajo la hipótesis nula, el valor del estadístico F es:

$$F = \frac{(8.013785 - 7.985369)/1}{7.985369/(52-5)} = 0.1672486$$

Este estadístico F también se puede calcular con el siguiente código:

((sum(modelo2_r_10$residuals^2)-sum(modelo2$residuals^2))/1)/(sum(modelo2$residuals^2)/(52-5))

La regla de rechazo es que si $F > F_{1,47;\alpha}$, se rechaza la hipótesis nula. Consideramos los tamaños del contraste 0.01, 0.05 y 0.10. Por tanto, los valores críticos son: $F_{1,47;0.01} = 7.206839$, *qf(0.99,1,47)*; $F_{1,47;0.05} = 4.0471$, *qf(0.95,1,47)*; y $F_{1,47;0.10} = 2.815438$, *qf(0.90,1,47)*. Por lo tanto, la hipótesis nula no se rechaza a ningún nivel de significatividad. Este resultado significa que no hay evidencia suficiente de que la restricción bajo la hipótesis nula no se cumple.

Para hacer este contraste de una manera más directa con R utilizamos el operador *linearHypothesis*. La Figura 4.1.13 muestra el resultado del contraste. Notar que el código hace referencia al modelo no restringido, modelo (2), a la restricción y al tipo de contraste, F.

Figura 4.1.13: linearHypothesis(modelo2,c("log(DENSIDAD)=log(HOTELES)"),test="F")

```
Linear hypothesis test

Hypothesis:
log(DENSIDAD) - log(HOTELES) = 0

Model 1: restricted model
Model 2: log(TURISMO) ~ log(DENSIDAD) + log(HOTELES) + KMCO + IPC

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     48 8.0138
2     47 7.9854  1  0.028416 0.1672 0.6844
```

El resultado del contraste, Figura 4.1.13, informa del valor del estadístico F del contraste, 0.1672, y su p-valor 0.6844. Notar que el valor del estadístico F coincide con el obtenido anteriormente. Dado que el p-valor es 0.6844, no rechazamos la hipótesis nula a ningún nivel de significatividad. Este resultado coincide con el obtenido anteriormente.
Como la hipótesis nula no se rechaza conviene incorporar la restricción en el modelo y continuar la práctica con el modelo en el que se impone la restricción $\beta_1 = \beta_2$. La estimación de este "modelo restringido" es más precisa que la del modelo sin restringir.

12) Lleve a cabo una predicción del logaritmo de las pernoctaciones hoteleras para una provincia hipotética que está situada en el interior, cuya densidad de población es igual a 4, cuenta con 300 establecimientos hoteleros, tiene una tasa de paro del 20%, un índice de PIB per cápita del 90% sobre la media nacional y presenta un IPC igual a 103. ¿Cuál es la predicción para el número de pernoctaciones hoteleras? Obtenga un intervalo de confianza para cada una de las predicciones al 95%.

Primero, vamos a realizar una predicción puntual del logaritmo de las pernoctaciones hoteleras. Después, una predicción por intervalos de dicho logaritmo. Finalmente, realizaremos una predicción puntual y por intervalos del número de pernoctaciones hoteleras. Estas predicciones las llevamos a cabo utilizando el "modelo restringido" estimado en el apartado anterior:

$$\log(turismo_i) = \beta_0 + \beta_2\left(\log(densidad_i) + \log(hoteles_i)\right) + \beta_3 kmco_i + \beta_4 ipc_i + \varepsilon_i$$

Para realizar la predicción puntual del logaritmo de las pernoctaciones hoteleras sustituimos el valor de las variables explicativas en la estimación del modelo anterior, Figura 4.1.12, y obtenemos que esa predicción es 14.9318243. Otra forma de realizar esta predicción es estimar el siguiente modelo reparametrizado:

$$\log(turismo_i) = \theta + \beta_2\left(\log(densidad_i) + \log(hoteles_i) - \log(4) - \log(300)\right) + \beta_3 kmco_i + \beta_4(ipc_i - 103) + \varepsilon_i$$

lm(log(TURISMO)~I(log(DENSIDAD)+log(HOTELES)-log(4)-log(300))+KMCO+I(IPC-103),data = Practica41)

A la estimación de este modelo la llamamos *modelo12*. Como podemos ver en la Figura 4.1.13, la predicción puntual del logaritmo de las pernoctaciones hoteleras coincide con la obtenida anteriormente.

Figura 4.1.13: summary(modelo12)

```
Call:
lm(formula = log(TURISMO) ~ I(log(DENSIDAD) + log(HOTELES) -
    log(4) - log(300)) + KMCO + I(IPC - 103), data = Practica41)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77265 -0.23058 -0.03407  0.18834  1.05909

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    14.9318243  0.1208688 123.537  < 2e-16 ***
I(log(DENSIDAD) + log(HOTELES) - log(4) - log(300))  0.5483432  0.0390313  14.049  < 2e-16 ***
KMCO                                            0.0005182  0.0002253   2.300  0.02586 *
I(IPC - 103)                                   -0.2595274  0.0913039  -2.842  0.00656 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4086 on 48 degrees of freedom
Multiple R-squared:  0.8549,    Adjusted R-squared:  0.8459
F-statistic:  94.3 on 3 and 48 DF,  p-value: < 2.2e-16
```

A continuación calculamos el intervalo del logaritmo de las pernoctaciones hoteleras al 95% de confianza, cuya expresión es la siguiente:

$$\hat{\log}(turismo) \pm t_{48;0,025} \cdot \sqrt{\hat{V}(e^0/X^P)}$$

El valor crítico $t_{48;0.025}$ = 2.010635, *qt(0.975,48)* y $\hat{V}(e^0/X^P) = s^2 + \hat{V}(\hat{y}^0/X^P)$, donde *s* corresponde al ítem *Residual standard error* (0.4086) y $\sqrt{\hat{V}(\hat{y}^0/X^P)}$ al *Std. Error* del término constante (0.1208688). Por lo tanto, $\hat{V}(e^0/X^P) = 0.4086^2 + 0.1208688^2 = 0{,}1815632$.[2] Sustituyendo estos datos en la fórmula del intervalo de predicción obtenemos que el intervalo es (14.07509; 15.78856).

La predicción puntual y del intervalo al 95% de confianza del logaritmo de las pernoctaciones hoteleras se puede calcular también directamente con R. Para ello, ejecutamos el siguiente código y obtenemos la Figura 4.1.14:

Figura 4.1.14: predict(modelo12,data.frame(DENSIDAD=4,HOTELES=300, KMCO=0, IPC=103),interval = "prediction")

```
       fit      lwr      upr
14.93182 14.07509 15.78856
```

[2] s^2 también se puede calcular con el siguiente código: *sum(modelo12$residuals^2)/(52-4).*

En la Figura 4.1.14, *fit* se refiere a la predicción puntual y, *lwr* y *upr* a los extremos inferior y superior del intervalo. Por defecto, el operador *predict* calcula intervalos al 95% de confianza. De modo que para calcular intervalos a otro nivel de confianza debemos indicarlo. Por ejemplo, para calcular el intervalo al 99% de confianza debemos ejecutar el siguiente código:

Figura 4.1.15: *predict(modelo12,data.frame(DENSIDAD=4,HOTELES=300, KMCO=0, IPC=103), interval = "prediction",level = 0.99)*

```
       fit      lwr      upr
14.93182 13.78893 16.07472
```

La predicción puntual del número de pernoctaciones hoteleras se obtiene a través de la exponencial de la predicción puntual del logaritmo del número de pernoctaciones hoteleras.

$$\hat{turismo} = e^{14,9318243} = 3053577$$

Por otro lado, la predicción por intervalos del número de pernoctaciones hoteleras se obtiene a través de la exponencial de los extremos del intervalo de confianza del logaritmo de las pernoctaciones hoteleras. De este modo, la predicción por intervalos al 95% es:

$$(e^{14.07509}; e^{15.78856}) = (1296383; 7192578)$$

Este intervalo contiene el verdadero valor de la variable con un 95% de confianza. El intervalo de confianza es muy amplio por lo que la predicción realizada es muy poco precisa. Este resultado no es sorprendente dada la elevada dispersión que tienen algunas variables en la muestra considerada.

Si a la predicción de la Figura 4.1.14 le asignamos el nombre *pred_95*, las predicciones puntual y por intervalos al 95% de confianza del número de pernoctaciones hoteleras también se obtienen ejecutando el código *exp(pred_95)*.

Figura 4.1.15: exp(pred_95)

```
     fit      lwr      upr
3053577 1296383 7192578
```