Regresión PEC3

Diego VZ

2023-03-22

```
library(faraway)
data("prostate")
model <- lm(lpsa ~ ., data = prostate)
summary(model)
```

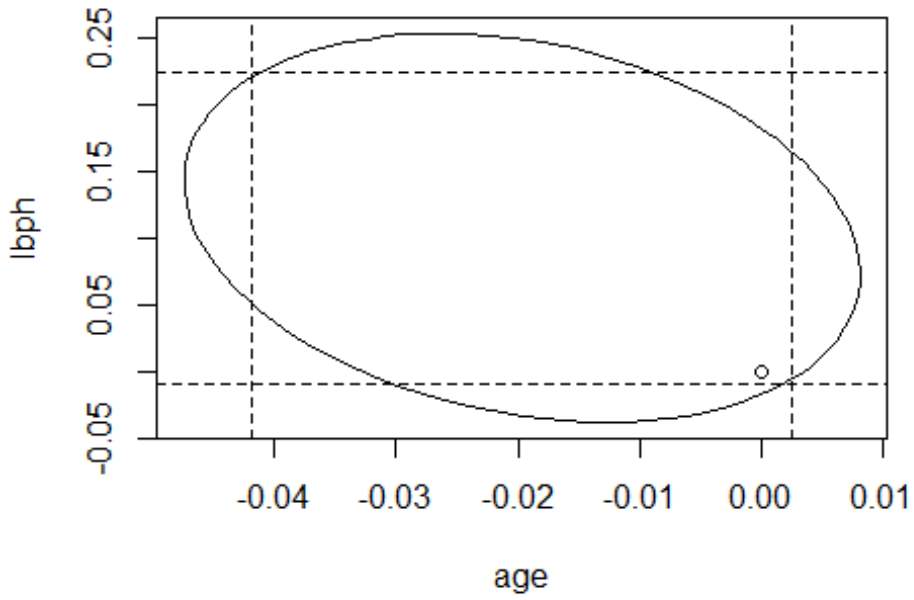
```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

```
confint(model, "age", level = 0.9)
```

```
##             5 %         95 %
## age -0.0382102 -0.001064151
```

```
#Al observar el intervalo de confianza del 90% para el parámetro de edad (-0,038 a -0,001), podemos ver que el intervalo no incluye cero.Esto sugiere que es probable que la edad pueda ser un predictor significativo de lpsa al nivel de significancia del 10%.
confint(model, "age", level = 0.95)
```

```
##           2.5 %      97.5 %
## age -0.04184062 0.002566267

#De manera similar, al observar el intervalo de confianza del 95% para el parámetro de edad (-0.042 a 0.003), podemos ver que el intervalo contiene cero. Esto sugiere que la edad puede o no ser un predictor significativo de lpsa, dependiendo del nivel de significancia elegido.

#Sin embargo, no podemos concluir con certeza que la edad sea un predictor significativo de lpsa ya que el valor p asociado con la variable de edad en la salida de regresión (0,08229) es mayor que 0,05 (el umbral típico de significancia estadística).

library(ellipse)

## Warning: package 'ellipse' was built under R version 4.2.3

##
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':
##
##     pairs

plot(ellipse(model, c('age', 'lbph')), type = "l")
points(0, 0, pch = 1)
abline(v= confint(model)['age',], lty = 2)
abline(h= confint(model)['lbph',], lty = 2)
```



```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'forcats' was built under R version 4.2.3

## ── Attaching core tidyverse packages ──────────────────────
tidyverse 2.0.0 ──
## ✓ forcats   1.0.0     ✓ readr     2.1.4
## ✓ ggplot2   3.4.1     ✓ stringr   1.5.0
## ✓ lubridate 1.9.2     ✓ tibble    3.1.8
## ✓ purrr     1.0.1     ✓ tidyr     1.3.0

## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to
force all conflicts to become errors

library(tidyverse)
t_value <- summary(model) %>% coef() %>% .['age', 't value']
permute_tmod <- function(nsims) {
map_dbl(1:nsims,
~ lm(sample(lpsa) ~ ., data = prostate) %>%
summary() %>%
coef() %>%
.['age', 't value'])
}
mean(abs(permute_tmod(100)) > abs(t_value))

## [1] 0.03

mean(abs(permute_tmod(1000)) > abs(t_value))

## [1] 0.084

mean(abs(permute_tmod(10000)) > abs(t_value))

## [1] 0.085
```

```r
model2 <- update(model, . ~ lcavol + lweight + svi)
anova(model, model2)
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
## Model 2: lpsa ~ lcavol + lweight + svi
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     88 44.163
## 2     93 47.785 -5   -3.6218 1.4434 0.2167
```

```r
#Ejercicio 2
model_ch <- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
summary(model_ch)
```

```
##
## Call:
## lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Acetic        0.3277     4.4598   0.073  0.94198
## H2S           3.9118     1.2484   3.133  0.00425 **
## Lactic       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

```r
#Acetic tiene un valor de p de 0.94198, lo que es mayor a 0.05. Por lo
tanto, Acetic no es un predictor estadísticamente significativo al nivel
del 5%.
#H2S tiene un valor de p de 0.00425, lo que es menor a 0.05. Por lo
tanto, H2S es un predictor estadísticamente significativo al nivel del
5%.
#Lactic tiene un valor de p de 0.03108, lo que es menor a 0.05. Por lo
tanto, Lactic también es un predictor estadísticamente significativo al
nivel del 5%.

model_ch1 <- lm(taste ~ I(exp(1)^Acetic) + I(exp(1)^H2S) + Lactic,
data=cheddar)
summary(model_ch1)
```

```
##
## Call:
## lm(formula = taste ~ I(exp(1)^Acetic) + I(exp(1)^H2S) + Lactic,
##     data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.209  -7.266  -1.651   7.385  26.335
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.897e+01  1.127e+01  -1.684   0.1042
## I(exp(1)^Acetic)  1.891e-02  1.562e-02   1.210   0.2371
## I(exp(1)^H2S)     7.668e-04  4.188e-04   1.831   0.0786 .
## Lactic            2.501e+01  9.062e+00   2.760   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.19 on 26 degrees of freedom
## Multiple R-squared:  0.5754, Adjusted R-squared:  0.5264
## F-statistic: 11.75 on 3 and 26 DF,  p-value: 4.746e-05

#El resultado es que Lactic es el único predictor estadísticamente
significativo al nivel del 5%.

# La prueba F es un método comúnmente utilizado para comparar dos
modelos, uno de los cuales es un subconjunto del otro. En este caso, los
modelos no comparten la misma estructura por lo que no pueden ser
comparados con un prueba F. En estos dos modelo vemos que el que se
ajusta en escala logarítmica natural tiene un ajuste mejor al conjunto de
datos basado en el criterio R^2.

#Para el modelo ajustado en la parte a), la estimación del coeficiente
para H2S es de 3.9118. Esto significa que, manteniendo todas las demás
variables constantes, un aumento de 0.01 en H2S se espera que aumente la
variable taste en 0.01 x 3.9118 = 0.039118 unidades.

# %cambio = (exp(0.01) - 1) x 100 = (1.01005 - 1) x 100 = 1.005%

#Por lo tanto, un aumento aditivo de 0.01 en la escala logarítmica
(natural) corresponde a un cambio porcentual del 1.005% en la escala
original de H2S.

#EJERCICIO 3

data(teengamb)

modelg <- lm(gamble ~ sex+status+income+verbal, data=teengamb)
summary(modelg)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

#Las variables que son estadísticamente significativas al nivel del 5% son el sexo y el ingreso, ya que sus valores de p son menores a 0.05.

```
modelgi <- lm(gamble ~ income, data=teengamb)
summary(modelgi)
```

```
##
## Call:
## lm(formula = gamble ~ income, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.020 -11.874  -3.757  11.934 107.120
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.325      6.030  -1.049      0.3
## income         5.520      1.036   5.330 3.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.95 on 45 degrees of freedom
## Multiple R-squared:  0.387,  Adjusted R-squared:  0.3734
## F-statistic: 28.41 on 1 and 45 DF,  p-value: 3.045e-06
```

```
anova(modelg, modelgi)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: gamble ~ sex + status + income + verbal
## Model 2: gamble ~ income
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     42 21624
## 2     45 28009 -3   -6384.8 4.1338 0.01177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# La variable $sexo está codificada como 0=masculino, 1=femenino y su
coeficiente = -22,118$.
#Esto significa que cuando todas las demás variables se mantienen
constantes y el sexo cambia de masculino a femenino, hay un cambio de -
22.118$ en las apuestas.

#Basándonos en el valor p de la estadística F, tenemos suficiente
evidencia para rechazar la hipótesis nula de que los modelos son
equivalentes en la varianza explicada a través de la estadística RSS.
Afirmamos que el modelo completo es mejor basándonos en el criterio RSS.

#En resumen, esto significa que después de realizar un análisis
estadístico, hemos encontrado suficiente evidencia para decir que el
modelo completo es mejor que otros modelos más simples en términos de
cuánta varianza puede explicar. El valor p es una medida que nos indica
si la evidencia que tenemos en contra de la hipótesis nula es lo
suficientemente fuerte como para rechazarla. En este caso, el valor p fue
lo suficientemente pequeño como para rechazar la hipótesis nula. La
estadística RSS es una medida de cuánto error hay en un modelo, por lo
que si el modelo completo tiene un valor RSS menor que otros modelos,
esto sugiere que el modelo completo es mejor.

#EJERCICIO 4

model_sat <- lm(total ~ expend+ratio+ salary, data=sat)
summary(model_sat)

##
## Call:
## lm(formula = total ~ expend + ratio + salary, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1069.234    110.925   9.639 1.29e-12 ***
## expend        16.469     22.050   0.747   0.4589
## ratio          6.330      6.542   0.968   0.3383
## salary        -8.823      4.697  -1.878   0.0667 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

```
model_sat1 <- lm(total ~ expend+ratio, data=sat)
anova(model_sat, model_sat1)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ expend + ratio + salary
## Model 2: total ~ expend + ratio
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     46 216812
## 2     47 233443 -1    -16631 3.5285 0.06667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Vemos que el estadístico F tiene un valor p de 0.0667. Esto es lo mismo que el valor p para el estadístico t dado arriba para el coeficiente salario
```

```
model_sat0 <- lm(total ~ 1, data=sat)
anova(model_sat0, model_sat)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ 1
## Model 2: total ~ expend + ratio + salary
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     49 274308
## 2     46 216812  3     57496 4.0662 0.01209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Basándonos en el estadístico F, tenemos suficiente evidencia para rechazar la hipótesis nula de que todos los coeficientes son cero. Afirmamos que al menos un predictor tiene un efecto sobre la respuesta.
```

```
model_t <- lm(total ~ expend+ratio+ salary + takers, data=sat)
summary(model_t)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## expend         4.4626    10.5465   0.423    0.674
## ratio         -3.6242     3.2154  -1.127    0.266
## salary         1.6379     2.3872   0.686    0.496
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16

model_t0 <- lm(total ~ expend+ratio+salary, data=sat)
summary(model_t0)

##
## Call:
## lm(formula = total ~ expend + ratio + salary, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1069.234    110.925   9.639 1.29e-12 ***
## expend        16.469     22.050   0.747   0.4589
## ratio          6.330      6.542   0.968   0.3383
## salary        -8.823      4.697  -1.878   0.0667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209

anova(model_t0, model_t)

## Analysis of Variance Table
##
## Model 1: total ~ expend + ratio + salary
## Model 2: total ~ expend + ratio + salary + takers
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     46 216812
## 2     45  48124  1    168688 157.74 2.607e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Al igual que arriba, vemos que el estadístico F para el modelo reducido
tiene un valor p que es el mismo que el valor p del estadístico t dado
anteriormente para el coeficiente takers

#EJERCICIO 4

y<-c(17,34,26,10,19,17,8,16,13,11,
+ 17,41,26,3,-6,-4,11,16,16,4,
+ 21,20,11,26,42,28,3,3,16,-10,
+ 10,24,32,26,52,28,27,28,21,42)
alpha<-c(rep(1,10),rep(0,10),rep(0,10),rep(1,10))
beta<-c(rep(0,10),rep(1,10),rep(1,10),rep(0,10))
gamma<-c(rep(0,10),rep(1,10),rep(0,10),rep(1,10))
crossover.lm<-lm(y ~ alpha+beta+gamma)
crossover.lm0<-lm(y~gamma)
anova(crossover.lm0,crossover.lm)

## Analysis of Variance Table
##
## Model 1: y ~ gamma
## Model 2: y ~ alpha + beta + gamma
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     38 6931.2
## 2     37 6147.9  1    783.23 4.7137 0.03641 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

gamma1<-c(rep(0,10),rep(1,10),rep(0,10),rep(0,10))
gamma2<-c(rep(0,10),rep(0,10),rep(0,10),rep(1,10))
crossover.lm1<-lm(y ~ alpha+beta+gamma1+gamma2)
anova(crossover.lm,crossover.lm1)

## Analysis of Variance Table
##
## Model 1: y ~ alpha + beta + gamma
## Model 2: y ~ alpha + beta + gamma1 + gamma2
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     37 6147.9
## 2     36 5547.3  1    600.62 3.8978 0.05606 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```