Actividad 3: Inferencia

María Sánchez Paniagua

2024-04-2

Ejercicios del libro de Faraway

1. (Ejercicio 1 cap. 3 pág. 48)

For the prostate data, fit a model with lpsa as the response and the other variables as predictors:

(a) Compute 90 and 95% CIs for the parameter associated with age. Using just these intervals, what could we have deduced about the p-value for age in the regression summary?

```r
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
```

```r
confint(model, c("age"), .95)
```

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

```r
confint(model, c("age"), .90)
```

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

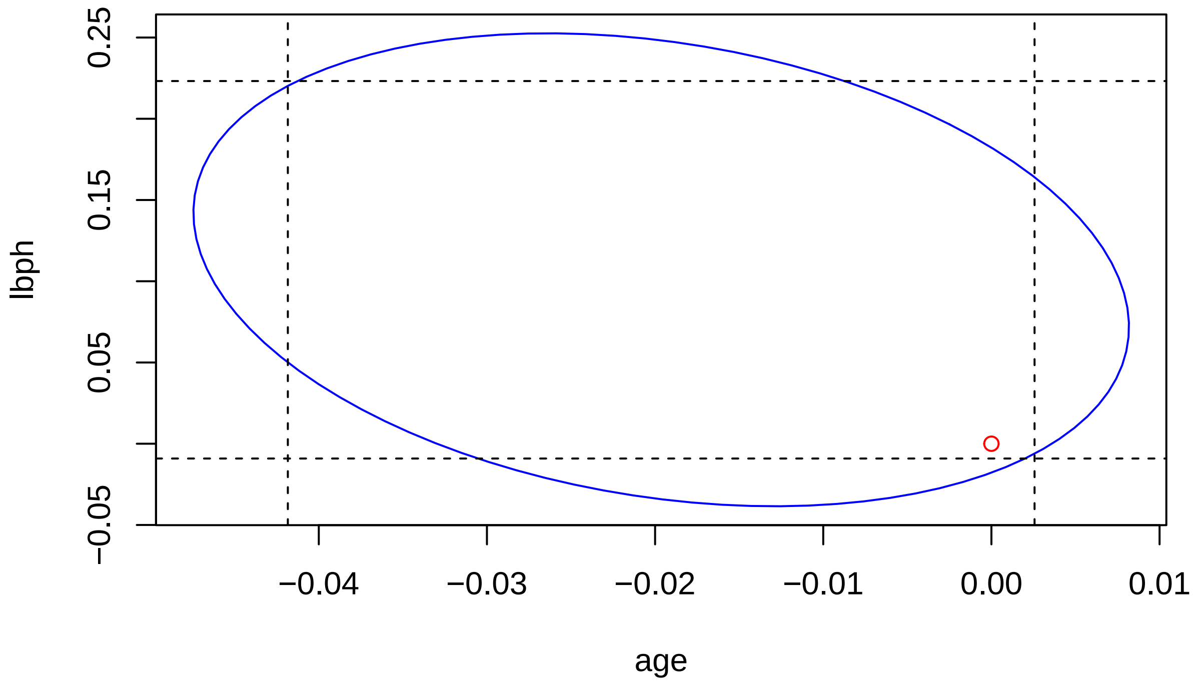
El p-valor con una significancia del 5% es del 0,08229, por lo que según este, no es significativa a este nivel.

Por otro lado, el 0 se encuentra en el intervalo de confianza del 95% pero no al 90%. Por lo que en el caso del 95% no es signfocativamente diferente de 0.

(b) Compute and display a 95% joint confidence region for the parameters associated with age and lbph. Plot the origin on this display. The location of the origin on the display tells us the outcome of a certain hypothesis test. State that test and its outcome.

```r
#coef(model)
plot(ellipse(model, c("age", "lbph"), level = 0.95), type = "l", col = "blue")
points(0, 0, pch = 1, col = "red")

abline(v= confint(model)['age',], lty = 2)
abline(h= confint(model)['lbph',], lty = 2)
```



La hipótesis nula podría ser *Ho: edad = lbph = 0*, pues el origen hace referencia a que ambos valores deberían ser 0.

Como el punto (0,0) está dentro de la elipse, indica que no hay evidencia para rechazar la hipótesis nula, ya que el valor cero para age y pbph están en el intervalo de confianza de la elipse.

(c) In the text, we made a permutation test corresponding to the F-test for the significance of all the predictors. Execute the permutation test corresponding to the t-test for age in this model. (Hint: summary(g)$coef[4,3] gets you the t-statistic you need if the model is called g.)

```r
t_statistic <- summary(model)$coef["age", "t value"]
p_value <- 2 * pt(abs(t_statistic), df = length(model$residuals) - length(model$coef), lower.tail = FALS
p_value # Valor real
```

```
## [1] 0.08229321
```

```r
set.seed(13)
t_value <- summary(model)$coefficients['age', 't value'] #summary(g)$coef[4,3]


permute_tmod <- function(nsim) {
  results <- numeric(nsim)  # Vector para almacenar los resultados

  for (i in 1:nsim) {
    mod_perm <- lm(sample(lpsa) ~ ., data = prostate)
    results[i] <- summary(mod_perm)$coefficients['age', 't value']  # Obtengo el valor t y lo vpy guard
    }
  return(results)  # Devolver los resultados
}

mean(abs(permute_tmod(100)) > abs(t_value))
```

```
## [1] 0.12
```

```r
mean(abs(permute_tmod(500)) > abs(t_value))
```

```
## [1] 0.08
```

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

```
## [1] 0.074
```

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

```
## [1] 0.0814
```

Mediante el test de permutaciones se puede ver que se va acercando la valor real del estadístico.

(d) Remove all the predictors that are not significant at the 5% level. Test this model against the original model. Which model is preferred?

```r
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
```

```r
modelo0 <- update(model, . ~ lcavol + lweight + svi)
anova(model, modelo0)
```

```
## 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
```

Este nuevo modelo no es mejor que el anterior.

2. (Ejercicio 2 cap. 3 pág. 49)

Thirty samples of cheddar cheese were analyzed for their content of acetic acid, hydrogen sulfide and lactic acid. Each sample was tasted and scored by a panel of judges and the average taste score produced. Use the cheddar data to answer the following:

(a) Fit a regression model with taste as the response and the three chemical contents as predictors. Identify the predictors that are statistically significant at the 5% level.

```r
library(faraway)
data(cheddar)
model_cheddar <- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
summary(model_cheddar)
```

```
##
## 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
```

Las variables ‘H2S’ y ‘Lactic’ son estadísticamente significativas al 5%.

(b) Acetic and H2S are measured on a log scale. Fit a linear model where all three predictors are measured on their original scale. Identify the predictors that are statistically significant at the 5% level for this model.

```r
model_cheddar_original <- lm(taste ~ exp(Acetic) + exp(H2S) + Lactic, data = cheddar)
summary(model_cheddar_original)
```

```
##
## Call:
## lm(formula = taste ~ exp(Acetic) + exp(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
## exp(Acetic)  1.891e-02  1.562e-02   1.210   0.2371
## exp(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
```

En este caso, Lactic es el único predictor que es estadísticamente significativo al 5%.

(c) Can we use an F-test to compare these two models? Explain. Which model provides a better fit to the data? Explain your reasoning.

```
anova(model_cheddar, model_cheddar_original)
```

```
## Analysis of Variance Table
##
## Model 1: taste ~ Acetic + H2S + Lactic
## Model 2: taste ~ exp(Acetic) + exp(H2S) + Lactic
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     26 2668.4
## 2     26 3253.6  0    -585.2
```

En este caso, el estadístico F es de 585.2 y el valor p es 0. Por tanto, hay una diferencia significativa entre dos modelos.

Para decidir cuál de los dos modelos ajusta mejor los datos hay que fijarse en el R cuadrado ajustado y el error estándar residual.

```
summary(model_cheddar)
```

```
##
## 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
```

```
summary(model_cheddar_original)
```

```
##
## Call:
## lm(formula = taste ~ exp(Acetic) + exp(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
## exp(Acetic)  1.891e-02  1.562e-02   1.210   0.2371
## exp(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 modelo con las variables originales tiene un R cuadrado ajustado más alto y un error estándar residual más bajo, es decir, mejor ajuste del modelo a los datos.

(d) If H2S is increased 0.01 for the model used in (a), what change in the taste would be expected?

```r
H2S_2 <- 0.01
coef_H2S <- coef(model_cheddar)["H2S"]
(taste_change <- coef_H2S * H2S_2)
```

```
##        H2S
## 0.03911841
```

(e) What is the percentage change in H2S on the original scale corresponding to an additive increase of 0.01 on the (natural) log scale?

```r
exp(H2S_2) - 1 # Paso a la escala original
```

```
## [1] 0.01005017
```

3. (Ejercicio 3 cap. 3 pág. 49)

Using the teengamb data, fit a model with gamble as the response and the other variables as predictors.

(a) Which variables are statistically significant at the 5% level?

```r
library(faraway)
data(teengamb)
model_teengamb <- lm(gamble ~ sex + status + income + verbal + gamble, data = teengamb)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on
## the right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 5 in
## model.matrix: no columns are assigned
```

```
summary(model_teengamb)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal + gamble,
##     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
```

La variables significativas al 5% son sex e income.

(b) What interpretation should be given to the coefficient for sex?

Al tener un valor negativo (-22), indica que los individuos de género femenino tienden a tener un gasto en juegos de azar menor.

(c) Fit a model with just income as a predictor and use an F-test to compare it to the full model.

```
model_income <- lm(gamble ~ income, data = teengamb)
anova(model_income, model_teengamb)
```

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

En este caso, la diferencia entre los dos modelos es estadísticamente significativa con un valor p de 0.01177. Esto sugiere que al menos uno de los predictores adicionales en el Modelo 2 (además de 'income') contribuye a la variabilidad de la variable de respuesta.

Por lo tanto, podemos concluir que el Modelo 2 tiene un mejor ajuste.

4. (Ejercicio 4 cap. 3 pág. 49)

Using the sat data:

(a) Fit a model with total sat score as the response and expend, ratio and salary as predictors.

Test the hypothesis that salary = 0.

Test the hypothesis that salary = ratio = expend = 0.

Do any of these predictors have an effect on the response?

```r
library(faraway)
data(sat)
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
```

El p-valor para la variable salary es 0.0667, por lo que no se puede rechazar la hipótesis nula h0: salary = 0. Por otro lado, el p-vaor del modelo general es 1.29e-12, por lo que el modelo es significativo y por tanto se rechaza la ho: salary = ratio = expend = 0.

(b) Now add takers to the model. Test the hypothesis that takers = 0. Compare this model to the previous one using an F-test. Demonstrate that the F-test and t-test here are equivalent.

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

anova(model_sat_t, model_sat) #Primero la hipótesis nula
```

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

El p-valor es de 2.607e-16, lo que significa que la diferencia es significativa, por lo que ha hipótesis nula (takers = 0) se rechaza.

A continuación voy a demostrar que el F-valor es una generalización del t-valor:

```r
# Estadístico t
t_stat <- summary(model_sat_t)$coefficients['takers', 't value']
t_stat^2
```

```
## [1] 157.7379
```

```r
# Estadístico f
f_est <- anova(model_sat, model_sat_t)[2, 'F']
f_est
```

```
## [1] 157.7379
```

Como se observa, el valor al cuadrado del estadístico t es igual al estadístico F.

#Otros ejercicios

###1.

En los ejemplos 5.3.2 y 5.6.3 del libro de Carmona y con los datos del diseño cross-over simplificado considerar el modelo en el que el efecto de la interacción es distinto cuando primero se administra el tratamiento a y a continuación el tratamiento b, que cuando se hace al revés. Es decir, hay dos parámetros distintos: ab y ba. Contrastar en ese modelo la hipótesis H0 : ab = ba. Comprobar primero que es una hipótesis contrastable.

```r
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))
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))
gammasum <- c(rep(0,10),rep(1,10),rep(0,10),rep(1,10))

lm0<-lm(y~alpha+beta+gammasum) #Modelo de la hipótesis nula
lm1<-lm(y~alpha+beta+gamma1+gamma2)

contraste <- anova(lm0, lm1)
contraste[2, 'Pr(>F)']
```

```
## [1] 0.05605847
```

```
contraste
```

```
## Analysis of Variance Table
##
## Model 1: y ~ alpha + beta + gammasum
## 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
```

El p valor obtenido, 0.05605847 indica que no se puede rechazar la hipótesis nula y por lo tanto, no hay diferencias entre los dos modelos (no importa el orden de administración de los fármacos).