Inferencia ejercicios

Laura Sudupe Medinilla

19/3/2021

Ejercicios 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:

```
lm1 <- lm(lpsa  ~ ., data = prostate)
summary(lm1)
```

```
##
## 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
```

(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
confint(lm1, level=0.9)
```

```
##                     5 %         95 %
## (Intercept) -1.485718237  2.824391633
## lcavol       0.440867156  0.733176497
## lweight      0.171846568  0.737088281
## age         -0.038210200 -0.001064151
## lbph         0.009890745  0.204217317
## svi          0.360029029  1.172285623
## lcp         -0.256770899  0.045822373
## gleason     -0.216620186  0.306903382
## pgg45       -0.002824333  0.011874796
```

```r
confint(lm1, level=0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) -1.906960983 3.245634379
## lcavol       0.412298699 0.761744954
## lweight      0.116603435 0.792331414
## age         -0.041840618 0.002566267
## lbph        -0.009101499 0.223209561
## svi          0.280644232 1.251670420
## lcp         -0.286344443 0.075395916
## gleason     -0.267786053 0.358069248
## pgg45       -0.004260932 0.013311395
```

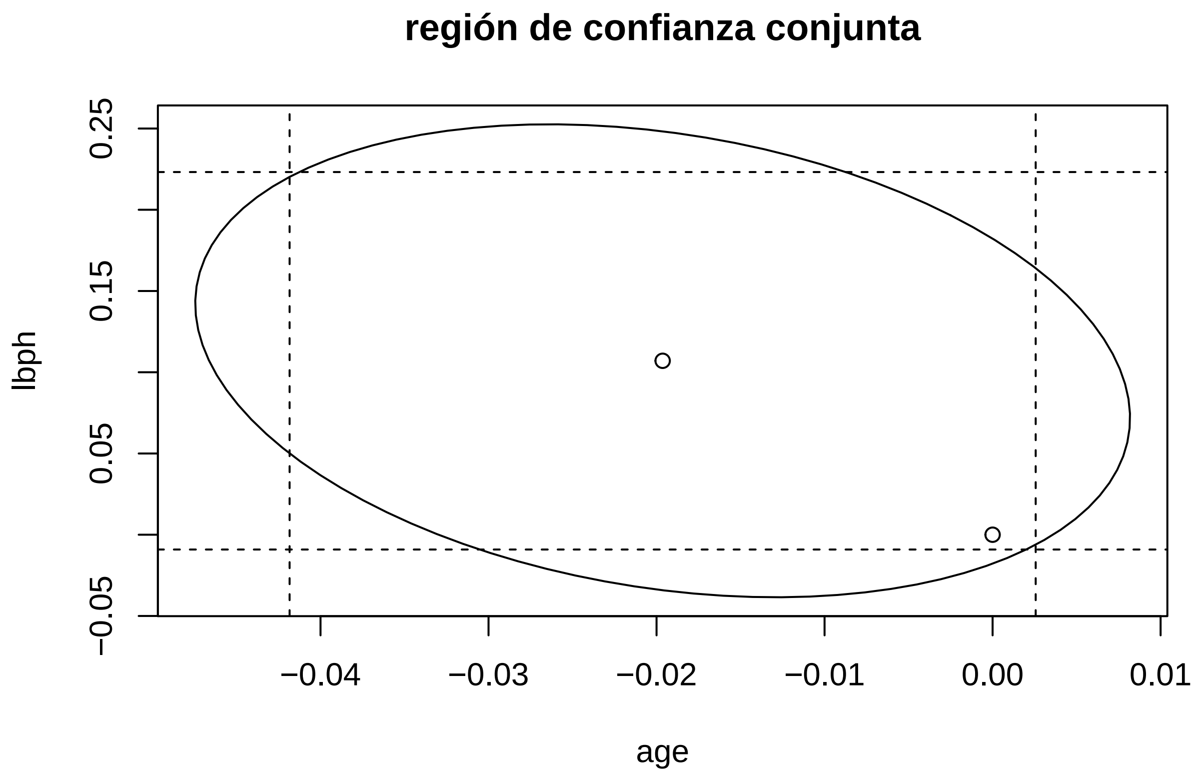
Cuando calculamos el intervalo de confianza al 90%, no esta incluido el 0. Al calcularlo al 95%, mas restrictivo, el 0 si esta incluido. En el primer caso entonces diremos que el parametro `age` es significativo pero para el segundo caso no. En el `summary()` del modelo vemos que su p-value es 0.08229, lo que reafirma lo visto con los intervalos.

(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.

Tenemos que calcular la región de confianza conjunta de los parametros `age` y `lbph`. La hipotesis a estudiar;

$$H_0 : \beta_{age} = \beta_{lbph} = 0$$

```r
plot(ellipse(lm1, c("age","lbph")), type="l", main = "región de confianza conjunta")
points(coef(lm1)["age"], coef(lm1)["lbph"])
points(0,0)
abline(v = confint(lm1)["age",], lty=2)
abline(h = confint(lm1)["lbph",], lty=2)
```



El centro del elipse es la estimación puntual de los parametros. El origen esta cerca de salirse de la elipse, por lo tanto, no tenemos evidencias para rechazar la hipotesis nula.

(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.)

Tenemos que hacer un contraste de hipotesis

$$H_0 : \beta_{age} = 0$$

Y no igual a 0

Con el test, vamos a obtener un valor similar al p-value calculado para la variable `age`. Estimaremos varias veces el modelo e iremos guardando los valores t-value y p-value.

```
#Para el test de permutaciones establecemos una semilla de aleatorizacion
set.seed(123)

t_value <- summary(lm1) %>% coef() %>% .['age', 't value']

#Funcion para permutaciones
permute_tmod <- function(nsims) {
    map_dbl(1:nsims,
            ~ lm(sample(lpsa) ~ ., data = prostate) %>%
            summary() %>%
            coef() %>%
```

```
            .['age', 't value'])
}
mean(abs(permute_tmod(1000)) > abs(t_value))
```

```
## [1] 0.085
```

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

```
lm1.2 <- lm(lpsa ~ lcavol + lweight + svi, data = prostate)
anova(lm1, lm1.2)
```

```
## 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
```

No podemos rechazar la hipotesis nula porque el valor es superior a 0.05. No hay mucha diferencia entre modelos.

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.

```
lm2 <- lm(taste ~ ., data = cheddar)
summary(lm2)
```

```
##
## Call:
## lm(formula = taste ~ ., 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
```

Los predictores estadisticamente significativos son `H2S` y `Lactic`. En el `summary()` podemos ver el contraste de hipotesis y el p-value por cada estadistico t, indicando con alpha=5% si la hipotesis se rechaza o no.

$$H_0 : \beta_{variable} = 0$$

(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
lm2.2 <- lm(taste ~ I(exp(Acetic)) + exp(H2S) + Lactic, data = cheddar)
summary(lm2.2)
```

```
##
## Call:
## lm(formula = taste ~ I(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
## I(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
```

Aqui podemos ver que el único predictor significativo es `Lactic`

(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.

No. Las variables tienen que ser las mismas. Para decir que modelo se ajusta mejor a los datos podemos mirar el valor de R^2 de cada modelo

```r
summary(lm2)$r.squared
```

```
## [1] 0.6517747
```

```r
summary(lm2.2)$r.squared
```

```
## [1] 0.575407
```

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

Multiplicamos el coeficiente de esta variable por 0.01

```r
coef(lm2)[3]*0.01
```

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

Vemos un aumento de 0.039 en la respuesta `taste`

(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?

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.

```r
lm3 <- lm(gamble ~ ., data = teengamb)
```

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

```r
summary(lm3)
```

```
##
## Call:
## lm(formula = 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
```

`sex` e `income`.

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

Con el coeficiente medimos el aumento de la prediccion de la variable explicada por cada aumento de unidad de este manteniendo constantes los demas predictores. O es para el hombre y 1 para la mujer, en este caso. Si multiplicamos 0 por el coeficiente de Sex obtenemos 0, que significa que no hay variación.

En cambio, si tomamos las mujeres obtendríamos -22.11. Por lo tanto, una disminución de 22.11 en la variable respuesta.

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

Este es nuestro contraste de hipotesis

$$H_0 : \beta_{sex} = \beta_{status} = \beta_{verbal} = 0$$

Si no rechazamos la hipotesis nula, entonces el modelo con solo `income` como predictora es mejor que el modelo inicial con todas las variables como predictoras

```
lm3.1 <- lm(gamble ~ income, data = teengamb)
anova(lm3, lm3.1)
```

```
## 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
```

Tenemos un p-value de 0.01177 por lo que rechazamos la hipotesis nula. El modelo con todas variables se ajusta mejor

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 betasalary = 0. Test the hypothesis that betasalary = betaratio = betaexpend = 0. Do any of these predictors have an effect on the response?

```
lm4 <- lm(total ~ expend + ratio + salary, data = sat)
summary(lm4)
```

```
##
## 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
```

En el caso de la primera hipotesis, $\beta_{salary} = 0$, tenemos que mirar en `summary()` el contraste individual de la variable `salary` que obtenemos con el t-test. El coeficiente no sera significativo, por lo tanto diremos que esta variable no es estadisticamente significativa, con una confianza del 95%

En este test, el valor del estadístico F es de 4.066, dejando un p-valor de 0.01209 que implica rechazar la $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$

(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
lm4.2 <- lm(total ~ 1, data = sat)
anova(lm4.2, lm4)
```

```
## 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
```

Vemos que el resultado con el t-test rechaza la hipotesis nula, entonces, `takers` es significativa.