El Model de Regressió Lineal

1

Lectura de la informació

Fitxer en format EXCEL: GPA1.xls Font: Wooldridge (5th edition), Ejemplo 3.1., p. 75 141 observacions
colGPA: nota mitjana en la universitat hsGPA: nota mitjana en l'institut ACT: resultat prova accés a la universitat SKIPPED : número de classes perdudes per setmana

```
# Llegim les dades
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.0.5
```

```
dades <- read_xls("gpa1.xls")
```

```
# Seleccionem aquelles variables que ens interessin i generem una nova BD
myvars <- c("colGPA", "hsGPA", "ACT", "skipped")
mydata <- dades[myvars]
```

```
mydata
```

```
## # A tibble: 141 x 4
##   colGPA hsGPA   ACT skipped
##   <dbl> <dbl> <dbl>   <dbl>
## 1     3     3    21     2
## 2    3.4    3.2    24     0
## 3     3     3.6    26     0
## 4    3.5    3.5    27     0
## 5    3.6    3.9    28     0
## 6     3     3.4    25     0
## 7    2.7    3.5    25     0
## 8    2.7     3    22     3
## 9    2.7     3    21     2
## 10   3.8     4    27    0.5
## # ... with 131 more rows
```

2

Especificació del model

$$\text{colGPA} = \beta_1 + \beta_2 * \text{hsGPA} + \beta_3 * \text{ACT} + \beta_4 * \text{SKIPPED} + U$$

2.1. Estimació MQ

```
result <- lm(colGPA ~ hsGPA + ACT + skipped, data = mydata)

output <- summary(result) # Resultats de l'estimació
output

##
## Call:
## lm(formula = colGPA ~ hsGPA + ACT + skipped, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.85698 -0.23200 -0.03935  0.24816  0.81657
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.38955     0.33155   4.191 4.95e-05 ***
## hsGPA         0.41182     0.09367   4.396 2.19e-05 ***
## ACT           0.01472     0.01056   1.393  0.16578
## skipped      -0.08311     0.02600  -3.197  0.00173 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3295 on 137 degrees of freedom
## Multiple R-squared:  0.2336, Adjusted R-squared:  0.2168
## F-statistic: 13.92 on 3 and 137 DF,  p-value: 5.653e-08
```

Alguns comentaris:

- t value fa referència a l'estadístic de significació parcial. Si ens fixem en els p-valors associats a cadascun dels estadístic t , veiem que el paràmetre que acompanya la variable ACT no és significatiu (p-valor > 0.05). Per tant, la variable ACT no és rellevant en el model.
- F -static fa referència a l'estadístic de significació conjunta. $F = 13.92$ i el p-valor associat és inferior a 0.05, la qual cosa significa que al menys un dels paràmetres és diferent de 0 estadísticament parlant. Així doncs, el model es significatiu en el seu conjunt.
- El R^2 no és massa elevat. En realitat no podem esperar un R^2 massa alt donat que tenim moltes variables, però tot i així, l'ajust no és massa bo (hi ha una variable no rellevant).

Interpretació paramètrica (tant signe com valor):

- hsGPA: signe positiu, així doncs, increments en la nota mitjana de l'institut es traslladen en increments de la nota mitjana de la universitat, la qual cosa té sentit. La estimació del coeficient és 0.41, la qual cosa significa que 1 punt més en la nota mitjana de l'institut suposa un increment de 0.41 punts en la mitjana de la universitat. Fixa't que si la estimació fos igual o superior a 1, aquest valor no seria lògic. Si mirem l'estadístic t i el p-valor associat conclouem que el coeficient és estadísticament diferent de 0 i, en conseqüència, la variable hsGPA és una variable rellevant per explicar les variacions de la variable dependent.

- ACT: signe positiu, així doncs, esperem que els estudiants que tinguin més nota a la Selectivitat també tinguin una nota mitjana de la universitat més alta. El valor estimat per aquest coeficient és 0.014, la qual cosa significa que un increment d'1 punt en la nota de Selectivitat suposa un increment de 0.014 en la nota mitjana de la universitat. Així doncs, la influència de la variable ACT sobre la variable dependent és molt petita. Per una altra part, el p-valor > 0.05 , la qual cosa indica que el paràmetre és significativament igual a 0 i, conseqüentment, la variable ACT no és rellevant per explicar la variable dependent.
- Skipped: signe negatiu, així doncs, com més hores perdudes de classe, més baixa és la nota mitjana de la universitat d'un cert estudiant, la qual cosa és coherent. La estimació del coeficient és de -0.0831, la qual cosa significa que un dia més de pèrdua de classe es trasllada en una reducció de 0.08 punts de la nota mitjana de la universitat. Per una altra part, el p-valor < 0.05 , la qual cosa indica que el paràmetre és significatiu i, en conseqüència, aquesta variable és rellevant per explicar la nota mitjana de l'expedient de la universitat.

2.2. Parts del model

Obtenim diferents parts del model de regressió que ens interessin.

```
vcov(result) # Matriu de variances i covariances dels paràmetres estimats
```

```
##           (Intercept)          hsGPA          ACT          skipped
## (Intercept)  0.1099277478 -0.0216112087 -1.437745e-03 -8.394934e-04
## hsGPA        -0.0216112087  0.0087748585 -3.562852e-04  3.386380e-04
## ACT          -0.0014377454 -0.0003562852  1.116165e-04 -4.305563e-05
## skipped      -0.0008394934  0.0003386380 -4.305563e-05  6.759236e-04
```

```
SQE <- deviance(result) # Suma de Quadrats dels Errors, únicament en els models lineals.
```

```
DegreesOfFreedom <- result$df # Graus de llibertat del model
```

```
Yhat <- result$fitted.values # Y ajustada del model
```

```
Resid <- result$residuals # Residus del model
```

```
Coef <- result$coefficients # Estimacions dels paràmetres
```

```
s <- output$sigma # Arrel quadrada de l'estimació de la variància del terme de pertorbació
```

```
RSquared <- output$r.squared # R quadrat
```

```
ARSquared <- output$adj.r.squared # R quadrat ajustat
```

3

Prediccions

Suposem que ens interessa conèixer el valor de la variable dependent y per a una observació individual.

El primer que fem és crear una nova base de dades que contingui els valors de la observació.

```
newdata <- data.frame(hsGPA = 3, ACT = 24, skipped = 2)
```

Utilitzem la funció `predict` per obtenir la predicció.

```
pred.puntual <- predict(result, newdata)
pred.puntual
```

```
##          1
## 2.812062
```

Si en comptes d'obtenir el valor puntual ens interessa conèixer quin és l'interval de confiança al 95% el codi és el següent. Cal remarcar que existeixen dos tipus d'interval:

- Interval per a un cas concret
- Interval per a la predicció mitjana

```
prova <- predict(result, newdata, interval = "prediction") # interval de confiança 95% de la predicció c
prova <- predict(result, newdata, interval = "confidence") # interval de confiança 95% de la mitjana
```

4

Constrast d'hipòtesis

4.1. Contrast d'hipòtesi 1

$$H_0 : \beta_2 + \beta_4 = 0.3$$

Construïm el model genèric, aquell que considera totes les variables, tal i com hem fet anteriorment. Posteriorment, utilitzem la funció `linearHypothesis` indicant la hipòtesi nul·la. Fixa't que fem servir el nom de les variables i no pas els paràmetres.

```
library(car)

# Construcció del model.
result <- lm(colGPA ~ hsGPA + ACT + skipped, data = mydata)

linearHypothesis(result, c("hsGPA + skipped "), rhs = 0.3, test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## hsGPA + skipped = 0.3
##
## Model 1: restricted model
## Model 2: colGPA ~ hsGPA + ACT + skipped
##
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     138 14.882
## 2     137 14.873  1 0.0088309 0.0813 0.7759
```

El p-valor obtingut > 0.05 , per tant, acceptem la hipòtesi nul·la. La restricció és certa.

4.2. Contrast d'hipòtesi 2

$H_0 : \beta_2 + \beta_4 = 0.3$ i $\beta_3 = 0.05$

```
linearHypothesis(result, c("hsGPA + skipped", "ACT"), rhs = c(0.3, 0.05), test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## hsGPA + skipped = 0.3
## ACT = 0.05
##
## Model 1: restricted model
## Model 2: colGPA ~ hsGPA + ACT + skipped
##
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     139 16.202
## 2     137 14.873  2    1.3293 6.1222 0.00284 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor < 0.05 , la qual cosa suposa que hem de rebutjar la hipòtesi nul·la. Així doncs, podem concloure que **les dues restriccions aplicades de manera simultània** no són certes.