Análisis para el dataset Caschool (Parte 3)

Joel Alejandro Zavala Prieto

Contents

Información de contacto

Mail: alejandro.zavala1001@gmail.com

Facebook: https://www.facebook.com/AlejandroZavala1001

Git: https://github.com/AlejandroZavala98

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

Modelando el dataset caschool

Descripción

En esta parte se hara un análisis del conjunto de datos "Caschool". Cuya descripción citare

> "La base de datos caschool.RData contiene información de las calificaciones de estudiantes de puntaje de prueba de California

dist_code:	district Code;
Read_scr:	avg Reading Score;
Math_scr:	avg Math Score;
County :	county;
District:	District;
gr_span:	grade span of district;
enrl_tot :	total enrollment;
teachers:	number of teachers;
computer:	number of computers;
testscr:	avg test score (= (read_scr+math_scr)/2);
comp_stu:	computers per student (= computer/enrl_tot);
expn_stu:	expentitures per student ($'s);
str:	NA
el_pct:	percent of English Learners;
Meal_pct:	Percent qualifying for reduced-price lunch;
cAlw_pct:	Percent qualifying for CalWorks;
avGinc:	district average income (in $1000's);

Visualización

Se muestra a continuación las variables a tomar para el analisis:

Las calificaciones obtenidas en los exámenes (testscr) con la ratio estudiantes-maestros (str) y el porcentaje de estudiantes que estudian inglés (el_pct).

Mostrando las primeras 10 observaciones

testscr	str	el_pct
690.80	17.88991	0.000000
661.20	21.52466	4.583334
643.60	18.69723	30.000002
647.70	17.35714	0.000000
640.85	18.67133	13.857677
605.55	21.40625	12.408759
606.75	19.50000	68.717949
609.00	20.89412	46.959461
612.50	19.94737	30.079157
612.65	20.80556	40.275921

Propuesta de modelo

Se propone el modelo

$$testscr_i = \beta_0 + \beta_1 str_i + \beta_2 elpct_i + \epsilon_i$$

Cuya función ajustada es:

$$\hat{testscr_i} = \hat{\beta_0} + \hat{\beta_1} str_i + \hat{\beta_2} elpct_i$$

Ajuste por forma matricial

Recordando el modelo $Y = X\beta$

La matrix X^tX

```
##          [,1]       [,2]       [,3]
## [1,]  420.000   8248.979   6622.625
## [2,] 8248.979 163513.029 132790.993
## [3,] 6622.625 132790.993 244529.766
```

La matrix $(X^tX)^{-1}$

```
##               [,1]          [,2]          [,3]
## [1,]  0.2625332419 -1.336368e-02  1.468819e-04
## [2,] -0.0133636794  6.911896e-04 -1.341804e-05
## [3,]  0.0001468819 -1.341804e-05  7.398080e-06
```

Para obtener finalmente

```
##             [,1]
## [1,] 686.0322487
## [2,]  -1.1012959
## [3,]  -0.6497768
```

Que por linea de comando

```
##
## Call:
## lm(formula = testscr ~ str + el_pct, data = caschool)
##
## Coefficients:
## (Intercept)          str       el_pct
##    686.0322      -1.1013      -0.6498
```

Cuya función ajustada finalmente es:

$$\hat{testscr_i} = 686.0322 - 1.1013 str_i - 0.6498 elpct_i$$

Medidas extras

Observaciones ajustadas

Las observaciones ajustadas de forma matricial son (mostrando primeros 5 observaciones):

```
## [1] 666.3302 659.3491 645.9478 666.9169 656.4652
```

Las observaciones ajustadas por linea de comando (mostrando primeras 5 observaciones):

```
##        1        2        3        4        5
## 666.3302 659.3491 645.9478 666.9169 656.4652
```

Residuales

Los residuales de forma matricial son (mostrando primeros 5 observaciones):

```
## [1]  24.469823   1.850931  -2.347791 -19.216886 -15.615218
```

Los residuales por linea de comando son (mostrando primeros 5 observaciones):

```
##          1          2          3          4          5
##  24.469823   1.850931  -2.347791 -19.216886 -15.615218
```

Error Estándar de la regresión

El error estándar de la regresión de forma matricial es:

```
##          [,1]
## [1,] 14.46448
```

El error estándar de la regresión por linea de comando es:

```
## [1] 14.46448
```

Suma de los residuos al cuadrado

SRC

```
##          [,1]
## [1,] 87245.29
```

Suma explicada de cuadrados

SEC

```
##         [,1]
## [1,] 64864.3
```

Suma total de cuadrados

STC

```
##          [,1]
## [1,] 152109.6
```

Coeficiente de determinación

De forma manual

```
##           [,1]
## [1,] 0.4264314
```

Por linea de comando

```
## [1] 0.4264314
```

Coeficiente de determinación ajustado

De forma manual

```
##           [,1]
## [1,] 0.4236804
```

Por linea de comando

```
## [1] 0.4236804
```

Inferencias a los parámetros obtenidos

Test de Shapiro-Wilk

```
##
##  Shapiro-Wilk normality test
##
## data:  ols_caschool$residuals
## W = 0.99634, p-value = 0.4515
```

Test de Breusch-Pagan

```
##
##  studentized Breusch-Pagan test
##
## data:  ols_caschool
## BP = 29.501, df = 2, p-value = 3.925e-07
```

Matrix de varianza-covarianza homocedástica

De forma matricial la matrix de varianza-covarianza homocedástica es:

```
##           [,1]      [,2]      [,3]
## [1,] 54.927553 -2.795967  0.030731
## [2,] -2.795967  0.144612 -0.002807
## [3,]  0.030731 -0.002807  0.001548
```

Por linea la matrix de varianza-covarianza homocedástica es:

```
##             (Intercept)         str       el_pct
## (Intercept) 54.92755274 -2.79596671  0.030730824
## str         -2.79596671  0.14461160 -0.002807340
## el_pct       0.03073082 -0.00280734  0.001547836
```

Significancia para el intercepto

Sabemos que el parámetro estimado del intercepto es:

```
##          [,1]
## [1,] 686.0322
```

De forma matricial el error estándar es:

```
## [1] 7.411312
```

Cuyo estadístico "T" es:

```
##          [,1]
## [1,] 92.56555
```

Significancia para el coeficiente de str

Sabemos que el parámetro estimado del coeficiente de la variabñe str es:

```
##           [,1]
## [1,] -1.101296
```

De forma matricial el error estándar es:

```
## [1] 0.3802783
```

Cuyo estadístico "T" es:

```
##           [,1]
## [1,] -2.896026
```

Significancia para el coeficiente de el_pct

Sabemos que el parámetro estimado del coeficiente de la variable el_pct es:

```
##            [,1]
## [1,] -0.6497768
```

De forma matricial el error estándar es:

```
## [1] 0.03934255
```

Cuyo estadístico "T" es:

```
##           [,1]
## [1,] -16.51588
```

Por linea de comando

```
##
## Call:
## lm(formula = testscr ~ str + el_pct, data = caschool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.845 -10.240  -0.308   9.815  43.461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 686.03225    7.41131  92.566  < 2e-16 ***
## str          -1.10130    0.38028  -2.896  0.00398 **
## el_pct       -0.64978    0.03934 -16.516  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.46 on 417 degrees of freedom
## Multiple R-squared:  0.4264, Adjusted R-squared:  0.4237
## F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```