Caschool Parte 2

Joel Alejandro Zavala Prieto

Contents

Información de contacto

Mail: alejandro.zavala1001@gmail.com

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

Git: https://github.com/AlejandroZavala98

Descripción del problema

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

Una pequeña descripción de las variables de la base de datos se da a continuación

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

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

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);

Modelo

Se propone el modelo

$$testscr_i = \beta_0 + \beta_1 str_i + u_i$$
$$i = 1, 2, ..., n$$

El nombre de columnas de la base de datos se muestra a continuación

```
##  [1] "Observation Number" "dist_cod"           "county"
##  [4] "district"           "gr_span"            "enrl_tot"
##  [7] "teachers"           "calw_pct"           "meal_pct"
## [10] "computer"           "testscr"            "comp_stu"
## [13] "expn_stu"           "str"                "avginc"
## [16] "el_pct"             "read_scr"           "math_scr"
```

Mostrando las primeras observaciones de la tabla para las variables requeridas

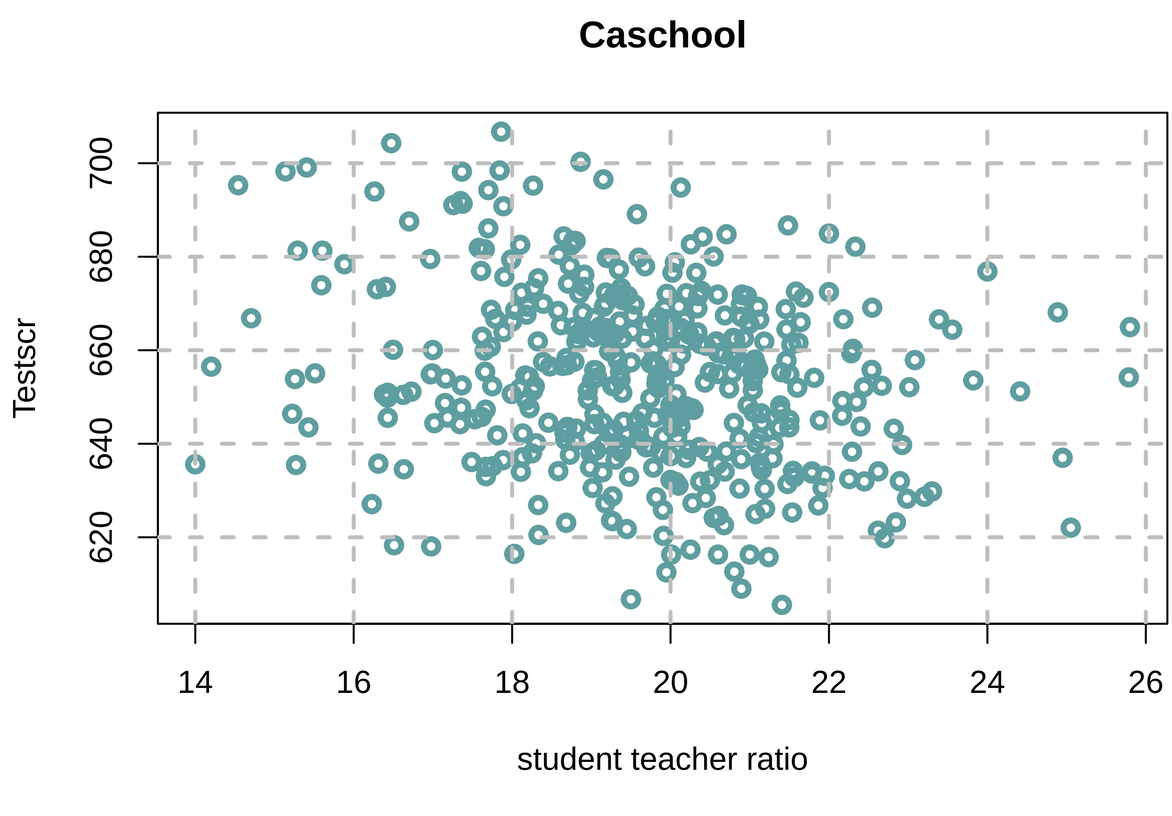
dist_cod	testscr	str
75119	690.80	17.88991
61499	661.20	21.52466
61549	643.60	18.69723
61457	647.70	17.35714
61523	640.85	18.67133
62042	605.55	21.40625

El modelo ajustado es

$$
\begin{aligned}
\widehat{testscr}_i &= \hat{\beta}_0 + \hat{\beta}_1 str_i \\
i &= 1, 2, ..., n
\end{aligned}
$$

Visualización de los datos

Una visualización previa de los datos



La regresión del modelo es

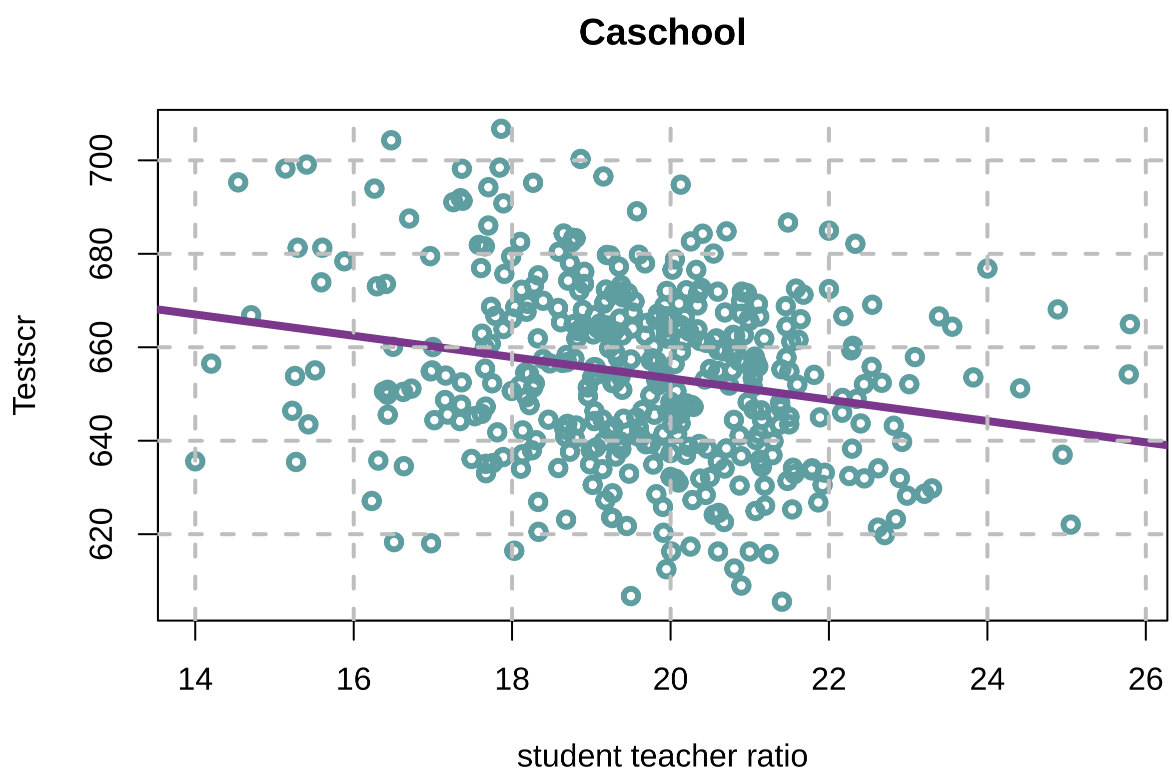
```
##
## Call:
## lm(formula = testscr ~ str, data = caschool)
##
## Coefficients:
## (Intercept)          str
##      698.93        -2.28
```

Modelo ajustado

El modelo ajustado es

$$\hat{testscr_i} = 698.93 - 2.28 str_i$$
$$i = 1, 2, ..., n$$

De tal forma



Inferencias respecto a los parámetros estimados

El tamaño de la muestra es:

```
## [1] 420
```

Recordemos que la varianza de los estimadores se estima por:

$$
\begin{aligned}
V[\hat{\beta}_0] &= c_{00}\sigma^2 \\
c_{00} &= \frac{\sum_{i=1}^{n} x_i}{nS_{xx}} \\
V[\hat{\beta}_1] &= c_{11}\sigma^2 \\
c_{11} &= \frac{1}{S_{xx}}
\end{aligned}
$$

Para la prueba de hipótesis

$$H_0 : \beta_i = \beta_{i0}$$

Se usa el estadístico de prueba

$$Z = \frac{\hat{\beta}_i - \beta_{i0}}{\sigma\sqrt{c_{ii}}}$$

Que a su vez

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{S\sqrt{c_{ii}}}$$

Ahora queremos probar

$$H_0 : \beta_1 = 0 \;\; \text{contra} \;\; H_a : \beta_1 \neq 0$$

Que tiene una distribución t de Student con n-2 grados de libertad

Recordemos que un estimador insesgado para σ^2 es S^2

$$S^2 = \frac{SRC}{n-2}$$

Recordemos que la suma total de cuadrados "STC"

$$STC = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

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

Recordemos que la suma explicada de cuadrados "SEC"

$$SEC = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

```
## [1] 7794.11
```

Recordemos que la suma de residuales cuadrados "SRC"

$$SRC = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

```
## [1] 144315.5
```

De tal modo nuestro estimador S

```
## [1] 18.58097
```

Calculando S_{xx}

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

```
## [1] 1499.581
```

Ahora para c_{00}

```
## [1] 0.259617
```

Ahora para c_{11}

```
## [1] 0.000666853
```

Mas adelante se vera que los coeficientes c_{ii} se obtienen de $(X^t X)^{-1}$

Volviendo a la prueba de hipótesis

Para el estimador β_1 el estadístico T es:

```
##       str
## -4.751327
```

Notemos ademas que el valor p:

$$
\begin{aligned}
&p = 2P(t > t_{estimado}) \\
&p = 2P(t > -4.75) \;\; \text{es una t con 420-2=118 grados de libertad}
\end{aligned}
$$

Cuyo valor p es:

```
##          str
## 2.020858e-06
```

Si probamos

$$H_0 : \beta_0 = 0 \;\; \text{contra} \;\; H_a : \beta_0 \neq 0$$

Nuestro estadistico t seria:

```
## (Intercept)
##    73.82451
```

Cuyo valor p es:

```
## (Intercept)
##           0
```

Por linea de comando

```
##
## Call:
## lm(formula = testscr ~ str, data = caschool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.727 -14.251   0.483  12.822  48.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
## str          -2.2798     0.4798  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.58 on 418 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
## F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Intervalo de confianza para los parámetros estimados

Un intervalo de confianza al nivel 95% para β_i

$$\beta_i \pm t_{\alpha/2} S \sqrt{c_{ii}}$$

Para β_0

```
##                  inf      sup
## (Intercept) 680.3767 717.4892
```

Para β_1

```
##           inf      sup
## str -3.220266 -1.33935
```

Por linea de comando

```
##                 2.5 %     97.5 %
## (Intercept) 680.32313 717.542779
## str          -3.22298  -1.336637
```

Aplicación de forma matricial

Desarrollando en forma matricial se tiene que la matrix $A = (X^t X)^{-1}$

```
y_value <- as.matrix(caschool$testscr);
n_m <- length(y_value);
X_matrix <- matrix(c(rep(1,n),caschool$str),nrow = n_m);
# K <- ncol(X);
A_matrix <- solve(t(X_matrix) %*% X_matrix);A_matrix# A=(X^t X)^-1
```

```
##              [,1]         [,2]
## [1,]  0.25961704 -0.013097277
## [2,] -0.01309728  0.000666853
```

Nuestro estimador $S^2 \sim \sigma^2$ es:

```
estimador_s^{2}
```

```
## [1] 345.2524
```

La matrix $(X^t X)^{-1}$:

```
A_matrix
```

```
##              [,1]         [,2]
## [1,]  0.25961704 -0.013097277
## [2,] -0.01309728  0.000666853
```

Nuestros estimadores son:

```
ols_beta <- A_matrix %*% (t(X_matrix) %*% y_value) ;ols_beta
```

```
##             [,1]
## [1,] 698.932952
## [2,]  -2.279808
```

La matrix de varianza covarianza $S^2(X^t X)^{-1}$

```
cov_var <- estimador_s^{2} * A_matrix ;cov_var
```

```
##            [,1]       [,2]
## [1,] 89.633394 -4.5218657
## [2,] -4.521866  0.2302326
```

Que por linea de comando

```
##             (Intercept)        str
## (Intercept)   89.633394 -4.5218657
## str           -4.521866  0.2302326
```

Creación de variables Dummy

Una variable binaria (que toma o transforma información en una o mas categorias) se denomina asimismo variable indicador o a veces variable ficticia o variable dummy.

Veamos las primeras 30 obervaciones haciendo:

$$D_i = \begin{cases} 1 & \text{Si str del distrito i-ésimo es menor a 20} \\ 0 & \text{Si str del distrito i-ésimo es mayor o igual a 20} \end{cases}$$

testscr	str	dummy
690.80	17.88991	1
661.20	21.52466	0
643.60	18.69723	1
647.70	17.35714	1
640.85	18.67133	1
605.55	21.40625	0
606.75	19.50000	1
609.00	20.89412	0
612.50	19.94737	1
612.65	20.80556	0
615.75	21.23809	0
616.30	21.00000	0
616.30	20.60000	0
616.30	20.00822	0
616.45	18.02778	1
617.35	20.25196	0
618.05	16.97787	1
618.30	16.50980	1
619.80	22.70402	0
620.30	19.91111	1

El modelo de regresión poblacional con D_i

$$testscr_i = \beta_0 + \beta_1 D_i$$
$$i = 1, 2, 3, ..., n$$

Si la variale str es alta, entonces $D_i = 0$. De tal forma la ecuación se reduce a

$$testscr_i = \beta_0 + \epsilon_i$$

Pero si la variale str es baja, entonces $D_i = 1$. De tal forma la ecuación se reduce a

$$testscr_i = \beta_0 + \beta_1 + \epsilon_i$$

Regresión con variables dummy

Si realizamos una regresión al modelo, se obtiene

```
##
## Call:
## lm(formula = testscr ~ dummy, data = df_dummy)
##
## Coefficients:
## (Intercept)        dummy
##     649.979        7.372
```

El modelo ajustado es:

$$\hat{testscr_i} = 649.979 + 7.372 D_i$$

Y un resumen general

```
##
## Call:
## lm(formula = testscr ~ dummy, data = df_dummy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.601 -14.047  -0.451  12.841  49.399
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  649.979      1.388 468.380  < 2e-16 ***
## dummy          7.372      1.843   3.999 7.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.72 on 418 degrees of freedom
## Multiple R-squared:  0.03685,    Adjusted R-squared:  0.03455
## F-statistic: 15.99 on 1 and 418 DF,  p-value: 7.515e-05
```

Y un intervalo de confianza para los estimadores

```
##                  2.5 %    97.5 %
## (Intercept) 647.251075 652.70662
## dummy         3.748774  10.99605
```