Caschool

Joel Alejandro Zavala Prieto

Contents

Informacion 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 informacion 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

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

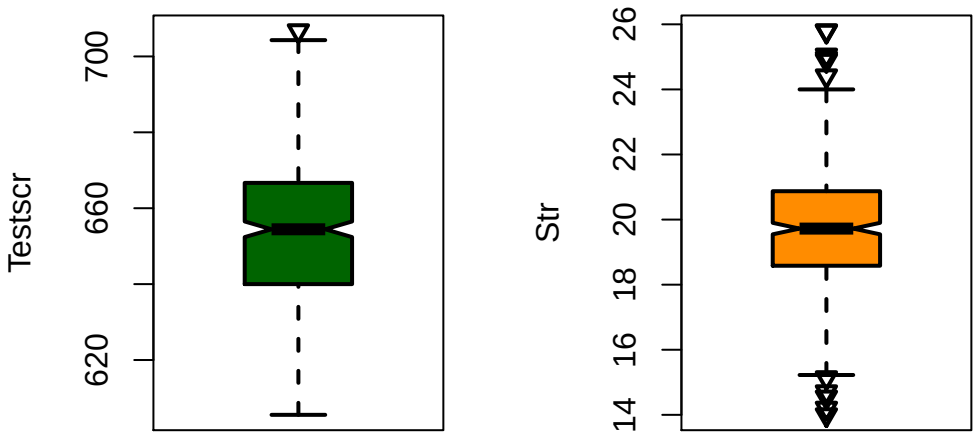
dist_cod	testscr	str
75119	690.80	17.88991
61499	661.20	21.52466
61549	643.60	18.69723
61457	647.70	17.35714
61523	640.85	18.67133
62042	605.55	21.40625
68536	606.75	19.50000
63834	609.00	20.89412
62331	612.50	19.94737
67306	612.65	20.80556
65722	615.75	21.23809
62174	616.30	21.00000
71795	616.30	20.60000
72181	616.30	20.00822
72298	616.45	18.02778
72041	617.35	20.25196
63594	618.05	16.97787
63370	618.30	16.50980
64709	619.80	22.70402
63560	620.30	19.91111

El modelo ajustado es

$$\hat{testscr}_i = \hat{\beta}_0 + \hat{\beta}_1 str_i$$
$$i = 1, 2, ..., n$$

Cuantiles y percentiles de las variables

Veamos sus diagramas de caja para las variables a analizar

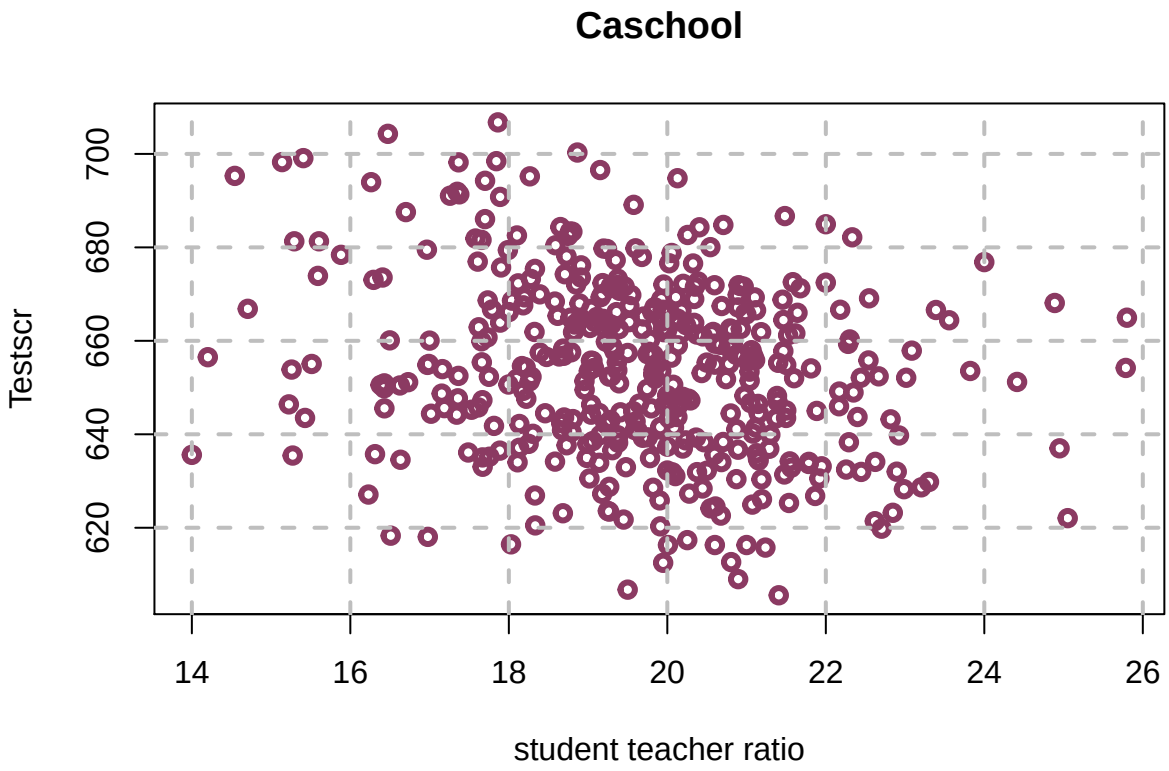


Veamos algunos de sus cuantiles y percentiles mas comunes

	percentil.10.	percentil.25.	percentil.40.	percentil.50.	percentil.60.	percentil.75.	percentil.90.
quantil_str	17.3486	18.58236	19.26618	19.72321	20.0783	20.87181	21.8674
quantil_testscr	630.3950	640.04999	649.07000	654.44998	659.4000	666.66251	678.8600

Visualización de los datos

Una visualización previa de los datos



La regresión del modelo es

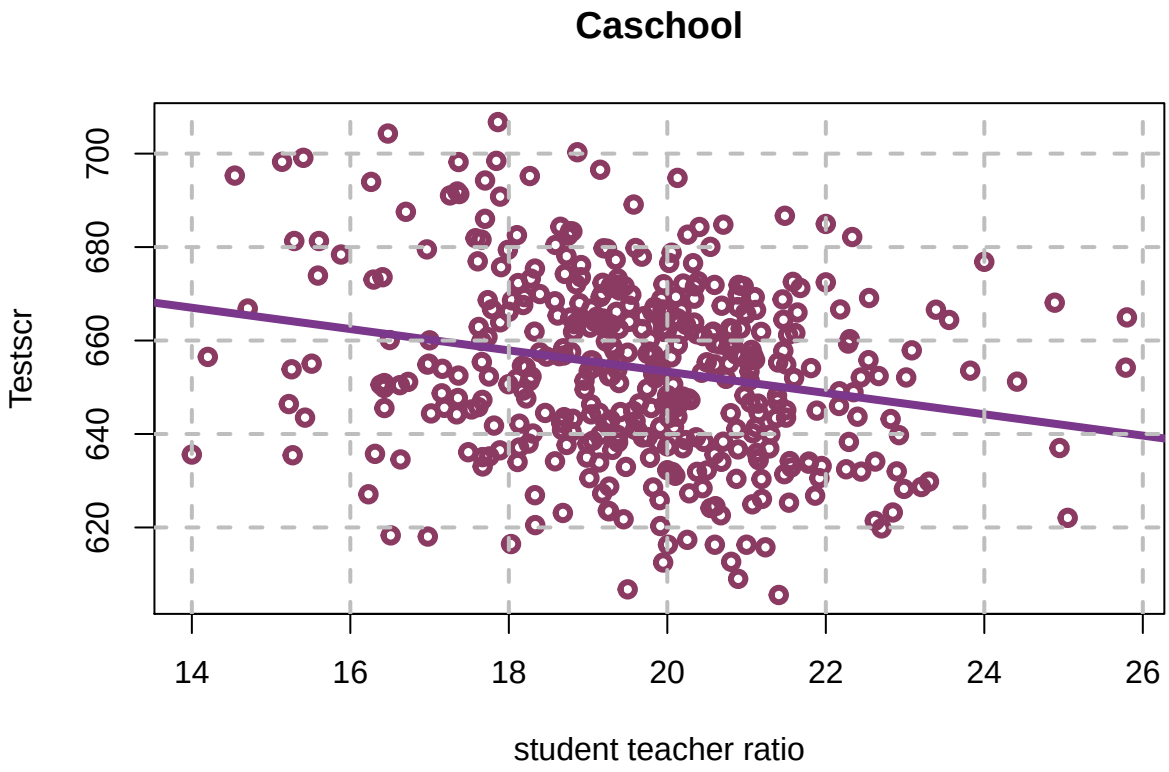
```
##
## Call:
## lm(formula = testscr ~ str, data = caschool)
##
## Coefficients:
## (Intercept)          str
##      698.93        -2.28
```

Modelo ajustado

El modelo ajustado es

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

De tal forma



Predicción

Se hacen unas estimaciones para valores de **str**

```
nuevas.str <- data.frame(str = c(17,18,19))
predict(ols_caschool, nuevas.str)
```

```
##        1        2        3
## 660.1762 657.8964 655.6166
```

Medidas estadisticas extras

Cantidad de datos totales

```
n <- length (caschool$str); n
```

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

Suma total de cuadrados

```
stc <- sum((caschool$testscr - mean(caschool$testscr))^2); stc
```

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

Suma explicada de cuadrados

```
sec <- sum((ols_ajustados - mean(caschool$testscr))^2); sec
```

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

Suma de residuales al cuadrado

```
errores <- caschool$testscr - ols_ajustados
src <- sum(errores^2); src
```

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

Error estándar de la regresión

```
esr <- sqrt (src/(n-2)) ; esr
```

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

Coeficiente de determinación R^2

```
R_2 <- sec/stc; R_2
```

```
## [1] 0.0512401
```

Lo que nos dice este modelo es que ajusta los datos en un 5.124%

Por linea de comando

```
summary(ols_caschool)$sigma
```

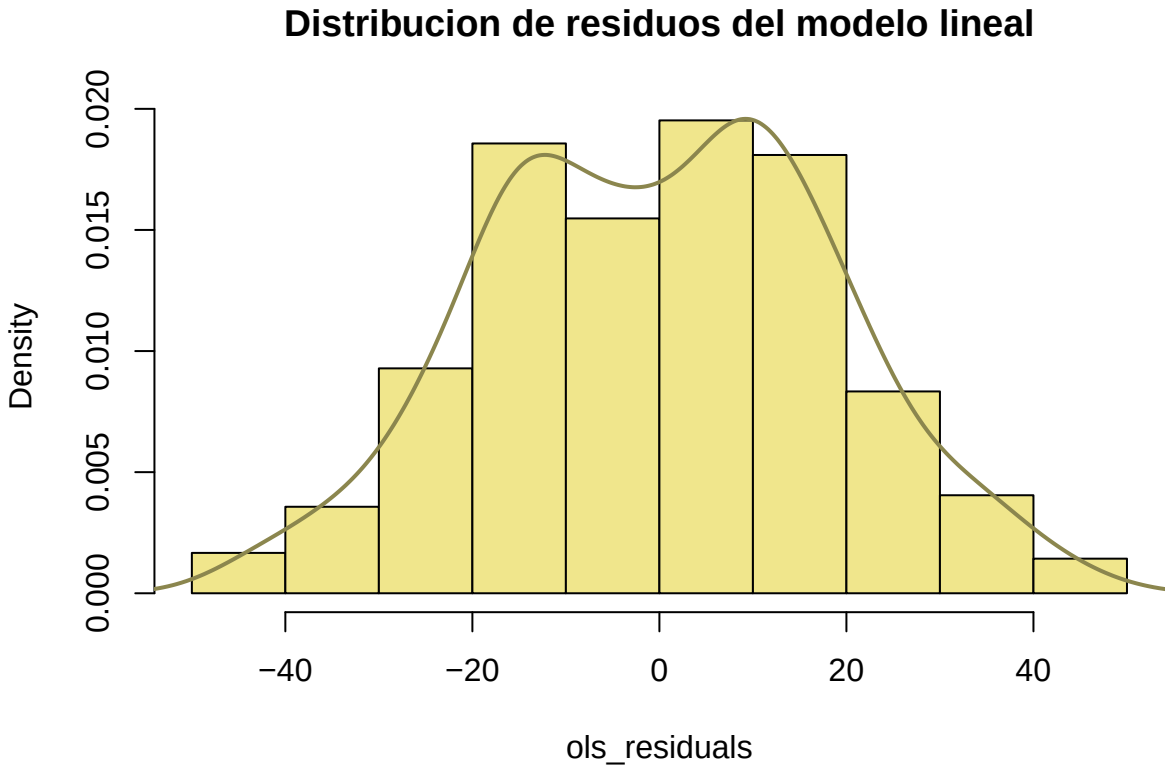
```
## [1] 18.58097
```

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

```
## [1] 0.0512401
```

Distribución de los residuales

Observando la distribución de los errores para estos datos



Resumen general

Testscr	Testscr_Ajustados	Residuales
690.80	658.1474	32.65260
661.20	649.8608	11.33917
643.60	656.3069	-12.70689
647.70	659.3620	-11.66198
640.85	656.3659	-15.51592
605.55	650.1308	-44.58076
606.75	654.4767	-47.72669
609.00	651.2984	-42.29837
612.50	653.4568	-40.95678
612.65	651.5003	-38.85025
615.75	650.5142	-34.76417
616.30	651.0570	-34.75699
616.30	651.9689	-35.66891
616.30	653.3181	-37.01807
616.45	657.8331	-41.38306