Estudio de análisis de regresión

Rocío Vecino Torres

29/12/2021

Estudio de análisis de regresión

Vamos a trabajar con un Excel, llamado StudentsPerformance, que contendrá datos de 1.000 observaciones y con 8 variables que tendrá la información de: genero, raza, nivel de estudios de los padres, si comieron, si hicieron el exámen de preparación, y los resultados de los exámenes en Writting, Reading y Matemáticas.

- Género: tiene valor de mujer o masculino.
- Raza: tiene valores de Grupo A, B, C, D, E.
- Nivel de educación de los padres: bacherlor´s degree (licenciatura), some college cd (algún grado), master´s degree (maestría), associate´s degree (grado asociado), high school (instituo), some high school.
- Comida: completa o reducida.
- Test de preparación del exámen: completa o ninguna.
- Resultados en los examenes de Writting, Reding y Matemáticas: Valores numéricos del 1 al 100.

Para comenzar, cargamos las librerias que serán necesarias.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dplyr)
```

Y cargamos los datos del fichero .csv:

```r
#En un primer lugar cargaremos los datos:
students_data <- read.csv("StudentsPerformancePreprocesado.csv", sep = ',', head = TRUE)
students_data$id <- NULL
colnames(students_data)
```

```
## [1] "gender"                       "race.ethnicity"
## [3] "parental.level.of.education" "lunch"
## [5] "test.preparation.course"      "math.score"
## [7] "reading.score"                "writing.score"
```

Estudio 1: Regresión multivariable.

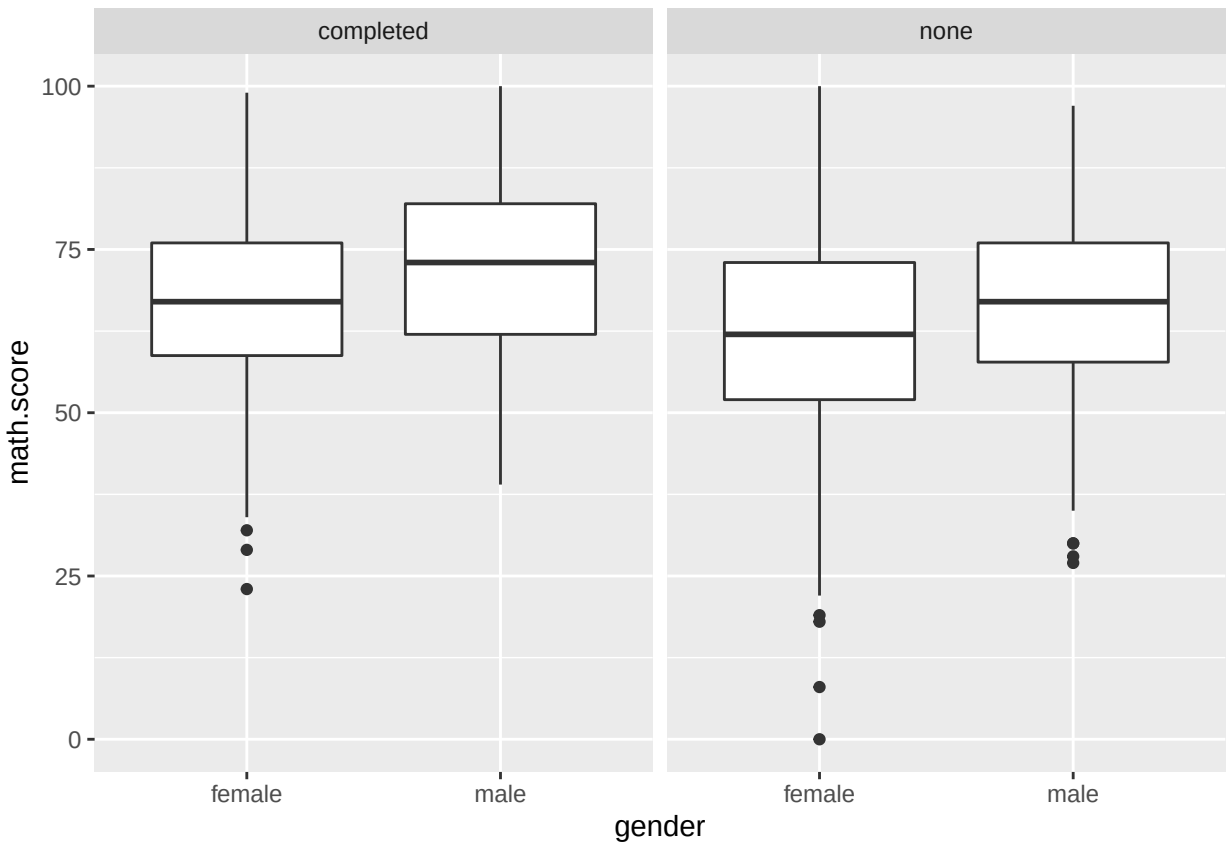
Para realizar la regresión multivariable, se deben de realizar los siguientes puntos:

1. Dibujar los resultados de los exámenes de Matemáticas, Reading y Writing en función del resto de variables para observar si hay una relación lineal.

Resultados de los exámenes de matemáticas:

Se estudia la relación entre los resultados del examen de matemáticas y dos variables: el género del estudiante y la preparación para el examen.

```r
#relación entre los resultados del examen de matemáticas y dos variables: el género del estudiante y la
ggplot(students_data, aes(x = gender, y = math.score)) + geom_boxplot() + facet_grid(~test.preparation.
```

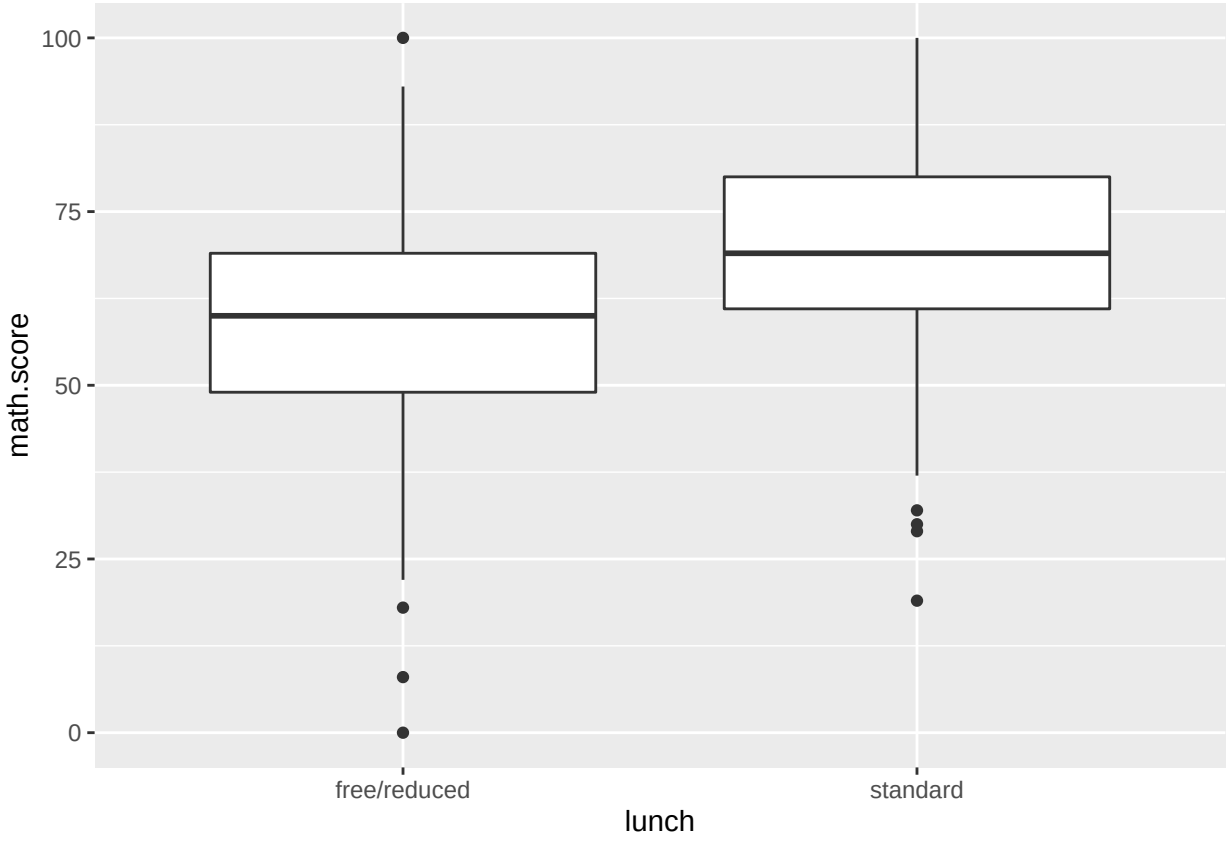


Como resultado, se puede ver que:

- Los hombres tienen un puntuaje un poco mayor que las mujeres de media. Los hombres tienen menos valores atípicos, sus datos difieren menos entre ellos.
- El curso de preparación ayudo a los estudiantes a lograr mejores resultados. Obteniendo un promedio más alto y valores menos atípicos.

Se estudia la relación entre los resultados del examen de matemáticas y lo que han comido los estudiantes.

```
#relación entre los resultados del examen de matemáticas y la comida del estudiante
ggplot(students_data, aes(x = lunch, y = math.score))+ geom_boxplot()
```

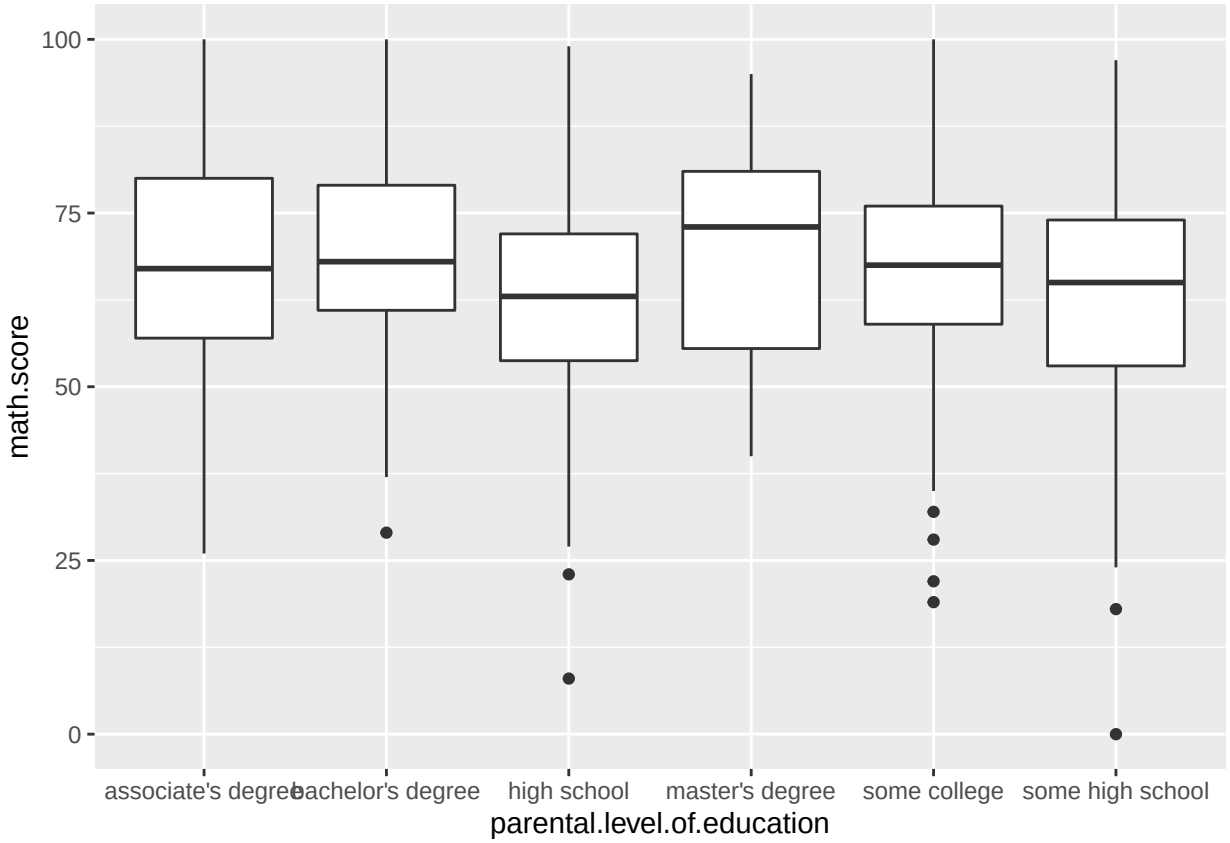


Como resultado de la gráfica, se observa que:

- Los estudiantes que comieron una comida standard obtuvieron mayor nota que los que comieron una comida reducida. Aunque, ambas tienen valores atípicos.

Se estudia la relación entre los resultados del examen de matemáticas y el nivel academico de los padres de los estudiantes.

```
#relación entre los resultados del examen de matemáticas y el nivel de estudios de los padres
ggplot(students_data, aes(x = parental.level.of.education, y = math.score))+ geom_boxplot()
```



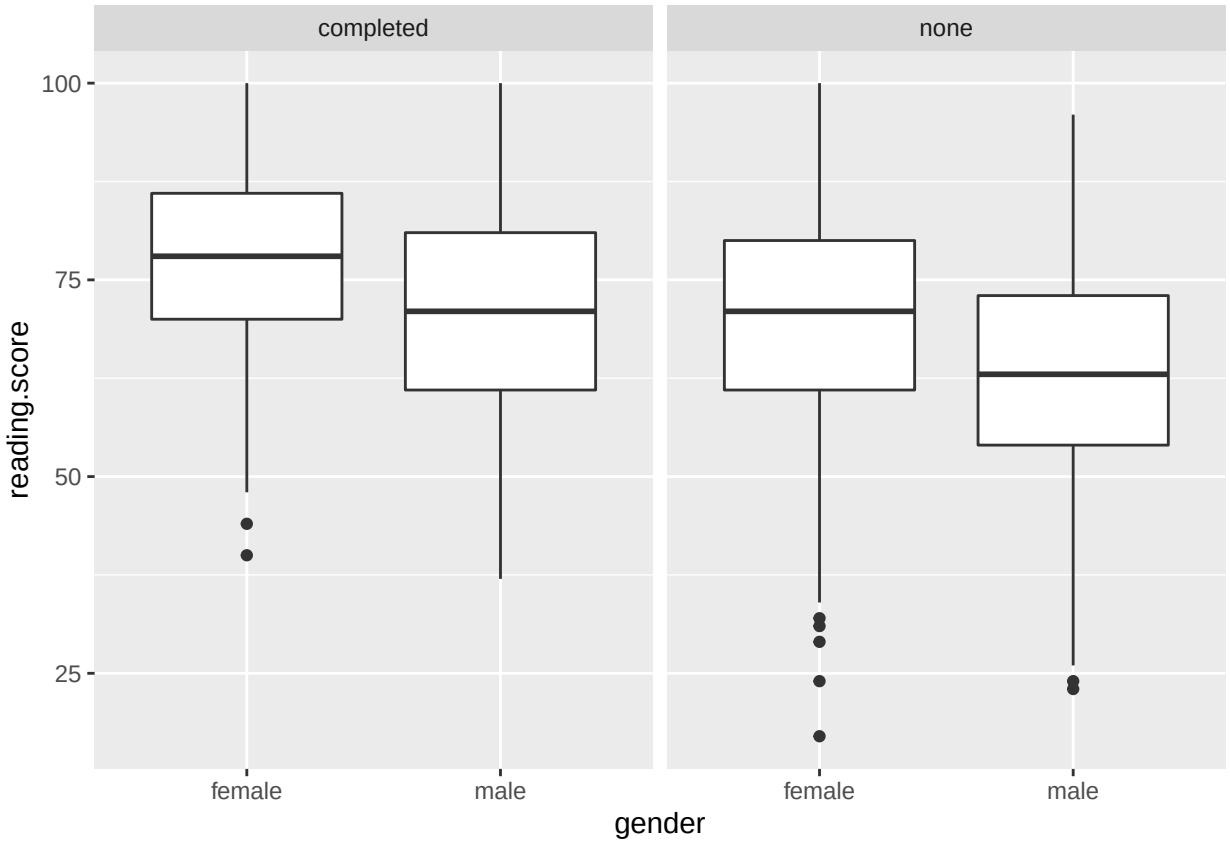
Como resultado, se puede ver lo siguiente:

- Los estudiantes cuyos padres tienen un nivel de educación de maestría (Master´s degree) tiene un promedio un poco más alto con respecto a los demás. El promedio de los que tiene un nivel de estudios de licenciado o de grado (associate´s degree, bachelor´s degree, some college) están muy igualados entre sí. Tienen más bajo promedio aquellos cuyo nivel educativo de los padres es de instituto (high school). Se puede ver también que los datos del nivel educativo de los padres de Master y grado asociado (associate´s degree) no tienen valores atípicos.

Resultados de los exámenes de Lectura (Reading):

Se estudia la relación entre los resultados del examen de lectura y dos variables: el género del estudiante y la preparación para el exámen.

```
#relación entre los resultados del examen de lectura y dos variables: el género del estudiante y la prep
ggplot(students_data, aes(x = gender, y = reading.score)) + geom_boxplot() + facet_grid(~test.preparatio
```

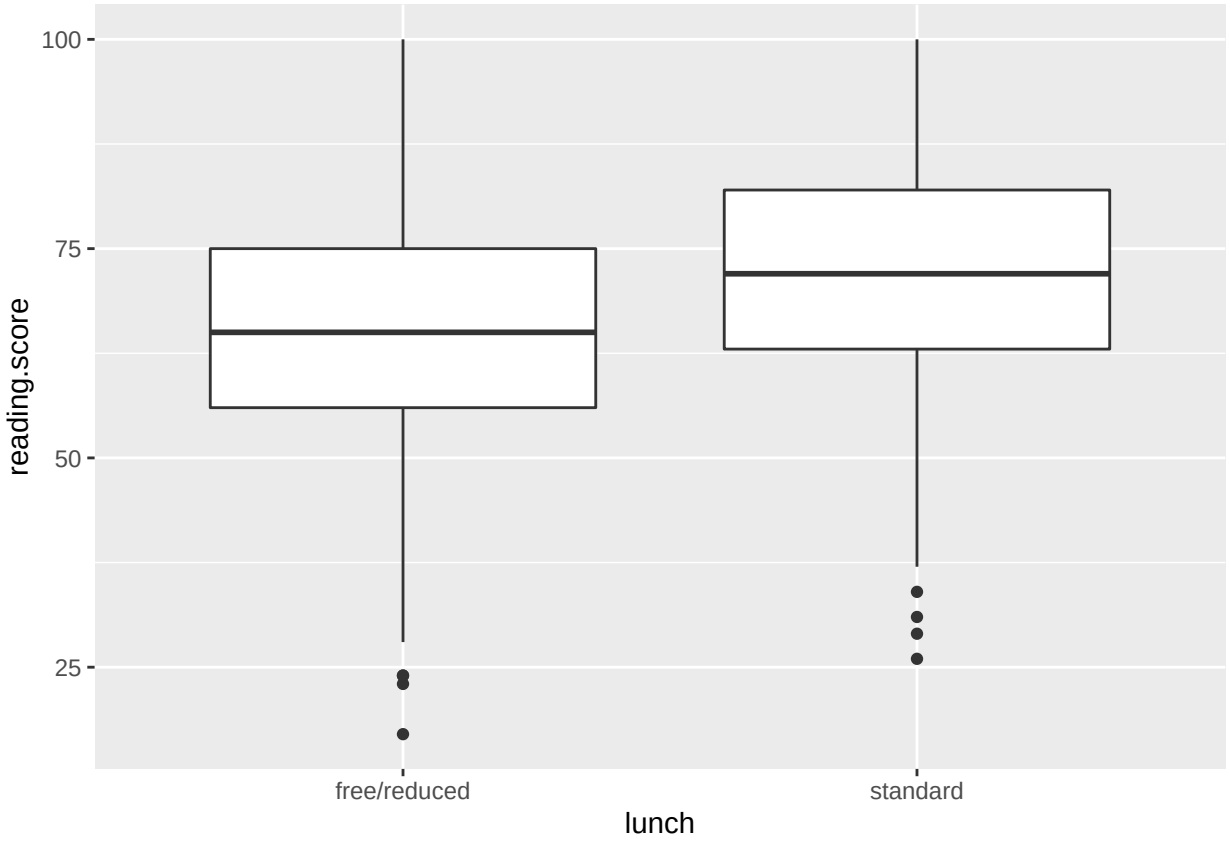


Como resultado, se observa lo siguiente:

- Las mujeres tienen un puntuaje de lectura más alto en promedio, pero tiene más valores atípicos.
- Una vez más el curso de preparación hace que los estudiantes obtengas mejores resultados.

Se estudia la relación entre los resultados del examen de lectura y lo que han comido los estudiantes.

```
#relación entre los resultados del examen de lectura y la comida del estudiante
ggplot(students_data, aes(x = lunch, y = reading.score))+ geom_boxplot()
```

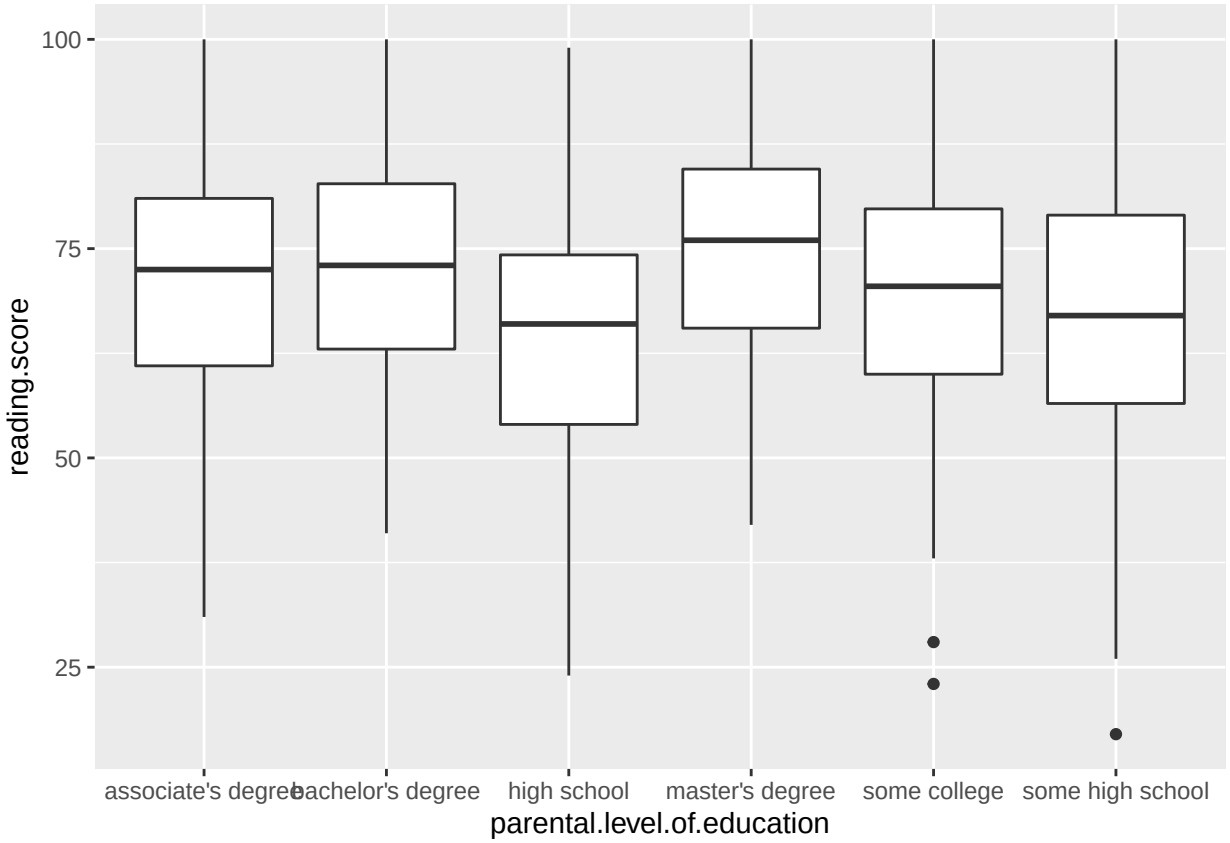


Como resultado, se observa lo siguiente:

- Nuevamente, los estudiantes que comieron una comida standard obtuvieron mejores resultados en el exámen.

Se estudia la relación entre los resultados del examen de lectura y el nivel academico de los padres de los estudiantes.

```
#relación entre los resultados del examen de lectura y el nivel de estudios de los padres
ggplot(students_data, aes(x = parental.level.of.education, y = reading.score))+ geom_boxplot()
```



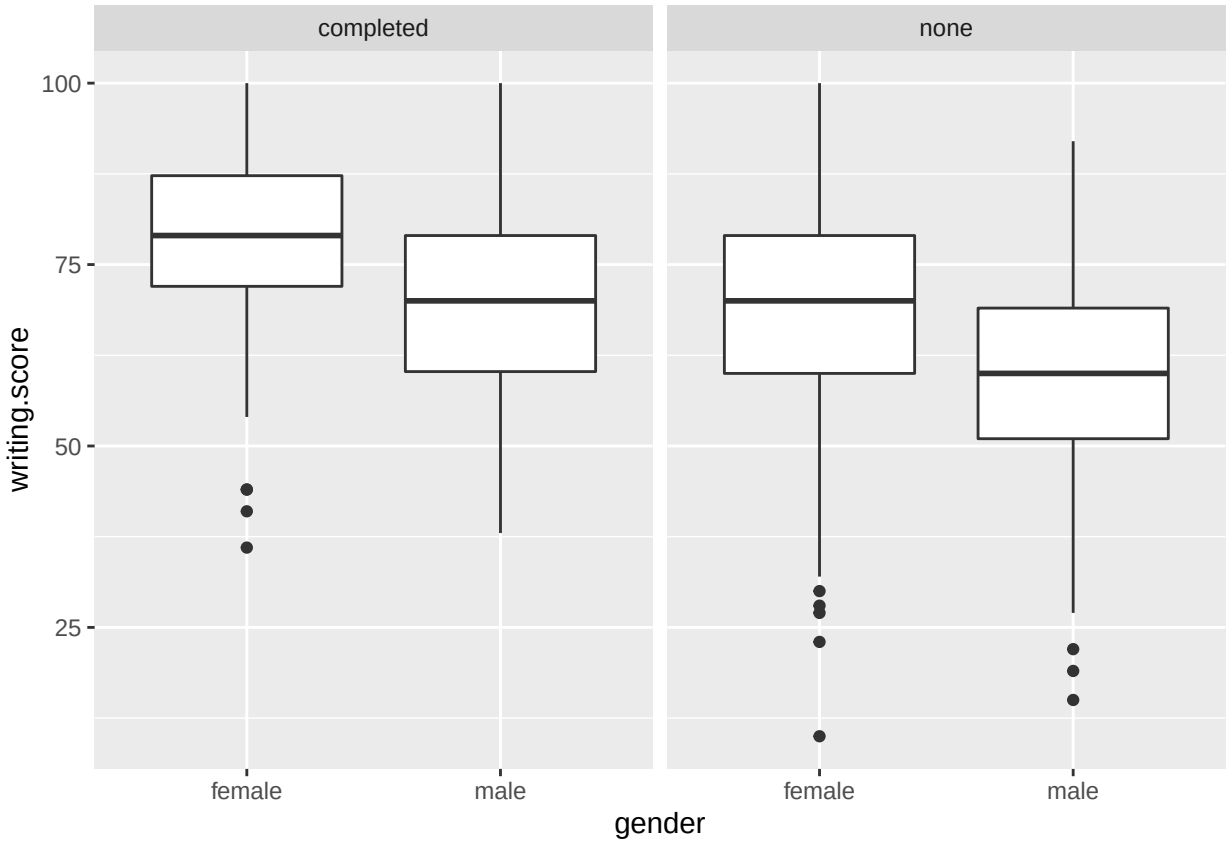
Como resultado, se puede ver lo siguiente:

- Se observa aproximádamente lo mismo que en los resultados de matemáticas, con la diferencia de que hay menos valores atípicos.

Resultados de los exámenes de Escritura (Writing):

Se estudia la relación entre los resultados del examen de escritura y dos variables: el género del estudiante y la preparación para el exámen.

```
#relación entre los resultados del examen de escritura y dos variables: el género del estudiante y la p
ggplot(students_data, aes(x = gender, y = writing.score)) + geom_boxplot() + facet_grid(~test.preparati
```

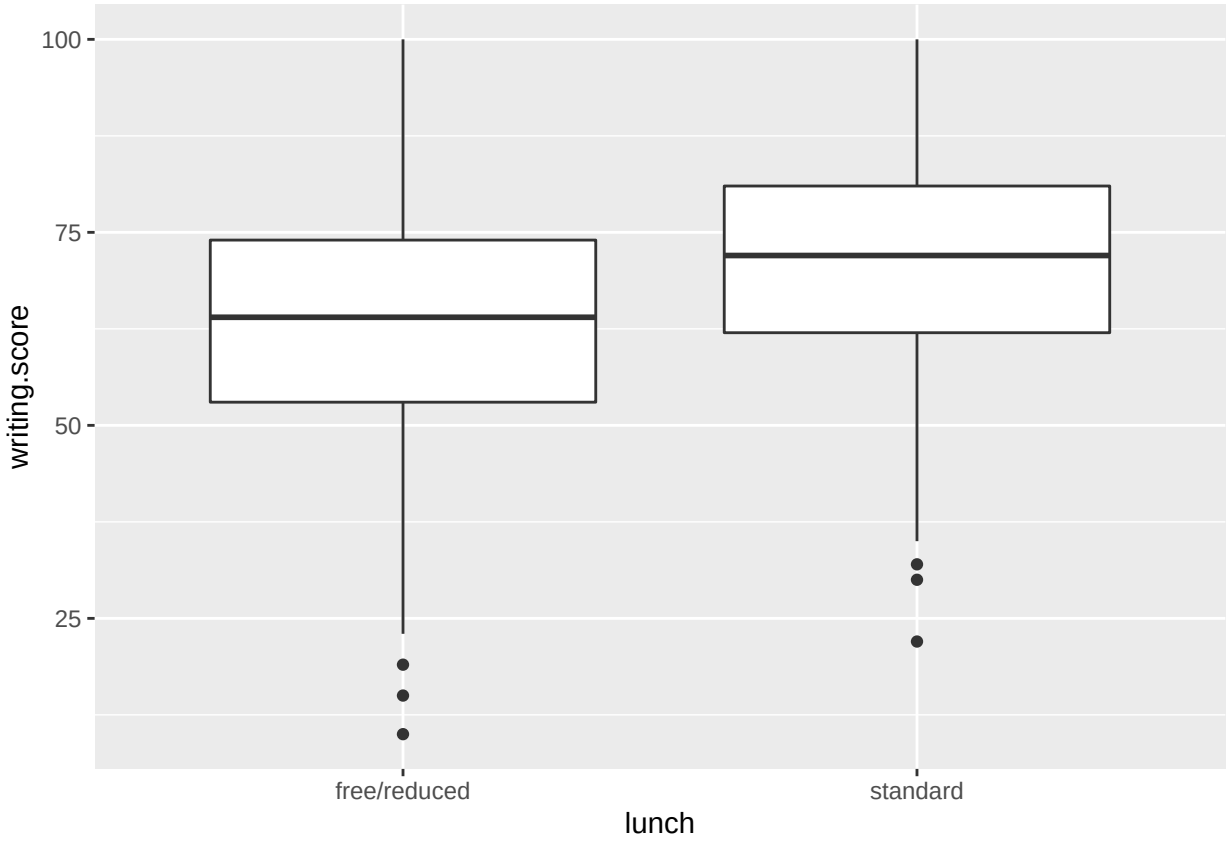


Como resultado, se observa que:

- Las mujeres tienen un puntuaje de escritura más alto en promedio, pero tiene más valores atípicos.
- Una vez más el curso de preparación hace que los estudiantes obtengas mejores resultados.

Se estudia la relación entre los resultados del examen de escritura y lo que han comido los estudiantes.

```
#relación entre los resultados del examen de escritura y la comida del estudiante
ggplot(students_data, aes(x = lunch, y = writing.score))+ geom_boxplot()
```

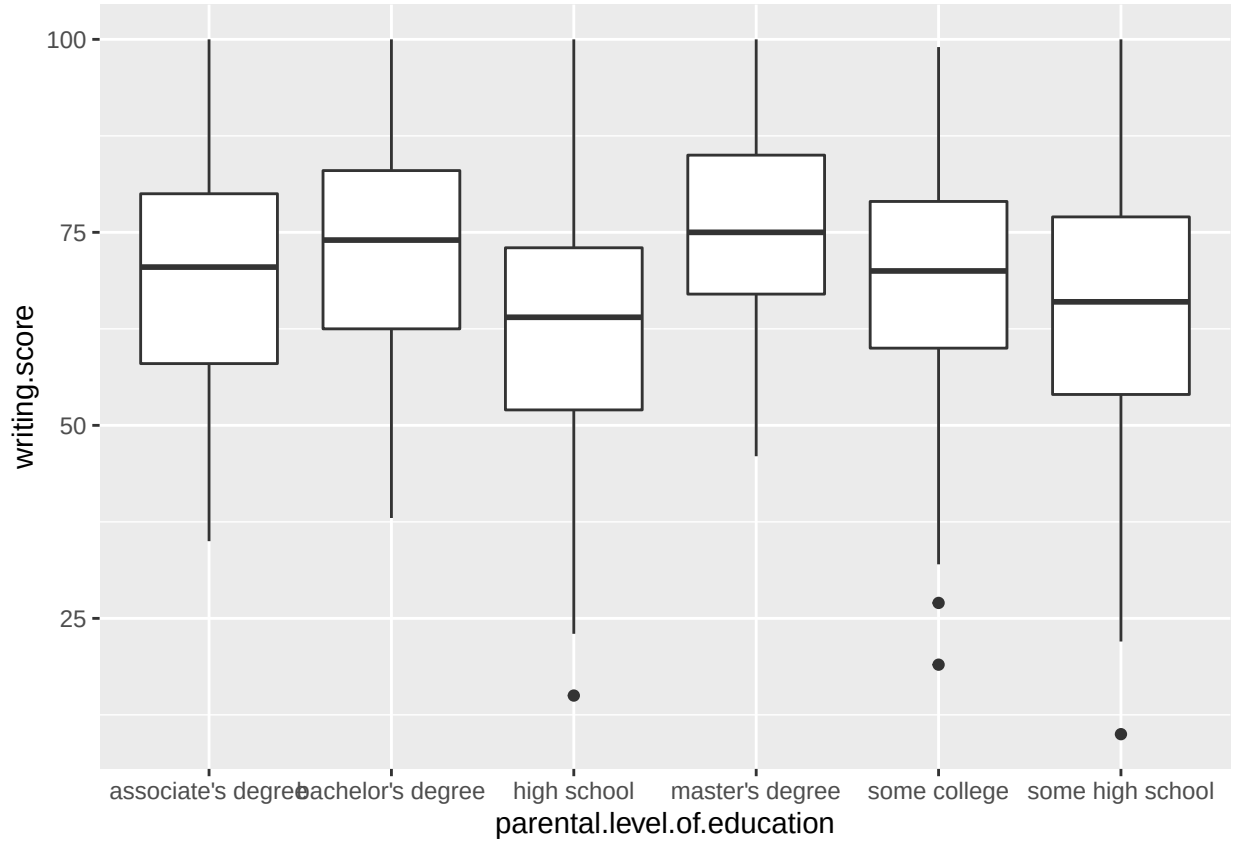


Como resultado, se puede ver que:

- Nuevamente, los estudiantes que comieron una comida standard obtuvieron mejores resultados en el exámen.

Se estudia la relación entre los resultados del examen de escritura y lo que han comido los estudiantes.

```
#relación entre los resultados del examen de escritura y el nivel de estudios de los padres
ggplot(students_data, aes(x = parental.level.of.education, y = writing.score))+ geom_boxplot()
```



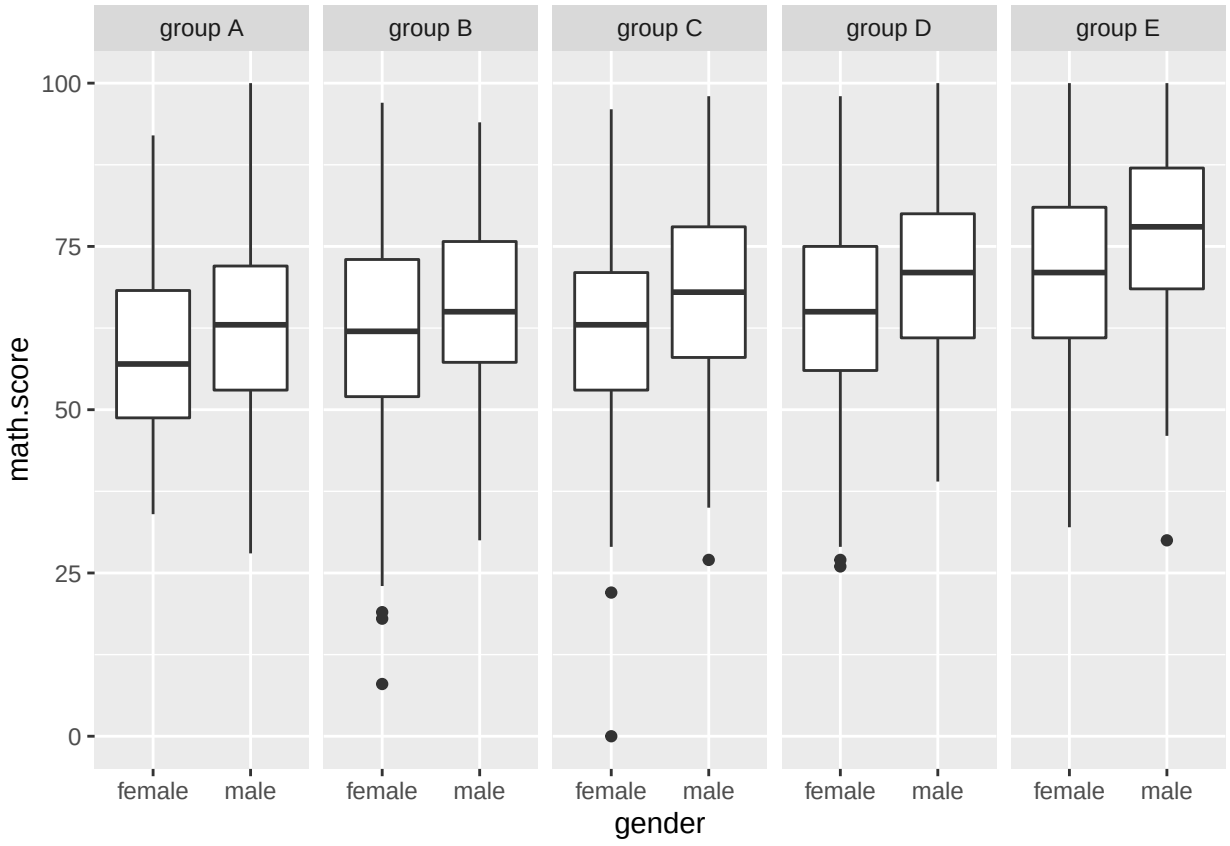
Como resultado se puede ver que:

- Se observa aproximadamente lo mismo que en los resultados de matemáticas y reading, con la diferencia de que hay menos valores atípicos. Pero en general en los tres, se ve que a mayor nivel educativo de los padres mayor puntuaje en las notas de los tres exámenes.

Ahora, realizamos una comparación de los resultados de los examenes (Escritura, Lectura y Matemáticas) en función del género y la raza/etnia.

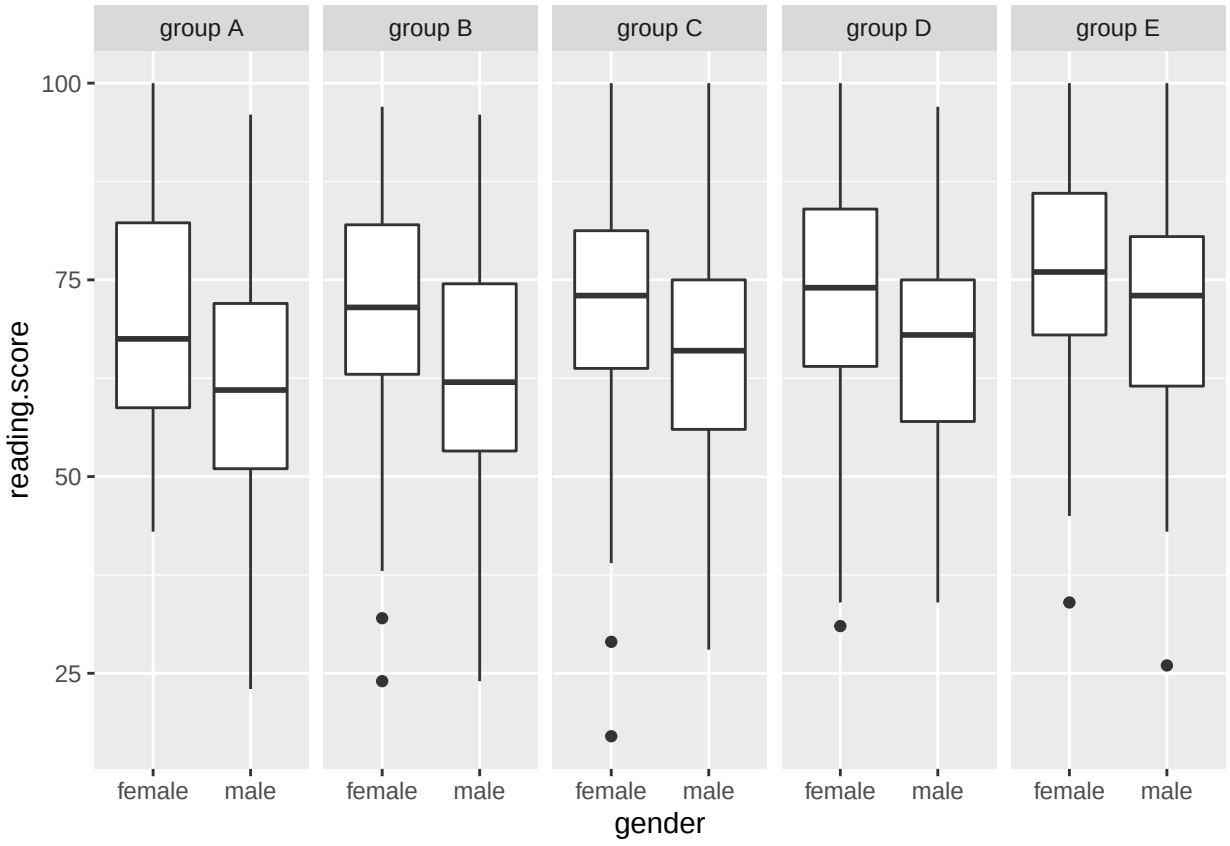
Exámenes de Matemáticas: Los estudiantes del grupo E son los que obtienen mejores resultados. El grupo A es el que tiene peor resultado en el examen. Se sigue viendo que los hombres tienen mejores resultados en el examen de matemáticas.

```
ggplot(students_data, aes(x = gender, y = math.score)) + geom_boxplot() + facet_grid(~race.ethnicity)
```



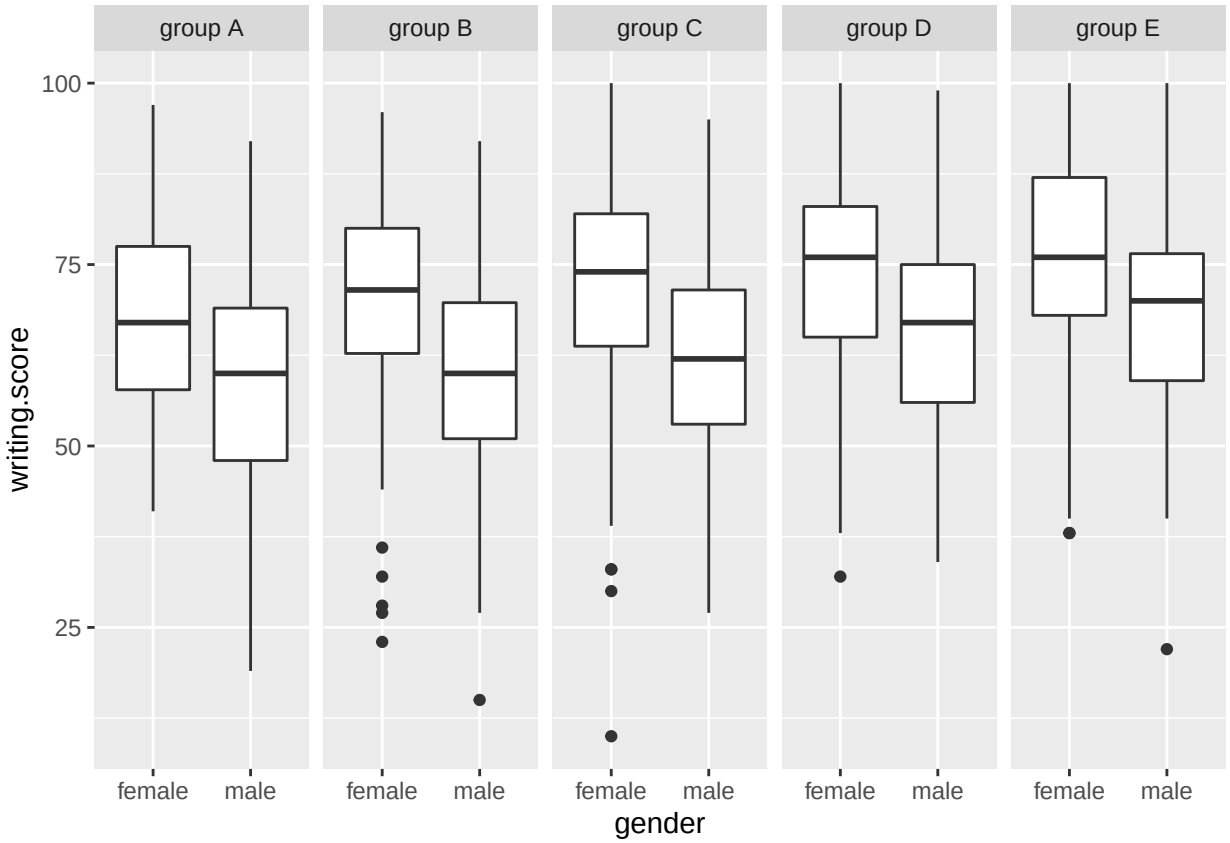
Exámenes de Lectura (Reading): Nuevamente, se ve que el grupo E obtiene mejores resultados. Pero no hay tanta diferencia entre las notas de los alumnos como ocurrió con los exámenes de matemáticas. Lo que sí se puede ver significativamente que las mujeres (como se vio anteriormente) que las mujeres obtienen mejores resultados en los exámenes de Lectura.

```
ggplot(students_data, aes(x = gender, y = reading.score)) + geom_boxplot() + facet_grid(~race.ethnicity
```



Exámenes de Escritura (Writing): Se observa lo mismo que ha ocurrido en la anterior gráfica. Los estudiantes del grupo E son los que obtienen mejores resultados en escritura.

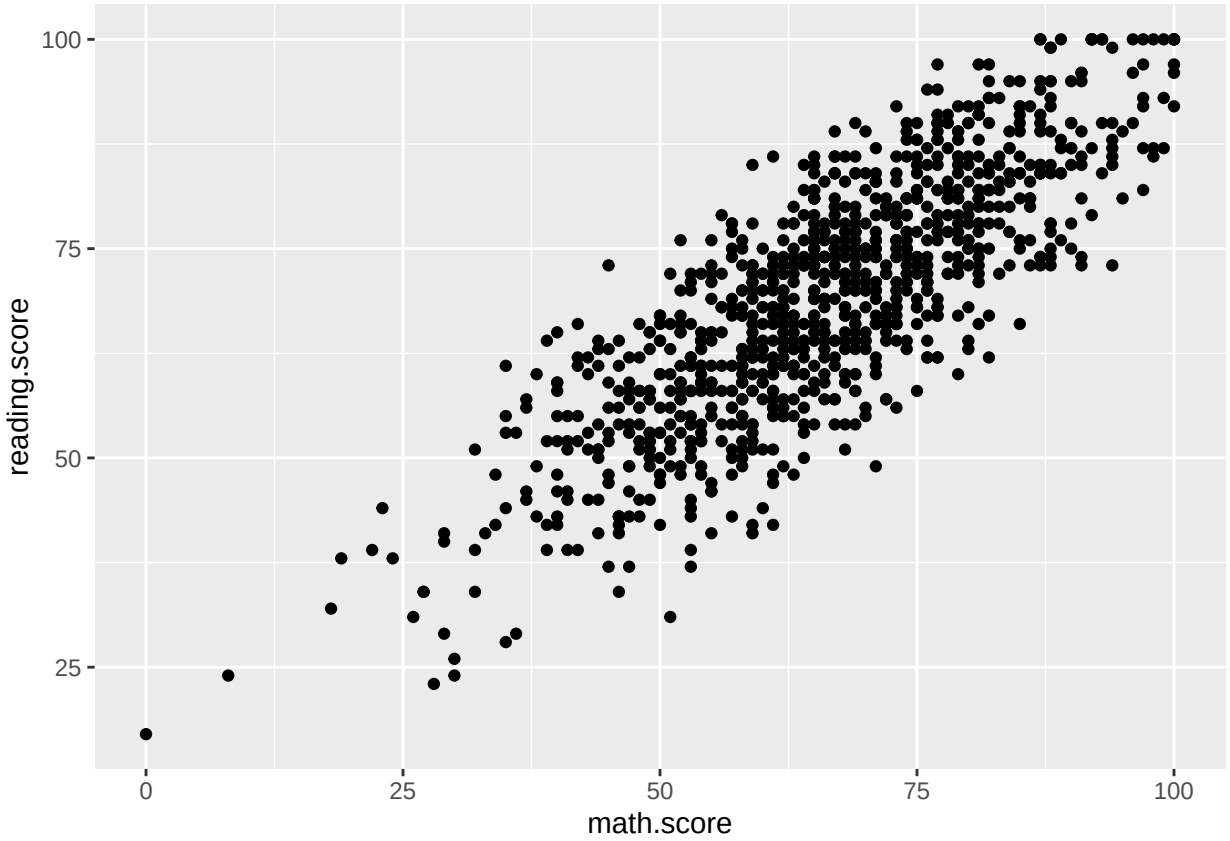
```
ggplot(students_data, aes(x = gender, y = writing.score)) + geom_boxplot() + facet_grid(~race.ethnicity
```



Por último, se hace una comparación entre los resultados de los exámenes:

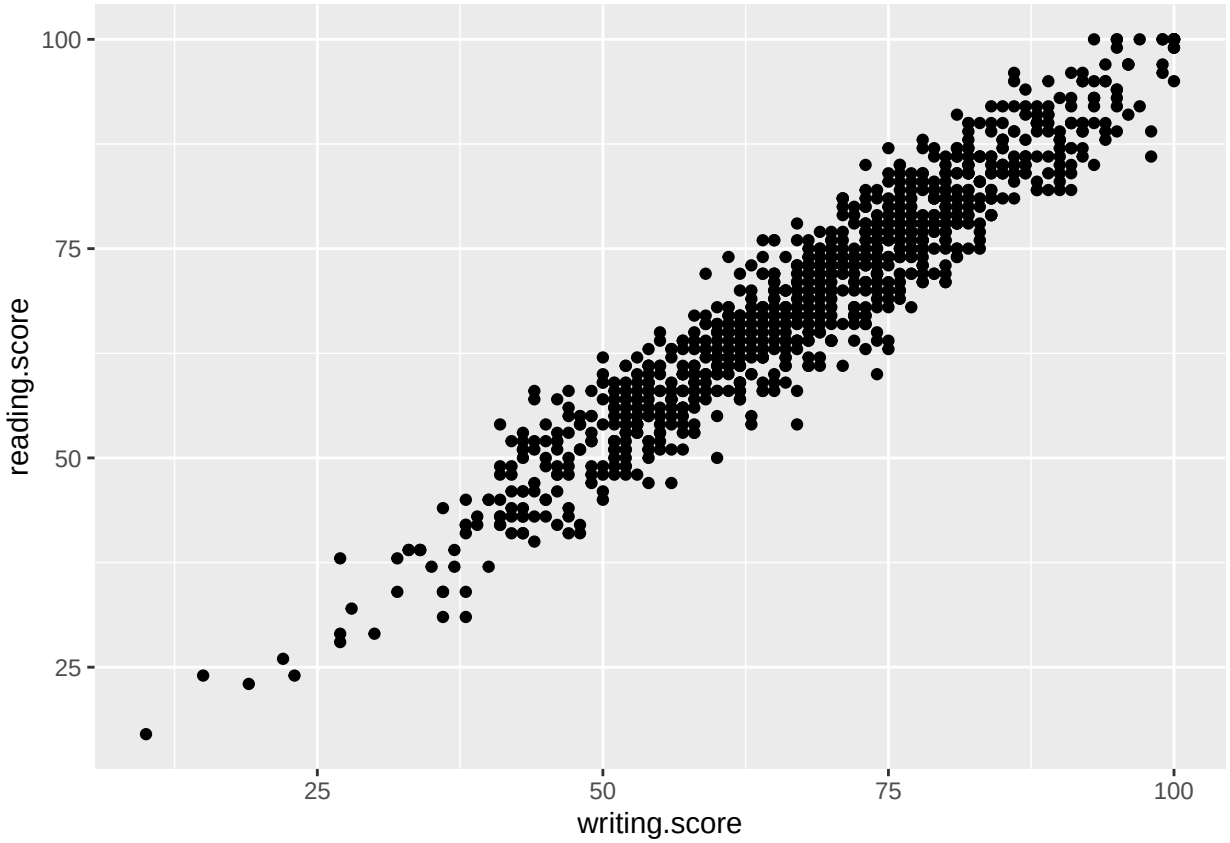
Resultados de Matemáticas Vs. Resultados de Lectura (Reading): Se observa que si los estudiantes obtuvieron una nota alta en matemáticas hay mucha posibilidad de que también lo hagan en lectura.

```
ggplot(students_data, aes(x = math.score, y = reading.score)) + geom_point()
```



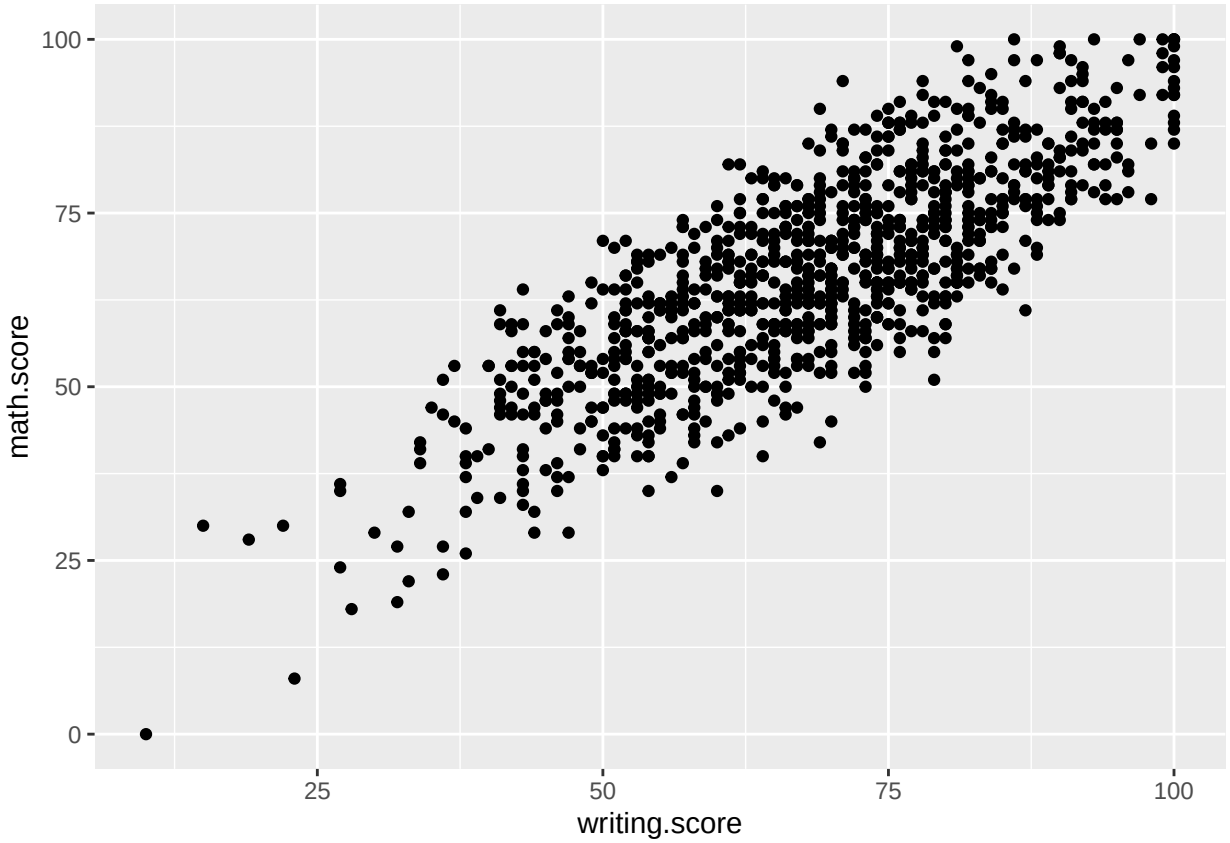
Resultados de Escritura (Writing) Vs. Resultados de Lectura (Reading): Este gráfico muestra que existe una correlación más alta que en el gráfico anterior.

```
ggplot(students_data, aes(x = writing.score, y = reading.score)) + geom_point()
```



Resultados de Matemáticas Vs. Resultados de Escritura (Writing): Se observa que a mayor nota de Matemáticas mayor nota de Escritura (Writing). Pero tiene menos correlación que en el anterior.

```
ggplot(students_data, aes(x = writing.score, y = math.score)) + geom_point()
```



2. Construimos el modelo lineal para predecir el resultado de los examenes en funcion del resto de variables.

   Creamos los datasets de entrenamiento y de pruebas (test). De tal forma que usamos 800 observaciones de entrenamiento y 200 para las pruebas de test.

```
set.seed(1)
sample = sample(nrow(students_data), floor(nrow(students_data) * 0.8))
datos_entrenamiento = students_data[sample,]
datos_test = students_data[-sample,]

dim(datos_entrenamiento)
```

```
## [1] 800   8
```

```
dim(datos_test)
```

```
## [1] 200   8
```

Lo siguiente es construir el modelo predictivo para cada uno de los exámenes (Matemáticas, Lectura y Escritura).

Modelo para Matemáticas:

```r
modelo_matematicas = lm(math.score~.,data = datos_entrenamiento)
summary(modelo_matematicas)
```

```
##
## Call:
## lm(formula = math.score ~ ., data = datos_entrenamiento)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6851  -3.6447   0.0742   3.5112  12.9402
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                                -12.24882    1.36009  -9.006
## gendermale                                  13.58150    0.41005  33.122
## race.ethnicitygroup B                        1.03306    0.76500   1.350
## race.ethnicitygroup C                        0.15328    0.71777   0.214
## race.ethnicitygroup D                        0.43548    0.74068   0.588
## race.ethnicitygroup E                        4.96000    0.80696   6.147
## parental.level.of.educationbachelor's degree -0.33711    0.66998  -0.503
## parental.level.of.educationhigh school        0.94797    0.58239   1.628
## parental.level.of.educationmaster's degree   -1.14865    0.89655  -1.281
## parental.level.of.educationsome college       0.83541    0.55949   1.493
## parental.level.of.educationsome high school   0.87754    0.61107   1.436
## lunchstandard                                 2.94867    0.40891   7.211
## test.preparation.coursenone                   3.52140    0.43508   8.094
## reading.score                                 0.25954    0.04738   5.478
## writing.score                                 0.70841    0.04902  14.452
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## gendermale                                   < 2e-16 ***
## race.ethnicitygroup B                          0.177
## race.ethnicitygroup C                          0.831
## race.ethnicitygroup D                          0.557
## race.ethnicitygroup E                       1.26e-09 ***
## parental.level.of.educationbachelor's degree   0.615
## parental.level.of.educationhigh school         0.104
## parental.level.of.educationmaster's degree     0.201
## parental.level.of.educationsome college        0.136
## parental.level.of.educationsome high school    0.151
## lunchstandard                               1.31e-12 ***
## test.preparation.coursenone                 2.20e-15 ***
## reading.score                               5.79e-08 ***
## writing.score                                < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.283 on 785 degrees of freedom
## Multiple R-squared:  0.8805, Adjusted R-squared:  0.8784
## F-statistic: 413.1 on 14 and 785 DF,  p-value: < 2.2e-16
```

Se observan las siguientes observaciones:

- R cuadrado es del 0,878. Esto quiere decir que el 87,8% de los resultados de los exámenes de matemáticas está sujeto al comportamiento de las variables. Todas las variables son importantes en el modelo.
- El coeficiente de estimación para el género Masculino es de 13,58. Esto quiere decir que la nota promedia del sexo masculino es mayor que la nota promedia del sexo femenino.
- El coeficiente de estimación para la raza/etnia del grupo E es del 4,96. Esto quiere decir que el promedio de nota para la etnia del Grupo E es mayor que para el grupo A.
- El coeficiente de estimación para la comida de tipo standard es de 2,94. Esto quiere decir que el promedio de nota para los estudiantes que comieron comida de tipo standard es mayor que para los que comieron reducida.
- El coeficiente de estimación para la variable de test de preparación es de 3,52. Esto quiere decir que los estudiantes que no realizaron un test de preparación para el examen tienen mayor nota promedia que los que si lo realizaron.

Modelo para Lectura(Reading):

```
modelo_lectura = lm(reading.score~.,data = datos_entrenamiento)
summary(modelo_lectura)
```

```
##
## Call:
## lm(formula = reading.score ~ ., data = datos_entrenamiento)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3815  -2.6946   0.1004   2.7468   9.8922
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                               2.23001    1.05322   2.117
## gendermale                               -0.13214    0.46939  -0.282
## race.ethnicitygroup B                     0.29423    0.56615   0.520
## race.ethnicitygroup C                    -0.14477    0.53066  -0.273
## race.ethnicitygroup D                    -1.55594    0.54490  -2.855
## race.ethnicitygroup E                    -0.19814    0.61076  -0.324
## parental.level.of.educationbachelor's degree -1.54503    0.49234  -3.138
## parental.level.of.educationhigh school    0.73581    0.43050   1.709
## parental.level.of.educationmaster's degree   -1.07008    0.66244  -1.615
## parental.level.of.educationsome college      -0.53842    0.41379  -1.301
## parental.level.of.educationsome high school   1.03874    0.45086   2.304
## lunchstandard                            -1.33764    0.30850  -4.336
## test.preparation.coursenone               1.82083    0.32845   5.544
## math.score                                0.14187    0.02590   5.478
## writing.score                             0.84828    0.02732  31.055
##                                          Pr(>|t|)
## (Intercept)                               0.03455 *
## gendermale                                0.77839
## race.ethnicitygroup B                     0.60341
## race.ethnicitygroup C                     0.78507
## race.ethnicitygroup D                     0.00441 **
## race.ethnicitygroup E                     0.74571
## parental.level.of.educationbachelor's degree  0.00176 **
```

```
## parental.level.of.educationhigh school        0.08781 .
## parental.level.of.educationmaster's degree    0.10663
## parental.level.of.educationsome college       0.19357
## parental.level.of.educationsome high school   0.02149 *
## lunchstandard                                1.64e-05 ***
## test.preparation.coursenone                  4.05e-08 ***
## math.score                                   5.79e-08 ***
## writing.score                                 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.906 on 785 degrees of freedom
## Multiple R-squared:  0.9296, Adjusted R-squared:  0.9284
## F-statistic: 740.7 on 14 and 785 DF,  p-value: < 2.2e-16
```

Se observa lo siguiente:

- R cuadrado es de 0,9284. Esto quiere decir que el 92,8% de los resultados de los exámenes de Lectura (Reading) esta sujeto al comportamiento de las variables.
- También, destacan los coeficientes de estimación de: Test de preparación para el exámen y el nivel academico de los padres.

Modelo para Escritura (Writing):

```
modelo_escritura = lm(writing.score~.,data = datos_entrenamiento)
summary(modelo_escritura)
```

```
##
## Call:
## lm(formula = writing.score ~ ., data = datos_entrenamiento)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0617  -2.3396   0.0104   2.2032  10.4499
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                   8.52057    0.87304   9.760
## gendermale                                   -6.05700    0.34939 -17.336
## race.ethnicitygroup B                        -0.48064    0.49532  -0.970
## race.ethnicitygroup C                         0.33784    0.46434   0.728
## race.ethnicitygroup D                         1.48770    0.47646   3.122
## race.ethnicitygroup E                        -1.45012    0.53211  -2.725
## parental.level.of.educationbachelor's degree  1.60956    0.42981   3.745
## parental.level.of.educationhigh school       -1.37610    0.37430  -3.676
## parental.level.of.educationmaster's degree    1.53540    0.57819   2.656
## parental.level.of.educationsome college       0.01993    0.36257   0.055
## parental.level.of.educationsome high school  -1.55478    0.39205  -3.966
## lunchstandard                                 0.41040    0.27285   1.504
## test.preparation.coursenone                  -3.52384    0.26470 -13.312
## math.score                                    0.29666    0.02053  14.452
## reading.score                                 0.64988    0.02093  31.055
```

```
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## gendermale                                   < 2e-16 ***
## race.ethnicitygroup B                       0.332172
## race.ethnicitygroup C                       0.467088
## race.ethnicitygroup D                       0.001860 **
## race.ethnicitygroup E                       0.006569 **
## parental.level.of.educationbachelor's degree 0.000194 ***
## parental.level.of.educationhigh school       0.000253 ***
## parental.level.of.educationmaster's degree   0.008079 **
## parental.level.of.educationsome college      0.956174
## parental.level.of.educationsome high school  7.98e-05 ***
## lunchstandard                               0.132942
## test.preparation.coursenone                  < 2e-16 ***
## math.score                                   < 2e-16 ***
## reading.score                                < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.419 on 785 degrees of freedom
## Multiple R-squared:  0.9503, Adjusted R-squared:  0.9494
## F-statistic:  1072 on 14 and 785 DF,  p-value: < 2.2e-16
```

Se puede observar que:

- R cuadrado es de 0,9494. Es decir, que le 94,94% de los resultados de los exámenes de Escritura (Writing) está sujeto al comportamiento de las variables.
- Destacan los coeficientes de estimación de: raza o etnia, nivel academico de los padres (master y bachelor degree) y test de preparación del examen.

3. Predicción del modelo y rendimiento.

Una vez construidos los 3 modelos, es tiempo de usar la funcion predict() para predecir dado un modelo los resultados con el dataset de pruebas que hicimos anteriormente. Y calcular el rendimiento del modelo, calculando el RMSE (Root Mean Square Error - Raíz del error cuadratico medio)

Predicción modelo del examen de Matemáticas:

```
pred_mates = predict(modelo_matematicas, newdata = datos_test)
sqrt(mean((datos_test$math.score - pred_mates)^2))
```

```
## [1] 5.717545
```

Predicción modelo del examen de Lectura(Reading):

```
pred_reading = predict(modelo_lectura, newdata = datos_test)
sqrt(mean((datos_test$reading.score - pred_reading)^2))
```

```
## [1] 4.308895
```

Predicción modelo del examen de Escritura (Writing):

```
pred_writing = predict(modelo_escritura, newdata = datos_test)
sqrt(mean((datos_test$writing.score - pred_writing)^2))
```

```
## [1] 3.794674
```

El error de predicción del modelo es relativamente pequeño.

Estudio 2: Arbol de regresión.

En este estudio se tratara de predecir las notas de matemáticas, lectura (Reading) y escritura (Writing) en función del resto de variables: género, raza, nivel académico de los padres, Comida (lo que comieron: reducida o standard), si hicieron el test de preparación para el examen.

Para ello, primero cargamos las libreria necesaria que es rpart:

```
library(rpart)
```

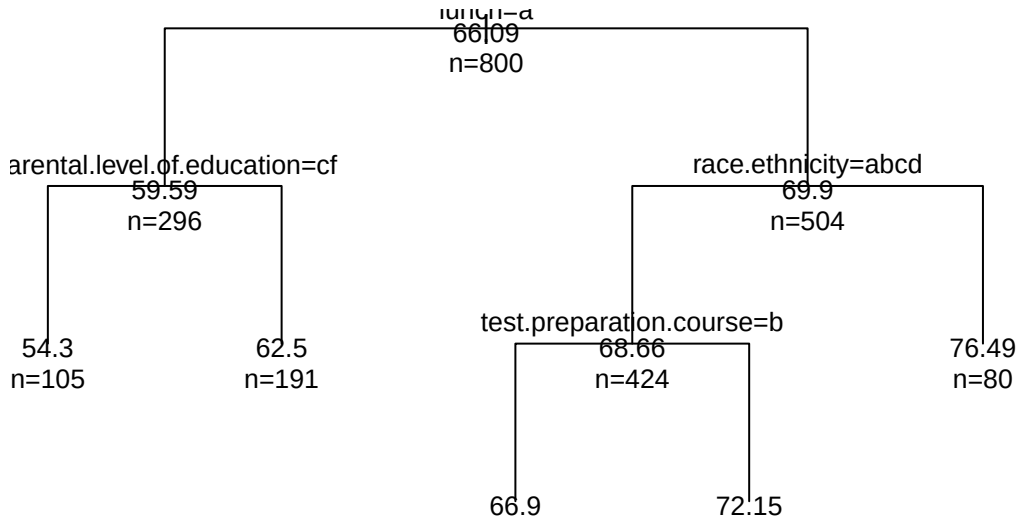
Árbol de Regressión para predecir la nota de Matemáticas:

Haciendo uso de la función rpart() de R, construimos el árbol.

```
tree_mates <- rpart(math.score~gender+race.ethnicity+parental.level.of.education+lunch+test.preparation
```

Visualizamos el árbol:

```
plot(tree_mates, uniform=TRUE,
   main="Árbol de regresión calificaciones de Matemáticas")
text(tree_mates, use.n=TRUE, all=TRUE, cex=.8)
```



Mostramos los resultados del árbol contruido:

```
printcp(tree_mates)
```

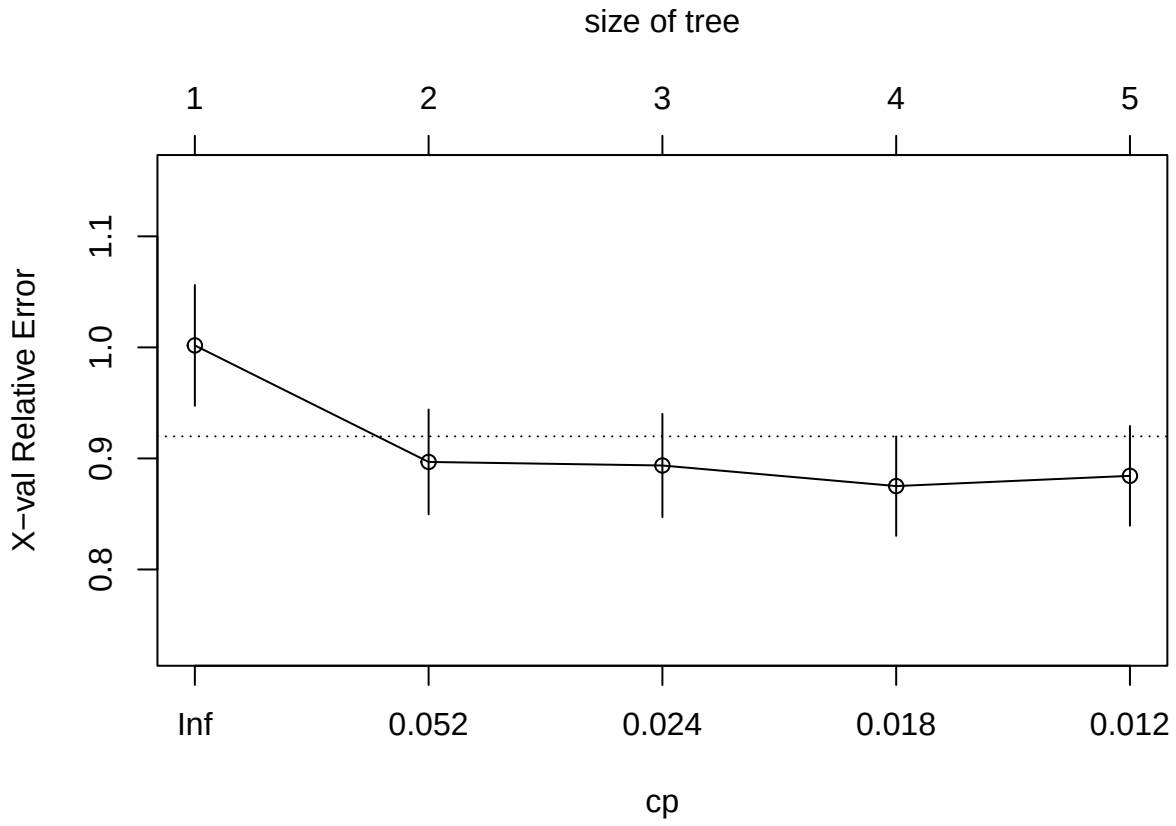
```
##
## Regression tree:
## rpart(formula = math.score ~ gender + race.ethnicity + parental.level.of.education +
##     lunch + test.preparation.course, data = datos_entrenamiento,
##     method = "anova")
##
## Variables actually used in tree construction:
## [1] lunch                       parental.level.of.education
## [3] race.ethnicity              test.preparation.course
##
## Root node error: 183325/800 = 229.16
##
## n= 800
##
##         CP nsplit rel error  xerror     xstd
## 1 0.108088      0   1.00000 1.00178 0.054402
## 2 0.024838      1   0.89191 0.89683 0.047201
## 3 0.022491      2   0.86707 0.89364 0.046462
## 4 0.014222      3   0.84458 0.87509 0.044769
## 5 0.010000      4   0.83036 0.88433 0.044875
```

Podemos comprobar tanto por el árbol dibujado anteriormente como por resultados ahora obtenidos que las

variables que utiliza para decidir son las siguientes: comida, test de preparación del curso, nivel educativo de los padres y raza.

Visualizamos los resultados de la validación cruzada:

```
plotcp(tree_mates)
```



Predecimos y vemos el error del Árbol de regresión:

```
pred_tree_mates = predict(tree_mates, newdata = datos_test)
sqrt(mean((datos_test$math.score - pred_tree_mates)^2))
```

```
## [1] 13.4669
```

El error es de 13,4669. Por lo que el error es relativamente alto.

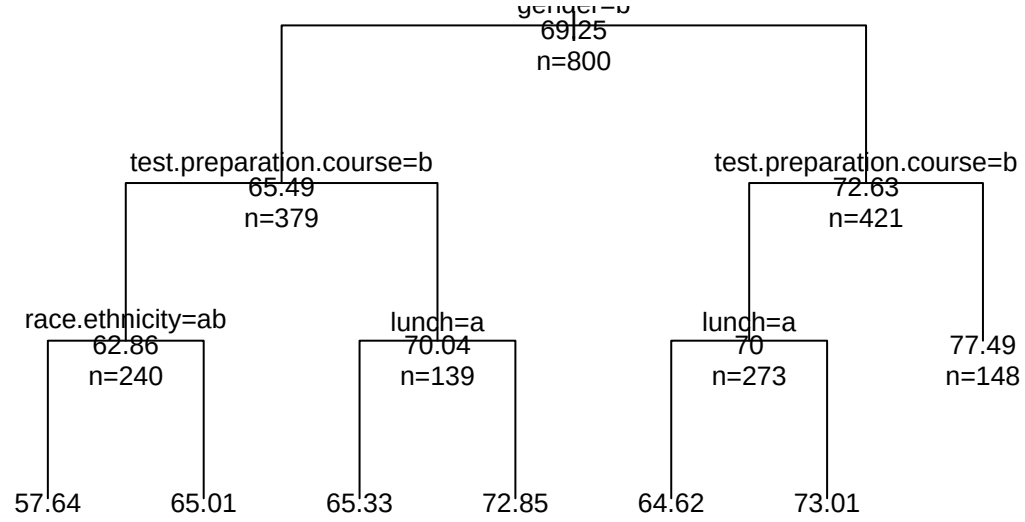
Árbol de Regressión para predecir la nota de Lectura (Reading):

Haciendo uso de la función rpart() de R, construimos el árbol.

```
tree_lectura <- rpart(reading.score~gender+race.ethnicity+parental.level.of.education+lunch+test.prepara
```

Visualizamos el árbol:

```
plot(tree_lectura, uniform=TRUE,
   main="Árbol de regresión calificaciones de Lectura(Reading)")
text(tree_lectura, use.n=TRUE, all=TRUE, cex=.8)
```



Mostramos los resultados del árbol contruido:

```
printcp(tree_lectura)
```

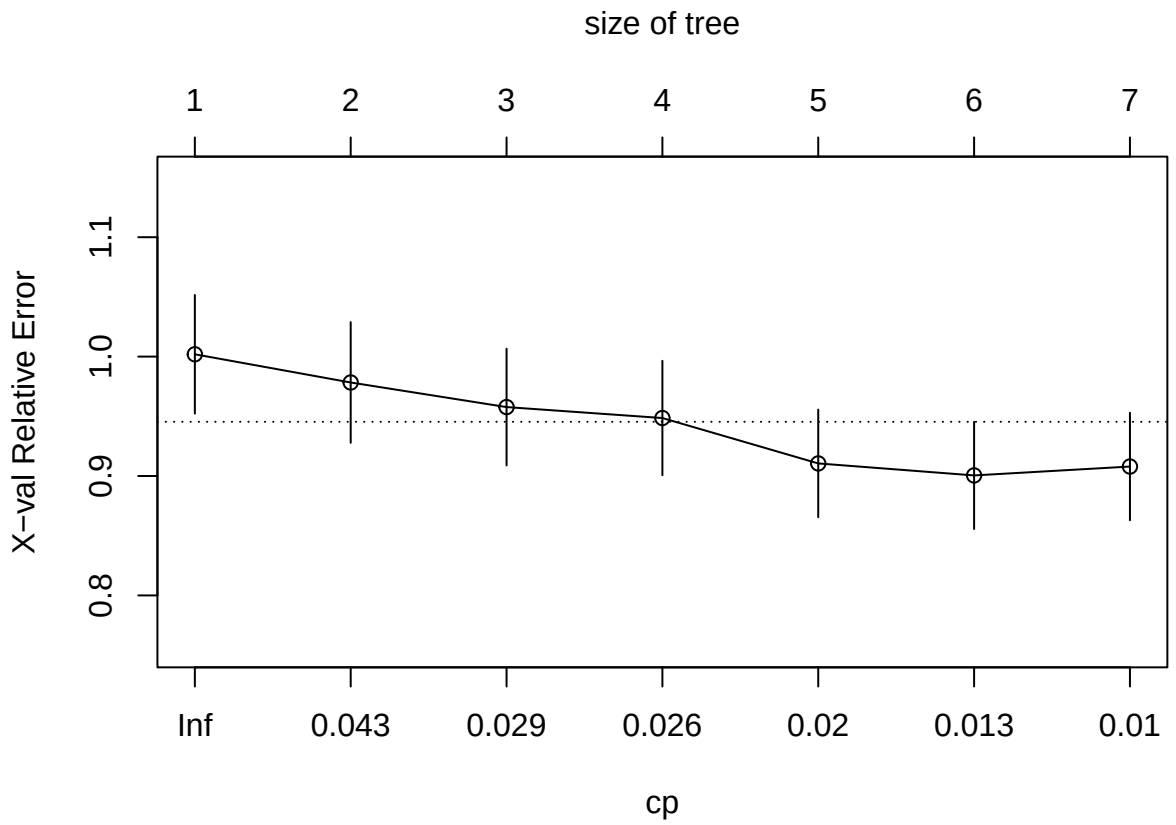
```
##
## Regression tree:
## rpart(formula = reading.score ~ gender + race.ethnicity + parental.level.of.education +
##     lunch + test.preparation.course, data = datos_entrenamiento,
##     method = "anova")
##
## Variables actually used in tree construction:
## [1] gender                  lunch                   race.ethnicity
## [4] test.preparation.course
##
## Root node error: 170165/800 = 212.71
##
## n= 800
##
##         CP nsplit rel error  xerror     xstd
## 1 0.059686      0   1.00000 1.00196 0.049685
## 2 0.031641      1   0.94031 0.97832 0.050549
```

```
## 3 0.026618      2   0.90867 0.95774 0.048890
## 4 0.025945      3   0.88205 0.94856 0.047933
## 5 0.015822      4   0.85611 0.91055 0.045097
## 6 0.010827      5   0.84029 0.90050 0.044897
## 7 0.010000      6   0.82946 0.90794 0.045027
```

Podemos comprobar tanto por el árbol dibujado anteriormente como por resultados ahora obtenidos que las variables que utiliza para decidir son las siguientes: comida, genero, si hizo el test de preparación para el curso y raza.

Visualizamos los resultados de la validación cruzada:

```
plotcp(tree_lectura)
```



Predecimos y vemos el error del Árbol de regresión:

```
pred_tree_lectura = predict(tree_lectura, newdata = datos_test)
sqrt(mean((datos_test$reading.score - pred_tree_lectura)^2))
```

```
## [1] 13.2552
```

El error es de 13,2552. Por lo que el error es relativamente alto.

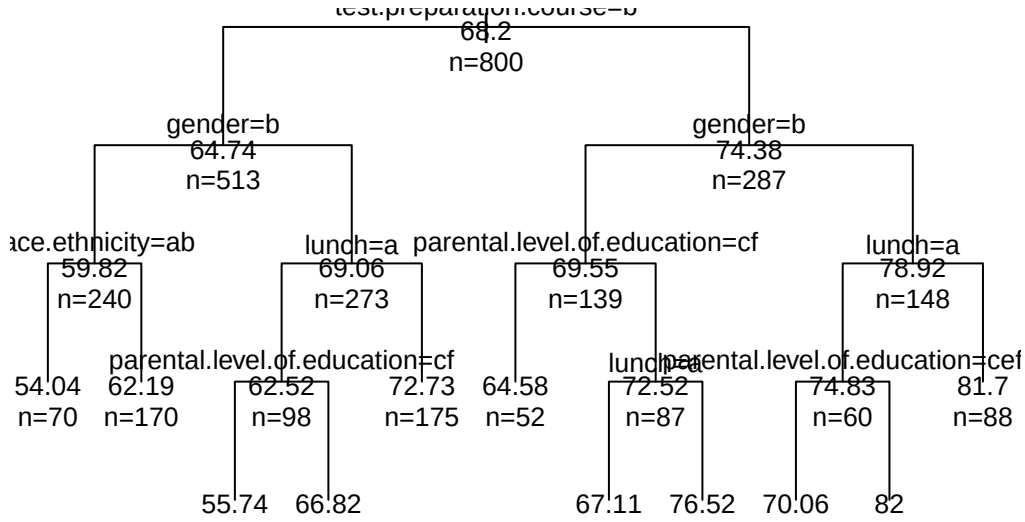
Árbol de Regressión para predecir la nota de Escritura (Writing):

Haciendo uso de la función rpart() de R, construimos el árbol.

```
tree_escritura <- rpart(writing.score~gender+race.ethnicity+parental.level.of.education+lunch+test.prepa
```

Visualizamos el árbol:

```
plot(tree_escritura, uniform=TRUE,
   main="Árbol de regresión calificaciones de Escritura (Writing")
text(tree_escritura, use.n=TRUE, all=TRUE, cex=.8)
```



Mostramos los resultados del árbol contruido:

```
printcp(tree_escritura)
```

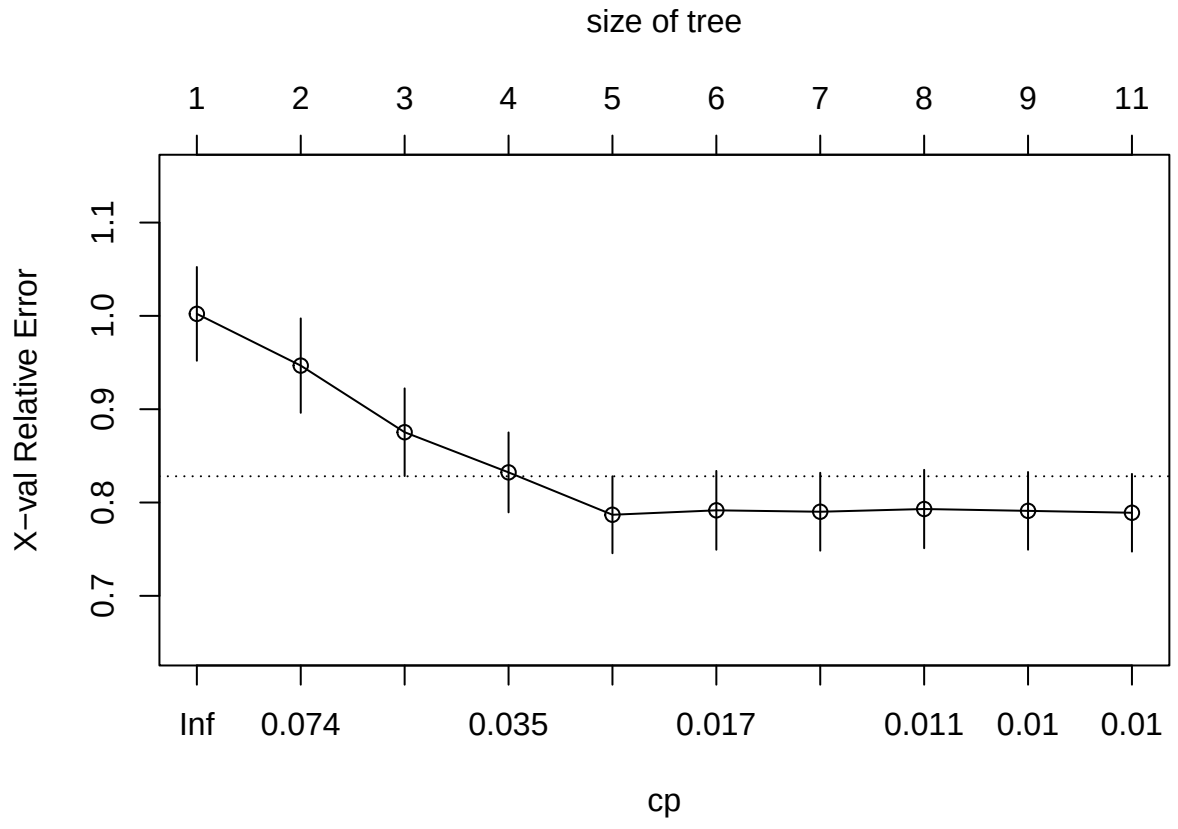
```
##
## Regression tree:
## rpart(formula = writing.score ~ gender + race.ethnicity + parental.level.of.education +
##     lunch + test.preparation.course, data = datos_entrenamiento,
##     method = "anova")
##
## Variables actually used in tree construction:
## [1] gender                       lunch
## [3] parental.level.of.education race.ethnicity
## [5] test.preparation.course
##
## Root node error: 184606/800 = 230.76
##
```

```
## n= 800
##
##          CP nsplit rel error  xerror     xstd
## 1  0.092701      0   1.00000 1.00217 0.050221
## 2  0.059140      1   0.90730 0.94672 0.050576
## 3  0.035441      2   0.84816 0.87534 0.046989
## 4  0.034106      3   0.81272 0.83229 0.042812
## 5  0.017846      4   0.77861 0.78690 0.041236
## 6  0.015471      5   0.76077 0.79164 0.042239
## 7  0.011116      6   0.74530 0.79013 0.041737
## 8  0.010204      7   0.73418 0.79304 0.042009
## 9  0.010126      8   0.72398 0.79108 0.041629
## 10 0.010000     10   0.70372 0.78903 0.041639
```

Podemos comprobar tanto por el árbol dibujado anteriormente como por resultados ahora obtenidos que las variables que utiliza para decidir son las siguientes: comida, genero, si hizo el test de preparación para el curso, nivel eduactivo de los padres y raza. Vemos que para este árbol se toma más variables para tomar la decisión.

Visualizamos los resultados de la validación cruzada:

```
plotcp(tree_escritura)
```



Predecimos y vemos el error del Árbol de regresión:

```
pred_tree_escritura = predict(tree_escritura, newdata = datos_test)
sqrt(mean((datos_test$writing.score - pred_tree_escritura)^2))
```

```
## [1] 12.72905
```

El error es de 12,72905. Por lo que el error es relativamente alto. Aunque menor que en los dos árboles anteriores.

Árbol de Regressión para predecir la nota de Escritura (Writing):

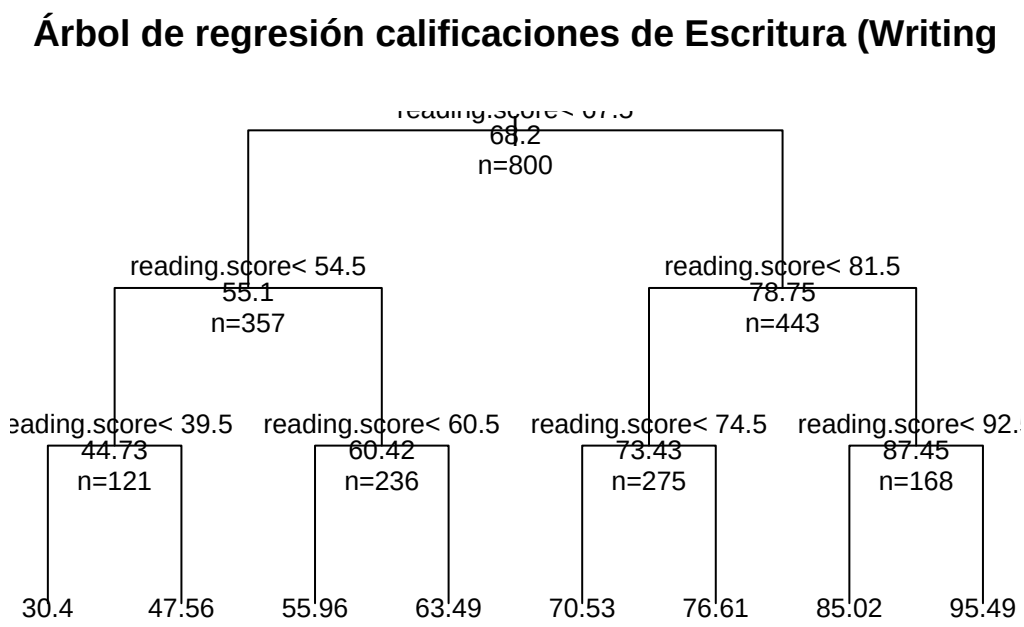
Volvemos ha hacer este arbol de regresión pero metiendole otra variable: reading.score. Ya que vimos en el estudio de regresión multivariable una correlación alta entre Writing(Escritura) y Reading(Lectura).

Haciendo uso de la función rpart() de R, construimos el árbol.

```
tree_escritura1 <- rpart(writing.score~gender+race.ethnicity+parental.level.of.education+lunch+test.prep
```

Visualizamos el árbol:

```
plot(tree_escritura1, uniform=TRUE,
   main="Árbol de regresión calificaciones de Escritura (Writing")
text(tree_escritura1, use.n=TRUE, all=TRUE, cex=.8)
```



Mostramos los resultados del árbol contruido:

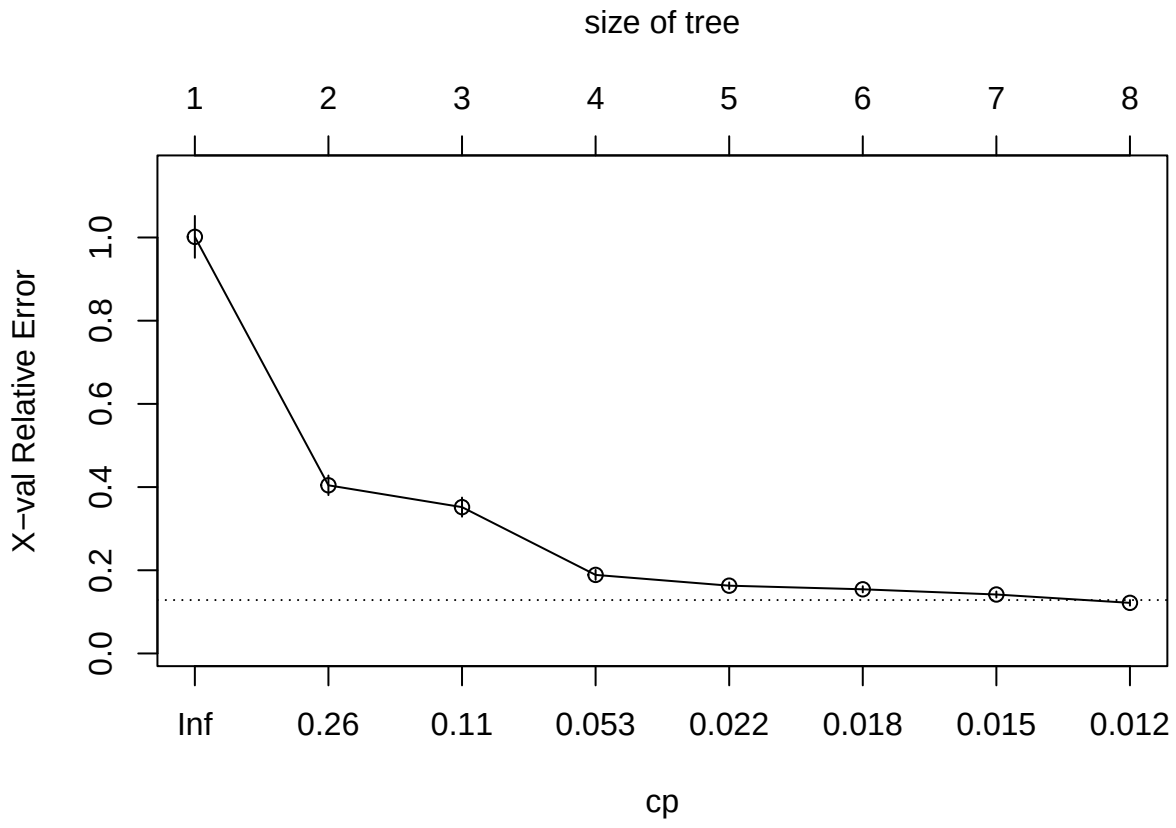
```
printcp(tree_escritura1)
```

```
##
## Regression tree:
## rpart(formula = writing.score ~ gender + race.ethnicity + parental.level.of.education +
##     lunch + test.preparation.course + reading.score, data = datos_entrenamiento,
##     method = "anova")
##
## Variables actually used in tree construction:
## [1] reading.score
##
## Root node error: 184606/800 = 230.76
##
## n= 800
##
##         CP nsplit rel error  xerror      xstd
## 1 0.598633      0   1.00000 1.00151 0.0502901
## 2 0.111094      1   0.40137 0.40427 0.0235477
## 3 0.106754      2   0.29027 0.35178 0.0229848
## 4 0.026642      3   0.18352 0.18889 0.0118953
## 5 0.017762      4   0.15688 0.16297 0.0079147
## 6 0.017479      5   0.13911 0.15425 0.0074531
## 7 0.013718      6   0.12163 0.14178 0.0072381
## 8 0.010000      7   0.10792 0.12169 0.0067254
```

Podemos comprobar tanto por el árbol dibujado anteriormente como por resultados ahora obtenidos que unicamente se usa la variable de Reading score.

Visualizamos los resultados de la validación cruzada:

```
plotcp(tree_escritura1)
```



Predecimos y vemos el error del Árbol de regresión:

```
pred_tree_escritura1 = predict(tree_escritura1, newdata = datos_test)
sqrt(mean((datos_test$writing.score - pred_tree_escritura1)^2))
```

```
## [1] 5.611069
```

El error es de 5.61108. Por lo que el error es relativamente bajo.

Conlusiones: En los tres árbole de regresión se toma como variables de decisión en común: genero, raza, comida y el test de preparación del curso.

Vemos que si se añade como variable de decisión el resultado del lectura para el árbol de regresión de escritura, solo se usa esa variable como decisión y el error de predicción disminuye por lo que vemos nuevamente una correlación alta.

Estudio 3: Uso de Random Forest.

Random Forest es un algoritmo de clasificación que consiste en el uso de muchos árboles de decisión. Utiliza bagging y usa características aleatorias para construir cada árbol de forma individual. Trata de crear un bosque de árboles no correlacionado entre ellos. Un bosque de árboles puede ser más preciso en la predicción que un árbol individual.

Antes de comenzar cargamos la libreria ranger:

```
library(ranger)
```

Construimos el modelo de Random forest con el conjunto de datos de entrenamiento que hemos estado usando para todos los casos:

Para Matemáticas:

Creamos en un primer momento la semilla aleatoria para reproducir los resultados del modelo de random forest.

```
semilla <- set.seed(1234)
```

Declaramos la variable de salida que queremos predecir: el resultado del examen de Matemáticas.

```
salida_mates <- "math.score"
```

Declaramos las variables de entrada, que sera común para los tres modelos.

```
entrada<- c("gender","race.ethnicity","parental.level.of.education","lunch","test.preparation.course")
```

Creamos la fórmula de la salida (resultado de mates) en función de las variables de entrada.

```
formula_mates <- paste(salida_mates, "~", paste(entrada, collapse = " + "))
```

Construimos, entrenamos y visualizamos el modelo de random forest, con un número de arbóles del bosque de 200.

```
(modelo_rf_mates <- ranger(formula_mates, datos_entrenamiento, num.trees = 200, respect.unordered.facto
```

```
## Ranger result
##
## Call:
##  ranger(formula_mates, datos_entrenamiento, num.trees = 200, respect.unordered.factors = "order",
##
## Type:                             Regression
## Number of trees:                  200
## Sample size:                      800
## Number of independent variables:  5
## Mtry:                             2
## Target node size:                 5
## Variable importance mode:         none
## Splitrule:                        variance
## OOB prediction error (MSE):       190.9703
## R squared (OOB):                  0.1676775
```

Vemos la predección del modelo de random forest y el error (RMSE).

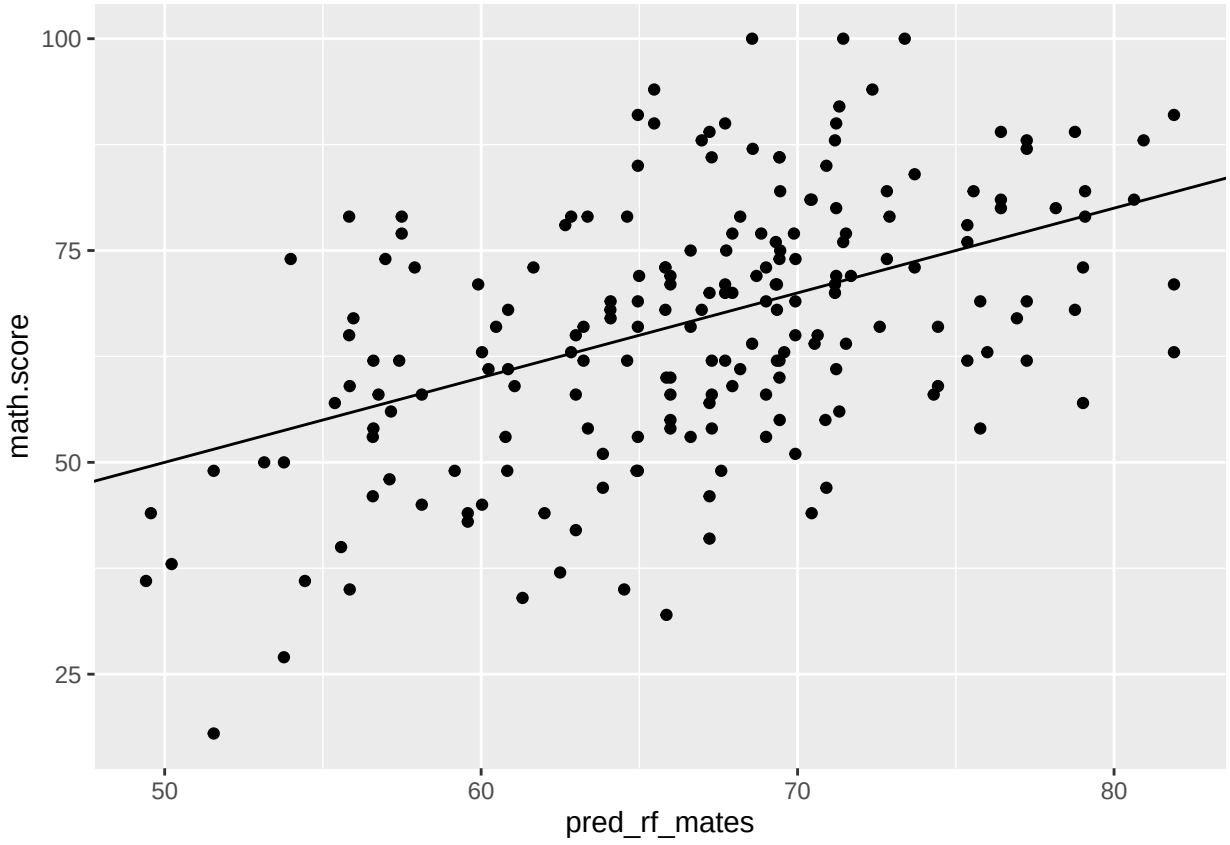
```
datos_test$pred_rf_mates <- predict(modelo_rf_mates, datos_test)$prediction
sqrt(mean((datos_test$math.score - datos_test$pred_rf_mates)^2))
```

```
## [1] 13.1413
```

El error es de 13.1413. Por lo que el error es alto.

Dibujamos una gráfica que representa la predicción de los datos de entrenamiento de los resultados de matemáticas y los datos actuales de los resultados.

```
ggplot(datos_test, aes(pred_rf_mates, math.score)) + geom_point() + geom_abline()
```



Vemos que efecitvamente hay mucho error en el la predición usando el modelo de random forest.

Para Lectura (Reading):

Declaramos la variable de salida que queremos predecir: el resultado del examen de Lectura.

```
salida_lectura <- "reading.score"
```

Creamos la fórmula de la salida (resultado de lecutura) en función de las variables de entrada.

```
formula_lectura <- paste(salida_lectura, "~", paste(entrada, collapse = " + "))
```

Construimos, entrenamos y visualizamos el modelo de random forest, con un número de arbóles del bosque de 200.

```
(modelo_rf_lectura <- ranger(formula_lectura, datos_entrenamiento, num.trees = 200, respect.unordered.fa
```

```
## Ranger result
##
```

```
## Call:
##  ranger(formula_lectura, datos_entrenamiento, num.trees = 200,      respect.unordered.factors = "ord
##
## Type:                             Regression
## Number of trees:                  200
## Sample size:                      800
## Number of independent variables:  5
## Mtry:                             2
## Target node size:                 5
## Variable importance mode:         none
## Splitrule:                        variance
## OOB prediction error (MSE):       182.0851
## R squared (OOB):                  0.1450323
```

Vemos la predección del modelo de random forest y el error (RMSE).

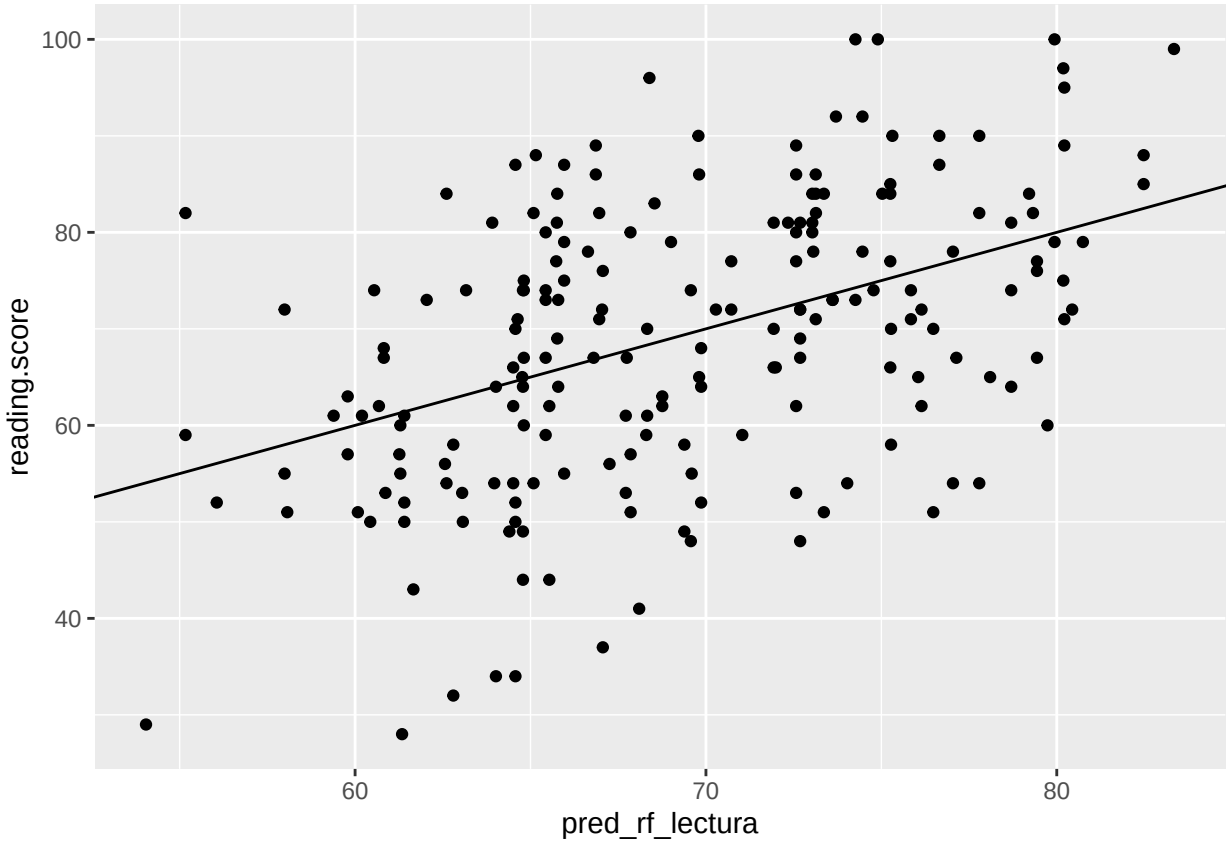
```
datos_test$pred_rf_lectura <- predict(modelo_rf_lectura, datos_test)$prediction
sqrt(mean((datos_test$reading.score - datos_test$pred_rf_lectura)^2))
```

```
## [1] 12.87443
```

El error es de 12.92341. Por lo que el error es alto, aunque un poco más bajo que el modelo de random forest para matemáticas.

Dibujamos una gráfica que representa la predicción de los datos de entrenamiento de los resultados de lectura y los datos actuales de los resultados.

```
ggplot(datos_test, aes(pred_rf_lectura, reading.score)) + geom_point() + geom_abline()
```



Vemos que efecitvamente hay mucho error en el la predición usando el modelo de random forest.

Para Escritura(Writing):

Declaramos la variable de salida que queremos predecir: el resultado del examen de Escritura.

```
salida_escritura <- "writing.score"
```

Creamos la fórmula de la salida (resultado deescritura) en función de las variables de entrada.

```
formula_escritura <- paste(salida_escritura, "~", paste(entrada, collapse = " + "))
```

Construimos, entrenamos y visualizamos el modelo de random forest, con un número de arbóles del bosque de 200.

```
(modelo_rf_escritura <- ranger(formula_escritura, datos_entrenamiento, num.trees = 200, respect.unorder
```

```
## Ranger result
##
## Call:
##  ranger(formula_escritura, datos_entrenamiento, num.trees = 200,      respect.unordered.factors = "o
##
## Type:                             Regression
## Number of trees:                  200
## Sample size:                      800
## Number of independent variables:  5
```

```
## Mtry:                           2
## Target node size:               5
## Variable importance mode:       none
## Splitrule:                      variance
## OOB prediction error (MSE):     170.1688
## R squared (OOB):                0.2634869
```

Vemos la predección del modelo de random forest y el error (RMSE).

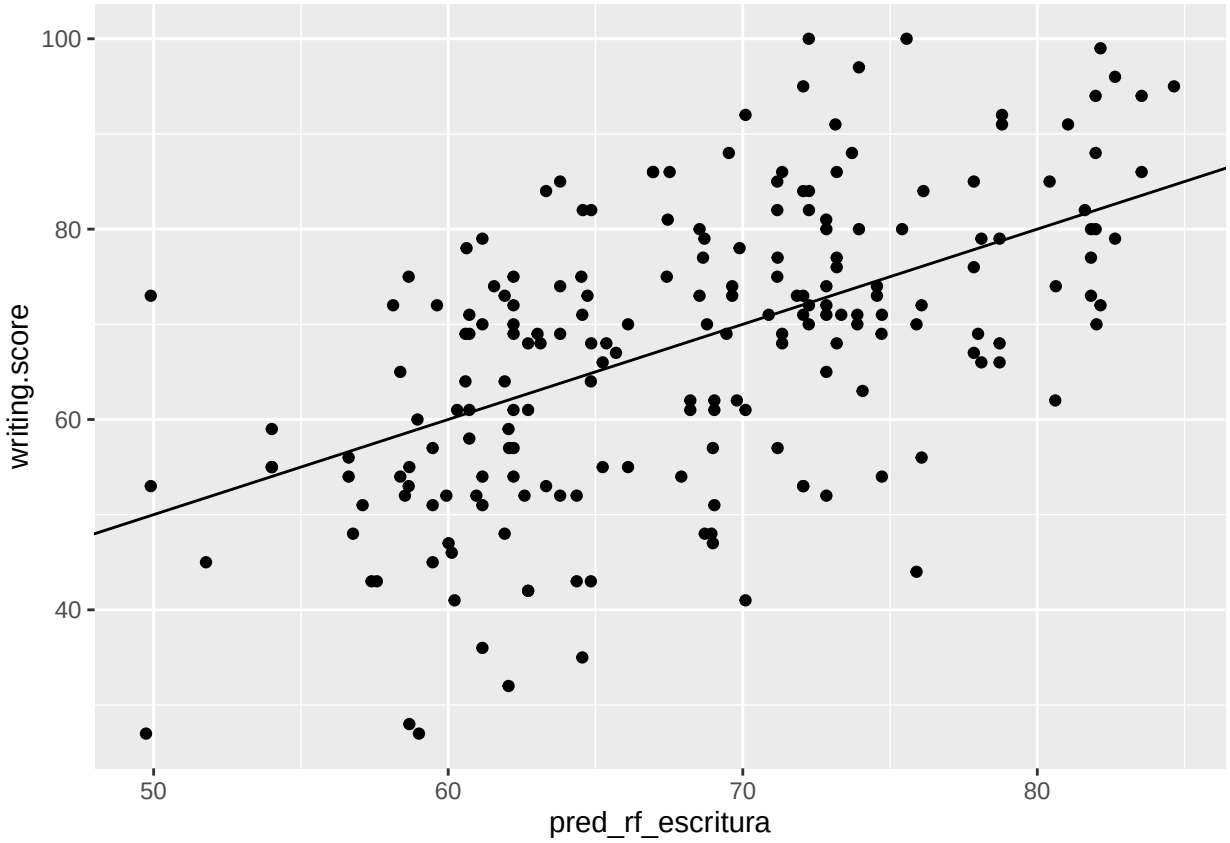
```
datos_test$pred_rf_escritura <- predict(modelo_rf_escritura, datos_test)$prediction
sqrt(mean((datos_test$writing.score - datos_test$pred_rf_escritura)^2))
```

```
## [1] 12.44285
```

El error es de 12.41713. Por lo que el error es alto, aunque un poco más bajo que el modelo de random forest para matemáticas.

Dibujamos una gráfica que representa la predicción de los datos de entrenamiento de los resultados de lectura y los datos actuales de los resultados.

```
ggplot(datos_test, aes(pred_rf_escritura, writing.score)) + geom_point() + geom_abline()
```



Vemos que efecitvamente hay mucho error en el la predición usando el modelo de random forest.

Referencias.

https://fhernanb.github.io/libro_regresion/predict.html

https://www.kaggle.com/yervandtadevosyan/linear-regression-students-performance

https://www.statmethods.net/advstats/cart.html

https://www.kaggle.com/sofiaabielmi/predicting-studentsperformance-with-random-forest#Random-Forest