

Proyecto Médodos

Mauricio Caicedo Palacio

2022-12-06

Introducción

Los datos que se usaran en este trabajo, son los resultados de las pruebas ICFES, que se realizaron en el segundo semestre de 2011. El ICFES es un acrónimo que se refiere a Instituto Colombiano para la Evaluación de la Educación Superior. Se trata de una entidad gubernamental de Colombia encargada de llevar a cabo exámenes de ingreso a la educación superior y de evaluar la calidad de la educación en el país. Los resultados de estos exámenes son utilizados por las instituciones de educación superior para tomar decisiones sobre la admisión de estudiantes y para mejorar la calidad de la educación que ofrecen. También se le suelen llamar prueba Saber 11.

En este trabajo usaremos un dataset perteneciente a una librería llamada “saber” desarrollada por Julian Cruz y Daniel. Esta librería nos será de mucha ayuda ya que contiene un número importante de datos que nos permitirán realizar nuestro análisis.

El código lo realizaremos en R usando el IDE. Instalamos y cargamos la librería y datos:

```
#install.packages('Rtools')
#library('Rtools')

#install.packages('devtools')
#devtools::install_github('https://github.com/nebulae-co/saber', force =T)

library('saber')
#SB11_2011 # Tiene alrededor de unos 37 mil datos.
data(SB11_20112)
```

Análisis Descriptivo Básico

Para ilustrar esta parte del trabajo, utilizaré los resultados en matemáticas de las pruebas del ICFES. Esto se debe a que usar todas las variables de los datos podría dificultar la comprensión de los conceptos que se pretenden transmitir con las visualizaciones. Por esta razón, me centraré en los resultados en matemáticas como ejemplo para ilustrar los conceptos que se abordan en este trabajo.

En esta ocasión, las calificaciones van de 0 a 100 puntos y la nota media es: 45.75. Las calificaciones en este rango permiten evaluar el desempeño de los estudiantes en una escala que va desde el insuficiente hasta el sobresaliente.

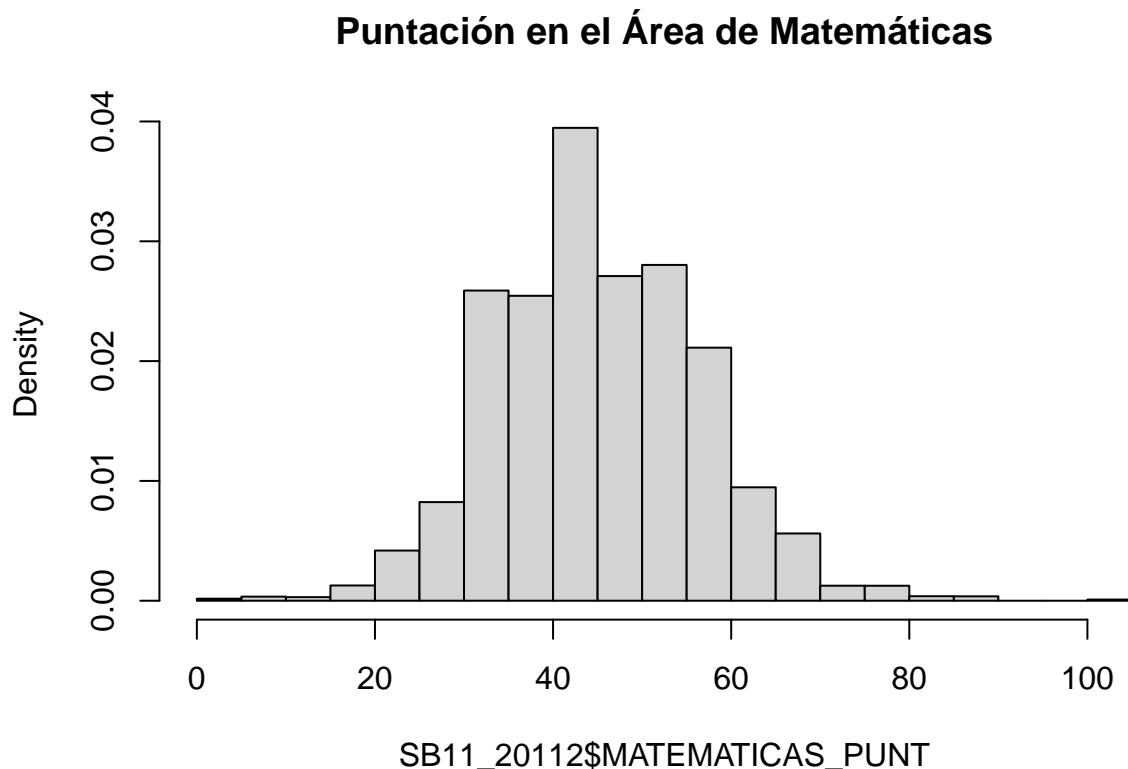
```
mean(SB11_20112$MATEMATICAS_PUNT); median(SB11_20112$MATEMATICAS_PUNT)
```

```
## [1] 45.74986
```

```
## [1] 45
```

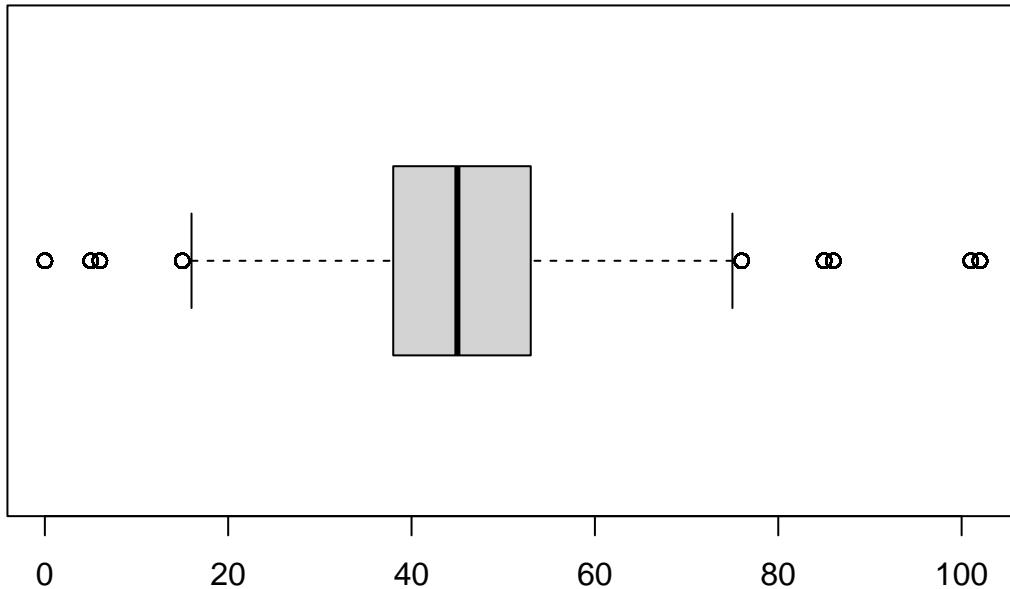
Una idea interesante es revisar en un histograma la densidad de las notas obtenidas en las pruebas del ICFES. Esto nos permitirá entender en qué rango se han dado las calificaciones y cómo se distribuyen en la población de estudiantes que tomó las pruebas. De esta manera, podremos comprender mejor el comportamiento de las notas en matemáticas en estas pruebas y obtener información valiosa para analizar la calidad de las notas del grupo de este periodo.

```
hist(SB11_20112$MATEMATICAS_PUNT, prob=T, main = 'Puntación en el Área de Matemáticas', nclass = 35)
```



De acuerdo con la información del histograma, las notas tienen un comportamiento aproximadamente normal, lo que significa que se distribuyen de manera similar a una curva de campana. Esto significa que la mayoría de las notas se encuentran cerca del valor promedio, y hay pocas notas muy altas o muy bajas en comparación con el resto. Para validar esta idea inicial y comprobar si realmente los datos están distribuidos de manera uniforme entorno a su media, realizaremos un boxplot que nos permite visualizar de una manera diferente los datos y evaluar la distribución de las notas.

```
boxplot(SB11_20112$MATEMATICAS_PUNT, horizontal = TRUE)
```



Como se puede observar en el boxplot, el 50% de los datos están en el intervalo de 45 puntos más o menos 8 puntos. Además, la media y la mediana son iguales en su parte entera, lo que indica que la distribución de las notas es bastante uniforme y no hay una gran desigualdad en el desempeño de los estudiantes. Una medida adicional que podemos usar para evaluar la distribución de los datos es la desviación típica, también conocida como desviación estándar. Esta medida nos permite determinar cuán dispersos están los datos en relación a su media. Revisaremos el valor de la desviación típica para obtener más información sobre la distribución de las notas de matemáticas en las pruebas del ICFES.

```
sd(SB11_20112$MATEMATICAS_PUNT); var(SB11_20112$MATEMATICAS_PUNT)
```

```
## [1] 11.83343
```

```
## [1] 140.0301
```

Aunque el 50% de los valores están relativamente cerca a la media, las medidas de desviación estándar y varianza nos permiten entender mejor los valores atípicos y aislados que se mostraron en el boxplot anterior. Estas medidas nos ayudan a identificar los valores que se alejan mucho del promedio y nos permiten determinar cuán dispersos están los datos en relación a su media.

Ahora, revisaremos los valores extremos y cuartiles de las notas de matemáticas en este período del ICFES. Como ya mencionamos, estos valores se visualizaron de manera gráfica en el boxplot que realizamos, pero no se mostró de manera clara dónde se encuentran exactamente estos puntos en el gráfico. Para entender mejor la distribución de los datos, vamos a calcular y analizar estos valores en detalle.

```
summary(SB11_20112$MATEMATICAS_PUNT)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      0.00   38.00  45.00  45.75  53.00 102.00
```

De acuerdo con los datos, hay calificaciones extremas en torno a 0 y otras en torno a 102 que podrían estar aumentando significativamente los valores de la varianza y la desviación estándar. Esto indica que hay algunas notas muy bajas y otras muy altas que se alejan mucho del promedio y que podrían estar afectando la distribución de los datos. Al calcular las medidas de dispersión, estos valores extremos pueden tener un gran impacto en el resultado y pueden distorsionar la información que obtenemos sobre el rendimiento de los estudiantes en las pruebas del ICFES. Por esta razón, es importante tener en cuenta estos valores y analizarlos con cuidado al evaluar la calidad de la educación en el campo de las matemáticas, dado que en casos como este cero es probable que haya algún error o que la persona ni siquiera presentó la prueba.

Inferencia

En esta parte, vamos a realizar una estimación de parámetros en torno a la media de las notas de lenguaje en la prueba ICFES. La nota media de lenguaje en la prueba ICFES es un indicador importante que nos permite conocer el rendimiento general de los estudiantes en esta materia. Vamos a realizar una estimación precisa de la media, como sigue:

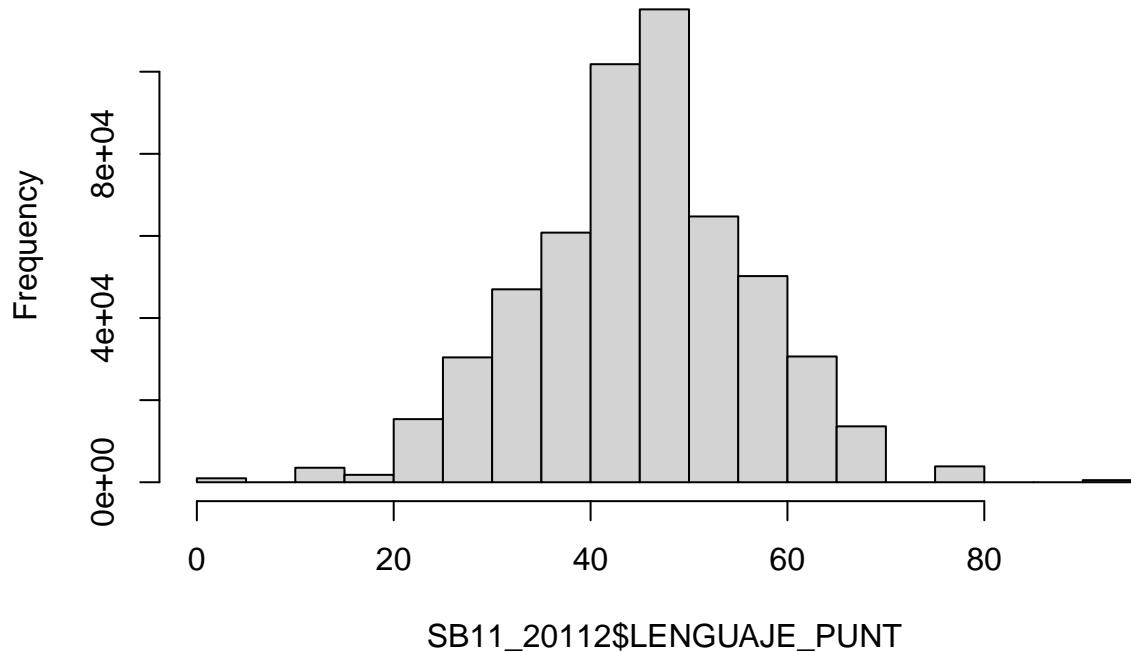
```
mean(SB11_20112$LENGUAJE_PUNT); sum(SB11_20112$LENGUAJE_PUNT)/length(SB11_20112$LENGUAJE_PUNT)
```

```
## [1] 45.79262  
## [1] 45.79262
```

Que ambos casos es 45.79. Ya sea usando la función ‘mean’ de R o sumando todos los valores y dividiendo por la cantidad de datos. Datos que al igual que el puntaje de matemáticas parecen tener una distribución normal.

```
hist(SB11_20112$LENGUAJE_PUNT)
```

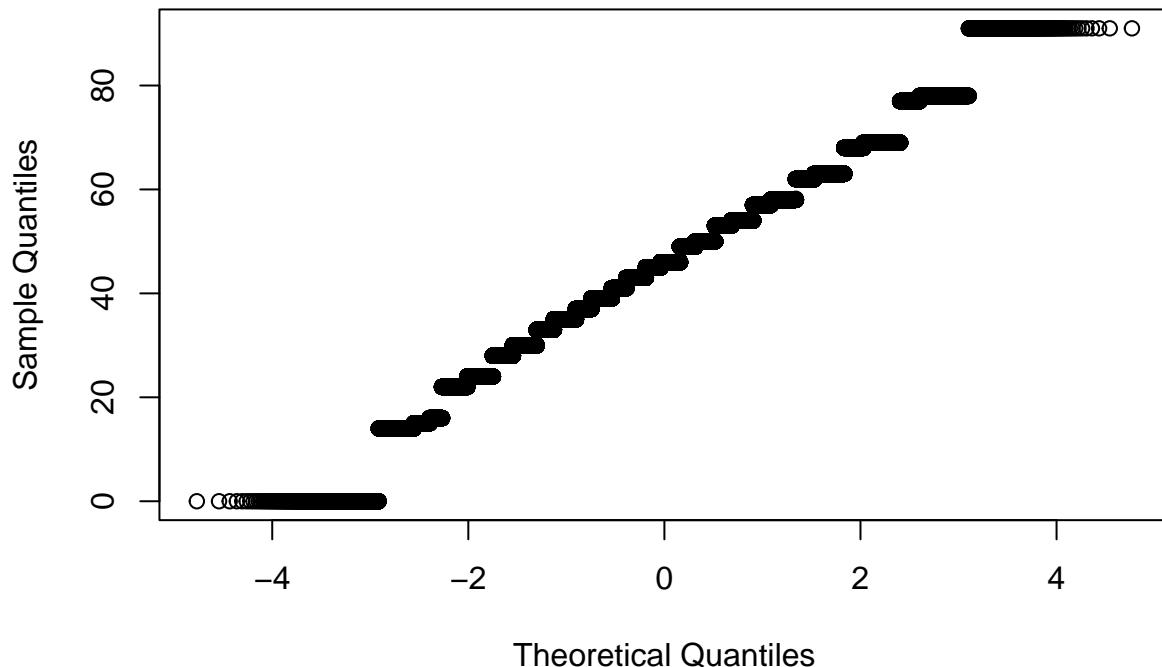
Histogram of SB11_20112\$LENGUAJE_PUNT



Veamos si el qqplot nos permite entender un poco mejor los datos de lenguaje, con la idea de que tienen una distribución normal.

```
qqnorm(SB11_20112$LENGUAJE_PUNT)
```

Normal Q-Q Plot



En la parte central, cercana a la media, parece que se asemeja a una distribución normal. Pero, se tienen saltos en las colas, cercanos a valores muy altos o muy bajos, se desdibuja la continuidad de esa normalidad. Como ya se expreso en el caso de matemáticas esos valores extremos, 0 y 100, podrían ser errores a la hora de realizar el computo de las pruebas o, personas que se les evaluó con cero, a pesar de no presentar el examen.

Vamos realizar un muestreo de los datos y los visualizaremos con el fin de adquirir más información sobre la hipótesis de normalidad de la distribución de las notas en el área de lenguaje en relación a su media.

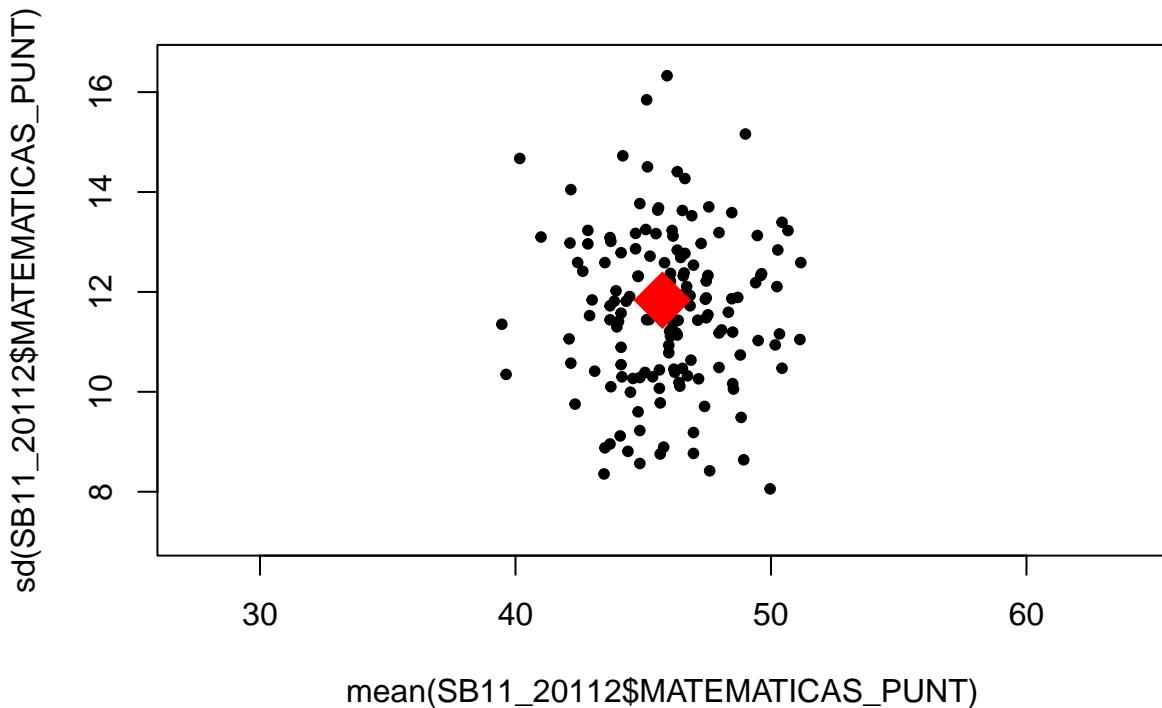
```
iteraciones <- 150 #La cantidad de la muestra es indiscriminada.  
tamano_muestral <- 30 #El tamaño de la muestra es indiscriminada.  
  
plot(  
  mean(SB11_20112$MATEMATICAS_PUNT),  
  sd(SB11_20112$MATEMATICAS_PUNT),  
  pch = 20,  
  cex = 4,  
  col = "white"  
)  
  
for(i in seq_len(iteraciones)){  
  points(  
    mean(sample(SB11_20112$MATEMATICAS_PUNT, tamano_muestral)),  
    sd(sample(SB11_20112$MATEMATICAS_PUNT, tamano_muestral)),  
    pch = 20  
)
```

```

}

points(
  mean(SB11_20112$MATEMATICAS_PUNT),
  sd(SB11_20112$MATEMATICAS_PUNT),
  pch = 18,
  cex = 4,
  col = 'red'
)

```



Como vemos, la nube de puntos importantes es entorno a la media. Juguemos un poco con el tamaño de la muestra y veamos que pasa.

```

iteraciones <- 150 #La cantidad de la muestra es indiscriminada.
tamano_muestral <- 150 #El tamaño de la muestra es indiscriminada.

plot(
  mean(SB11_20112$MATEMATICAS_PUNT),
  sd(SB11_20112$MATEMATICAS_PUNT),
  pch = 20,
  cex = 4,
  col = "white"
)

for(i in seq_len(iteraciones)){

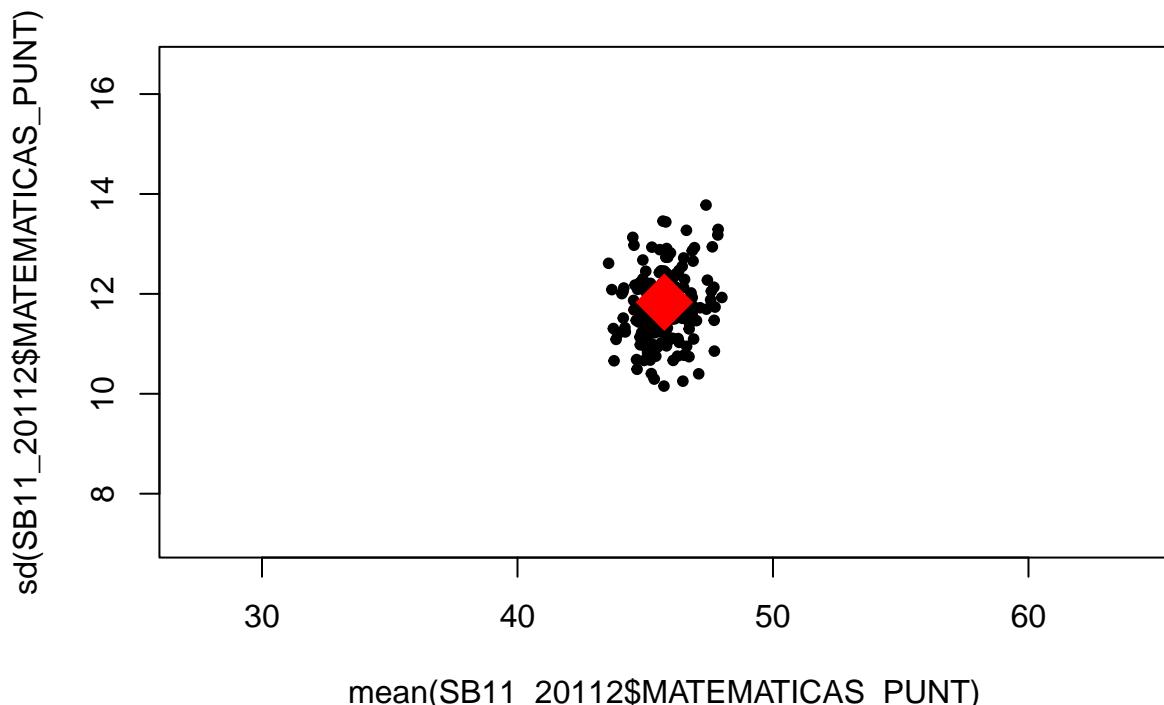
```

```

    points(
      mean(sample(SB11_20112$MATEMATICAS_PUNT, tamano_muestral)),
      sd(sample(SB11_20112$MATEMATICAS_PUNT, tamano_muestral)),
      pch = 20
    )
  }

  points(
    mean(SB11_20112$MATEMATICAS_PUNT),
    sd(SB11_20112$MATEMATICAS_PUNT),
    pch = 18,
    cex = 4,
    col = 'red'
)

```



Como se muestra la gráfica, los datos convergen a la media poblacional. Es decir, a medida que aumentemos el tamaño de la muestra, nuestras medias muestrales convergen a la media poblacional.

Para realizar un contraste de la media del ICFES en lenguaje, utilizaremos el test de t de Student para comparar la media de una muestra con la media poblacional esperada. Aún más vamos a separar a las personas que tienen carro y los que no tienen carro y, ver si esto tiene alguna incidencia a la hora presentar la prueba.

```
tamano_muestral <- 3000
iteraciones <- 100
```

```

poba_A <- SB11_20112$LENGUAJE_PUNT[SB11_20112$ECON_SN_AUTOMOVIL == 0]
poba_B <- SB11_20112$LENGUAJE_PUNT[SB11_20112$ECON_SN_AUTOMOVIL == 1]
media_pob_A <- mean(poba_A, na.rm = TRUE)
media_pob_B <- mean(poba_B, na.rm = TRUE)

plot(media_pob_A, media_pob_B, col = 4, pch = 20)
abline(0,1)

pval_A <- vector()
pval_B <- vector()

for(i in seq_len(iteraciones)){

  muestra <- sample(seq_len(nrow(SB11_20112)), tamano_muestral)

  cuales_A <- seq_len(nrow(SB11_20112)) %in% muestra & SB11_20112$ECON_SN_AUTOMOVIL == 0
  muestra_A <- SB11_20112$LENGUAJE_PUNT[cuales_A]

  cuales_B <- seq_len(nrow(SB11_20112)) %in% muestra & SB11_20112$ECON_SN_AUTOMOVIL == 1
  muestra_B <- SB11_20112$LENGUAJE_PUNT[cuales_B]

  media_muestra_A <- mean(muestra_A, na.rm = TRUE)
  t_test_A <- t.test(muestra_A)
  intervalo_A <- t_test_A$conf.int
  LI_A <- min(intervalo_A)
  LS_A <- max(intervalo_A)

  media_muestra_B <- mean(muestra_B, na.rm = TRUE)
  t_test_B <- t.test(muestra_B, na.rm = TRUE)
  intervalo_B <- t_test_B$conf.int
  LI_B <- min(intervalo_B)
  LS_B <- max(intervalo_B)

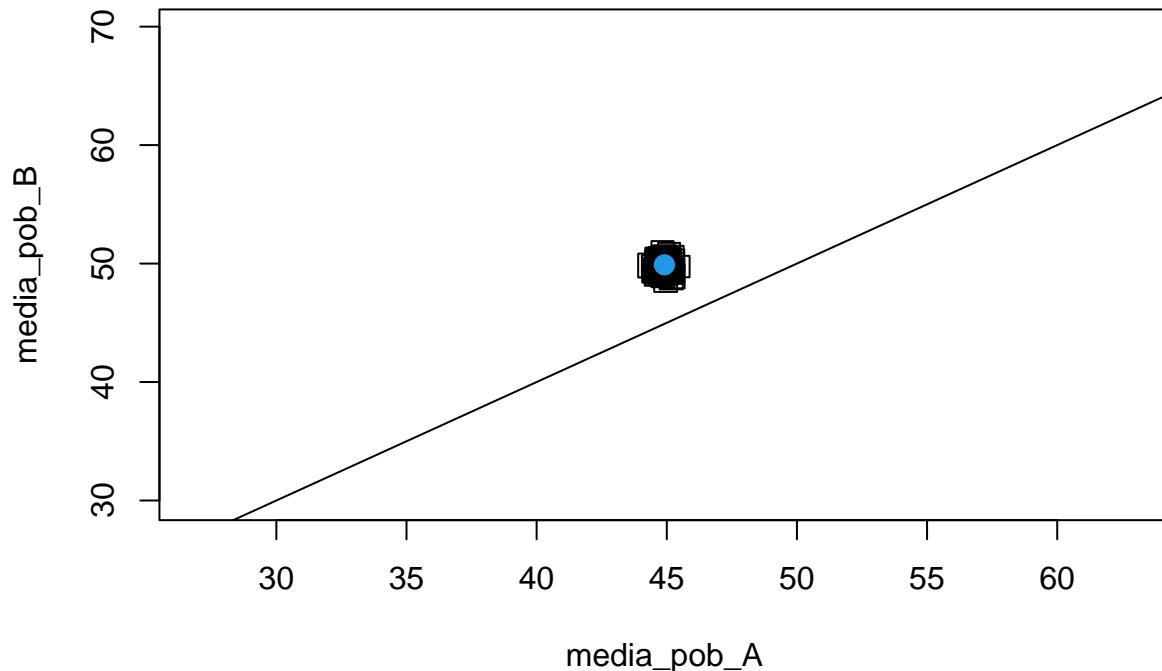
  pval_A[i] <- t_test_A$p.value
  pval_B[i] <- t_test_B$p.value

  # points(media_muestra_A, media_muestra_B, col = 2, pch = 20)
  rect(LI_A, LI_B, LS_A, LS_B)

}

points(media_pob_A, media_pob_B, col = 4, pch = 20, cex = 2)

```



La recta en la gráfica es la recta $y = x$, el punto azul la media, lo que esto nos muestra es que en general si hay una mejora en lenguaje a la hora de presentar la prueba si tienes un carro. Los rectángulos en la gráfica son los intervalos de confianza de la media para cada una de las muestras con la cantidad de iteraciones tomadas.

```
mean(pval_A); mean(pval_B); max(pval_A); max(pval_B)

## [1] 0

## [1] 4.09741e-318

## [1] 0

## [1] 4.054804e-316
```

Note que en cada uno de los casos, el valor máximo de los p-valores es cero. Es decir, que no hay información significativa, desde el punto de vista estadístico, para no decir que los intervalos de confianza están bien y que finalmente nos permite concluir, que hay una mejora en las resultados de lenguaje de pruebas ICFES si una persona tiene vehículo.

Veamos que se tiene de las varianzas de ambas poblaciones:

```
var.test(pobla_A, pobla_B)
```

```

## 
## F test to compare two variances
## 
## data: poblA and poblB
## F = 0.95803, num df = 444614, denom df = 95769, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.9494679 0.9666296
## sample estimates:
## ratio of variances
## 0.9580298

```

Usamos el F-test para comparar las dos varianzas y determinar si hay una diferencia significativa entre ellas. En este caso, después de realizar la prueba F en los datos de las dos poblaciones, poblA y poblB, y se ha obtenido un valor F de 0.95803.

El valor p es menor que 2.2e-16, lo que indica que hay una diferencia significativa entre las varianzas de los dos grupos de datos. Además, el intervalo de confianza del 95% para la razón de varianzas está entre 0.9494679 y 0.9666296, lo que confirma que las varianzas de las dos poblaciones son diferentes en un 95% de confianza. Por lo tanto, se puede concluir que la hipótesis alternativa de que la razón de varianzas verdadera no es igual a 1 es correcta.

Por otra lado, ¿que podemos afirmar de la normalidad de estas muestras?. Realizamos un test de Lilliefors, que nos ayudara a dar respuesta a esto.

```

install.packages("nortest")
library(nortest)
lillie.test(poblA)

```

```

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data: poblA
## D = 0.057262, p-value < 2.2e-16

```

Dado que el p-valor es aproimadamente cero, lo que nos dice que hay una diferencia significativa entre como se distribuyen los datos en esta muestra y una distribución normal. Es decir, los datos de la poblA no siguen una distribución normal. Ahora, poblA que es una muestra de las notas, de las personas que presentaron el examen ICFES, muestra que tiene la condición que estas personas no tiene vehículo. Si la muestra no tiene una distribución normal, lo tendrá la población?

```
wilcox.test(SB11_20112$LENGUAJE_PUNT)
```

```

## 
## Wilcoxon signed rank test with continuity correction
## 
## data: SB11_20112$LENGUAJE_PUNT
## V = 1.4555e+11, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 0

```

Lo que nos dice el test de Kolmogorov-Smirnov, en este caso, es que no. Tenemos un p-valor muy cercano a cero lo que nos dice la hipótesis nula de la supuesta normalidad de los datos es rechazada. Es decir, las muestras no son normales, porque en general la población no es normal.

Regresión Lineal

Finalmente, llevaremos a cabo un análisis de regresión múltiple y uno de regresión simple. Con el segundo podremos visualizar la recta de regresión. Con esto, vamos a predecir el puntaje del ICFES en matemáticas a partir de variables socioeconómicas. Las variables que usaremos son las siguientes:

- “ECON_PERSONAS_HOGAR”: Cuantas personas viven en el hogar.
- “ECON_CUARTOS”: Número de cuartos en el hogar.
- “ECON_SN_LAVADORA”: Si tiene lavadora.
- “ECON_SN_NEVERA”: Si tiene nevera.
- “ECON_SN_HORNO”: Si tiene horno.
- “ECON_SN_DVD”: Si tiene DVD.
- “ECON_SN_MICROHONDAS”: Si tiene microondas
- “ECON_SN_AUTOMOVIL”: Si tiene automóvil.
- “MATEMATICAS_PUNT”: Puntaje en matemáticas.

```
#Creamos un dataframe con las variables que vamos a usar.
df <- data.frame (
  X1 <- SB11_20112$ECON_PERSONAS_HOGAR,
  X2 <- SB11_20112$ECON_CUARTOS,
  X3 <- SB11_20112$ECON_SN_LAVADORA,
  X4 <- SB11_20112$ECON_SN_NEVERA,
  X5 <- SB11_20112$ECON_SN_HORNO,
  X6 <- SB11_20112$ECON_SN_DVD,
  X7 <- SB11_20112$ECON_SN_MICROHONDAS,
  X8 <- SB11_20112$ECON_SN_AUTOMOVI,
  X9 <- SB11_20112$ECON_SN_INTERNET,
  Y <- SB11_20112$MATEMATICAS_PUNT
)
#df2 <- na.omit(df)
```

```
#Creamos la regresión linea.
model <- lm(Y ~ ., data = df)
```

```
summary(model)

##
## Call:
## lm(formula = Y ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -4.517e-11  0.000e+00  0.000e+00  0.000e+00  7.200e-14 
## 
## Coefficients:
## (Intercept)          Estimate Std. Error t value Pr(>|t|)    
##                (Intercept) 2.323e-13 4.751e-16 4.890e+02 < 2e-16
```

```

## X1....SB11_20112.ECON_PERSONAS_HOGAR -2.797e-15 5.363e-17 -5.215e+01 < 2e-16
## X2....SB11_20112.ECON_CUARTOS      1.749e-15 9.391e-17 1.862e+01 < 2e-16
## X3....SB11_20112.ECON_SN_LAVADORA 5.623e-15 2.052e-16 2.741e+01 < 2e-16
## X4....SB11_20112.ECON_SN_NEVERA   7.163e-15 2.728e-16 2.625e+01 < 2e-16
## X5....SB11_20112.ECON_SN_HORNO   4.500e-15 1.896e-16 2.374e+01 < 2e-16
## X6....SB11_20112.ECON_SN_DVD     1.490e-15 1.935e-16 7.702e+00 1.35e-14
## X7....SB11_20112.ECON_SN_MICROHONDAS 1.422e-15 2.248e-16 6.327e+00 2.50e-10
## X8....SB11_20112.ECON_SN_AUTOMOVI 1.450e-14 2.465e-16 5.884e+01 < 2e-16
## X9....SB11_20112.ECON_SN_INTERNET 2.156e-14 2.101e-16 1.026e+02 < 2e-16
## Y....SB11_20112.MATEMATICAS_PUNT 1.000e+00 7.393e-18 1.353e+17 < 2e-16
##
## (Intercept) ***
## X1....SB11_20112.ECON_PERSONAS_HOGAR ***
## X2....SB11_20112.ECON_CUARTOS ***
## X3....SB11_20112.ECON_SN_LAVADORA ***
## X4....SB11_20112.ECON_SN_NEVERA ***
## X5....SB11_20112.ECON_SN_HORNO ***
## X6....SB11_20112.ECON_SN_DVD ***
## X7....SB11_20112.ECON_SN_MICROHONDAS ***
## X8....SB11_20112.ECON_SN_AUTOMOVI ***
## X9....SB11_20112.ECON_SN_INTERNET ***
## Y....SB11_20112.MATEMATICAS_PUNT ***
##
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.146e-14 on 540215 degrees of freedom
## (264 observations deleted due to missingness)
## Multiple R-squared: 1, Adjusted R-squared: 1
## F-statistic: 2.003e+33 on 10 and 540215 DF, p-value: < 2.2e-16

```

Teniendo en cuenta que ‘R-squared’ y el ‘Adjusted R-squared’ son medidas para evaluar la precisión del modelo de regresión lineal. En este caso, ambos tienen un valor de 1, es decir indica un ajuste perfecto del modelo a nuestros datos. Sin embargo, este modelo es demasiado perfecto, lo que puede ser un indicio de que está sobreajustando los datos y no será capaz de generalizar bien a datos nuevos. Por lo tanto, examinaremos el modelo con más detalle, analizaremos un poco más la relación de las otras variables con la variable objetivo. Podemos realizar una regresión con una o dos variables y ver los resultados del modelo.

Esto nos ayudará a entender que está pasando en el modelado y si el modelo es el adecuado para nuestro problema.

```

#pairs(df,
#       upper.panel = NULL
#       diag.panel = panel.hist)

```

Realizar una gráfica de correlación es muy poco útil, aunque podría intentarse, ya que dada la cantidad de datos que hay, solo se ven nubes de puntos, lo cual poco nos deja para interpretación.

```

#Eliminamos todos los datos que tienen na en nuestro df.
df2 <- na.omit(df)

```

```
cor(df2)
```

```
## X1....SB11_20112.ECON_PERSONAS_HOGAR
```

## X1....SB11_20112.ECON_PERSONAS_HOGAR	1.00000000
## X2....SB11_20112.ECON_CUARTOS	0.46224581
## X3....SB11_20112.ECON_SN_LAVADORA	-0.05240400
## X4....SB11_20112.ECON_SN_NEVERA	-0.06333156
## X5....SB11_20112.ECON_SN_HORNO	-0.05488611
## X6....SB11_20112.ECON_SN_DVD	-0.01268964
## X7....SB11_20112.ECON_SN_MICROHONDAS	-0.08423803
## X8....SB11_20112.ECON_SN_AUTOMOVI	-0.05973826
## X9....SB11_20112.ECON_SN_INTERNET	-0.10807852
## Y....SB11_20112.MATEMATICAS_PUNT	-0.09684992
##	
X2....SB11_20112.ECON_CUARTOS	
## X1....SB11_20112.ECON_PERSONAS_HOGAR	0.46224581
## X2....SB11_20112.ECON_CUARTOS	1.00000000
## X3....SB11_20112.ECON_SN_LAVADORA	0.15261973
## X4....SB11_20112.ECON_SN_NEVERA	0.11125616
## X5....SB11_20112.ECON_SN_HORNO	0.12909461
## X6....SB11_20112.ECON_SN_DVD	0.13024146
## X7....SB11_20112.ECON_SN_MICROHONDAS	0.14349865
## X8....SB11_20112.ECON_SN_AUTOMOVI	0.14324161
## X9....SB11_20112.ECON_SN_INTERNET	0.16496249
## Y....SB11_20112.MATEMATICAS_PUNT	0.04837664
##	
X3....SB11_20112.ECON_SN_LAVADORA	
## X1....SB11_20112.ECON_PERSONAS_HOGAR	-0.05240400
## X2....SB11_20112.ECON_CUARTOS	0.1526197
## X3....SB11_20112.ECON_SN_LAVADORA	1.00000000
## X4....SB11_20112.ECON_SN_NEVERA	0.3348223
## X5....SB11_20112.ECON_SN_HORNO	0.3065270
## X6....SB11_20112.ECON_SN_DVD	0.3236858
## X7....SB11_20112.ECON_SN_MICROHONDAS	0.3631712
## X8....SB11_20112.ECON_SN_AUTOMOVI	0.2806587
## X9....SB11_20112.ECON_SN_INTERNET	0.4345005
## Y....SB11_20112.MATEMATICAS_PUNT	0.1768446
##	
X4....SB11_20112.ECON_SN_NEVERA	
## X1....SB11_20112.ECON_PERSONAS_HOGAR	-0.06333156
## X2....SB11_20112.ECON_CUARTOS	0.11125616
## X3....SB11_20112.ECON_SN_LAVADORA	0.33482228
## X4....SB11_20112.ECON_SN_NEVERA	1.00000000
## X5....SB11_20112.ECON_SN_HORNO	0.23200545
## X6....SB11_20112.ECON_SN_DVD	0.24182670
## X7....SB11_20112.ECON_SN_MICROHONDAS	0.19076413
## X8....SB11_20112.ECON_SN_AUTOMOVI	0.14719835
## X9....SB11_20112.ECON_SN_INTERNET	0.25312409
## Y....SB11_20112.MATEMATICAS_PUNT	0.12735575
##	
X5....SB11_20112.ECON_SN_HORNO	
## X1....SB11_20112.ECON_PERSONAS_HOGAR	-0.05488611
## X2....SB11_20112.ECON_CUARTOS	0.12909461
## X3....SB11_20112.ECON_SN_LAVADORA	0.30652703
## X4....SB11_20112.ECON_SN_NEVERA	0.23200545
## X5....SB11_20112.ECON_SN_HORNO	1.00000000
## X6....SB11_20112.ECON_SN_DVD	0.23357318
## X7....SB11_20112.ECON_SN_MICROHONDAS	0.33151982
## X8....SB11_20112.ECON_SN_AUTOMOVI	0.28023349
## X9....SB11_20112.ECON_SN_INTERNET	0.32698957
## Y....SB11_20112.MATEMATICAS_PUNT	0.14769268

```

## X6....SB11_20112.ECON_SN_DVD
## X1....SB11_20112.ECON_PERSONAS_HOGAR -0.01268964
## X2....SB11_20112.ECON_CUARTOS 0.13024146
## X3....SB11_20112.ECON_SN_LAVADORA 0.32368582
## X4....SB11_20112.ECON_SN_NEVERA 0.24182670
## X5....SB11_20112.ECON_SN_HORNO 0.23357318
## X6....SB11_20112.ECON_SN_DVD 1.00000000
## X7....SB11_20112.ECON_SN_MICROHONDAS 0.28039276
## X8....SB11_20112.ECON_SN_AUTOMOVI 0.21446948
## X9....SB11_20112.ECON_SN_INTERNET 0.26844025
## Y....SB11_20112.MATEMATICAS_PUNT 0.11180059
## X7....SB11_20112.ECON_SN_MICROHONDAS
## X1....SB11_20112.ECON_PERSONAS_HOGAR -0.08423803
## X2....SB11_20112.ECON_CUARTOS 0.14349865
## X3....SB11_20112.ECON_SN_LAVADORA 0.36317116
## X4....SB11_20112.ECON_SN_NEVERA 0.19076413
## X5....SB11_20112.ECON_SN_HORNO 0.33151982
## X6....SB11_20112.ECON_SN_DVD 0.28039276
## X7....SB11_20112.ECON_SN_MICROHONDAS 1.00000000
## X8....SB11_20112.ECON_SN_AUTOMOVI 0.34913471
## X9....SB11_20112.ECON_SN_INTERNET 0.39945285
## Y....SB11_20112.MATEMATICAS_PUNT 0.15308637
## X8....SB11_20112.ECON_SN_AUTOMOVI
## X1....SB11_20112.ECON_PERSONAS_HOGAR -0.05973826
## X2....SB11_20112.ECON_CUARTOS 0.14324161
## X3....SB11_20112.ECON_SN_LAVADORA 0.28065874
## X4....SB11_20112.ECON_SN_NEVERA 0.14719835
## X5....SB11_20112.ECON_SN_HORNO 0.28023349
## X6....SB11_20112.ECON_SN_DVD 0.21446948
## X7....SB11_20112.ECON_SN_MICROHONDAS 0.34913471
## X8....SB11_20112.ECON_SN_AUTOMOVI 1.00000000
## X9....SB11_20112.ECON_SN_INTERNET 0.35521295
## Y....SB11_20112.MATEMATICAS_PUNT 0.18563794
## X9....SB11_20112.ECON_SN_INTERNET
## X1....SB11_20112.ECON_PERSONAS_HOGAR -0.1080785
## X2....SB11_20112.ECON_CUARTOS 0.1649625
## X3....SB11_20112.ECON_SN_LAVADORA 0.4345005
## X4....SB11_20112.ECON_SN_NEVERA 0.2531241
## X5....SB11_20112.ECON_SN_HORNO 0.3269896
## X6....SB11_20112.ECON_SN_DVD 0.2684402
## X7....SB11_20112.ECON_SN_MICROHONDAS 0.3994528
## X8....SB11_20112.ECON_SN_AUTOMOVI 0.3552129
## X9....SB11_20112.ECON_SN_INTERNET 1.0000000
## Y....SB11_20112.MATEMATICAS_PUNT 0.2527238
## Y....SB11_20112.MATEMATICAS_PUNT
## X1....SB11_20112.ECON_PERSONAS_HOGAR -0.09684992
## X2....SB11_20112.ECON_CUARTOS 0.04837664
## X3....SB11_20112.ECON_SN_LAVADORA 0.17684457
## X4....SB11_20112.ECON_SN_NEVERA 0.12735575
## X5....SB11_20112.ECON_SN_HORNO 0.14769268
## X6....SB11_20112.ECON_SN_DVD 0.11180059
## X7....SB11_20112.ECON_SN_MICROHONDAS 0.15308637
## X8....SB11_20112.ECON_SN_AUTOMOVI 0.18563794
## X9....SB11_20112.ECON_SN_INTERNET 0.25272376

```

```
## Y....SB11_20112.MATEMATICAS_PUNT 1.00000000
```

En general, la correlación entre las variables es muy baja. Esto es positivo, ya que nos permite descartar la idea de que hay información redundante en nuestro modelo inicial. Ahora podemos usar una regresión con dos variables independientes para intentar predecir nuestra nota.

```
model2 <- lm(Y~X8+X9, data = df)
```

```
summary(model2)
```

```
##  
## Call:  
## lm(formula = Y ~ X8 + X9, data = df)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -51.712  -8.126  -0.126   7.688  58.874  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 43.12616    0.02005 2151.25 <2e-16 ***  
## X8          3.39983    0.04339   78.36 <2e-16 ***  
## X9          5.18609    0.03397  152.65 <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 11.38 on 540382 degrees of freedom  
##   (105 observations deleted due to missingness)  
## Multiple R-squared:  0.07438,   Adjusted R-squared:  0.07437  
## F-statistic: 2.171e+04 on 2 and 540382 DF,  p-value: < 2.2e-16
```

```
summary(model2)$r.squared; model2$coefficients
```

```
## [1] 0.07437502
```

```
## (Intercept)          X8          X9  
## 43.126159    3.399834    5.186088
```

En este caso, el resumen de los datos nos muestra que a pesar de que los coeficientes beta de las variables seleccionadas son muy relevantes para ajustar el modelo, el modelo generado tiene una precisión muy baja, ya que su R-cuadrado y su R-cuadrado ajustado no superan el 0.075. Esto indica que el modelo no es muy eficiente en la predicción de las notas, y podría ser necesario explorar otras variables o técnicas de modelado para mejorar su rendimiento, incluso modelos más sofisticados.

```
anova(model2)
```

```
## Analysis of Variance Table  
##  
## Response: Y  
##              Df  Sum Sq Mean Sq F value    Pr(>F)  
## X8            1 2607409 2607409   20117 < 2.2e-16 ***
```

```

## X9          1 3020338 3020338   23303 < 2.2e-16 ***
## Residuals 540382 70039426      130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Aún más, este resultado de ANOVA indica que las variables X8 y X9 tienen un efecto significativo en la respuesta Y. Esto se confirma por los valores de F y Pr(>F) obtenidos en la tabla, que son cero y sugieren una alta significancia estadística. Sin embargo, también se puede observar que hay un alto residuo en el modelo, lo que indica que hay una gran cantidad de variabilidad en los datos que no se explica por el modelo. Esto puede deberse a la presencia de otros factores que influyen en la respuesta Y y que no se han considerado en el análisis, como se mencionó en el párrafo anterior.

```
anova(model)
```

```

## Warning in anova.lm(model): ANOVA F-tests on an essentially perfect fit are
## unreliable

## Analysis of Variance Table

## Response: Y

##                                     Df  Sum Sq Mean Sq F value
## X1....SB11_20112.ECON_PERSONAS_HOGAR 1 709586 709586 1.8784e+32
## X2....SB11_20112.ECON_CUARTOS        1 834685 834685 2.2096e+32
## X3....SB11_20112.ECON_SN_LAVADORA    1 1797217 1797217 4.7576e+32
## X4....SB11_20112.ECON_SN_NEVERA      1 283589 283589 7.5072e+31
## X5....SB11_20112.ECON_SN_HORNO       1 474285 474285 1.2555e+32
## X6....SB11_20112.ECON_SN_DVD         1 85612 85612 2.2663e+31
## X7....SB11_20112.ECON_SN_MICROHONDAS 1 231997 231997 6.1415e+31
## X8....SB11_20112.ECON_SN_AUTOMOVI    1 767168 767168 2.0309e+32
## X9....SB11_20112.ECON_SN_INTERNET    1 1352332 1352332 3.5799e+32
## Y....SB11_20112.MATEMATICAS_PUNT     1 69113153 69113153 1.8296e+34
## Residuals                           540215      0      0
##                                     Pr(>F)
## X1....SB11_20112.ECON_PERSONAS_HOGAR < 2.2e-16 ***
## X2....SB11_20112.ECON_CUARTOS        < 2.2e-16 ***
## X3....SB11_20112.ECON_SN_LAVADORA    < 2.2e-16 ***
## X4....SB11_20112.ECON_SN_NEVERA      < 2.2e-16 ***
## X5....SB11_20112.ECON_SN_HORNO       < 2.2e-16 ***
## X6....SB11_20112.ECON_SN_DVD         < 2.2e-16 ***
## X7....SB11_20112.ECON_SN_MICROHONDAS < 2.2e-16 ***
## X8....SB11_20112.ECON_SN_AUTOMOVI    < 2.2e-16 ***
## X9....SB11_20112.ECON_SN_INTERNET    < 2.2e-16 ***
## Y....SB11_20112.MATEMATICAS_PUNT     < 2.2e-16 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

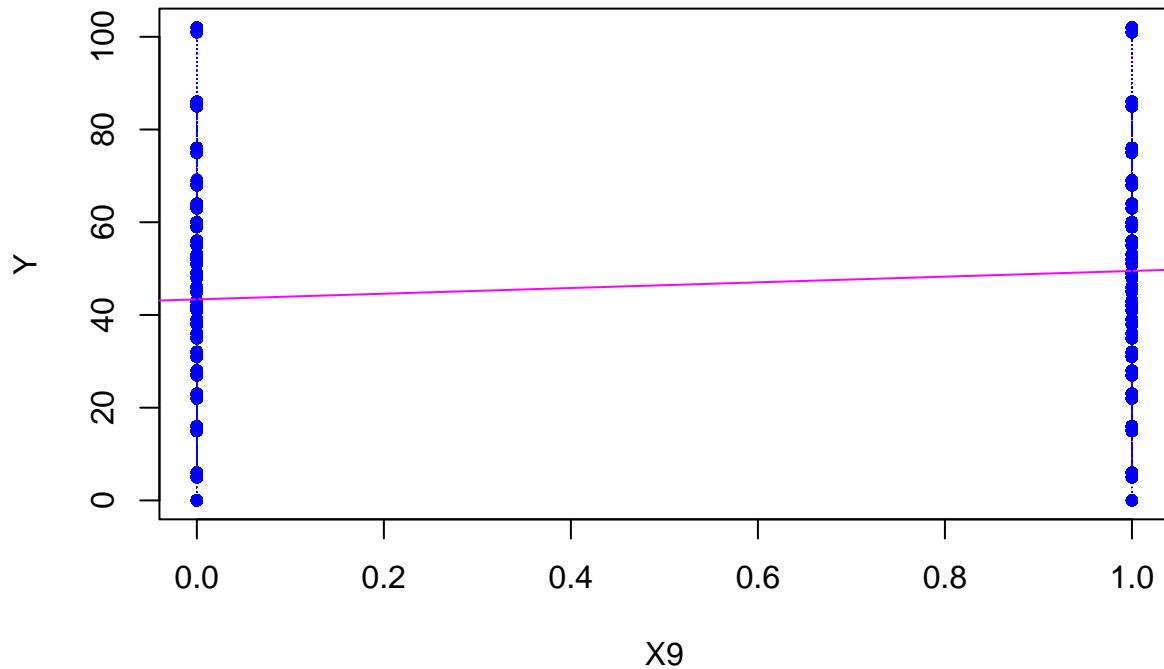
Aunque en el anova realizado al modelo original, con todas las variables, tiene un ajuste perfecto y un residuo demasiado bajo, como ya se había dicho es muy poco confiable, tal es así que R, lanza una advertencia hacia lo poco confiable que podría ser.

Finalmente realizaremos una regresión lineal simple y posteriormente realizaremos una gráfica que nos va a permitir ver de manera visual nuestros datos.

```

plot(X9,Y,col='blue',type='p',pch=19,cex=0.7)
mod<-lm(Y~X9)
abline(mod,col='magenta')
points(mean(X9),mean(Y),cex=0.7,col='blue')
text(173.5,63,expression(paste("(",bar(X9)," , ",bar(Y)," )",sep="")),cex=0.7)
segments(X9,Y,X9,fitted(mod),lty='dotted',col='blue',lwd=0.5)

```



```

b0<-round(coef(mod)[1],2)
#text(160,85,expression(paste(hat(beta)[0],"=", -84.05,sep="")),pos=1,cex=1)
#text(160,80,expression(paste(hat(beta)[1],"=", 0.86,sep="")),pos=1,cex=1)

```

Como se esperaba, la recta de regresión no es el modelo más adecuado para estos datos, como se puede observar en la gráfica. Por lo tanto, sus resultados deberían ser consistentes con lo que se ve en la gráfica.

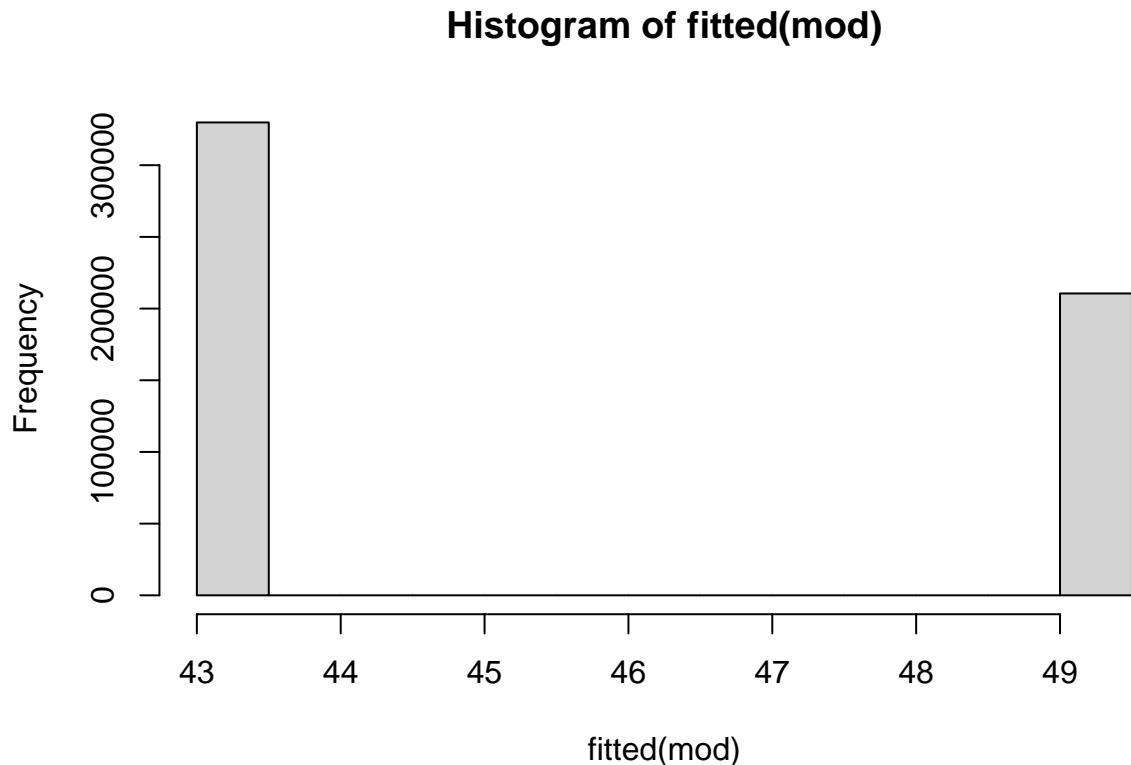
```

anova(mod)

## Analysis of Variance Table
##
## Response: Y
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## X9          1 4831860 4831860   36861 < 2.2e-16 ***
## Residuals 540383 70835313      131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

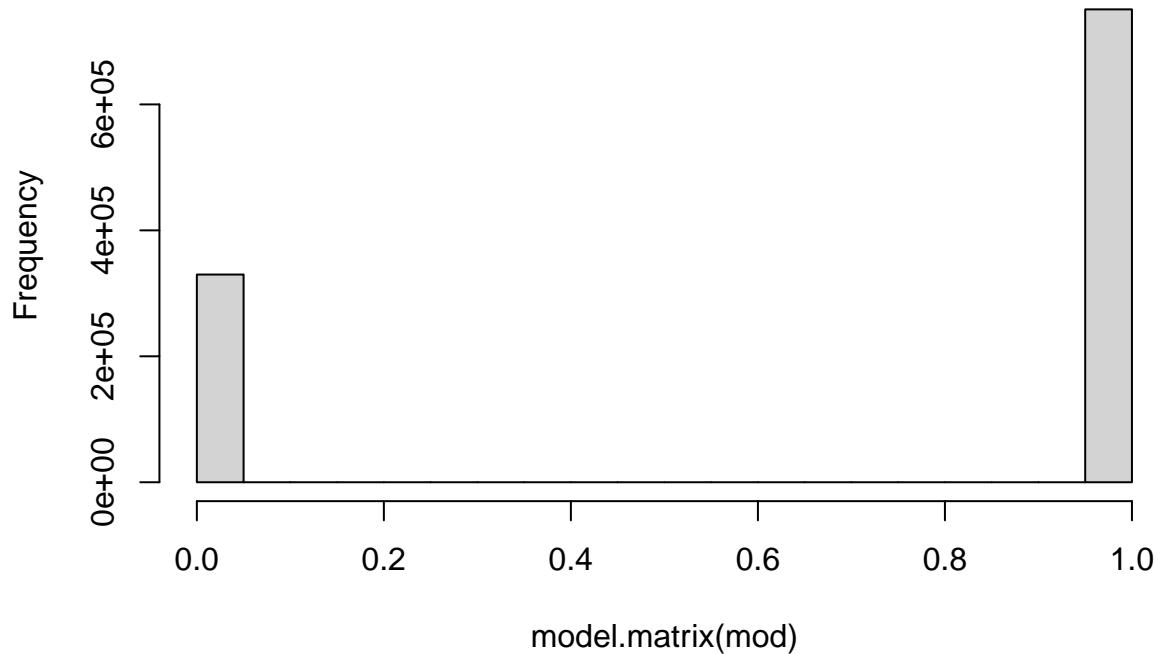
```
hist(fitted(mod))
```



Como el modelo lineal no es suficientemente preciso para predecir los resultados en matemáticas de cada persona, se utilizan dos valores de referencia: el 43 para aquellos que no tienen internet y el 49 para aquellos que sí tienen internet. Esto claramente muestra que la falta de variables en el modelo genera un gran sesgo y un alto residuo. Una de las principales debilidades del modelo es que, aunque parece haber una pequeña mejora en los resultados de matemáticas del ICFES en aquellos casos en los que las personas tienen internet (en promedio, 6 puntos más), el modelo no es capaz de capturar adecuadamente esta relación y por lo tanto no es confiable para hacer predicciones precisas, como se esperaba.

```
hist(model.matrix(mod))
```

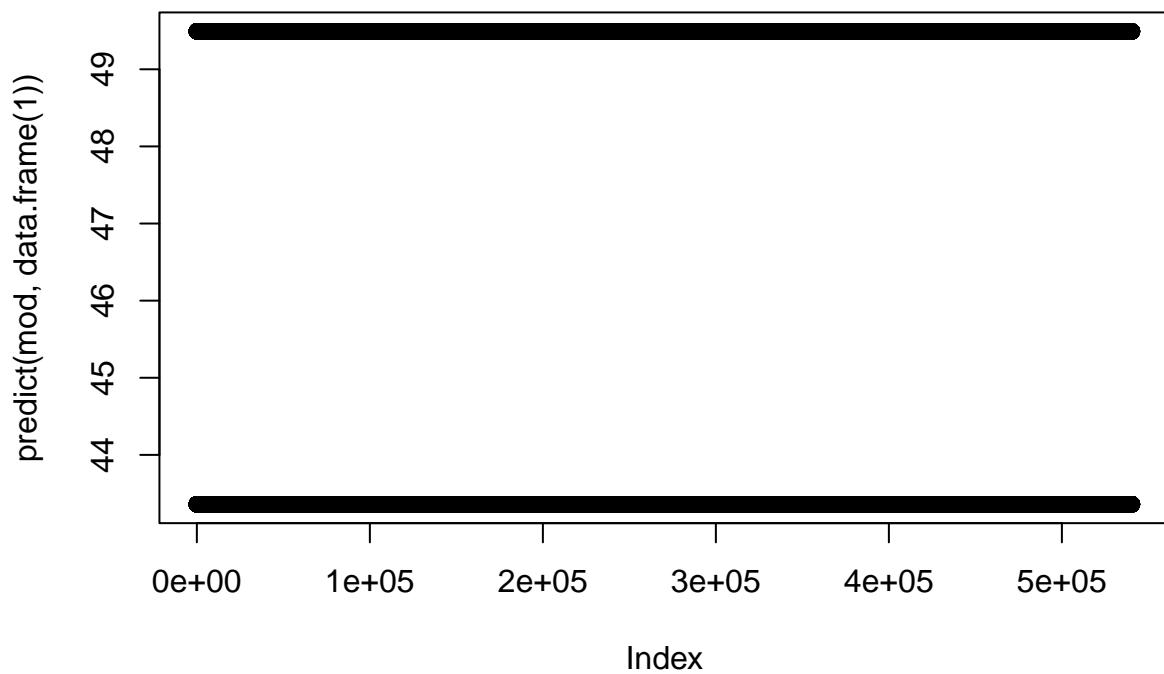
Histogram of model.matrix(mod)



Como era de esperarse, la gráfica refleja los valores que toma la variable, cero en caso de no tener internet y uno en caso de tener.

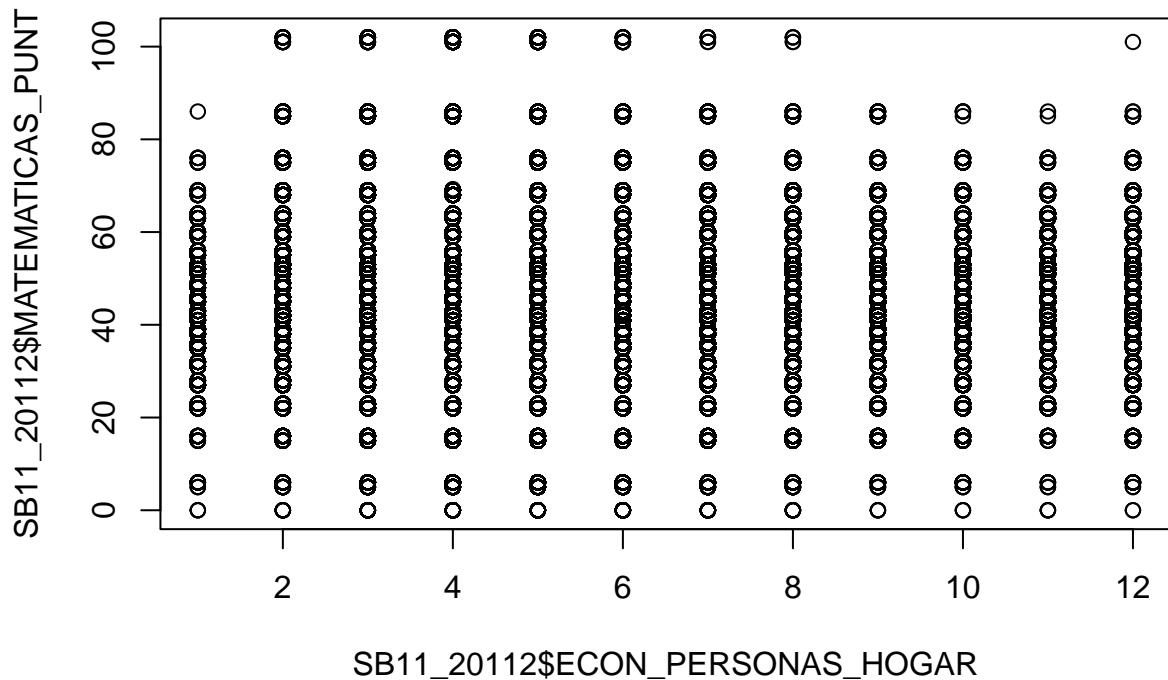
```
plot(predict(mod,data.frame(1)))
```

```
## Warning: 'newdata' had 1 row but variables found have 540490 rows
```

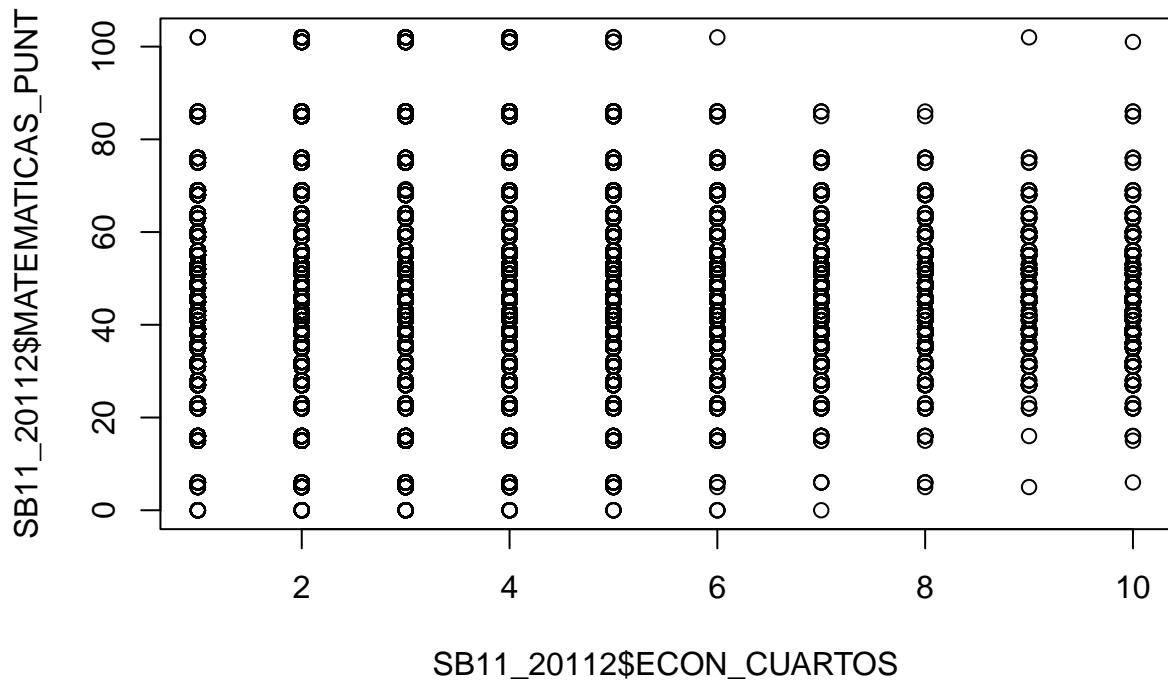


Como se ha reiterado este modelo solo puede predecir dos valores.

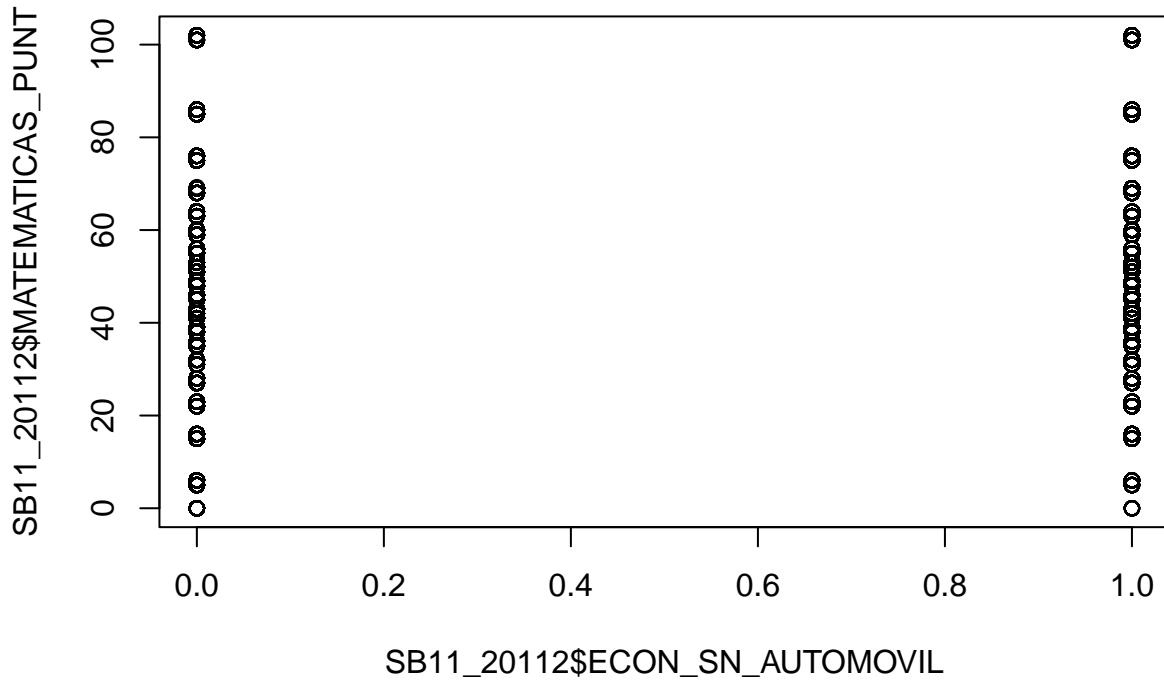
```
plot(SB11_20112$MATEMATICAS_PUNT ~ SB11_20112$ECON_PERSONAS_HOGAR)
```



```
plot(SB11_20112$MATEMATICAS_PUNT ~ SB11_20112$ECON_CUARTOS)
```



```
plot(SB11_20112$MATEMATICAS_PUNT ~ SB11_20112$ECON_SN_AUTOMOVIL)
```



Si se fija una variable y un valor de esa variable, es posible observar una distribución normal de los datos con una ligera inclinación hacia aquellos con mejores condiciones económicas, como ya se mencionó en el caso del acceso a internet. En ese sentido, estos datos podrían ser una buena opción para predecir el puntaje en matemáticas en las pruebas, pero requerirían un modelo más sofisticado para hacer predicciones precisas. Como se ha mencionado anteriormente, el modelo lineal no es suficientemente preciso y podría ser necesario explorar otras técnicas de modelado para mejorar su rendimiento.

Bibliografía

- [1] W. Larry, *All of Statistics*, 1st ed. Springer Science, 2004.
- [2] R Core Team, 'R: A Language and Environment for Statistical Computing'. Vienna, Austria, 2018. [Online]. Available: <https://www.R-project.org/>
- [3] Rs. Team, 'RStudio: Integrated Development Environment for R'. RStudio, Inc., 2022.
- [4] A. C. J. Luis, *Notas de Clase*. 2020.
- [5] ICFES, '<https://www.icfes.gov.co/>', Dec. 07, 2022.
- [6] Cruz Julian and Daniel, '<https://github.com/nebulae-co/saber>', Aug. 27, 2015.