Proyecto 1

Agustin Riquelme y Heriberto Espino

Para este modelo de regresión líneal se usó la base de datos lung cap, la cuál recolecta información de un grupo de personas acerca de si son fumadoras o no, las edades respectivas, además de la altura, el género y su FEV(Forced Expiratory Volumen), que mide la habilidad de expedir aire de los pulmones. En este caso, la variable que se tomará como predictora será la edad, sin emabrgo, de igual manera se seccionará en dos, en fumadores y no fumadores para un mejor análisis de los datos.

Para el caso general, donde se opta por juntar ambos grupos sin diferencia, tenemos los siguientes datos estadísticos, donde podemos observar que las personas que fueron parte de esta muestra van desde los 3 hasta los 19 años de edad con un promedio de 10 años y misma mediana, además, la edad que más se repite es de 9 años y además su varianza no difiere mucho de la media.

Edad

Datos.estadísticos	Valores
Mínimo	3.000
Primer cuantil	8.000
Mediana	10.000
Media	9.931
Tercer cuantil	12.000
Máximo	19.000
Moda	9.000
Varianza	8.726
Desviacion estándar	2.954

Para el caso del FEV, el mínimo es de 0.8 redondeado con un máximo de 3.8, y la media de este índice es de 2.6 con una varianza mucho más pequeña que la media en términos unitarios y porcentuales. Esta variable de FEV corresponderá a la variable de respuesta, pues se cree que el FEV va a depender de la edad de la persona en cuestión pero igualmente de si es fumador o no, pero esos datos se desglorazán a continuación.

Datos.estadísticos	Valores
Mínimo	0.791
Primer cuantil	1.981
Mediana	2.547
Media	2.637
Tercer cuantil	3.118
Máximo	5.793
Moda	3.082
Varianza	0.752
Desviacion estándar	0.867

Para el casode los fumadores, estos inician a partir de la edad de 9 años y termina en el mismo rango que los datos en conjunto, siendo 19 con un promedio de 14 años pero con 13 años como la cifra más repetida. Y a pesar de no ser mostrado, el porcentaje de fumadores dentro de la base corresponde a 10%.

Datos.estadísticos	Valores
Mínimo	9.000
Primer cuantil	12.000
Mediana	13.000
Media	13.523
Tercer cuantil	15.000
Máximo	19.000
Moda	13.000
Varianza	5.472
Desviacion estándar	2.339

Para el caso del índice FEV, el mínimo corresponde a 1.7 mientras que el máximo a 4.8, teniendo una media por encima de los datos totales sin embargo una varianza menor.

Datos.estadísticos	Valores
Mínimo	1.694
Primer cuantil	2.795
Mediana	3.169
Media	3.277
Tercer cuantil	3.751
Máximo	4.872
Moda	3.297
Varianza	0.562
Desviacion estándar	0.750

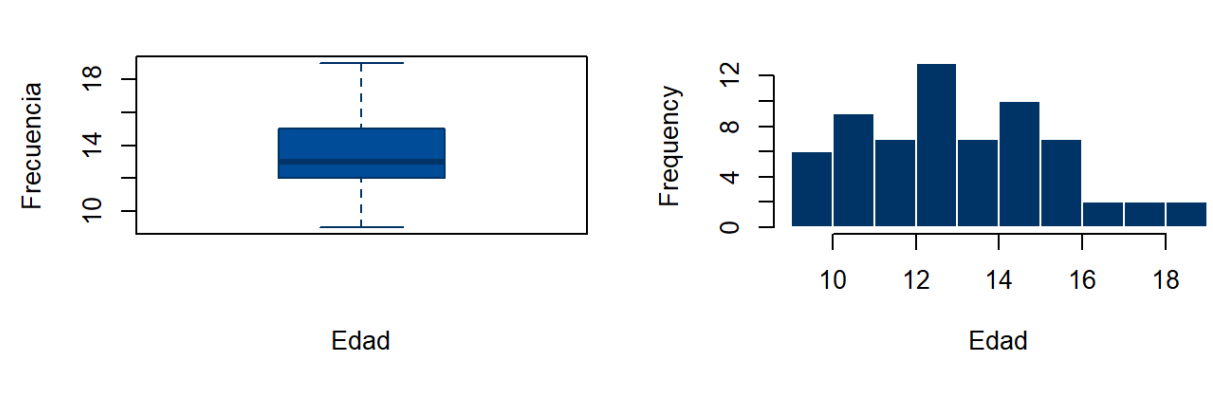
Ahora para el caso de los no fumadores, la edad más pequeña registrada es de 3 años, mientras que la máxima de igual manera corresponde a la máxima conjunta, en este caso, la media es menor, de 9 años al igual que su moda y mediana. Estos datos tienen una mayor varianza que los fumadores pero menor a la conjunta.

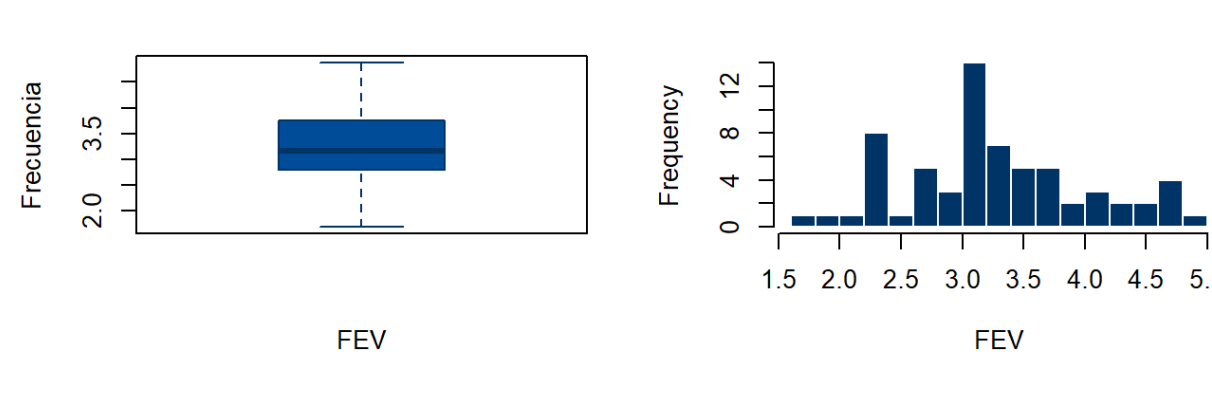
Datos.estadísticos	Valores
Mínimo	3.000
Primer cuantil	8.000
Mediana	9.000
Media	9.535
Tercer cuantil	11.000
Máximo	19.000
Moda	9.000
Varianza	7.511
Desviacion estándar	2.741

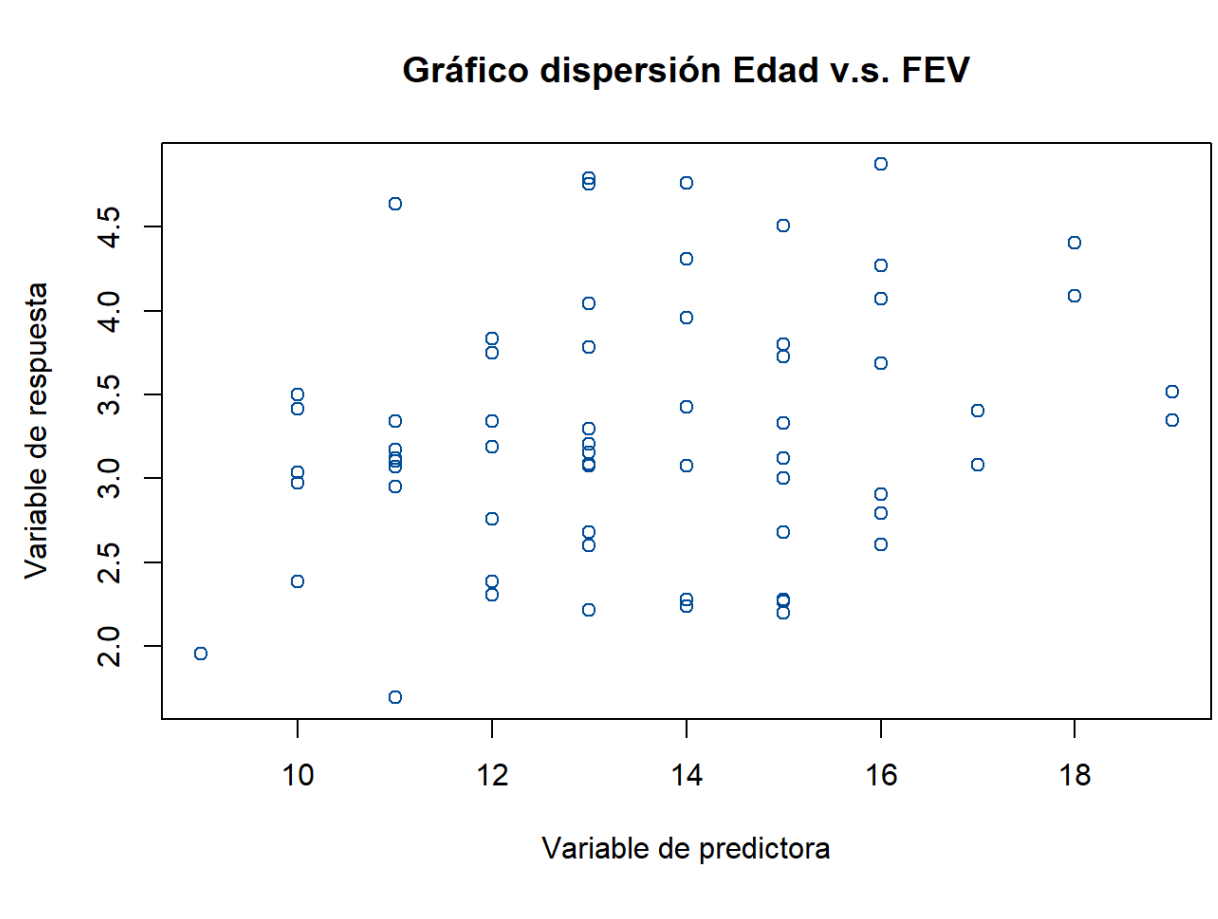
En el caso del índice FEV, se registró un mínimo de 0.8 y un máximo de 5.8. mayor que el caso de los fumadores. Sin embargo, los demás valores parecen ser más pequeños que los datos de fumadores, por lo que analizarlo de esta manera en conjunto y de dos grupos podrá determinar si el fumar tiene alguna consecuencia en este índice al igual que la edad.

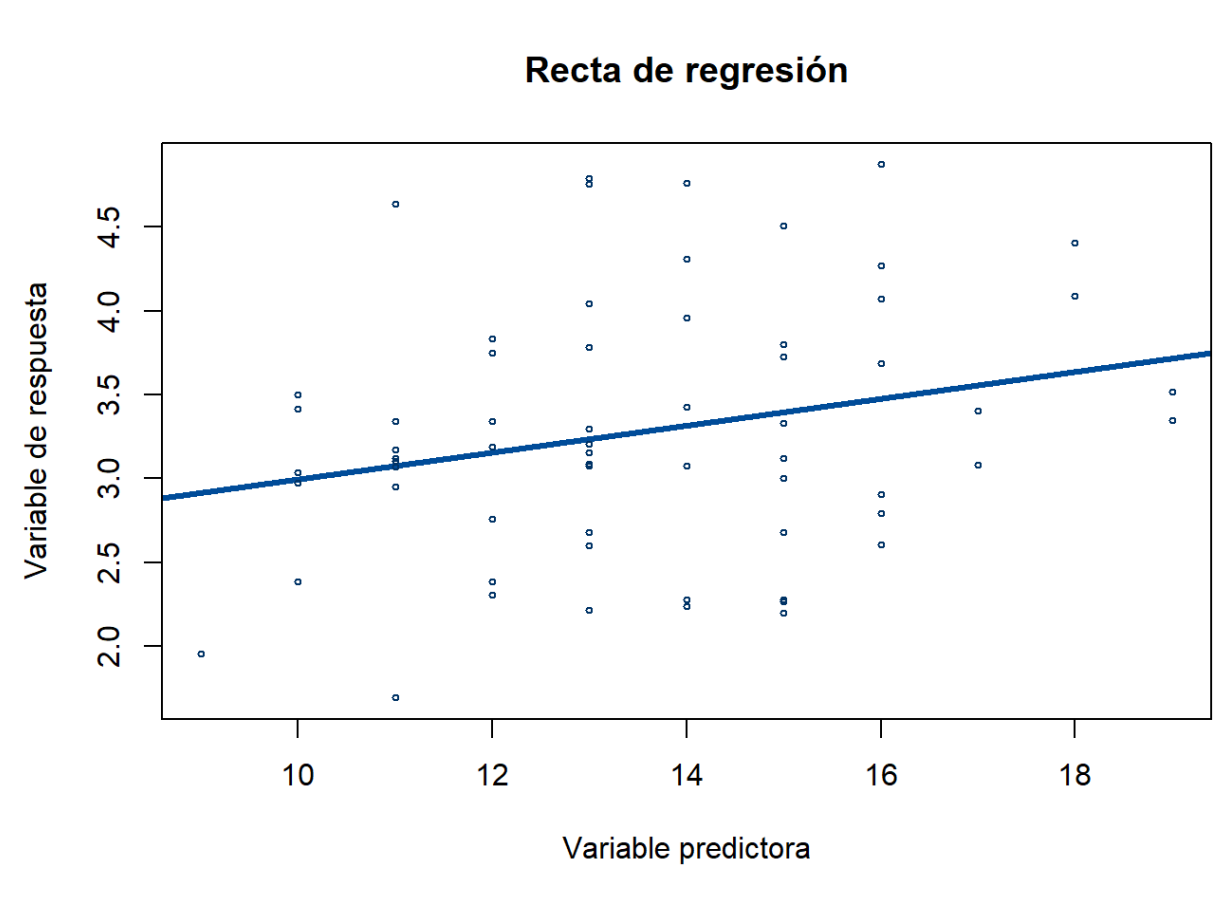
Datos.estadísticos	Valores
Mínimo	0.791
Primer cuantil	1.920
Mediana	2.465
Media	2.566
Tercer cuantil	3.048
Máximo	5.793
Moda	1.624
Varianza	0.723
Desviacion estándar	0.851

Entonces, para un mejor análisis, a partir de este momento, los datos serán seccionados en dos grupos: fumadores y no fumadores. En el primer caso, tendremos los siguientes histogramas y boxplot para el caso de las edades y su índice FEV. Visualmente no observamos ningún outlier en los datos, y no parece haber un comportamiento similar a una distribución normal.

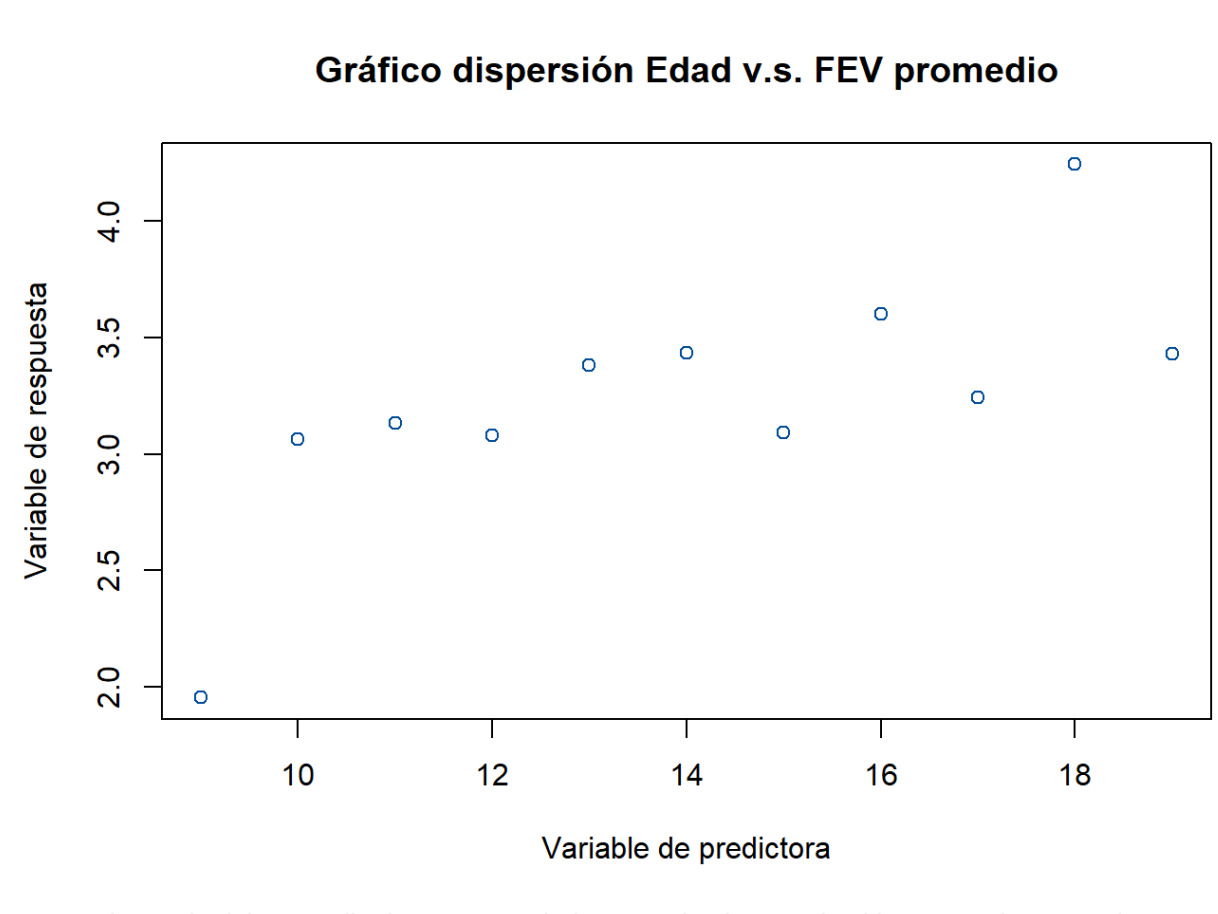




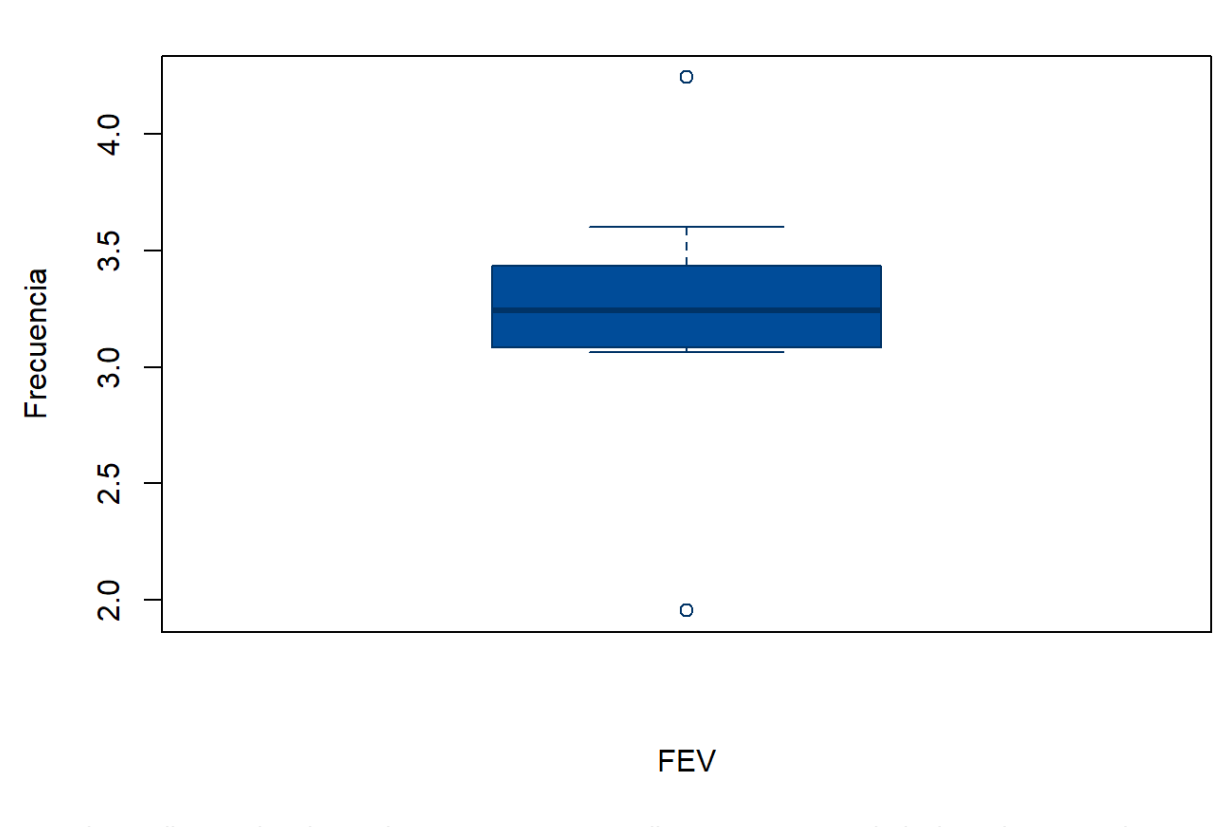




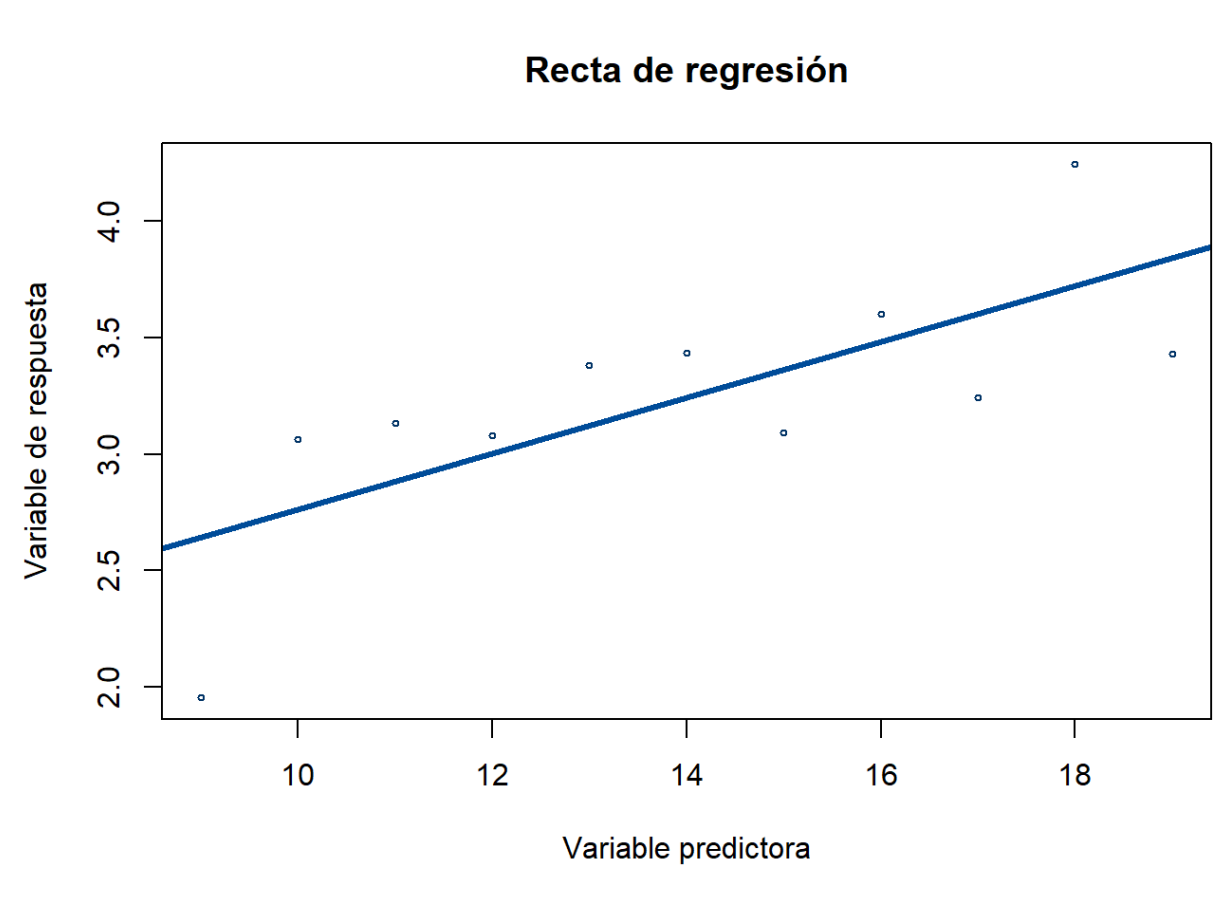
Como podemos observar, la recta de regresión pasa casi al centro de la nube de puntos dispersos, dicha relación puede ser vista de manera más fácil a través de un gráfico de dispersión donde incluyamos el FEV promedio por edad, el cuál es una medida más precisa para poder describir el impacto de la edad sobre este índice, podemos observarlo a continuación:

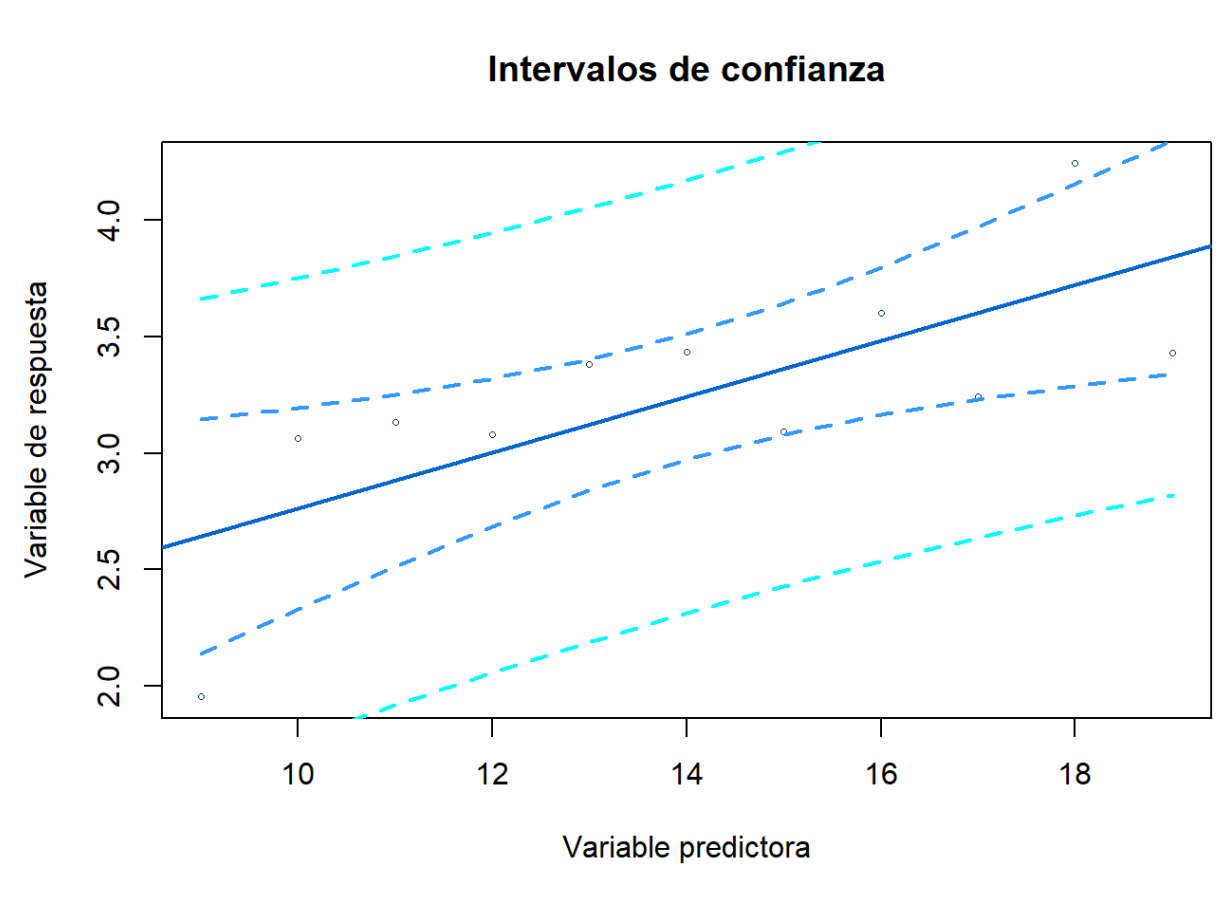


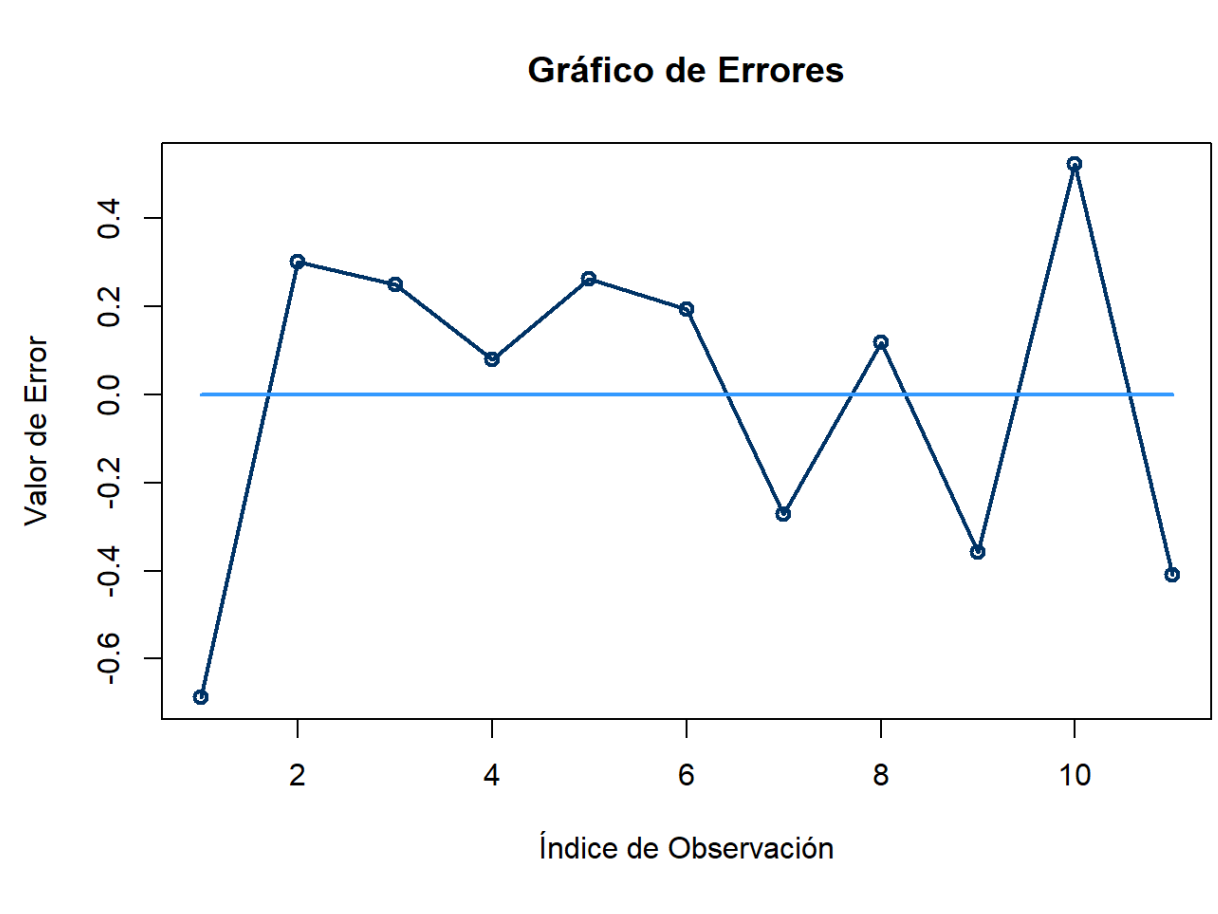
De esta otra manera, con el manejo del promedio de FEV por edad, se puede observar intuitivamente de una mejor manera la forma de la recta de regresión, pudiendo suponer que tiene relación positiva y con una pendiente menor a 1. Además podemos observar un posible outlier a la edad de 9 años, por lo que realizaremos de nuevo un boxplot para su análisis.

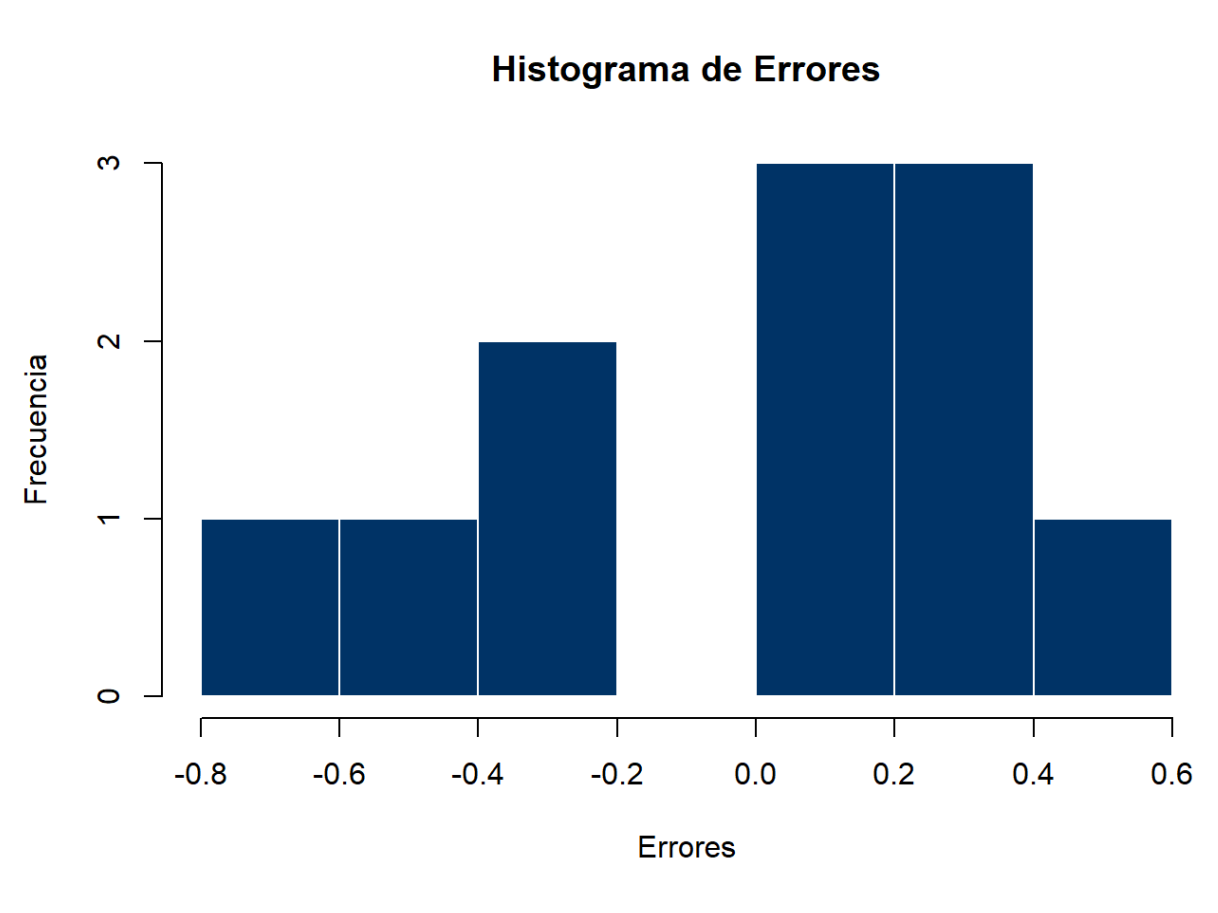


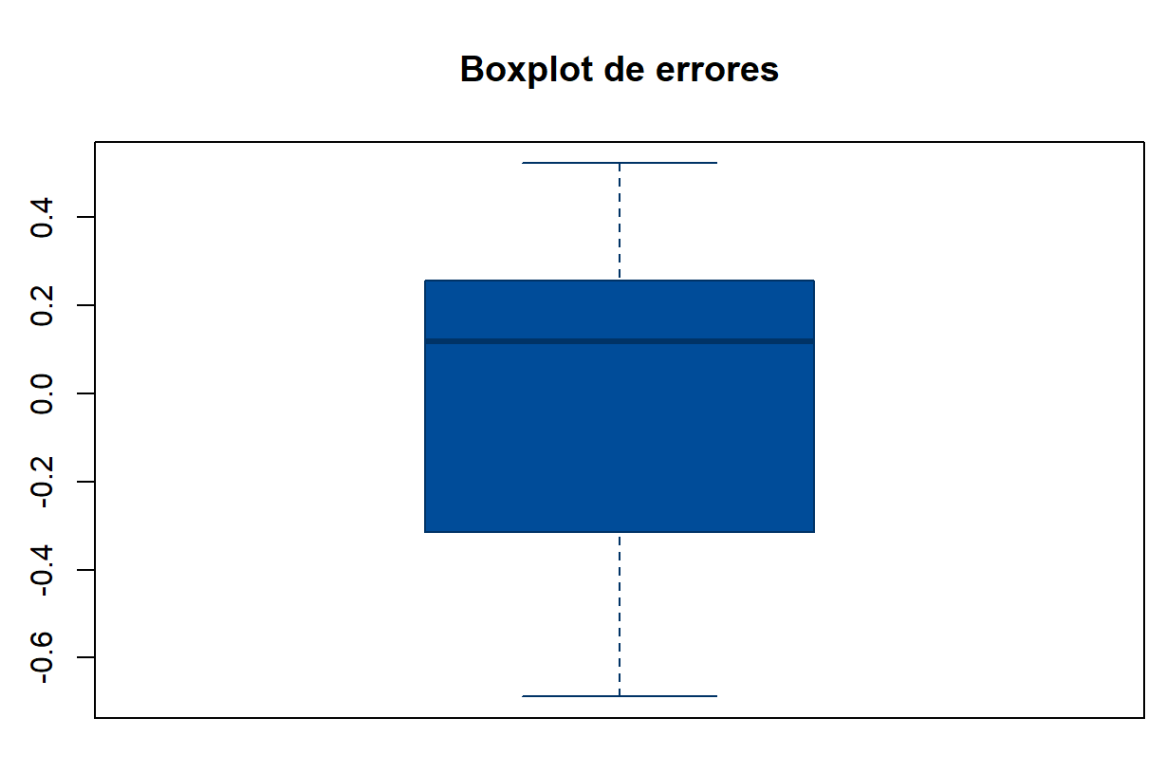
Aquí podemos observar dos outliers en los datos al sacar su FEV promedio respecto a su edad, sin embargo, podemos ver menos dispersos los datos para una mejor aproximación líneal que minimice la distancia entre todos los puntos y por ende los errores totales, esto se verá a continuación:

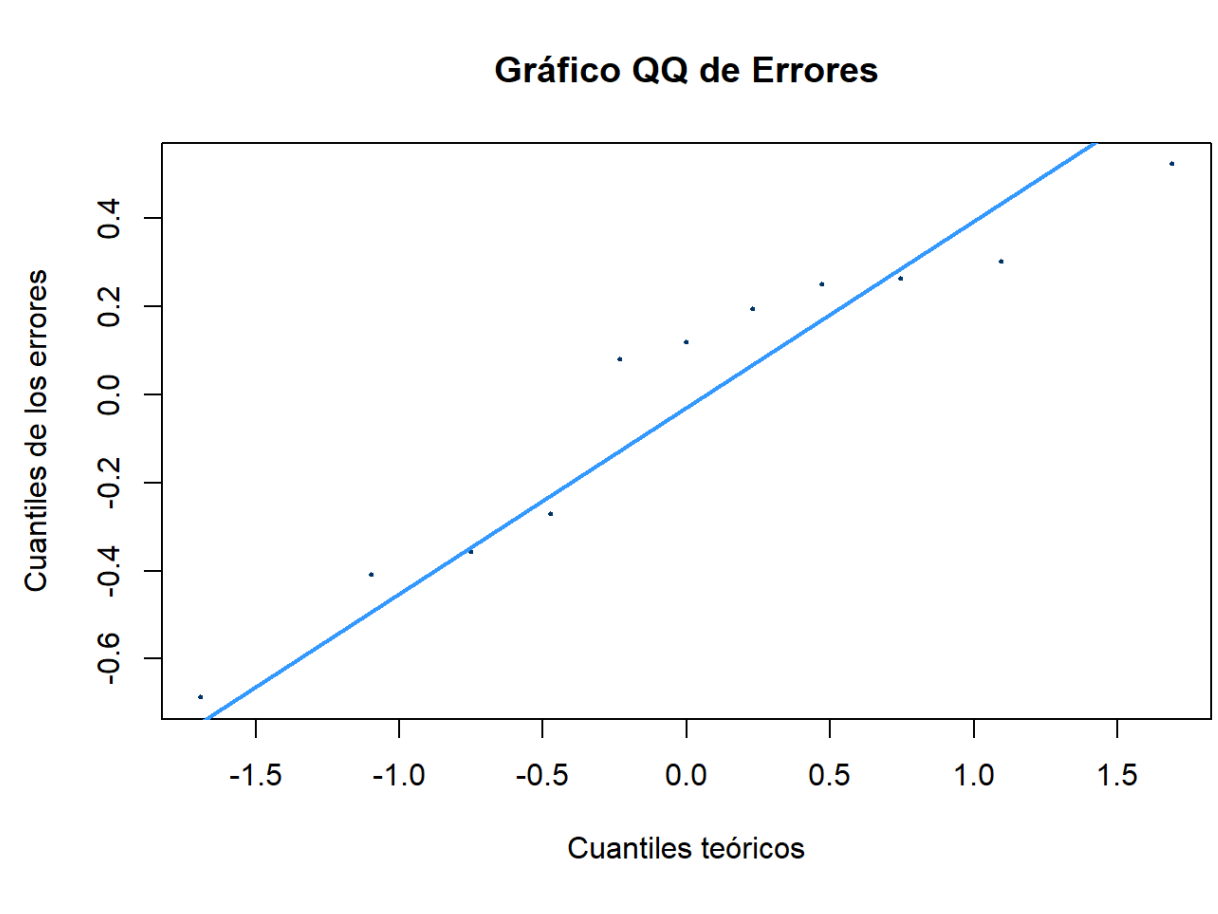












Informacion del modelo

Datos	Valores
α	0.05
β0	1.560257005217
β1	0.120099224109224
S	0.393903024150109
S^2	0.155159592434601
CV	12.1513270358503
df	9
qti	-2.2621571627982
qtd	2.2621571627982
tc	3.19777003936184
P valor	0.010870134685864
Mín. β1	0.0351389809644712
Máx. β1	0.205059467253977
Prom. err.	-3.5325278056255e-17
Var. error	0.139643633191141
Shap. P valor	0.366242866344028
Anderson P valor	0.218479236573656
Rho Pearson	0.729299260696424
Rho Spearman	0.772727272727273
dw	2.35620794814095

Inferencia para β1 H0: β1 = 0 vs H1: β1 ≠ 0

Prueba.de.hipótesis	Conclusión
tc	Rechazamos H0. β1 ≠ 0
p valor	Rechazamos H0. β1 ≠ 0

Coeficiente de correlación

Rho	Conclusión
Pearson	Hay una correlación fuerte
Spearman	Hay una correlación fuerte

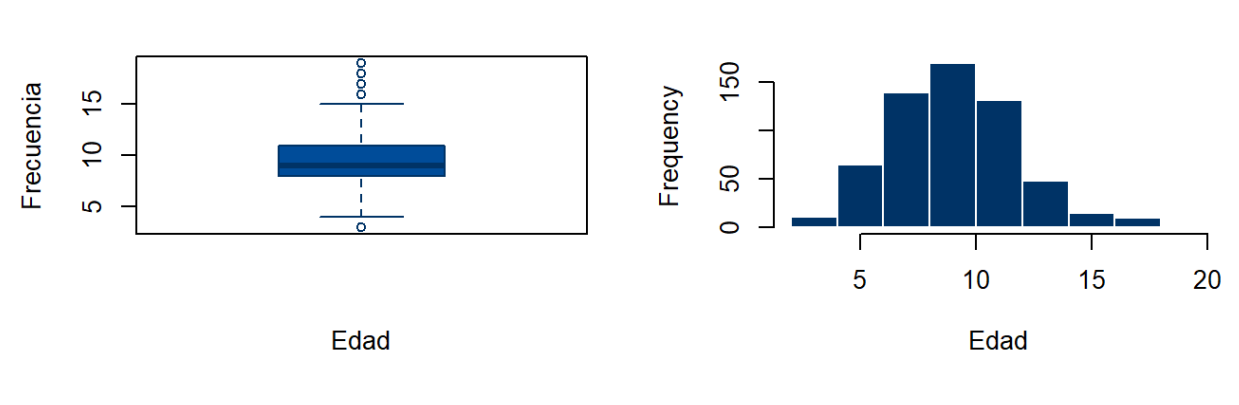
Test de normalidad H0: Los datos son normales vs H1: Los datos no son normales

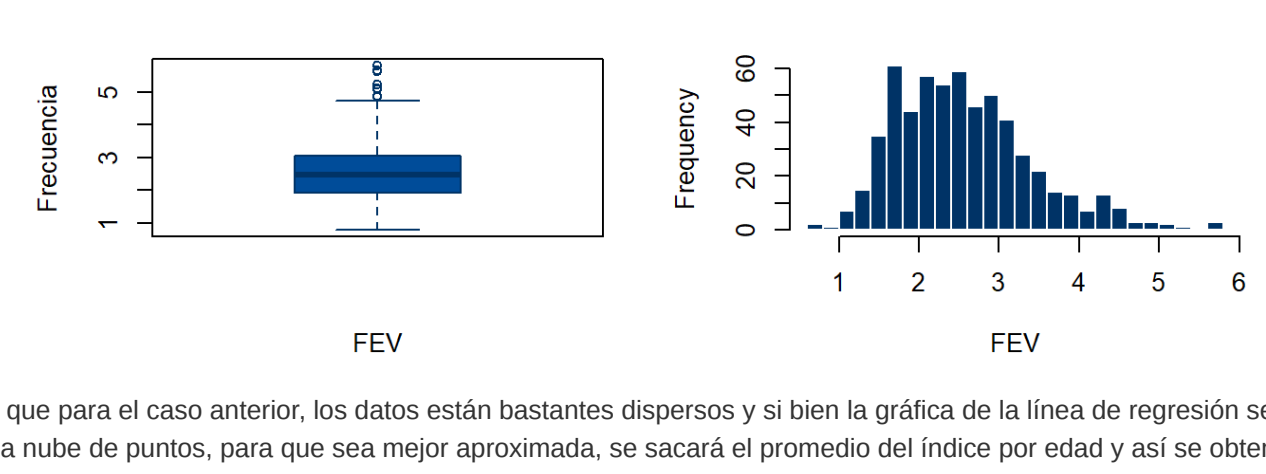
Test.de.normalidad	Valor
Normalidad Shapiro	pvalor ≥ α, no rechazamos H0
Normalidad Anderson	pvalor ≥ α, no rechazamos H0

Test Durbin-Watson

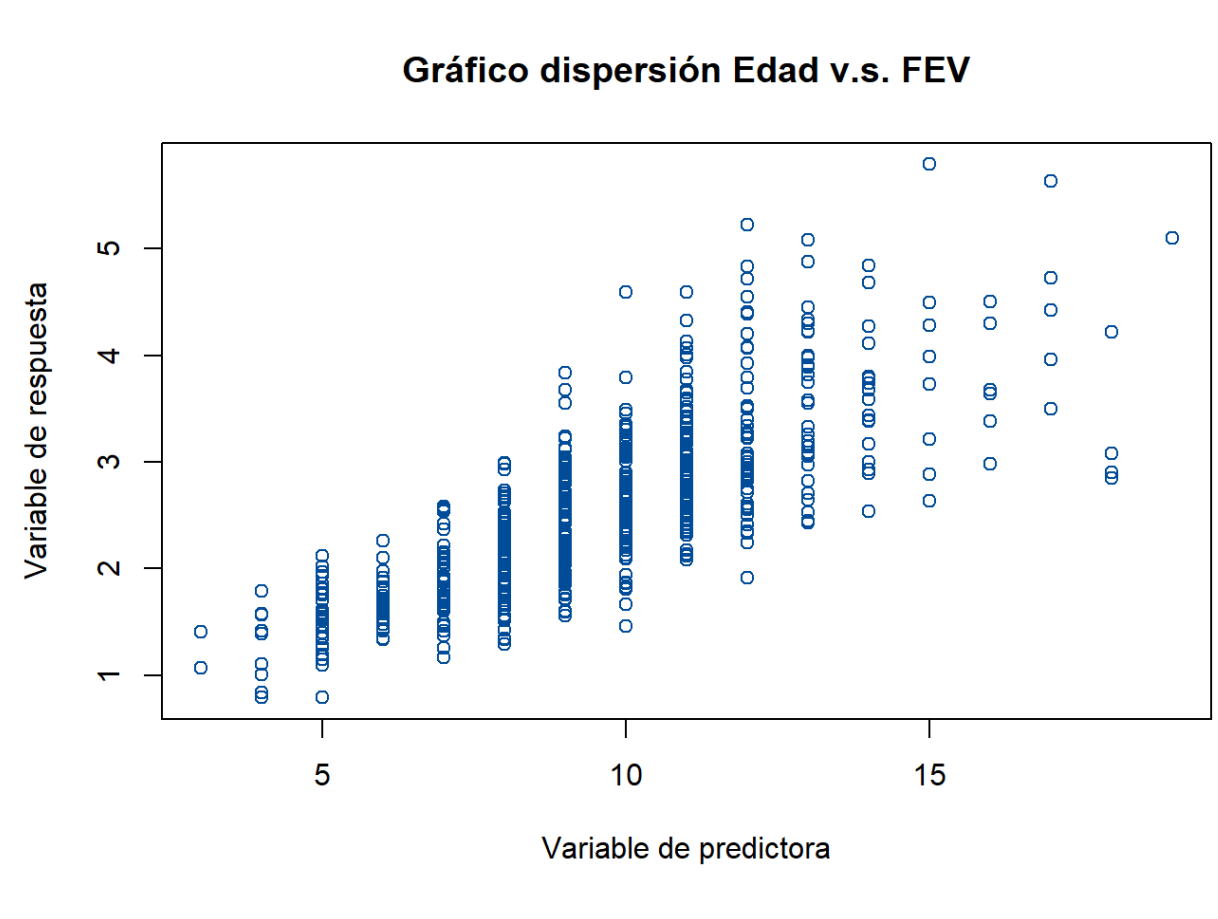
Test	Conclusión
dw	No hay una correlación en los residuos

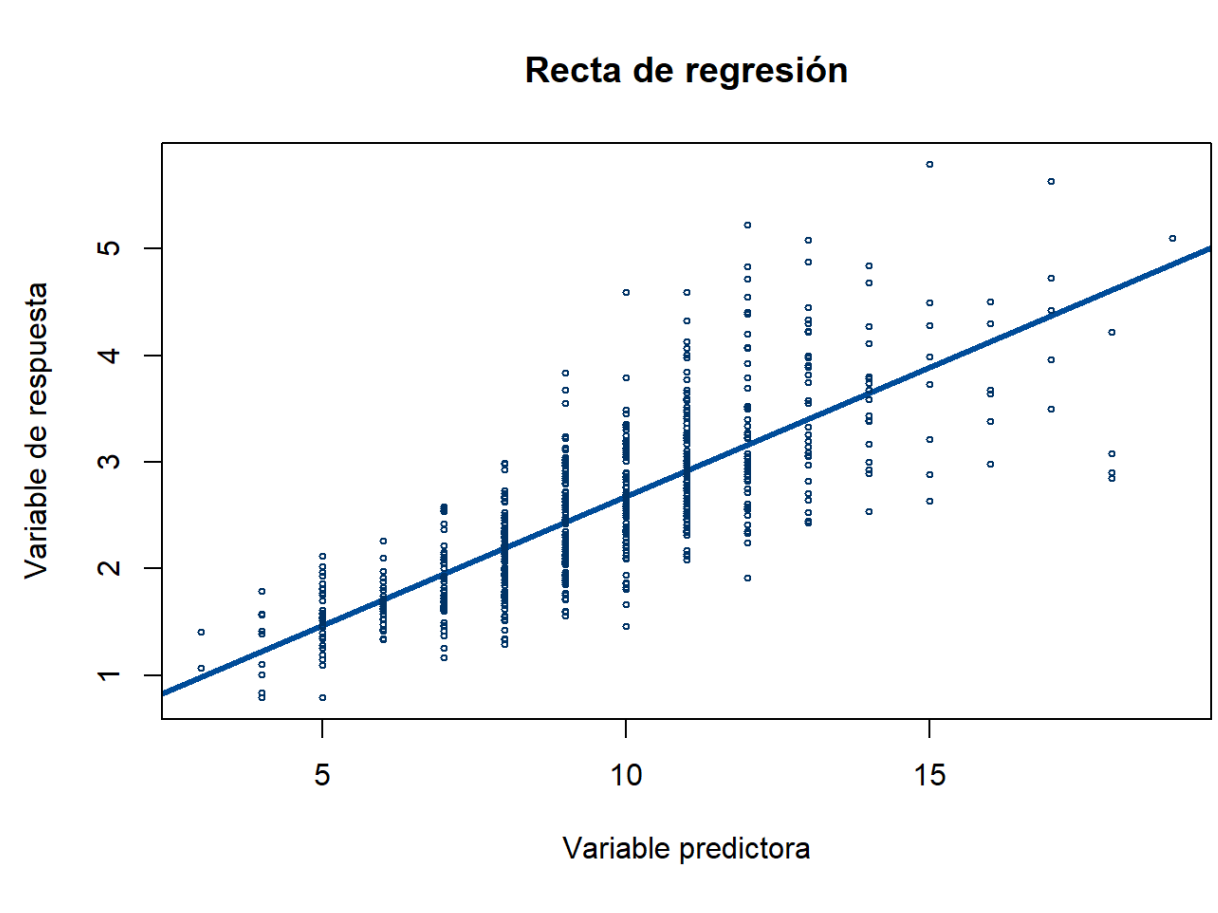
Para el caso de los no fumadores, tendremos muchos más outliers para casos de la edad como del FEV, esto podría ser esperado al saber que la mayoría de los datos se concentra en este grupo de no fumadores. Lo que sí podemos observar es que en el histograma ambos datos parecen tener un sesgo a la derecha.



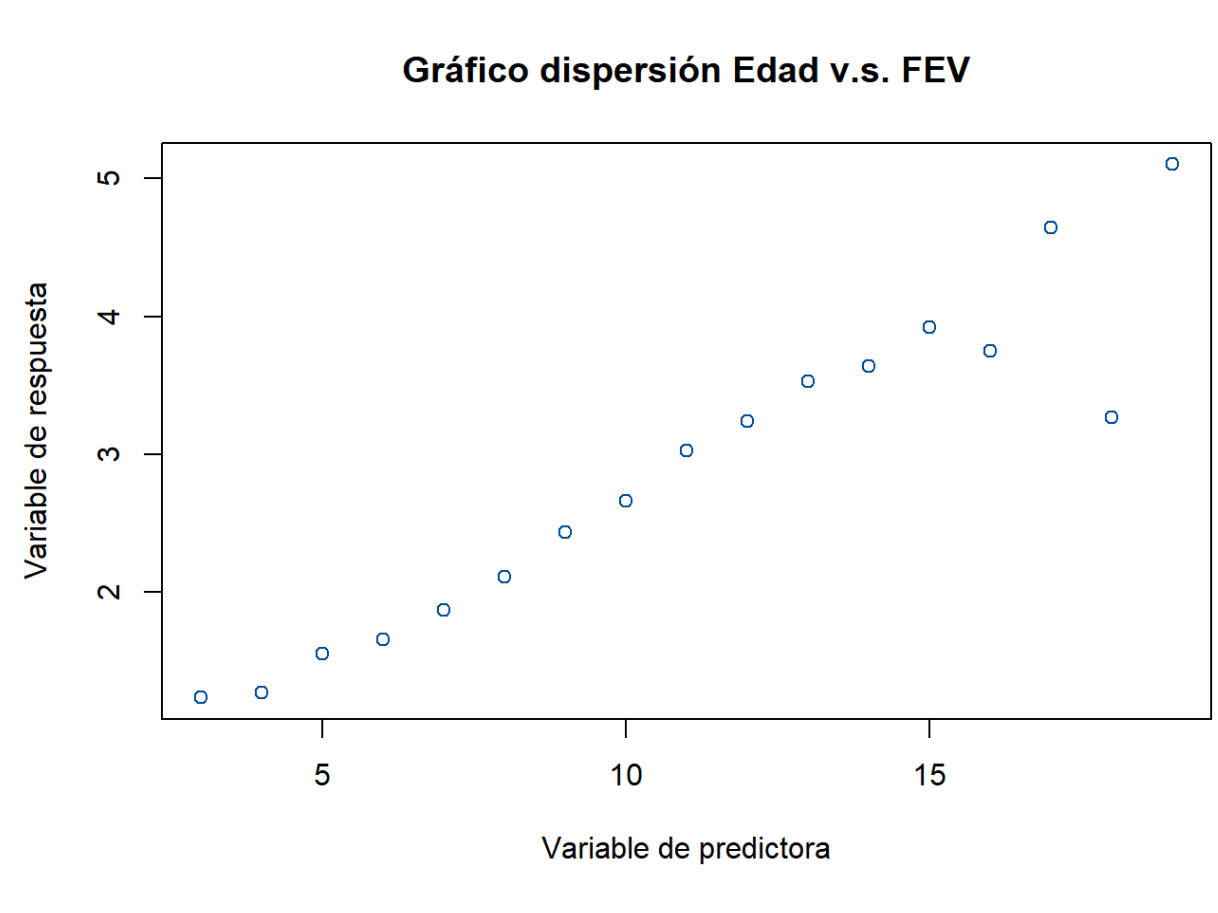


De igual manera que para el caso anterior, los datos están bastantes dispersos y si bien la gráfica de la línea de regresión se encuentra en un punto medio de la nube de puntos, para que sea mejor aproximada, se sacará el promedio del índice por edad y así se obtendrá un número de puntos según el número de edades.

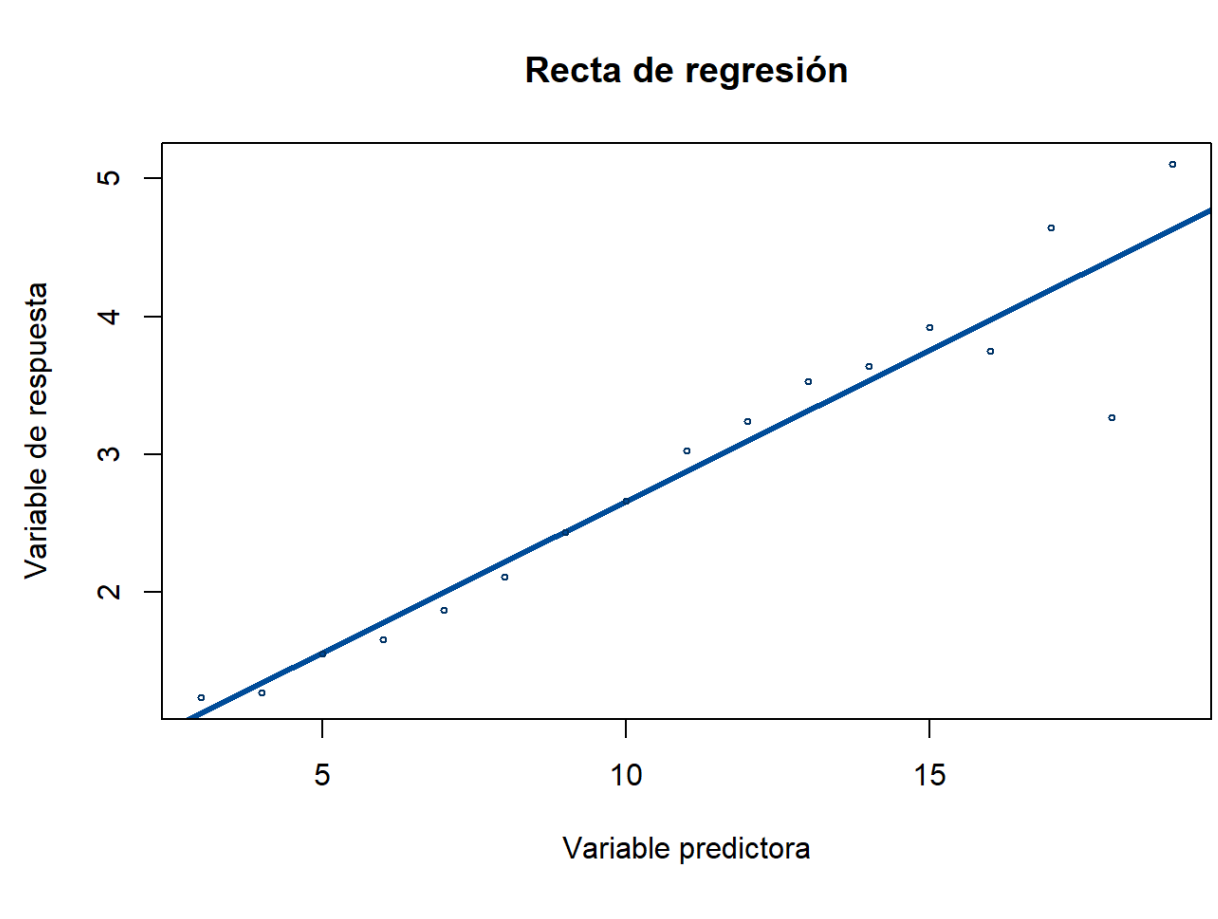


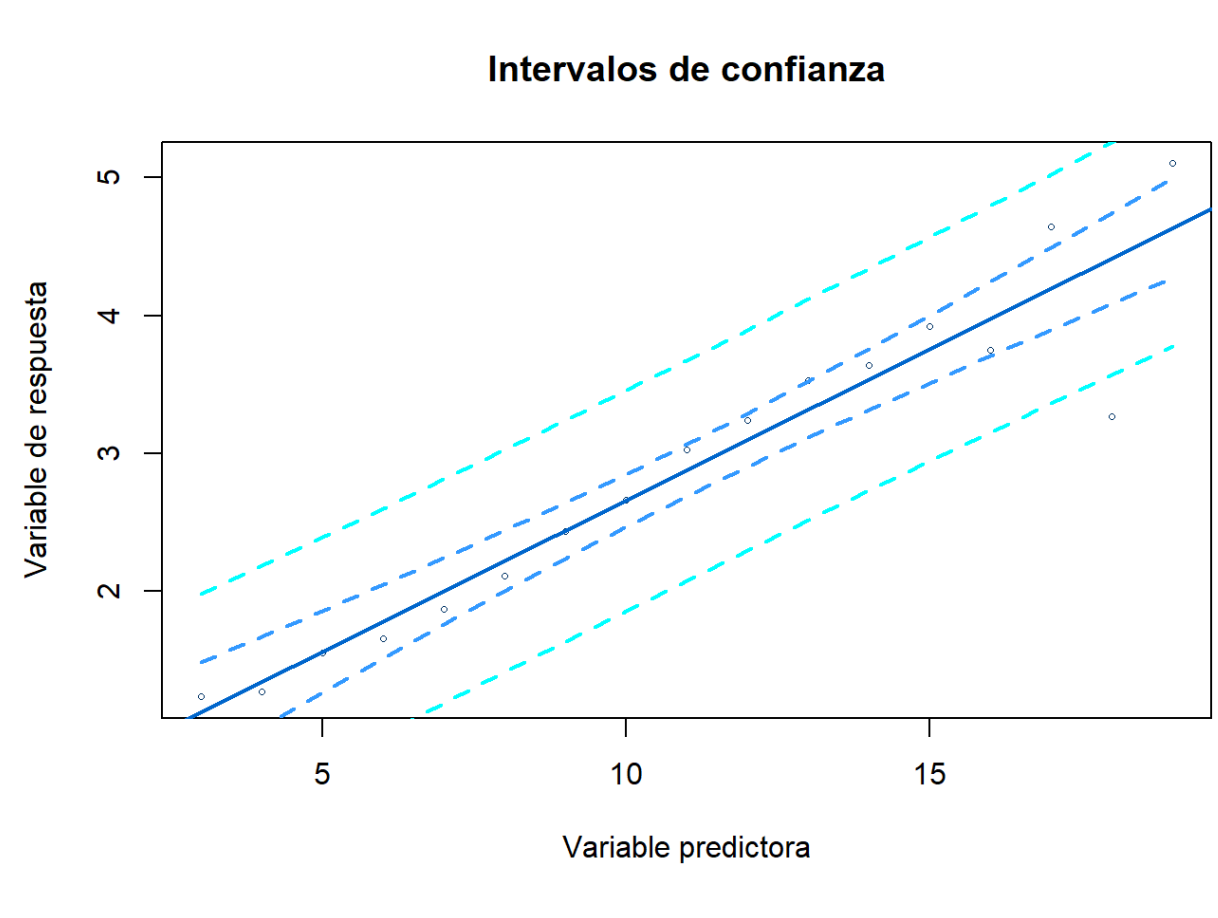


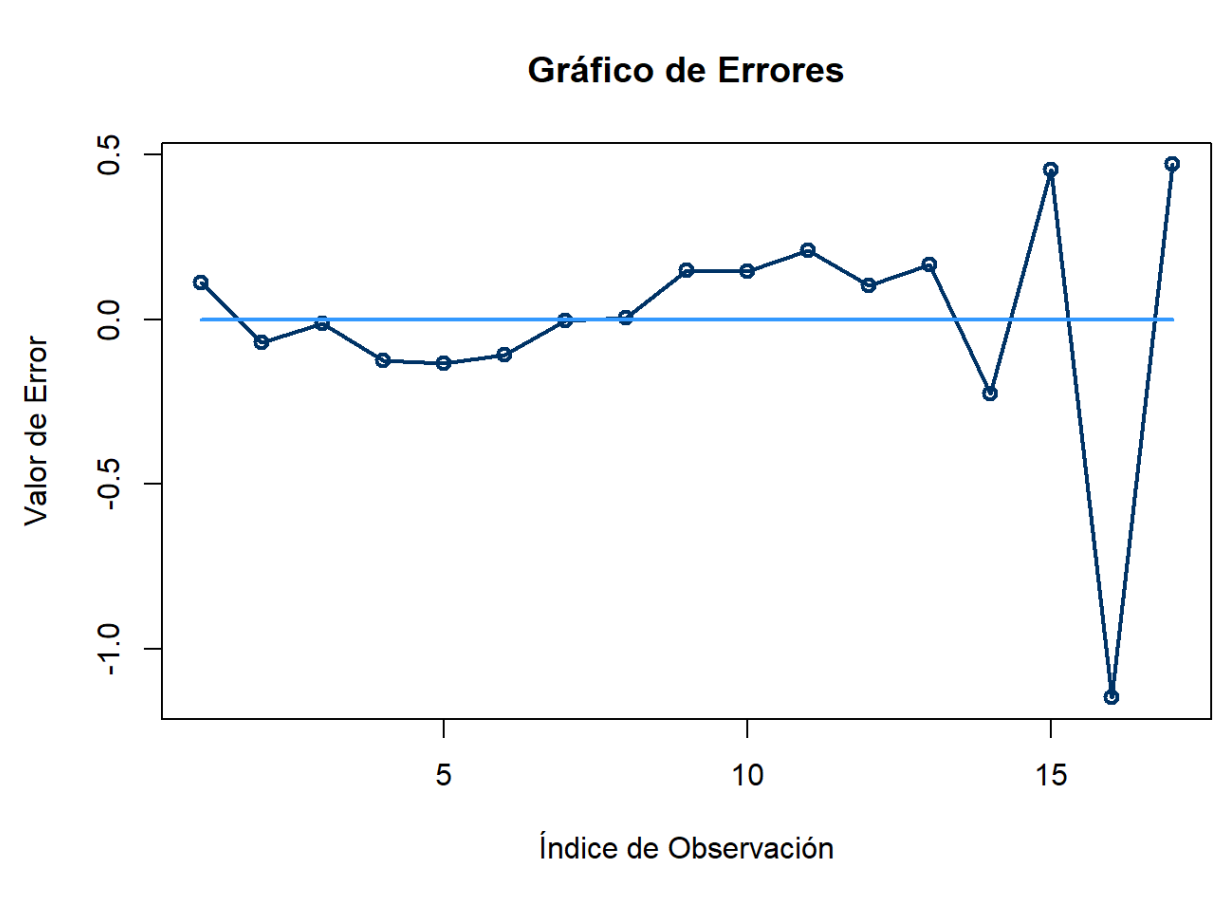
Aquí los puntos se encuentran mucho mejor posicionados para una mejor interpretación gráfica donde sí vemos una relación positiva entre las variables, incluso de mejor manera que el caso anterior.

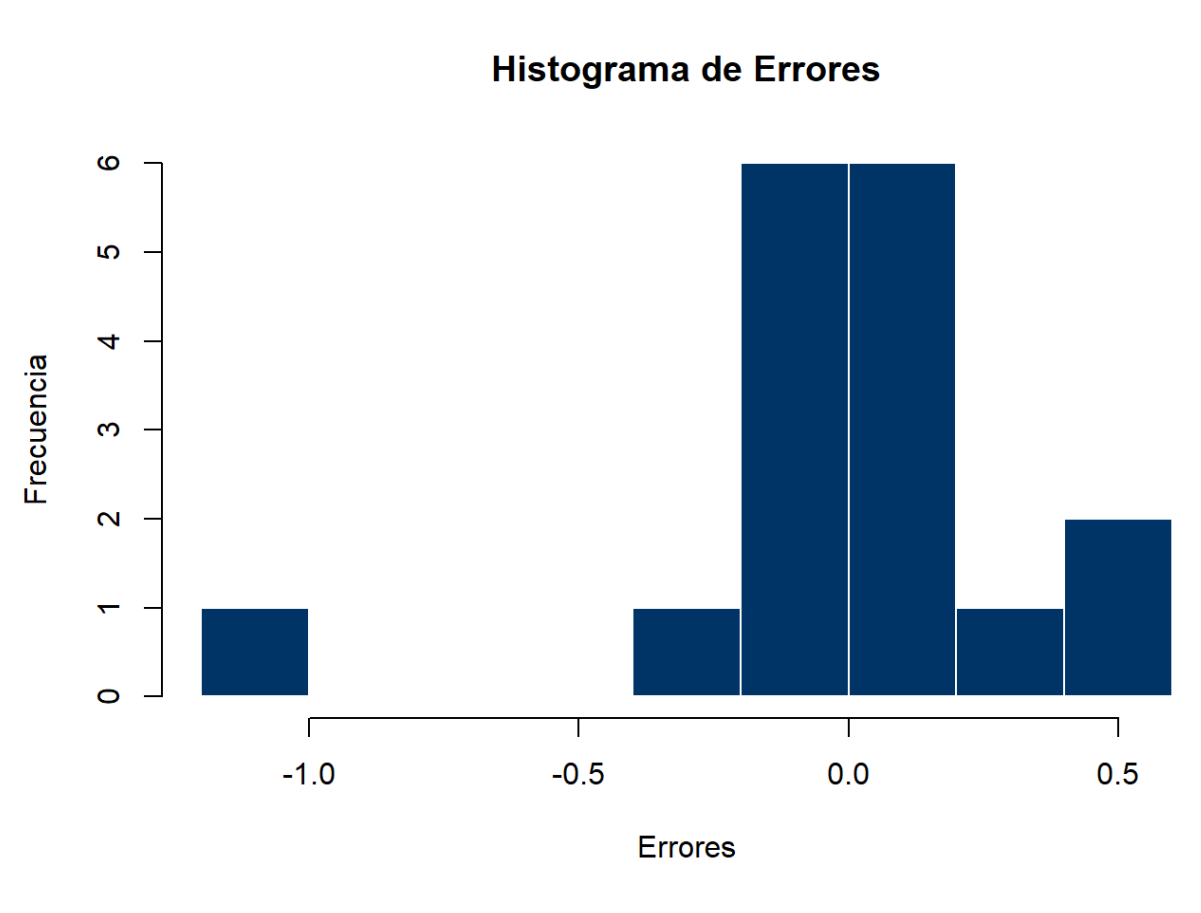


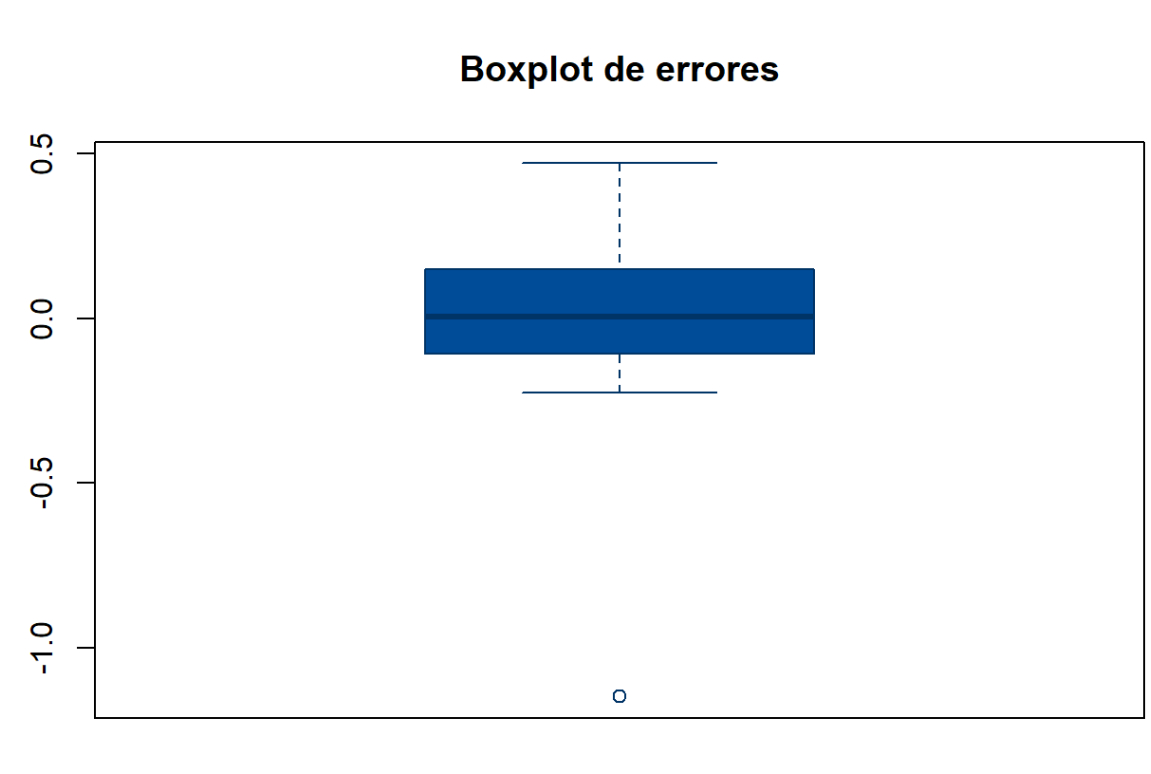
Al tener un entorno mucho más limpio, el análisis es más sencillo y podemos observar que estos datos se comportan mucho más como la línea de regresión a excepción de dos puntos, lo que podría indicar que los errores de este modelo podrían ser menores que el anterior y sugiriendo que el fumar pueda alterar los resultados normales a los que se debería de comportar el índice FEV según la edad.

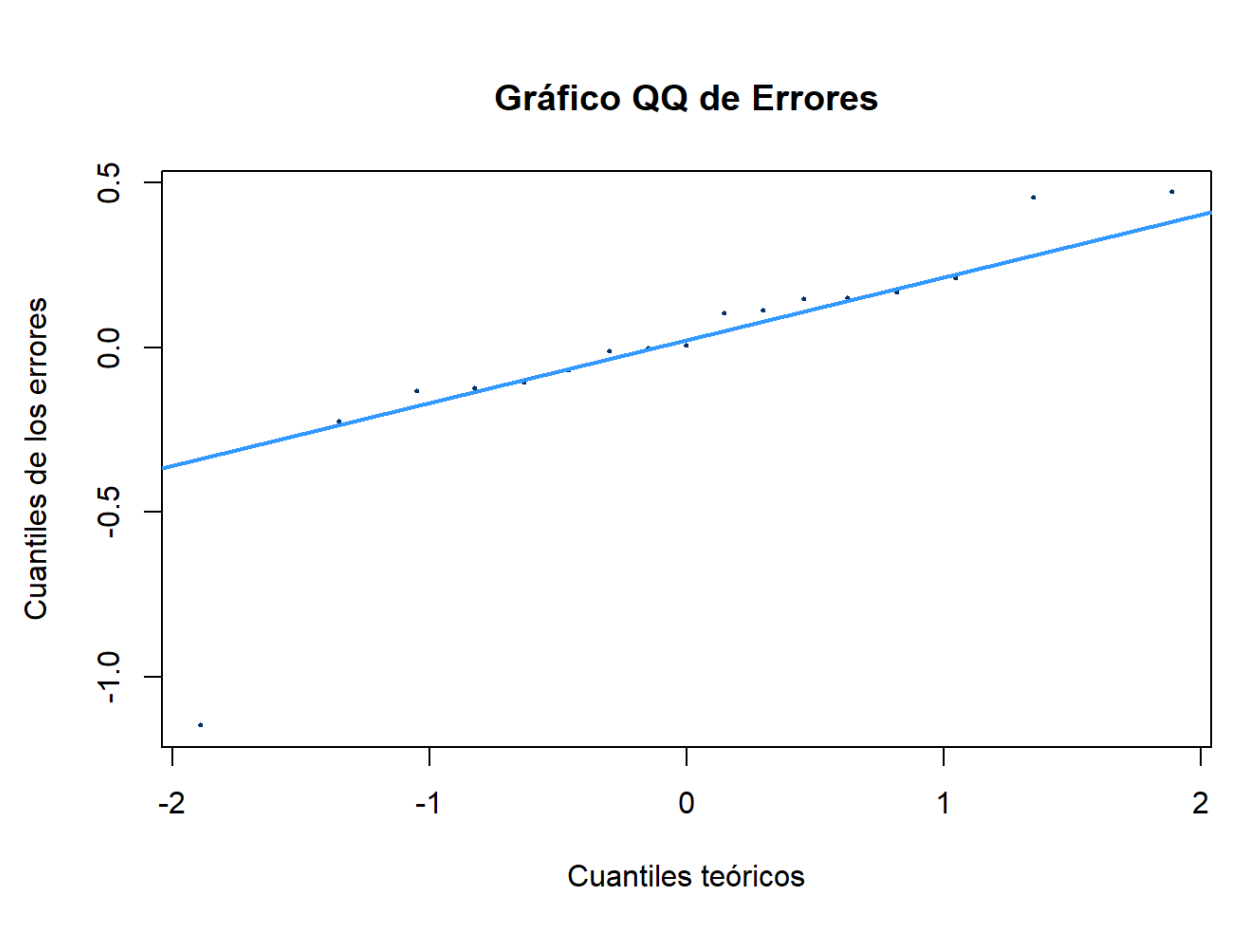












Informacion del modelo

Datos	Valores
α	0.05
β0	0.468252357996954
β1	0.219122650140799
S	0.363613080465725
S^2	0.132214472285774
CV	12.6315879172559
df	15
qti	-2.13144954555978
qtd	2.13144954555978
tc	12.1724459657473
P valor	0.00000000355599872489347
Mín. β1	0.180753297392903
Máx. β1	0.257492002888695
Prom. err.	-0.0000000000000000289848695494323
Var. error	0.123951067767913
Shap. P valor	0.00124135991555556
Anderson P valor	0.0035356766036633
Rho Pearson	0.952927310022899
Rho Spearman	0.96078431372549
dw	2.97416176068524

Inferencia para β1 H0: β1 = 0 vs H1: β1 ≠ 0

Prueba.de.hipótesis	Conclusión
tc	Rechazamos H0. β1 ≠ 0
p valor	Rechazamos H0. β1 ≠ 0

Coeficiente de correlación

Rho	Conclusión
Pearson	Hay una correlación fuerte
Spearman	Hay una correlación fuerte

Test de normalidad H0: Los datos son normales vs H1: Los datos no son normales

Test.de.normalidad	Valor
Normalidad Shapiro	pvalor < α, rechazamos H0. No son normales
Normalidad Anderson	pvalor < α, rechazamos H0. No son normales

Test Durbin-Watson

Test	Conclusión
dw	No hay una correlación en los residuos