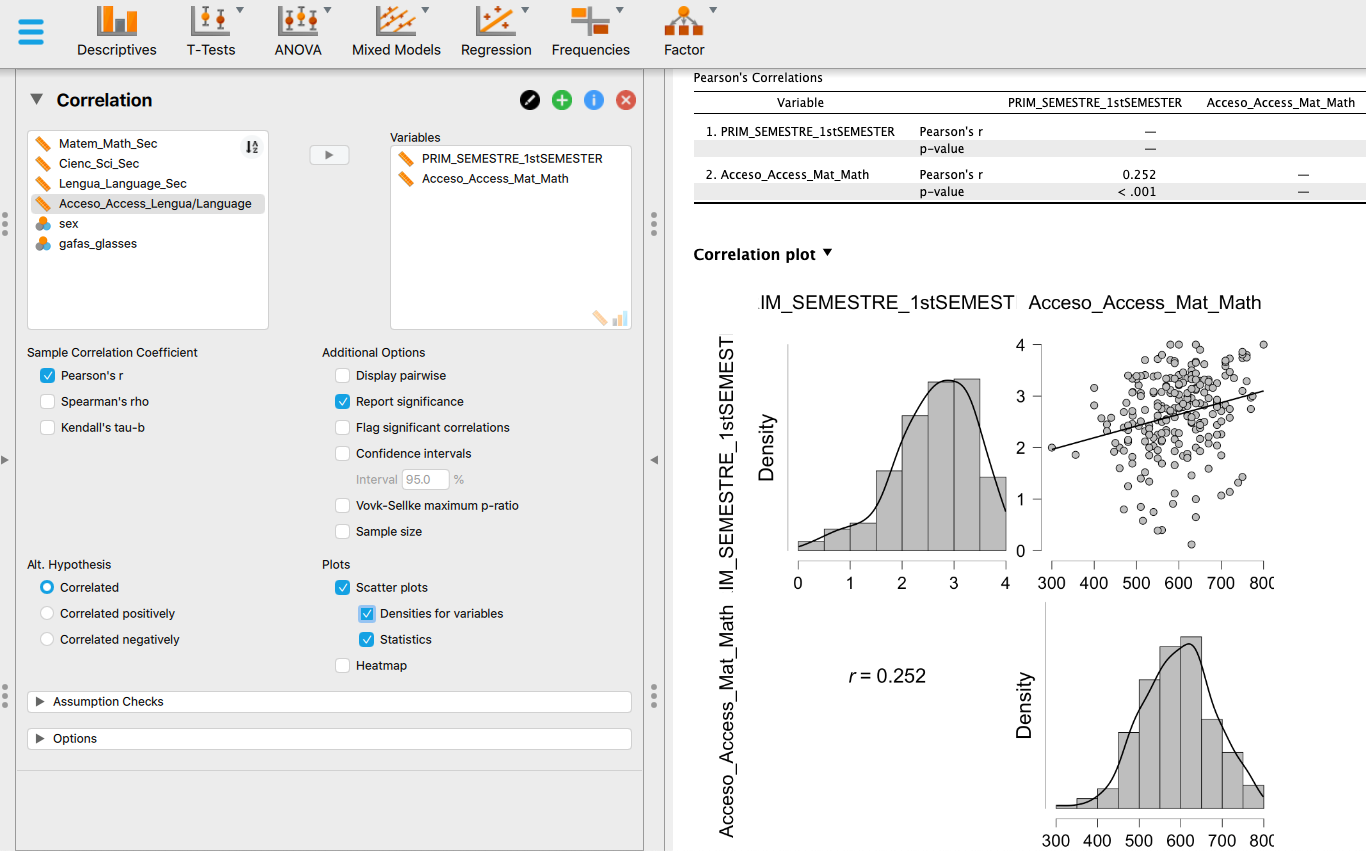
**Ejercicio 5. Estadística. Correlación y regresión**

Tenemos un conjunto de datos (<https://www.uv.es/mperea/PRIM_SE.jasp>), que contiene información de las notas en el primer cuatrimestre de un grupo de estudiantes en una carrera de ciencias, así como información de las notas promedio en instituto de matemáticas, ciencias y lengua. También tenemos las notas de selectivo (PAU) de matemáticas y lengua, así como el sexo y si llevan gafas.

La idea (global) es ver cuán predictivas son las notas del instituto cuando el alumnado llega a la universidad en su primer cuatrimestre y ha de examinarse.

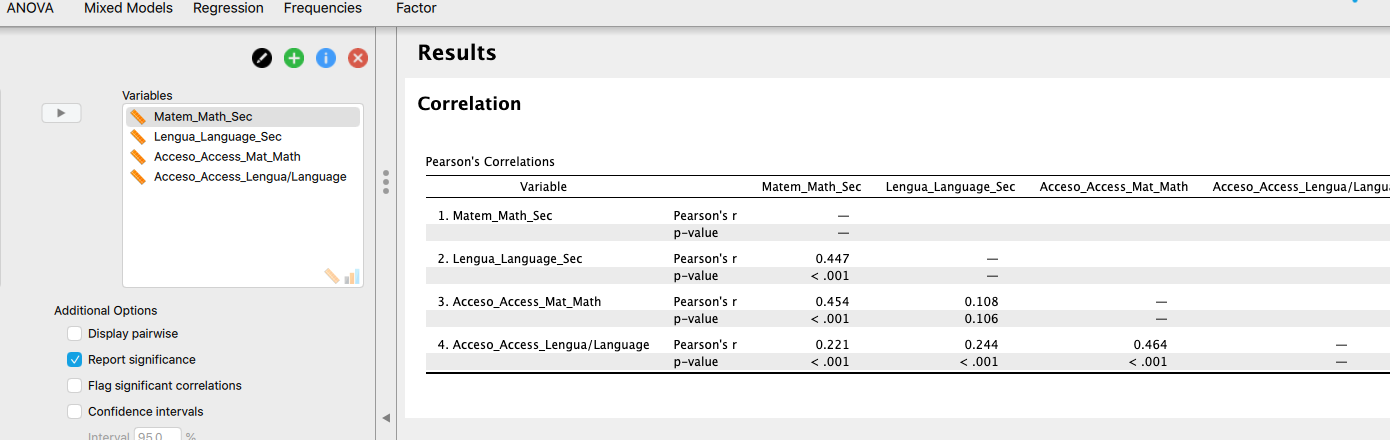
Correlación

*--¿Qué relación aprecias (signo y fuerza) entre las notas de primer cuatrimestre (Y) y las notas en selectivo de matemáticas (X)? Efectúa el gráfico adecuado y calcula el índice adecuado.*



Hay una relación positiva pero relativamente débil entre ambas variables, como atestigua el diagrama de dispersión y el valor de coeficiente de correlación de Pearson (r = 0.252).

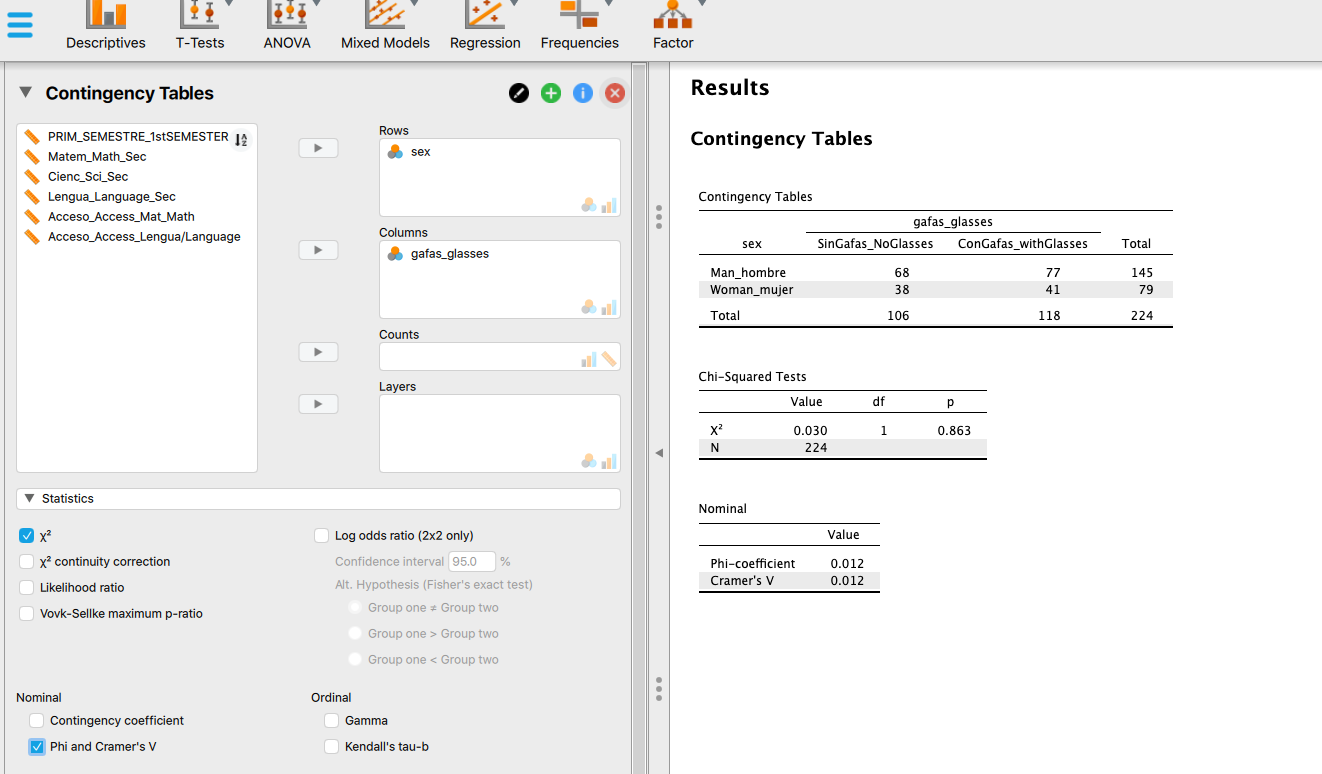
*--Pensemos que tenemos una matriz de correlaciones con las siguientes variables: notas promedio de instituto en matemática, notas promedio de instituto en lengua, notas de la PAU en matemáticas y notas de la PAU en lengua. ¿En qué caso la relación es más fuerte? ¿Y más débil? ¿Pensáis que hubiera sido posible encontrar relaciones negativas entre estas variables?*



La relación más fuerte es en el caso del mayor coeficiente de correlación de Pearson en valor absoluto: las puntuaciones de matemáticas y de lengua de la PAU (r= 0.464). La más débil es entre las notas de Lengua-Secundaria y PAU-Matemáticas.

Todas las relaciones han sido positivas (en diferente grado) dado que se están midiendo aspectos similares: los buenos estudiantes suelen serlo a través de las asignaturas.

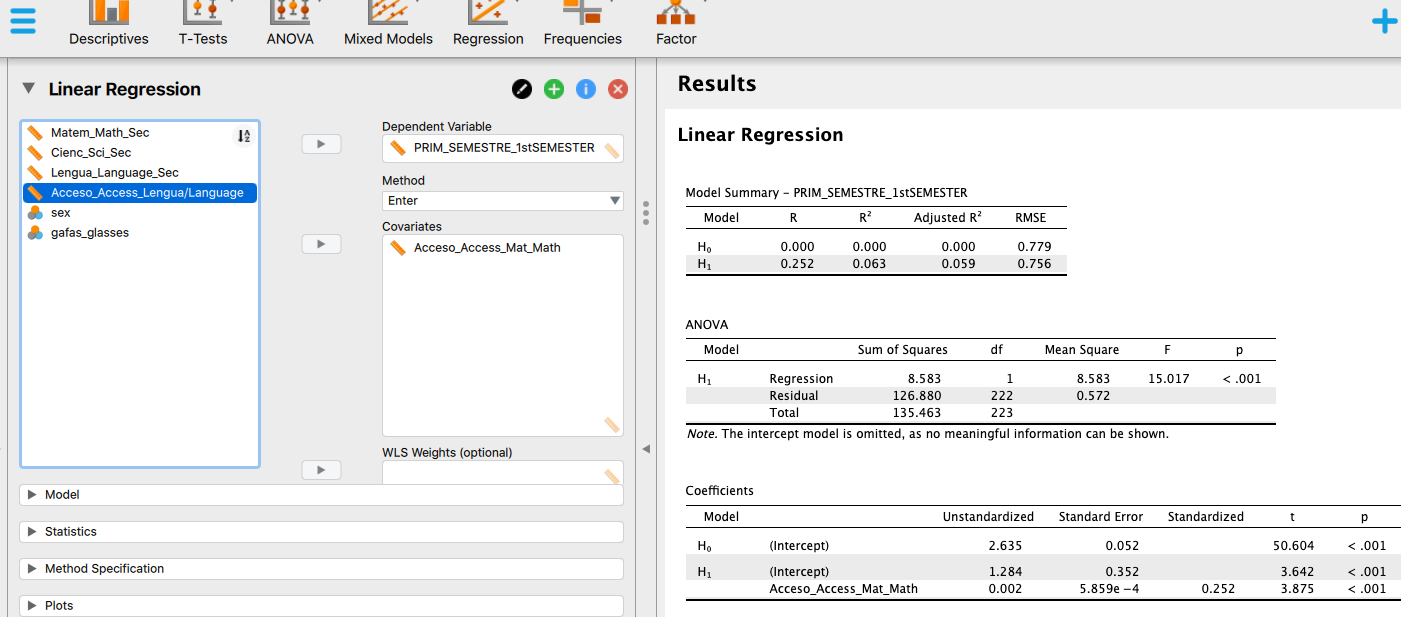
*--Queremos saber si la proporción de personas con gafas es mayor en chicas que en chicos (o en otras palabras, si hay relación entre sexo y el llevar o no gafas). Efectúa la tabla de contingencia y calcula el índice apropiado.*



Se trata de obtener la tabla de contingencia (menú de “Frequencies”) y el índice V de Cramer. No se observa relación alguna (V=0.012): la proporción de hombres con gafas es similar a la de mujeres con gafas en la muestra.

Regresión simple

*--¿Cuál es la ecuación de la recta de regresión de las notas de primer cuatrimestre (Y) y a partir de las notas de la PAU en matemáticas (X) (en puntuaciones directas)? ¿Qué quiere decir la ordenada en el origen y la pendiente en este caso?*



NOTA\_UNI ‘ = 1.284 + 0.002 PAU\_MAT

1.284 (A, ordenada en el origen) es lo que se predice a alguien con un CERO en PAU\_MAT

0.002 (B, pendiente) indica cuanto sube en la variable predicha (NOTA\_UNI) por cada unidad de la variable X (la predictora).

*--Idem en puntuaciones típicas. ¿El valor de la pendiente coincide necesariamente con…?*

Es la columna “standarized”. En regresión simple, la pendiente en puntuaciones típicas coincide necesariamente con el coeficiente de correlación de Pearson.

*--¿Qué porcentaje de varianza de las notas de primer cuatrimestre (Y) explica la recta de regresión?*

Es R^2. Si miramos el R^2 ajustado, el valor es 0.059. O sea, la recta de regresión puede explicar el 5.9% de la varianza de notas de primer cuatrimestre.

*--¿Qué puntuación en notas de primer cuatrimestre predirías a la persona de la primera fila del archivo con la recta de regresión?*

Es sustituir en la fórmula…

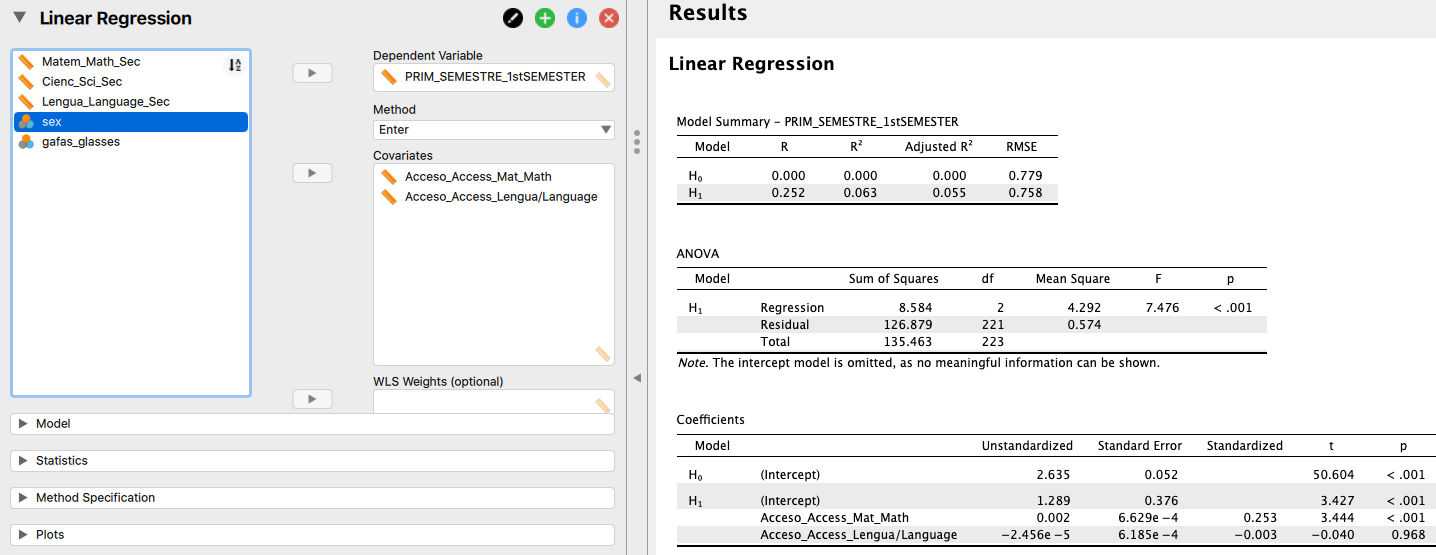
NOTA\_UNI ‘ = 1.284 + 0.002 PAU\_MAT

Y donde dice PAU\_MAT, indicáis el valor de la persona primera en esa variable. Hacéis el cálculo y ya.

Podríais también crear la columna entera de “puntuaciones predichas” con la fórmula anterior. Y así tenéis no solamente la primera persona, sino también el resto.

Regresión múltiple

*--¿Cuál es la ecuación de regresión de las notas de primer cuatrimestre (Y) y a partir de los predictores nota de la PAU en matemáticas y nota de la PAU en lengua? ¿Qué quieren decir los componentes de la ecuación de regresión en puntuaciones directas?*



NOTA\_UNI ‘ = 1.289 + 0.002 PAU\_MAT + 0.002 PAU\_LENGUA

Tenemos el intercepto, como antes, y ahora tenemos coeficientes para cada predictor.

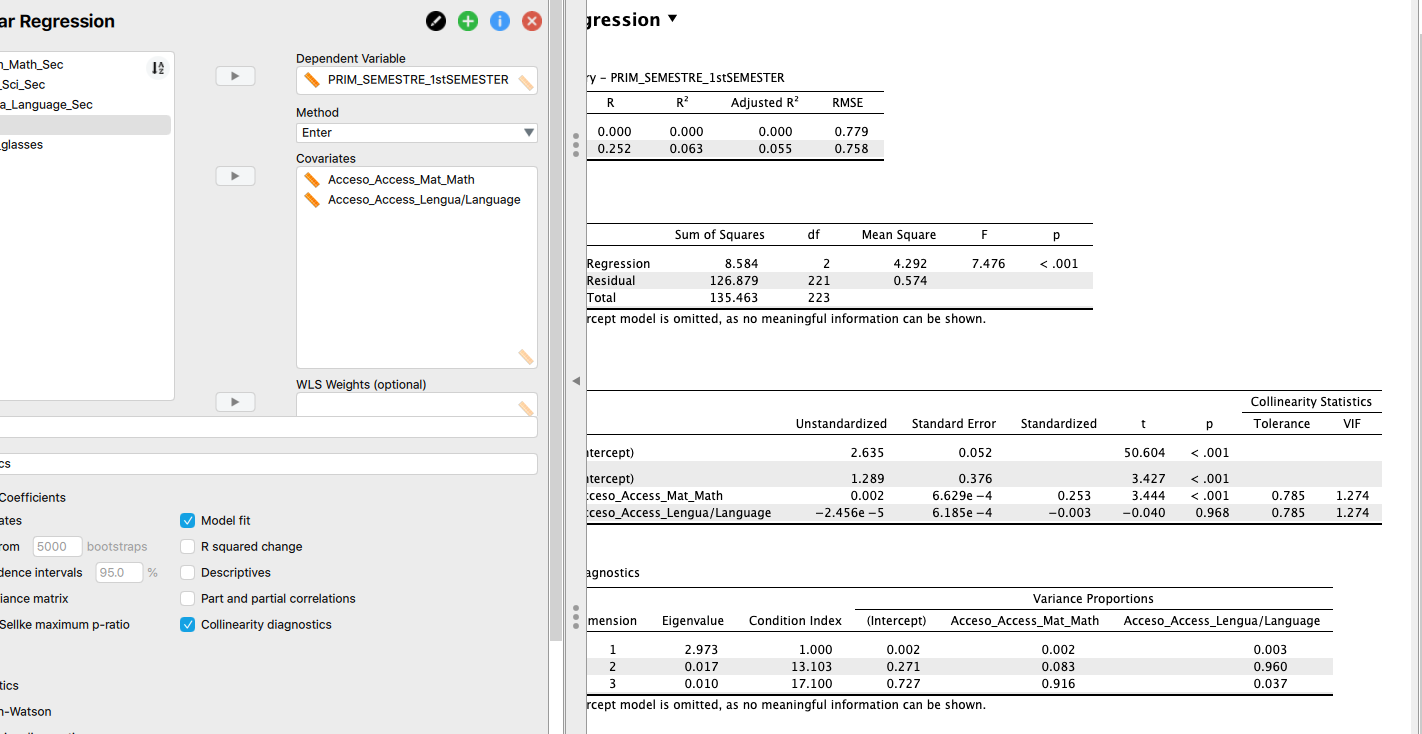
*--¿Qué porcentaje de varianza de las notas de primer cuatrimestre explica la ecuación de regresión de la pregunta anterior?*

Es R^2. Si miramos el R^2 ajustado es 0.055. O sea, la ecuación de regresión puede explicar el 5.5% de la varianza de notas de primer cuatrimestre.

*--¿Cuál es el mejor predictor de la ecuación anterior y por qué?*

El mejor predictor en la ecuación es aquel con un valor estandarizado mayor (en términos absolutos, sin mirar el signo). En este caso PAU\_MAT

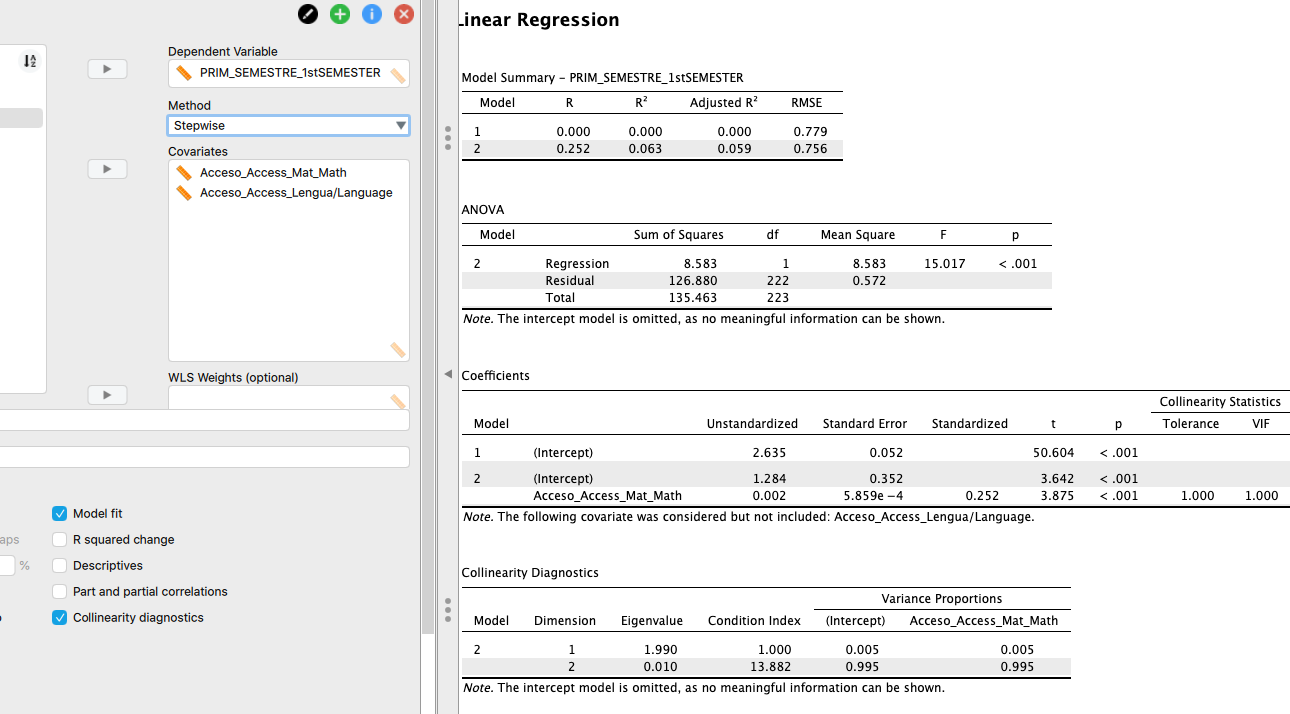
*--¿Crees que ha habido problemas de colinealidad entre los predictores? (recordad: mirar el VIF [Variance Inflation Factor], en el caso ideal sería 1.0)*



VIF ha sido 1.274. Lo idea es 1.0. Dado que el límite se indicó en 10…se puede concluir:

No ha habido graves problemas de colinealidad (VIF = 1.27).

*--Pensemos que hiciéramos la regresión “stepwise” (paso a paso). ¿Qué ecuación resultante tendríamos? ¿Por qué?*



El predictor PAU\_LENGUA no ha entrado en la ecuación. De hecho, en la pregunta anterior se apreciaba que apenas aportaba nada a la ecuación (su coeficiente estandarizado era -0.003, o sea, que es virtualmente nulo).

Por tanto, lo más parsimonioso es excluir a PAU\_Lengua de la ecuación de regresión.