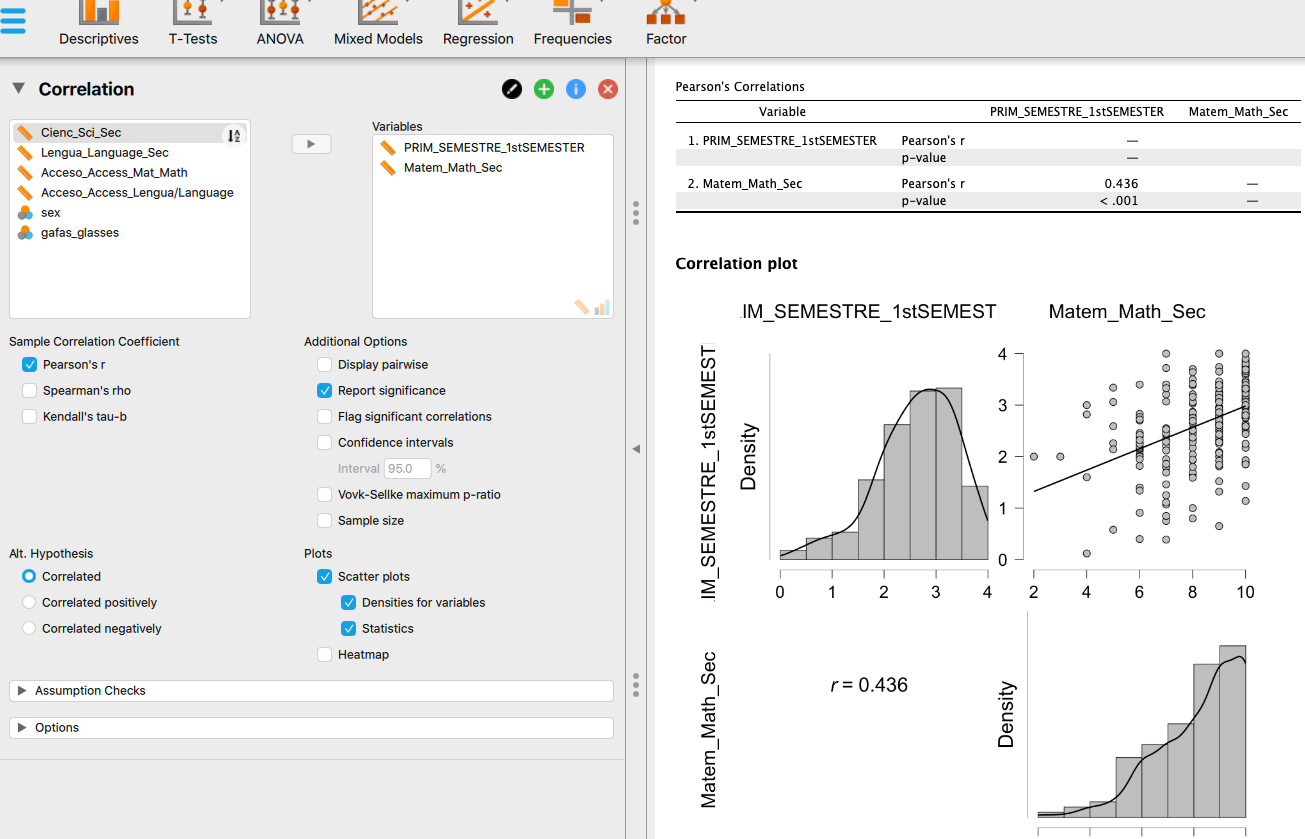
**Ejercicio de correlación y regresión, 23 noviembre 2020**

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 de matemáticas y lengua, así como el sexo.

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.

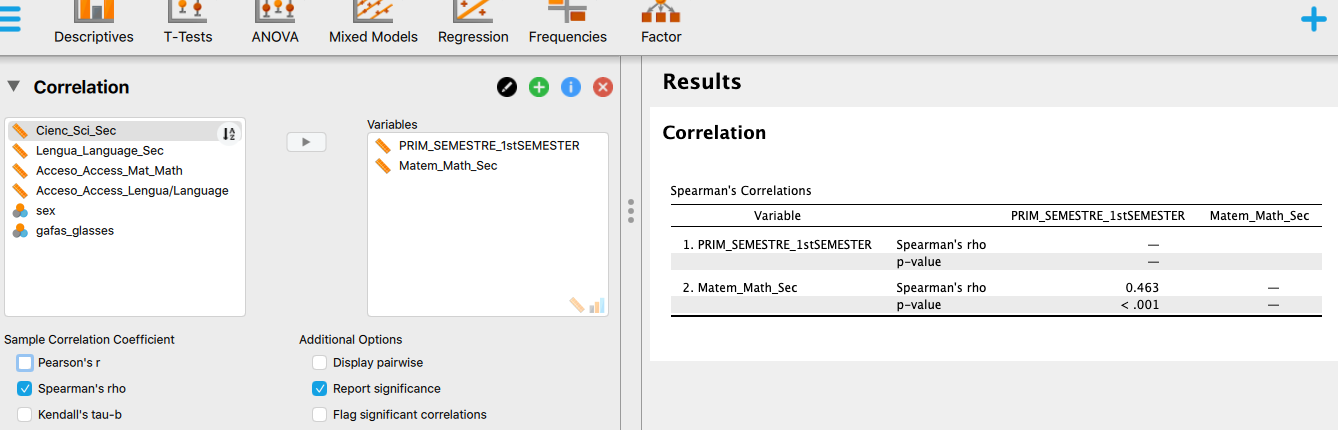
Correlación

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



El diagrama de dispersión muestra una relación positiva moderada. El coeficiente de correlación de Pearson es 0.436.

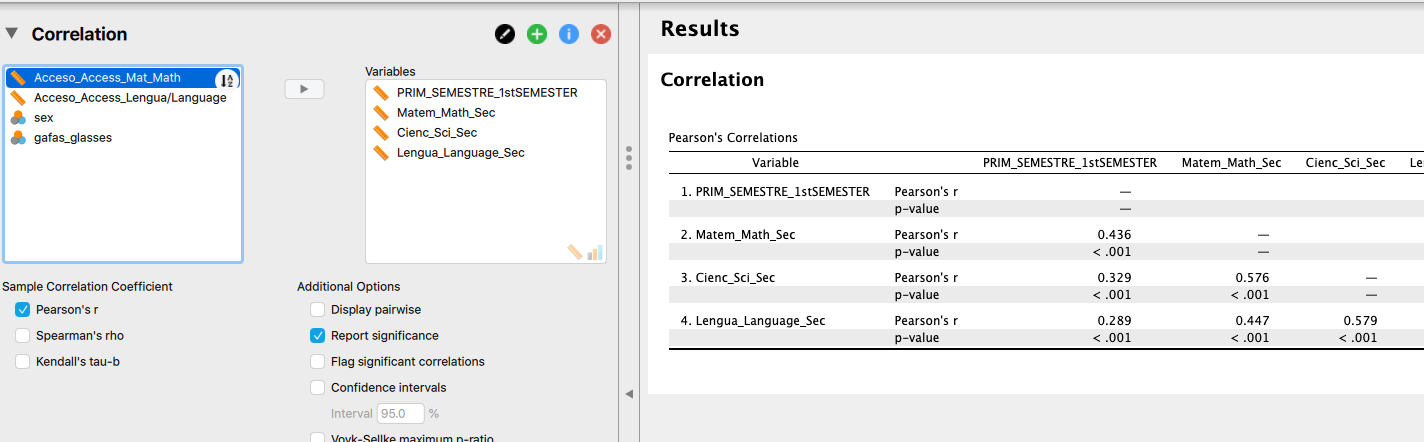
*--Si bien no es el caso, pero si hubiéramos considerado como “ordinales” las variables anteriores, ¿qué índice calcularías y cuál es su valor? (Una vez hecho el ejercicio, recordad indicar esas variables como cuantitativas.)*



En este caso, habríamos de calcular el coeficiente de correlación de Spearman, que se interpreta de manera análoga al de Pearson; claro está, no es emplea el término “lineal” dado que se asume que la relación sería ordinal.

El coeficiente de Spearman es 0.463.

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

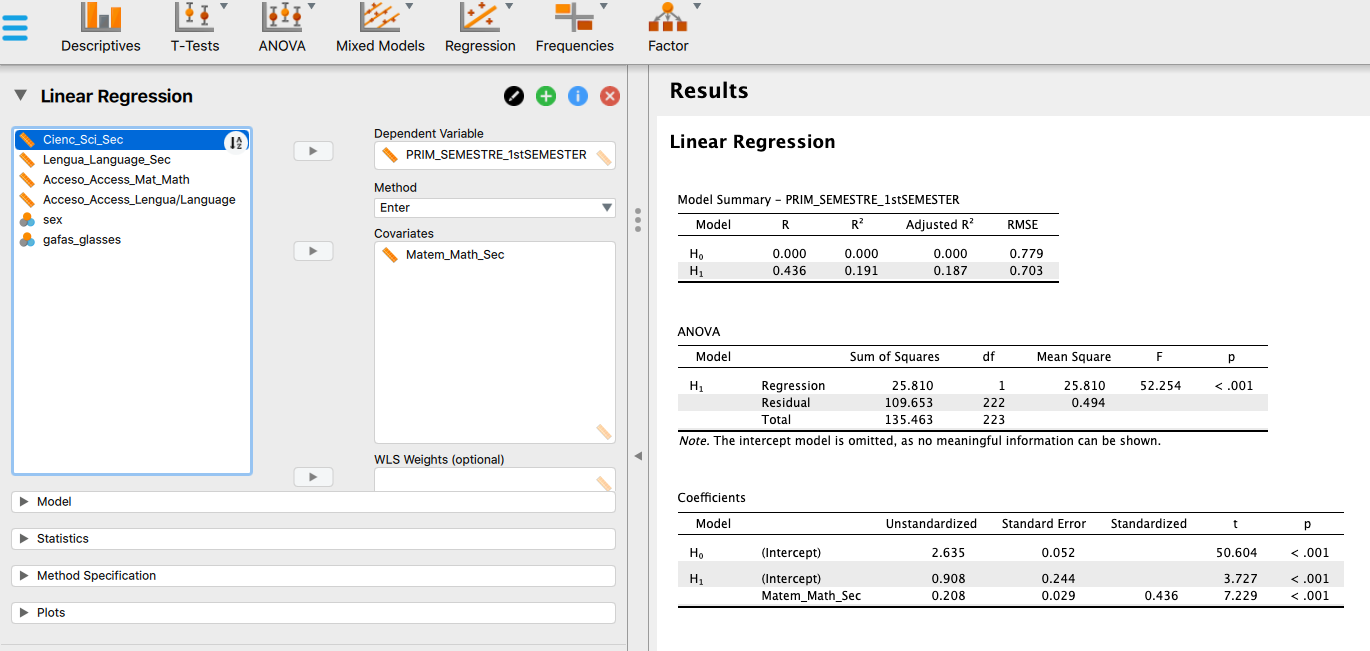


La relación más fuerte es aquella con el mayor coeficiente de correlación de Pearson en valor absoluto (recordar que el signo solamente da la dirección; el signo no informa de la fuerza): Notas de lengua en secundaria y notas de ciencia de secundaria. La más débil es notas de Lengua en Secundaria con las notas de primer semestre en la universidad.

Lo normal es que las personas que sean buenas estudiantes pues lo sean en general en todas las asignaturas, por lo que lo esperable (como así ocurrió) es que las relaciones entre las variables sean positivas. Y así ha sido.

Regresión simple

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



NOTA\_UNIV ‘ = 0.908 + 0.208\*NOTA\_MATEM\_SEC

0.908 (A; ordenada en origen) es el valor predicho en la nota de la universidad si la nota de matemáticas hubiera sido cero

0.208 (B) es la pendiente. Cada incremento de una unidad en las notas de matemáticas en segundaria va acompañadas por un incremento de 0.208 en la nota de la universidad.

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

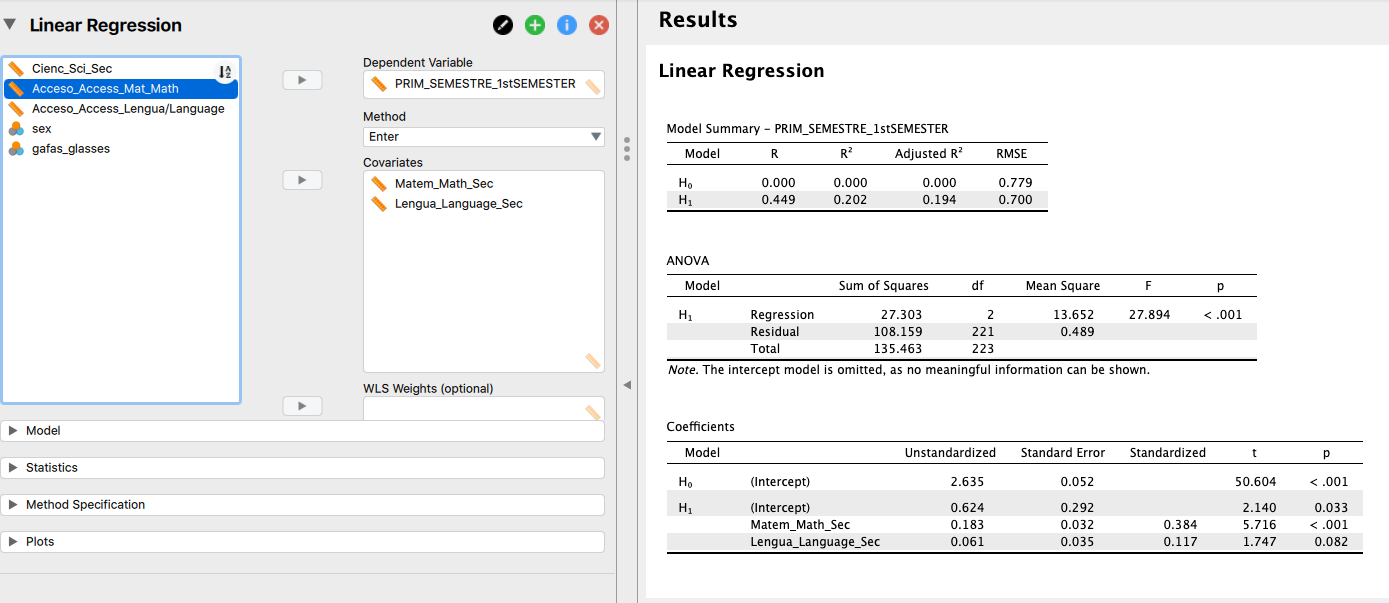
Es el coeficiente de correlación de Pearson: 0.436.

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

Es R^2. 0.187 (si miramos el índice R^2 ajustado). Al indicarlo en porcentaje: la ecuación de regresión explica el 18.7% de las notas del primer cuatrimestre.

Regresión múltiple

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



NOTA\_UNIV ‘ = 0.624 + 0.183\*NOTA\_MATEM\_SEC+ 0.061\*NOTA\_LENGUA\_SEC

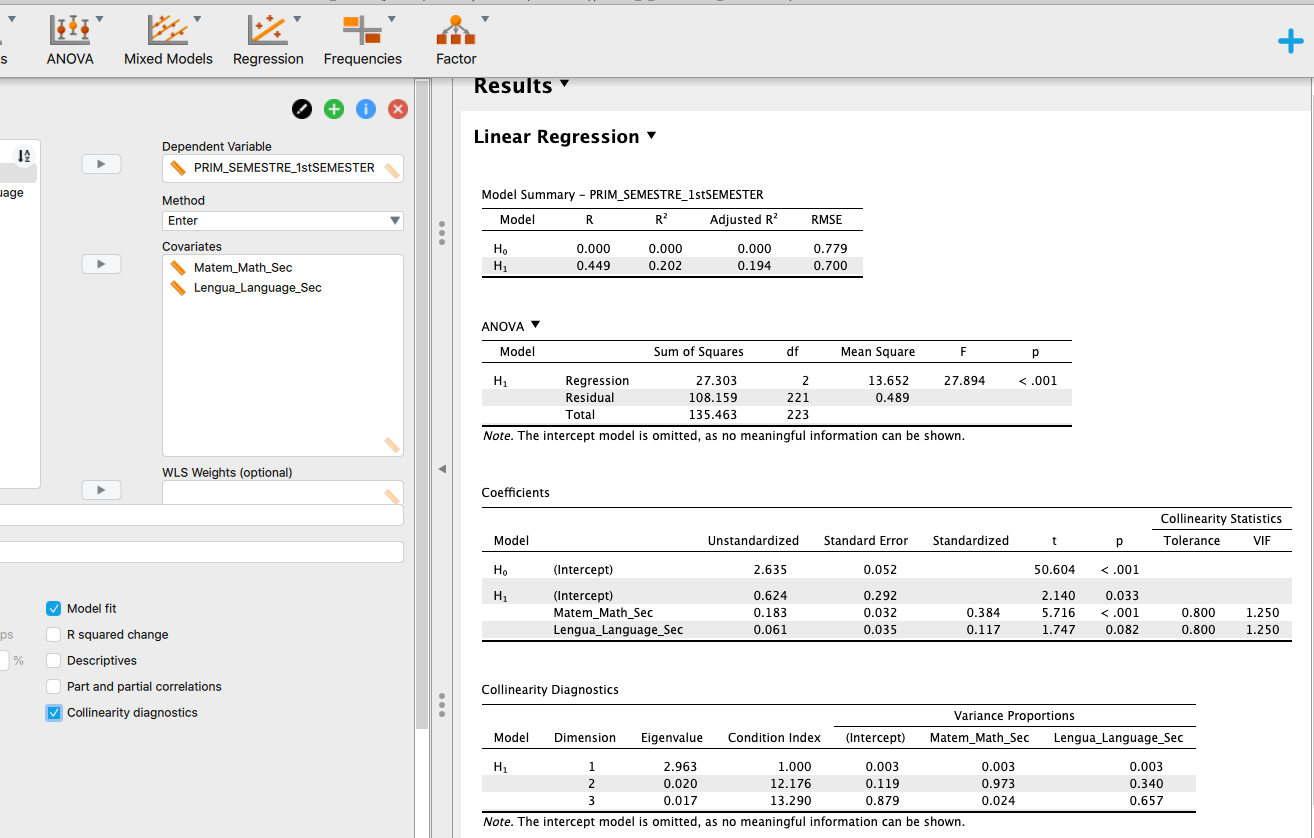
*--¿Qué porcentaje de varianza de las notas de primer cuatrimestre explica la ecuación de regresión de la pregunta anterior?*

Es R^2. 0.194 (si miramos el índice R^2 ajustado). Al indicarlo en porcentaje: la ecuación de regresión explica el 19.4% de las notas del primer cuatrimestre.

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

El mejor predicto con mayor valor (en valor absoluto) en su coeficiente estandarizado (estos coeficientes se pueden interpretar al estilo de los coeficientes de correlación, aunque no son exactamente el coeficiente de Pearson). O sea, las notas de matemáticas.

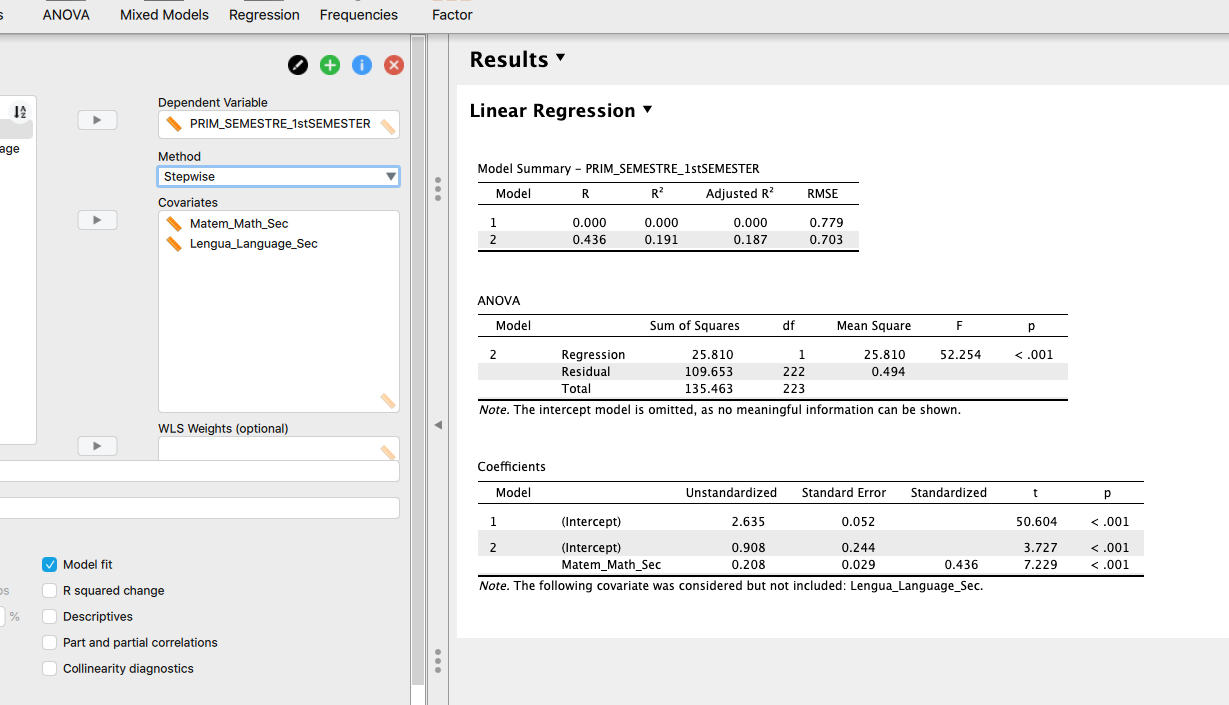
*--¿Crees que ha habido problemas de colinealidad entre los predictores? (recordad: mirar el VIF, en el caso ideal sería 1.0)*



El valor de VIF es 1.25. No es lo más perfecto (que es 1.0) pero está lejos del límite de VIF de 10.

Por tanto, no parece haber problemas graves de colinealidad en los análisis.

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



Solamente un predictor entra en la ecuación, las notas en matemáticas. Observad en las preguntas anteriores que el coeficiente estandarizado de las notas de lengua era muy pequeño, y no ha entrado en la ecuación.

Con este procedimiento “stepwise” solamente entrarán en la ecuación de regresión aquellos predictores que aporten algo sustancial para explicar la variable dependiente (notas de primer cuatrimestre en la universidad), como es el caso de las notas de matemáticas. Pero no las notas en lengua.