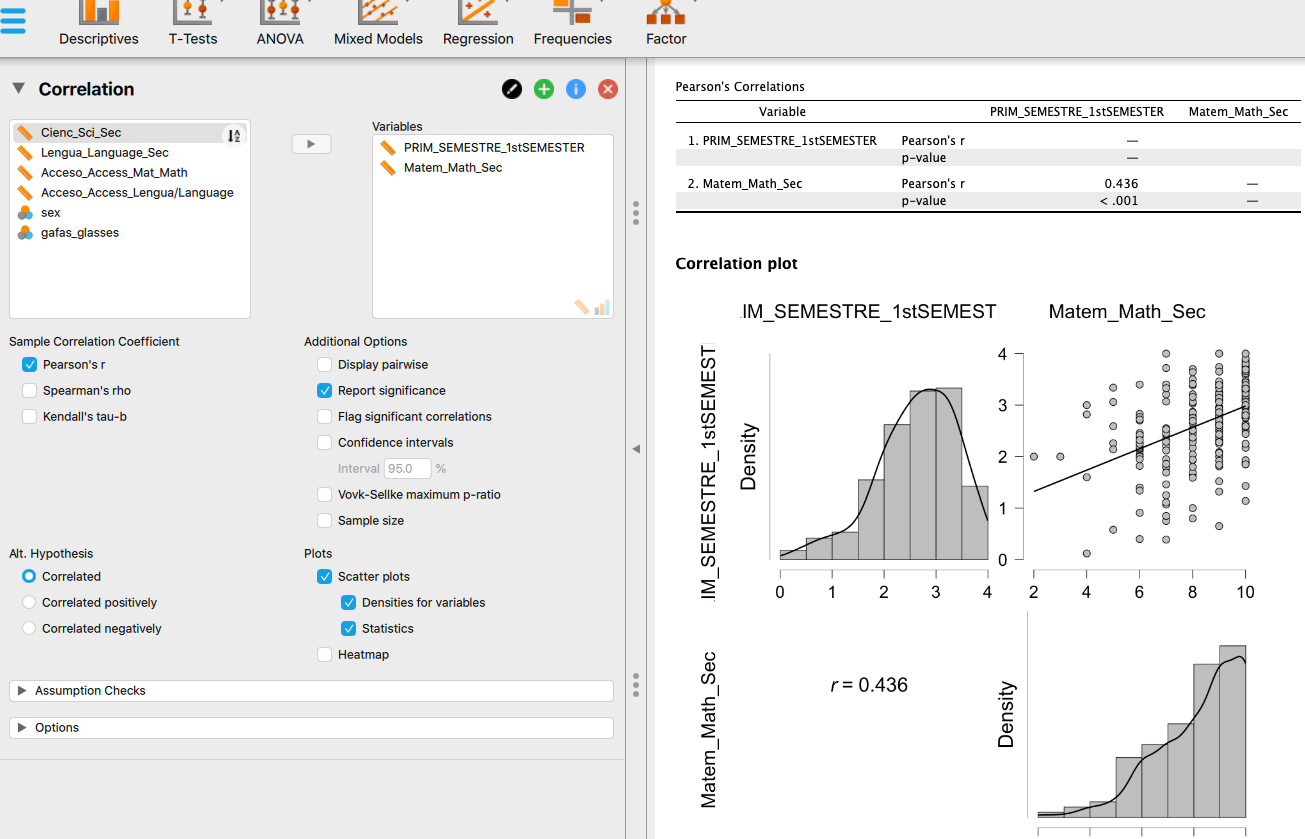
**Correlation and regression, November 23, 2020**

We have the following data set (<https://www.uv.es/mperea/PRIM_SE.jasp>). This file contains information on the grades in the first semester of a group of students in a science career, as well as information on the average grades (in secondary school) in math, science, and language. We also have their math and language test scores (PAU; “prueba de acceso a la Universidad”), as well as sex.

The idea (overall) is to see how predictive the grades in high school are once the students arrive at college (i.e., the grades in their first semester in college).

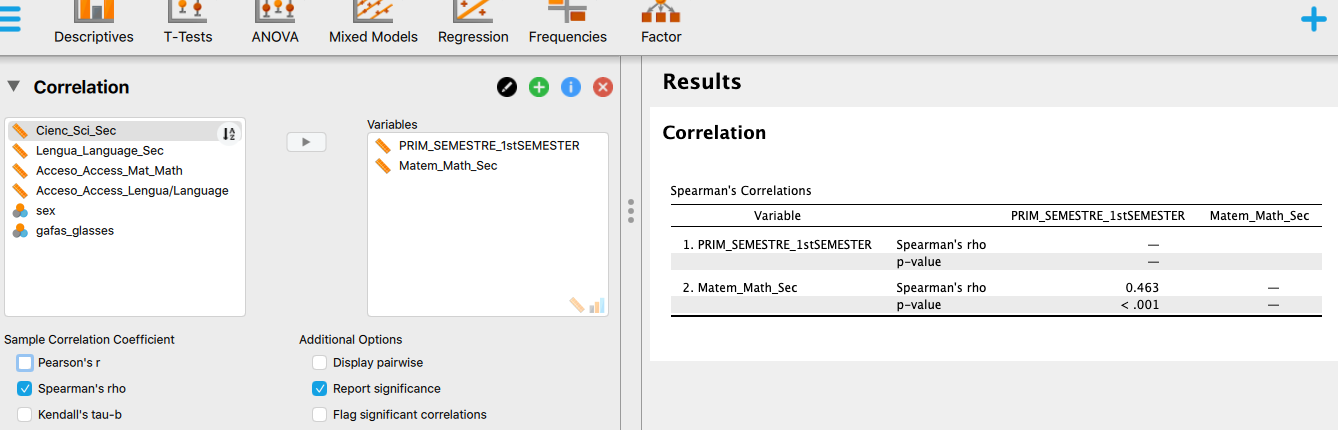
Correlation

*--What is the relationship (sign and strength) between the grades in the first semester (Y) and average high school grades in math (X)? Plot the appropriate graph and compute the appropriate index.*



The scatterplot shows a moderate positive relationship. The Pearson correlation coefficient is 0.436.

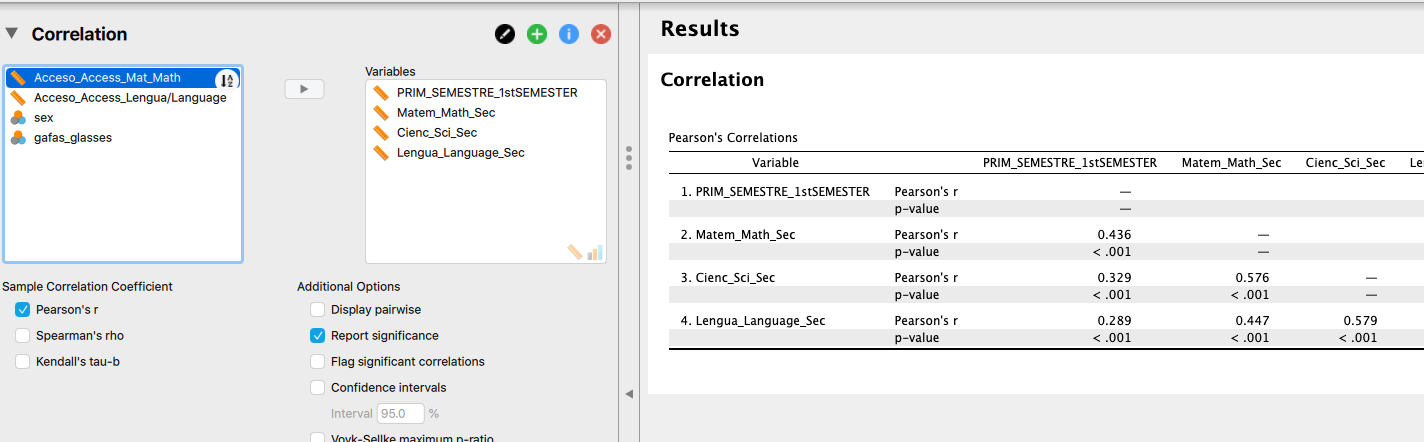
*--If we had considered the previous variables as "ordinals" (they are actually quantitative!), which index would you calculate and what would be its value?*



In this case, we would have to calculate Spearman's correlation coefficient, which is interpreted in an analogous way to Pearson's; of course, it is not "linear" since it is assumed that the relationship would be ordinal.

Spearman's coefficient is 0.463.

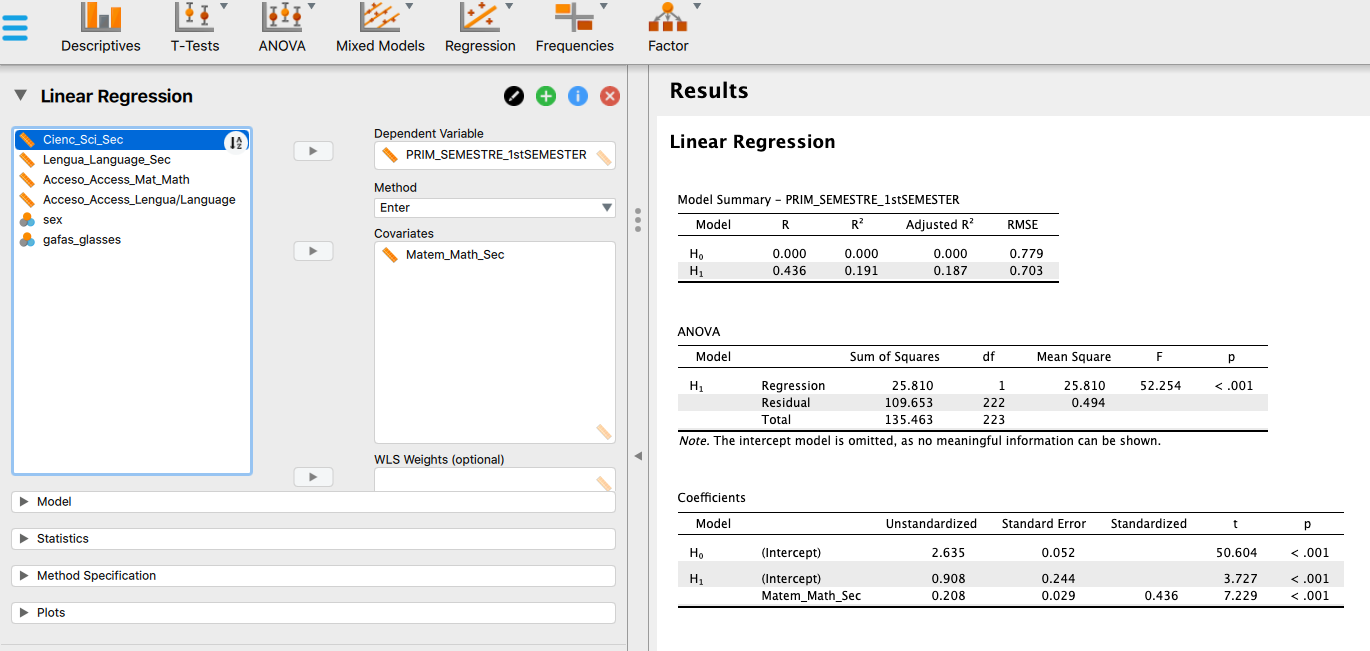
*--We want to have a correlation matrix with the following variables: grades of the first semester, average high school grades in mathematics, average high school grades in science, and average high school grades in language. For which variables was the relationship strongest? And weakest? Do you think it would have been possible to find negative relationships between these variables?*



The strongest relationship is the one with the highest Pearson's correlation coefficient in absolute value (remember that the sign only provides the direction; the sign does not inform about the strength): language grades in high school and science grades high school. The weakest is language grades in high school and grades in the first semester in the university.

Simple Regression

*--What is the regression line for the grades in the first semester (Y) when using the high school average grades in mathematics (X) as a predictor (in direct scores)? What is the meaning of the intercept and slope in that equation?*



GRADE\_UNIV ' = 0.908 + 0.208\*GRADE\_MATH\_SEC

0.908 (A; intercept) is the predicted value in the grade in the university if the math grade had been zero

0.208 (B) is the slope. Each one unit increase in math grades in high school is accompanied by a 0.208 increase in the grade in the university.

*--Idem in typical (standardized) scores. Does the value of the slope necessarily coincide with...?*

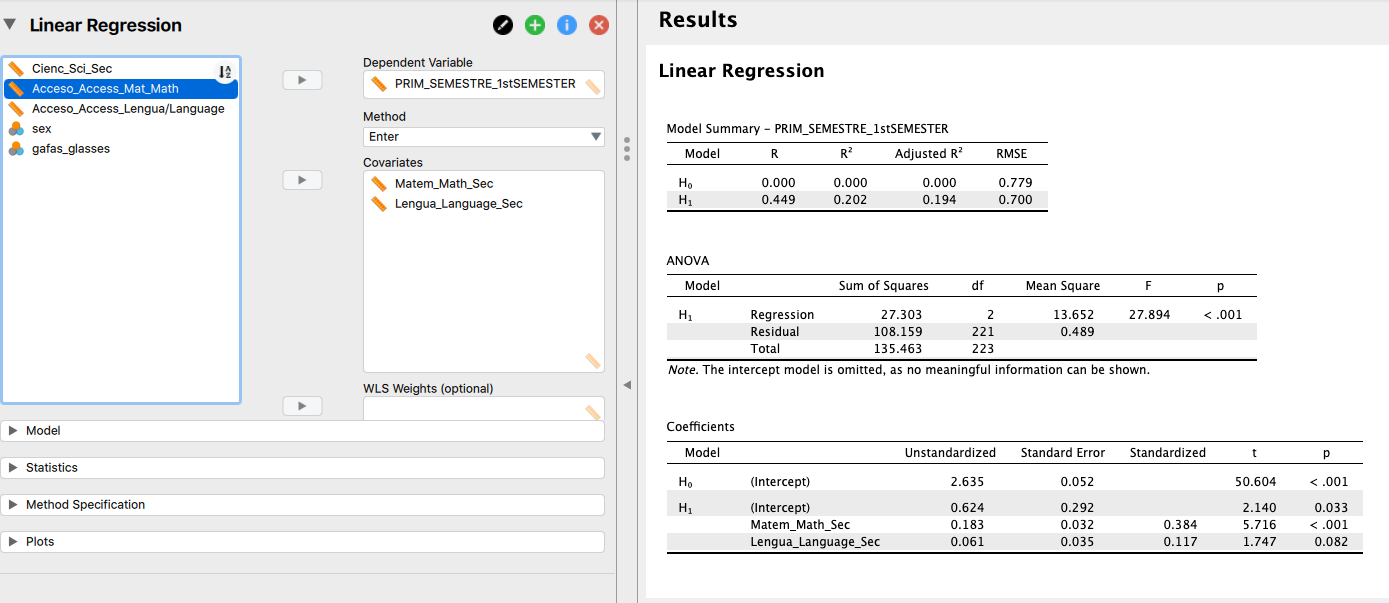
It is Pearson's correlation coefficient: 0.436.

*--What percentage of variance of the scores in the first semester (Y) can explain the regression line?*

It's R^2. 0.187 (if we look at the adjusted R^2 index). When stated as a percentage: the regression equation explains 18.7% of the grades in the first semester at the uni.

Multiple regression

*--What is the regression equation for the grades in the first semester (Y) when the predictors are the average high school grades in math and the average high school grades in language? What are the components of the regression equation in direct scores?*



GRADE\_UNIV ' = 0.624 + 0.183\*GRADE\_MATH\_SEC+ 0.061\*GRADE\_LANGUAGE\_SEC

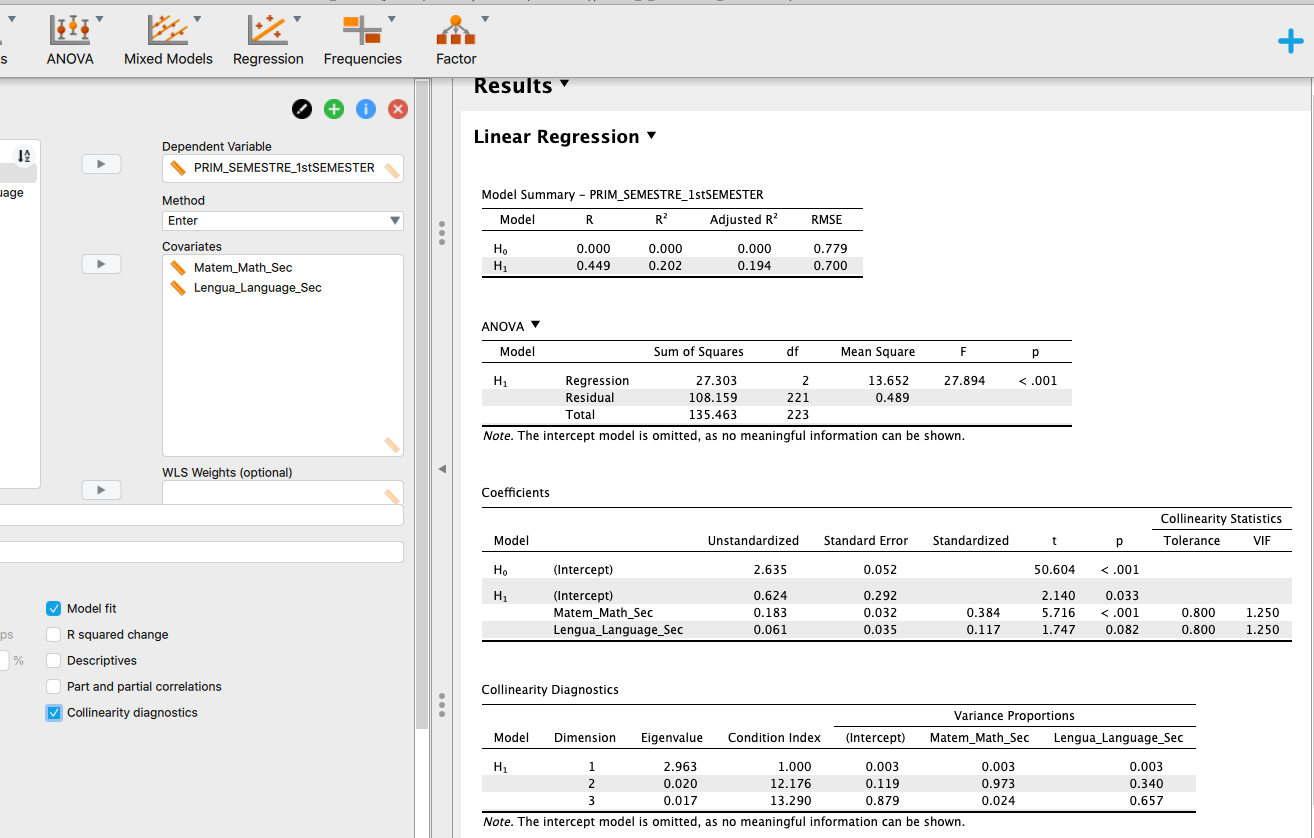
*--What percentage of variance of the grades in the first semester can explain the regression equation in the previous question?*

It's R^2. 0.194 (if we look at the adjusted R^2 index). To put it in percent: the regression equation explains 19.4% of the grades in the university.

*--What is the best predictor of the above equation and why? What is the worst predictor of the above equation and why?*

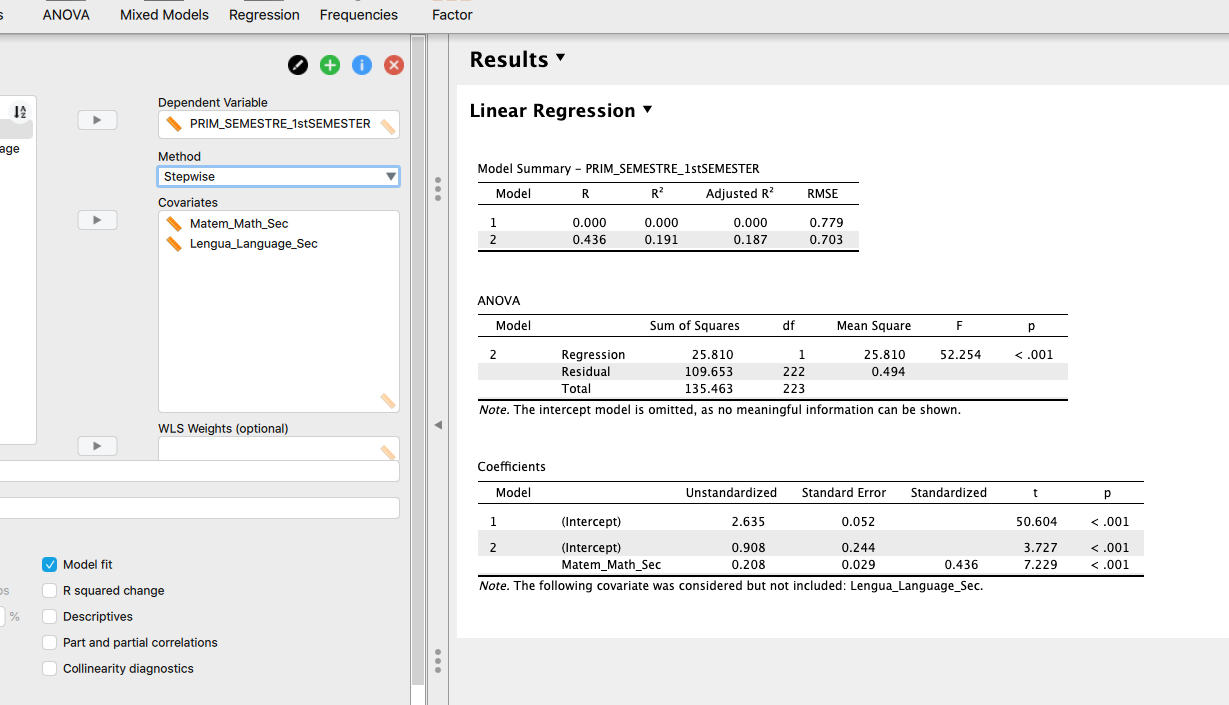
The best predictor is that with the highest value (in absolute value) in its standardized coefficient (these coefficients can be interpreted in the style of correlation coefficients, although they are not exactly Pearson's coefficient). In our case, the answer is the grades in math.

*--Do you think there were collinearity problems in the equation? (remember: look at the VIF, in the ideal case it would be 1.0)*



The VIF value is 1.25. It is not the most perfect (which is 1.0) but it is far from the VIF limit of 10.

*--Let's assume that, for parsimony, we want to compute a stepwise regression. What would be the resulting equation? Why?*



Only one predictor enters the equation: math scores in high school. Keep in mind that in the previous questions, the standardized coefficient of the grades of language in high school was very small. Thus, it did not enter the equation.

With this procedure "stepwise" only those predictors will enter the regression equation that contribute something substantial to explain the dependent variable (grades of the first semester in the university), as is the case of the grades in math (secondary school). But not the grades in language (secondary school). Indeed, the R^2 values were very similar when having one predictor and when having two predictors. So was this second predictor necessary? Clearly, the answer is “no”.