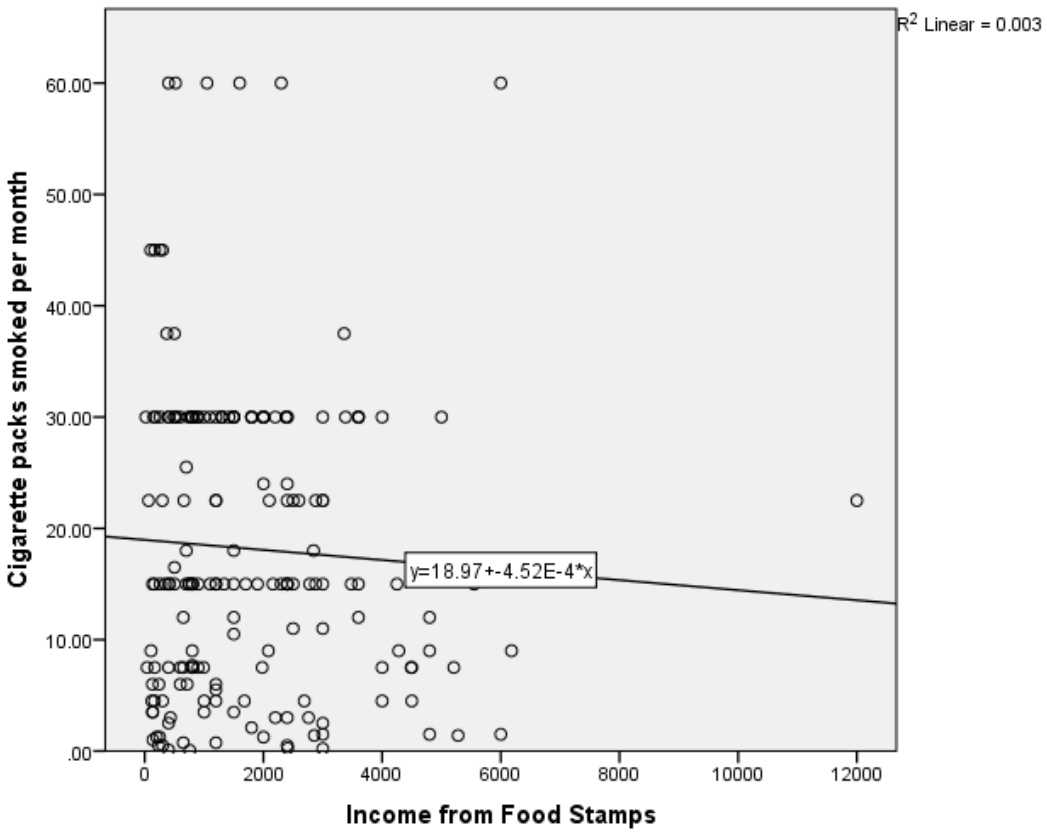
11. Correlation

Video Link:
https://www.youtube.com/watch?v=xs6j7OkDbqg&list=PL2fQHGEDK7Yyl1W9tgIo8wpYFTDumgcj&index=11

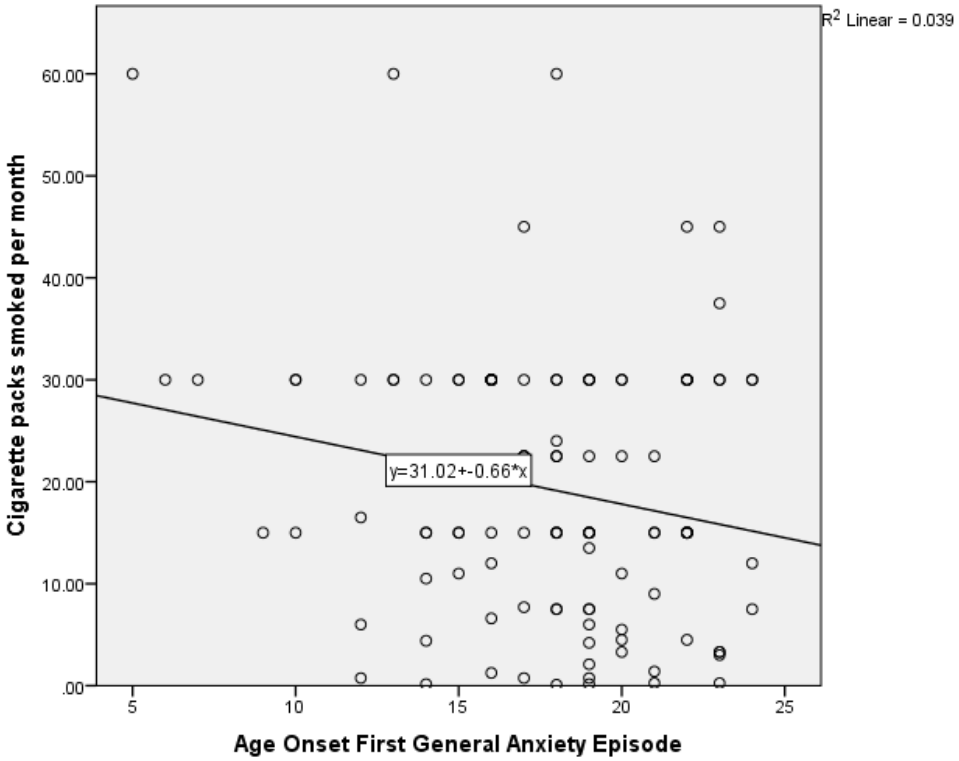
Section 11.1: Quantitative Explanatory Variable and Quantitative Response Variable
Section 11.2: R-Squared

Section 11.1: Quantitative Explanatory Variable and Quantitative Response Variable

To demonstrate how to request a correlation coefficient in SPSS, let's go back to the scatter plots we created for some of the NESARC variables. We used these scatter plots when visualizing the association between two quantitative variables. The first scatter plot shows Association between Income from Food Stamps by Number of Packs Smoked Per Month in Young Adult Smokers. From looking at the scatterplot we can guess this is a negative association, that is, as food stamps income goes up number of packs smoked per month goes down.

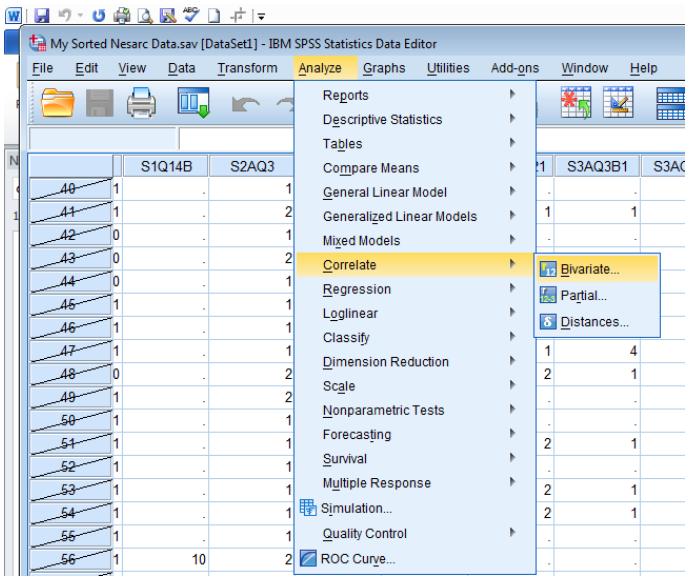


The second shows the Association between Age Onset First General Anxiety Episode by Number of Packs Smoked Per Month in Young Adult Smokers. From looking at the scatterplot we can guess this is a negative association, that is, the older onset of first general anxiety episode occurred the lower the number of packs smoked per month.

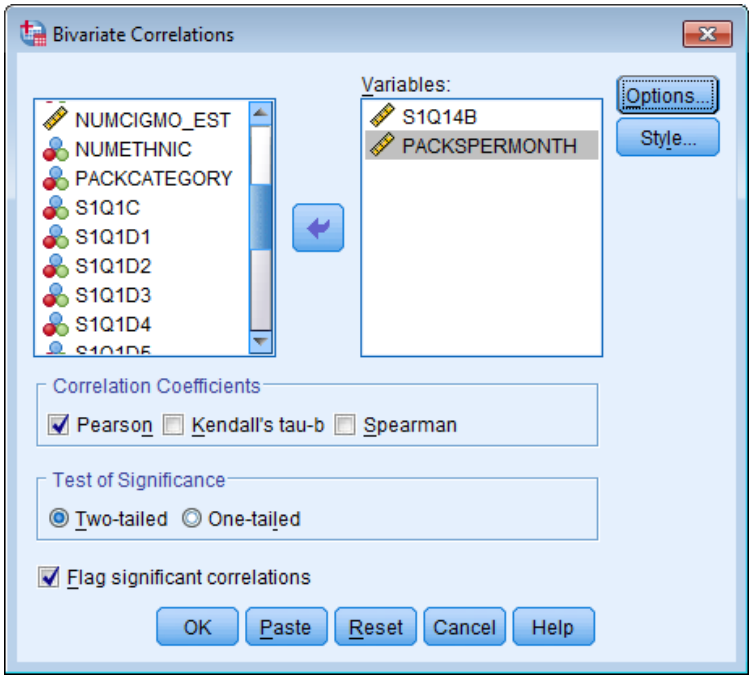


Now let's find the correlation coefficient.

1. Click **Analyze > Correlate > Bivariate**.



2. Select your Quantitative Explanatory Variable and Quantitative Response Variable from the left hand side using the arrow to move to the **Variables:** window. Click **OK**.



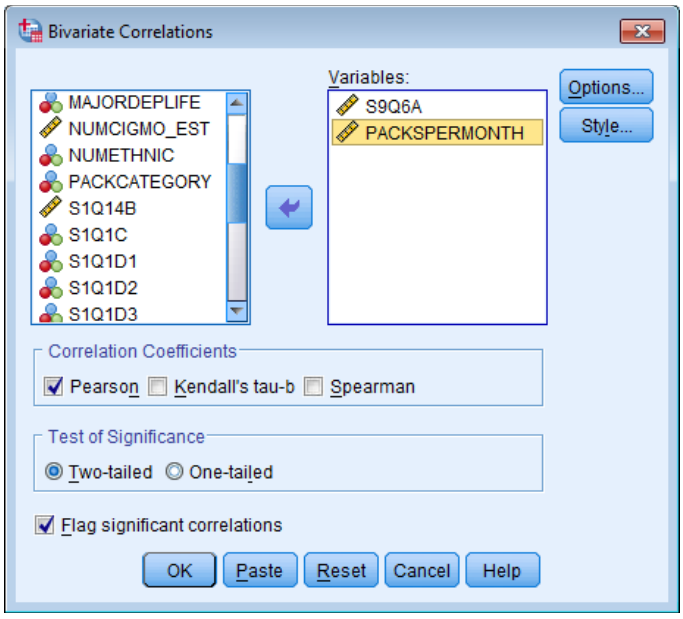
To locate the correlation coefficients of interest and the associated p values, we need to examine the Pearson Correlation Coefficient table here, and find the row and column where our two variables of interest intersect.

Correlations

		Income from Food Stamps	Cigarette packs smoked per month
Income from Food Stamps	Pearson Correlation	1	-.053
	Sig. (2-tailed)		.471
	N	187	186
Cigarette packs smoked per month	Pearson Correlation	-.053	1
	Sig. (2-tailed)	.471	
	N	186	1697

For the association between Income from Food Stamps and Cigarette Packs Smoked Per Month, the correlation coefficient is approximately -0.053 with a p-value of .471. This tells us that the relationship is not statistically significant. Now we can actually interpret the scatter plot and the coefficient together. There is no association essentially, as the scatter plot had already shown us through lack of a trend or pattern in the graph. That is, it's highly likely that a relationship of this is due to chance.

For our second example…



For the association between Income Age Onset First General Anxiety Episode and Cigarette Packs Smoked Per Month, the correlation coefficient is approximately -.198 with a p-value of .40 This tells us that the relationship is statistically significant.

Correlations

		Age Onset First General Anxiety Episode	Cigarette packs smoked per month
Age Onset First General Anxiety Episode	Pearson Correlation	1	-.198*
	Sig. (2-tailed)		.040
	N	108	108
Cigarette packs smoked per month	Pearson Correlation	-.198*	1
	Sig. (2-tailed)	.040	
	N	108	1697

*. Correlation is significant at the 0.05 level (2-tailed).

Now we can actually interpret the scatter plot and the coefficient together. The association between Income Age Onset First General Anxiety Episode and Cigarette Packs Smoked Per Month is fairly weak and it's also negative, as the scatter plot had already shown us. It is statistically significant (i.e. $p<.05$).

Section 11.2: R-Squared

Here's some good news. Post hoc tests are not necessary when conducting Pearson correlation. Post hoc tests are needed only when your research question includes a categorical explanatory variable with more than two levels. Because our explanatory variable and the context of correlation coefficient is quantitative, there's never a need to perform a post hoc test.

Another interesting and useful aspect of the correlation coefficient is if we square the correlation coefficient. That is, we multiply it by itself, we get a value that also helps our understanding of the association between the two quantitative variables.

Small r squared is the fraction of the variability of one variable that can be predicted by the other. For example, when looking at the relationship between Age Onset First General Anxiety Episode and Cigarette Packs Smoked Per Month, if we square our correlation coefficient of 0.198, we get 0.04. This could be interpreted the following way. If we know the Age Onset First General Anxiety Episode, we can predict 4% of the variability we will see in Cigarette Packs Smoked Per Month. This of course is a less than ideal r-squared but is too be expected with a small correlation coefficient that is practically insignificant.

Of course, that also means that 96% of the variability is unaccounted for.

Since the relationship between Income from Food Stamps and Cigarette Packs Smoked Per Month is non-significant it is inappropriate to discuss r-squared.

Again, correlation coefficients are commonly denoted with a lowercase r, and they're squared to determine the amount of variability that can be predicted.

You might be wondering how much variability in Internet use rates can be predicted if we simultaneously consider both urban rate and income per person. A multivariate inferential tool called multiple regression can be used to answer this question and we'll discuss that in a later tutorial.