Correlation

This week, we will use the Crime Rates Datasets of NSW Local Government Areas (NSW Crime). This workshop introduces how to 1) produce a matrix of bivariate correlations and 2) create scatterplots.

1. **Bivariate Correlations**

Suppose we want to know how robbery rates in NSW Local Government Areas (LGAs) are correlated with unemployment rates. To create a correlation table, go to Analyze > Correlate > Bivariate (see <Figure 1>) . In the popped-up box, 1) move all the variables of interest (in this case, *robbery* and *unemploy*) to the right-side box of Variables. And then 2) Choose *Pearson* as Correlation Coefficients. You can include asterisks (\*\*) in your output to highlight statistically significant correlations by 3) checking the box “Flag significant correlations) (see <Figure 2>). 4) click OK.



<Figure 1>



<Figure 2>

<Table 1> is the output. Due to the table structure, correlations are always shown twice. And you may notice different numbers of cases across cells. Correlations can be calculated only by cases having valid values on both of two variables. With Pearson’s r of .261 and its associated p-value of .014, we can say that robbery and unemployment rates are weakly correlated, and statistically significant at .05



<Table 1>

You can simultaneously run correlation tests for multiple variables. Let’s say we think of other variables that may be correlated with robbery and unemployment such as percentage of residents who rent dwelling (*pctrent*), income inequality measured by Gini coefficient (*giniinc*), median house prices (*housmedprice*), and median age of residents (*medage*). You can simply add all the variables of interest to the list of *Variables* (see <Figure 3>).





<Figure 3>

When you enter three or more variables into the correlation matrix, how to treat missing values can make a very large impact on your output. The default is pairwise deletion by which SPSS uses all the available cases for each pair of variables. This method will produce a correlation matrix with different size of cases across cells but uses maximum amount of information available. Another option is listwise deletion by which SPSS only includes cases that have complete information on all the variables listed. The output matrix will be produced based on the exact same dataset, but you could lose sizable amount of data depending on the pattern of missing values in your data. We will use listwise deletion for more consistent comparison across different analyses. To choose this option, 1) open *Options*, 2) tick *Exclude cases listwise* and then3) press *Continue* (see <Figure 3>).



<Table 2>

Carefully read the output <Table 2>. Which variables have significant correlations with robbery rates? Why do you think the correlation between unemployment and robbery becomes non-significant?

1. **Scatterplots**

Scatterplot is another effective way to examine bivariate association in a graph format. To create a scatterplot, go to Graphs > Legacy Dialogs > Scatter/Dot (see <Figure 4>). When you have only two variables, select the Simple Scatter (see <Figure 5>).



<Figure 4>



<Figure 5>

In a new popped-up window, 1) move your dependent variable (robbery) into Y Axis, 2) independent variable (unemploy) into X Axis, and 3) click OK.



<Figure 6>

<Figure 7> is the scatterplot of robbery rate by unemployment rate. Do you see a pattern?



<Figure 7>

You can create scatterplots for several different variables so that we can determine whether there are any relationships among those multiple variables. To do so, you should select Matrix Scatter (see <Figure 8>) and put all the variables of interests into *Matrix Variables* (see <Figure 9>). It will return a output seen as <Figure 10>.



<Figure 8>



<Figure 9>



**Workshop Activities**

Q1. Generate a scatterplot for the following pairs of variables. Which scatterplot suggests the strongest association?

1. Gini coefficient of total income by unemployment rate
2. Gini coefficient of total income by % of residents who rent dwelling
3. Gini coefficient of total income by median sale price of houses
4. Gini coefficient of total income by median age

Q2. Generate a scatterplot matrix with all the variable used in the Q1. (That is, giniinc, unemploy, pctrent, housmedprice, and medage). Which variables show negative association with median sale price of houses?

Q3. Now we go back to our examination about how various characteristics of LGAs are associated with crime rate, but we change our focus from robbery to sexual offences (*sexoff*). We continue use the same sets of independent variables about LGA attributes we had examined for robbery rate.

1. Create a scatterplot using sexoff and unemploy and compare the output with the one with robbery (Figure 7). With which crime rate does the unemployment rate have stronger association? Robbery or sexual offence?
2. Construct a correlation matrix with variables of sexoff, unemploy, pctrent, giniinc, housmedprice, and medage (For the convenience of comparison we will do in the following question Q3-3, put the variables in the exact order listed). Which variables are significantly correlated with sexoff? And how are they correlated with sexoff? (interpret both direction and strength of the correlation).
3. Now compare your answer to the question above (Q3-2) with the <table 2> on robbery rate. Do you find variables are differently associated with robbery and sexual offences? (You may want to focus on significant correlations only). How would you explain the different patterns of correlation?