**CORRELATION AND REGRESSION ANALYSIS: OVERVIEW**

**by Simon Moss**

 Correlation and regression analysis are utilized to compare the association between numerical measures—such as the amount of food that animals consume and their weight. Several variants need to be considered. To decide which variants to utilize, you need to answer four questions

|  |
| --- |
| **Number of variables in each set** |

 Consider a researcher who wants to assess whether the economy of a nation is associated with sporting prowess. Three variations can be differentiated. First, the researcher could use more than one variable to assess economy—such as inflation and unemployment rate—and more than one variable to assess sporting prowess—such as number of Olympic medals and ranking of the national soccer team. This design will be called multiple predictors and multiple outcomes.

 Second, the researcher could use one variable to assess economy but more than one variable to assess sporting prowess or vice versa. This design will be called multiple predictors and one outcome.

 Third, the researcher could use one variable to assess economy and one variable to assess sporting prowess. This design will be called one predictor and one outcome. Decide which of these three alternatives correspond to the analysis you want to undertake. Do not be concerned if you are not sure which variables are the predictors and which variables are the outcomes.

|  |
| --- |
| **Control variables** |

 Next, you might want to include what are called control variables. To illustrate, consider the table below, in which each row represents one nation

|  |
| --- |
| Data from this study |
|  GDP | Number of Olympic medals each year | Population  |
| $21.2 trillion | 61 | 120 million |
| $28.2 trillion | 40 | 105 million |
| $24.2 trillion | 29 | 92 million |
| $19.5 trillion | 38 | 201 million |
| $31.3 trillion | 51 | 83 million |
| $26.9 trillion | 62 | 81 million |
| $0.5 trillion | 9 | 2 million |
| $1.2 trillion | 8 | 3 million |
| $1.4 trillion | 5 | 7 million |
| $0.2 trillion | 7 | 3 million |
| $1.9 trillion | 11 | 9 million |
| $0.1 trillion | 4 | 11 million |

 This table generates some interesting conclusions

* If you scan the first two columns, you will conclude that nations with a strong economy—as measured by GDP—are more likely to excel in sport
* But another explanation is possible. Perhaps population affects both the GDP and number of Olympic medals
* So, to assess whether the economy really enhances sporting prowess, the researcher needs to control population. For example, the researcher needs to examine only nations with a large or small population.
* Indeed, as the following table shows, if you examine only nations with a small population, the association between GDP and sporting prowess is not as apparent
* In short, we should control variables that could affect both the predictor and outcome, such as population—called spurious variables

|  |
| --- |
| Data from this study |
|  Which books they prefer | Social skills—as measured by a special test | Age group  |
| $21.2 trillion | 61 | 120 million |
| $28.2 trillion | 40 | 105 million |
| $24.2 trillion | 29 | 92 million |
| $19.5 trillion | 38 | 201 million |
| $31.3 trillion | 51 | 83 million |
| $26.9 trillion | 62 | 81 million |
| $0.5 trillion | 9 | 2 million |
| $1.2 trillion | 8 | 3 million |
| $1.4 trillion | 5 | 7 million |
| $0.2 trillion | 7 | 3 million |
| $1.9 trillion | 11 | 9 million |
| $0.1 trillion | 4 | 11 million |

 Besides spurious variables, researchers might also want to control variables for other reasons. In particular, the measures are sometimes contaminated or confounded with other variables. To illustrate, perhaps the measure of Winter Olympic medals is also confounded with the magnitude of snow that falls in the nation.

 In short, at times, you might want to control variables, such as population or magnitude of snow. You can apply two approaches to control variables:

* You can examine only a subset of participants, such as only nations with a population of around 5 million people
* Or you can utilize statistical tests to predict what the results would be if you controlled variables—such as if the nations were average in population.

 So, when should you control variables? You should control variables whenever you have collected information about a variable, such as population, that is likely to be strongly associated with the measures. Population is likely to be associated with Olympic medal tally, so population, should be controlled if possible. Language is not as likely to be associated with Olympic medal tally, so language might not need to be controlled.

|  |
| --- |
| **Categorical variables** |

 Imagine you ask 10 males and 10 females to specify their confidence on a scale from 1 to 10 and then to indicate whether they prefer to read online books or actual books. In this example, confidence is a numeric variable: Each person is assigned a number. In contrast, reading preference is a categorical variable: Each person is assigned a category, such as online or actual.

 Even if we coded online books as 0 and actual books as 1, reading preference would still be a categorical variable. The numbers 0 and 1 are merely symbols to represent categories. They are not true numbers. For example, actual books, or 1, are not actually higher than online books, or 0.

 Some categorical variables, called dichotomous variables, comprise two categories only, such as whether people live in the northern hemisphere or southern hemisphere. Some categorical variables comprise more than two categories, such as hair color: blond, brown, black, and red.

 You need to ascertain whether any of your variables are categorical. Then you need to determine whether these categorical variables are dichotomous or not.

|  |
| --- |
| **Moderators** |

 Finally, the relationship between two variables is often dependent upon another variable. For example, whether GDP is related to number of Olympic medals might depend on whether the population values sport. If the population values sport, GDP may be strongly associated with Olympic medals: Nations with a higher GDP will tend to earn more medals. In contrast, if the population does not value sport, GDP may not be strongly associated with Olympic medals: Nations with a higher GDP might not earn more medals because they might not invest this money into sport.

GDP of nation

Number of Olympic medals

Nations that

value sport

Nations that do

not value sport

 Variables that affect the level of association between two other variables are called moderators. In this example, the degree to which the population values sport is a moderator. You should consider whether some moderator could affect the association between your variables. If so, determine whether or not you have measured this moderator.

|  |
| --- |
| **Table of alternatives** |

 Once you have answered the previous questions, you should utilize the following table to ascertain which technique to utilize.

|  |
| --- |
| Table to decide which technique to utilize |
| Design  | Do you want to control variables as well | Does the analysis include categorical variables | Does the analysis include moderators |  Technique |
| One predictor and one outcome | No | No  | No  | Correlation—strictly called Pearson product moment correlation.  |
| One predictor and one outcome | Yes | No | No | Multiple regression—also called ordinary least squares regression or linear regression |
| Many predictors and one outcome | Does not matter | No | No | Multiple regression |
| Many predictors and one outcome | Does not matter | Yes | No | Multiple regression with dummy variables |
| Many predictors and one outcome | Does not matter | No | Yes |  Moderated regression |
| Many predictors and one outcome | Does not matter | Yes | Yes | Moderated regression with dummy variables |
| Many predictors and many outcomes | No | No | No | Canonical correlation. Or, more commonly, simply conduct a separate multiple regression for each outcome, such as each measure of sporting prowess  |
| Many predictors and many outcomes | Yes | If yes, include dummy variables | If yes, include moderators in the design | Conduct a separate multiple regression for each outcome, such as each measure of sporting prowess |