

## Chapter 5

### 5.1

Causality is a big issue in statistics and philosophy. One reason for this is because there is little agreement as to exactly what the word causality means. It is usually accepted that causation can be established if we satisfy three criteria: association, direction of influence, and isolation. We will discuss these three points in more detail

#### Association

The first thing we learn is that correlation does not mean causation. However you should also remember that causation does imply correlation. If two variables are causally related, a change in one must produce a change in the other. Therefore, a statistical association (be it a regression coefficient or a correlation) is necessary but not sufficient to make a claim of causality.

#### Direction of causality

If two variables (we call them A and B) are associated, there are three possible reasons for the association:

- It is possible that A is a cause of B.
- It is possible that B may be a cause of A.
- Another variable, C is a cause of both A and B.

Therefore, for any correlation between two variables, it is not evident what the underlying causal mechanism is giving rise to the observed association. In other words, we do not know the direction of the causality. How can we tell whether A causes B, or B causes A? The answer is that we always expect the cause to come first in time, before the effect. If A causes B, a change in A should result in a change in B after a particular time period. So we need to be able to demonstrate temporal priority; that is, that changes in the dependent variable must be observed after a change in the independent variable, in other words A always precedes B.

The time interval between cause and effect may vary widely depending on the variable in question. The notion of temporal priority is central to the basics of experiment design because the manipulation of the independent variable always precedes the measurement of the dependent variable. This temporal precedence cannot be observed in non-experimental or cross-sectional research where all data are usually collected at one point in time. In this case we rely on mental experiments. Mental experiments are decisions about the direction of causality based on theory, previous research and, in many situations, common sense.

#### Isolation

To be certain that an independent variable, A, is a cause of a dependent variable, B, it is necessary to isolate the dependent variable (B) from all influences other than the hypothesised cause (A). This isolation is what experimentation attempts to achieve. In practice experimentation can only approximate isolation, or achieve what is called pseudo-isolation.

Whereas experimental control can be used to isolate the independent variable, in non-experimental research the independent variable(s) cannot be isolated. This needs to be done in another way. Remember that a regression slope indicates the effect of the independent variable on the dependent variable while holding constant the effect of all other independent variables in the equation. Therefore regression models can be used to isolate the influence of an independent variable.

To tie together the three criteria discussed above we need to go beyond statistical analysis, we need to involve theory. When psychologists collect data, they want to collect data that are both accurate and useful. This costs a lot of time. Theory has a most important function in successful use of regression and should not be underestimated.

A series of findings, all related to correlations can collectively be evidence of a causal process. This series of findings is referred to as the signature of that process.

## 5.2

We will talk about what the effect of the sample size is on the regression analysis. The main idea regarding sample size is that, the bigger the sample size is the better. The standard error of a mean is equal to:

$$se(\bar{x}) = \sqrt{\frac{sd^2}{n}} \quad (50)$$

From this equation you can see that the larger the  $n$ , sample size is, the larger the denominator is and the standard error will therefore be smaller. This means that your parameter estimates will be more accurate. Besides that a lower standard error will increase the chance of finding a significant association. However there are problems with samples that are too large. When a sample is unnecessarily large, collecting more data would have wasted expense and time. Also ethical boards spend more time in reviewing the sample sizes. They say that participants donate their time in the hope that they may be doing some good.

There are two ways to determine an appropriate sample size. The first is the rules of thumb that represent simple rules that suggest minimum sample sizes. The second is a power analysis. We will now discuss both methods.

Rule of thumbs are mostly very simple. Green has proposed a method for determining the minimum sample size to test the  $R^2$  of a regression model. He said that the minimum sample should be greater than  $50 + 8k$ , where  $k$  is equal to the number of independent variable. And if you want to carry out significance tests on regression slopes, the size should be greater than  $104 + k$ . However such rules of thumb do not take into account the expected effect size, or desired power of the test. So such rules lack generality and may even mislead.

To use the power analysis we need to have the following pieces of information:

- The value of alpha, this is the level of significance that we use as the criterion for determining whether or not we have a significant effect. This value is mostly set at 0.05 (5%). When alpha is larger, the more chance we have of finding a significant effect, however the change of setting a spurious result also increases. The probability of a type I error is equal to the value of alpha.

- The size of the effect in the population that we would be interested in. The effect size is in a multiple regression equal to the  $R^2$ . The larger the effect size, the greater the chance of finding it. However if the effect size is sufficiently small, then finding it would not be useful. The effect size can be determined in three ways:
  1. Effect can be based on substantive knowledge.
  2. To base an estimate of the effect size based on previous research.
  3. Use conventions for determining expected effect size. Cohen has defined the amount of effect size for values of  $R^2$ . When  $R^2 = 0.02$  it is a small effect size, when  $R^2 = 0.13$  it is a medium effect size and when  $R^2 = 0.26$  it is a large effect size.
- An appropriate level of power must be selected. Power is the probability of finding a result given that the effect does exist in the population. By convention, power is set to 0.80. That gives an 80% chance of finding a significant result if there is an effect of the specified size in the population from which the sample is taken. The probability of a Type II error is 1- power, so in this case,  $1 - 0.80 = 0.20$  (20%).

From his information we can calculate the number of participants required. You can use programs such as G\*power to calculate this. From this program you can get some graphs, which show the difference in the amount of participants. From this you can draw some conclusions, but you cannot use it for accurate power calculation. (See graphs on page 122-125).

## 5.3

The next issue we will examine is collinearity, sometimes referred to as Multicollinearity. Collinearity refers to the size of correlations among the independent variables in a regression calculation. It happens because two (or more) independent variables correlate. This means that it is difficult for the regression calculation to determine which of them is actually the more important one of the two; it could be either, so we have increased uncertainty (standard errors) and inaccuracy (slope coefficient). So we cannot decide which variable is important in determining the outcome. The regression calculations imply that they take account of this uncertainty, and call it a larger standard error. In more formal terms, when the correlation between two independent variables is one (or very close to one) or the multiple correlation between any independent variable is one (or very close to one), this is referred to as perfect or complete collinearity. It is an assumption of regression that perfect collinearity is not present, and if perfect collinearity does occur, most statistical packages will stop and produce an error message. However, perfect collinearity is a very infrequent occurrence with real data, unless a data entry or manipulation error of some sort has been made. If this does occur, it is most likely that two or more independent variables have been summed to create an additional variable that is then also used as an independent variable.

A common occurrence that our collinearity is high enough to cause some problems, but not actually high enough to violate the assumptions of regression.

When you find regression coefficients that are not significant, when the overall equation is significant, you should suspect that collinearity might have played a part. What may have happened in this case is that the regression equation 'knows' that a high proportion of the variance can be explained by the independent variables, but it does not know what size parameters estimates to assign to which independent variable. If you suspect collinearity problems, there are several methods of determining the severity of the problem. We will discuss three of these methods.

Inspect visually the matrix of correlations amongst the independent variables. If one variable correlates fairly high with the other two variables that can be a clue that collinearity is a problem. However, low correlation does not indicate that there is no problem. The correlation tells us how much variance two variables share. The value we need to know is the proportion of variance in each independent variable, which is shared by all of the other independent variables. We can find this out to conduct a multiple regression analysis. We need to do a separate regression analysis for every independent variable, which is tedious. However, most statistical analysis software will give two other diagnostic statistics to help you to diagnose collinearity. These are tolerance and the variance inflation factor (VIF).

Tolerance is a very slight extension of  $R^2$ ; the tolerance of an independent variable is the extent to which that independent variable cannot be predicted by the other independent variables. Tolerance is calculated by  $1 - R^2$ . If there are only two independent variables in the regression analysis, the value of  $R$  (the multiple correlation) will be equal to the value of  $r$  (the bivariate correlation), and tolerance can therefore be calculated from the bivariate correlation matrix. Tolerance varies between zero and one. A tolerance of 0 for a variable means that it is completely predictable from the other independent variables. Therefore is perfect collinearity. If a variable has a tolerance value of 1, this means that the variable is complete uncorrelated with the other independent variables.

The variance inflation factor (VIF) is closely related to tolerance. When there are more than two independent variables the VIF is calculated using the formula:

$$VIF = 1/\text{tolerance} \quad (110)$$

This value relates to the amount that the standard error of the variable has been increased because of collinearity. The increase in standard error is equal to the square root of the VIF.

However what do you need to do when you find collinearity in your dataset? The best thing to do if collinearity is a serious problem is to discard the old data and go and collect new data, which avoids the problem in some way. But there are also other options:

1. Collect more data. It is not much of an improvement on collecting new data. Collinearity causes the standard errors to increase in size. Increasing the sample size has smaller standard errors, so a larger dataset, will in some way, make up for some of the effects of collinearity. It will not help when you have perfect collinearity
2. Remove or combine variables. If variables are highly correlated this implies that they are measuring similar constructs and that the information measured by those variables may be, at least partially, redundant. If you are dealing with a large number of independent variables, a more technical way of reducing the number of variables is to use principal components analysis (PCA). This is similar to factor analysis, in that it groups the original variables into a smaller number of uncorrelated 'factors'.
3. Stepwise entry. This is a form of hierarchical regression. It can be used when collinearity is a problem to select variables for analysis. However the only time we might temper this advice is when analysis is being carried out purely to predict a dependent variable.
4. Ridge regression. When collinearity is so high that the regression procedure cannot continue, a possible solution that has been proposed is ridge regression. This is a complex procedure, which is beyond the scope of this book. It is also hard to interpret, therefore is this rarely used.

## 5.4

This paragraph is about the one-factor repeated measures ANOVA (“ORMA”) model (OMRA is not an actual abbreviation, but used here for convenience).

First of all it is important to discuss this model’s characteristics: The one factor repeated measures model makes it possible to examine two or more measurements, forming an extension of the dependent t-test. The “repeated” part of ORMA means that each subject responds to each level of factor A (also referred to as within-subjects design). In this way subjects function as their own controls, as individual differences are taken into account. This has as a consequence, however, that subjects’ scores are not independent across the levels of factor A.

Due to subjects and variations caused by the interaction between A and subjects in ORMA, the residual variation is further decomposed into variation. This means a reduction in the residual sum of squares, a stronger model, and more accuracy in the estimation of the effects on A (which means less subjects are needed).

The ORMA is a mixed model. This is due to the fact that the subject-factor is a random effect, while the A factor is generally a fixed effect. The ORMA can also be seen as a special case of the two-factor mixed-effects design, only with one subject ( $n=1$ ) per cell.

The more negative aspects of the ORMA include that there is some risk of carryover effects from one level of A to another, due to the fact that each subject responds to all levels of A. These effects can be minimized by (1) counterbalancing the administration order of the levels of A in such a way that each subject does not receive the same order of the levels of A; (2) letting time pass between the administration of the levels; or (3) matching/blocking similar subjects with the assumption that subjects within a block are randomly assigned to a level of A (aka randomized block design).

The lay-out for the ORMA model is seen here:

	Level of Factor A (Repeated Factor)				
Level of Factor S	1	2	...	J	Row mean
1	Y <sub>11</sub>	Y <sub>12</sub>	...	Y <sub>1J</sub>	$\bar{Y}_{1.}$
2	Y <sub>21</sub>	Y <sub>22</sub>	...	Y <sub>2J</sub>	$\bar{Y}_{2.}$
....	...	...	...	...	...
n	Y <sub>n1</sub>	Y <sub>n2</sub>	...	Y <sub>nJ</sub>	$\bar{Y}_{n.}$
Column mean	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	...	$\bar{Y}_{.J}$	$\bar{Y}_{..}$

Here it can be seen as well that the ORMA model is a form of the two-factor model, but with only one observation per cell.

Columns are shown as the levels of factor A, and the rows as the subject (factor S). Columns here thus represent the different measurements. Subject means are shown, but rarely used.

The formula of model is as follows (written in terms of population parameters):

$$Y_{ij} = \mu + \alpha_j + \sigma_i + (\sigma\alpha)_{ij} + \epsilon_{ij}$$

$Y_{ij}$  is the observed score on the dependent variable for individual  $i$  responding to level  $j$  of factor A.  $\mu$  stands for the grand population mean.  $\alpha$  is the fixed effect for level  $j$  of factor A.  $s_i$  is the random effect for subject  $i$  of the subject factor.  $(s\alpha)_{ij}$  is the interaction between subject  $i$  and level  $j$ .  $\epsilon_{ij}$  is the random residual error for individual  $i$  in level  $j$ .

Measurement error and/or other unconsidered factors can lead to the residual error.

For this model the null hypothesis indicates that the means for each measurement are the same. This hypothesis is in the terms of means due to the fact that factor A is a fixed effect.

When it comes to assumptions, the ORMA is again very similar to the two-factor mixed-effects model. Just like in this model the assumptions of ORMA are mainly about the distribution of the random effects and the dependent variable scores. ORMA contains only two new assumptions that the two-factor mixed-effects model does not have:

- Compound symmetry: This assumption states that the co-variances between the subject scores remains constant across the levels of the repeated factor A. This assumption is often broken in ANOVA, particularly when factor A is time (as the continual change means the co-variances are not constant). When this assumption is violated there are three options: (1) limit the levels of factor A to (a) those that meet the assumption, or (b) to having only 2 repeated measures; (2) use adjusted F-tests; or (3) use MANOVA (multiple analysis of variance), which may be less powerful but does not carry the compound symmetry assumption. Interesting to note here is that the first counter-measure against carryover effects can also minimize problems with this assumption.
- Sphericity: This assumption states that for each pair of factor levels the variance of the difference scores is the same. This is the necessary and sufficient condition for the validity of the F-test (compound symmetry is sufficient but not necessary).

The ANOVA summary table for this model is as follows:

Source	SS	df	MS	F
A	SSA	$J - 1$	MSA	MSA/MSSA
S	SSS	$n - 1$	MSS	
SA	SSSA	$(J - 1)(n - 1)$	MSSA	
Total	SS <sub>total</sub>	$N - 1$		

With sources of variation the ORMA is again similar to the two-factor model, with the exception that ORMA has no within-cell variation. As the table shows the sources of variation here are: A (the repeated measure), S (the subjects), SA (the interaction between A and S), and the total. Even though this means we can compute three main squares terms, there is only an R-ratio result for factor A. This shows that there is no appropriate error term for the subjects' effect, and this cannot be tested.

The sum of squares for ORMA also needs to be considered. Decomposing the total sum of squares is done as follows:

$$SS_{\text{total}} = SSA + SSS + SSSA$$

The expected mean squares are important for the formation of the proper F-ratio. The expected mean squares are, however, dependent on whether the null hypothesis (means are the same for each of the measures) is true or not. When  $H_0$  is true the expected mean squares are:

- $E(MSA) = \sigma_\epsilon^2$
- $E(MSS) = \sigma_\epsilon^2$
- $E(MSSA) = \sigma_\epsilon^2$

Here  $\sigma_\epsilon^2$  is the population variance of the residual errors.

When  $H_0$  is false, the expected mean squares are:

$$E(MSA) = \sigma_\epsilon^2 + \sigma_{s\alpha}^2 + n \left( \frac{\sum_{j=1}^J \alpha_j^2}{J-1} \right)$$

$$E(MSS) = \sigma_\epsilon^2 + J\sigma_s^2$$

$$E(MSSA) = \sigma_\epsilon^2 + \sigma_{s\alpha}^2$$

Here  $\sigma_s^2$  is the variability due to subjects, and  $\sigma_{s\alpha}^2$  is the interaction of factor A and subjects.

The proper F-ratio is formed using this formula:

$$F = \frac{\text{systematic variability} + \text{error variability}}{\text{error variability}}$$

Due to the earlier discussed compound symmetry assumption, the following procedural sequence is recommended for the test of factor A: (1) Do the usual F-test, even though it regularly rejects  $H_0$  too often. (2a) If  $H_0$  is not rejected, stop. (2b) If  $H_0$  is rejected, use the Geisser and Greenhouse (1958) conservative F-test. The degrees of freedom for the F-critical-value are adjusted to be “1” and “ $n-1$ ” in this model. (3a) If  $H_0$  is rejected, stop, as this is an indication that both tests have reached the same conclusion. (3b) If  $H_0$  is not rejected a further test needs to be used, in order to break the tie. This is an adjusted F-test, with the adjustment being known as Box’s (1954b) correction (aka the Huyn & Feldt procedure). The numerator degrees of freedom are “ $(J-1)\epsilon$ ”, and the denominator degrees of freedom are “ $(J-1)(n-1)\epsilon$ ”. The “ $\epsilon$ ” here is a correction factor (thus not the same “ $\epsilon$ ” that refers to residuals).

If there are more than two levels of the repeated factor A, and the  $H_0$  (for that repeated factor) gets rejected, it may be of interest which means are different from one another. This can be assessed by using multiple comparison procedures (MCP). Different MCP’s are outlined in chapter 2, and most can be used for an ORMA.

However, a violation of the compound symmetry assumption severely affects MCP’s. Thus there are two alternatives if this is the case. The first is to use a spare error term for each contrast tested (instead of using the same error term, namely MSSA). The second alternative is to use multiple dependent t-tests, in which the  $\alpha$ -level is adjusted similarly to the way it is adjusted in the Bonferonni procedure.

There are several alternatives to the ORMA model. Most notable is the Friedman test, which is a nonparametric procedure based on ranks (much like the Kruskal-Wallis test), but can also be used in a repeated measures model. The Friedman test is conducted as follows:

- Scores are ranked within subject (if there are 6 levels of factor A, the scores for each subject are ranked from 1 to 6).
- From these ranks a mean ranking for each level of factor A can be computed.
- Then  $H_0$  is a test of whether the mean rankings for the levels of A are equal.
- The test statistic ( $\chi^2$ ) is compared to the critical value of  $\alpha\chi^2_{J-1, n}$ . If the test statistic exceeds the critical value, the  $H_0$  will be rejected.

This test does hold the problem of the test statistic not being precisely distributed as  $\chi^2$  when the n or J is small (specifically when either is below 6). Thus it is important to consult the table of critical values in Marascuilo and McSweeney (1997, Table A-22, p. 521).

Just like the Kruskal-Wallis test, the Friedman test assumes that the shape and variability of the population distributions is the same, and that the dependent measure is continuous.

For the Friedman test multiple MCP's can be used. A multiple-matched-pair Wilcoxon tests in a Bonferonni form are best in the case of a planned pairwise comparison.