Common Errors in Interpretation of Correlation, Causation, and Association in Research

VEDANTH NANDIVADA

"Don't confuse correlation with causation. Almost all great records eventually dwindle."

– Charlie Munger

During research and data analysis, several questions arise regarding the relationships between variables. For example: In what way are the two variables related? Are they dependent on each other? Is there a cause-and-effect relationship? It is easy to misinterpret the relationships between variables from an experiment (Bewick et al., 2003). For example, consider the terms correlation, causation and association; they refer to the nature of relationships between variables. While correlation between two variables can imply association, it may not always lead to a causal effect of an independent variable on the dependent variable.

Though statistical formulae are objective, their interpretation is often subjective, so understanding the significance of relationships between variables is foundational for building statistical skills and communicating results in a scientific manner. In recent years, scholars have found that many research experiments could not be replicated, thereby questioning the credibility of research (Hope et al., 2021). Roughly half of all research in the natural and social science fields were considered false as they could not be replicated. This is referred to as

Keywords: Statistics, subjectivity, objectivity, association, cause-and-effect, errors, misinterpretation.

the ‘replication crisis’ (Loannidis, 2005). This is partly due to inappropriate and inaccurate use of statistics in research. There is significant research on various types of errors and misinterpretations in using statistical techniques in research, and the urgent need for improving statistical training to address this crisis (Makin, Orban de Xivry, 2019).

This article summarizes the differences between *causation, correlation* and *association* between variables, common pitfalls in using the terminology through real world examples, and the statistical techniques to be used in future research.

Let us take a scenario where everyone in Section A of 6th grade played football for two hours a day and got 85% in their Math examination. All students in Section B of the same grade played football for half an hour each day, and everyone got 55% in the same exam. Soon, a student in Section B found out Section A’s secret to getting good grades; he excitedly told his classmates, “Guys, we must play football for more time and we will get better grades in our exam!” In the next exam, Section A again got 85%, while Section B got 30%, despite playing football for an increased number of hours. The students scratched their heads, wondering how the secret formula had failed them.

The above example illustrates the common misunderstandings between correlation and causation. Just because two events occurred together, it cannot be said that one happened because of the other.

Definitions

It is important to understand the definitions of key statistical terms used in research before delving into the common errors in interpretation.

- An **independent variable** is the quantity that is changed or manipulated in a research experiment. Its changes in value do not depend on other variables in the experiment.
- The **dependent variable** is the quantity that is measured and the value generally depends on an independent variable.
- **Discrete** data is data that can only take a finite or countable set of values. Examples of this type of data are the number of children in a family: 0, 1, 2, 3, 4, ….
- **Continuous** data is data that can take an infinite number of values over a continuum. Examples include the height and weight of individuals.
- **Linear correlation** indicates the extent of *linear relationship* between two or more variables. The direction of the relationship between two variables is captured by the terms ‘positive’ or ‘negative’ that are attached to the word ‘correlation’. If there is a correlation, the pattern of correlation between two variables can be seen in a scatterplot. The *r*-value, often referred to as the Pearson correlation coefficient, measures the strength of the correlation and ranges between –1 and 1. The closer the value is to 1 or –1, the higher the strength of the correlation.
- **Association** indicates that two variables are dependent on one another. The terms association and dependence are used interchangeably to convey the message that a change in the independent variable *influences* a change in the dependent variable.
- **Causation** is a phenomenon where a change in a dependent variable is the *result* of a change in an independent variable.
- **Prediction** is the ability to predict (‘guess’) the value of the dependent variable for a given value of the independent variable given that an association has been confirmed between the variables. A regression model is used for the prediction.
- **A regression model** is a statistical technique to estimate a relationship between a dependent variable and one or more independent variables. In this paper, we will focus on linear regression models to determine the relationship between two variables. Linear regression models follow the equation $y = \beta_0 + \beta_1 x$, where y is the dependent variable, x is the independent variable, β_0 is the value of the dependent variable when x is equal to 0,

and β_1 is the change in the dependent variable for every unit change in the independent variable. Regression models are developed and enhanced by training and testing. During the training phase, the model uses the given x and y values to calibrate the β_0 and β_1 values in the model. During the testing phase, test values of x are inputted to the model and values of y are predicted by the model. These values are compared by the true value of y in the testing data to assess model accuracy. In this paper, when the term 'beta value' is mentioned, we refer to the β_1 coefficient. The accuracy of the fit of the model and hence the degree of association is determined using a regression model by looking at its r^2 value. This ranges between 0 and 1, with an r^2 closer to 1 depicting a higher degree of accuracy of the prediction model.

- **The null hypothesis** suggests that there is **no** statistical relationship between two variables in an experiment. This hypothesis is assumed to be true until statistical analysis suggests otherwise. For example, in a science experiment to assess the effect of increasing concentrations of Vitamin C on plant shoot growth, the null hypothesis would be "Increasing Vitamin C concentration does not have a significant effect on plant shoot growth."
- **The alternate hypothesis** suggests that there **is** a statistical relationship between two variables in an experiment. This hypothesis is the opposite of the null hypothesis and is considered valid only when the null hypothesis has been discarded. Continuing the same example as above, in an experiment to assess the effect of increasing concentrations of Vitamin C on plant shoot growth, the alternate hypothesis would be "Increasing Vitamin C concentration has a significant effect on plant shoot growth."
- **The p-value** is a number that measures the evidence against the null hypothesis. More precisely, it tells us the probability of obtaining a result that is as bad as the result observed, assuming that the null hypothesis is true. For example, consider an experiment designed to assess the effect of increasing concentrations of Vitamin C on plant shoot growth. As noted above, the null hypothesis would be "Increasing Vitamin C concentration does not have a significant effect on plant shoot growth." Suppose we obtain a p-value of 0.03. As this is less than 0.05, it means that the observed result is unlikely to have occurred by chance alone (assuming the null hypothesis). Therefore, we reject the null hypothesis in such a case.

Let us walk through three important common errors in correlation, association and causation.

Error 1 - "Correlation always implies causation"

Linear correlation answers the following question: Is there a linear relationship between two or more variables? Take for instance, a study that showed a significant correlation between yearly chocolate consumption and the number of Nobel Laureates per country ($r = 0.79$). This finding has led to the suggestion that an increased chocolate intake leads to an increase in the number of Nobel Laureates due to the cognitive effect of cocoa flavanols (Mourage et al., 2013). This incorrect assumption is predominantly due to the misunderstanding of the terms, 'correlation' and 'causation.' Correlation by itself does not have enough proof to infer causation. A strong correlation could simply be due to a sampling error or a random chance coincidence. Had a different sample been chosen, there is a possibility that a different r value could have been obtained, leading to a different conclusion. Hence, a high correlation does not always imply that a change in one variable truly *causes* the change in the other.

Avoiding the error

1. *Research Question:* Are the two variables related?
2. *Statistical Technique:* If the variables are continuous, correlation analysis can be done using Pearson's or Spearman's coefficient.

3. *Statistical Analysis:* While analyzing the results, focus needs to be on the relationship between the variables, direction of the relationship, and strength of the relationship. If the *r* value is positive, there is a positive correlation. If the *r* value is close to 1, the correlation is strong.
4. *Recommended Terminology:* Key terms to be used in the interpretation of the correlation analysis are "correlated" or "related." Terms that should not be used are "caused by. . ." or "influenced by…." For example, if you are studying a correlation between height and weight, you could summarize as "Height and weight are positively correlated." Avoid the use of language such as "Weight seems to be caused by the height of the individual." Ensure that the terminology does not refer to that of causation if no further analysis is performed beyond correlation (Makin, Orban de Xivry, 2019).

Error 2 - "Association always implies causation"

Association conveys that there is a dependency between an independent variable and a dependent variable. Significance of the association between two variables can be determined by reviewing the *p*-value of the beta coefficients in a regression model. A positive beta coefficient depicts positive association whereas a negative beta coefficient depicts a negative association between variables. For example, several studies in recent decades have found an association between root canal treatment and protection against cardiovascular diseases. Nevertheless, there is not enough proof to deduce a causal relationship between the two variables (Jiménez-Sánchez et al., 2020).

Association by itself does not provide sufficient proof to infer causation. To infer causation, association needs to be backed up by consistency and specificity where the association is replicable in different studies, thus reducing its variability (Hill, 1965). Hence, a strong association does not necessarily imply that there is a causal relationship between two variables.

Avoiding the error

1. *Research Question:* Are the two variables dependent on each other?
2. *Statistical Technique:* Linear regression analysis can be undertaken for understanding the dependency or association between a dependent variable and one or more independent variables. Variables can be discrete or continuous.
3. *Statistical Analysis:* While analyzing the results, focus needs to be on the significance of the influence of one variable over the other by verifying that the p-value is less than 0.05 given that the confidence interval is set at 95%. This indicates that for 95% of the experiments, the result falls under the alternate hypothesis.
4. *Recommended Terminology:* Key terms to be used in the interpretation of regression analysis are "factors were influenced" or "factors were dependent" or "factors were associated." Terms that should not be used are "factor caused" or "causal analysis." For example, if you are studying the influence of parental income on nutrition of children using regression analysis, you could say "It was observed that income influences children's nutrition" or "Income and nutrition were found to be associated." Avoid the use of language such as "Malnutrition seems to be caused by the parental income of the students." (Nandivada, Gurtoo, in communication.)

Error 3 - "Prediction is the same as causation"

Regression models have two predominant purposes which are often confused with one another. One is prediction, and the other is causation. Prediction involves deriving a formula based on the observed independent and dependent variable values in a training set in a study. Using the training set and the regression model, dependent variables can be predicted by inputting new values of the independent variable in the prediction model.

Prediction does not always imply causation. Causation can also be determined using a regression model by understanding whether an independent variable causes the change in the dependent variable (Allison, 1999). One important requirement to determine causation is to conduct randomized controlled experiments. This involves identifying two homogenous groups and treating one group as a control group and the other as a treatment group to compare the effect of treatment using various statistical tests (Gianicolo et al., 2020). To determine causation, there are multiple important steps which focus on ensuring both qualitative and quantitative proof that a change in the dependent variable is caused by a change in the independent variable in addition to determining an associative effect (Hill, 1965).

While a regression model can be used to predict a dependent variable, causation cannot be implied till additional statistical methods and analysis are accompanied.

Avoiding the error

1. *Research Question:* Can the relationship between two variables be predicted? OR: Is there a cause-and-effect relationship between variables?
2. Statistical Technique:

- *Prediction:* Regression analysis can be used to build a predictive model using a training data set. Variables can be discrete or continuous.
- *Causation:* Use of control study and establishing two treatment groups will help in collecting the data for building a causation. Regression analysis and other probabilistic models can be developed to determine the causation if controlled randomized experiments are conducted to test for causality of a treatment.

3. Statistical Analysis:

- *Prediction:* While analyzing the results for prediction, focus needs to be on the prediction model and the factors that could influence the accuracy of the prediction of dependent variables. To increase the r^2 value of the model, one can use moderately increased training data to train the model and moderately reduced testing data.
- *Causation:* To determine causation, it is important to ensure homogeneity across groups before subjecting one group to a treatment. Regression analysis and additional statistical tests can be used for confirming the causation.

4. Recommended Terminology:

- *Prediction:* Key terms to be used in prediction analysis are "model predicts the factors influencing…." or "factors were dependent…" or "factors were associated…" Terms that should not be used are "factor caused" or "causal analysis." If you are building only a prediction model based on regression analysis for education and income, you could summarize this as "Income changes are dependent on education" or "Income changes can be predicted based on education." Avoid the use of language such as "Education causes income changes" unless it is observed through causal analysis.
- *Causation:* If controlled experiments are conducted on two homogenous groups and causation is proved through statistical tests, then the term 'causation' can be used.

Conclusion

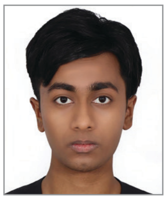
Though statistical formulae are objective, the interpretation is often subjective. Therefore, the interpretation of relationships between variables is foundational for building statistical skills and communicating results scientifically. Additionally, accurate understanding of statistical techniques is important to produce research that can be replicated. In this paper we have summarized three common statistical errors in research (there are other types of errors, of course, but they are much less common). We hope further attempts will be made by the research community to improve the understanding of common errors in interpreting statistical techniques in research and by the public at large.

Acknowledgements

The author would like to thank Dr.Thomas Britz and an anonymous referee of Parabola for their feedback on an earlier version of this paper. This greatly improved the focus and clarity of the content of this article.

Bibliography

1. Bewick, Viv, Liz Cheek, and Jonathan Ball. "Statistics review 7: Correlation and regression." *Critical care* 7. 6 (2003): 1-9.
2. Hope, David, Avril Dewar, and Christopher Hay. "Is there a replication crisis in medical education research?" *Academic Medicine* 96. 7 (2021): 958-963.
3. Ioannidis, John P. A. "Why most published research findings are false." *PLoS medicine* 2. 8 (2005): e124.
4. Makin, Tamar R., and Jean-Jacques Orban de Xivry. "Ten common statistical mistakes to watch out for when writing or reviewing a manuscript." Elife 8 (2019).
5. Maurage, Pierre, Alexandre Heeren, and Mauro Pesenti. "Does chocolate consumption really boost Nobel award chances? The peril of over-interpreting correlations in health studies." The Journal of Nutrition 143. 6 (2013): 931-933.
6. Jiménez-Sánchez, María Carmen, et al. "Cardiovascular diseases and apical periodontitis: association not always implies causality." Medicina Oral, Patología Oral y Cirugía Bucal 25. 5 (2020):
7. Hill, Austin Bradford. "The environment and disease: association or causation?" (1965): 295-300.
8. Nandivada, Vedanth and Gurtoo, Anjula. "Socio-economic factor association and effect size analysis of malnutrition in South Indian Government School children." Journal of Emerging Investigators, In communication
9. Allison, Paul D. *Multiple regression: A primer*. Pine Forge Press, 1999.
10. Gianicolo, Emilio AL, et al. "Methods for Evaluating Causality in Observational Studies: Part 27 of a Series on Evaluation of Scientific Publications" *Deutsches Ärzteblatt International* 117. 7 (2020): 101.



VEDANTH NANDIVADA is a grade 12 student who is passionate about Math and Statistics. He published multiple research articles in high school journals such as 'Parabola', at The University of Sydney, and 'The Journal of Emerging Investigators'. He believes that there is a need to bridge the gap in the foundational knowledge and understanding of statistical techniques among high school students across the world. He enjoys reading about Bayesian statistics and hopes to pursue a career involving data and its applications. He may be contacted at ved.nandivada@gmail.com