Luis Ignacio Pulido

Identifying sources of bias in statistical models

The notion of Causal Inference

In professional domains such as pure data science, statistics, and academic environments with a strong emphasis on research, encompassing fields from Economics to medicine, the concept of causal inference stands out as a prominent and contemporary approach. Whether engaged in developing machine learning models, assessing the effectiveness of monetary policies, or investigating the impact of a novel medicine on a group of individuals, understanding and mitigating bias is essential. This enables analysts to establish meaningful connections between the exposure variables and the observed outcomes.

This paper aims to elaborate on these crucial concepts, providing a comprehensive understanding of the emergence of bias in statistical models. The exploration will delve into both the mathematical intricacies and a qualitative perspective. The focus will specifically revolve around explanatory research questions, where the goal is to identify the potential impact of a given variable on a specific target, thus establishing causality. This contrasts with the predictive approach, currently prevalent, where the objective is to construct a model that accurately forecasts outcomes by evaluating its predictive power on a test set—observations the model has not encountered before. In the context of explanatory research, the paper will concentrate on the intricacies of causality. It will shed light on how one can assess the introduction or removal of information into the model to strike a balance between system complexity and the variability to which predictions are subject.

The importance of causal inference lies in its ability to provide not only predictive models but, essentially, offer an understanding of the underlying relationships between variables. By establishing causality, practitioners can inform decision-making processes, guide interventions, and contribute to advancements in diverse fields.

Explanatory Research Design

The exploration of challenges in establishing direct causality in various scenarios is simplified by addressing the fundamental problem of causal inference, commonly known, and particularly pertinent to our study, as the omitted variable bias problem. While randomized assignment serves to maintain an average even distribution of confounding variables across treatment and control groups, the practical feasibility of such randomization is not always attainable, leading to persistent issues of bias.

To illustrate, consider the scenario in which we aim to assess the causal effect of an educational program on a group of students by observing outcomes following its implementation. The attribution of any observed changes to the program lies on the availability of a counterfactual scenario—namely, the hypothetical performance of these students in the absence of the program. Consequently, there is a need to isolate the effect of the educational program by controlling for external factors.

For instance, assume the program demonstrates greater efficacy among students engaged in mathematics and statistics, compared to a group focused on history and languages due to the program's tailored emphasis on certain activities. Furthermore, take that the program is conducted after school hours, limiting attendance for students with additional responsibilities. If performance is measured exclusively for this subgroup, an inherent bias is introduced into the coefficient accounting for the impact of the educational program on student accomplishment. The qualitative rationale behind this bias lies in the partial exclusion of students with restricted time availability from participating in the learning initiative. Consequently, the measured performance of students who have undergone the course is opposed with those who have not, creating a disparity influenced by these conditions.

In statistical terms, this introduces bias into our coefficient estimation, as the comparison fails to account for the systematic differences in characteristics between the two groups:

$$cov(Performance,\ afterschool_{responsibilities}) \neq 0 \ (1)$$

and

$$cov(\mathrm{Pr}\,o\,gram_{attendance},\ afterschool_{responsibilities}) \neq 0 \ (2)$$

The following example will help us get a grip on what we mean by coefficient bias in a real-world scenario. We are diving into the impact of social media on mental health, focusing on individuals under 30. Let´s depart from a basic look and then ramp it up to a more complex model, considering all kinds of factors that might disturb our results. Our hypothesis will require us to assume that the addressed theory on how the constant employment of social networks results in unpleasant mental outcomes is true.

One should consider the expected implications of such a theory in order to hypothesise what we could observe in the real world. For instance, a proxy to consider the degree of mental health harm could be to measure restlessness or anxiety levels among young people with a contrasting amounts of social media daily use. We will be focusing on an explanatory research design such that the purpose of our analysis will be to establish the link between social media operation and (negative) mental side effects. In other words, the centre of attention is to verify the causal effect of one variable on the other, which means that we will be able to make some conjecture about the expected value of an individual´s anxiety level conditional to being aware of their use of social networks. Our simplest model (with the exclusion of any bias-contemplation):

$$Anxiety_i = \beta_0 + \beta_1 Social_{media_i} + u_i (3)$$

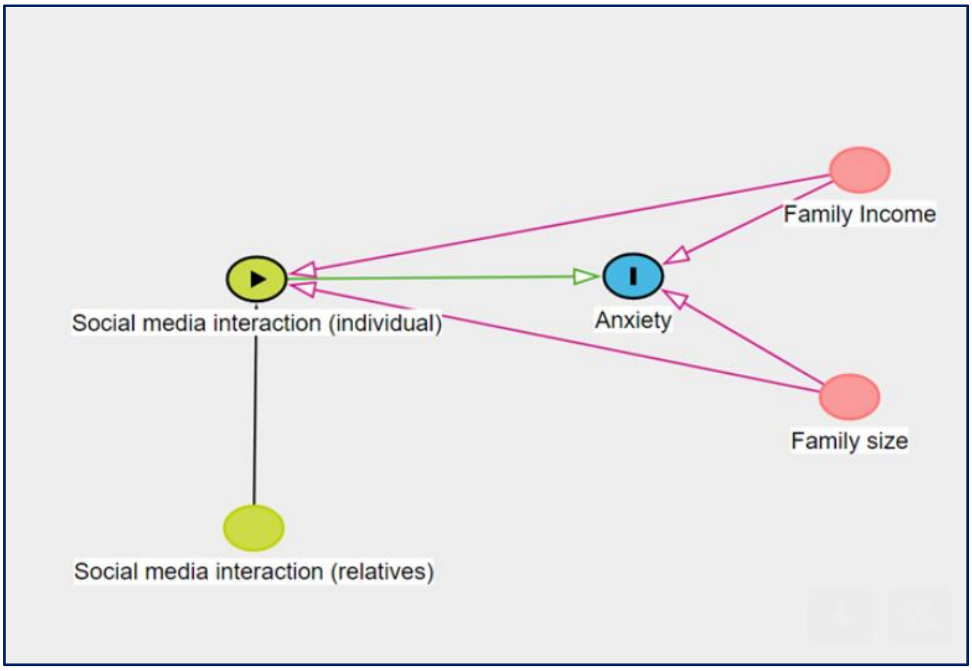
One potentially interesting observation in equation (3) would be to account for the number of members in the family of a person. One could declare that individuals with brothers/sisters or many cohabitants have less time to make use of social media (and vice versa) while at the same time being less prone to falling into anxious/depressive disorders due to having more people to rely on or a more appropriate intimate surrounding. It follows that we would be comparing the response of mental health to social media use between two non-comparable groups, leading to a positive bias in our measurement such that we would be overestimating the true causal effect.

A recent investigation conducted by a Washington D.C research centre examined the existing relationship between the number of social media that an average individual employs subject to its salary income for several years. The outcome reveals the positive and significant correlation between these two measures, supplying us with another source of bias for our current research design. One could claim that wealthier families can afford psychological aid and treatment to their children whenever required, which suggests that, if we weren´t to hold constant these socioeconomic variables, we would be recording anxiety levels among individuals which are not comparable.

Assume we examined two individuals, one of them reporting a relatively low daily use on social media whereas the second one admits using many of these platforms and as a frequent routine and recorded their anxiety levels. Imagine that the first individual belongs to a much wealthier family than our second person and that we observed that the difference among these levels is not remarkable. The existing gap between these levels might have been narrowed by the fact that the first individual could have had straight access to some kind of treatment, which means that, opposite to the bias direction depicted for the number of family members, we would be underestimating the causal effect. We will discuss over/under estimation of the coefficients in the following section.

Let us illustrate a Directed Acyclic Graph (DAG), a graphical representation that captures the directional relationships among variables in a model without forming cycles. DAGs, also known as causal diagrams, not only depict interacting variables but also reveal the direction of their causal influence. The initial step in this endeavour involves identifying all paths through which social media impacts mental health. Tracing these paths, represented by arrows in any direction, we may account for variables such as family income and family size to reach the dependent variable. Social media interactions from relatives influence our own social media usage, possibly due to contagion effects or the fear of missing out, but do not necessarily imply a direct effect on anxiety levels.

Paths where the arrows point inward towards the treatment variable are termed "backdoors." It is imperative to "close" these backdoors to prevent bias in our analysis. Closing a path involves employing statistical methods to account for the variation introduced by variables along the path. In our case, this calls for incorporating the relevant attribute into the regression model. A path is considered closed if at least one variable along the path exhibits no variation. It is crucial not to control for "channel" variables in our model, as doing so might eliminate the variation in the variable acting as the link between our treatment and the target. This could obscure any potential partial effects, impeding our ability to discern the primary focus of our inquiry: the causal effect.


Family Income
Social media interaction (individual)
Anxiety
Family size
Social media interaction (relatives)

Bias vs Variance

Suppose that the true population model is depicted by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \qquad (4)$$

yet, we were to estimate the following incomplete model:

$$\dot{y} = \dot{\beta}_0 + \dot{\beta}_1 x_1 + u \text{ (5)}$$

Let´s replace the true value of y in the following expression accounting for the fact that have already described its behaviour in equation (4). Hence, the estimation of the coefficient (arising from the Simple Linear Model) becomes:

$$E(\dot{\beta}) = E\left(\frac{cov(x_1,y)}{Var(x_1)}\right) = E\left(\frac{cov(x_1, \beta_0+\beta_1x_1+\beta_2x_2+u)}{Var(x_1)}\right) \text{ (6)}$$

Hence, and considering that the expected value between x_1 and u is null because $E(x_1, u) = E(x_1)E(u) = E(x_1) \times 0$, we are left with:

$$E\left(\frac{\beta_1 Var(x_1) + \beta_2 cov(x_1,x_2)}{Var(x_1)}\right) = E\left(\beta_1 + \frac{\beta_2 cov(x_1,x_2)}{Var(x_1)}\right) \text{(7)}$$

So, we obtain that the expected value of the coefficient of the restricted model is:

$$E(\dot{\beta}_1) = \beta_1 + \frac{\beta_2 cov(x_1,x_2)}{Var(x_1)} \text{ (8)}$$

Expression (8) points to the fact that the expected value of the coefficient is influenced by a term proportionate to two pivotal components in any regression analysis. The initial component pertains to the magnitude of the marginal effect of the omitted regressor on the target outcome. The second term is reminiscent of a fundamental aspect in linear regression familiar to those versed in the discipline, encapsulating the coefficient of the regression of the incumbent explanatory attribute against the omitted one.

The bias introduced in our model due to the exclusion of a pertinent attribute is directly proportional to the product of the correlation between the variables within the model and those outside, as well as the correlation between the latter and the dependent variable. Taking into consideration the hypothesized scenario involving family size, where we introduced a negative association between family size and anxiety patterns and assumed that individuals from larger families exhibit reduced inclination towards social media usage (implying negative covariance in statistical terms), the product of these measures suggests a positive bias (negative times negative yields a positive result). The intuition behind these statements is like the ones introduced in expressions (1) and (2).

This outcome is intuitively sensible. Consider smaller families more predisposed to using social media, yet concurrently likely to report poorer mental health compared to larger families. This additional factor functions as an "extra" handicap, leading to a potential overestimation of the impact of social media usage on emotional well-being and a deviation from the true causal effect. Consequently, this underscores the imperative inclusion of such variables in our model to preclude bias and accurately capture the genuine causal relationships.

Let´s devote a section to elucidate the implications for the variance of the coefficient estimation when introducing a new regressor, considering the presence of bias, as previously discussed. For the sake of illustration, we will continue with the example of family size. Assuming that family size significantly contributes to explaining the variation in our target variable, it is plausible to observe a reduction in the magnitude of the residuals. By incorporating additional terms such as family size, we effectively diminish the error term within our model. This reduction in the error term allows for an enhancement in the R-squared (fitting) of the model, indicating an improved explanatory power. Our specific focus will be directed towards examining the variance of the coefficient associated with social media usage – the variable representing our exposure of interest. The variance of the coefficient is depicted by the following expression:

$$var(\beta) = \frac{\sigma_u^2}{n\sigma_x^2\left(1-R_j^2\right)} \quad (9)$$

Expression (9) encapsulates a nuanced trade-off between bias and variance as we input additional information into the model. The act of introducing more variables is akin to reducing the numerator, primarily addressing bias (reduction of population noise). However, there's a twist – the newly introduced variable is correlated with the existing regressor, a necessary condition to bias. The R-squared term in the denominator, crucial for understanding the model's fit, responds by expanding. Consequently, the entire equation experiences an elevation. This dynamic highlights the complex dance between cutting bias and cranking up variance as we try to develop our statistical models.

Alternative Computational Methods

In addition to the methodologies explored in previous sections, we shall now delve into supplementary approaches for establishing causality that extend beyond the scope of our discussion. The initial method involves an examination of the direct mechanism within the system. If we acknowledge a scientifically validated or proven relationship between our explanatory variable and the outcome, we can assert the existence of a causal effect irrespective of our statistical analyses. While this method may appear less conventional, its credibility is rooted in the fundamental understanding that certain causal relationships can be substantiated through scientific inquiry. This approach offers a robust means of accounting for marginal effects, albeit in a manner that may be less quantifiable. Consider the example of smoking and its impact on health. If biological studies unequivocally demonstrate, through an understanding of cellular and physiological mechanisms, that smoking is detrimental to health, the link is established independent of statistical analyses.

An alternative method employed to attribute causality involves taking advantage of instrumental variables. These variables serve as a mechanism to identify an exogenous source of variation within the system, enabling the isolation of the "clean part" of the regressor. The term "clean" denotes the variation component of the exposure variable devoid of confounding effects stemming from external factors. To achieve this, an instrumental variable is sought, characterized by a non-null correlation with the explanatory variable in the model and a minimal correlation with the error term (representing external factors impacting the dependent variable). The instrumental variable acts as a conduit through which the marginal effect on the target variable is exclusively transmitted via the exposure variable.

Various computational methods have been devised for the seamless integration of instrumental variables into the model, automatically yielding the desired results. One of the most renowned techniques for this purpose is the Two-Stage Least Squares (2SLS) method.

Another method employed to establish causality, particularly in the context of dynamic systems where bias avoidance is crucial, is the Difference-in-Differences (Diff-in-Diff) model. The foundational premise involves the division of the study population into a control group, unaffected by the event, and a treatment group, subject to the event. By examining the differences in outcomes between these two groups while accounting for the slopes, the Diff-in-Diff model seeks to discount the biasing or confounding effects originating from non-treatment factors. This approach isolates the marginal effect of the exposure by considering the control group as a reference, thereby enhancing the reliability of causal inferences within the dynamic system under investigation.

Diff in diff is a suitable technique especially when we are dealing with panel (temporal) data where a randomized controlled trial (RCT) is not feasible or ethical, and researchers want to estimate the causal effect of an intervention. The latter is commonly used in observational studies when there is a clear distinction between an exposed group and a control group, provided both groups experience changes in the dependent variable over time. The main assumption, besides the one established above, is the fact that, had not been the intervention, the observed differences in our control and treatment groups would be null (parallel trends) such that we are actually dealing with a proper counterfactual.

References:

- Huntington-Klein, N. (n.d.). Chapter 8 - Causal Paths and Closing Back Doors | The Effect. In theeffectbook.net. from https://theeffectbook.net/ch-CausalPaths.html

- Jeffrey M. Wooldridge. (2012). Introductory Econometrics: A Modern Approach

- https://www.dagitty.net/dags.html

- Libretexts. (n.d.). 4.8: Expected Value and Covariance Matrices. Statistics LibreTexts. https://stats.libretexts.org/Bookshelves/Probability_Theory/Probability_Mathematical_Statistics_and_Stochastic_Processes_(Siegrist)/04%3A_Expected_Value/4.08%3A_Expected_Value_and_Covariance_Matrices