Cleaning Up Confounding: Accounting for Endogeneity Using Instrumental Variables and Two-Stage Models

LORENZ GRAF-VLACHY, TU Dortmund University, Dortmund, Germany and University of Stuttgart, Institute of Software Engineering, Stuttgart, Germany
STEFAN WAGNER, Technical University of Munich, TUM School of Computation, Information and Technology, Heilbronn, Germany

Studies in empirical software engineering are often most useful if they make causal claims because this allows practitioners to identify how they can purposefully influence (rather than only predict) outcomes of interest. Unfortunately, many non-experimental studies suffer from potential endogeneity, for example, through omitted confounding variables, which precludes claims of causality. In this conceptual tutorial, we aim to transfer the proven solution of instrumental variables and two-stage models as a means to account for endogeneity from econometrics to the field of empirical software engineering. To this end, we discuss causality and causal inference, provide a definition of endogeneity, explain its causes, and lay out the conceptual idea behind instrumental variable approaches and two-stage models. We also provide an extensive illustration with simulated data and a brief illustration with real data to demonstrate the approach, offering Stata and R code to allow researchers to replicate our analyses and apply the techniques to their own research projects. We close with concrete recommendations and a guide for researchers on how to deal with endogeneity.

CCS Concepts: • **Software and its engineering**; • **General and reference** → **Empirical studies**;

Additional Key Words and Phrases: Regression, endogeneity, confounder, two-stage least squares, 2SLS, instrumental variables

ACM Reference format:
Lorenz Graf-Vlachy and Stefan Wagner. 2024. Cleaning Up Confounding: Accounting for Endogeneity Using Instrumental Variables and Two-Stage Models. *ACM Trans. Softw. Eng. Methodol.* 33, 8, Article 199 (November 2024), 31 pages.
https://doi.org/10.1145/3674730

1 Introduction

Imagine a researcher who discovers a bug in a software program she uses. The software is critical to her next grant application, so she would like the bug to be fixed as soon as possible. The researcher begins to draft a bug report and starts to wonder how polite the language in the bug report should be. Would a more or a less politely written bug report get the bug fixed faster?

Authors' Contact Information: Lorenz Graf-Vlachy (Corresponding author), TU Dortmund University, Dortmund, Germany and University of Stuttgart, Institute of Software Engineering, Stuttgart, Germany; e-mail: lorenz.graf-vlachy@tu-dortmund.de; Stefan Wagner, Technical University of Munich, TUM School of Computation, Information and Technology, Heilbronn, Germany; e-mail: stefan.wagner@tum.de.

ACM 1557-7392/2024/11-ART199
https://doi.org/10.1145/3674730

To make a particularly informed choice, the researcher decides to perform an empirical study on this topic herself. She thus begins to collect data from software repositories on the politeness of the language of bug reports (e.g., as other software engineering researchers have done on posts on Stack Exchange [26, 92] or Jira issues [27]) and the corresponding bug-fix time (which other software engineering researchers have studied, for example, in prediction models [1, 11, 93, 99]).

And in fact, the researcher runs a simple regression analysis and discovers a statistically significant positive correlation between politeness and bug-fix time. She thus finds an empirical association between greater politeness in bug reports and a longer time until a bug gets fixed. She is puzzled, though. Can it really be that politeness *increases* bug-fix time?

Then, she begins to wonder if there might potentially be a confounding variable missing from the analysis that could explain the relationship. It would be conceivable, for example, that bugs that are particularly hard to fix may, on average, come with a more polite bug report than simple bugs because a bug reporter himself might be unsure about the precise nature of the bug, for instance because it is hard to reproduce and describe. If this were the case, econometricians would explain that the researcher's analysis suffered from endogeneity, i.e., the independent variable (politeness in the bug report) was correlated with the error term in the regression model.

While the researcher's initial findings are thus interesting, and they allow her to *predict* bug-fix time based on the politeness of an observed bug report, she cannot credibly claim that there is a causal link between the two variables. This, however, would be of particular interest because it would allow the particularly useful insight that users can favorably *influence* how quickly bugs get fixed by writing bug reports in a less polite way.

Scientific studies in empirical software engineering often face such challenges. Researchers might find correlations between different variables, but it is frequently not clear whether such findings represent causal relationships or if they are artifacts of endogeneity.

However, there are ways to deal with endogeneity that are also applicable to empirical software engineering. Specifically, one popular technique that has been developed in econometrics and that has found widespread acceptance across scientific disciplines is the use of instrumental variables in two-stage regression models.

With this tutorial article, we aim to acquaint empirical software engineering researchers with this technique and its application to empirical software engineering research, guide the reader through illustrative analyses, and provide recommendations regarding the use of instrumental variables and two-stage models. In the article itself, we shall remain on a conceptual and statistical level, but we discuss specific commands and packages used in the statistics software suites Stata and R in an online supplement, which is available at https://doi.org/10.6084/m9.figshare.22315174.

2 The Challenge of Causal Inference

Some research in empirical software engineering is explicitly purely descriptive in nature (e.g., [53]). Other research is concerned only with predicting outcomes based on some inputs (e.g., [99]). However, in many cases, what researchers in empirical software engineering are really interested in—and what would often be most useful—are statements about causality. In the following, we will therefore first discuss the nature of causality and how to think about it, then discuss experiments as an ideal way of establishing causality, and finally consider widely used quasi-causal approaches to causal inference using observational data.

2.1 Causality and Counterfactuals

The nature of causality is a thorny issue that has been treated at length by various philosophers and scientists [69]. In fact, as Heckman puts it, causality is "a very intuitive notion that is difficult to

make precise without lapsing into tautology" [47, p. 1]. Antonakis et al., therefore, build on others before them [51, 52, 59] and define a causal relationship between two variables X and Y as the fulfillment of three conditions [5]:

(1) X precedes Y temporally.
(2) As X changes, Y changes as well, and this relationship is statistically significant (i.e., it is not due to chance).
(3) No other causes explain the relationship between X and Y.

The first condition is rather intuitive and unproblematic. If X occurs only after Y, it would be hard to argue that X could have caused Y, after all (although this becomes more challenging if X and Y mutually cause each other, as we will briefly discuss in Section 3.2.3). The second condition is similarly unproblematic and essentially simply requires a researcher interested in causality to have empirical data that allow for statistical analyses. The third condition is the greatest challenge in identifying causality in practice (which we will discuss further in Section 3.2.4).

Especially related to this third condition, one key perspective on causality that has found widespread acceptance is the idea of understanding causality through "counterfactuals." As Lewis argues, "We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects—some of them, at least, and usually all—would have been absent as well" [56, p. 557]. This perspective is formalized in the potential outcomes framework or the Neyman–Rubin model [77, 82] and has found renewed attention and further development recently [69], including in software engineering research [81].

At a simple level, counterfactuals are essentially the answer to the question: "What would have happened to Y if X had been different?" Naturally, to really know this, one would need to observe the same unit once "treated," i.e., subject to a change in the independent variable X, and once untreated. For instance, in our introductory example, we might want to see the effects of the same bug report once written in a polite way, and once written impolitely. In real-world data, however, a unit is at one point in time either treated or untreated and can thus only be observed in one of the two states. In other words, the counterfactual cannot be observed. Researchers are therefore confronted with what Holland calls the "fundamental problem of causal inference" [51, p. 947].[1]

Before we proceed in our discussion, it is worth clarifying *why* establishing causality between two variables is important in the first place, when having correlations may seem potentially sufficient. It is in fact true that correlations may be useful in many cases. In all situations in which a researcher is merely interested in making predictions about a system without interfering with it, a causal understanding of relationships between variables is not necessary and statistical associations, i.e., correlations, are sufficient [68]. If, for instance, we want to predict the likelihood of failure of a system from, for example, its complexity, then we may be able to perform a simple regression analysis using historic data on systems of the same type for which we know both the complexity and have failure information. Critically, however, this is only the case if we do not wish to *intervene* in any way. In many practical applications, however, the goal of the researcher is to enable targeted interventions to favorably influence some outcome. In this case, the causal relationships between the variables of interest must be understood. Similarly, social scientists are typically fundamentally interested in causal relationships because they help *explain* the world. In other words, while correlations are interesting and useful for predictions, causal relationships are the

[1]Pearl takes a very strict definition of counterfactuals, arguing that counterfactuals entail a retrospective element and that thus even randomized experiments alone cannot provide full causal information [70]. We take a somewhat looser view of counterfactuals here.

building blocks of scientific theories that help to understand the world and allow to manipulate it in a purposeful way.

2.2 Experimental Methods as the "Gold Standard" of Causal Inference

Experiments are typically considered the "gold standard" when it comes to inferring causality [5, p. 1088][81, p. 2]. Ideally, experiments allow the researcher to manipulate one independent variable of interest (the "treatment"), control for all other potentially confounding factors, and observe potential changes in a dependent variable of interest. If a change in the dependent variable is observed, the researcher knows that it must have been *caused* by the treatment, i.e., the change in the independent variable.

To do this, however, the researcher must actually control for all potentially confounding factors. The best way to do so is through randomization. In this context, randomization means that the researcher randomly assigns the units (e.g., patients in a clinical trial) either to a group that receives the treatment (the "treatment group") and to another group that does not (the "control group"). In this way, assuming a sufficiently large number of units, the researcher can be almost certain that, on average, the units in the treatment group do not differ from the units in the control group. In particular, randomization has the convenient property that it controls for *all* possible confounding factors, irrespective of whether the researcher is even aware of them.

Since the two groups are essentially identical before the treatment group receives the treatment, this allows the researcher to tackle the fundamental problem of causal inference by considering the control group as the counterfactual of the treatment group. The control group tells the researcher what would have happened to the treatment group had the researcher not administered the treatment.

Alas, in many situations, performing experiments is impossible, impractical, or otherwise not satisfactory. For instance, some experiments cannot be conducted for ethical reasons, like inducing a mental disorder in programmers to study its effects. Others cannot be performed due to resource constraints, like hiring experienced professional programmers to perform extensive programming tasks.

Further, experiments are often performed only in the laboratory because randomization is often not possible in field settings. For instance, it is hardly conceivable that a study to address the question of the relationship between politeness in bug reports and bug-fix time would do so in a perfect field experiment. Such a large-scale field study would require finding a large number of genuine bugs (in one or multiple projects) and then submitting bug reports that differ randomly and *only* in the degree of politeness to observe whether more or less polite ones get addressed more quickly.

Unfortunately, while experiments performed in a laboratory are excellent for establishing internal validity, they frequently suffer from potentially poor external validity, i.e., any findings might not generalize to the world outside of the laboratory, for example, because programmers behave differently under intense observation or because their real-world surroundings differ from the lab in important ways [18].

2.3 Quasi-Causal Approaches to Inferring Causality from Observational Data

Despite the conceptual superiority of experiments, researchers—as the one in our introductory example—therefore frequently have to resort to other data sources, such as surveys or software repositories, where they face a particular challenge when it comes to inferring causality. Specifically, in such observational data, there is no counterfactual—and it is much harder to create a credible one than it is to do so in randomized experiments.

Luckily, emerging from scholarship on the potential outcomes framework [77, 82], researchers have developed several quasi-causal approaches that (while making some critical assumptions) allow us to perform causal inference from data that does not come from randomized experiments. Four particularly relevant ones are **regression discontinuity design (RDD)**, **difference-in-differences (DiD)** approaches, **propensity score matching (PSM)**, and—the focus of this article—two-stage models using instrumental variables [5, 54].

First, RDD identifies the causal effect of a treatment variable by assigning a cutoff point on a continuous assignment variable to differentiate between units who receive the treatment and those who do not. It exploits the idea that units just above and below the cutoff are similar in all respects except for the receipt of the treatment. By comparing the outcomes of units just above and below this cutoff, RDD isolates the treatment's effect from other factors, assuming that the only systematic difference between the groups is the treatment itself. This design is particularly useful when random assignment is not feasible, providing a credible estimate of the treatment's impact. Despite its broader applicability, this design has been used in software engineering mostly to study the effects of changes in a system or its environment, for instance, the introduction of a new technology or a software engineering method, over time, resting on the assumption that units before and after the change are similar except for the fact that they did or did not experience the change [19, 40, 60, 94, 101, 102].

Second, DiD is a statistical technique used to estimate the causal effect of a treatment by comparing the changes in outcomes over time between a group that is exposed to the treatment and a group that is not [10]. The key assumption of DiD is that, in the absence of the treatment, the average change in outcomes for both groups would have been the same. By calculating the difference in outcome changes between the treatment and control groups, DiD controls for unobserved factors that are constant over time, isolating the effect of the intervention. The technique has already found application in software engineering [31, 58].

Third, PSM is a technique used to estimate the effect of a treatment by accounting for the covariates that predict receiving the treatment [75]. It involves calculating a propensity score for each unit in the study, which is the probability of being assigned to the treatment group given their observed characteristics. Units in the treatment and control groups are then matched based on their propensity scores, creating pairs or groups of unit with similar characteristics except for the treatment status. This matching process aims to mimic randomization, reducing selection bias and enabling a more accurate estimation of the treatment's causal effect on the outcome by comparing outcomes between matched pairs or groups. Like the other techniques, this technique has already been applied in empirical software engineering research [91].

Elements of the individual methods can also be combined. PSM methods, for instance, can be used within DiD designs to construct treatment and control groups that are likely similar [31].

A fourth approach to causal inference—and the one that is arguably most widely used across scientific disciplines—is the use of regular regression analysis paired with statistical corrections through the use of two-stage models with instrumental variables. Such corrections are necessary because of the potential for endogeneity in regression analyses.

3 The Problem of Endogeneity in Regression Models

3.1 Endogeneity in Regression Models

Endogeneity is unfortunately hard to describe precisely and intuitively at the same time. We thus begin by providing a mathematical definition and will provide intuition later. Consider the example for a regression model shown in Equation (1). Y denotes the dependent variable (i.e., an effect to be explained—bug-fix time in our introductory example), X denotes an independent variable

(or "regressor") (which we may hypothesize to be a cause of Y—the politeness of a bug report in our introductory example), and ε denotes the error (or "disturbance") term. β_0 and β_1 are regression coefficients, with β_0 representing the intercept and β_1 the slope of the regression line that is to be fitted to a sample of data from an overall population about which one would want to make causal claims.

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

In a regression analysis, an estimator selects values for all coefficients β_i such that a particular objective is fulfilled in a sample of data. In the most common type of linear regression, i.e., "**ordinary least squares (OLS)** regression," the objective is to minimize the squared deviations (or "residuals") of the fitted regression line from the observed sample data. In the following, we will focus on OLS as a point of reference due to its relative simplicity and its importance in practice. We acknowledge that there are, of course, other estimators, for example, for dependent variables with non-normal distributions like binary or count data.

Notably, OLS operates on the assumption that there is a linear data-generating process in the real world that is specified by what are called the true population parameters. In the thought world of regression analysis, this data-generating process can be thought of as being a "true line" that expresses the true relation of the dependent variable to the independent variables in the entire population (not merely what is observed in the researcher's sample). OLS assumes that every time a data point is generated, it is taken from a normal distribution that centers on this "true line." Consequently, any individual data point the researcher may observe will likely not be perfectly on that "true line," but exhibit some difference from it. Such differences between the observed sample of data points and the "true line" are captured in the error term ε.

Importantly, since the researcher cannot observe the "true line"—it is, after all, precisely what she is trying to approximate with the regression line fitted on the observed sample—she can also not know the error term ε. Note that the phrase "error term" may be considered a bit of a misnomer because every regression analysis will have an error term. There being an error term does not signal that something went wrong or that a mistake was made, it just means that the regression line is not a perfect approximation of the "true line."

Figure 1 illustrates this. The top of the figure shows the data-generating process, i.e., the "true line," defined by the true *population parameters* β_0 and β_1 and shown as a dashed line. From the data generated by the process, the researcher samples six data points, as shown in the diagram. Every data point is a little bit off the "true" line, creating an error. At the bottom, the researcher uses the data points to estimate a regression line, described by the corresponding *estimated* coefficients β_0 and β_1, and shown as a solid line. Again, each individual data point is a little off this regression line, which constitutes a residual. As the estimated line and the "true" line are very likely to not be identical, errors and residuals are not the same.

In the context of a regression analysis like that shown in Equation (1), the above has an important implication. Namely, any variance in Y that is not explicitly accounted for (i.e., that is not in X) is being captured in the error term ε. In other words, everything about Y that is not explained by X becomes part of the unobservable ε. Or, in yet other words, any causes of Y that are not X become part of ε. Consider again our introductory example. If there are any causes of bug-fix time that go beyond the politeness of the bug report (which would appear quite likely), then these would all be part of the error term.

As stated by the Gauss–Markov theorem, OLS is the best linear unbiased estimator when a variety of assumptions hold [97]. One key assumption is the so-called exogeneity condition, i.e., the assumption that the independent variable is exogenous, meaning it is not correlated with the error term. This is to say that ε has an expected value of zero for any X, or $\mathbb{E}(\varepsilon|X) = 0$ [97].

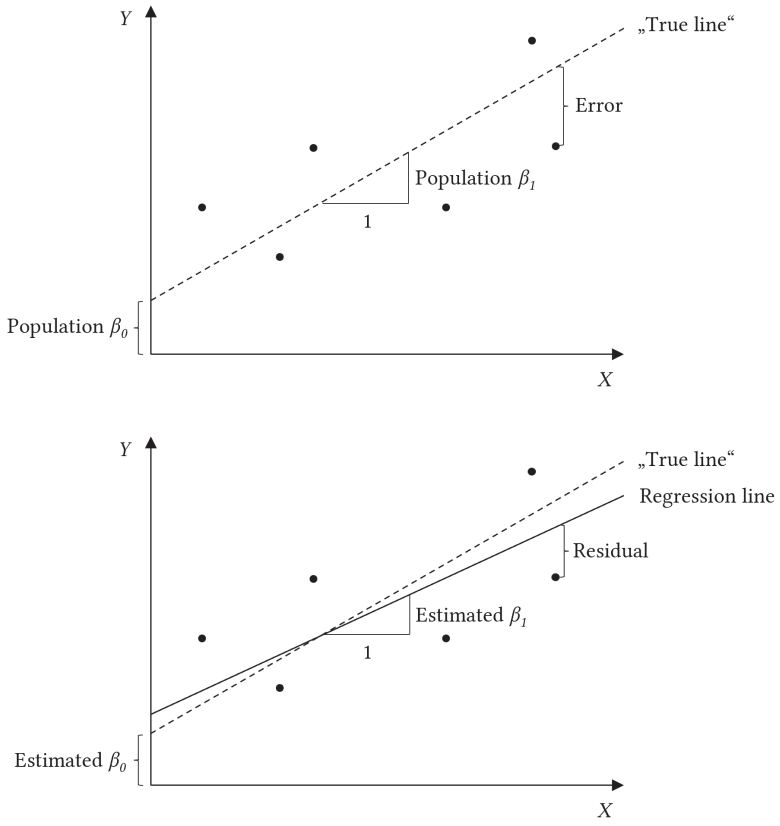

Fig. 1. Unobservable errors versus observable residuals.

Endogeneity is defined as the violation of this assumption, i.e., a situation in which the expected value of the error is dependent on X. In other words, endogeneity exists when *systematic* information from the regressor X is captured in the error term and there is thus a non-zero covariance between X and the errors, i.e., $cov(X, \varepsilon) \neq 0$. This situation can be described by saying that "X is endogenous." In the context of our introductory example, this means that if any causes of bug-fix time beyond the politeness of the bug report have entered the error term, and if any of these causes are correlated with the politeness of the bug report, then the politeness of the bug report is endogenous.

Endogeneity is a problem because it renders the OLS estimator inconsistent. This means that, even for very large data samples, the estimated coefficients do not converge on the true population parameters, and the regression line, therefore, does not become an approximation of the "true line." In other words, the estimated coefficient β_1 is not trustworthy.

A key concern about endogeneity that is critical to keep in mind is that it cannot be tested for without making additional assumptions. Since the true population parameters are unknown (all a researcher knows is the part of the population she observes, i.e., her sample), *it is impossible to know the error term* ε, and thus impossible to assess its covariance with X. Importantly, the residuals cannot be used as a proxy for the errors, because when the OLS estimator creates the regression line (and thus determines the residuals), it *assumes* the absence of endogeneity, and the correlation between X and the residuals is therefore always zero by construction.

3.2 Causes of Endogeneity in Regression Models

At least four causes of endogeneity exist [6, 7, 50], namely measurement error, selection, simultaneity, and omitted confounding variables. They are not mutually exclusive and can accumulate in a research project.[2] All can, at least to some degree, be addressed with instrumental variables.

3.2.1 Measurement Error. One cause of endogeneity are errors in variables or measurement error [7]. Many measurements in software engineering exhibit errors since researchers in the field are often dealing with latent variables, i.e., they cannot directly measure the true phenomenon of interest, but they have to measure something that represents it. Such measurement error may cause endogeneity when a regressor is measured with error, and this error is correlated with the regressor itself.[3] This could be, for example, the case if a measure for programming skill is very reliable for highly skilled programmers but is systematically less reliable for assessing lower levels of skill. In our introductory example, it would be a problem if the reliability of the politeness measurement varies with different levels of politeness. Another example could be found in studies on the effect of programmers' health on performance. Programmers with poor performance might (consciously or unconsciously) misreport their health to have a legitimate explanation for their poor performance [17].

3.2.2 Selection. There exist two primary types of selection that can cause endogeneity [21]. In both, certain variables are only observable when other variables take on particular values, because inclusion in the sample is contingent on those variables [96]. For one, there is sample selection, i.e., situations in which a researcher does not observe a random sample of the population. For instance, a researcher may only be able to observe software repositories of projects that were successful enough so they were not shut down and the repositories deleted. Related to our introductory example, the researcher may only observe bug reports of a certain range of politeness in case there is a moderation system implemented that removes comments that are extremely impolite or outright hostile. For another, there may be self-selection of individuals or organizations. If a researcher, for instance, observes programmers contributing to open-source projects, she cannot readily generalize to the entire population of programmers because those who contribute may be systematically different (e.g., in terms of skills or personal values) from those who do not. This is particularly critical to studies of the effectiveness of tools or techniques in software engineering "in the wild." Programmers and managers may use certain techniques precisely because they (rightly or wrongly) believe that these techniques will be effective [20, 43]. If a researcher, for instance, were to study the performance effects of agile development methods, she would only be able to observe teams that self-selected into the use of such techniques. Because it is possible that this decision is influenced, for example, by the prior experience of team members, this means that the researcher cannot readily make inferences about the true effect of the technique. Importantly, including prior experience as an additional independent variable does not solve this issue because more experienced teams may be systematically more or less likely to select into the sample. Notably, non-response in surveys also represents a case of selection.

[2] Some authors list additional threats to validity related to endogeneity such as inconsistent inference (e.g., the use of non-robust standard errors) or model misspecifications [5]. These threats to validity have a more technical rather than conceptual link to endogeneity and are thus not treated here.

[3] Measurement error can be viewed as a special case of an omitted variable problem (see Section 3.2.4) because if the measurement error itself could be included in the regression equation of interest, the endogeneity problem would be resolved. Under certain assumptions, it can be possible to use one error-ridden measurement of a variable as an instrumental variable for another error-ridden measurement of the same variable [7, endnote 5].

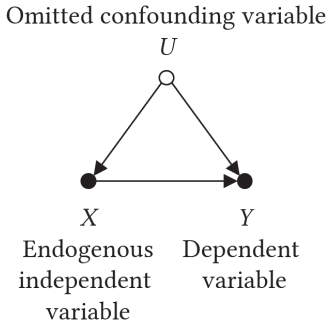

Fig. 2. Problematic omitted confounding variable.

3.2.3 Simultaneity. The third cause of endogeneity is simultaneity, sometimes also referred to as simultaneous causality. Simultaneity denotes the situation in which X causes Y, but Y also causes X at the same time, thus leading both to be determined simultaneously. Consider, for instance, the relationship between the number of users on a platform like Stack Exchange and the quality of the content of such a platform. The more users frequent the platform, the higher the chance that some will leave questions and provide answers. At the same time, the more questions and answers are posted, the more useful the platform becomes and the more likely users are to visit it. In archival data, it is very hard to determine whether one caused the other, or the other way around. Consequently, and irrespective of which variable is considered the independent and the dependent variable in a regression, there will be endogeneity [96].

3.2.4 Omitted Confounding Variables. In many important cases, however, endogeneity is the result of confounding variables that are omitted from the model. Omitting a variable is sometimes a conscious choice (when one erroneously believes that the variable is irrelevant to the model), sometimes simply an oversight (when a relevant variable is forgotten), and sometimes impossible to avoid (when a variable cannot be measured).

Figure 2 is a directed acyclic graph visualizing the causal structural relationships between variables in such a situation. Black nodes represent observed variables, whereas white nodes represent unobserved variables. X and U are two variables that influence Y, as indicated by the arrows between them. If U is omitted from the regression model (e.g., because it was not observed), the shared variance between U and Y is not accounted for and thus enters the error term ε. Because U is also correlated with X, this means that ε will now also be correlated with X, making X endogenous. This will make estimations of β_1 in Equation (1) inconsistent.

The introductory example about politeness and bug-fix time illustrated this problem. As more complex bugs may be described using more polite language, one might suspect the complexity of the bug to drive the results. Specifically, if more complex bugs take longer to fix and are systematically reported with greater politeness in the bug reports, "bug complexity" would be an omitted confounding variable in the model. This is because it is not accounted for in the model, and thus enters the error term ε. Because bug complexity is correlated with bug-fix time, the error term ε is now also correlated with the dependent variable, creating endogeneity. For a purely mathematical illustration of how omitted variables cause endogeneity, we refer the reader to more technical literature [5].

Figure 3 shows the same situation using intuition about the variances of the variables, which are represented by the circles. The researcher is interested in identifying the shared variance between X and Y because it represents the causal effect of X on Y. Since U, however, shares variance with both

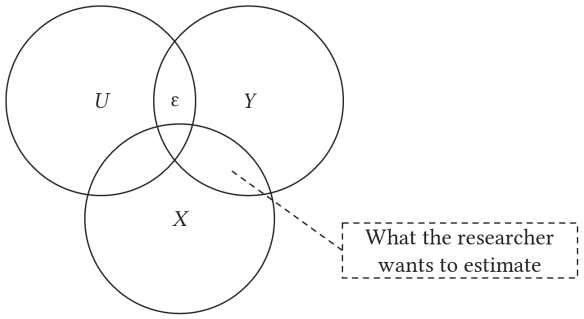

Fig. 3. Shared variances with omitted confounding variable.

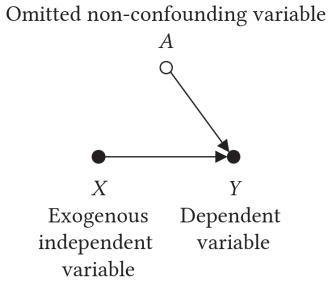

Fig. 4. Unproblematic omitted non-confounding variable.

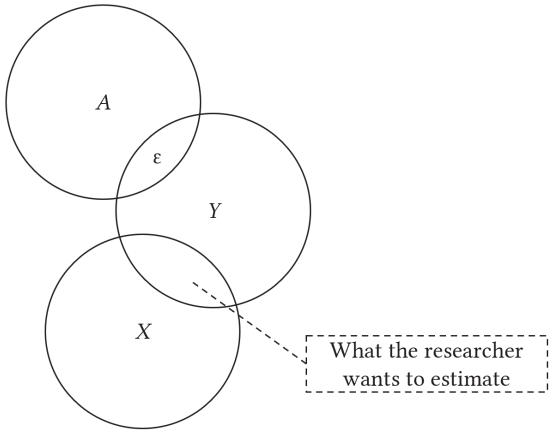

Fig. 5. Shared variances with omitted non-confounding variable.

X and Y, the researcher's estimate will not only include the true component of shared variance, but it will also include an error component in the form of a part of ε that is also related to X, causing endogeneity and tainting the estimate.

Note that not every variable that is not included in a model causes endogeneity. Figure 4 visualizes such a situation. The omitted variable A is not confounding because it is not related to X. In this case, there is no shared variance between A and X that ends up in the error term ε (see Figure 5), leaving X exogenous, and not creating the problem of endogeneity. In our introductory example,

the number of developers working on the module in which the bug appears could be such a variable. While it may be systematically related to bug-fix time, there is no reason to assume that it would be related to the politeness of a corresponding bug report.

4 Consideration of Endogeneity in Empirical Software Engineering

Many empirical software engineering researchers are aware of potential problems for causal claims through omitted variables. The "threats to validity" sections of papers frequently discuss potentially "confounding factors" or "confounders" (e.g., [33]). Consequently, authors often choose their wording and the claims they make very carefully. Some simply avoid mentioning causality at all, for instance referring to "correlations" instead (e.g., [33]), and others explicitly highlight that they merely study associations between variables and do not make claims of causality (e.g., [72]).

But even where causal relationships are quite clearly the primary interest of the researcher performing regression analyses, the statistical notion of endogeneity itself is not widely discussed in the empirical software engineering literature. Further, in the few cases where it is, and where instrumental variables approaches are employed to address it, it is often hard to judge if these approaches were implemented correctly. A complete ACM Digital Library search of the *ACM Transactions on Software Engineering and Methodology* and the *IEEE Transactions on Software Engineering* and the proceedings of the three CORE A*-ranked software engineering conferences *ASE*, *FSE*, and *ICSE* (including co-located events) yielded only four articles for the keyword "endogeneity." Of these articles, one claims absence of endogeneity based on a Durbin–Wu–Hausmann test [74], while another one claims to find endogeneity and uses an instrumental variable approach [73]. Disconcertingly, neither reports the instrumental variable used or any test results. The third article also claims to address potential endogeneity using instrumental variables but does again not disclose them [22]. Only a single article performs an instrumental variable regression while disclosing the instruments and reporting some information about test statistics (albeit without making any theoretical argument for the exogeneity of the used instruments, instead relying on empirical clues that can necessarily only be indicative at best) [8].

Generalizing the search string to "endog*" yields 187 articles, the vast majority of which do not actually contain any such phrase in the full text at all. In almost all other cases, the term is used in different contexts, for instance in discussions of "endogenous variables" in structural equations models, where the term corresponds roughly to what users of regression models would call dependent variables, or in discussions of "endogenous model transformations," i.e., situations in model-driven engineering where source and target models belong to the same meta model. Only in four previously unconsidered cases was a mention of endogeneity related to causality. One case [90], however, centered on Granger causality tests, which do not actually test causality but only the ability of one time series of data to forecast another [39, 51], and two cases deal with defect repair or fairness improvement in neural networks [87, 100]. The last case [38] uses a type of multi-stage regression model to address simultaneity but not omitted confounding variables, and the authors do again not justify the employed instruments.

The picture is similar with regard to the specific remedy for endogeneity this article focuses on. Searches in the aforementioned outlets for "instrumental variable*," "two-stage least squares," or the common abbreviation "2SLS" yielded only a single relevant result the prior searches had not already uncovered [2]. In this work, a two-stage model was again used to resolve simultaneity, yet without any explanation or justification of instruments. Further searches for "two-stage model*" and "two-stage regression*" yielded no additional relevant articles.

5 Remedies to Implement Before Using Instrumental Variables

Before researchers resort to instrumental variable approaches, they should consider other measures. This is because the use of instrumental variable approaches is not free of challenges, as we will discuss later.

Focusing on endogeneity caused by omitted variable bias, there are two helpful oft-used approaches. The first is to identify, measure, and include (i.e., control for) potentially relevant omitted variables. Returning to our introductory example, we might, for instance, wonder if the length of a bug report might not also be a confounding omitted variable. Longer reports may be a characteristic of bugs that are particularly hard to reproduce and describe (potentially associated with higher politeness) and that may also be very hard to fix (potentially associated with longer bug-fix time). Because of its potential to introduce endogeneity into a regression analysis, the researcher should thus control for the length of the bug report.

Similarly, academic researchers like Raghuraman et al., who study whether using the unified modeling language is associated with lower defect proneness in software projects, explicitly control "for known confounding factors" and suggest that, because of this, their "results should be relatively robust" [72, p. 103] (although they notably do not claim to have demonstrated causality). A study by Papadakis et al. shows the importance and effectiveness of such control variables [67]. These authors take on the question of whether the association between mutant scores and real fault detection in the testing literature holds when controlling for test suite size, and they find that the association disappears once this control variable is accounted for. This suggests that their initial analysis was distorted by endogeneity.

While controlling for relevant variables is always the preferred option because it rules out confounding effects of omitted variables without making any further assumptions [76], it is not always feasible. In particular, an omitted variable may be simply be impossible to measure (at least with sufficient precision).

The second approach only works with what is called "panel data," i.e., data where the same "panel unit" is observed multiple times. This could be, for example, a dataset of programmer behaviors that includes multiple observations for the same programmer working on different tasks or a dataset of software projects that include multiple observations for each projects (e.g., for every release of a project). In such situations, it is advisable to include so-called "fixed effects" for the panel unit. In the programmer example, this would entail including dummy variables in a regression, i.e., including one binary variable for each programmer, with each of these binary variables set to one only for one specific programmer, and to zero for all others. Doing so effectively controls for all stable characteristics of the programmer (i.e., the panel unit) that do not change between tasks (i.e., within the panel unit across the entire panel). Note that using panel fixed effects only works when there is variance in the independent variable of interest within a panel unit. This is because in the absence of such variance, no coefficient for this independent variable can be estimated since any effect of the variable gets absorbed into the coefficients of the dummy variables. These coefficients, however, are hardly interpretable because they capture the effects of *all* characteristics of each panel unit that are constant in the entire panel (e.g., a programmer's personality or intelligence over time).

Since the inclusion of a large number of dummy variables quickly makes regression models computationally costly, there exist specific estimators for fixed-effects models. They are more computationally efficient while yielding the same results for the variables of interest. They derive their computational efficiency from the fact that they "demean" the data, i.e., subtract the mean from each variable within a panel unit, and then perform the estimation without actually estimating coefficients for panel unit dummy variables at all.

Importantly, the use of fixed effects in panel data should be accompanied by the use of standard errors that are clustered at the panel unit. Computing such standard errors accounts for the fact that the standard errors of multiple observations might be correlated because the observations belong to the same panel unit.

6 Two-Stage Models with Instrumental Variables

In many situations, the approaches mentioned above will not be sufficient to address all endogeneity concerns. In such cases, researchers might additionally resort to instrumental variable regressions, which are, according to Turing Award laureate Pearl "the workhorses of causal inference in practice" [71] or, in the words of economics Nobel laureate Angrist and collaborator Pischke, the "most powerful weapon in [the] arsenal" [4, p. 114]. The most common kind of such instrumental variable regression is the "two-stage least squares" approach. The employed instrumental variables are also frequently referred to as "instruments."

6.1 The Two-Stage Model Approach

As the name suggests, the key idea behind two-stage models is to perform coefficient estimations in two separate stages. For example, assume that we want to estimate the following model, which we might take to represent our introductory example, with Y being bug-fix time, X_1 representing politeness, and X_2 being the length of a bug report

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon. \tag{2}$$

Further assume that X_1 is endogenous, i.e., correlated with the error term ε, for instance due to an omitted variable U, and X_2 is exogenous, i.e., uncorrelated with the error term ε. In the first-stage regression, we would regress the endogenous regressor X_1 on all exogenous regressors (only X_2 in our case) and all instrumental variables (in this example, we will only have one instrument, Z_1). We would thus estimate the following model (with ζ as its error term)

$$X_1 = \gamma_0 + \gamma_1 X_2 + \gamma_2 Z_1 + \zeta. \tag{3}$$

This allows us to calculate a predicted version of X_1, which we denote $\hat{X_1}$

$$\hat{X_1} = \hat{\gamma}_0 + \hat{\gamma}_1 X_2 + \hat{\gamma}_2 Z_1. \tag{4}$$

In the second stage, we then estimate the model we are actually interested in (specified in Equation (2)), but replace X_1 with its predicted version $\hat{X_1}$

$$Y = \beta_0 + \beta_1 \hat{X_1} + \beta_2 X_2 + \varepsilon. \tag{5}$$

If we estimate this model, we will obtain consistent estimates for β_1. The intuition behind the approach is that we use information from the exogenous instrumental variable Z_1 to estimate a version of X_1 that has no correlation with the error term anymore and is thus also exogenous. We might say that we use the first stage to "partial out" variance that the endogenous regressor X_1 and the instrument Z_1 share, so that the predicted $\hat{X_1}$ does not include any variance that is shared with the error term in the second stage. Therefore, the second stage is "cleaned up" and does not suffer from endogeneity anymore.

Figure 6 shows the intuition again using shared variances between variables (we omit the exogenous X_2 for simplicity). In the two-stage approach, the estimator relies on $\hat{X_1}$, i.e., the variance that the instrument Z_1 shares with X_1, to estimate the relationship between X_1 and Y. To this end, the estimator can only use a part of this variance, specifically the part that is also overlapping with Y. This way, it only uses a part of the overall overlapping variance between X_1 and Y (making 2SLS less efficient than OLS), but it can now estimate the relationship consistently because it does so

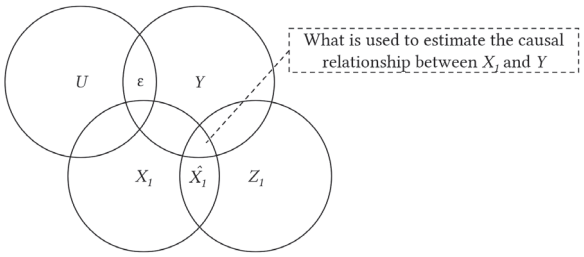

Fig. 6. Shared variances with instrumental variable.

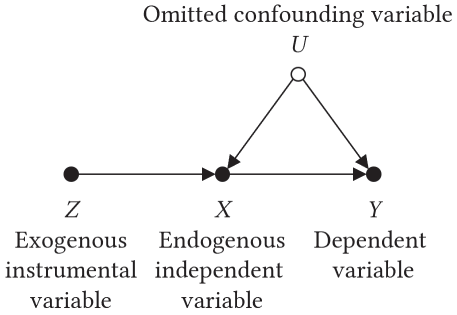

Fig. 7. Requirements for instrumental variables.

without being affected by any shared variance with the omitted variable U. This is to say that the coefficient β_1 will be correctly estimated despite using less information. Remember that $\hat{X}_1$ is the predicted value of X_1 that is due to Z_1, which we obtain from the first-stage regression. The critical characteristic of this value is that it has no overlap with U and is thus exogenous. This is exactly what allows us to use $\hat{X}_1$ as the independent variable in the second stage.

Note that the standard errors (and associated p-values) would not be accurate if we were to actually manually estimate the model this way [4]. Although often done, simply reporting robust standard errors does not address this issue [43]. Instead, the standard errors must be adjusted, either analytically or via bootstrapping. All major statistics suites offer commands to estimate the first stage and the second stage jointly and automatically adjust standard errors. We strongly recommend using such commands (see the online supplement at https://doi.org/10.6084/m9.figshare.22315174).

6.2 Requirements for Instrumental Variables

The critical challenge in instrumental variable approaches is the choice of instruments. For one, it is important to understand that to ensure that an instrumental variable regression can cure the problem of endogeneity, one needs at least one instrument for every endogenous regressor. For another, all instruments must fulfill certain requirements. Specifically, they must be relevant and exogenous.

6.2.1 Relevance. The first requirement for an instrument is that it be relevant, or "strong" [4]. Simply put, this means that the instrument must be clearly related to the endogenous regressor (see left side of Figure 7, where you see the direct influence of Z on X, and Figure 6, which shows how the variance of the instrument overlaps with that of the endogenous variable). Researchers should do two things to assess the relevance of an instrument. First, they should ideally be able to identify a theoretical reason that would support a strong correlation with the endogenous regressor. In case

of our introductory example, we might for instance choose the typical linguistic politeness of a bug reporter's country of origin as an instrument because it is likely to substantially influence how polite—on average—a message from this bug reporter is through cultural imprinting.

Second, and ultimately the decisive criterion, they should empirically establish that there is a correlation between the instrument and the endogenous regressor, conditional on all other (exogenous) control variables [64, 65].

Specifically, researchers should consider at least two aspects when assessing the relevance of instruments empirically. For one, the instruments should be statistically significant predictors in the first-stage model [55]. The signs of their coefficients should also be of the expected direction, i.e., positive if a positive relationship between instrument and independent variable was expected and negative if a negative relationship was expected [55] (also see Section 7, where we demonstrate this using our introductory example). For another, researchers should consider the partial F statistic of the instruments in the first stage to rule out "weak identification." The intuition behind this test statistic is that it represents a joint test of the significance of the instruments. The larger the F statistic, the stronger the instruments. In a way, the F statistic can thus be thought of as a measure of how much variance in the endogenous variables the instruments can explain [80]. The exact threshold values that allow telling if instruments are sufficiently strong differ depending on various factors like, for example, the number of instruments used [83]. Stock and Yogo offer a corresponding table with relevant thresholds [84, p. 100]. As a rule of thumb, and with all the limitations that naturally come with such rules, it is commonly accepted that F statistics should always be greater than 10 for the case of a single endogenous regressor and a single instrument, although higher values are invariably better [83, 86]. Note that there are two different relevant F statistics, depending on whether regular or clustered standard errors are used. For regular standard errors, the Cragg–Donald Wald F statistic is appropriate, while the Kleibergen–Paap rk Wald F statistic is appropriate in case of clustered standard errors.

6.2.2 Exogeneity. The second requirement for an instrument is that it be exogenous [4]. Exogeneity requires that the instrument is not correlated with the error term. This implies that any effect the instrument has on the dependent variable must be through the endogenous regressor, and not through any other paths (see Figure 7, where you can see that Z influences X but the only way in which it affects Y is *through* X, and Figure 6, which shows how the instrument and U do not share any variance). If this is fulfilled, the instrument meets the "exclusion restriction" because the instrument can be said to be excluded from the causal model of interest.

If there are exactly as many instruments as endogenous regressors, the model is said to be "just-identified" or "exactly identified." In this case, a justification of the exogeneity of instruments must rely exclusively on theoretical arguments. This means that there needs to be a justification that cannot rest on statistical analyses at all for why the instrument is exogenous, consider, for instance, the instrument candidate of the typical level of politeness in a bug reporter's country of origin in our introductory example. There is simply no compelling reason for why it should be related to whether a specific bug would be fixed faster or slower, except through the politeness of the bug report for this specific bug.

If there are more instruments than endogenous regressors, a model is "overidentified" and one may perform so-called tests of overidentifying restrictions [4]. The arguably most popular one is the Sargan–Hansen test that uses Sargan's statistic/Hansen's J statistic. This test assesses the exogeneity of all provided instruments simultaneously. Note that this means that in case one or more instruments are not exogenous, the test cannot identify *which* individual instruments are problematic. The intuition behind the test is that it regresses the residuals of a two-stage model on the instruments used in that model. If there is a significant relationship between the residuals and

the instruments, this indicates problematic instruments. A significant test result is thus a warning signal to the researcher and casts doubt on the exogeneity of the instruments. It is important to understand, though, that the Sargan–Hansen test makes the assumption that there is at least one exogenous instrument (which again, cannot be statistically verified but only theoretically argued). If this is not the case, the test may provide misleading results and incorrectly suggest exogenous instruments [61].

When there are two more instruments than endogenous regressors, it may be possible to test the exogeneity of an individual instrument. Specifically, the C statistic (or difference-in-Sargan statistic) can be computed, which is the difference between two Sargan or J statistics [7, 45]. Again, a significant test result indicates endogeneity. Note that this test also operates on the assumption that the other instruments are exogenous.

6.3 Selecting Instrumental Variables

Finding good instruments is essential to conducting successful two-stage analyses. Unfortunately, finding good instruments is not trivial. In fact, it takes "a combination of institutional knowledge and ideas about processes determining the variable of interest" [4, p. 117]. The likely most useful way to think about instruments is to revisit Figure 7 and ask: *Which exogenous variable Z affects X, but only affects Y through X and not through any other causal path?* [4]

6.3.1 Examples for Instrumental Variables. Since finding good instruments is often a challenge, we provide a few examples in the following. These examples stem from disciplines outside software engineering but may give an impression of the tactics and thinking of researchers from fields that have a longer tradition of using instrumental variable regressions.

Finance scholars Bennedsen et al. study the relationship between the number of directors on the boards of small firms (which are often family firms) and firms' financial performance [9]. Since they are concerned about omitted confounding variables that may influence both the size of the board and the performance of a firm, they resort to the use of an instrumental variable. Specifically, they use the number of children of the **Chief Executive Officer (CEO)** as an instrument since it is empirically highly predictive of the number of family members, and thus ultimately the number of directors overall, on a small firm's board. This establishes the relevance of the instrument. At the same time, the authors argue that there is no plausible way in which the fertility of the CEO impacts the firm's performance. This suggests that the instrument is exogenous.

Garretsen et al. study the relationship between the prevalence of the personality trait neuroticism and economic growth in regions and cities across the UK using survey data [36]. Since they were concerned about potential endogeneity, they perform instrumental variable regression to check the robustness of their results from a non-instrumented model. The instrument they choose is traumatic childhood experiences as reported by the surveyed individuals. Even though personality is often deemed largely stable across adult life, childhood experiences can affect personality. Specifically, in the sample of Garretsen et al., it is correlated with neuroticism, making the instrument relevant. Further, the authors argue that such individual-specific experiences should be uncorrelated with the overall economic conditions in the geographic area of residence of each survey participant, showing the instrument's exogeneity.

Marketing scholars Germann et al. study whether firms' adoption of a **Chief Marketing Officer (CMO)** position has a causal effect on firms' financial performance [37]. Since it is conceivable that an omitted variable, i.e., a focal firm's culture, might cause both CMO adoption and firm

[4] Sometimes, researchers use a rule of thumb to identify instruments by seeking out variables that correlate with the endogenous independent variable but not with the dependent variable (e.g., [8]). However, this rule can be deceiving because these correlations can themselves be affected by endogeneity [5].

performance, there is an endogeneity concern. They select the prevalence of CMOs in competitors, i.e., firms from the same industry, as an instrument. This instrument can be expected to be relevant since firms in similar market conditions are likely to behave similarly in terms of CMO adoption. Additionally, this instrument is likely exogenous because there is no apparent link between it and the omitted variable (and thus the error term). Specifically, the potentially omitted variable of a focal firm's culture is unlikely to be observable for competitors, and even if it were, it is unlikely that most or all competitors would derive a joint strategy to react to the focal firm's culture by adopting CMOs.

In a study in strategic management, Gupta et al. examine how a CEO's political ideology influences how they allocate capital within their firm [41]. Again, this relationship might be endogenous because some omitted properties of a firm might influence both what kind of a CEO is appointed and how capital will be allocated. The authors use the political ideology of the electorate of the US state in which the firm is headquartered and that of the predecessor CEO as instrumental variables. They argue and empirically show that both instruments are relevant in that they directly influence the ideology of the person chosen to be CEO. In addition, they argue that the instruments are exogenous because there is no direct effect of either instrument on a firm's capital allocation that does not flow through the current CEO's ideology.

A seemingly exotic but very popular instrument in conflict research is rainfall. Bohlken and Sergenti, for example, study the effect of economic growth on conflict in different states in India [12]. This relationship may be endogenous because conflict may also influence economic growth. The authors thus use rainfall as an instrument that is clearly exogenous, i.e., neither economic growth nor conflict nor any other variable in their model credibly influence rainfall. Additionally, it is hard to see a causal path from rainfall to conflict that does not go through economic growth. Finally, they argue for the relevance of the instrument by positing that economic growth in the relevant regions depends on agriculture, which in turn is affected by rainfall.

Economists Licht et al. study the effects of how individualistic or collectivist a national culture is on how firms are governed across different countries [57]. They employ an instrumental variable design and propose to exploit differences in the use of pronouns between the languages of firms' home countries. Specifically, some languages allow speakers to omit certain pronouns and this possibility to drop pronouns, especially "I," blurs the contrast between person and context. For this reason, one might expect pronoun drop to be more frequent in societies which focus more on the contextualization of individuals (i.e., are more collectivist) rather than the individuals themselves (i.e., are individualistic). The researchers present F statistics for the first stage to demonstrate the instrument's relevance. They then argue for its exogeneity by explaining that "it does not exert an influence on governance other than through culture" [57, p. 673].

Other researchers use yet other instruments for which exogeneity is essentially guaranteed. Economist Angrist, for instance, uses the results of the US draft lottery at the time of the Vietnam War to study the impact of military service on individuals' life-time earnings [3]. Given that the draft lottery is by definition both random, i.e., exogenous, and has a direct connection to serving in the military, i.e., is relevant, he has a credible instrument. This helps allay concerns that the direct relationship between military service and earnings might be endogenous due to the fact that people who voluntarily serve in the military might be systematically different from those who do not, and that this difference may impact earnings.

As is evident from some of these instrumental variable choices, it may be beneficial to consider which instrumental variable one might want to select before a study's data collection begins. Once a theoretical model is developed and hypotheses are formed, it may make sense to identify potential endogeneity concerns, consider possible instruments, and collect data on the instruments together with data collection for the variables of ultimate interest.

6.3.2 The Role of Theory. Methods researchers frequently highlight the role of prior theory in the selection of instrumental variables. This pertains to both the identification of potential instruments and the assessment of their exogeneity [5, 13]. It is easy to see how extant theory can be helpful in identifying potential instruments, as this is essentially about the identification of variables that are likely strongly correlated with the endogenous independent variable. Scientific theory—but also lay theories and mere hunches of an informed researcher or even practitioner—may provide ideas for such variables and thus for potential instruments. There is little risk of error, as the relevance of instruments can be easily tested.

In contrast, it may seem curious to use theory in the justification of and instrument's exogeneity, given that theory (in the social sciences, at least) is typically understood as—ideally empirically corroborated—conceptual relationships between constructs. In the case of selecting and justifying an instrument, however, a researcher needs to plausibly suggest the *absence* of a relationship between variables, specifically between the instrument and the dependent variable (except through the potentially endogenous independent variable). Researchers in this case are thus best advised to review prior theory from the literature to identify if there are any apparent violations of the exogeneity assumption of the instrument. As an example, consider the following statement made by economists Licht et al.: "We find no claims that link [instrument candidate] to [dependent variable] or to any of the factors mentioned as relevant to [dependent variable] in the literature" [57, p. 673]. In other words, it is their review of prior literature that gives them confidence in their instrument. Naturally, researchers should not blindly rely on extant work, but will further need to think for themselves, and might additionally incorporate critical feedback from other researchers into their considerations of instrumental variables.

Ultimately, however, finding instrumental variables is a creative act—with some methods experts even explaining that we must expect good instruments to feel "weird" [24]—and making a decision on an instrument's exogeneity must typically remain a judgment. We concur with econometricians Stock and Watson, who write about the exogeneity assumption that it is "incumbent on both the empirical analyst and the critical reader to use their own understanding of the empirical application to evaluate whether this assumption is reasonable" [85, p. 417].

6.4 Testing for Endogeneity with Instrumental Variables

As laid out in Section 3.1, endogeneity is by definition not directly detectable and thus cannot be tested for. However, if one is certain that one has a strong and exogenous instrument, the Durbin–Wu–Hausman test for endogeneity can be performed [28, 44, 98]. The test assesses whether there is a significant difference between the coefficients of an OLS model and a corresponding 2SLS model with instruments.

If the test result is significant and a significant difference in coefficients thus exists, the researcher may conclude that there is endogeneity in the model (because the correction for endogeneity actually changed the results). Consequently, the researcher should report the results from the two-stage model, because these results are consistent in this case. If there is no significant difference, the researcher should report the OLS results because the OLS estimator is not only consistent in this case but also more efficient, leading to smaller standard errors.

It is worth emphasizing two points. First, recall again that the exogeneity of an instrument cannot be tested for without assuming the exogeneity of another instrument. Second, this test may incorrectly suggest the presence of endogeneity if instruments are used that do not fulfill the requirements detailed in Section 6.2. In this case, however, reporting the two-stage results would be no better than reporting OLS results because the two-stage results would be impacted by the unsuitable instruments. The test should thus only be used when one can reasonably assume the exogeneity of an instrument.

7 Illustration Using Simulated Data

To illustrate how to apply the instrumental variables approach, let us return to our simple and fictitious introductory example. The researcher is interested in studying whether the linguistic politeness of the natural language in bug reports will influence the time it takes until a bug gets fixed. Prior research in software engineering has studied politeness, for example, in Stack Exchange posts [26, 92], Jira issues [27], or GitHub comments [66], or related constructs like incivility in code review discussion [32]. Similarly, there is a host of prior research on bug-fix time (e.g., in the form of prediction models [1, 11, 93, 99]).

For this intentionally simple illustration, we will make several assumptions. First, we assume that the data-generating process, which describes how the true relationships in the overall population give rise to any data that one might sample, is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon. \tag{6}$$

Y is bug-fix time, β_0 is the intercept (or "constant") and takes a value of 1, β_1 is the coefficient for politeness (X_1) and takes a value of -0.2, and β_2 is the coefficient for the length of a bug report (X_2) and takes a value of -0.3. This means that both politeness and length of a bug report are causally linked to reduced bug-fix time.

However, we shall also assume that X_1 is endogenous, i.e., that there is a correlation between politeness and the error term. This endogeneity could, for instance, be driven by a variable that influences both politeness and bug-fix time. A conceivable variable with these properties might be the severity of the bug or the complexity of fixing it. To keep things simple, we will assume that X_2 is exogenous, i.e., the length of the bug report is not related to the error term.

We further consider three variables that might serve as instruments. The first instrument (Z_1) is the linguistic politeness typically exhibited in the poster's country of origin. To obtain such data for different countries, the researcher might study text corpora with English-language contributions of individuals from different countries. She might do so, for example, by applying automated tools for identifying linguistic aspects of politeness [26] to online Q&A websites or forums where the country of origin of a poster is available, such as Stack Exchange [26, 92]. The typical politeness in a reporter's home country must be expected to be—on average—related to how polite a reporter is in communication with others, for instance when submitting a bug report. At the same time, there is no obvious reason why the typical level of politeness in a bug reporter's country of origin would influence whether a specific bug would be fixed faster or slower, except through the politeness of the bug report of this very bug. The second instrument (Z_2) is the politeness of prior bug reports by others. The "tone" that prior bug reporters struck in their reports might influence the focal bug reporter. At the same time, it seems unlikely that other bug reports' politeness would have an effect on how long a focal bug takes to get fixed. The third instrument (Z_3) is the number of prior bug reports by the same reporter. With an increasing number of bug reports, a bug reporter might begin to strike a less polite tone, for example, because they feel more at ease reporting a new bug. This is a doubtful instrument since, of course, the number of prior bug reports could conceivably also have an effect on bug-fix time, for example, if a would-be fixer takes a reported bug more seriously if it comes from a regular contributor.

We follow prior endogeneity researchers and simulate data with the exact relationships between the variables we just described [5]. Thus, we know what the results of a correct analysis *should* be, and we can compare the results of different analytical approaches to this ground truth. A replication package with Stata and R code to simulate the data and perform all analyses, as well as an online supplement discussing the specific Stata and R commands used, is available at https://doi.org/10.6084/m9.figshare.22315174.

Table 1. Simulated Data: Correlations

	Bug-fix time (Y)	Politeness (X_1)	Length (X_2)	Error term (ε)	Politeness of reporter's home country (Z_1)	Politeness of prior reports from others (Z_2)	Num. of reporter's prior reports (Z_3)
Bug-fix time (Y)	1.00						
Politeness (X_1)	0.11^{***}	1.00					
Length (X_2)	-0.28^{***}	-0.00	1.00				
Error term (ε)	0.94^{***}	0.30^{***}	0.01	1.00			
Politeness of reporter's home country (Z_1)	-0.17^{***}	0.68^{***}	0.16^{***}	0.00	1.00		
Politeness of prior reports from others (Z_2)	-0.08^{*}	0.06	0.16^{***}	-0.02	0.43^{***}	1.00	
Num. of reporter's prior reports (Z_3)	0.27^{***}	0.60^{***}	0.05	0.40^{***}	0.19^{***}	0.18^{***}	1.00

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Num., Number.

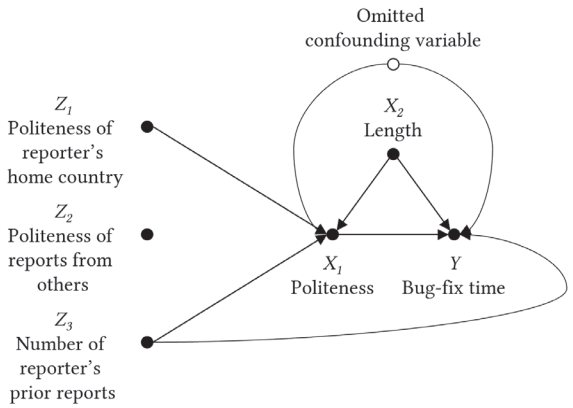

Fig. 8. Relationships between variables (unknown to researcher).

Specifically, we make the described assumptions about the relationships and correlations in the population and draw a sample of 1,000 observations from that population. Table 1 shows the correlations in the sample between the dependent variable, the two independent variables, the error term, and the instruments. Figure 8 illustrates the relationships (that are not known to the researcher). Note that our first instrument has a strong correlation with politeness and none with the error term. It is thus a relevant and exogenous instrument. The second instrument has a weak correlation with politeness and none with the error term. It is thus not relevant, yet exogenous. The third instrument is strongly correlated with politeness, but also has a correlation with the error term. It is thus, although relevant, endogenous and not exogenous.

Say the researcher from our introductory example conjectures that politeness reduces bug-fix time because greater politeness in bug reports might show social consideration and avoid a loss of face [15] on the part of the would-be bug-fixers (who might be the developers who introduced a bug in the first place), and instead bestow social rewards on them [88]. Consequently, politeness might compel would-be bug fixers to prioritize such bug reports. If the researcher could establish this causal link, she would know that users who want their bugs fixed quickly should use more polite language in the bug reports—a potentially very useful insight for herself and others.

The researcher thus collects the following data for 1,000 randomly sampled bugs (the exact observations underlying Table 1): bug-fix time, politeness of the corresponding bug report, and

Table 2. Simulated Data: OLS and 2SLS Models

	OLS	2SLS: Z_1	2SLS: Z_2	2SLS: Z_3
Politeness (X_1)	0.119^{***}	-0.198^{***}	-0.592	0.497^{***}
	(0.032)	(0.048)	(0.653)	(0.056)
Length (X_2)	-0.287^{***}	-0.288^{***}	-0.289^{***}	-0.286^{***}
	(0.031)	(0.032)	(0.038)	(0.033)
Constant	0.961^{***}	0.967^{***}	0.975^{***}	0.953^{***}
	(0.031)	(0.032)	(0.040)	(0.033)
Observations	1,000	1,000	1,000	1,000

Dependent variable: Bug-fix time.
Standard errors in parentheses.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

the length of the bug report. The researcher could run a simple OLS model, obtain the coefficient estimates as reported in the first model in Table 2, and would—based on the results—incorrectly conclude that there is a positive relationship between politeness and bug-fixing time.

If the researcher suspects X_1 to be endogenous, she might collect additional data to use as an instrument. If she uses the first instrument Z_1 (relevant and exogenous), she would recover the actual coefficient for X_1 (which we had set to -0.2 in the parameters of our simulation) almost perfectly, i.e., see that there is in fact a negative effect of politeness on bug-fix time. If she chooses Z_2 (not relevant and endogenous), the estimated coefficient is very different from the true coefficient, and it is insignificant. The researcher would thus incorrectly conclude that there is no statistically significant effect. Finally, if she chooses Z_3 (relevant and endogenous) as an instrument, she would incorrectly conclude that her OLS results are understated, and that there is an even stronger positive relationship between politeness and bug-fixing time. In sum, a model using a relevant and exogenous instrument yields essentially perfect results for the coefficient for X_1, while an OLS model or problematic instruments lead to incorrect conclusions. The coefficients for the unproblematic variable X_2 are close to the true value (which we had set to -0.3 in our simulation) in all cases.

To not distract from the key results, we do not tabulate the results of various further analyses, but they can be reproduced with the code we provide in the online supplement. First, the researcher may assess the relevance of the instruments. Used separately, all three would be significant predictors in a first stage. However, Z_2 can be identified as not relevant because it offers an F value that is below any relevant cut-off value. Second, the researcher may wish to assess the exogeneity of the instruments. Since she already knows that instrument Z_2 is not relevant, she must discard it. She can perform a test for overidentifying restrictions with the remaining instruments Z_1 and Z_3. The Sargan statistic is highly significant, suggesting that at least one of the instruments is not exogenous. As the researcher cannot identify which instrument is problematic from this test, she must rely on theoretical arguments to select the exogenous instrument. When the researcher uses the relevant and exogenous instrument Z_1, she can also demonstrate that her OLS model is affected by endogeneity. Running a Durbin–Wu–Hausman test produces a highly significant result, showing that the estimates of the OLS model and the 2SLS model differ significantly. Note that this test can of course not produce interpretable results for the cases in which the instruments are either not relevant or not exogenous.

8 Illustration Using Real Data

To further show the applicability of the technique, we now demonstrate its use on a real dataset. In contrast to the prior illustration, we will also use this as an opportunity to extend our tutorial by introducing some additional intricacies of instrumental variable regressions. Specifically, we will use a different type of dependent variable, and we will include fixed effects in our analysis.

Our objective shall be to determine if there is a causal relationship between the number of **lines of code (LOC)** in a class and whether there are any defects in it. Note, however, that our focus lies on the application of the technique, and not on a substantial contribution to the literature on bugs, which is why we will keep our analysis deliberately simple again.[5] For this illustration, we use the GitHub Bug DataSet, a large publicly available dataset of bugs collected from more than 100 releases of major Java projects hosted on GitHub [89] (version 1.1). It contains information on bugs and various code metrics at the class level. The replication package with complete Stata code to reformat the dataset and perform all analyses, as well as an online supplement discussing the specific commands used, is available at https://doi.org/10.6084/m9.figshare.22315174.

Prior research found that code metrics like LOC are associated with defects, but these studies were not typically conducted with a focus on causal relationships [23, 25].[6] Where regressions have been used, there was typically no correction for potential endogeneity [42, 62].

To test for a causal link, we will perform a regression with a binary indicator of whether a class is defective (i.e., contains at least one bug) as the dependent variable and the class's LOC as one of multiple independent variable. Since the dependent variable is binary, we cannot use OLS and 2SLS, but instead must use a probit model and a corresponding instrumented version.

As we are concerned that a corresponding regression analysis might be tainted by endogeneity through omitted variables, we first include several potentially relevant control variables to capture complexity, cohesion, inheritance, and coupling. Specifically, we add **cyclomatic complexity (CC)**, **lack of cohesion in methods 5 (LCOM5)**, **depth of inheritance tree (DIT)**, **coupling between object classes (CBOs)**, and **coupling between object classes inverse (CBOI)**. For simplicity, we keep this selection deliberately limited.

In addition, as there might be unobserved systematic differences between different projects and releases, we also include fixed effects at the release level, i.e., a dummy variable for every individual release of every project, and we cluster the standard errors accordingly. Note that since these dummy variables identify each release for every project, they control for *all* time-invariant characteristics of every project and release. Figure 9 shows the assumed causal relationships between all variables, with the control variables grouped together for simplicity.

If we are still concerned that there might be one or more omitted variables that might render LOC endogenous (e.g., unobserved characteristics of the main author of the focal class), we may choose to use instrumental variable regression. To do so, we need an instrument. A potentially viable instrument that is not related to whether a given class is defective (exogeneity) but might predict this class's LOC (relevance) would be the mean LOC for all other classes in a release. The argument for exogeneity of this instrument is that there is no apparent reason for why the defectiveness of a given class should depend on the mean number of LOC in the *other* classes in a project's release (except through the LOC of the focal class). Again, the causal relationships are shown in Figure 9. To test relevance, we consider the sign and significance of the instrument's coefficient in the first stage, as well as the associated F statistic. Unfortunately, the probit instrumental variable estimator

[5] Various authors have highlighted the "noisiness" of bug datasets, for example, due to dormant bugs [30] and otherwise incorrectly labeled files [48, 49]. However, since "there is still noise" [48, p. 42] even in datasets that were specifically designed to fix such issues, and because our aim is not to definitively solve a fundamental question on bugs but merely to illustrate an econometric technique, we intentionally use the small and simple GitHub Bug DataSet.

[6] Some works test for Granger causality (e.g., [23]), which, however, is not causality in the sense of this article [39, 51].

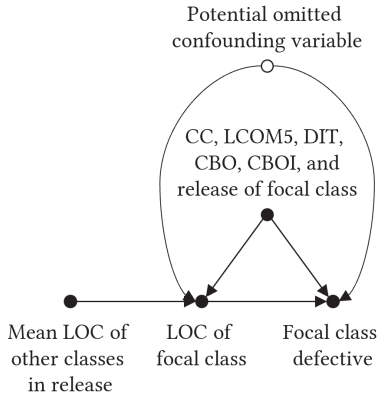

Fig. 9. Relationships between variables (control variables grouped for simplicity).

Table 3. Real Data: Correlations

	Defective	KLOC (log)	CC	LCOM5	DIT	CBO	CBOI
Defective	1.00						
KLOC (log)	0.19^{***}	1.00					
CC	-0.02^{***}	-0.02^{***}	1.00				
LCOM5	0.06^{***}	0.29^{***}	0.00	1.00			
DIT	0.08^{***}	0.07^{***}	-0.01^{***}	0.03^{***}	1.00		
CBO	0.22^{***}	0.54^{***}	0.04^{***}	0.19^{***}	0.25^{***}	1.00	
CBOI	0.09^{***}	0.21^{***}	-0.05^{***}	0.09^{***}	-0.01^{**}	0.13^{***}	1.00

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

we use in this illustration does not produce an F statistic for its first stage. However, the first stage is essentially identical to the one used in a 2SLS model. Hence, we run such a model, and only consider its first stage. The first stage we obtain in this way is shown in Table 4. We can see that the coefficient for the instrument is negative (because if the focal class contains many LOC, all other classes within the same release are likely to contain relatively less, and vice versa) and significant, suggesting a relevant instrument. The 2SLS first stage also provides us with a Kleibergen–Paap rk Wald F value that exceeds the canonical threshold of 10 (remember that since we use clustered standard errors, the appropriate F value is not the much higher Cragg–Donald Wald F).

For our substantive analysis, we divide LOC by 1,000 to make the coefficient more interpretable (yielding KLOC), and take the logarithm to correct for skew. Table 3 reports descriptive statistics. We also compared the correlations obtained between LOC and other variables to those found in prior work. We find that our correlations with LCOM5, DIT, and CBO are extremely similar to those obtained by prior researchers [29]. Table 4 shows the results of a non-instrumented and the instrumental-variable model. The magnitude of the coefficients and their significance are very similar. If we perform a Wald test (which is for probit regression what the Durbin–Wu–Hausman test is for OLS), we do not reject the null hypothesis of no endogeneity.

Overall, our results thus suggest that the independent variable of LOC was not tainted by endogeneity in our analysis. Therefore, we can be confident that there exists indeed a positive causal relationship between a class's LOC and whether or not it is defective.

Table 4. Real Data: Probit and Instrumental Variable Probit Models

	Non-instrumented probit	2SLS first stage	Instrumental variable probit
KLOC (log)	0.241***		0.227***
	(0.017)		(0.025)
CC	−0.272***	−0.153*	−0.273***
	(0.065)	(0.067)	(0.065)
LCOM5	−0.004	0.076***	−0.003
	(0.003)	(0.006)	(0.003)
DIT	0.044**	−0.043*	0.043*
	(0.017)	(0.016)	(0.017)
CBO	0.019***	0.076***	0.020***
	(0.002)	(0.003)	(0.002)
CBOI	0.002***	0.009***	0.002***
	(0.000)	(0.001)	(0.000)
Instrument		−0.653***	
		(0.169)	
Constant	−2.671***	12.331**	−1.560***
	(0.064)	(4.426)	(0.116)
Observations	160,240	160,240	160,240

Dependent variable: Class defective.
Standard errors in parentheses.
Release fixed effects included in all models.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

9 Challenges and Alternatives

9.1 Challenges of Two-Stage Models with Instrumental Variables

Although theoretically appealing and very useful in practice, there are several challenges when using two-stage models. First, instruments must be actually relevant. If they are not relevant, and thus "weak," 2SLS yields unreliable results in that the coefficients will be biased [14, 63, 83]. In fact, particularly in smaller samples, weak instruments lead to estimates that are just as biased as OLS estimates [14]. Importantly, the 2SLS estimator becomes more biased toward the corresponding OLS estimate if there are *many* weak instruments [4]. This means that weak instruments cannot be compensated for by simply more weak instruments [86]. As just-identified 2SLS is approximately median-unbiased, it is generally advisable to select the single best instrument and use only this one instrument [4].

Second, as mentioned above, instruments must be exogenous. If this is not the case, results from 2SLS may be biased even more than OLS results [79].

Third, 2SLS should be used only with larger samples. Whereas OLS is consistent and unbiased (i.e., the estimates are centered on the true population coefficients) even for small samples, 2SLS is consistent but biased (i.e., it needs large samples for the coefficient estimates to converge on the population parameters) [4]. Note that samples are typically considered large for statistical purposes if they exceed 100 observations [85], but the precise bias of 2SLS depends on the number and

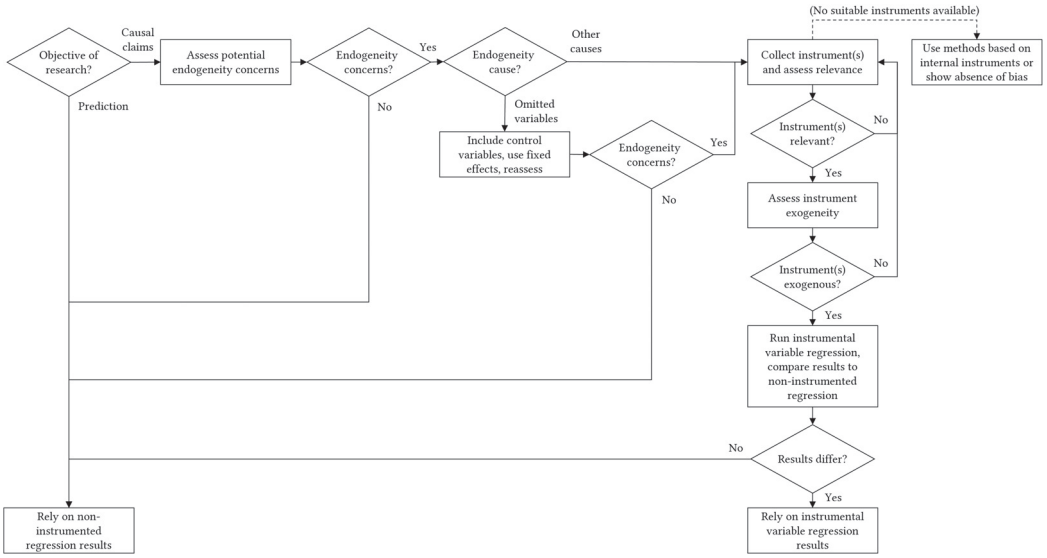

Fig. 10. Guide for how to deal with endogeneity concerns.

strength of instruments used. 2SLS is approximately unbiased in the case that it is just-identified, i.e., that there are exactly as many instruments as endogenous regressors. As the number of instruments increases and as the instruments get weaker, the needed sample size increases [4].

Finally, the discussed 2SLS approach cannot immediately be applied in cases in which endogenous regressors are binary [4]. Instead, a probit 2SLS model is needed [96].

9.2 Alternative Approaches to Resolving Endogeneity Concerns

Researchers may attempt to demonstrate that there is simply likely no bias due to endogeneity. One such way is to perform an "**Impact Threshold of a Confounding Variable" (ITCV)** analysis, which determines how strongly a suspected omitted variable would have to be correlated with the independent and dependent variable to invalidate a focal causal inference [34]. This can be convincing if it can be argued that it is implausible that an omitted variable would exhibit such high correlations. This may be the case, for example, if the required correlation exceeded that of all other known confounders which are already included as control variables in a model. A closely related technique is the "**Robustness of an Inference to Replacement" (RIR)** [35]. Here, the researcher can determine which percentage of observations would have to be replaced with null hypothesis cases to invalidate the focal causal inference.

Depending on the source of the endogeneity concern, researchers may choose yet different methods or research designs [5]. For example, if selection is the key concern, Heckman selection models [20, 21, 46] are likely a good choice, although they require a variable with properties very similar to those of an instrument [16, 95].

10 Recommendations

10.1 Recommendations for Dealing with Potential Endogeneity

The flowchart in Figure 10 represents a guide for researchers on how to deal with endogeneity in regression analysis. It is based on advice in prior literature [68] and has additionally been validated with an expert who has published repeatedly on the subject of endogeneity and its remedies.

If researchers are not interested in making causal claims, they may readily resort to analytical methods, such as OLS or similar non-instrumented regression, and may largely stop worrying about potential endogeneity. The same is true if they want to make causal claims, but there is simply no reasonable endogeneity concern. If there are endogeneity concerns, however, they need to be addressed. If they can be addressed through control variables or fixed effects, non-instrumented regression can still be used. If not, researchers will need to find instruments. Each instrument must be shown to be relevant (e.g., by performing weak identification tests) and exogenous (by theoretical argument and possibly tests of overidentifying restrictions). If no such instruments can be found, researchers may resort to either attempting to demonstrate that there is likely no bias due to endogeneity (using, e.g., ITCV or RIR) or they may use methods that do not require external instruments (such as the use of Gaussian copulas or heteroskedasticity-based techniques) which are beyond the scope of this article (but which we briefly discuss in the online supplement). If researchers have credible instruments, they may estimate both an non-instrumented regression and an instrumental variable model and compare them (e.g., using a Durbin–Wu–Hausman or Wald test). If the results differ, they should report the instrumental variable results. If the results are similar, researchers should report the non-instrumented regression results for efficiency reasons.

10.2 Recommendations for Reporting

When researchers decide to use instrumental variable approaches, we encourage them to provide certain pieces of information to allow readers to assess if endogeneity concerns were addressed properly. Specifically, we recommend to report:

- *Endogeneity concerns*: What are the potential endogeneity concerns? If possible, report specific concerns about which variables might be omitted.
- *Control variables*: Which endogeneity concerns are addressed through control variables or fixed effects? In case of fixed effects, report if standard errors were clustered.
- *Assumption to identify causality*: Report if the assumption is that the independent variable is uncorrelated with the error term (no endogeneity) or if the assumption is that one or more instruments are uncorrelated with the error term.
- *Relevance of instruments*: Why are the instruments relevant? Provide both theoretical rationale and empirical test results (in particular full first-stage results and F statistics for all instruments).
- *Exogeneity of instruments*: Why are the instruments exogenous? Provide both theoretical rationale and empirical test results (if more than one credible instrument at hand).
- *Robustness*: How robust are the results? Report original non-instrumented results and results of a model with only the strongest instrument. Report results of a test for endogeneity (if strong and exogenous instruments exist).

11 Conclusion

We explicated the technique of two-stage regressions with instrumental variables and provided practical recommendations to researchers in the field of empirical software engineering. Despite its limitations, it is the dominant method to make credible causal inferences in many scientific fields, including economics, which notoriously has "an obsession with establishing causal relationships" [78, p. 125]. We hope that our article can serve as a guide for software engineering researchers to adequately account for the phenomenon of endogeneity in their studies, especially as caused by omitted confounding variables.

Acknowledgments

The first author would like to acknowledge many useful insights on the topic from Dominik Papies during a seminar. We would also like to thank John Busenbark for feedback on an earlier draft and the validation of the flowchart. We further thank three anonymous reviewers and Associate Editor Bogdan Vasilescu for valuable insights and guidance. Finally, we thank FrontEndART Software for providing us with documentation for the SourceMeter software.

References

[1] W. Abdelmoez, Mohamed Kholief, and Fayrouz M. Elsalmy. 2012. Bug fix-time prediction model using naïve Bayes classifier. In *Proceedings of the 22nd International Conference on Computer Theory and Applications (ICCTA '12)*. IEEE, 167–172. DOI: https://doi.org/10.1109/ICCTA.2012.6523564

[2] Manish Agrawal and Kaushal Chari. 2007. software effort, quality, and cycle time: A study of CMM level 5 projects. *IEEE Transactions on Software Engineering* 33, 3 (2007), 145–156. DOI: https://doi.org/10.1109/TSE.2007.29

[3] Joshua D. Angrist. 1990. lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review* 80, 3 (1990), 313–336.

[4] Joshua David Angrist and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton and Oxford.

[5] John Antonakis, Samuel Bendahan, Philippe Jacquart, and Rafael Lalive. 2010. On making causal claims: A review and recommendations. *The Leadership Quarterly* 21, 6 (2010), 1086–1120. DOI: https://doi.org/10.1016/j.leaqua.2010.10.010

[6] John Antonakis, Samuel Bendahan, Philippe Jacquart, and Rafael Lalive. 2014. causality and endogeneity: Problems and solutions. In *The Oxford Handbook of Leadership and Organizations*. David V. Day (Ed.), Oxford University Press, 1–53. DOI: https://doi.org/10.1093/oxfordhb/9780199755615.013.007

[7] Guilhem Bascle. 2008. Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization* 6, 3 (2008), 285–327. DOI: https://doi.org/10.1177/1476127008094339

[8] Michel Benaroch and Kalle Lyytinen. 2022. How much does software complexity matter for maintenance productivity? The link between team instability and diversity. *IEEE Transactions on Software Engineering* 49, 4 (2022), 1–17. DOI: https://doi.org/10.1109/TSE.2022.3222119

[9] Morten Bennedsen, Hans Christian Kongsted, and Kasper Meisner Nielsen. 2008. The causal effect of board size in the performance of small and medium-sized firms. *Journal of Banking & Finance* 32, 6 (2008), 1098–1109. DOI: https://doi.org/10.1016/j.jbankfin.2007.09.016

[10] M. Bertrand, E. Duflo, and S. Mullainathan. 2004. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics* 119, 1 (2004), 249–275. DOI: https://doi.org/10.1162/003355304772839588

[11] Pamela Bhattacharya and Iulian Neamtiu. 2011. Bug-fix time prediction models. In *Proceeding of the 8th Working Conference on Mining Software repositories—MSR'11*. Arie van Deursen, Tao Xie, and Thomas Zimmermann (Eds.), ACM, New York, NY, 207–210. DOI: https://doi.org/10.1145/1985441.1985472

[12] Anjali Thomas Bohlken and Ernest John Sergenti. 2010. Economic growth and ethnic violence: An empirical investigation of Hindu—Muslim riots in India. *Journal of Peace Research* 47, 5 (2010), 589–600. DOI: https://doi.org/10.1177/0022343310373032

[13] Grégoire Bollmann, Serguei Rouzinov, André Berchtold, and Jérôme Rossier. 2019. Illustrating instrumental variable regressions using the career adaptability—Job satisfaction relationship. *Frontiers in Psychology* 10 (2019), 1481. DOI: https://doi.org/10.3389/fpsyg.2019.01481

[14] John Bound, David A. Jaeger, and Regina M. Baker. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 430 (1995), 443–450. DOI: https://doi.org/10.1080/01621459.1995.10476536

[15] Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Studies in interactional sociolinguistics, Vol. 4. Cambridge University Press, Cambridge and New York, NY, USA.

[16] Shawn Bushway, Brian D. Johnson, and Lee Ann Slocum. 2007. Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology. *Journal of Quantitative Criminology* 23, 2 (2007), 151–178. DOI: https://doi.org/10.1007/s10940-007-9024-4

[17] J. S. Butler, Richard V. Burkhauser, Jean M. Mitchell, and Theodore P. Pincus. 1987. Measurement error in self-reported health variables. *The Review of Economics and Statistics* 69, 4 (1987), 644. DOI: https://doi.org/10.2307/1935959

[18] D. T. Campbell. 1957. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin* 54, 4 (1957), 297–312. DOI: https://doi.org/10.1037/h0040950

[19] Nathan Cassee, Bogdan Vasilescu, and Alexander Serebrenik. 2020. The silent helper: The impact of continuous integration on code reviews. In *Proceedings of the IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER '20)*. IEEE, 423–434. DOI: https://doi.org/10.1109/SANER48275.2020.9054818

[20] S. Trevis Certo, John R. Busenbark, Hyun-soo Woo, and Matthew Semadeni. 2016. Sample selection bias and Heckman models in strategic management research. *Strategic Management Journal* 37, 13 (2016), 2639–2657. DOI: https://doi.org/10.1002/smj.2475

[21] Joseph A. Clougherty, Tomaso Duso, and Johannes Muck. 2016. Correcting for self-selection based endogeneity in management research. *Organizational Research Methods* 19, 2 (2016), 286–347. DOI: https://doi.org/10.1177/1094428115619013

[22] Stefano Comino, Fabio M. Manenti, and Franco Mariuzzo. 2016. To upgrade or not to upgrade? The release of new versions to survive in the hypercompetitive app market. In *Proceedings of the International Workshop on App Market Analytics*. Meiyappan Nagappan, Federica Sarro, and Emad Shihab (Eds.), ACM, New York, NY, 37–42. DOI: https://doi.org/10.1145/2993259.2993261

[23] Cesar Couto, Christofer Silva, Marco Tulio Valente, Roberto Bigonha, and Nicolas Anquetil. 2012. Uncovering causal relationships between software metrics and bugs. In *Proceedings of the 16th European Conference on Software Maintenance and Reengineering*. IEEE, 223–232. DOI: https://doi.org/10.1109/CSMR.2012.31

[24] Scott Cunningham. 2021. *Causal inference: The mixtape*. Yale University Press, New Haven and London. https://doi.org/10.12987/9780300255881

[25] Marco D'Ambros, Alberto Bacchelli, and Michele Lanza. 2010. On the impact of design flaws on software defects. In *Proceedings of the 10th International Conference on Quality Software*. IEEE, 23–31. DOI: https://doi.org/10.1109/QSIC.2010.58

[26] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. 250–259.

[27] Giuseppe Destefanis, Marco Ortu, Steve Counsell, Stephen Swift, Michele Marchesi, and Roberto Tonelli. 2016. Software development: Do good manners matter? *PeerJ Computer Science* 2 (2016), e73. DOI: https://doi.org/10.7717/peerj-cs.73

[28] James Durbin. 1954. Errors in variables. *Review of the International Statistical Institute* 22, 1/3 (1954), 23–32. DOI: https://doi.org/10.2307/1401917

[29] K. El Emam, S. Benlarbi, N. Goel, and S. N. Rai. 2001. The confounding effect of class size on the validity of object-oriented metrics. *IEEE Transactions on Software Engineering* 27, 7 (2001), 630–650. DOI: https://doi.org/10.1109/32.935855

[30] Davide Falessi, Aalok Ahluwalia, and Massimiliano Di Penta. 2022. The impact of dormant defects on defect prediction: A study of 19 apache projects. *ACM Transactions on Software Engineering and Methodology* 31, 1 (2022), 1–26. DOI: https://doi.org/10.1145/3467895

[31] Hongbo Fang, Hemank Lamba, James Herbsleb, and Bogdan Vasilescu. 2022. "This is damn slick!". In *Proceedings of the 44th International Conference on Software Engineering*. Matthew B. Dwyer, Daniela Damian, and Andreas Zeller (Eds.), ACM, New York, NY, 2116–2129. DOI: https://doi.org/10.1145/3510003.3510121

[32] Isabella Ferreira, Jinghui Cheng, and Bram Adams. 2021. The "shut the f**k up" phenomenon: Characterizing incivility in open source code review discussions. In *Proceedings of the ACM on Human-Computer Interaction,* Vol. 5. 1–35. DOI: https://doi.org/10.1145/3479497

[33] Armstrong Foundjem, Ellis Eghan, and Bram Adams. 2021. Onboarding vs. diversity, productivity and quality—Empirical study of the openstack ecosystem. In *Proceedings of the IEEE/ACM 43rd International Conference on Software Engineering (ICSE'21)*. IEEE, 1033–1045. DOI: https://doi.org/10.1109/ICSE43902.2021.00097

[34] Kenneth A. Frank. 2000. Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research* 29, 2 (2000), 147–194. DOI: https://doi.org/10.1177/0049124100029002001

[35] Kenneth A. Frank, Spiro J. Maroulis, Minh Q. Duong, and Benjamin M. Kelcey. 2013. What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis* 35, 4 (2013), 437–460. DOI: https://doi.org/10.3102/0162373713493129

[36] Harry Garretsen, Janka I. Stoker, Dimitrios Soudis, Ron Martin, and Jason Rentfrow. 2019. The relevance of personality traits for urban economic growth: Making space for psychological factors. *Journal of Economic Geography* 19, 3 (2019), 541–565. DOI: https://doi.org/10.1093/jeg/lby025

[37] Frank Germann, Peter Ebbes, and Rajdeep Grewal. 2015. The chief marketing officer matters! *Journal of Marketing* 79, 3 (2015), 1–22. DOI: https://doi.org/10.1509/jm.14.0244

[38] A. Gopal, M. S. Krishnan, T. Mukhopadhyay, and D. R. Goldenson. 2002. Measurement programs in software development: determinants of success. *IEEE Transactions on Software Engineering* 28, 9 (2002), 863–875. DOI: https://doi.org/10.1109/TSE.2002.1033226

[39] C. W. J. Granger and Paul Newbold. 1977. *Forecasting Economic Time Series*. Academic Press, New York, NY, USA and London.

[40] Yunfang Guo and Philipp Leitner. 2019. Studying the impact of CI on pull request delivery time in open source projects-a conceptual replication. *PeerJ Computer science* 5, e245. DOI: https://doi.org/10.7717/peerj-cs.245

[41] Abhinav Gupta, Forrest Briscoe, and Donald C. Hambrick. 2018. Evenhandedness in resource allocation: Its relationship with CEO ideology, organizational discretion, and firm performance. *Academy of Management Journal* 61, 5 (2018), 1848–1868. DOI: https://doi.org/10.5465/amj.2016.1155

[42] Tracy Hall, Min Zhang, David Bowes, and Yi Sun. 2014. Some code smells have a significant but small effect on faults. *ACM Transactions on Software Engineering and Methodology* 23, 4 (2014), 1–39. DOI: https://doi.org/10.1145/2629648

[43] Barton H. Hamilton and Jackson A. Nickerson. 2003. Correcting for endogeneity in strategic management research. *Strategic Organization* 1, 1 (2003), 51–78. DOI: https://doi.org/10.1177/1476127003001001218

[44] Jerry A. Hausman. 1978. Specification tests in econometrics. *Econometrica* 46, 6 (1978), 1251–1271. DOI: https://doi.org/10.2307/1913827

[45] Fumio Hayashi. 2000. *Econometrics*. Princeton University Press, Princeton.

[46] James J. Heckman. 1979. Sample selection bias as a specification error. *Econometrica* 47, 1 (1979), 153–161. DOI: https://doi.org/10.2307/1912352

[47] James J. Heckman. 2005. The scientific model of causality. *Sociological Methodology* 35, 1–97.

[48] Steffen Herbold, Alexander Trautsch, Fabian Trautsch, and Benjamin Ledel. 2022. Problems with SZZ and features: An empirical study of the state of practice of defect prediction data collection. *Empirical Software Engineering* 27, 2 (2022), 42. DOI: https://doi.org/10.1007/s10664-021-10092-4

[49] Kim Herzig, Sascha Just, and Andreas Zeller. 2013. It's not a bug, it's a feature: How misclassification impacts bug prediction. In *Proceedings of the 35th International Conference on Software Engineering (ICSE '13)*. IEEE, 392–401.

[50] Aaron D. Hill, Scott G. Johnson, Lindsey M. Greco, Ernest H. O'Boyle, and Sheryl L. Walter. 2021. Endogeneity: A review and agenda for the methodology-practice divide affecting micro and macro research. *Journal of Management* 47, 1 (2021), 105–143. DOI: https://doi.org/10.1177/0149206320960533

[51] Paul W. Holland. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81, 396 (1986), 945–960.

[52] David A. Kenny. 1979. *Correlation and Causality*. John Wiley & Sons, New York, NY, USA.

[53] Sunghun Kim and E. James Whitehead. 2006. How long did it take to fix bugs?. In *Proceedings of the 2006 international workshop on Mining software repositories—MSR'06*. Stephan Diehl, Harald Gall, and Ahmed E. Hassan (Eds.), ACM Press, New York, NY, 173–174. DOI: https://doi.org/10.1145/1137983.1138027

[54] Yongnam Kim and Peter Steiner. 2016. Quasi-experimental designs for causal inference. *Educational psychologist* 51, 3–4 (2016), 395–405. DOI: https://doi.org/10.1080/00461520.2016.1207177

[55] David F. Larcker and Tjomme O. Rusticus. 2010. On the use of instrumental variables in accounting research. *Journal of Accounting and Economics* 49, 3 (2010), 186–205. DOI: https://doi.org/10.1016/j.jacceco.2009.11.004

[56] David Lewis. 1973. Causation. *The Journal of Philosophy* 70, 17 (1973), 556. DOI: https://doi.org/10.2307/2025310

[57] Amir N. Licht, Chanan Goldschmidt, and Shalom H. Schwartz. 2007. Culture rules: The foundations of the rule of law and other norms of governance. *Journal of Comparative Economics* 35, 4 (2007), 659–688. DOI: https://doi.org/10.1016/j.jce.2007.09.001

[58] Danaja Maldeniya, Ceren Budak, Lionel P. Robert Jr., and Daniel M. Romero. 2020. Herding a deluge of good Samaritans: How GitHub projects respond to increased attention. In *Proceedings of The Web Conference 2020*. Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.), ACM, New York, NY, 2055–2065. DOI: https://doi.org/10.1145/3366423.3380272

[59] John Stuart Mill. 1843. *A System of Logic, Ratiocinative and Inductive,* Vol. 1. John W. Parker, London.

[60] Ambarish Moharil, Dmitrii Orlov, Samar Jameel, Tristan Trouwen, Nathan Cassee, and Alexander Serebrenik. 2022. Between JIRA and GitHub. In *Proceedings of the 19th International Conference on Mining Software Repositories*. David Lo, Shane McIntosh, and Nicole Novielli (Eds.), ACM, New York, NY, 112–116. DOI: https://doi.org/10.1145/3524842.3528528

[61] Michael P. Murray. 2006. Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives* 20, 4 (2006), 111–132. DOI: https://doi.org/10.1257/jep.20.4.111

[62] Nachiappan Nagappan, Thomas Ball, and Andreas Zeller. 2006. Mining metrics to predict component failures. In *Proceedings of the 28th International Conference on Software Engineering*. Leon J. Osterweil, Dieter Rombach, and Mary Lou Soffa (Eds.), ACM, New York, NY, 452–461. DOI: https://doi.org/10.1145/1134285.1134349

[63] A. L. Nagar. 1959. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica* 27, 4 (1959), 575–595. DOI: https://doi.org/10.2307/1909352

[64] Charles R. Nelson and Richard Startz. 1990a. The distribution of the instrumental variables estimator and its T-ratio when the instrument is a poor one. *The Journal of Business* 63, 1 (1990), S125–S140.

[65] Charles R. Nelson and Richard Startz. 1990b. Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica* 58, 4 (1990), 967–976. DOI: https://doi.org/10.2307/2938359

[66] Marco Ortu, Tracy Hall, Michele Marchesi, Roberto Tonelli, David Bowes, and Giuseppe Destefanis. 2018. Mining communication patterns in software development. In *Proceedings of the 14th International Conference on Predictive*

Models and Data Analytics in Software Engineering. Burak Turhan, Ayse Tosun, and Shane McIntosh (Eds.), ACM, New York, NY, 70–79. DOI: https://doi.org/10.1145/3273934.3273943

[67] Mike Papadakis, Donghwan Shin, Shin Yoo, and Doo-Hwan Bae. 2018. Are mutation scores correlated with real fault detection?. In *Proceedings of the 40th International Conference on Software Engineering*. Michel Chaudron, Ivica Crnkovic, Marsha Chechik, and Mark Harman (Eds.), ACM, New York, NY, 537–548. DOI: https://doi.org/10.1145/3180155.3180183

[68] Dominik Papies, Peter Ebbes, and Harald J. van Heerde. 2017. Addressing endogeneity in marketing models. In *Advanced Methods for Modeling Markets*. Peter S. H. Leeflang, Jaap E. Wieringa, Tammo H. A. Bijmolt, and Koen H. Pauwels (Eds.), Springer International Publishing, Cham, 581–627. DOI: https://doi.org/10.1007/978-3-319-53469-5_18

[69] Judea Pearl. 2000. *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press, Cambridge.

[70] Judea Pearl. 2016. The Three Layer Causal Hierarchy. Retrieved from https://web.cs.ucla.edu/kaoru/3-layer-causal-hierarchy.pdf

[71] Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, NY, USA.

[72] Adithya Raghuraman, Truong Ho-Quang, Michel R. V. Chaudron, Alexander Serebrenik, and Bogdan Vasilescu. 2019. Does UML modeling associate with lower defect proneness?: A preliminary empirical investigation. In *Proceedings of the IEEE/ACM 16th International Conference on Mining Software Repositories (MSR '19)*. IEEE, 101–104. DOI: https://doi.org/10.1109/MSR.2019.00024

[73] Narayan Ramasubbu and Rajesh Krishna Balan. 2007. Globally distributed software development project performance. In *Proceedings of the 6th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC-FSE'07)*. Ivica Crnkovic and Antonia Bertolino (Eds.), ACM, New York, NY, 125. DOI: https://doi.org/10.1145/1287624.1287643

[74] Narayan Ramasubbu, Marcelo Cataldo, Rajesh Krishna Balan, and James D. Herbsleb. 2011. Configuring global software teams. In *Proceedings of the 33rd International Conference on Software Engineering*. Richard N. Taylor, Harald Gall, and Nenad Medvidović (Eds.), ACM, New York, NY, 261–270. DOI: https://doi.org/10.1145/1985793.1985830

[75] Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55. DOI: https://doi.org/10.1093/biomet/70.1.41

[76] Peter E. Rossi. 2014. Even the rich can make themselves poor: A critical examination of IV methods in marketing Applications. *Marketing Science* 33, 5 (2014), 655–672. DOI: https://doi.org/10.1287/mksc.2014.0860

[77] Donald B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 5 (1974), 688–701. DOI: https://doi.org/10.1037/h0037350

[78] John Rust. 2016. Mostly useless econometrics? Assessing the causal effect of econometric theory. *Foundations and Trends in Accounting* 10, 2–4 (2016), 125–203. DOI: https://doi.org/10.1561/1400000049

[79] Matthew Semadeni, Michael C. Withers, and S. Trevis Certo. 2014. The perils of endogeneity and instrumental variables in strategy research: Understanding through simulations. *Strategic Management Journal* 35, 7 (2014), 1070–1079. DOI: https://doi.org/10.1002/smj.2136

[80] John Shea. 1997. Instrument relevance in multivariate linear models: A simple measure. *The Review of Economics and Statistics* 79, 2 (1997), 348–352.

[81] Julien Siebert. 2023. Applications of statistical causal inference in software engineering. *Information and Software Technology* 159 (2023), 107198. DOI: https://doi.org/10.1016/j.infsof.2023.107198

[82] Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science* 5, 4 (1990), 465–472. Retrieved from https://www.jstor.org/stable/2245382

[83] Douglas Staiger and James H. Stock. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65, 3 (1997), 557–586. DOI: https://doi.org/10.2307/2171753

[84] James Stock and Motohiro Yogo. 2005. Testing for weak instruments in linear IV regressions. In *Identification and Inference for Econometric Models*. Donald W. K. Andrews, James H. Stock, and Thomas J. Rothenberg (Eds.), Cambridge University Press, Cambridge, 80–108.

[85] James H. Stock and Mark W. Watson. 2019. *Introduction to Econometrics* (4 ed.). Pearson, New York, NY, USA.

[86] James H. Stock, Jonathan H. Wright, and Motohiro Yogo. 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* 20, 4 (2002), 518–529. DOI: https://doi.org/10.1198/073500102288618658

[87] Bing Sun, Jun Sun, Long H. Pham, and Jie Shi. 2022. Causality-based neural network repair. In *Proceedings of the 44th International Conference on Software Engineering*. Matthew B. Dwyer, Daniela Damian, and Andreas Zeller (Eds.), ACM, New York, NY, 338–349. DOI: https://doi.org/10.1145/3510003.3510080

[88] Kazunori Terada, Mitsuki Okazoe, and Jonathan Gratch. 2021. Effect of politeness strategies in dialogue on negotiation outcomes. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents*. ACM, New York, NY, 195–202. DOI: https://doi.org/10.1145/3472306.3478336

[89] Zoltán Tóth, Péter Gyimesi, and Rudolf Ferenc. 2016. A public bug database of GitHub projects and its application in bug prediction. In *Computational Science and Its Applications—ICCSA 2016*. Osvaldo Gervasi, Beniamino Murgante, Sanjay Misra, Ana Maria A. C. Rocha, Carmelo M. Torre, David Taniar, Bernady O. Apduhan, Elena Stankova, and Shangguang Wang (Eds.), Lecture Notes in Computer Science, Vol. 9789, Springer International Publishing, Cham, 625–638. DOI: https://doi.org/10.1007/978-3-319-42089-9_44

[90] Asher Trockman, Rijnard van Tonder, and Bogdan Vasilescu. 2019. Striking gold in software repositories? An econometric study of cryptocurrencies on GitHub. In *Proceedings of the IEEE/ACM 16th International Conference on Mining Software Repositories (MSR '19)*. IEEE, 181–185. DOI: https://doi.org/10.1109/MSR.2019.00036

[91] Masateru Tsunoda and Sousuke Amasaki. 2017. On software productivity analysis with propensity score matching. In *Proceedings of the ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '17)*. IEEE, 436–441. DOI: https://doi.org/10.1109/ESEM.2017.59

[92] Yi Wang. 2021. The price of being polite: politeness, social status, and their joint impacts on community Q & A efficiency. *Journal of Computational Social Science* 4, 1 (2021), 101–122. DOI: https://doi.org/10.1007/s42001-020-00068-7

[93] Cathrin Weiss, Rahul Premraj, Thomas Zimmermann, and Andreas Zeller. 2007. How long will it take to fix this bug?. In *Proceedings of the 4th International Workshop on Mining Software Repositories (MSR '07: ICSE Workshops '07)*. IEEE, 1–8. DOI: https://doi.org/10.1109/MSR.2007.13

[94] Mairieli Wessel, Alexander Serebrenik, Igor Wiese, Igor Steinmacher, and Marco A. Gerosa. 2020. Effects of adopting code review bots on pull requests to OSS projects. In *Proceedings of the IEEE International Conference on Software Maintenance and Evolution (ICSME '20)*. IEEE, 1–11. DOI: https://doi.org/10.1109/ICSME46990.2020.00011

[95] Sarah E. Wolfolds and Jordan Siegel. 2019. Misaccounting for endogeneity: The peril of relying on the Heckman two-step method without a valid instrument. *Strategic Management Journal* 40, 3 (2019), 432–462. DOI: https://doi.org/10.1002/smj.2995

[96] Jeffrey M. Wooldridge. 2010. *Econometric Analysis of Cross Section and Panel Data* (2 ed.). MIT Press, Cambridge and London.

[97] Jeffrey M. Wooldridge. 2020. *Introductory Econometrics: A Modern Approach* (7 ed.). Cengage, Boston.

[98] De-Min Wu. 1973. Alternative tests of independence between stochastic regressors and disturbances. *Econometrica* 41, 4 (1973), 733–750. DOI: https://doi.org/10.2307/1914093

[99] Hongyu Zhang, Liang Gong, and Steve Versteeg. 2013. Predicting bug-fixing time: An empirical study of commercial software projects. In *Proceedings of the 35th International Conference on Software Engineering (ICSE '13)*. IEEE, 1042–1051. DOI: https://doi.org/10.1109/ICSE.2013.6606654

[100] Mengdi Zhang and Jun Sun. 2022. Adaptive fairness improvement based on causality analysis. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. Abhik Roychoudhury, Cristian Cadar, and Miryung Kim (Eds.), ACM, New York, NY, 6–17. DOI: https://doi.org/10.1145/3540250.3549103

[101] Yangyang Zhao, Alexander Serebrenik, Yuming Zhou, Vladimir Filkov, and Bogdan Vasilescu. 2017. The impact of continuous integration on other software development practices: A large-scale empirical study. In *Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE '17)*. IEEE, 60–71. DOI: https://doi.org/10.1109/ASE.2017.8115619

[102] Theo Zimmermann and Annali Casanueva Artis. 2019. Impact of switching bug trackers: A case study on a medium-sized open source project. In *Proceedings of the IEEE International Conference on Software Maintenance and Evolution (ICSME '19)*. IEEE, 13–23. DOI: https://doi.org/10.1109/ICSME.2019.00011

Received 7 September 2023; revised 8 February 2024; accepted 23 May 2024