Endogeneity, Instruments, and Two-Stage Models

Lorenz Graf-Vlachy
lorenz.graf-vlachy@iste.uni-stuttgart.de
University of Stuttgart, Institute of Software Engineering
Stuttgart, Germany
TU Dortmund University
Dortmund, Germany

Stefan Wagner
stefan.wagner@iste.uni-stuttgart.de
University of Stuttgart, Institute of Software Engineering
Stuttgart, Germany

ABSTRACT

Background: Studies in software engineering are often particularly useful if they make causal claims because this allows practitioners to identify how they can influence outcomes of interest. Unfortunately, many non-experimental studies suffer from potential endogeneity through omitted confounding variables, which precludes claims of causality. *Aims and Method:* We introduce instrumental variables and two-stage models as a means to account for endogeneity to the field of empirical software engineering. *Results and Conclusions:* We define endogeneity, explain its primary cause, and lay out the idea behind instrumental variable approaches and two-stage models.

CCS CONCEPTS

• **General and reference** → **Empirical studies**.

KEYWORDS

regression, endogeneity, confounder, two-stage least squares, 2SLS

ACM Reference Format:
Lorenz Graf-Vlachy and Stefan Wagner. 2024. Endogeneity, Instruments, and Two-Stage Models. In *2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion '24), April 14–20, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3639478.3643064

1 INTRODUCTION

In 2017, a *Stack Overflow* blog post showed that there was a strong relationship between whether programmers used spaces or tabs for indentation and these programmers' salaries [3]. Of course, observers were quick to argue that there were likely some confounding variables missing from the analysis that could explain the relationship. In econometric terms, the analysis likely suffered from endogeneity. While the findings were thus interesting, and while they may allow to *predict* programmers' salaries if their indentation preferences are known, one cannot credibly claim that there is a causal link between the two variables.

Scientific studies in empirical software engineering often face such challenges. Researchers might find relationships between variables, but it is frequently not clear whether such findings represent causal relationships or if they are artifacts of endogeneity. One technique to deal with endogeneity that has been developed in econometrics and that has found widespread acceptance is the use of instrumental variables in two-stage regression models.

ICSE-Companion '24, April 14–20, 2024, Lisbon, Portugal

ACM ISBN 979-8-4007-0502-1/24/04…$15.00
https://doi.org/10.1145/3639478.3643064

2 ENDOGENEITY IN REGRESSION MODELS

Endogeneity as a general phenomenon is hard to describe intuitively. We must thus rely on a mathematical definition. Consider the simplest possible regression model shown in Equation 1. Y denotes the dependent variable, X denotes an independent variable (or "regressor") (which we may hypothesize to be a cause of Y), and ε denotes the error term. All these variables are scalars, and we omit indices for convenience. 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 "ordinary least squares regression" (OLS) analysis, the estimator selects values for all coefficients (β_i) such that the squared deviations (or "residuals") of the fitted regression line from the observed sample data is minimized. It is important to understand that in Equation 1, any variance in Y that is not explicitly accounted for (i.e., that is not in X) is captured in the error term ε.

OLS is the best linear unbiased estimator when a variety of assumptions hold [4]. 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 $\mathbb{E}(\varepsilon|X) = 0$.

Endogeneity is the violation of this assumption, i.e., a situation in which the expected value of the error is dependent on X. Endogeneity exists when systematic information from the regressor X is captured in the error term ε and there is thus a non-zero covariance between X and the errors, i.e., $cov(X, \varepsilon) \neq 0$ (see Figure 1).

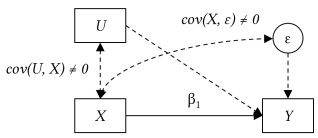

Figure 1: Endogeneity due to omitted confounding variable.

Endogeneity is a problem because it renders OLS inconsistent. This means that, even for very large samples, the estimated values do not converge on the true population parameters. Therefore, the estimates of the coefficients—β_1 in our case—are not trustworthy.

Note that endogeneity *cannot be tested for*. Since the true population parameters are unknown (a researcher only knows the observed sample, not the population), it is impossible to know the error term ε, and thus impossible to assess its covariance with X.

In many cases, endogeneity is the result of confounding variables that are omitted from the model. The "tabs vs. spaces" example illustrates this. As some languages mandate the use of spaces for indentation, one might suspect that the used programming languages might drive the results. If, for example, Python programmers were systematically better-paid than others, programming language would be an omitted confounding variable in the model.

Again, Figure 1 visualizes such a situation. X and U are two variables that influence Y. If U is omitted from the regression model, the shared variance between U and Y is not accounted for and thus enters the error term ε. If 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 inconsistent.

Many empirical software engineering researchers are aware of such endogeneity. The "threats to validity" sections of papers frequently discuss potentially "confounding factors" or "confounders" (e.g., [2]). Nevertheless, such awareness often appears intuitive rather than technical and is typically not accompanied by a convincing implementation of countermeasures.

3 INSTRUMENTAL VARIABLE REGRESSION

To address endogeneity, researchers can resort to instrumental variable regressions. The most common kind is the "two-stage least squares" (2SLS) approach. The employed instrumental variables are also frequently simply referred to as "instruments".

3.1 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 simple linear model with two independent variables X_1 and X_2:

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

Further assume that X_1 is endogenous, i.e., correlated with the error term ε, and X_2 is exogenous, i.e., uncorrelated with 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 use only one instrument, Z_1. We would thus estimate the following model (with ζ as its error term) as the first stage:

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

This allows us to calculate a predicted version of X_1, i.e., $\hat{X_1}$:

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

In the second stage, we then estimate the model we are actually interested in (specified in Equation 2), but we 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 exogenous instrument Z_1 share, so that the predicted $\hat{X_1}$ does not include any variance shared with the error term in the second stage.

Note that the standard errors would not be accurate if we were to actually manually estimate the model this way [1]. Instead, the standard errors must be adjusted, which all major statistics packages do automatically when using appropriate commands.

3.2 Requirements for Instrumental Variables

While the statistical technique is fairly easy to implement, a challenge in instrumental variable approaches is the choice of instruments. For one, for an instrumental variable regression to cure the problem of endogeneity, one needs at least one instrument for every endogenous regressor. Second, all instruments must fulfill certain requirements [4]. Specifically, they must be relevant and exogenous (see Figure 2). Relevance means that *the instrument must be clearly related to the endogenous regressor*. This means that there should be a strong correlation between the instrument and the endogenous regressor, conditional on all other (exogenous) control variables. 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*. In case of endogeneity due to an omitted confounding variable, this is the case when the instrument is unrelated to the omitted variable, because then the instrument does not share variance with the error term.

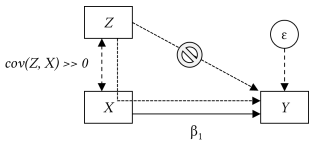

Figure 2: Requirements for instrumental variables.

4 CONCLUSION

In this work, we introduce instrumental variables and two-stage models to software engineering research. In the poster, we provide examples using data, demonstrating the problem of endogeneity and the efficacy of the proposed remedy. We also provide explicit guidelines for software engineering researchers.

REFERENCES

[1] Joshua David Angrist and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press, Princeton and Oxford.
[2] Armstrong Foundjem, Ellis Eghan, and Bram Adams. 2021. Onboarding vs. Diversity, Productivity and Quality — Empirical Study of the OpenStack Ecosystem. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE).* IEEE, 1033–1045. https://doi.org/10.1109/ICSE43902.2021.00097
[3] David Robinson. 2017. Developers Who Use Spaces Make More Money Than Those Who Use Tabs. https://stackoverflow.blog/2017/06/15/developers-use-spaces-make-money-use-tabs/
[4] Jeffrey M. Wooldridge. 2020. *Introductory econometrics: A modern approach* (7 ed.). Cengage, Boston.