Internal and External Validity [1]

Jasmine(Yu) Hao

Faculty of Business and Economics
Hong Kong University

August 17, 2021

[1] This section is based on Stock and Watson (2020), Chapter 9.

Internal and External Validity I

- The concepts of internal and external validity, defined in Key Concept 9.1, provide a framework for evaluating whether a statistical or econometric study is useful for answering a specific question of interest.
- Internal and external validity distinguish between the population and setting studied and the population and setting to which the results are generalized.
- The population studied is the population of entities—people, companies, school districts, and so forth—from which the sample was drawn. The population to which the results are generalized, or the population of interest, is the population of entities to which the causal inferences from the study are to be applied. For example, a high school (grades 9 through 12) principal might want to generalize our findings on class sizes and test scores in California elementary school districts (the population studied) to the population of high schools (the population of interest).

Internal and External Validity II

- By setting, we mean the institutional, legal, social, physical, and economic environment. For example, it would be important to know whether the findings of a laboratory experiment assessing methods for growing organic tomatoes could be generalized to the field—that is, whether the organic methods that work in the setting of a laboratory also work in the setting of the real world. We provide other examples of differences in populations and settings later in this section.

Threats to Internal Validity I

- Internal validity has two components. First, the estimator of the causal effect should be unbiased and consistent. For example, if bn STR is the OLS estimator of the effect on test scores of a unit change in the student–teacher ratio in a certain regression, then bn STR should be an unbiased and consistent estimator of the population causal effect of a change in the student–teacher ratio, bSTR.
- Second, hypothesis tests should have the desired significance level (the actual rejection rate of the test under the null hypothesis should equal its desired significance level), and confidence intervals should have the desired confidence level.

For example, if a confidence interval is constructed as bn $STR \pm 1.96SE(\hat{\beta}STR)$, this confidence interval should contain the true population causal effect, βSTR, with 95% probability over repeated samples drawn from the population being studied.

Threats to Internal Validity II

- In regression analysis, causal effects are estimated using the estimated regression function, and hypothesis tests are performed using the estimated regression coefficients and their standard errors.
- Accordingly, in a study based on OLS regression, the requirements for internal validity are that the OLS estimator is unbiased and consistent and that standard errors are computed in a way that makes confidence intervals have the desired confidence level.
- For various reasons, these requirements might not be met, and these reasons constitute threats to internal validity. These threats lead to failures of one or more of the least squares assumptions in Key Concept 6.4.
- For example, one threat that we have discussed at length is omitted variable bias; it leads to correlation between one or more regressors and the error term, which violates the first least squares assumption.

Threats to Internal Validity III

- If data are available on the omitted variable or on an adequate control variable, then this threat can be avoided by including that variable as an additional regressor.

Threats to External Validity I

Potential threats to external validity arise from differences between the population and setting studied and the population and setting of interest.

1. Differences in populations.
   1.1 Differences between the population studied and the population of interest can pose a threat to external validity. For example, laboratory studies of the toxic effects of chemicals typically use animal populations like mice (the population studied), but the results are used to write health and safety regulations for human populations (the population of interest). Whether mice and men differ sufficiently to threaten the external validity of such studies is a matter of debate.
   1.2 More generally, the true causal effect might not be the same in the population studied and the population of interest. This could be because the population was chosen in a way that makes it different from the population of interest, because of differences in characteristics of the populations, because of geographical differences, or because the study is out of date.

Threats to External Validity II

2. Differences in settings.
    2.1 Even if the population being studied and the population of interest are identical, it might not be possible to generalize the study results if the settings differ. For example, a study of the effect on college binge drinking of an antidrinking advertising campaign might not generalize to another, identical group of college students if the legal penalties for drinking at the two colleges differ. In this case, the legal setting in which the study was conducted differs from the legal setting to which its results are applied.
    2.2 More generally, examples of differences in settings include differences in the institutional environment (public universities versus religious universities), differences in laws (differences in legal penalties), and differences in the physical environment (tailgate-party binge drinking in southern California versus Fairbanks, Alaska).

Application to test scores and the student–teacher ratio I

- Chapters 7 and 8 reported statistically significant, but substantively small, estimated improvements in test scores resulting from reducing the student–teacher ratio. This analysis was based on test results for California school districts. Suppose for the moment that these results are internally valid. To what other populations and settings of interest could this finding be generalized?
- The closer the population and setting of the study are to those of interest, the stronger the case is for external validity.
- For example, college students and college instruction are very different from elementary school students and instruction, so it is implausible that the effect of reducing class sizes estimated using the California elementary school district data would generalize to colleges.

Application to test scores and the student–teacher ratio II

- On the other hand, elementary school students, curriculum, and organization are broadly similar throughout the United States, so it is plausible that the California results might generalize to performance on standardized tests in other U.S. elementary school districts.

How to assess the external validity of a study I

- External validity must be judged using specific knowledge of the populations and settings studied and those of interest.
- Important differences between the two will cast doubt on the external validity of the study. Sometimes there are two or more studies on different but related populations.
- If so, the external validity of both studies can be checked by comparing their results.
- For example, in Section 9.4, we analyze test score and class size data for elementary school districts in Massachusetts and compare the Massachusetts and California results. In general, similar findings in two or more studies bolster claims to external validity, while differences in their findings that are not readily explained cast doubt on their external validity.

How to design an externally valid study I

- Because threats to external validity stem from a lack of comparability of populations and settings, these threats are best minimized at the early stages of a study, before the data are collected. Study design is beyond the scope of this textbook, and the interested reader is referred to Shadish, Cook, and Campbell (2002).

Threats to Internal Validity of Multiple Regression Analysis

- Studies based on regression analysis are internally valid if the estimated regression coefficients are unbiased and consistent for the causal effect of interest and if their standard errors yield confidence intervals with the desired confidence level.
- This section surveys five reasons why the OLS estimator of the multiple regression coefficients might be biased, even in large samples: omitted variables, misspecification of the functional form of the regression function, imprecise measurement of the independent variables ("errors in variables"), sample selection, and simultaneous causality.
- All five sources of bias arise because the regressor is correlated with the error term in the population regression, violating the first least squares assumption in Key Concept 6.4. For each, we discuss what can be done to reduce this bias. The section concludes with a discussion of circumstances that lead to inconsistent standard errors and what can be done about it.

Omitted Variable Bias I

- Recall that omitted variable bias arises when a variable that both determines Y and is correlated with one or more of the included regressors is omitted from the regression.
- This bias persists even in large samples, so the OLS estimator is inconsistent.
- How best to minimize omitted variable bias depends on whether or not variables that adequately control for the potential omitted variable are available.
- Solutions to omitted variable bias when the variable is observed or there are adequate control variables. If you have data on the omitted variable, then you can include that variable in a multiple regression, thereby addressing the problem.
- Alternatively, if you have data on one or more control variables and if these control variables are adequate in the sense that they lead to conditional mean independence [Equation (6.18)], then including those control variables eliminates the potential bias in the coefficient on the variable of interest.

Omitted Variable Bias II

- Adding a variable to a regression has both costs and benefits. On the one hand, omitting the variable could result in omitted variable bias. On the other hand, including the variable when it does not belong (that is, when its population regression coefficient is 0) reduces the precision of the estimators of the other regression coefficients.
- In other words, the decision whether to include a variable involves a trade-off between bias and variance of the coefficient of interest. In practice, there are four steps that can help you decide whether to include a variable or set of variables in a regression.
    - The first step is to identify the key coefficient or coefficients of interest in your regression. In the test score regressions, this is the coefficient on the student–teacher ratio because the question originally posed concerns the effect on test scores of reducing the student–teacher ratio.

Omitted Variable Bias III

◇ The second step is to ask yourself: What are the most likely sources of important omitted variable bias in this regression? Answering this question requires applying economic theory and expert knowledge, and should occur before you actually run any regressions; because this step is done before analyzing the data, it is referred to as a priori ("before the fact") reasoning. In the test score example, this step entails identifying those determinants of test scores that, if ignored, could bias our estimator of the class size effect. The results of this step are a base regression specification, the starting point for your empirical regression analysis, and a list of additional, "questionable" control variables that might help to mitigate possible omitted variable bias.

Omitted Variable Bias IV

- The third step is to augment your base specification with the additional, questionable control variables identified in the second step. If the coefficients on the additional control variables are statistically significant and/or if the estimated coefficients of interest change appreciably when the additional variables are included, then they should remain in the specification and you should modify your base specification. If not, then these variables can be excluded from the regression.
- The fourth step is to present an accurate summary of your results in tabular form.

This provides "full disclosure" to a potential skeptic, who can then draw his or her own conclusions. Tables 7.1 and 8.3 are examples of this strategy. For example, in Table 8.3, we could have presented only the regression in column (7) because that regression summarizes the relevant effects and nonlinearities in the other regressions in that table. Presenting the other regressions, however, permits the skeptical reader to draw his or her own conclusions.

Solutions to omitted variable bias when adequate control variables are not available I

- Adding an omitted variable to a regression is not an option if you do not have data on that variable and if there are no adequate control variables. Still, there are three other ways to solve omitted variable bias. Each of these three solutions circumvents omitted variable bias through the use of different types of data.
- The first solution is to use data in which the same observational unit is observed at different points in time. For example, test score and related data might be collected for the same districts in 1995 and again in 2000. Data in this form are called panel data.
- As explained in Chapter 10, panel data make it possible to control for unobserved omitted variables as long as those omitted variables do not change over time.

Solutions to omitted variable bias when adequate control variables are not available II

- The second solution is to use instrumental variables regression. This method relies on a new variable, called an instrumental variable. Instrumental variables regression is discussed in Chapter 12.
- The third solution is to use a study design in which the effect of interest (for example, the effect of reducing class size on student achievement) is studied using a randomized controlled experiment. Randomized controlled experiments are discussed in Chapter 13.

Misspecification of the Functional Form of the Regression Function I

- If the true population regression function is nonlinear but the estimated regression is linear, then this functional form misspecification makes the OLS estimator biased.
- This bias is a type of omitted variable bias, in which the omitted variables are the terms that reflect the missing nonlinear aspects of the regression function. For example, if the population regression function is a quadratic polynomial, then a regression that omits the square of the independent variable would suffer from omitted variable bias.
- Solutions to functional form misspecification. When the dependent variable is continuous (like test scores), this problem of potential nonlinearity can be solved using the methods of Chapter 8. If, however, the dependent variable is discrete or binary (for example, if Yi equals 1 if the ith person attended college and equals 0 otherwise), things are more complicated. Regression with a discrete dependent variable is discussed in Chapter 11.

Measurement Error and Errors-in-Variables Bias I

- Suppose that in our regression of test scores against the student–teacher ratio we had in advertently mixed up our data, so that we ended up regressing test scores for fifth graders on the student–teacher ratio for tenth graders in that district.
- Although the student–teacher ratio for elementary school students and tenth graders might be correlated, they are not the same, so this mix-up would lead to bias in the estimated coefficient. This is an example of errors-in-variables bias because its source is an error in the measurement of the independent variable. This bias persists even in very large samples, so the OLS estimator is inconsistent if there is measurement error.

Measurement Error and Errors-in-Variables Bias II

- There are many possible sources of measurement error. If the data are collected through a survey, a respondent might give the wrong answer. For example, one question in the Current Population Survey involves last year's earnings. A respondent might not know his or her exact earnings or might misstate the amount for some other reason. If instead the data are obtained from computerized administrative records, there might have been errors when the data were first entered.
- To see that errors in variables can result in correlation between the regressor and the error term, suppose there is a single regressor Xi (say, actual earnings) which is measured imprecisely by Xi (the respondent's stated earnings). Because X_i, not X_i, is observed, the regression equation actually estimated is the one based on Xi.

Measurement Error and Errors-in-Variables Bias III

Written in terms of the imprecisely measured variable Xi, the population regression equation $Y_i = \beta_0 + \beta_1 X_i + u_i$

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1(X_i - \tilde{X}_i) + u_i] \\
&= \beta_0 + \beta_1 \tilde{X}_i + v_i,
\end{aligned}
$$

where $v_i = \beta_1(X_i - \tilde{X}_i) + u_i$.

- the population regression equation written in terms of X_i has an error term that contains the measurement error, the difference between Xi and Xi.
- If this difference is correlated with the measured value Xi, then the regressor Xi will be correlated with the error term, and $\hat{\beta}_{n,1}$ will be biased and inconsistent

$$\hat{\beta}_1 \to_p \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1.$$

Solutions to errors-in-variables bias I

- The best way to solve the errors-in-variables problem is to get an accurate measure of X. If this is impossible, however, econometric methods can be used to mitigate errors-in-variables bias.
- One such method is instrumental variables regression. It relies on having another variable (the instrumental variable) that is correlated with the actual value Xi but is uncorrelated with the measurement error. This method is studied in Chapter 12.
- A second method is to develop a mathematical model of the measurement error and, if possible, to use the resulting formulas to adjust the estimates. For example, if a researcher believes that the classical measurement error model applies and if she knows or can estimate the ratio s2w $>$ s2 X, then she can use Equation (9.2) to compute an estimator of b1 that corrects for the downward bias.

Solutions to errors-in-variables bias II

- Because this approach requires specialized knowledge about the nature of the measurement error, the details typically are specific to a given data set and its measurement problems, and we shall not pursue this approach further in this text.

Missing Data and Sample Selection I

- Missing data are a common feature of economic data sets. Whether missing data pose a threat to internal validity depends on why the data are missing. We consider three cases: when the data are missing completely at random, when the data are missing based on X, and when the data are missing because of a selection process that is related to Y beyond depending on X.
- When the data are missing completely at random—that is, for random reasons unrelated to the values of X or Y—the effect is to reduce the sample size but not introduce bias. For example, suppose you conduct a simple random sample of 100 classmates, then randomly lose half the records. It would be as if you had never surveyed those individuals. You would be left with a simple random sample of 50 classmates, so randomly losing the records does not introduce bias.

Missing Data and Sample Selection II

- When the data are missing based on the value of a regressor, the effect also is to reduce the sample size but not to introduce bias. For example, in the class size/ student–teacher ratio example, suppose we used only the districts in which the student–teacher ratio exceeds 20. Although we would not be able to draw conclusions about what happens when $STR \leq 20$, this would not introduce bias into our analysis of the class size effect for districts with $STR > 20$.
- In contrast to the first two cases, if the data are missing because of a selection process that is related to the value of the dependent variable (Y) beyond depending on the regressors (X), then this selection process can introduce correlation between the error term and the regressors. The resulting bias in the OLS estimator is called sample selection bias. An example of sample selection bias in polling was given in the box "Landon Wins!" in Section 3.1. In that example, the sample selection method(randomly selecting phone numbers of automobile owners) was related to the dependent variable (who the

Missing Data and Sample Selection III

individual supported for president in 1936) because in 1936 car owners with phones were more likely to be Republicans.

- The sample selection problem can be cast either as a consequence of nonrandom sampling or as a missing data problem. In the 1936 polling example, the sample was a random sample of car owners with phones, not a random sample of voters. Alternatively, this example can be cast as a missing data problem by imagining a random sample of voters but with missing data for those without cars and phones. The mechanism by which the data are missing is related to the dependent variable, leading to sample selection bias.

Simultaneous Causality I

- So far, we have assumed that causality runs from the regressors to the dependent variable (X causes Y). But what if causality also runs from the dependent variable to one or more regressors (Y causes X)? If so, causality runs "backward" as well as forward; that is, there is simultaneous causality. If there is simultaneous causality, an OLS regression picks up both effects, so the OLS estimator is biased and inconsistent.
- For example, our study of test scores focused on the effect on test scores of reducing the student–teacher ratio, so causality is presumed to run from the student–teacher ratio to test scores. Suppose, however, a government initiative subsidized hiring teachers in school districts with poor test scores. If so, causality would run in both directions: For the usual educational reasons, low student–teacher ratios would arguably lead to high test scores, but because of the government program, low test scores would lead to low student–teacher ratios.

Simultaneous Causality II

- Simultaneous causality leads to correlation between the regressor and the error term. In the test score example, suppose there is an omitted factor that leads to poor test scores; because of the government program, this factor that produces low scores in turn results in a low student–teacher ratio. Thus a negative error term in the population regression of test scores on the student–teacher ratio reduces test scores, but because of the government program, it also leads to a decrease in the student–teacher ratio. In other words, the student–teacher ratio is positively correlated with the error term in the population regression.
- This in turn leads to simultaneous causality bias and inconsistency of the OLS estimator. This correlation between the error term and the regressor can be made mathematically precise by introducing an additional equation that describes the reverse causal link. For convenience, consider just the two variables X and Y, and ignore other possible regressors.

Simultaneous Causality III

Accordingly, there are two equations, one in which X causes Y and one in which Y causes X:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$X_i = \gamma_0 + \gamma_1 Y_i + v_i.$$

- Equation (9.3) is the familiar one in which b1 is the effect on Y of a change in X, where u represents other factors. Equation (9.4) represents the reverse causal effect of Y on X. In the test score problem, Equation (9.3) represents the educational effect of class size on test scores, while Equation (9.4) represents the reverse causal effect of test scores on class size induced by the government program.

Simultaneous Causality IV

- Simultaneous causality leads to correlation between Xi and the error term ui in Equation (9.3). To see this, imagine that ui is positive, which increases Yi. However, this higher value of Yi affects the value of Xi through the second of these equations, and if g1 is positive, a high value of Yi will lead to a high value of Xi. In general, if g1 is nonzero, Xi and ui will be correlated.
- Because it can be expressed mathematically using two simultaneous equations, simultaneous causality bias is sometimes called simultaneous equations bias. Simultaneous causality bias is summarized in Key Concept 9.6.

Correlation of the error term across observations I

- In some settings, the population regression error can be correlated across observations. This will not happen if the data are obtained by sampling at random from the population
- Sometimes, however, sampling is only partially random.
  - The most common circumstance is when the data are repeated observations on the same entity over time, such as the same school district for different years. If the omitted variables that constitute the regression error are persistent (like district demographics),
  - "serial" correlation is induced in the regression error over time. Serial correlation in the error term can arise in panel data (e.g., data on multiple districts for multiple years) and in time series data (e.g., data on a single district for multiple years).
  - Another situation in which the error term can be correlated across observations is when sampling is based on a geographical unit. If there are omitted variables that reflect geographic influences, these omitted variables could result in correlation of the regression errors for adjacent observations.

Correlation of the error term across observations II

- Correlation of the regression error across observations does not make the OLS estimator biased or inconsistent, but it does violate the second least squares assumption. The consequence is that the OLS standard errors—both homoskedasticity-only and heteroskedasticity-robust—are incorrect in the sense that they do not produce confidence intervals with the desired confidence level.
- In many cases, this problem can be fixed by using an alternative formula for standard errors. We provide formulas for computing standard errors that are robust to both heteroskedasticity and serial correlation in Chapter 10 (regression with panel data) and in Chapter 16 (regression with time series data).

Internal and External Validity When the Regression Is Used for Prediction I

- When regression models are used for prediction, concerns about external validity are very important, but concerns about unbiased estimation of causal effects are not.

Chapter 4 began by considering two problems.

- A school superintendent wants to know how much test scores will increase if she reduces **class sizes** in her school district; that is, the superintendent wants to know the causal effect on test scores of a change in class size.
- A father, considering moving to a school district for which test scores are not publicly available, wants a **reliable prediction about test scores in that district**, based on data to which he has access. The father does not need to know the causal effect on test scores of class size—or, for that matter, of any variable. What matters to him is that the prediction equation estimated using the California districtlevel data provides an accurate and

Internal and External Validity When the Regression Is Used for Prediction II

reliable prediction of test scores for the district to which the father is considering moving.

Reliable prediction using multiple regression has three requirements.

- The first requirement is that the data used to estimate the prediction model and the observation for which the prediction is to be made are drawn from the same distribution. This requirement is formalized as the first least squares assumption for prediction, given in Appendix 6.4 for the case of multiple predictors. If the estimation and prediction observations are drawn from the same population, then the estimated conditional expectation of Y given X generalizes to the out-of-sample observation to be predicted. This requirement is a mathematical statement of external validity in the prediction context. In the test score example, if the estimated regression line is useful for other districts in California, it could well be useful for elementary school districts in other states, but it is unlikely to be useful for colleges.

Internal and External Validity When the Regression Is Used for Prediction III

- The second requirement involves the list of predictors. When the aim is to estimate a causal effect, it is important to choose control variables to **reduce the threat of omitted variable bias**. In contrast, for prediction the aim is to have an accurate out-of-sample forecast. For this purpose, the predictors should be ones that substantially contribute to explaining the variation in Y, whether or not they have any causal interpretation. The question of choice of predictor is further complicated when there are time series data, for then there is the opportunity to exploit correlation over time (serial correlation) to make forecasts—that is, predictions of future values of variables. The use of multiple regression for time series forecasting is taken up in Chapters 15 and 17.

Internal and External Validity When the Regression Is Used for Prediction IV

- The third requirement concerns the estimator itself. So far, we have focused on OLS for estimating multiple regression. In some prediction applications, however, there are very many predictors; indeed, in some applications the number of predictors can exceed the sample size. If there are very many predictors, then there are—surprisingly—some estimators that can provide more accurate out-of-sample predictions than OLS. Chapter 14 takes up prediction with many predictors andexplains these specialized estimators.

References I

Stock, J. H. and Watson, M. W. (2020). *Introduction to econometrics*, volume 4. Pearson New York.