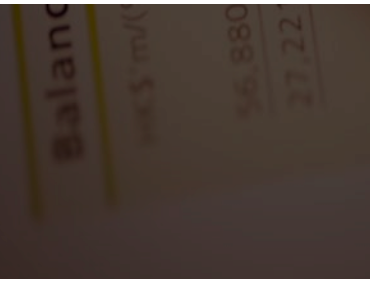
Economics Collection
Philip J. Romero and Jeffrey A. Edwards, *Editors*

Building Better Econometric Models Using Cross Section and Panel Data

Jeffrey A. Edwards



businessexpert Press
www.businessexpertpress.com

Building Better Econometric Models Using Cross Section and Panel Data

Building Better Econometric Models Using Cross Section and Panel Data

Jeffrey A. Edwards

businessexpert Press

First published in 2014 by
Business Expert Press, LLC
222 East 46th Street, New York, NY 10017
www.businessexpertpress.com

ISBN-13: 978-1-60649-974-0 (paperback)
ISBN-13: 978-1-60649-975-7 (e-book)

Business Expert Press Economics Collection

Collection ISSN: 2163-761X (print)
Collection ISSN: 2163-7628 (electronic)

Cover and interior design by Exeter Premedia Services Private Ltd, Chennai, India

First edition: 2014

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America.

This book is dedicated to my lovely wife, Catherine, who has endured everything I've done in my life to get to where I am today. It is also dedicated to our feline children, Dora, Hope, Snaggle, and Weezie, whose antics keep us laughing.

Abstract

Many empirical researchers yearn for an econometric model that better explains their data. Yet these researchers rarely pursue this objective for fear of the statistical complexities involved in specifying that model. This book is intended to alleviate those anxieties by providing a practical methodology that anyone familiar with regression analysis can employ—a methodology that will yield a model that is both more informative and is a better representation of the data.

Most empirical researchers have been taught in their undergraduate econometrics courses about statistical misspecification testing and respecification. But the impact these techniques can have on the inference that is drawn from their results is often overlooked. In academia, students are typically expected to explore their research hypotheses within the context of theoretical model specification while ignoring the underlying statistics. Company executives and managers, by contrast, seek results that are immediately comprehensible and applicable, while remaining indifferent to the underlying properties and econometric calculations that lead to these results.

This book outlines simple, practical procedures that can be used to specify a better model; that is to say, a model that better explains the data. Such procedures employ the use of purely statistical techniques performed upon a publicly available data set, which allows readers to follow along at every stage of the procedure. Using the econometric software Stata (though most other statistical software packages can be used as well), this book shows how to test for model misspecification, and how to respecify these models in a practical way that not only enhances the inference drawn from the results, but adds a level of robustness that can increase the confidence a researcher has in the output that has been generated. By following this procedure, researchers will be led to a better, more finely tuned empirical model that yields better results.

Keywords

cross-sectional data, inference, misspecification testing, panel data, regression, regression models, respecification, Stata, statistical adequacy

Contents

Preface

As the title states, this book is about using relatively easy-to-perform methods in order to find a better econometric model—but methods that don't rely on any theory that is specific to a particular field. The methods employed in this book rely entirely upon the statistical assumptions underlying any ordinary least squares (OLS) type of econometric model, and, therefore, transcend all disciplines.

Regression analysis is perhaps the most widely used method for evaluating data. It is used in academics, large and small companies, and some may even use it to forecast their household budgets (although these individuals are probably rare). One of the issues some have had over the years, however, is the fact that sometimes their models provide a poor explanation of the data, or generate results that do not make sense within the context of the analyst's respective area. And even if the models these analysts employ do in fact generate results that make sense, good researchers frequently wonder whether they can generate even better results with regard to the robustness of the estimates, and the overall fit of the model to the data. This is where this book tries to help out.

This book starts with a basic outline of the concept of statistical adequacy and what it means within the context of an econometric model. And even though this topic is reviewed within the realm of OLS regressions, all models whether they are ordered logistic, probit, and so on, employ similar probabilistic assumptions that researchers shouldn't ignore.

Basically speaking, statistical adequacy of an econometric model means that in the finite sample, the model embodies what it needs to in order to satisfy the underlying distributional properties imposed upon it, a priori, by the researcher. It is all too common for analysts to ignore these properties and simply run regressions. Their excuse is usually that the central limit theorem absolves all of their responsibility toward attaining statistical adequacy in the finite sample because everything's normal at the limit, right? But what they fail to realize is the fact that for years scientists have constructed finite sample misspecification tests for these

probabilistic assumptions. The reason is that if the assumptions are satisfied in the finite sample, there is no need to employ the central limit theorem in this context. In other words, if a distribution is normal in the finite sample, of course it's normal at the limit! From this then, it seems to make sense that if the assumptions are satisfied in the finite sample, the results the model generates will be more trustworthy than the results generated by a model that relies solely upon the properties of the central limit theorem. The critical dichotomy in thought here is the idea that the latter group of scientists actually realize that no one ever has an infinite number of observations, nor will anyone ever approach an infinite number of observations. The former group of scientists must then live on another planet, one for which infinity is less a concept and more a reality. But here on the Earth, infinity is still just a concept.

From this area of discussion, I generate baseline results using cross-section and panel models that the reader can also employ with publically available data from the World Bank. In fact, throughout the book, the same data, same cross-sectional model, and panel model will be used letting the reader follow along with their own econometric software; I, however, employ the use of Stata. We then pursue misspecification testing and respecification for the cross-sectional model in the first half of the book, and panel model misspecification testing and respecification in the second half of the book. Since it is rare that a researcher simultaneously uses cross section and panel models in the same research piece, each of these sections are written in a way that is mostly independent from the other. For the most part, if the reader is using a panel model, they will not have to read the chapters on cross-sectional specification in order to interpret what is said in the panel section—they can just jump straight from Chapter 4 to Chapter 7. And lastly, within each section, if misspecification is found to be present the models are respecified accordingly and a discussion takes place that compares the newly respecified model to the baseline results. In the end, we will have a "better" model than we started with.

Having said all this, the purpose of this book isn't necessarily to fully attain statistical adequacy in the finite sample. In fact, it's unlikely that any empirical researcher can achieve complete adequacy especially if their data is sociologically generated data collected from multiple sources such as most economic data. But my purpose in this book is to rely upon the

concept of statistical adequacy in order to come as close as possible to satisfying our underlying probabilistic assumptions. To that end, this book serves its purpose; and hopefully, the researchers who read this book will find that it does as well. I sincerely hope you enjoy reading it as much as I enjoyed writing it!

CHAPTER 1

What Is a Statistically Adequate Model and Why Is It Important?

For an ordinary least squares (OLS)-type regression, the researcher assumes that the errors are normally, identically, and independently distributed (NIID). In practice, it is the residuals (the estimated errors) that must satisfy these assumptions in order for the researcher to draw valid inference from the results of their regression. In other words, if the residuals are not NIID in the finite sample, the model is considered to be misspecified, and at least some of the inference drawn from the results cannot be trusted. Sometimes it is the standard errors of the coefficients that are biased, and sometimes it is the coefficients themselves, or both. Either way, any conclusions made from the results are likely to be tenuous at best.

Many would argue that the central limit theorem absolves any responsibility from the researcher to assure that a model is statistically adequate; after all, everything is normal at the limit, right? First, no its not. A non-normal distribution is non-normal period! Secondly, the central limit theorem only holds under certain conditions—conditions that can be violated. Thirdly, researchers like me have shown that assuming asymptotic normality produces substantially different results than assuming normality in the finite sample (Edwards and McGuirk 2004; Edwards et al. 2006). Lastly, no one ever has an infinite number of observations. Think about it, if you could *approach* infinity, just how far from infinity would you still be? You would still be an infinite number of observations away from infinity!

But if normality is established in the finite sample then one does not have to rely on the central limit theorem. This is why for many years econometricians have developed a variety of ways to test for NIID in the

finite sample; unfortunately, these tests are usually ignored by empirical researchers even though all of them learned these testing methods in their undergraduate days.

It certainly seems to be common sense that if NIID is established in the finite sample, the inference you draw from your results are more reliable and robust, giving your research far more validity (Edwards et al. 2006; Spanos 1999; McAleer 1994; McAleer et al. 1985). So why don't more researchers make conscious efforts to satisfy the NIID assumptions in finite samples? This is a good question that has four possible answers: (1) they do not know any better, which is a scary thought in itself; (2) it is simply too difficult, time consuming, or both, to test for mispecification; (3) if mispecification is discovered, it is too difficult to respecify the models; or (4) they are afraid that the results from their respecified model would not support their research hypotheses. To the extent that I do not want to lose complete faith in the empirical research community, I will go with the excuse that most researchers do not pursue more statistically adequate models because it is too difficult to respecify them and actually attain NIID residuals, and the models that one may end up with may not make *sense* given the researcher's objective. Let me clarify this statement.

Some data sets are fairly easy to work with, are relatively homogeneous in their observations across individuals or over time, or both, and are collected by relatively few sources allowing someone to more easily justify merging them together. Other data sets are the complete opposite. Some are collected from hundreds, if not thousands of different sources, each with seemingly different collection criteria even when collecting the same variable(s). Some have extreme levels of heterogeneity both over time and across individuals. Some are simply so poorly constructed that they are almost impossible to work with, much less model. And some data sets have all three issues. In essence, when modeling some data sets, one may not be able to attain NIID in the finite sample, no matter what the specification of their model is. This may be a good reason why so many researchers simply regress (pardon the pun) to the simple linear model and ignore any misspecification issues. However, this is not a wise thing to do because the closer one comes to NIID, even if not exactly conforming to all of the NIID assumptions, the more robust and reliable

the inference will be that they draw from their results. This is what we set out to do in this book, beginning with this chapter. We will do our best to achieve NIID given our finite sample data sets; but in the end, even if we do not quite get there, our results will be more reliable than they otherwise would have been had we just ignored our probabilistic assumptions in the first place.

To the extent that I do not want to get too technical in this book and simply outline a practical methodology to attain a model that more accurately reflects the data, I will henceforth focus on a list of typical misspecification issues. Correcting these problems may or may not make your model satisfy *all* of the underlying NIID conditions, but it will allow you to have far more confidence with the inference you draw from your results. In addition, the reader can easily test for these using most econometric software packages as the simpler tests are usually canned in the package as prewritten commands that can be engaged simply by typing one or two words; and if they are not canned, the construction of the uncanned tests are also quite easy.

A more sophisticated reader will notice that the paragraph they just read gives the impression that we will be testing for misspecification. I should probably clarify what I mean as one never tests for misspecification, he or she actually tests for correct specification. For instance, one never tests for heteroskedasticity, he or she tests for homoskedasticity; therefore, for a test such as this, homoskedasticity would be the null hypothesis. If the null is not satisfied, then it is *possible* that heteroskedasticity exists. So, when someone tests for misspecification, he or she is actually testing for correct specification; any test results that come back showing that the modeler does not have the correct specification, then the modeler can assume that the misspecification *might* lie in the area of the probabilistic assumption of that specific test. Having said this, in the context of this book and its purpose, which is to provide the reader with basic methods both to find misspecifications and correct for them, I focus on the (possible) misspecification instead. I have found that students of empirical work better understand deviations from probabilistic assumptions and how to correct for them when put in this context. Therefore, I do hope that the expert will ultimately allow (or better yet, ignore) this rhetorical faux pas of mine throughout the remainder of the book.

Types of Misspecification Covered in This Book

The basic misspecification issues that we will focus on are as follows:

1. Heteroskedasticity
2. Intercept heterogeneity
3. Dependent variable dynamics in panel data
4. Slope heterogeneity
5. Statistical omitted variable bias

For the more advanced reader and modeler, we can add to this list the following issues:

6. Variance heterogeneity
7. Consistent and balanced panels
8. Dynamic parametric heterogeneity

Misspecifications 1 through 5 are the most important for empirical research. If one's objective is to draw fairly reliable inference but not get into the quagmire of sophisticated econometric techniques, the bare minimum of misspecification issues one should correct are these. A less novice researcher may find excluded from the misspecification issues listed earlier endogeneity between x and y, dependence among the residuals, and spatial dependence—these omissions are not by accident.

One of the underlying assumptions made in regression analysis is that any variable that lies on the right-hand side of the equation causes y to change (or not) and not the other way around; in other words, x determines y (this is where the word *determinant* comes from). However, if y causes any of the right-hand side variables to change, then we say the relationship between x and y is endogenous implying that information contained in y feeds back into the right-hand side. In this book, endogeneity is not addressed as a theoretically anticipated feedback from y to x; but it is addressed by misspecification (3) as a purely statistical issue when performing dynamic regressions using panel data.

In the theoretical case, if we assume that y does feedback to x, we must get that assumption from some theoretical relationship we anticipate

must exist *given that particular data set*. For instance, if we run a regression of the growth rate in gross domestic product (GDP) for the United States on gross domestic investment in the United States, an underlying assumption is that investment *causes* growth. However, it can also be argued that increased growth will in turn cause firms to invest more. This is a purely theoretical argument and not a statistical one.

In this book, I only focus on the statistical aspects of finding a better model and to a great extent avoid theoretical implications. Having said all this, I will actually use models that already satisfy the exogeneity assumption by lagging the variables in the conditioning set by one period. This is a common way of circumventing endogeneity when using longitudinally collected data. The cross-sectional model will utilize averages of two time spans over years of data. I use more recent year-span averages for the left-hand-side variable and less recent year-span averages for the right-hand-side variables. The panel model will use one-year lags of all variables on the right-hand side. This variable construction procedure will necessarily correct any endogeneity issues simply because an event that happens today cannot cause something to occur yesterday. Therefore, by definition our right-hand-side variables must *cause* our left-hand-side and not the other way around.

With regard to error and spatial dependence, we partially address the former (again in misspecification (3)), and simply ignore the latter. We address error dependence by including dependent-variable dynamics into our panel regression model. To go much beyond that would involve a level of sophistication that exceeds the intentions of this book—that is, quick and easy ways to attain a better model. The same argument is applied to spatial dependence. However, the suggested readings at the end of this chapter will help those interested in these two particular misspecification issues.

Suggestions for Further Reading

Readers interested in furthering their knowledge of statistical adequacy in econometric modeling should read:

Edwards, J.A.; and A. McGuirk. "Statistical Adequacy and the Reliability of Inference." *Econ Journal Watch* 1, no. 2 (August 2004), pp. 244–59.

Edwards, J.A., A. Sams; and B. Yang. "A Refinement in the Specification of Empirical Macroeconomic Models as an Extension to the EBA Procedure." *The BE Journal of Macroeconomics* 6, no. 2 (October 2006), pp. 1–26.

McAleer, M. "Sherlock Holmes and the Search for Truth: A Diagnostic Tale." *Journal of Economic Surveys* 8, no. 4 (December 1994), pp. 317–70.

Spanos, A. *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge: Cambridge University Press, 1999.

Readers interested in the basics of misspecification issues, including spatial dependence, should read:

Anselin, L. *Spatial Econometrics: Methods and Models*. New York, NY: Springer Publishing, 1988.

Gujarati, D.N. *Basic Econometrics*. New York, NY: McGraw-Hill Publishing, 2003.

CHAPTER 2

Basic Misspecifications

In explaining each of the misspecification issues 1 through 5, I start from a basic model of the form

$$y = a_0 + a_1 x + e \tag{2.1}$$

Typically the errors, e, are assumed NIID. However, if any one of the misspecifications 1 through 5 exist, some area of this assumption is violated. It is these violations that we will identify and, if encountered, will correct for. Correcting for these basic forms of misspecification will allow the empirical researcher to draw more accurate inference from their results and, therefore, have more confidence that the outcomes they reach are some of the best attainable.

The order with which I describe, test, and correct for each misspecification is not by accident. The ordering of these misspecification issues is one that I've personally developed, and one that is most likely to result in an adequately specified model with the least amount of effort. Many misspecifications can mask themselves in ways that lead the researcher away from the true cause. For instance, a quadratic relationship could potentially mask itself as slope and intercept heterogeneity (issues discussed in length later). In other words, one could test for a quadratic relationship and find that one exists according to the test used; but the actual misspecification issue at hand is a totally different one from what the test was supposed to catch. Furthermore, respecifying a model may actually produce misspecification in another probabilistic assumption. In essence, finding an adequate model can sometimes become a circular process and be quite frustrating at times. The order I outline in this book is one that I've found reduces (although doesn't eliminate) this frustration.

Heteroskedasticity

One of the implications of identically distributed errors is that the conditional variance of y is not a function of x. If it is, we have heteroskedasticity; if it isn't, we have homoskedasticity. Mathematically, homoskedasticity is represented by

$$E(e^2) = \sigma^2 \tag{2.2}$$

where heteroskedasticity is represented by

$$E(e^2) = \sigma^2(x) \tag{2.3}$$

The reader will notice a slight deviation here from what they probably learned earlier in their academic career as there is no subscript on e or x. A heteroskedastic variance is different from a heterogeneous variance. The former is a function of x, whereas the latter is a function of some deterministic change in σ that may or may not be a function of x. We address variance heterogeneity later in this book.

An easy way to spot if heteroskedasticity might be present is to plot the absolute value of the residuals, r, over increasing values of x. (If one remembers from undergraduate econometrics courses, the residual is the estimated error. Therefore, the residual is the empirical representation of the error, and hence the variable one uses to test for misspecification.) Heteroskedasticity exists if there is an obvious correlation between $|r|$ and x. Figure 2.1 shows a homoskedastic relationship and Figure 2.2 a heteroskedastic one. The reader should keep in mind, however, that the heteroskedastic plot reflects only one version of this phenomenon. But any correlation or pattern other than no correlation would be indicative of heteroskedasticity.

There are many quantitative tests for homoskedasticity; however, most researchers these days don't even test for it, they simply assume heteroskedasticity exists and by default correct for it. They do this mostly because it is easy and the procedure for correction is built into all programming packages and usually consists of adding a word at the end of the regression's programming line. (For instance, in Stata a researcher would simply

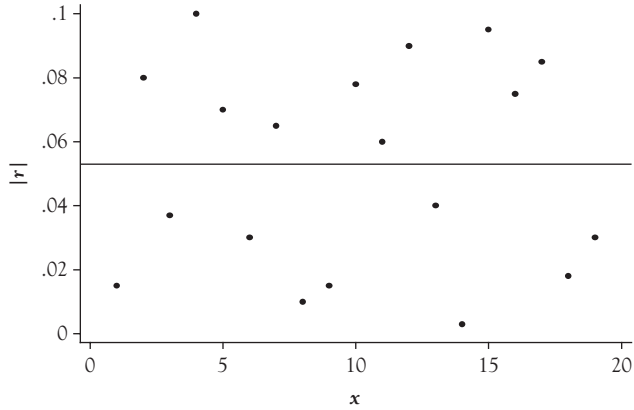


Figure 2.1 Homoskedastic residuals scattered uniformly across x

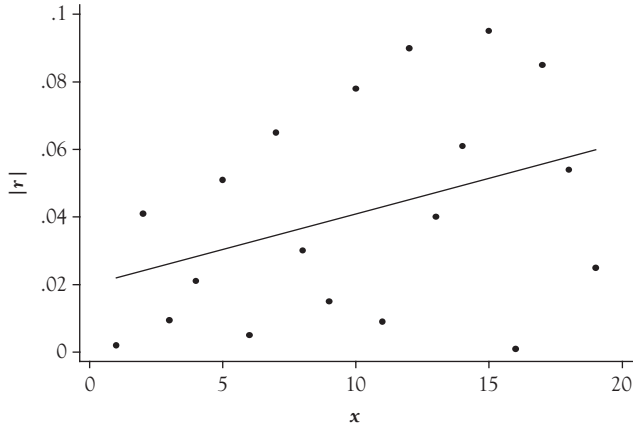


Figure 2.2 Heteroskedastic residuals getting larger with x

add the word "robust" to the end of the line containing the regression command.) Their argument for automatically applying the correction procedure is that if the errors don't need correcting, then the respecification procedure used to correct for heteroskedasticity will not create substantial problems when drawing accurate inference from the coefficient estimates. At worst, correcting for heteroskedasticity when none is present will err on the side of an insignificant finding resulting in no statistically significant relationship between x and y (at least in practice). On the other hand, one can easily argue that correcting for something

that isn't there seems a flawed concept in itself and doesn't coincide with conducting "good econometrics."

Intercept Heterogeneity

Another concept of identically distributed errors is the idea that if we cluster or order the errors in "some meaningful way" they don't reveal any obvious patterns. An intercept with no heterogeneity would be mathematically represented as

$$a_0(it) = a_0 \tag{2.4}$$

where i represents some obvious clustering of the data, and t represents a time dimension within each i. On the other hand, if heterogeneity did exist we would have

$$a_0(it) = a_{i0} \tag{2.5a}$$

or

$$a_0(it) = a_{t0} \tag{2.5b}$$

where (2.5a) reflects clustering heterogeneity, and (2.5b) reflects time heterogeneity. Heterogeneity of the form in (2.5a) can exist when using either cross section or panel data, whereas that of the form (2.5b) can only exist with panel data. Of course, we could have both simultaneously if panel data is used. If not explicitly modeled, each of these phenomena would be identified as either shifting or trending residuals when ordered over clusters, time, or both.

When modeling intercept heterogeneity, we have to remember that the residual is all the information contained in y that is not accounted for by whatever is on the right-hand side. Potentially, there are an infinite number of variables that can be included to fully explain y. Think about it, for any given complex variable like economic growth, would we ever be able to run a regression of growth on a set of variables and get an adjusted

R^2 = 1.00? I think not. This is because not all explanatory variables are available to the researcher at the time the research is being conducted. However, there are variables that can be *constructed* by the researcher with very little effort. These are dummy variables and time trend variables; and it is these variables that we use to attempt to pick up patterns in meaningfully ordered residuals. As we probably already know, a dummy variable is a binary, zero-one variable that takes the value 1 for some characteristic of y and 0 otherwise. A time trend variable is a variable that is ordinal on a discreet interval, say, 1, 2, 3, and so on, within each i.

And while the construction of the time trend variable is obvious, clusters are more complicated. Examples of such clustering could be gender, regions in the United States, types of manufacturing, race, and so on. Even continuous variables, such as income, population growth rates, countries' development levels, and so on, can be made discreet and used as clustering variables as long as these clusters make intuitive sense. As an example, while income data is usually continuous, we could cluster this data to reflect high income, middle income, and low income. Whereby the development of a nation is usually determined by its per capita wealth (or some proxy of it), which is also a continuous variable, we could cluster countries by developed, developing, and emerging economies by delineating each at some predetermined level of wealth. But one thing is for sure, not accounting for clustering can make any inference drawn from an estimation, tenuous.

To make the concept of intercept heterogeneity clearer, especially in the case of clustering, assume that we run a regression of economic growth on domestic investment for all of the states in the United States and Mexico. Also assume that our data weren't sorted by country. Since a regression minimizes the sum of the squared residuals without reference to heterogeneity, because it's the researcher who must define likely areas of heterogeneity, our plot of the residuals over the individual states, i, would probably look like those in Figure 2.3. Notice how all of the residuals are nicely centered around zero and their spread is fairly random. But this is not the case if we sort these same residuals by country. Figure 2.4 tells us that when the residuals are sorted by country, conditional mean growth rates in the United States were underestimated, and those for the Mexican states were overestimated! Hence, inference drawn from the model that

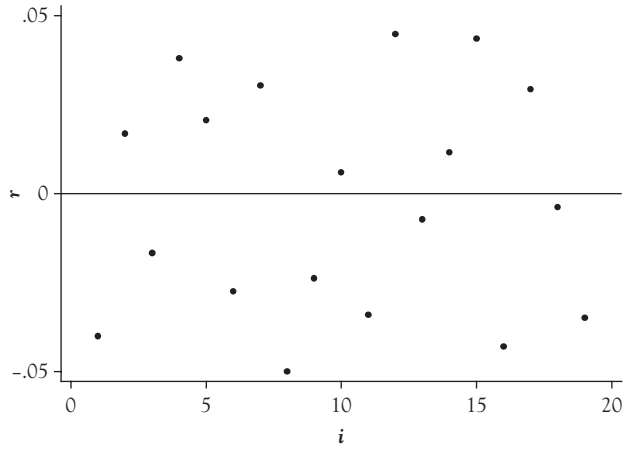


Figure 2.3 Unsorted residuals that appear to be homogeneous

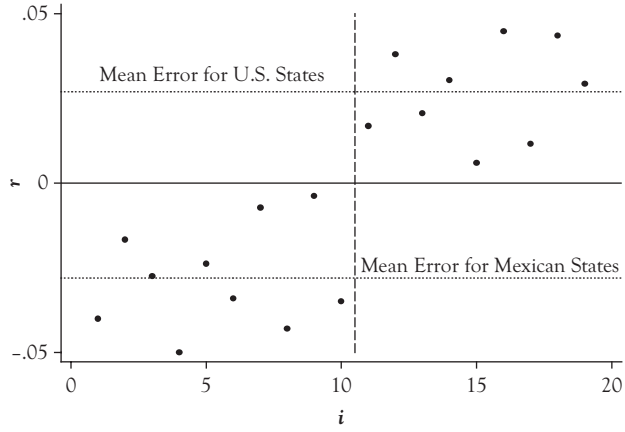


Figure 2.4 Same residuals but clustered by country

produced Figure 2.3 would not represent the data as well as a model that accounted for the variation in the residuals like in Figure 2.4.

Dependent Variable Dynamics in Panel Data

A very common issue that exists when using panel data are time dynamics within the left-hand-side variable. Whatever your "dependent" variable is, as long as it is continuous (it doesn't have to be however, but we won't go into that here) and has a relatively short time interval between

observations, it is usually safe to assume that this period's observation of y is dependent upon last periods' observation of y. Mathematically, this would be depicted as

$$y_{it} = a_0 + a_1 y_{it-1} + e_{it} \tag{2.6}$$

Alternatively, we would see it in a plot of y like in Figure 2.5. Look familiar? This is what a business cycle looks like (of course this is a perfect business cycle as the data was generated; a typical business cycle would not be as "smooth" so to speak). Since so much data, economic or not, is influenced by business cycles, it makes sense that if there is a time dimension to your data, it will probably exhibit this pattern. And since the errors are simply the information in y not accounted for by the right-hand side, it is easy to see that the violation that would occur is of the independence assumption—that is, the first I in NIID.

In theory, this sort of misspecification issue does not produce bias in the slope coefficient estimates, only their standard errors and, therefore, inference drawn from them. Having said that, if what is occurring is indeed business cycle dynamics, an argument could be made that without holding the short-run dynamics of y constant, the effect that x has on y could be muddled as you would be including both short- and long-run effects in x's slope coefficient.

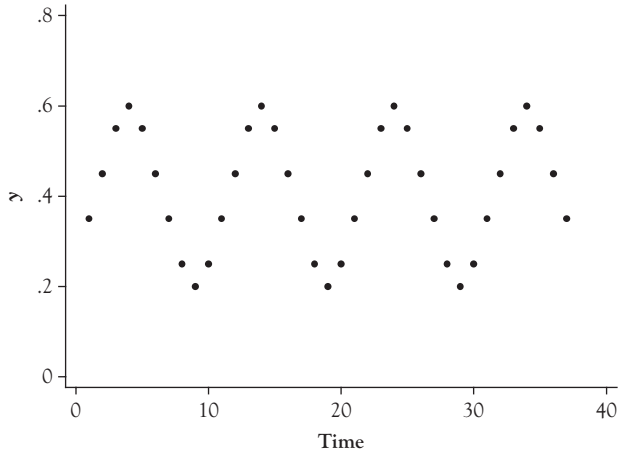


Figure 2.5 Dependency in y over time

Slope Heterogeneity

Like intercept heterogeneity whereby the intercept changes as a function of something that is structurally determined, slope heterogeneity is when the slope coefficient, a_1 in equation (2.1), changes because of some structural characteristic. These characteristics are exactly the same as those possible for intercept heterogeneity, such as gender, region, income level, and so on. This type of heterogeneity can be mathematically represented as

$$y = a_0 + a_1 x + a_2 Dx + e \tag{2.7}$$

where D is a dummy variable representing the predetermined clustering of the data that the researcher chooses. In many cases, both slope and intercept heterogeneity must be modeled. If this is the case, (2.7) would look like

$$y = a_0 + a_1 D + a_2 x + a_3 Dx + e \tag{2.8}$$

To get a visual idea of slope heterogeneity and the misleading inference drawn from it, assume we have two variables, x and y, as depicted in Figure 2.6. In this plot, there is an obvious positive correlation between x and y; the estimated relationship we see in Figure 2.7. However, if y could be clustered into two groups, for instance, we could actually have

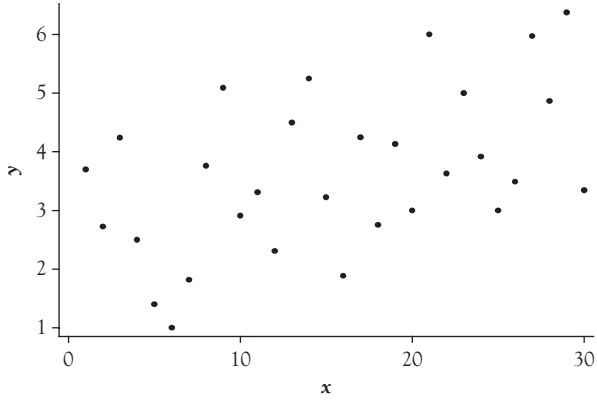


Figure 2.6 Scatter plot showing positive relationship between x and y

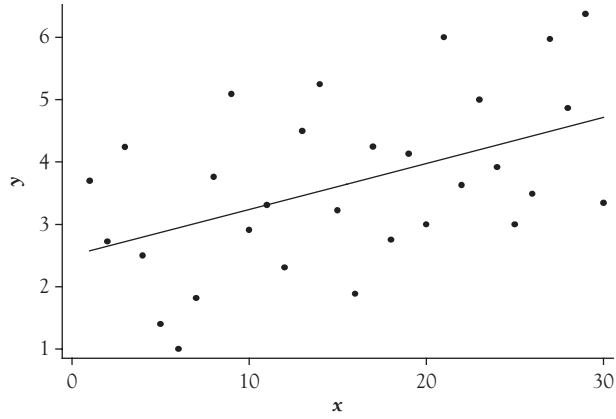


Figure 2.7 Same relationship but with estimated regression line

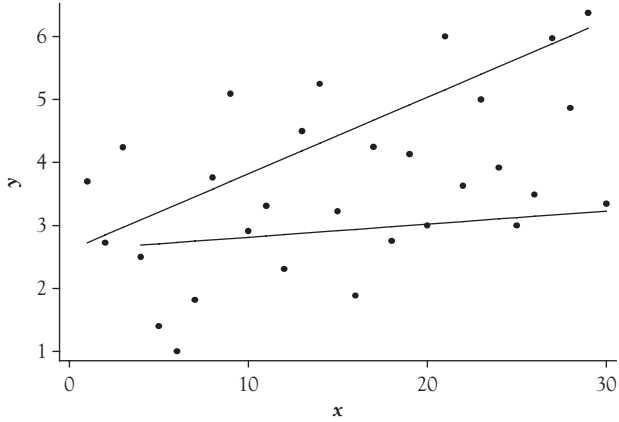


Figure 2.8 Same relationship but accounting for slope heterogeneity

the relationship depicted in Figure 2.8. One can obviously see that the depiction in Figure 2.8 is considerably different than that in 2.7. In fact, it could be the case that the effect *x* has on *y* for one group is significant, but that for the other group is not—emphasizing the need to control for different slopes across clusters. To this end, it behooves the researcher to investigate possible clustering to account for any slope heterogeneity that may be present in the data.

Statistical Omitted Variable Bias

Omitted variable bias will take two forms—a theoretically driven bias, and a statistically driven bias. Assume we have the functional form

$$y = a_0 + a_1 x + a_2 z + e \tag{2.9}$$

One of the fundamental properties of an OLS-type regression is that the errors are not correlated with the other variables on the right-hand side. Theoretical omitted variable bias (a term I coined) is when z is not actually modeled like it is in (2.9), but is contained in e and is correlated with x; but z must also be a completely *different* variable from x. An example of two such variables would be housing starts in the United States and yearly rainfall amounts. If, for instance, lumber prices were regressed on housing starts (i.e., lumber prices on the left and housing starts on the right), the coefficient representing the effect housing starts has on lumber prices would be biased if rainfall amounts were not also included in the data (I'm obviously ignoring the endogeneity as well as the supply and demand aspects of this argument—this example is solely for exposition purposes). This is because rainfall amounts are obviously correlated with housing starts (can't start a house if it's raining a lot), and with lumber prices (can't timber land if it's raining a lot). But this argument implies that rainfall should be included only because there exists a purely *theoretical* reason that it should be included.

For another example on a macro level, assume y is growth in GDP for a broad cross section of countries, and x is foreign direct investment for these countries. It could certainly be argued that economic growth in an economy is affected by that country's political stability; the more unstable a country's political system is, the lower its growth rate should be. The same argument holds for foreign direct investment as outside investors are unlikely to send much capital to an economy if their political structure is in disarray—there would be too much risk involved in such an investment. Therefore, to get an accurate picture (or unbiased picture) of the effect that foreign investment has on growth, one must control for political instability as well; hence, according to this purely theoretical argument, z should be included.

This book makes no theoretical assumptions on these grounds as they can only be made case by case (i.e., depending upon what the dependent variable and the conditioning set is). Furthermore, this argument could be made for all models that have any amount of error! Couldn't it always be the case that an argument exists for the inclusion of some z-variable? Because of this, we only approach the idea of omitted variable bias as a statistical concept and not a theoretical one.

Statistical omitted variable bias (another term I coined) exists when z is a direct function of x. In the typical case, this is when

$$Z = x^2 \tag{2.10}$$

It is obvious that if (2.10) holds and (2.9) is the true model, but instead we run the regression without z, we will be leaving out a variable that is directly correlated with x and y. Now its not the case that we couldn't construct a theoretical argument why the relationship between x and y should be a quadratic one (think of a production function for example), but there doesn't have to be such an argument. For instance, assume we have the relationship between x and y as depicted in Figure 2.9. Now assume we run a regression without z; in other words, an additively linear regression of the basic form of (2.1) from earlier, that is

$$y = a_0 + a_1 x + e \tag{2.11}$$

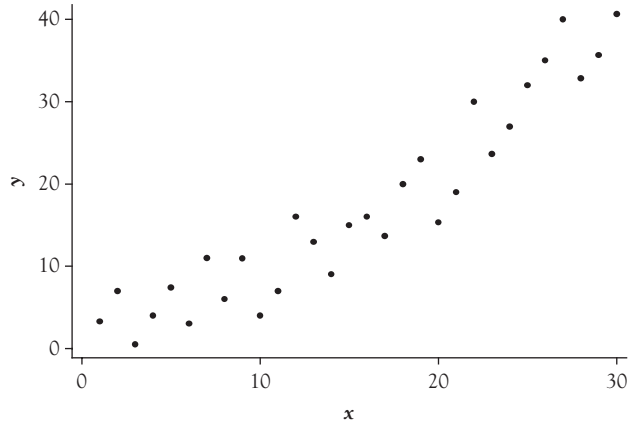


Figure 2.9 Scatter plot of nonlinear relationship between x and y

If we did this we would get the estimated relationship as in Figure 2.10.

But this would clearly be an inaccurate regression line. The observations in the center are below the line, while those at the ends are above the line (generally speaking). A more accurate regression model to estimate would be

$$y = a_0 + a_1 x + a_2 x^2 + e \tag{2.12}$$

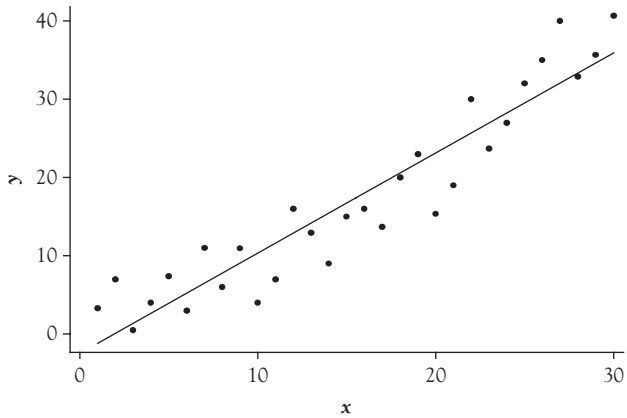


Figure 2.10 Estimated linear relationship between x and y

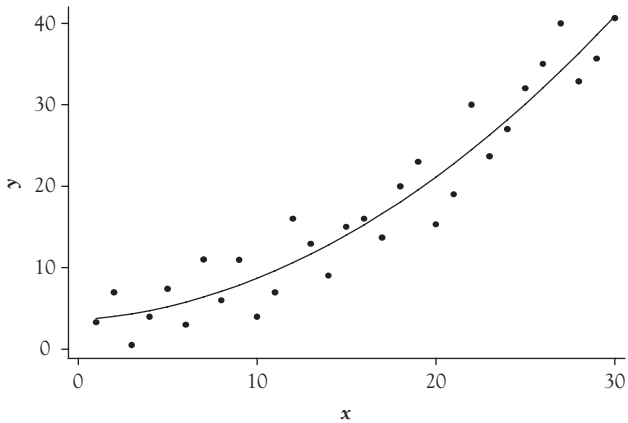


Figure 2.11 Estimated nonlinear regression line

If we fitted this line to the scatter plot we would get Figure 2.11. Obviously, regression model (2.12) is a more accurate reflection of the relationship between x and y. Did we have to make any theoretical arguments to come to this conclusion? Certainly not—that's why I call this *statistical* omitted variable bias and differentiate it from other forms of omitted variable bias.

CHAPTER 3

Misspecifications for the More Advanced Reader

As implied earlier, the next three topics are not for the faint of heart but often rear their ugly heads in most regression-based studies. The problem is that most researchers either don't know these issues could exist, or if they believe these issues can exist they don't know how to check for them outside of time-series analyses, and they certainly never correct for these problems if they believe they can exist and know how to check for them.

Variance Heterogeneity

Variance heterogeneity is a strange beast. It is perhaps the most widely misinterpreted and misrepresented misspecification issue in the history of econometrics. Here's why. From earlier we know that heteroskedasticity is represented as

$$E(e^2) = \sigma^2(x) \tag{3.1}$$

Implying that the conditional variance is a function of x. If you also remember, the subscript was purposely left off to indicate that this is *purely* a function of x and nothing else. On the other hand, variance heterogeneity takes the form

$$E(e^2) = \sigma^2(i, t) \tag{3.2}$$

Hence, variance heterogeneity means that the conditional variance is a function of the y variable's subscript. Like intercept heterogeneity, this implies that the conditional variance is a function of some sort of clustering, ordering, or both of the squared residuals.

Imagine a case whereby we are regressing GDP growth on a set of determinants for a broad cross section of countries. Eventually the question should be asked, do growth rates vary more in developing countries than they do in developed countries? Of course they do. But this DOES NOT mean that the variation in growth is a function of x; on the contrary, it is a function of a dummy variable that would equal one for developing countries, and zero otherwise. The variance can also be a function of time. Have growth rates become more or less volatile since the great depression? I think it would be reasonable enough to argue that on average, they have become less volatile, possibly in a fairly linear and deterministic way. Of course, both of these hypothetical arguments would have to be tested empirically, but the reader should get the gist of what I am saying.

To make the comparison to heteroskedasticity clearer, assume we have the model (2.1) from Chapter 2 as before, repeated in the following:

$$y = a_0 + a_1 x + e \tag{3.3}$$

We estimate this model and obtain the residuals, r. Roughly speaking, we could test for heteroskedasticity by running the following regression

$$r^2 = b_0 + b_1 x + u \tag{3.4}$$

If $b_1 = 0$, then $r^2 = b_0$, which is a constant, meaning heteroskedasticity is not present. To test for variance heterogeneity, we would run the regression

$$r^2 = c_0 + c_1 D + \mu \tag{3.5}$$

whereby D is a dummy variable representing either a clustering of i or a time-dependent structural change. In this case, if c_1 is significant, it would indicate that we do have heterogeneity in the conditional variance. But unless this D is in the group of variables making up x, equation (3.4) would never pick up what equation (3.5) found; this means that tests for heteroskedasticity would never pick up this phenomenon and, therefore, the standard heteroskedasticity correction procedure would never correct for it!

The vast majority of researchers continue to believe that heteroskedasticity testing and subsequent correction procedures necessarily address variance heterogeneity as well. They think this, believe it or not, because the subscript for *x* tends to be the same subscript for *y*. Unfortunately, sophisticated graduate programs teaching econometrics have not done a good job deciphering the two for their students. We will do exactly this later in this text.

Consistent and Balanced Panels

Using balanced, consistent panels, or both in panel data regression analysis is a purely conceptual idea of misspecification, but an important one; it is also one that many researchers either overlook because they simply don't think about it being a problem, or ignore it to maximize the number of usable observations in their data set.

Balanced panels simply mean that exactly the same number of periods are covered for each *i*. Balanced panels are critical because to a great extent, panel data regression coefficients reflect the average of individual cross-sectional relationships. For instance, coefficient estimates obtained using panel data covering 6 months, that is, each *i* contains up to 6 observations, to a great extent reflect the average of the cross-sectional relationships between *x* and *y* over that period. Hence, if you were to run six separate monthly cross-sectional regressions, you would get six different values of the slope coefficient. If you were then to take the average of these six values, you would get a value close to the value obtained if you had just estimated the relationship using the entire panel data set. Because of this, it makes sense that the relationship between *x* and *y* should be "averaged" over the same number of periods for each *i*. In an extreme case you could consider two separate relationships between *x* and *y* whereby in one case each *i* has only one period, but for the other each *i* has six periods; obviously, the chance that the relationships are capturing similar "averages" is unlikely.

For a more intuitive example, assume that you are assembling a data set to study unemployment rates in the United States. You wish to regress these unemployment rates on the percentage of welfare aid that each state provides to its citizens as a fraction of state GDP. To increase the

likelihood that your regression results are capturing a similar cyclical average (i.e., a balanced empirical relationship between x and y), a researcher should use the same number of years across states to capture equal lengths of each states' business-cycle component.

Something else to consider would be consistency within this same data set. Obviously, business cycles are going to play an important role in this relationship as both unemployment and welfare aid tend to rise in contractions and fall in expansions. If on average a complete business cycle that covers peak to trough and back again lasts 12 years, then each state should have balance panels and contain 12 years worth of data. Furthermore, one should be consistent by starting that time dimension for each state in the same location—that is, all states should start and end in the same year. In other words, the data for each state would contain not only the same length of business cycle, but exactly the *same* business cycle.

Dynamic Parametric Heterogeneity

And our last misspecification issue that someone may want to address in order to draw more accurate inference from regression results is that of dynamic parametric heterogeneity. This is a misspecification issue that is rarely thought about by researchers; yet could have a huge impact on current dominant paradigms in many empirical disciplines if it was recognized and accounted for in regression modeling.

Normally, researchers assume that all slope coefficients are constant over time. This means that we assume all relationships between x and y are also constant. In other words, researchers typically assume that

$$a_1(t) = a_1 \tag{3.6}$$

But what if the relationship between x and y isn't constant over time? Let's think about what this means. (In 2009, I published a paper on this very topic [Edwards and Kasibhatla 2009]. The outline of this misspecification issue will be largely drawn from that piece of work.) Assume that a researcher estimates economic growth with one variable, for instance, the ratio of investment to GDP. Now assume that the relationship is either

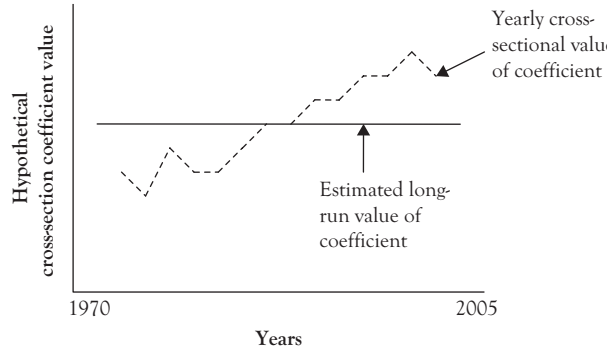


Figure 3.1 A trending coefficient estimate

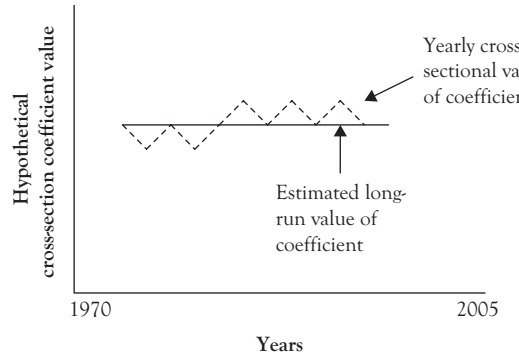


Figure 3.2 A shifting coefficient estimate

trending upward (or downward) over time, or has a structural shift partway through the time period covered. The former is depicted in Figure 3.1, and the latter in Figure 3.2. The graphs show hypothetical yearly cross-sectional estimates for the coefficient of investment, with the average of the estimates depicted by the solid line. Obviously, to base long-run inference on the horizontal line would be problematic. Since many studies tend to use data sets that range over different periods, one can easily see that the estimated long-run value could actually be relatively low or high depending upon whether the data set covers more of the earlier years, or the latter years. Furthermore, as the data set expands into the future, the long-run coefficient value would creep upward. This would imply that drawing inference from a single long-run point estimate may not be wise.

Mathematically, these phenomena can be represented within the marginal effect of the variable in question, for instance, *x*. If our regression is the simple one from earlier

$$y = a_0 + a_1 x + e \tag{3.7}$$

Then the marginal effect of *x* on *y* would be simply

$$\frac{\partial y}{\partial x} = a_1 \tag{3.8}$$

In the former case where the coefficient is trending in a deterministic fashion over time, our regression function would have to take the form

$$y = a_0 + a_1(t)x + e, \tag{3.9}$$

such that the marginal effect now becomes a function of a time trend variable

$$\frac{\partial y}{\partial x} = a_1(t), \tag{3.10}$$

where

$$a_1(t) = b_0 + b_1 t, \tag{3.11}$$

Now, if the marginal effect is not trending in a significant fashion over time, then

$$b_1 = 0 \tag{3.12}$$

and, therefore, the marginal effect is constant, or

$$a_1(t) = b_0 \tag{3.13}$$

and we are back to our original specification in (3.7). In the second case, (3.11) would instead take the form

$$a_1(t) = b_0 + b_1 D \tag{3.14}$$

where D is a dummy variable representing some break in the marginal effect, perhaps delineated by a significant one-time event in the history of the relationship. In this case, of course, we would be testing whether $b_1 = 0$. If it does, we are again back to a constant marginal effect delineated by the simple model in (3.7).

CHAPTER 4

Original Specification and Drawing Inference From It

Two Related Models

This chapter begins our attempt at understanding how to specify a better econometric model. We will be using two types of data—cross-sectional data and panel data. The data were obtained from the World Bank's, World Development Indicators Database, 2013 version, and is downloadable at http://data.worldbank.org/data-catalog/world-development-indicators. It covers 172 countries, and contains unbalanced panels with an average of 21.8 observations per country. These are the data sets that will be used for the remainder of this book; it would definitely behoove the reader to download these data and follow along. A brief word of caution, however. Keep in mind that you may find that as you follow along, not all of your estimates will be similar to mine. This is simply because this is a rather large data set that is continually updated. But, if you download the variables that we use here in their entirety, and follow the instructions I lay out subsequently, your estimates should be close. Regardless, it is the methodology that should interest you most, and not the duplication of the results.

At this point the reader is probably wondering why cross-country macroeconomic data are being used instead of say microeconomic census data or some other data. The reason is simple, I am very familiar with these data as this is my own area of expertise (i.e., growth and development) and it seems as though every econometrics book out there to a large extent uses microeconomic data, so I wanted to change it up some, so to speak. Having said that, it really doesn't matter what data set is used because the procedures are the same regardless of data. This book

addresses specification as a purely statistical phenomena and is applicable to all continuous variables used from any data set.

This chapter focuses on setting up the discussion for the remainder of the book. It probably doesn't do much good talking about misspecification issues and accurate inference if we don't first have base regressions to start from. And even though these base regressions may seem like simple additively linear regressions, which they are, they are perhaps the most widely used. We start with the following cross-sectional regression model:

$$Growth_i = a_0 + a_1 I_{i,-1} + a_2\ FDI_{i,-1} + a_3\ School_{i,-1} + a_3\ Trade_{i,-1} + a_4\ Pop_{i,-1} + a_5\ G_{i,-1} + e_i \quad (4.1)$$

while the panel model we subsequently use takes the form

$$Growth_{it} = a_0 + a_1 I_{it-1} + a_2\ FDI_{it-1} + a_3\ School_{it-1} + a_3\ Trade_{it-1} + a_4\ Pop_{it-1} + a_5\ G_{it-1} + e_{it} \quad (4.2)$$

The reader can quickly tell that the only real difference between the two models is the subscript. The *i* subscript indicates an individual country, and the *t* subscript indicates the time dimension. Obviously, cross-sectional data has no relevant time dimension and, therefore, lacks the *t* subscript. However, as mentioned earlier our base regressions already account for possible endogeneity problems. Therefore, the right-hand side variables in (4.1) are averages calculated over a block of years, which are prior to the block of years used to calculate the average for *Growth*; each calculation also uses about the same number of years for consistency's sake. For instance, if a country has observations that span the 20 years from 1991 to 2010, *Growth* would be a mean value calculated over the years 2001 to 2010, while, for example, foreign direct investment (*FDI*) will be a mean calculated over the years 1991 through 2000. This is why the subscript for the right-hand side variables in (4.1) include a –1 indicating that "last periods" averages were used. Furthermore, if the country covers an odd number of years the mean for the right-hand side variables will include one less year in their calculation than the mean for the left-hand side variable. In the panel data model (4.2), one period lags (a period is

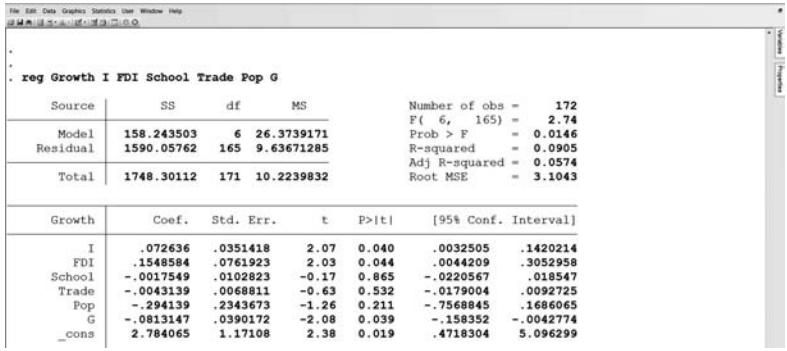
1 year in this case) of each variable are used on the right-hand side and this is represented by the $t-1$ in the subscript.

The variables and respective World Development Indicator codes from left to right are annual growth in per capita real GDP (*Growth*: NY.GDP.PCAP.KD.ZG), gross domestic investment as a percentage of GDP (*I*: NE.GDI.TOTL.ZS), foreign direct investment (inflows) as a percentage of GDP (*FDI*: BX.KLT.DINV.WD.GD.ZS), gross secondary school attendance in percentages (*School*: SE.SEC.ENRR), total trade as a percentage of GDP (*Trade*: NE.TRD.GNFS.ZS), annual population growth rates (*Pop*: SP.POP.GRDW), and general government final consumption expenditure as a percentage of GDP (*G*: NE.CON.GOVT.ZS).

All of the data have been purged of both missing observations and of country groups, such as Arab League Countries, OECD countries, and so on. The World Development Indicators not only gives data on specific countries, but also on predefined groups of countries. Obviously, if we want to avoid the double counting of countries, then we would want to remove all of the predetermined country groups and only use data on the individual countries themselves. As already implied, the cross-sectional regressions will contain 172 observations, while the panel regressions will contain 3,758 observations. The econometric software package used for this book will be Stata. Hence, screenshots of the output will be in Stata as well. However, any econometric software package should be able to conduct all of the tests and respecification procedures we outline in this book although a basic level of programming knowledge for your package will be required.

Base Regressions and Inference

The screenshot of Stata output for regression (4.1) is in Figure 4.1. The first outcome one will notice is the adjusted sample correlation coefficient, also known as the adjusted R^2. Given that this is a cross section of 172 countries (i.e., about 90% of all countries in the world), this simple model only explains about 6 percent of the variation in *Growth*. Neither *School*, *Trade*, nor *Pop* contribute to economic growth because the *p*-values are below 0.100—our cutoff criteria for statistical significance that we will use for the remainder of the book. The other variables, *I*, *FDI*, and

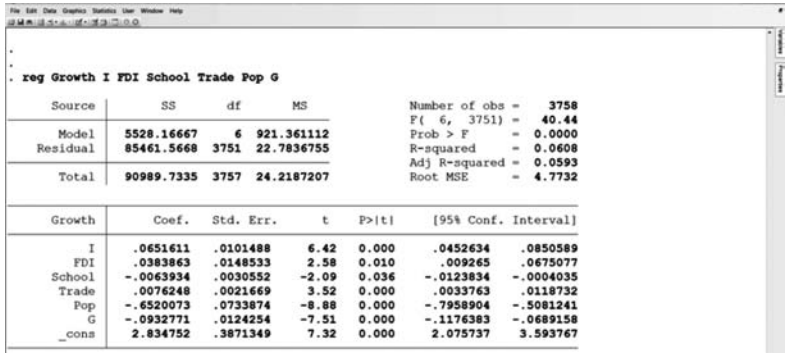


```
. reg Growth I FDI School Trade Pop G

      Source |       SS       df       MS              Number of obs =     172
-------------+------------------------------           F(  6,   165) =    2.74
       Model |  158.243503     6  26.3739171            Prob > F      =  0.0146
    Residual |  1590.05762   165  9.63671285            R-squared     =  0.0905
-------------+------------------------------           Adj R-squared =  0.0574
       Total |  1748.30112   171  10.2239832            Root MSE      =  3.1043

------------------------------------------------------------------------------
      Growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |    .072636   .0351418     2.07   0.040     .0032505    .1420214
         FDI |   .1548584   .0761923     2.03   0.044     .0044209    .3052958
      School |  -.0017549   .0102823    -0.17   0.865    -.0220567     .018547
       Trade |  -.0043139   .0068811    -0.63   0.532    -.0179004    .0092725
         Pop |   -.294139   .2343673    -1.26   0.211    -.7568845    .1686065
           G |  -.0813147   .0390172    -2.08   0.039    -.158352    -.0042774
       _cons |   2.784065   1.17108     2.38   0.019     .4718304    5.096299
------------------------------------------------------------------------------
```

Figure 4.1 Base regression using cross-sectional data



```
. reg Growth I FDI School Trade Pop G

      Source |       SS       df       MS              Number of obs =    3758
-------------+------------------------------           F(  6,  3751) =   40.44
       Model |  5528.16667     6  921.361112            Prob > F      =  0.0000
    Residual |  85461.5668  3751  22.7836755            R-squared     =  0.0608
-------------+------------------------------           Adj R-squared =  0.0593
       Total |  90989.7335  3757  24.2187207            Root MSE      =  4.7732

------------------------------------------------------------------------------
      Growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |   .0651611   .0101488     6.42   0.000     .0452634    .0850589
         FDI |   .0383863   .0148533     2.58   0.010      .009265    .0675077
      School |  -.0063934   .0030552    -2.09   0.036    -.0123834   -.0004035
       Trade |   .0076248   .0021669     3.52   0.000     .0033763    .0118732
         Pop |  -.6520073   .0733874    -8.88   0.000    -.7958904   -.5081241
           G |  -.0932771   .0124254    -7.51   0.000    -.1176383   -.0689158
       _cons |   2.834752   .3871349     7.32   0.000     2.075737    3.593767
------------------------------------------------------------------------------
```

Figure 4.2 Base regression using panel data

G, do contribute to *Growth*. Countries with high levels of *I* and *FDI* tend to have higher growth rates, whereas countries with high levels of *G* tend to have lower growth rates. Again, we do not go into the theoretical reasons for these outcomes and whether or not they jibe with conventional wisdom. All we did here was run a regression on a sample of data and the outcome is what it is. Figure 4.2 shows the panel regression outcome.

The panel estimates are certainly different in magnitude from the cross-sectional estimates, and there are more significant coefficients as well; in fact, all of the variables have a statistically significant effect on *Growth*. The inference for *I*, *FDI*, and *G* is the same as it was for the cross-sectional outcome, but in this case countries with high levels of *School* and *Pop* have lower growth in GDP, but countries with high levels of *Trade* will have higher growth in GDP.

It is at this point that we must make the reader aware of a couple of issues regarding each of the regressions. The first is that we have not corrected for heteroskedasticity, which as indicated in the previous chapter is nearly always automatically corrected for by researchers regardless of whether it is needed or not. However, these are base regressions from which we will test for misspecification and respecify *if needed*. We will not assume that a problem exists when it may not. The second issue is that we did not run a fixed-effects regression with our panel data; fixed-effects regressions are also quite commonly employed from the outset. But again, we are starting from the most basic setup to allow the reader to actually see how different the outcomes will be once a misspecification issue is corrected for. All of these we address in the next chapter.

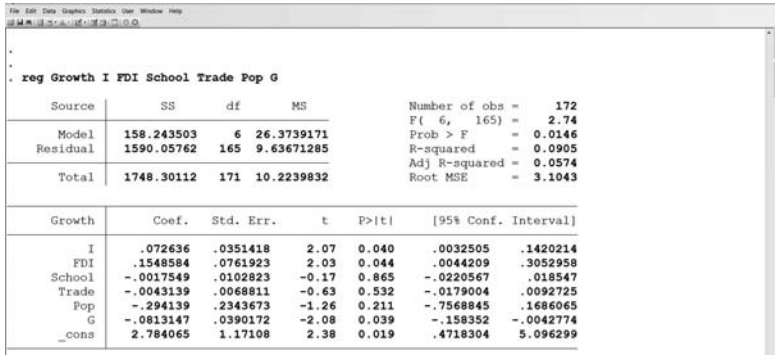
CHAPTER 5

Basic Misspecification Testing and Respecification

The Cross-Sectional Case

In this chapter we test for and correct (if needed) misspecification issues (1), (2), (4), and (5) from Chapter 1 specifically for cross-sectional cases. Issue (3) isn't addressed in this chapter because it only applies to models that use panel data. The algorithm outlined in this chapter (and for the panel models later on in the book) is one developed by myself over years of performing empirical research, analyzing other authors' research, and publishing many articles that address the topic of model specification.

The reason I call the step-by-step process an algorithm is because it can be recursive in nature. Many of the misspecifications listed in 1 through 5 can mask themselves as other issues. In fact, there have been instances in my past empirical work whereby I test for one misspecification issue, find that it exists, and correct for it, but have to go back and address it again after correcting for a completely different misspecification issue! And although we won't specifically address the recursive nature of this process in this book for the sake of brevity, one should not just perform the steps we are getting ready to outline and be done with it; after we think we have a good model, a researcher should reanalyze the early procedures to ensure that something else hasn't reared its ugly head. In addition, not all misspecification can be corrected in just one way. What I mean by this is that some form of misspecification, heterogeneity in particular, may need to be investigated from different angles, not just an angle with a quick fix in mind. More about this will be explained subsequently. For ease of reading and referencing, I have again displayed the base model results from the



```
. reg Growth I FDI School Trade Pop G

      Source |       SS       df       MS              Number of obs =     172
-------------+------------------------------           F(  6,   165) =    2.74
       Model |  158.243503     6  26.3739171            Prob > F      =  0.0146
    Residual |  1590.05762   165  9.63671285            R-squared     =  0.0905
-------------+------------------------------           Adj R-squared =  0.0574
       Total |  1748.30112   171  10.2239832            Root MSE      =  3.1043

------------------------------------------------------------------------------
      Growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |    .072636   .0351418     2.07   0.040     .0032505    .1420214
         FDI |   .1548584   .0761923     2.03   0.044     .0044209    .3052958
      School |  -.0017549   .0102823    -0.17   0.865    -.0220567     .018547
       Trade |  -.0043139   .0068811    -0.63   0.532    -.0179004    .0092725
         Pop |   -.294139   .2343673    -1.26   0.211    -.7568845    .1686065
           G |  -.0813147   .0390172    -2.08   0.039    -.158352    -.0042774
       _cons |   2.784065   1.17108     2.38   0.019     .4718304    5.096299
------------------------------------------------------------------------------
```

Figure 5.1 Duplicated base model results for cross section

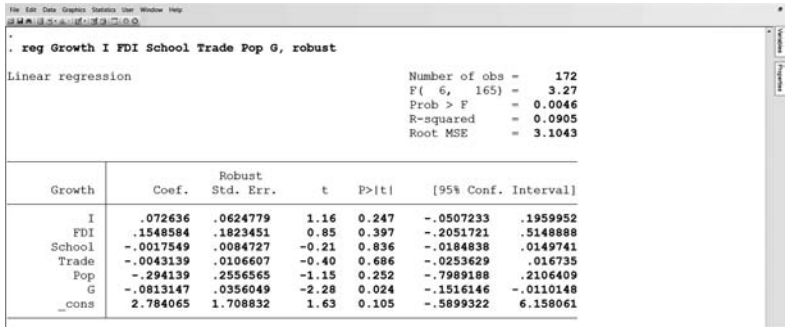
previous chapter in Figure 5.1. We will track these results and how they change as we move through this chapter.

Heteroskedasticity

Heteroskedasticity has always been considered a relatively minor problem with an easy fix if it exists. A visual test for heteroskedasticity was used in Chapter 2 whereby one can simply create a scatter plot of the absolute value of the residuals on x and look for a correlation. Another would be to generate a scatter plot of y on x and check for unequal spread as x changes. But both of these are cumbersome when the model has many right-hand-side variables. One test that is preprogrammed into nearly every software package is the White's test (White 1980), and one that is often preprogrammed is some form of the Breusch–Pagan test (Breusch and Pagan 1979).

Glossing over the gory details, both tests basically construct an auxiliary regression whereby a mathematical permutation of the residuals are regressed upon a conditioning set that is a function of the x's used in the original regression. If a correlation exists then the null of homoskedasticity must be rejected. In our case, both the White's, and Breusch–Pagan tests return p-values of 0.000, meaning that we can reject the null of homoskedasticity.

The most common way of correcting for heteroskedasticity is to use some form of a robust command in your statistical software package. In Stata this command is attached to the end of your regression command line. Most of these robust commands are constructed using some form of



```
. reg Growth I FDI School Trade Pop G, robust

Linear regression                               Number of obs =     172
                                                F(  6,   165) =    3.27
                                                Prob > F      =  0.0046
                                                R-squared     =  0.0905
                                                Root MSE      =  3.1043

------------------------------------------------------------------------------
             |               Robust
      Growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |    .072636   .0624779     1.16   0.247    -.0507233    .1959952
         FDI |   .1548584   .1823451     0.85   0.397    -.2051721    .5148888
      School |  -.0017549   .0084727    -0.21   0.836    -.0184838    .0149741
       Trade |  -.0043139   .0106607    -0.40   0.686    -.0253629     .016735
         Pop |   -.294139   .2556565    -1.15   0.252    -.7989188    .2106409
           G |  -.0813147   .0356049    -2.28   0.024    -.1516146   -.0110148
       _cons |   2.784065   1.708832     1.63   0.105    -.5899322    6.158061
------------------------------------------------------------------------------
```

Figure 5.2 Heteroskedasticity corrected cross-sectional results

White's heteroskedasticity corrected standard errors (White 1980). The nice thing about this procedure is that it only corrects the standard errors and doesn't affect the estimates of the coefficients like a feasible generalized least squares (FGLS) procedure would.

Rerunning the regression for the cross-sectional model with the robust command added to the end of our command line we find relatively large changes in the *p*-values of some of our coefficients. When comparing Figure 5.2 with Figure 5.1, the *p*-values for the coefficients of *I* and *FDI* have changed substantially. Before, the coefficient for *I* had a *p*-value of 0.040, and for *FDI* it was 0.044 indicating that these variables had a statistically significant effect on growth. Now they are 0.247 and 0.397 respectively. In fact, only *G* has a significant influence on growth. All this said, none of the coefficient values have changed. This is one of the nice properties of White's correction versus FGLS.

The reader will also notice that there is no more indication of an adjusted R^2 value like there was in the original base regression output. This is because when the robust command is employed using Stata, it drops this particular statistic from its output screen. To be honest, I'm not sure exactly why this happens. But it's not really relevant in this case anyway, and if truly interested it can still be retrieved through the e(r2_a) command.

Intercept Heterogeneity

We now move on to testing for intercept heterogeneity, and of course correcting for it if it exists. Testing for intercept heterogeneity on

a cross-sectional model is easy. You would simply include a dummy variable (or variables) that equals 1 for some quasi-obvious clustering of data across the *i*-dimension, and equals 0 otherwise.

In our case since the dependent variable is growth in real per capita GDP across a broad cross section of countries, an obvious way to cluster this data would be to delineate wealthier nations from those that are less wealthy. The reasoning for this grouping comes from economic theory whereby countries that have lower stocks of capital are not as far along on their production functions as countries with relatively high levels of capital stock; so these countries should be growing at a relatively higher rate because they are sitting at a point on their production function that is steeper than nations with larger levels of capital stock.

Another possible grouping of this data would be regions of the world. The question one would ask here to justify clustering our data in this fashion would be: Do countries in Asia tend to grow more on average than countries in Africa? Or, do countries in North America tend to grow more rapidly on average than countries in Europe? If the answer is yes, then clustering countries by region of the world may be appropriate as well.

Again, these groupings are based upon what makes sense regarding the dependent variable and its characteristics within the discipline for which it is being examined. For instance, assume our left-hand-side variable was trade in goods and services. Then delineating by region of the world might be better than economic status as trading within a region is typically cheaper in terms of transportation costs than trading with overseas countries. For quasi-micro data of a single country, you could cluster data by region or state within that country. For industry data one could cluster data by firm size, type of industry, regions, or states where firms in the industry reside, and so on.

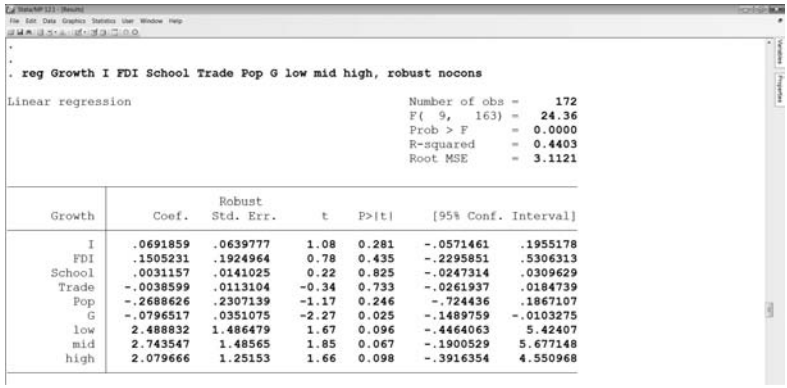
Now assume that there are K income groups in our data set. We would test for heterogeneity by constructing K dummies and including them into our heteroskedasticity corrected regression from the previous section. We can determine the presence of intercept heterogeneity by testing the equality of coefficients across dummy variables. However, note that since we are including all K groups in these tests, the constant will have to be dropped from the regression due to perfect collinearity; you can reinstate the constant after the tests are completed. A *p*-value greater than

0.100 would indicate that there is no significant difference between groups and therefore no difference in their intercepts. A *p*-value less than this would indicate differences in intercepts that should be addressed.

You could also test for statistically significant *differences* in intercepts by running a regression with *K*–1 dummies and retaining the constant. In this case, the coefficient for the constant would reflect the intercept for the control group of whatever delineation you choose, while the coefficients attached to the *K*–1 dummies would reflect the difference between the control group intercept and that for the other groups. The problem with this method is that the modeler must take into consideration the sign of the coefficient for the *K*–1 dummies relative to the coefficient's sign for the constant. If they are of opposite signs, it doesn't necessarily mean that the sum of the two coefficients is different from zero. Therefore, evaluating it relative to zero may be problematic.

The World Bank dataset roughly delineates three income groups as low, middle, and high. (Note that the middle-income countries are actually broken down into lower middle and upper middle; I do not make that distinction here and combine those two groups.) Not that in reality there aren't more than this, obviously there are. But since they have already done much of our work for us, let's just pretend that these are the only three that we conduct our test with. We represent each of these, respectively, as *low*, *mid*, and *high* in our regression output.

The reader can see in Figure 5.3 that there is not a substantial difference in magnitude between the coefficient estimates of the income dummy variables; in fact, they only range from 2.079 to 2.743. The *p*-value for a test measuring the simultaneous equality of all of the income coefficients is 0.727, indicating that we can accept the null that the coefficients are equal to each other as a group; we can thus conclude that intercept heterogeneity does not exist in a statistically significant sense. However, we will continue to work with this model through testing for slope heterogeneity before we consolidate any of the intercepts. The reason we want to do this is because many times if the slope differs so does the intercept, even though when tested on its own there is no apparent difference in intercepts. There is no theoretical or statistical justification for this relationship, it's simply one I've noticed to occur over the years.



```
. reg Growth I FDI School Trade Pop G low mid high, robust nocons

Linear regression                               Number of obs =     172
                                                F(  9,   163) =   24.36
                                                Prob > F      =  0.0000
                                                R-squared     =  0.4403
                                                Root MSE      =  3.1121

------------------------------------------------------------------------------
             |               Robust
      Growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |   .0691859   .0639777     1.08   0.281    -.0571461    .1955178
         FDI |   .1505231   .1924964     0.78   0.435    -.2295851    .5306313
      School |   .0031157   .0141025     0.22   0.825    -.0247314    .0309629
       Trade |  -.0038599   .0113104    -0.34   0.733    -.0261937    .0184739
         Pop |  -.2688626   .2307139    -1.17   0.246     -.724436    .1867107
           G |  -.0796517   .0351075    -2.27   0.025    -.1489759   -.0103275
         low |   2.488832   1.486479     1.67   0.096    -.4464063     5.42407
         mid |   2.743547    1.48565     1.85   0.067    -.1900529    5.677148
        high |   2.079666   1.25153     1.66   0.098    -.3916354    4.550968
------------------------------------------------------------------------------
```

Figure 5.3 Intercept heterogeneity regression

A minor point highlighted in Figure 5.3 that the reader should be aware of is the value of R^2. It is far higher than it was before—0.440 versus 0.090. Keep in mind that the calculation of this statistic requires that the regression contain a constant, and this one doesn't. The issue lies in the fact that R^2 is a function of the deviation in the actual data from the left-hand-side variables' mean, and this is where the constant plays a role. So, without the constant, the regression software uses an arbitrary mean, for lack of a better term, from which to calculate this deviation; hence, the value of R^2 will always be inflated, and should be ignored.

Slope Heterogeneity

As mentioned in Chapter 2, slope heterogeneity is much like intercept heterogeneity, but when the slope differs in a statistically significant way across some obvious clustering of the dependent variable. Injecting slope heterogeneity into equation (2.1) from Chapter 2 we would have

$$y = a_0 + a_1 D_1 x + a_2 D_2 x + e \tag{5.1}$$

where D_1 is a dummy variable that equals 1 when characterizing one group and 0 otherwise, and D_2 would represent, say, a second group if it equals 1 and 0 otherwise. And while (5.1) is the best for conducting the initial tests of differences in slopes as will be obvious subsequently, a more common form is

$$y = a_0 + a_1 x + a_2 Dx + e \tag{5.2}$$

where D is a dummy variable that is equal to 1 for some cluster and equal to 0 for some other cluster. Therefore, the slope for a control group would simply be a_1 but the slope for the delineated cluster would be $a_1 + a_2$. Hence, a_2 measures the difference between the control group's slope and that for the delineated cluster.

In our case, an obvious clustering of our dependent variable, Growth, would be by income group—the same clustering we explored for intercept heterogeneity. The reader should keep in mind that testing for slope and intercept heterogeneity across the *same* grouping is recommended. It would make little sense to test the intercepts for income, but test differences in slopes for groups delineated by geographical region.

As mentioned earlier, in the case of a cross-sectional model in particular, always leave the intercept heterogeneity components in the model when testing for slope heterogeneity, even if there is not a statistically significant difference between the intercepts when tested separately. This is because it is often the case that slopes and intercepts differ simultaneously, but when checked individually, they don't.

Figure 5.4 gives us our first look at simultaneously controlling for intercept and slope heterogeneity for the cross-sectional model. Even

File Edit Data Graphics Statistics User Window Help

Growth	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
I_high	.2154285	.1671306	1.29	0.199	-.114788	.545645
FDI_high	.2605446	.251976	1.03	0.303	-.2373094	.7583985
School_high	-.0709579	.0324656	-2.19	0.030	-.1351034	-.0068124
Trade_high	-.0087865	.0161787	-0.54	0.588	-.0407524	.0231795
Pop_high	-.2128313	.4157835	-0.51	0.609	-1.034336	.6086733
G_high	.0917256	.1033148	0.89	0.376	-.1124037	.2958549
I_mid	.0168272	.0385955	0.44	0.663	-.0594298	.0930841
FDI_mid	-.0426921	.1183682	-0.36	0.719	-.2765638	.1911796
School_mid	.0340134	.0124047	2.74	0.007	.0095041	.0585227
Trade_mid	-.0048122	.0106298	-0.45	0.651	-.0258146	.0161901
Pop_mid	-.3170557	.2690493	-1.18	0.240	-.8486431	.2145317
G_mid	-.0743136	.032006	-2.32	0.022	-.137551	-.0110762
I_low	.1250095	.1176248	1.06	0.290	-.1073934	.3574124
FDI_low	1.176926	.4337015	2.71	0.007	.3200195	2.033833
School_low	.0313956	.0221322	1.42	0.158	-.0123332	.0751244
Trade_low	-.0006048	.0278854	-0.02	0.983	-.0557008	.0544912
Pop_low	.1737549	.5273	0.33	0.742	-.8680839	1.215594
G_low	-.2594946	.0769471	-3.37	0.001	-.4115265	-.1074627
low	1.124048	2.211986	0.51	0.612	-3.246391	5.494488
mid	2.956464	1.18792	2.49	0.014	.6093716	5.303556
high	1.660136	4.640907	0.36	0.721	-7.509363	10.82964

Figure 5.4 Slope heterogeneity regression

though Figure 5.4 gives us the estimates of our growth model while controlling for intercept and slope heterogeneity, it doesn't really tell us much at this point. What we need to do now is test the equality of the intercepts and slopes for each variable across income groups. If we find that these do not differ for some variable(s), then we can drop that particular delineation from our regression. To this end, the *p*-value for equal intercepts is 0.753 indicating that like before, the intercepts do not differ across groups. For the slopes we find *p*-values of 0.375, 0.962, and 0.709 for the variables *I*, *Trade*, and *Pop*, respectively; this means that we can consolidate income groups for these variables as well. But testing the slopes for *FDI* we get a *p*-value of 0.020, *School* we get 0.010, and *G* we get 0.018. Therefore, we cannot consolidate the slopes of these groups.

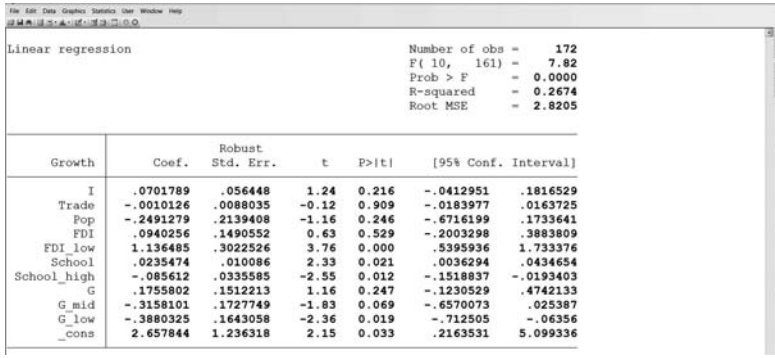
As you can see, accounting for differences in slope substantially changed the inference we can draw from our model. In this cross-sectional case we find that *FDI* has no effect on *Growth* for middle- and high-income countries, but has a large positive effect for low-income countries. And while *School* has a positive effect on middle- and low-income countries, it actually has a negative effect on high-income nations. *G*, or government spending, negatively affects growth in middle- and low-income countries, but doesn't have a significant effect on high-income economies. The model depicted in Figure 5.5, however, is not in its most parsimonious form. To come to that form, we must now test whether any two seemingly similar income-level effects are equal, and if so, combine

Linear regression

Number of obs = 172
F(12, 159) = 6.47
Prob > F = 0.0000
R-squared = 0.3017
Root MSE = 2.7709

Growth	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
I	.0697735	.0520097	1.34	0.182	-.0329454	.1724924
Trade	-.0074983	.0121671	-0.62	0.539	-.0315282	.0165317
Pop	-.2645145	.2091444	-1.26	0.208	-.6775739	.148545
FDI_high	.3129326	.2833281	1.10	0.271	-.2466393	.8725044
FDI_mid	-.0457913	.1102523	-0.42	0.678	-.2635392	.1719567
FDI_low	1.292061	.332193	3.89	0.000	.6359813	1.948141
School_high	-.0625313	.0305082	-2.05	0.042	-.1227849	-.0022777
School_mid	.0302894	.0113813	2.66	0.009	.0078113	.0527674
School_low	.0297883	.0166962	1.78	0.076	-.0031866	.0627632
G_high	.1682013	.1318029	1.28	0.204	-.092109	.4285116
G_mid	-.1190285	.0379113	-3.14	0.002	-.1939031	-.0441539
G_low	-.2072219	.052146	-3.97	0.000	-.31021	-.1042337
_cons	2.816502	1.098291	2.56	0.011	.6473819	4.985621

Figure 5.5 Regression corrected for slope heterogeneity



Linear regression

Number of obs	= 172
F(10, 161)	= 7.82
Prob > F	= 0.0000
R-squared	= 0.2674
Root MSE	= 2.8205

Growth	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
I	.0701789	.056448	1.24	0.216	-.0412951	.1816529
Trade	-.0010126	.0088035	-0.12	0.909	-.0183977	.0163725
Pop	-.2491279	.2139408	-1.16	0.246	-.6716199	.1733641
FDI	.0940256	.1490552	0.63	0.529	-.2003298	.3883809
FDI_low	1.136485	.3022526	3.76	0.000	.5395936	1.733376
School	.0235474	.010086	2.33	0.021	.0036294	.0434654
School_high	-.085612	.0335585	-2.55	0.012	-.1518837	-.0193403
G	.1755802	.1512213	1.16	0.247	-.1230529	.4742133
G_mid	-.3158101	.1727749	-1.83	0.069	-.6570073	.025387
G_low	-.3880325	.1643058	-2.36	0.019	-.712505	-.06356
_cons	2.657844	1.236318	2.15	0.033	.2163531	5.099336

Figure 5.6 Final model moving into next section

them. Also, when all is said and done, common practice dictates that our model be left in the form of equation (5.2), not equation (5.1).

Testing the equality of high-income *FDI* with middle-income *FDI*, we get a *p*-value of 0.214 indicating that we can combine these two, thereby separating these effects from the low-income effect. Testing low- and middle-income *School* we get a *p*-value of 0.977; again, we can combine these two effects. Testing middle- and low-income *G* we get a *p*-value of 0.062, indicating that we cannot combine these effects, so they will be left as is. Combining (or not) these slopes and converting our model to the more popular form of equation (5.2) as shown previously, the cross-sectional model we will move forward with will be that in Figure 5.6.

Statistical Omitted Variable Bias

I speak about statistical omitted variable bias in my other book (Edwards 2013), and how it differs from what I call theoretical omitted variable bias. I also described it back in Chapter 2 of this book. Most researchers lump them both together and call them simply omitted variable bias, but I like to differentiate theoretical bias from the more objective and purely statistical bias.

A perfect example of statistical omitted variable bias is the inclusion of a squared *x*. If a squared *x* is needed but not included on the right-hand side prior to running the regression, bias in the relationship of interest would result. And while we could hypothesize why there would exist

a nonlinear relationship between *x* and *y*, for the most part, testing for it is simply to paint a more statistically accurate picture of the relationship with *Growth*. Theoretical omitted variable bias is different. The idea of it relies only on theory. We may have such bias if (1) there exists a variable that we have access to that is correlated with *Growth*, and (2) it is simultaneously correlated with our variable of interest. Both (1) and (2), of course, are based purely on theory and are nearly always likely to be true with at least one variable that is not in your regression. In this sense, as long as any residual at all exists from our regression, that is, as long as our R^2 is less than 1.00, our estimators will always be biased! This is why, in my opinion, theoretical omitted variable bias is a concept that has tenuous argumentative support at best.

While Figures 2.9 to 2.11 show perfectly a hypothetical nonlinear relationship between *x* and *y*, when there are more than one *x*, graphically testing for statistical bias can be problematic. Furthermore, statistical omitted variable bias proper actually involves more than just the quadratic specification issue, it also involves interactions. A regression model addressing the former would look like equation (2.12) in Chapter 2, or

$$y = a_0 + a_1 x_1 + a_2 x_1^{\,2} + e \tag{5.3}$$

But a regression model addressing the latter would look like

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 + e \tag{5.4}$$

whereby x_1 and x_2 are different *x*'s.

The reason why I only presented the quadratic case in Chapter 2 is because I know from experience that empirically testing for the interaction of variables can lead to outcomes that are theoretically tenuous, and therefore difficult if not impossible for the researcher to explain. For instance, let's assume we found that domestic investment interacts with government spending such that the marginal effect of investment on growth is

$$\frac{\partial Growth}{\partial I} = a_1 + a_2 G \tag{5.5}$$

Equation (5.5) states that the marginal effect of *I* on *Growth* is a function of government spending. This is easy to explain since we know from basic macroeconomics that the two can affect each other through crowding out type of arguments—especially if *G* represents deficit spending. The deficit would push interest rates up reducing the quantity demanded of loanable funds, thereby reducing *I* and *Growth* as a result. But, what if we found that *Trade* and *School* interacted with each other, thereby affecting their own separate relationships with *Growth*? That would be much harder to explain. To this end, we only explore the quadratic issue in this book and not interactions. However, please be aware that this problem can indeed exist and should be explored. In my own work, I empirically explore these interactions only if there are strong theoretical reasons why they should exist.

The most common way of testing for quadratic relationships is to simply include squared *x*'s into the regression and evaluate the statistical significance of their coefficients. To this end, Figure 5.7 is just such a regression of our cross section. The coefficient estimates for each of the squared variables are indicated by an "sq" after the variable abbreviation.

What we find is that nearly all coefficients are insignificant. This is not uncommon with this operation. Just like when we try to force a linear model on a nonlinear relationship, we could get insignificance of the linear relationship because the relationship is actually quadratic. In the same sense, trying to force a nonlinear relationship on a linear one can cause

Growth	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
I	-.3169055	.2352951	-1.35	0.180	-.7818013	.1479902
Isq	.0071636	.0048416	1.48	0.141	-.0024025	.0167297
Trade	-.0143536	.0175614	-0.82	0.415	-.0490515	.0203442
Tradesq	.0000531	.0000507	1.05	0.297	-.000047	.0001532
Pop	.0036563	.5535024	0.01	0.995	-1.089953	1.097266
Popsq	-.0515583	.1140995	-0.45	0.652	-.276996	.1738795
FDI	.2642997	.2478324	1.07	0.288	-.2253672	.7539666
FDIsq	-.0114361	.0123145	-0.93	0.355	-.0357671	.012895
FDI_low	2.279046	.8142506	2.80	0.006	.6702509	3.887842
FDIsq_low	-.2384613	.167051	-1.43	0.156	-.5685206	.0915979
School	.0000819	.0418323	0.00	0.998	-.0825703	.082734
Schoolsq	.0003154	.0004074	0.77	0.440	-.0004896	.0011204
School_high	-.050043	.0758158	-0.66	0.510	-.1998397	.0997537
Schoolsq_high	-.0004385	.000533	-0.82	0.412	-.0014916	.0006146
G	.4021995	.3841781	1.05	0.297	-.3568592	1.161258
Gsq	-.0062347	.0069323	-0.90	0.370	-.0199316	.0074622
G_mid	-.3173092	.3333623	-0.95	0.343	-.9759661	.3413477
Gsq_mid	-.0000332	.0067947	-0.00	0.996	-.0134581	.0133918
G_low	-.4930795	.346533	-1.42	0.157	-1.177759	.1916
Gsq_low	.0038001	.0068027	0.56	0.577	-.0096406	.0172408
_cons	6.001799	2.026351	2.96	0.004	1.998136	10.00546

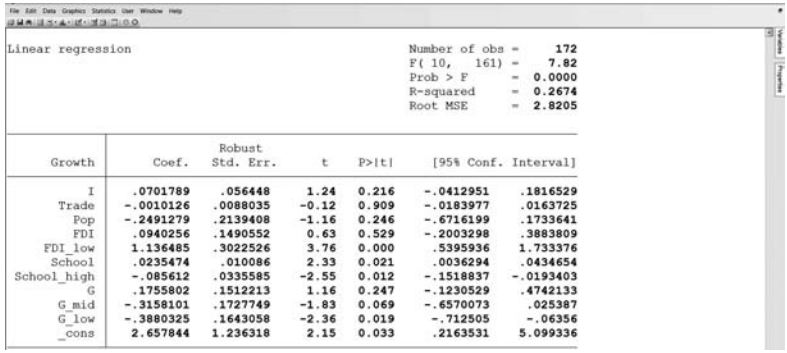
Figure 5.7 Regression checking for statistical omitted variable bias

that same issue—that is, the linear relationship can become insignificant when it is actually the correct relationship. Therefore, we can safely drop all of the squared right-hand-side variables and return to the output in Figure 5.6 as being the final cross-sectional model.

The Final Cross-Sectional Model and the Inference We CanDraw From It

At this point, I want to elaborate on the inference we can draw from these results. For ease of referencing, our final estimates are repeated in Figure 5.8, while the model and output we started with is repeated in Figure 5.9.

Since there is heterogeneity in three of the relationships, inference from those particular results must be drawn according to the delineated

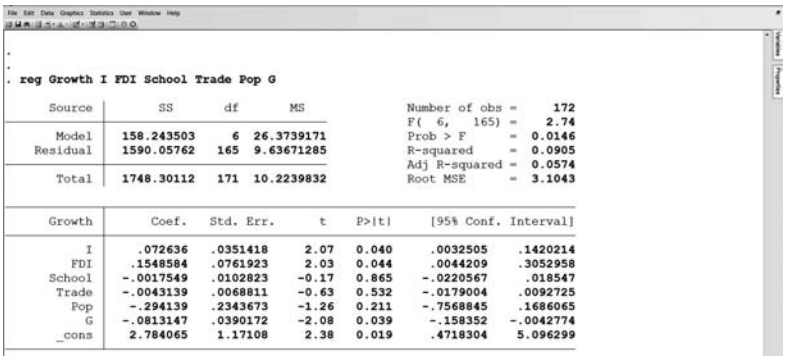


Linear regression

Number of obs = 172
F(10, 161) = 7.82
Prob > F = 0.0000
R-squared = 0.2674
Root MSE = 2.8205

Growth	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
I	.0701789	.056448	1.24	0.216	-.0412951	.1816529
Trade	-.0010126	.0088035	-0.12	0.909	-.0183977	.0163725
Pop	-.2491279	.2139408	-1.16	0.246	-.6716199	.1733641
FDI	.0940256	.1490552	0.63	0.529	-.2003298	.3883809
FDI_low	1.136485	.3022526	3.76	0.000	.5395936	1.733376
School	.0235474	.010086	2.33	0.021	.0036294	.0434654
School_high	-.085612	.0335585	-2.55	0.012	-.1518837	-.0193403
G	.1755802	.1512213	1.16	0.247	-.1230529	.4742133
G_mid	-.3158101	.1727749	-1.83	0.069	-.6570073	.025387
G_low	-.3880325	.1643058	-2.36	0.019	-.712505	-.06356
_cons	2.657844	1.236318	2.15	0.033	.2163531	5.099336

Figure 5.8 Final model using cross-sectional data



. reg Growth I FDI School Trade Pop G

Source	SS	df	MS
Model	158.243503	6	26.3739171
Residual	1590.05762	165	9.63671285
Total	1748.30112	171	10.2239832

Number of obs = 172
F(6, 165) = 2.74
Prob > F = 0.0146
R-squared = 0.0905
Adj R-squared = 0.0574
Root MSE = 3.1043

Growth	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
I	.072636	.0351418	2.07	0.040	.0032505	.1420214
FDI	.1548584	.0761923	2.03	0.044	.0044209	.3052958
School	-.0017549	.0102823	-0.17	0.865	-.0220567	.018547
Trade	-.0043139	.0068811	-0.63	0.532	-.0179004	.0092725
Pop	-.294139	.2343673	-1.26	0.211	-.7568845	.1686065
G	-.0813147	.0390172	-2.08	0.039	-.158352	-.0042774
_cons	2.784065	1.17108	2.38	0.019	.4718304	5.096299

Figure 5.9 Duplicated base model results for cross section

groups. However, for the other three right-hand-side variables, this distinction is not necessary. Regardless of income group, neither *I*, *Trade*, nor *Pop* have any effect on economic growth in real per capita GDP. However, for low-income groups, increasing injections of *FDI* will have a positive effect on *Growth*, but no effect for high- and middle-income countries. *School* also has a positive effect on *Growth* for both low- and middle-income countries. But unlike the income delineated outcome for *FDI*, we cannot infer the same for high-income *School*. The *FDI* coefficients are of the same sign, but the *School* coefficients estimates are of opposite signs. This means that we will actually have to test whether the effect of *School* in high-income economies is negative. A test such as this is necessary because the high-income coefficient estimate is larger in absolute value than the low and middle estimate, yet negative; therefore, we are testing whether the sum of the two estimates, which is approximately –0.0621, is a statistically significant negative coefficient. The *p*-value of this test is 0.064 which is well below our 0.100 threshold, and therefore we can conclude that *School* has a positive influence on growth in low- and middle-income countries, but a negative effect on high-income countries. For the astute reader, one would notice that this coefficient was also significant in Figure 5.5 where we separately delineated each income group instead of modeling the differences in groups. However, that *p*-value is misleading as we hadn't yet combined the low- and middle-income countries.

Moving on to *G*'s effect on *Growth*, we have the same issue just for different income groups. In this case we will have to test whether 0.1755 plus –0.3158 is different from zero, and whether 0.1755 plus –0.3880 are different from zero. Conducting the first test returns a *p*-value of 0.001, and the second returns a *p*-value of 0.000. Therefore, the inference for the effect *G* has on *Growth* is that it has no effect for high income economies, but does have a negative effect on low- and middle-income countries.

It should be obvious that these results are considerably different from those obtained at the beginning of this lesson—that is, the base model results repeated in Figure 5.9. It is also the case that the final model explains far more of the variation in *Growth* than the base model did. In fact, the R^2 for the base model is only 0.090, but is 0.267 for the respecified model. This is an increase of nearly 200 percent in explanatory power! And while many applied researchers correct for heteroskedasticity

regardless of whether they have it or not, this correction alone would not lead us to our final model. One can see then that the model generating the results in Figure 5.8 is in fact the better model. But for the more sophisticated reader, we still aren't finished. We must now check for variance heterogeneity—a concept we explore in the next chapter. But, for the less initiated, following the basic testing and correction procedures explored in Chapter 5 will at least provide the researcher with more reliable and robust inference, and produce a model that better explains the data.

CHAPTER 6

Variance Heterogeneity

The Cross-Sectional Case

If we recollect from Chapter 3, variance heterogeneity can rarely be detected with a test for homoskedasticity. To reiterate equation (3.5) from that chapter, we have

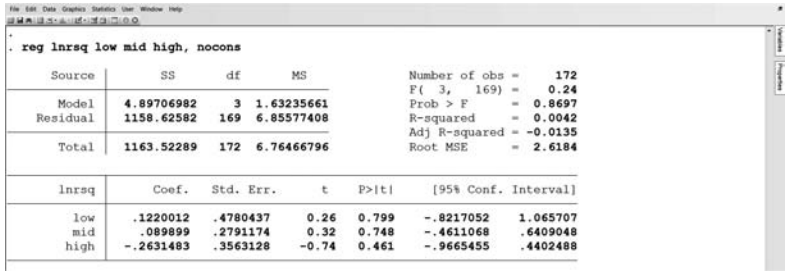
$$r^2 = c_0 + c_1 D + \mu, \tag{6.1}$$

which tells us that in the case of variance heterogeneity, a correlation would exist between the squared residuals and a set of dummy variables representing a particular clustering of the data like in intercept and slope heterogeneity. However, we run into an issue if we perform a regression specified exactly like (6.1). That is because the dependent variable is chi-squared distributed since it is a squared version of a normally distributed variable (at least we hope it's normally distributed). Park (1966) recommended using the natural log of the squared residuals instead, resulting in the regression

$$\ln(r^2) = c_0 + c_1 D + v. \tag{6.2}$$

This is the representation we will use.

Figure 6.1 shows the screenshot of our Stata output when we run a regression of the natural log of our squared residuals from the final model in Chapter 5 on the income level dummy variables we created previously. We find that coefficients for the dummy variables are all insignificant, but like intercept heterogeneity, it is not the actual values we are necessarily interested in, but the differences in these values from each other. After



```
. reg lnrsq low mid high, nocons

      Source |       SS       df       MS              Number of obs =     172
-------------+------------------------------           F(  3,   169) =    0.24
       Model |  4.89706982     3  1.63235661            Prob > F      =  0.8697
    Residual |  1158.62582   169  6.85577408            R-squared     =  0.0042
-------------+------------------------------           Adj R-squared = -0.0135
       Total |  1163.52289   172  6.76466796            Root MSE      =  2.6184

------------------------------------------------------------------------------
       lnrsq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         low |   .1220012   .4780437     0.26   0.799    -.8217052    1.065707
         mid |    .089899   .2791174     0.32   0.748    -.4611068    .6409048
        high |  -.2631483   .3563128    -0.74   0.461    -.9665455    .4402488
------------------------------------------------------------------------------
```

Figure 6.1 Regression testing for variance heterogeneity

conducting an equality test of the *low*, *mid*, and *high* coefficients we get a *p*-value of 0.702, indicating that the coefficients are not different from each other in a statistically significant sense, and therefore, no variance heterogeneity exists. But, what if the result of this test were different, exactly how would we correct for variance heterogeneity if it did exist?

Let's assume that variance heterogeneity did indeed exist. The correction for it would entail using a generalized least squares (GLS) type of method. Nearly all undergraduate textbooks address this methodology when used for heteroskedasticity correction. But, GLS is only good if you know the true nature of the heterogeneity in the variance; in our case, however, we are estimating it. To this end, we must use a variant called a feasible generalized least squares procedure (FGLS). This is perhaps the best way to correct for variance heterogeneity when it is known to be present (Edwards et al. 2006). The way someone would perform an FGLS in this case is as follows.

Again, let us assume that each of our *low*, *mid*, and *high* coefficients in Figure 6.1 are different from each other in a statistically significant sense. Then, given these estimates we would estimate the natural log of the squared residuals as

$$\widehat{\ln(r^2)} = \widehat{C_0} + \widehat{C_1}D \tag{6.3}$$

Specifically, in our case we would have the function

$$\widehat{\ln(r^2)} = 0.122\,low + 0.089\,mid - 0.263\,high \tag{6.4}$$

We would then take these estimates of $\widehat{\ln(r^2)}$, and convert them to simply $\hat{r}$ by first exponentiating, and then taking the square root of what's left. To correct for heterogeneity, we would then weight all of the variables in our model, including the constant. Hence, reflecting on model (2.1) in Chapter 2, our new model would look like

$$\frac{y}{\hat{r}} = a_0 \frac{1}{\hat{r}} + a_1 \frac{x}{\hat{r}} + \frac{e}{\hat{r}} \tag{6.5}$$

Completing this operation for the final regression from Chapter 5, and rerunning that regression, we get the output in Figure 6.2.

The "w" in front of each of the coefficient or variable names stands for that variable weighted by the $\hat{r}$ that we generated earlier. The reader will notice that none of the relevant *p*-values changed in any dramatic way—at least in a way that would cause a re-evaluation of the inference we drew from the results of our final model in Chapter 5. But we expected this. The reason is that since we didn't find any variance heterogeneity in the results from Figure 6.1, all of those standard errors, and therefore the *p*-values generated from them, were unbiased; remember, we only conducted this experiment for expositional purposes. Having said this, there are a couple of differences in this latest batch of output that the reader should be aware of.

The first difference is in the sample correlation coefficient—that is, the R^2. The R^2 depicted in Figure 5.8 was 0.267, but the one depicted

Linear regression

Number of obs	=	172				
F(11, 161)	=	26.38				
Prob > F	=	0.0000				
R-squared	=	0.5306				
Root MSE	=	2.9757				

wGrowth	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
wI	.0847894	.0665509	1.27	0.204	-.0466358	.2162146
wTrade	-.00185	.0099419	-0.19	0.853	-.0214833	.0177833
wPop	-.2489578	.2348451	-1.06	0.291	-.7127318	.2148162
wFDI	.1214842	.1660685	0.73	0.466	-.2064692	.4494376
wFDI_low	1.115638	.3045038	3.66	0.000	.5143016	1.716975
wSchool	.0219289	.0106421	2.06	0.041	.0009128	.042945
wSchool_high	-.0850391	.0323051	-2.63	0.009	-.1488356	-.0212427
wG	.1668387	.1416739	1.18	0.241	-.1129401	.4466174
wG_mid	-.3162949	.1674732	-1.89	0.061	-.6470223	.0144325
wG_low	-.3851943	.1571388	-2.45	0.015	-.6955133	-.0748754
wcons	2.55269	1.393706	1.83	0.069	-.1996123	5.304993

Figure 6.2 FGLS model if variance heterogeneity existed

in Figure 6.2 is 0.530, or nearly twice as high. This is a false reading. The reason is the same as it is for an R^2 that is generated from a regression that doesn't have a constant term such as that used to generate the results in Figure 5.3, which also had an inflated R^2. The calculation of the R^2 assumes that a constant exists, otherwise captured in $\overline{y}$, which is one of the main components in the calculation of the correlation coefficient itself; therefore, any calculation of this statistic when a constant doesn't exist is biased. In our case, the constant is divided by $\hat{r}$ which is a variable, and therefore the entire term itself, $a_0 \frac{1}{\hat{r}}$, is no longer constant.

The other difference in the two batches of output one should notice is that the coefficient estimates themselves have changed, albeit not by much. This is the main drawback of correcting for variance heterogeneity. Although asymptotically unbiased, after performing the FGLS procedure the coefficients do not maintain the nice finite sample properties that they had before the procedure was completed. To this end, the researcher must use discretion when evaluating the benefits of variance heterogeneity correction, such as the benefit of accounting for substantial bias in *p*-values versus the changes in the coefficient estimates themselves. The rule of thumb I use is that if the magnitudes of the coefficients are important for forecasting specific levels of the left-hand-side variable, then I will probably ignore the heterogeneity because any change in the coefficient values may greatly change those forecasts. But if the actual values of the coefficients aren't important, and only the sign of the coefficient is important, then I will probably correct for variance heterogeneity.

To be honest, I have absolutely nothing to support my reasoning. It's simply a personal condition I apply to my own work. But it doesn't matter; as long as the researcher is aware of these issues and weighs the pros and cons of variance correction using the FGLS procedure, then one should be able to make an educated decision whether to perform it or not.

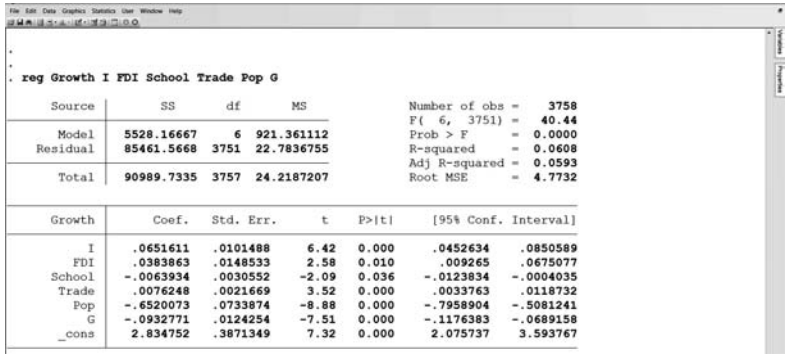
CHAPTER 7

Basic Misspecification Testing and Respecification

The Panel Data Case

In this chapter we test for and correct (if needed) misspecification issues 1 through 5 from Chapter 1 specifically for panel data regressions. The order of the misspecification testing and respecification procedures certainly works well for me when using cross-sectional models, and it seems to work particularly well with panel data models. Exactly why, I don't know; but, I suspect it has much to do with the fact that panel data has an extra dimension—a time dimension. Hence, we need to add misspecification (3) to our list of procedures—dependent variable dynamics in panel data. I've found that once (3) is addressed, the theoretical interpretation of the results from (4) and (5) tend to be less awkward.

To repeat myself from Chapter 5 (if you read that chapter, but it's not necessary if your only interest is in panel data modeling), the algorithm outlined in this chapter is one developed by myself over years of performing empirical research, analyzing other author's research, and publishing many articles that address the topic of model specification. The reason I call the step-by-step process an algorithm is because it can be recursive in nature. There have been instances in my past empirical work whereby I'll test for one misspecification issue, find that it exists and correct for it, but have to go back and address it again after correcting for a completely different misspecification issue! Therefore, one cannot just perform the steps we are getting ready to outline and be done with it, even though for brevity's sake, we will do just that in this book. But the reader should take it upon themselves to recheck their assumptions after any major respecification takes place.



```
. reg Growth I FDI School Trade Pop G

      Source |       SS       df       MS              Number of obs =    3758
-------------+------------------------------           F(  6,  3751) =   40.44
       Model |  5528.16667     6  921.361112            Prob > F      =  0.0000
    Residual |  85461.5668  3751  22.7836755            R-squared     =  0.0608
-------------+------------------------------           Adj R-squared =  0.0593
       Total |  90989.7335  3757  24.2187207            Root MSE      =  4.7732

------------------------------------------------------------------------------
      Growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |   .0651611   .0101488     6.42   0.000     .0452634    .0850589
         FDI |   .0383863   .0148533     2.58   0.010      .009265    .0675077
      School |  -.0063934   .0030552    -2.09   0.036    -.0123834   -.0004035
       Trade |   .0076248   .0021669     3.52   0.000     .0033763    .0118732
         Pop |  -.6520073   .0733874    -8.88   0.000    -.7958904   -.5081241
           G |  -.0932771   .0124254    -7.51   0.000    -.1176383   -.0689158
       _cons |   2.834752   .3871349     7.32   0.000     2.075737    3.593767
------------------------------------------------------------------------------
```

Figure 7.1 Base regression using panel model

Just to reiterate the base model results we obtained earlier in the book, Figure 7.1 is the same as that from Figure 4.2 in Chapter 4. We will track these results and how they change as we move through this chapter.

Heteroskedasticity

Heteroskedasticity has always been considered a relatively minor problem with an easy fix if it exists. A visual test for heteroskedasticity was used in Chapter 2 whereby one can simply create a scatter plot of the absolute value of the residuals on x and look for a correlation. Another would be to generate a scatter plot of y on x and check for unequal spread as x changes. But both of these are cumbersome when the model has many right-hand-side variables. Furthermore, using panel data incorporates an extra dimension to the data which, theoretically at least, adds another area for heteroskedastic errors to exist. In other words, not only could the squared residuals be a function of x for some contemporaneous time period like in a cross-sectional framework, but the squared residuals could also be a function of past observations of y as well.

The tests for heteroskedasticity in a panel framework are the same as those for cross-sectional data. (Of course, this isn't entirely true because of the added layers of difficulty in modeling panel data in general, such as the existence of possibly many panels, each with its own time series; but, at this level of sophistication, these tests are sufficient for one to determine whether heteroskedasticity proper exists or not.) One test that is preprogrammed into nearly every software package is the White's test (White 1980), and

one that is often preprogrammed is some form of the Breusch–Pagan test (Breusch and Pagan 1979). Both tests construct an auxiliary regression, whereby a mathematical permutation of the residuals is regressed upon a conditioning set that is a function of the *x*'s used in the original regression. If a correlation exists then the null of homoskedasticity must be rejected. In our case, both the White's, and Breusch–Pagan tests return *p*-values of 0.000, meaning that we can reject the null of homoskedasticity.

The most common way of correcting for heteroskedasticity is to use some form of a robust command in your statistical software package. In Stata this command is attached to the end of your regression command line. Most of these robust commands are constructed using some form of White's heteroskedasticity corrected standard errors (White 1980). The nice thing about this procedure is that it only corrects the standard errors and doesn't affect the estimates of the coefficients like a feasible generalized least squares (FGLS) procedure would.

Rerunning the regression for the panel model with the robust command added to the end of our command line we find relatively large changes in the *p*-values of some of our coefficients. When comparing Figures 7.1 and 7.2, there are substantial increases in the *p*-values of the coefficients for *FDI* and *School*. Even though still statistically significant, the *p*-value for *FDI* has increased by a factor of five, while that of *School* has nearly doubled. Having said this, the overall inference that was drawn back in Chapter 4 is still the same.

The reader will notice that there is no more indication of an adjusted R^2 value like there was in the original base regression output. This is

```
. reg Growth I FDI School Trade Pop G, robust

Linear regression                               Number of obs =    3758
                                                F(  6,  3751) =   22.08
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0608
                                                Root MSE      =  4.7732

------------------------------------------------------------------------------
             |               Robust
      Growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |   .0651611   .0174086     3.74   0.000     .03103    .0992923
         FDI |   .0383863   .0195279     1.97   0.049     .0000999  .0766727
      School |  -.0063934   .0033916    -1.89   0.059    -.0130431  .0002562
       Trade |   .0076248   .0024038     3.17   0.002     .002912   .0123376
         Pop |  -.6520073   .0967304    -6.74   0.000    -.8416565  -.462358
           G |  -.0932771   .0147447    -6.33   0.000    -.1221855  -.0643686
       _cons |   2.834752   .5107782     5.55   0.000     1.833322   3.836182
------------------------------------------------------------------------------
```

Figure 7.2 Heteroskedasticity-corrected panel results

because when the robust command is employed using Stata, it drops this particular statistic. To be honest, exactly why this is I'm not sure. But, it's not really relevant in this case anyway, and if truly interested, it can still be retrieved through the e(r2_a) command. But what the reader will also notice is that none of the coefficient values have changed. This is one of the nice properties of White's correction versus FGLS.

Intercept Heterogeneity

Moving on to intercept heterogeneity, while the general tests are the same for both the cross-sectional and panel cases, the construction of the correction variables is somewhat different. Therefore, the test construction we used in Chapter 5 is quite different from what we will be using in this chapter. Let us address the testing phase first.

As mentioned in Chapter 5, testing for intercept heterogeneity on a cross-sectional model is easy. You would simply include a dummy variable (or variables) that equals 1 for some quasi-obvious clustering of data within the *i* dimension, and 0 otherwise. Likewise, for panel data that has both a small *i* and *t* dimension, one should rely entirely on the cross-sectional procedure and cluster the *i*'s in some meaningful way as previously described. But for all other cases, for example, panel models with a small *i* dimension and at least several observations over *t* for each *i*, or a relatively large *i* and *t* dimension, and so on, you should construct a dummy variable for each *i* and run the same test of equality across the coefficients. In this latter case, however, because the *i* dimension can be large like our current case where *i* equals 172, testing the equality of coefficients can be quite tedious, all with the result that you will probably find significant heterogeneity across *i* (I know I always have). This is why most researchers, like myself, simply assume it exists and make the necessary corrections.

The only real difference between large panel intercept heterogeneity correction procedures and that for the cross section is that in the panel case, researchers correct for this issue at a smaller level than they would in a cross section. Ideally, it is always better to address intercept and slope heterogeneity at the smallest possible level as long as the researcher doesn't overspecify their model, thereby swamping the effects the *x*'s have on *y*. In

the cross-sectional case we checked for heterogeneity using income delineated dummies. The inclusion of these three dummy variables are not enough to swamp the effects *I*, *FDI*, and so on, have on *Growth*, yet they are enough that if intercept heterogeneity did exist, they would pick it up. Obviously, we wouldn't have been able to use country-level dummies simply because there aren't enough observations. But because of the relatively large number of observations we have to work within our panel data set, each *i* having over 20 observations on average, correcting intercept heterogeneity at the country level is now possible. To this end, when using panel data that contain at least several *t* observations for each *i*, researchers will commonly use a technique known as a within regression. Within regressions can be easily performed with most regression packages.

A within regression goes something like this. We begin with equation (7.1)

$$y_{it} = a_{i0} + a_1 x_{it} + e_{it}. \tag{7.1}$$

Now, if we assume that this relationship is consistent over time, then we can rewrite (7.1) as

$$\overline{y}_i = a_{i0} + a_1 \overline{x}_i + \overline{e}_i, \tag{7.2}$$

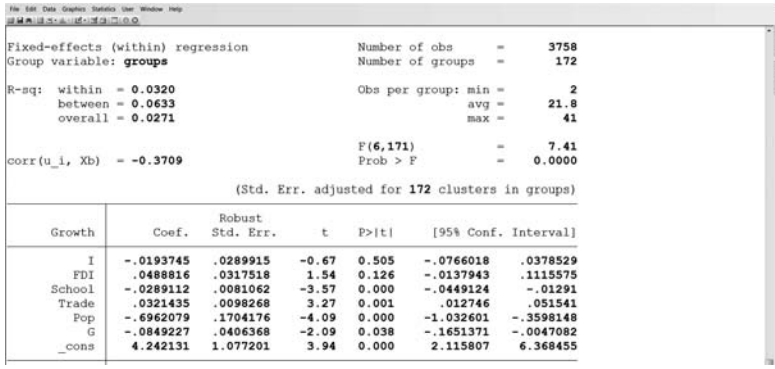
where the bar across the top of each variable stands for that variable's time mean within country *i*. Now, subtracting (7.2) from (7.1), we get,

$$(y_{it} - \overline{y}_i) = (a_{i0} - a_{i0}) + a_1 (x_{it} - \overline{x}_i) + (e_{it} - \overline{e}_i). \tag{7.3}$$

Since the constants do not vary over time, they cancel each other out, resulting in the equation

$$(y_{it} - \overline{y}_i) = a_1 (x_{it} - \overline{x}_i) + (e_{it} - \overline{e}_i). \tag{7.4}$$

Hence, there is no need to construct dummy variables for each country. We just transform the data as shown and we will get the same slope coefficient estimates as earlier. And since we are only interested in the a_1's anyway, we can still draw the inference needed to conduct our research.



```
Fixed-effects (within) regression               Number of obs      =      3758
Group variable: groups                          Number of groups   =       172

R-sq:  within  = 0.0320                         Obs per group: min =         2
       between = 0.0633                                        avg =      21.8
       overall = 0.0271                                        max =        41

                                                F(6,171)           =      7.41
corr(u_i, Xb)  = -0.3709                        Prob > F           =    0.0000

                                  (Std. Err. adjusted for 172 clusters in groups)
------------------------------------------------------------------------------
             |               Robust
      Growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |  -.0193745   .0289915    -0.67   0.505    -.0766018    .0378529
         FDI |   .0488816   .0317518     1.54   0.126    -.0137943    .1115575
      School |  -.0289112   .0081062    -3.57   0.000    -.0449124   -.01291
       Trade |   .0321435   .0098268     3.27   0.001     .012746     .051541
         Pop |  -.6962079   .1704176    -4.09   0.000    -1.032601   -.3598148
           G |  -.0849227   .0406368    -2.09   0.038    -.1651371   -.0047082
       _cons |   4.242131   1.077201     3.94   0.000     2.115807    6.368455
------------------------------------------------------------------------------
```

Figure 7.3 Within regression results

Applying this technique to the heteroskedasticity corrected panel model we get the output in Figure 7.3.

Comparing these results with those of Figure 7.2, we find that there was substantial bias in many of the estimates. With the exception of *Pop* and *G*, the other estimates changed substantially in terms of magnitude and significance. In fact, no longer are *I* and *FDI* significant determinants of economic growth across these countries.

There are three caveats that should be mentioned with regard to the within results. First, the reader will notice that although we performed the within transformation, the output in Figure 7.3 still has a coefficient estimate for _cons—that is, it still estimated an intercept term. The researcher can actually drop this from the regression with a nocons command at the end of the regression line; but, it's not necessary. The reason comes directly from a Stata webpage, "The results that (the appropriate Stata command) *xtreg, fe* reports have simply been reformulated so that the reported intercept is the average value of the fixed effects (if they were explicitly modeled) (http://www.stata.com/support/faqs/statistics/intercept-in-fixed-effects-model/)." Therefore, there is no need to worry in our case; the coefficient is reported simply for convenience.

Second, the sample correlation coefficient in Figure 7.3, also known as the R^2, is not entirely accurate; and to a great extent, this inaccuracy transcends regression packages. This R^2 is calculated using equation (7.4), where there are no dummy variables representing the countries—they've been subtracted out before hand. Hence, the within R^2 actually reflects

the fit of the regression as it pertains to the right-hand-side x's and not any intercepts. Because of this, if the researcher is truly interested in an accurate R^2, they will have to run a regression of the form (7.1) and use that R^2 value instead. The slope values will be the same as in Figure 7.3, but that R^2 will be the more accurate one.

The final caveat is the fact that the degrees of freedom are not obvious. For a researcher who wants to perform subsequent testing by hand (i.e., programmed testing that is not canned in the econometric software package), this can be a huge issue. Typically, the degrees of freedom for an OLS type of regression equal the number of observations minus the number of right-hand-side variables plus the intercept. In a cross-sectional model, if I equals the number of observations and k equals the number of right-hand-side variables, then the degrees of freedom would be $I- k -1$. The 1 in this case represents the intercept. In a rudimentary panel model, the number of observations would equal IT, where T reflects the total number of time-period observations and I reflects the number of i groups, in our case here these are individual countries. Therefore, without correcting for intercept heterogeneity, the degrees of freedom would be $IT - k -1$. But when running a regression of the form (7.1), we must subtract out all of the intercepts, not just one of them. So the degrees of freedom in this case would be $IT - k -I$. Believe it or not, the degrees of freedom for equation (7.4) is exactly the same as it is for (7.1) even though there are not any actual intercepts modeled in (7.4); this is why any subsequent programming by hand must account for these larger degrees of freedom.

Moving back to the purpose of this chapter, that is, heterogeneity testing, there is one last heterogeneity issue that must be tested for when using panel data that isn't an issue for cross-sectional data, and that is heterogeneity as a function of t. Time heterogeneity, as I like to call it, simply means that y conditioned on the x's is trending either upward or downward over time, on average, for each i. In this case of heterogeneity, equation (7.1) would look like

$$y_{it} = a_{i0} + a_1 x_{it} + a_2 t_i + e_{it}. \tag{7.5}$$

Testing for this form of misspecification is easy as all you need to do is construct a discreet ordinal variable that starts at one and continues to the

```
. xtreg Growth I FDI School Trade Pop G t, fe robust

Fixed-effects (within) regression               Number of obs      =      3758
Group variable: groups                          Number of groups   =       172

R-sq:  within  = 0.0322                         Obs per group: min =         2
       between = 0.0537                                        avg =      21.8
       overall = 0.0241                                        max =        41

                                                F(7,171)           =      6.39
corr(u_i, Xb)  = -0.3783                        Prob > F           =    0.0000

                                 (Std. Err. adjusted for 172 clusters in groups)
------------------------------------------------------------------------------
             |               Robust
      Growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |  -.0176229   .0291792    -0.60   0.547    -.0752206    .0399749
         FDI |   .0468228   .0311892     1.50   0.135    -.0147425    .1083882
      School |  -.0349342   .0130498    -2.68   0.008    -.0606937   -.0091748
       Trade |   .0309677   .0102923     3.01   0.003     .0106513     .051284
         Pop |  -.6983936   .1705763    -4.09   0.000      -1.0351   -.3616871
           G |  -.0837694   .0408519    -2.05   0.042    -.1644084   -.0031303
           t |   .0130129   .0230753     0.56   0.574    -.0325361    .0585619
       _cons |   4.467535   1.156746     3.86   0.000     2.184196    6.750875
------------------------------------------------------------------------------
```

Figure 7.4 Testing for time heterogeneity

end of each *i*'s time span. One then simply includes this in their heteroskedasticity and intercept-corrected regression and evaluates whether its coefficient is statistically significant or not. In our case, the coefficient for *t* in Figure 7.4 is highly insignificant indicating that on average, growth is not trending upward or downward over time. Therefore, we can drop it from our subsequent analyses. And it actually makes sense when one thinks about the dynamics of economic growth. *Growth* is not a stock variable, it is a flow variable. Trending would mean that *Growth* either increases in a linear way without bound, or decreases in the same fashion; but *Growth* (i.e., the percentage change in GDP) will always be bounded in some way whether by business cycle dynamics or because the long-run trend in GDP cannot be exponential. On the other hand, if we were to model a country's level of GDP, that is, a stock variable, it would indeed trend in a deterministic way. Hence, a researcher should be aware of the type of variable they are using and whether they would be expecting it to trend or not.

Dependent Variable Dynamics

Failing to control for dependent variable dynamics in panel regressions is perhaps the second most critical misspecification issue; the most critical being statistical omitted variable bias which we will investigate next. The reason that controlling for dynamics is so critical is simply that nearly

every continuous regressand, or *y*, related to economics has a cyclical component to it. In our case, annual growth in per capita real GDP actually defines, and in the short run is defined by, a business cycle. Even variables like migration have cyclical components to them. And even though ignoring the cyclical component itself does not result in biased estimators, if the right-hand-side variable of interest responds to this dynamic, then a correlation between the regressor and the error of the model would result. Furthermore, simply as a case of drawing the purest inference possible from the true relationship between the variable of interest and the dependent variable, subtracting out the effects of the dynamic component in *y* is critical.

Mathematically, dependent variable dynamics can be modeled as

$$y_{it} = a_{i0} + b_1 y_{it-1} + a_1 x_{it} + e_{it} \tag{7.6}$$

The dynamic component comes in through y_{it-1}. I've allowed for one lag to start our particular case. But more than 1 year should also be tested for if the coefficient estimate for the first lag, b_1, is larger than 0.500. For data that has a higher frequency, however, you may need to include even more lags regardless of the coefficient value of the first lag.

Modeling the dynamic component of y_{it} may look straightforward—that is, include one lag of itself on the right-hand side of the equation and simply run the regression; but one significant problem arises when this operation is performed using a within regression. Assume we run a standard within regression like (7.4), but include a lagged dependent variable on the right-hand side as in (7.7)

$$(y_{it} - \overline{y}_i) = b_1(y_{it-1} - \overline{y}_i) + a_1(x_{it} - \overline{x}_i) + (e_{it} - \overline{e}_i). \tag{7.7}$$

It is apparent from (7.7) that the $\overline{y}_i$ on the right-hand side is correlated with the $\overline{e}_i$ (revisit equation (7.2) from earlier in this chapter and the correlation is easy to see). This means that there will exist dependence between the residuals and transformed lagged dependent variable. Fortunately, overcoming this obstacle is easy although another issue arises, which we will address later.

Anderson and Hsaio (1982) recognized that the best way to treat this issue is to take first differences of (7.6) rather than use a within transformation method. Model (7.7) will then look like

$$(y_{it} - y_{it-1}) = b_1(y_{it-1} - y_{it-2}) + a_1(x_{it} - x_{it-1}) + (e_{it} - e_{it-1}) \tag{7.8}$$

But this form presents another issue—the fact that now y_{it-1} is correlated with e_{it-1}. One way to overcome this problem is to instrument $(y_{it-1} - y_{it-2})$ with y_{it-2} so that (7.8) now becomes

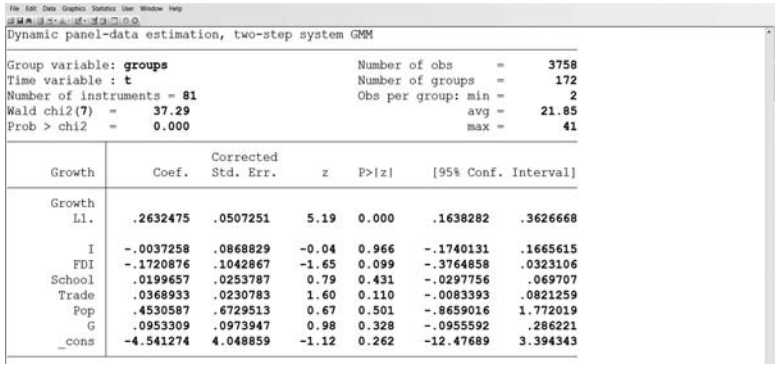
$$(y_{it} - y_{it-1}) = b_1 y_{it-2} + a_1(x_{it} - x_{it-1}) + (e_{it} - e_{it-1}). \tag{7.9}$$

Another way is to instrument $(y_{it-1} - y_{it-2})$ with $(y_{it-2} - y_{it-3})$ so that (7.8) now becomes

$$(y_{it} - y_{it-1}) = b_1(y_{it-2} - y_{it-3}) + a_1(x_{it} - x_{it-1}) + (e_{it} - e_{it-1}). \tag{7.10}$$

Most researchers performing this operation simply use the second lagged level y_{it-2} and run a regression of the form (7.9). Having said this, a more sophisticated method that takes advantage of the maximum number of moments is something called a dynamic panel generalized method of moments (GMM) operation that uses both lagged levels as well as lagged differences as instruments (Arellano and Bond 1991; Arellano and Bover 1995; Blundell and Bond 1998). There are a few technical issues when performing this operation that one should be aware of. I suggest that a researcher using this methodology consult Roodman 2006, 2009.

In our case, since I've been using this methodology many years now, I'm about as close to an expert in dynamic panel modeling as one can get (other than Roodman, of course). To this end, I will simply portray the output of our model and the reader can have confidence that the technical details are appropriately addressed during the operation. Having said that, even though many packages such as Stata now have this feature preprogrammed into their software, the programmer will need to remain vigilant that the technical details (most of which are outlined in Roodman's manuscripts) are within the bounds of reasonable statistical



Dynamic panel-data estimation, two-step system GMM

Group variable: **groups**		Number of obs	= 3758
Time variable : **t**		Number of groups	= 172
Number of instruments = **81**		Obs per group: min	= 2
Wald chi2(**7**) = **37.29**		avg	= 21.85
Prob > chi2 = **0.000**		max	= 41

Growth	Coef.	Corrected Std. Err.	z	P>|z|	[95% Conf.	Interval]
Growth L1.	.2632475	.0507251	5.19	0.000	.1638282	.3626668
I	-.0037258	.0868829	-0.04	0.966	-.1740131	.1665615
FDI	-.1720876	.1042867	-1.65	0.099	-.3764858	.0323106
School	.0199657	.0253787	0.79	0.431	-.0297756	.069707
Trade	.0368933	.0230783	1.60	0.110	-.0083393	.0821259
Pop	.4530587	.6729513	0.67	0.501	-.8659016	1.772019
G	.0953309	.0973947	0.98	0.328	-.0955592	.286221
_cons	-4.541274	4.048859	-1.12	0.262	-12.47689	3.394343

Figure 7.5 Dynamic panel regression

outcomes. Some of these preprogrammed versions of the GMM method are not as flexible as the user written versions are. Ultimately, if the researcher is not comfortable using this procedure, they should consider using the Anderson and Hsaio method instead.

Controlling for cyclical components in *Growth* we get the output in Figure 7.5. First, the coefficient for lagged *Growth*, labeled as *Growth L1*, is highly significant meaning that there is a dynamic component to *Growth*. Comparing the results in Figure 7.5 with those of Figure 7.3, we find that dependent variable dynamics did indeed play a role in influencing the estimates of the other coefficients. The coefficient for *FDI* changed signs and is now marginally statistically significant. On the other hand, whereas *School*, *Trade*, *Pop*, and *G* were significant before, no longer are they. What this means is that the other variables were simply picking up (or mimicking if you will) the dynamic component of *Growth*, obscuring their true long-run relationship with *Growth*. (For the more sophisticated reader, the Hansen test for overriding restrictions returned a *p*-value of 0.442, the AB Test for AR(1) returned a *p*-value of 0.000 as expected, and that for AR(2) is 0.201. All of these are well within the bounds of reasonable test statistics for the GMM process. Problems would exist if these test statistics approached their bounds of 0.000 and 1.000. Only one lag was used to instrument lagged *Growth*, resulting in 81 instruments; this is far less than the number of groups which is 172.)

At this point it would interest the reader to keep in mind that the dynamic panel GMM regression we just ran already controls for

endogeneity between all of the other variables and current growth (see model (4.2) in Chapter 4). Because we lagged *I*, *FDI*, *School*, *Trade*, *Pop*, and *G* prior to putting them into our regression, there is no need to use the GMM methodology with these variables. Using this method on all of the right-hand-side variables would lead to a large number of instruments, overspecification, and significant problems with test statistics and statistical inference. And even though the GMM methodology has variations of it that can address this issue (such as the *collapse* function which respecifies matrices into simpler forms), to realize the full potential of such a method, and unless you are very familiar with this routine, it is not recommended that these procedures be performed. However, if the researcher has a particular variable of interest, say *FDI*, it may behoove that person to perhaps keep that variable in its contemporaneous form and perform the GMM dynamic method to it as well, while keeping the other variables as simple one-period lags. The choice is the researchers to make; but again, be aware of using too many instruments and the problems it poses (see Roodman's work mentioned earlier for a full explanation of these issues).

Slope Heterogeneity

It's time now that we go back to exploring heterogeneity, but this time in the slope and not the intercepts. As mentioned in Chapter 2, slope heterogeneity is much like intercept heterogeneity but when the slope differs in a statistically significant way across some obvious clustering of the dependent variable. Rewriting equation (2.7) in Chapter 2 we would have

$$y = a_0 + a_1 D_1 x + a_2 D_2 x + e, \tag{7.11}$$

where D_1 is a dummy variable that equals 1 when characterizing one group and 0 otherwise, and D_2 another. And while (7.11) is the best for conducting the initial tests of differences in slopes as will be obvious below, a more common form is

$$y = a_0 + a_1 x + a_2 D x + e, \tag{7.12}$$

where D is a dummy variable that is equal to 1 for some cluster and equal to 0 for all other clusters. Therefore, the slope for a "control" group would simply be a_1 but the slope for the delineated cluster would be $a_1 + a_2$. Hence, a_2 measures the difference between the control group's slope and that for the delineated cluster.

In our case, an obvious clustering of our dependent variable, *Growth*, would be by income group—the same clustering we explored for intercept and slope heterogeneity in the cross-sectional case. Let us now perform the same procedure on the panel data model.

There are interesting results in Figure 7.6. First, as expected, economies of all developmental levels have robust business cycles as the coefficients on all of the lagged *Growth* variables are highly significant. Also, it seems as though the business cycle in high-income countries has more memory as the value of its lagged *Growth* coefficient is more than twice as large as middle and lower income economies. What's also interesting is the fact that only high-income *FDI*, high-income *Trade*, and middle-income *G*, have significant coefficients—only about one-half of the number of significant coefficients than the cross-sectional case. That said, we have to remember that this is a fixed effects regression that accounts for changes in intercepts at a far smaller level, that is, country level, than we accounted for in the cross-sectional model, that is, income groups. To that end, we are accounting for far more heterogeneity in the data and this could be washing out the differences in the effects of the individual

Prob > chi2 = 0.000 max = 41

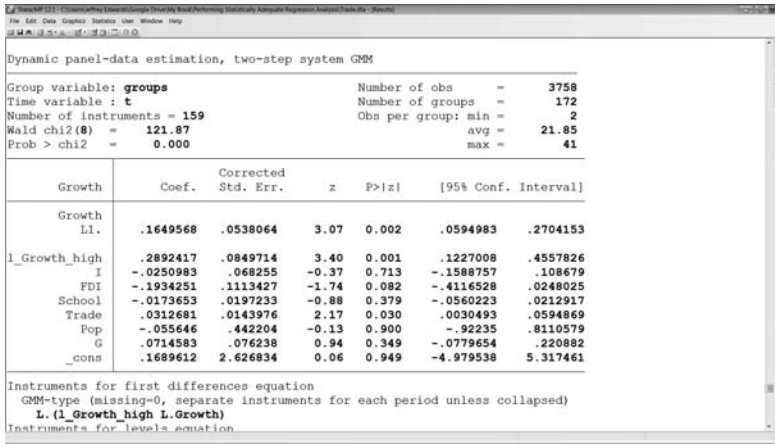
Growth	Coef.	Corrected Std. Err.	z	P>|z|	[95% Conf.	Interval]
l_Growth_high	.4572006	.0662274	6.90	0.000	.3273972	.587004
l_Growth_mid	.1761421	.0658846	2.67	0.008	.0470106	.3052736
l_Growth_low	.1463185	.0683204	2.14	0.032	.012413	.2802239
I_high	-.0637614	.070691	-0.90	0.367	-.2023133	.0747905
FDI_high	-.2468248	.1397715	-1.77	0.077	-.520772	.0271223
School_high	.0260575	.0279219	0.93	0.351	-.0286683	.0807834
Trade_high	.0541589	.024223	2.24	0.025	.0066827	.101635
Pop_high	-.2109107	.6945064	-0.30	0.761	-1.572118	1.150297
G_high	-.0299617	.1300767	-0.23	0.818	-.2849073	.2249839
I_mid	-.1082442	.0976448	-1.11	0.268	-.2996245	.083136
FDI_mid	.002523	.205658	0.01	0.990	-.4005593	.4056054
School_mid	.0289363	.0269886	1.07	0.284	-.0239604	.0818329
Trade_mid	.0144573	.0160119	0.90	0.367	-.0169254	.04584
Pop_mid	.9408488	.6551163	1.44	0.151	-.3431555	2.224853
G_mid	.2102305	.1110858	1.89	0.058	-.0074938	.4279548
I_low	.067466	.1565904	0.43	0.667	-.2394456	.3743776
FDI_low	.0239329	.2058047	0.12	0.907	-.3794369	.4273026
School_low	-.0705439	.0645364	-1.09	0.274	-.1970329	.0559451
Trade_low	.0568556	.0703844	0.81	0.419	-.0810952	.1948064
Pop_low	-.3364707	.5757211	-0.58	0.559	-1.464863	.7919219
G_low	.1245207	.1629143	0.76	0.445	-.1947855	.443827
_cons	-2.964906	3.140766	-0.94	0.345	-9.120694	3.190882

Figure 7.6 Slope heterogeneity regression

economic components like *I*, *School*, and so on. However, all of this discussion is meaningless until we actually test for equality across the slopes.

Testing the equality of the slope coefficients for the panel model yields *p*-values from 0.192 for the *Pop* coefficients to 0.623 for the *I* coefficients. These tests tell us that there is no statistically significant difference in slope coefficient estimates across the income groups for any of the *x* variables. However, the same *p*-value for the lagged *Growth* coefficients is 0.001, telling us that business cycle memory does indeed differ across income groups. The coefficient for lagged growth of the high-income group is 0.457, while those for the middle- and low-income groups, respectively, are 0.176, and 0.146; in other words, the middle- and low-income groups are quite close, while the high-income group, as mentioned earlier, is much larger. Therefore, performing a test of equality on the lagged *Growth* coefficients for the low- and middle-income groups we get a *p*-value of 0.757; hence, we will combine these countries. The final panel model we will move forward with is that in Figure 7.7.

One result we see in this permutation of the panel model estimates is probably expected, while the other is a direct result of correcting for slope heterogeneity. First, high-income nations generally have business cycles that contain more memory. In other words, this period's growth is more reliant on last period's growth than for middle and lower income



```
Dynamic panel-data estimation, two-step system GMM

Group variable: groups                          Number of obs      =      3758
Time variable : t                               Number of groups   =       172
Number of instruments = 159                     Obs per group: min =         2
Wald chi2(8)  =    121.87                                      avg =     21.85
Prob > chi2   =     0.000                                      max =        41

------------------------------------------------------------------------------
      Growth |      Coef.   Corrected Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Growth |
         L1. |   .1649568   .0538064     3.07   0.002     .0594983    .2704153
             |
l_Growth_high|   .2892417   .0849714     3.40   0.001     .1227008    .4557826
           I |  -.0250983    .068255    -0.37   0.713    -.1588757     .108679
         FDI |  -.1934251   .1113427    -1.74   0.082    -.4116528    .0248025
      School |  -.0173653   .0197233    -0.88   0.379    -.0560223    .0212917
       Trade |   .0312681   .0143976     2.17   0.030     .0030493    .0594869
         Pop |  -.055646    .442204    -0.13   0.900      -.92235    .8110579
           G |   .0714583   .076238     0.94   0.349    -.0779654     .220882
       _cons |   .1689612  2.626834     0.06   0.949    -4.979538    5.317461
------------------------------------------------------------------------------
Instruments for first differences equation
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L.(l_Growth_high L.Growth)
Instruments for levels equation
```

Figure 7.7 Regression corrected for slope heterogeneity

countries. This makes sense as higher income economies tend to be more developed and less erratic. Second, if we remember back in Figure 7.5 only the coefficient for *FDI* was significant, and marginally at that. Here, however, the *p*-value for this coefficient fell by about 17 percent, from 0.099 to 0.082, making it even more significant. Since adding the high income lagged *Growth* variable is all we did differently from the regression model that generated the results in Figure 7.5, this fall in *p*-value is a direct result of that misspecification issue. But that's not the only one. The reader will also notice that the coefficient estimate for *Trade* was 0.036 in Figure 7.5, now it's 0.031—not much of a change in estimates. However, the *p*-value has fallen substantially from 0.110 to 0.030—a fall of about 70 percent! Now, *Trade* does indeed have a significant effect on *Growth*.

Statistical Omitted Variable Bias

I speak about statistical omitted variable bias in my other book (Edwards 2013), and how it differs from what I call theoretical omitted variable bias; I also describe it earlier in Chapter 2 of this book. Most researchers lump them both together and call them simply omitted variable bias, but I like to differentiate theoretical bias from the more objective and purely statistical bias. A perfect example of statistical omitted variable bias is the inclusion of a squared *x*. If a squared *x* is needed but not included on the right-hand side prior to running the regression, bias in the relationship of interest would result. And while we could hypothesize why there would exist a nonlinear relationship between *x* and *y*, for the most part, testing for it is simply to paint a more statistically accurate picture of the relationship with *Growth*. Theoretical omitted variable bias is different. The idea of it only relies on theory. We may have such bias if (1) there exists a variable that we have access to that is correlated with *Growth*, and (2) it is simultaneously correlated with our variable of interest. Both (1) and (2), of course, are based purely on theory and are nearly always likely to exist with at least one variable that is not in your regression. In this sense, as long as any residual at all exists from our regression, our estimators will always be biased! This is why, in my opinion, theoretical omitted variable bias is a concept that has tenuous argumentative support at best.

Figures 2.9 to 2.11 show perfectly a hypothetical nonlinear relationship between *x* and *y*; but when there are more than one *x*, pictorially testing for statistical bias can be problematic. Furthermore, statistical omitted variable bias actually involves more than just the quadratic specification issue, it also involves interactions. A regression model addressing the former would look like equation (2.12) in Chapter 2, or

$$y = a_0 + a_1 x_1 + a_2 x_1^{\,2} + e. \tag{7.13}$$

But a regression model addressing the latter would look like

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 + e, \tag{7.14}$$

where by x_1 and x_2 are different *x*'s.

The reason why I only presented the quadratic case in Chapter 2 is because I know from experience that empirically testing for the interaction of variables can lead to outcomes that are theoretically tenuous, and therefore difficult if not impossible for the researcher to explain. For instance, assume we found that domestic investment interacts with government spending such that the marginal effect of *I* on *Growth* is

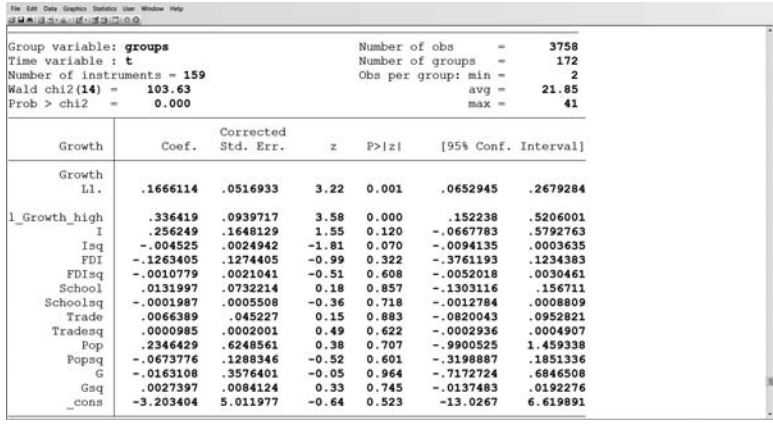
$$\frac{\partial Growth}{\partial I} = a_1 + a_2 G \tag{7.15}$$

meaning that the marginal effect is a function of government spending. Well, this is easy to explain since we know from basic macroeconomics that the two can affect each other through the crowding out type of arguments. Hence, these interactions can easily translate to domestic investment's effect on growth. But, what if we found that *Trade* and *School* interacted with each other, thereby affecting their own relationships with *Growth*? That would be much harder to explain. To this end, we only explore the quadratic issue in this book and not interactions. But again, please be aware that this problem can indeed exist and should be explored. In my own work, I empirically explore these interactions only if there are strong theoretical reasons why they should exist.

The most common way of testing for quadratic relationships is to simply include squared *x*'s into the regression and evaluate the statistical significance of their coefficients. Notice that I did not mention the inclusion of squared lagged *y*'s for the panel regression. This is because the coefficient to lagged *y* is already picking up a dynamic nonlinear relationship by definition as long as the coefficient estimate is less than 1.000 in absolute value. Furthermore, what exactly does it mean when we say that current economic growth is a function of squared lagged economic growth? I'm not sure anyone can answer that question. Hence, there is no need to include squared lagged dependent variables when performing this test.

Performing the same operation on the panel model, the output in Figure 7.8 tells us that only *I* has a significant nonlinear effect on *Growth*. But before we attempt to draw inference from this result, let us drop the other squared terms, leaving us with a far more parsimonious form. This we see in Figure 7.9. Investigating the simpler model in Figure 7.9, we find that *FDI* and *Trade* continue to have a significant effect on *Growth*. Furthermore, high-income economies continue to have a different business cycle to middle- and low-income economies. However, the interesting inference lies in the coefficient estimates for *I* and *Isq*.

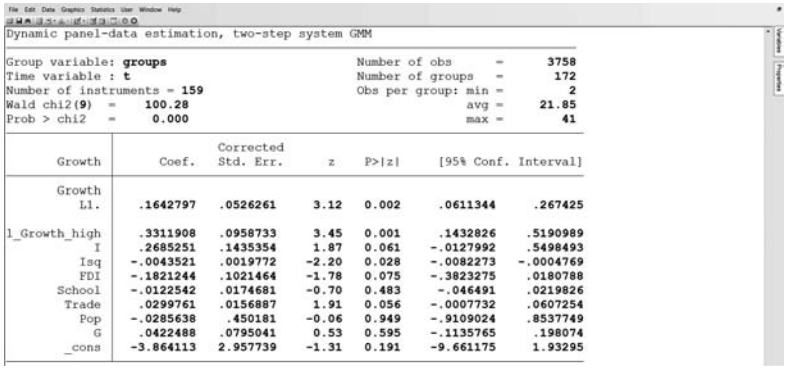
From Figure 7.7 we found that the effect *I* has on *Growth* is insignificant. The results in Figure 7.9 paint a different picture. Since the effect is quadratic and concave downward, there will exist a maximum. We know it's concave downward because the second derivative of *Growth* with



Group variable: **groups** Number of obs = **3758**
Time variable : **t** Number of groups = **172**
Number of instruments = **159** Obs per group: min = **2**
Wald chi2(**14**) = **103.63** avg = **21.85**
Prob > chi2 = **0.000** max = **41**

Growth	Coef.	Corrected Std. Err.	z	P>|z|	[95% Conf.	Interval]
Growth L1.	.1666114	.0516933	3.22	0.001	.0652945	.2679284
l_Growth_high	.336419	.0939717	3.58	0.000	.152238	.5206001
I	.256249	.1648129	1.55	0.120	-.0667783	.5792763
Isq	-.004525	.0024942	-1.81	0.070	-.0094135	.0003635
FDI	-.1263405	.1274405	-0.99	0.322	-.3761193	.1234383
FDIsq	-.0010779	.0021041	-0.51	0.608	-.0052018	.0030461
School	.0131997	.0732214	0.18	0.857	-.1303116	.156711
Schoolsq	-.0001987	.0005508	-0.36	0.718	-.0012784	.0008809
Trade	.0066389	.045227	0.15	0.883	-.0820043	.0952821
Tradesq	.0000985	.0002001	0.49	0.622	-.0002936	.0004907
Pop	.2346429	.6248561	0.38	0.707	-.9900525	1.459338
Popsq	-.0673776	.1288346	-0.52	0.601	-.3198887	.1851336
G	-.0163108	.3576401	-0.05	0.964	-.7172724	.6846508
Gsq	.0027397	.0084124	0.33	0.745	-.0137483	.0192276
_cons	-3.203404	5.011977	-0.64	0.523	-13.0267	6.619891

Figure 7.8 Regression checking for statistical omitted variable bias



Dynamic panel-data estimation, two-step system GMM

Group variable: **groups** — Number of obs = **3758**
Time variable : **t** — Number of groups = **172**
Number of instruments = **159** — Obs per group: min = **2**
Wald chi2(**9**) = **100.28** — avg = **21.85**
Prob > chi2 = **0.000** — max = **41**

Growth	Coef.	Corrected Std. Err.	z	P>|z|	[95% Conf.	Interval]
Growth L1.	.1642797	.0526261	3.12	0.002	.0611344	.267425
l_Growth_high	.3311908	.0958733	3.45	0.001	.1432826	.5190989
I	.2685251	.1435354	1.87	0.061	-.0127992	.5498493
Isq	-.0043521	.0019772	-2.20	0.028	-.0082273	-.0004769
FDI	-.1821244	.1021464	-1.78	0.075	-.3823275	.0180788
School	-.0122542	.0174681	-0.70	0.483	-.046491	.0219826
Trade	.0299761	.0156887	1.91	0.056	-.0007732	.0607254
Pop	-.0285638	.450181	-0.06	0.949	-.9109024	.8537749
G	.0422488	.0795041	0.53	0.595	-.1135765	.198074
_cons	-3.864113	2.957739	-1.31	0.191	-9.661175	1.93295

Figure 7.9 Regression after correcting for statistical omitted variable bias

respect to I is negative. The real question is whether the maximum lies within the relevant sample space of I. To this end, we take the derivative of the function

$$Growth_{it} = 0.268 I_{it} - 0.004 I_{it}^{2} \tag{7.16}$$

to get

$$\frac{\partial Growth}{\partial I} = 0.268 - 0.008 I_{it}. \tag{7.17}$$

We then set this equal to zero and solve for I to find our maximum. The maximum of this function is at I_{it} = 33.5, which is well within our relevant sample space. Since I is in percentages in our case, this means that increases in (lagged) investment positively impacts economic growth at a decreasing rate for countries that have investment to GDP below 33.5 percent, but further increases in (lagged) investment negatively impacts growth at an increasing rate for countries with higher levels of investment. What this means is that if we were to draw inference from the linear effects in Figure 7.7, that inference would be incorrect.

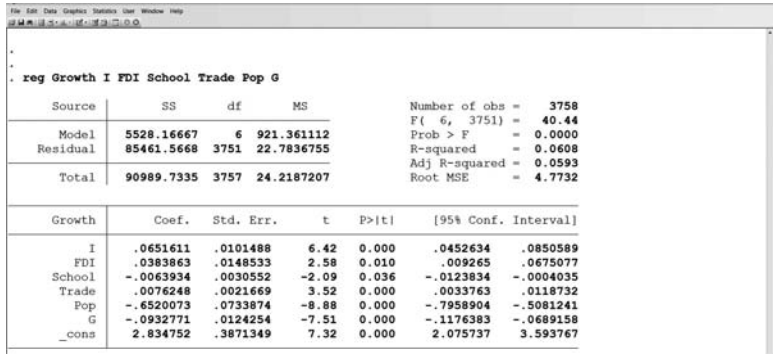
Final Panel Model and the Inference We Can Draw From It

To finish this chapter of the book, it is very apparent, just like it was in the cross-sectional case, that to have drawn accurate inference from the

original results depicted in Figure 7.1 would have been a premature task to perform. I put the two results together so we can analyze them simultaneously. Figure 7.1 is repeated in Figure 7.10, and Figure 7.9 is repeated in Figure 7.11.

If we were to stop drawing inference at the results from our base regression, Figure 7.10, we would have concluded that all of our right-hand-side variables affect economic growth in a statistically significant way. And they probably do; just not in the way we think they do, nor in the way our model was intending. To elaborate on this, I must bring in my own expertise in modeling cross-country growth in real per capita GDP.

When a researcher puts real per capita growth on the left-hand side of a regression function, the intention is to model long-run growth, not short-run growth. What the results in Figure 7.10 are capturing are the

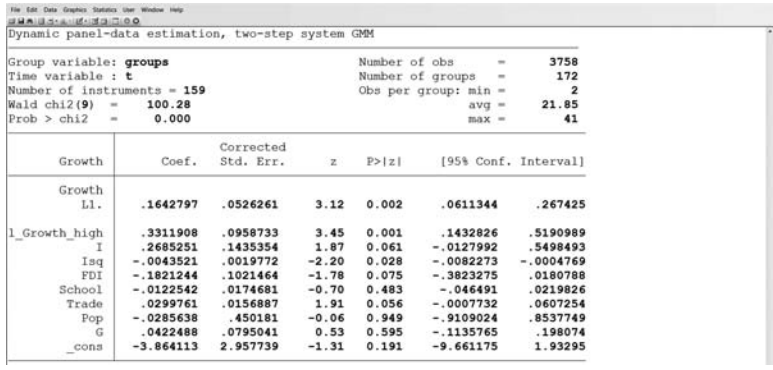


```
. reg Growth I FDI School Trade Pop G

      Source |       SS       df       MS              Number of obs =    3758
-------------+------------------------------           F(  6,  3751) =   40.44
       Model |  5528.16667     6  921.361112            Prob > F      =  0.0000
    Residual |  85461.5668  3751  22.7836755            R-squared     =  0.0608
-------------+------------------------------           Adj R-squared =  0.0593
       Total |  90989.7335  3757  24.2187207            Root MSE      =  4.7732

------------------------------------------------------------------------------
      Growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |   .0651611   .0101488     6.42   0.000     .0452634    .0850589
         FDI |   .0383863   .0148533     2.58   0.010      .009265    .0675077
      School |  -.0063934   .0030552    -2.09   0.036    -.0123834   -.0004035
       Trade |   .0076248   .0021669     3.52   0.000     .0033763    .0118732
         Pop |  -.6520073   .0733874    -8.88   0.000    -.7958904   -.5081241
           G |  -.0932771   .0124254    -7.51   0.000    -.1176383   -.0689158
       _cons |   2.834752   .3871349     7.32   0.000     2.075737    3.593767
------------------------------------------------------------------------------
```

Figure 7.10 Base regression results



```
Dynamic panel-data estimation, two-step system GMM

Group variable: groups                          Number of obs      =      3758
Time variable : t                               Number of groups   =       172
Number of instruments = 159                     Obs per group: min =         2
Wald chi2(9)  =    100.28                                      avg =     21.85
Prob > chi2   =     0.000                                      max =        41

------------------------------------------------------------------------------
      Growth |      Coef.   Corrected   z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Growth |
         L1. |   .1642797   .0526261     3.12   0.002     .0611344    .267425
             |
l_Growth_high |   .3311908   .0958733     3.45   0.001     .1432826    .5190989
           I |   .2685251   .1435354     1.87   0.061    -.0127992    .5498493
         Isq |  -.0043521   .0019772    -2.20   0.028    -.0082273   -.0004769
         FDI |  -.1821244   .1021464    -1.78   0.075    -.3823275    .0180788
      School |  -.0122542   .0174681    -0.70   0.483     -.046491    .0219826
       Trade |   .0299761   .0156887     1.91   0.056    -.0007732    .0607254
         Pop |  -.0285638   .450181     -0.06   0.949    -.9109024    .8537749
           G |   .0422488   .0795041     0.53   0.595    -.1135765    .198074
       _cons |  -3.864113   2.957739    -1.31   0.191    -9.661175    1.93295
------------------------------------------------------------------------------
```

Figure 7.11 Final regression results

long-run and short-run components of growth. But, after controlling for the short-run component through the lagged *Growth* variable, we can now determine what influences long-run growth; only domestic investment, foreign direct investment, and total trade volume as a percentage of GDP affect *Growth*; on the other hand, neither schooling, population growth, nor government spending influence long-run growth. But this begs the question, would we have expected *School*, *Pop*, or *G*, to impact short-run growth as the output in Figure 7.10 says they do? Absolutely not! Only *G* should influence short-run growth through fiscal policy. Both *School* and *Pop* are long-run determinants by definition because you can't change the amount of schooling a nation has, or its rate of population growth overnight like you can its level of government spending. This means that not only were the results in Figure 7.10 capturing both short-run and long-run influences on *Growth*, but they were also substantially biased meaning the inference a researcher would draw from them couldn't be trusted in the first place!

In general, unlike cross-sectional models, we have seen that panel models are far more dynamic and close attention must be paid to the plethora of possible misspecification issues contained within these models. To have simply stopped at the simple linear form would have been an injustice to the realm of macroeconomic growth and led policymakers down the wrong path for their countries. But it isn't just with this data set that one must explore these issues; unfortunately, all data sets are just as likely to contain them. Microeconomists, macroeconomists, those performing medical research in the fields of psychology, medicine, neuroscience, biology, or any empirical field for that matter, must pay very close attention to model misspecification, and at the very least, to the items outlined in this chapter. When one wonders why data doesn't hold up to theory, misspecification is usually the answer that is the most quickly ignored.

Moving forward throughout the remainder of the book we will use the model that generated the results found in Figure 7.11 only when considering the next two topics—variance heterogeneity and consistency in panels. As will also become apparent, the topics of consistency and dynamic parametric heterogeneity should not be taken in sequence as the data and conditioning sets will change substantially, rendering the final model found in Chapter 7 incomparable.

CHAPTER 8

Variance Heterogeneity

The Panel Data Case

If we remember from Chapter 3, variance heterogeneity can rarely be detected with a test for homoskedasticity. To reiterate equation (3.5) from that chapter, we have

$$r^2 = c_0 + c_1 D + \mu, \tag{8.1}$$

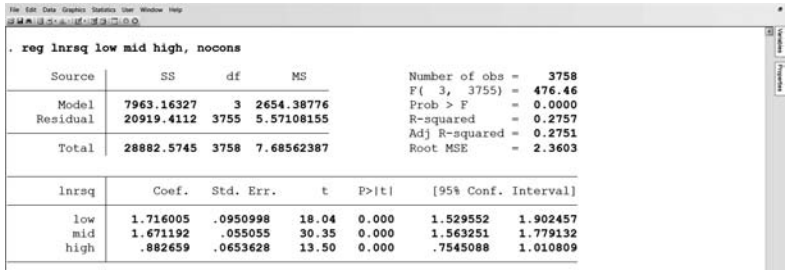
which tells us that in the case of variance heterogeneity, a correlation would exist between the squared residuals and a set of dummy variables representing a particular clustering of the data, much like in intercept and slope heterogeneity. However, we run into an issue if we perform a regression specified exactly like (8.1). That is because the dependent variable is a squared version of a normally distributed variable (at least we hope it's normally distributed). Park (1966) recommended using the natural log of the squared residuals instead, resulting in the regression

$$\ln(r^2) = c_0 + c_1 D + v. \tag{8.2}$$

This is the representation we will explore.

Figure 8.1 shows the screenshot of our Stata output when we run a regression of the natural log of our squared residuals from the final model in Chapter 7 on the income-level dummy variables we created previously.

A test for the overall equality of the coefficients for the three income levels returns a *p*-value of 0.000 indicating that the coefficients are not equal to one another. Viewing the estimates in Figure 8.1, it seems as though the coefficients for the *low-* and *mid-* income groups are nearly

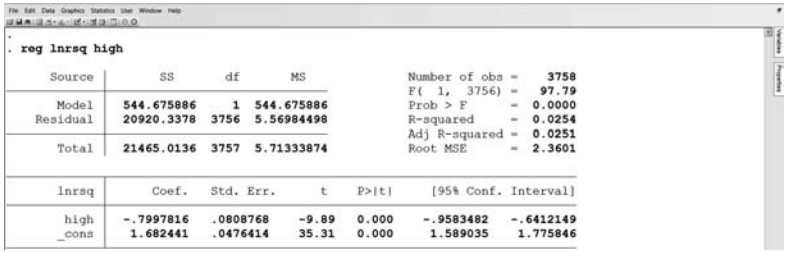


```
. reg lnrsq low mid high, nocons

      Source |       SS       df       MS              Number of obs =    3758
-------------+------------------------------           F(  3,  3755) =  476.46
       Model |  7963.16327     3  2654.38776            Prob > F      =  0.0000
    Residual |  20919.4112  3755  5.57108155            R-squared     =  0.2757
-------------+------------------------------           Adj R-squared =  0.2751
       Total |  28882.5745  3758  7.68562387            Root MSE      =  2.3603

------------------------------------------------------------------------------
       lnrsq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         low |    1.716005   .0950998    18.04   0.000     1.529552    1.902457
         mid |    1.671192    .055055    30.35   0.000     1.563251    1.779132
        high |     .882659   .0653628    13.50   0.000     .7545088    1.010809
------------------------------------------------------------------------------
```

Figure 8.1 Regression testing for variance heterogeneity



```
. reg lnrsq high

      Source |       SS       df       MS              Number of obs =    3758
-------------+------------------------------           F(  1,  3756) =   97.79
       Model |  544.675886     1  544.675886            Prob > F      =  0.0000
    Residual |  20920.3378  3756  5.56984498            R-squared     =  0.0254
-------------+------------------------------           Adj R-squared =  0.0251
       Total |  21465.0136  3757  5.71333874            Root MSE      =  2.3601

------------------------------------------------------------------------------
       lnrsq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        high |   -.7997816   .0808768    -9.89   0.000    -.9583482   -.6412149
       _cons |    1.682441   .0476414    35.31   0.000     1.589035    1.775846
------------------------------------------------------------------------------
```

Figure 8.2 Variance regression using* low *and* mid *as the control group

identical, while those for the *high* group are only about one-half of the others. This actually makes sense as it is well known that developed nations tend to have less volatile economies than emerging or developing countries.

Conducting a separate test on the equality of the *low-* and *mid-* income countries, we get a *p*-value of 0.683. This means that as expected, the *low-* and *mid-* coefficients are not significantly different from one another and their groups can be combined. Rerunning the same regression but lumping the *low-* and *mid-* countries together and using them as the control group, that is, letting that group act as the overall constant for the regression, we get the results in Figure 8.2.

While the results in Figure 8.1 reflected the actual levels of volatility for each income group with the *high* group having the lowest volatility, the constant in Figure 8.2 reflects the weighted average level of volatility across the *low* and *mid* groups, with the *high* group coefficient being the difference between the control group and the high-income countries. This means that on average, high-income countries have 0.799 percentage

points of lower volatility in real per capita GDP growth than volatility in lower income countries. The procedure now is to correct for this misspecification by performing a feasible generalized least squares procedure (FGLS) regression on the final model from Chapter 7.

Correcting for Variance Heterogeneity

The correction for variance heterogeneity will entail using a generalized least squares (GLS) type of method. Nearly all undergraduate textbooks address this methodology when used as an alternative to heteroskedasticity correction. But, GLS is only good if you know the true nature of the heterogeneity in the variance; in our case, however, we are estimating it. To this end, we must use a variant called an FGLS. This is perhaps the best way to correct for variance heterogeneity when it is known to be present (Edwards et al. 2006). The way someone would perform an FGLS in this case is as follows.

Using the procedure we just performed, that is, estimating the natural log of the squared residuals as

$$\widehat{\ln(r^2)} = \widehat{c_0} + \widehat{c_1}D \tag{8.3}$$

and specifically in our case we would have the function

$$\widehat{\ln(r^2)} = 1.682\,low\,mid - 0.799\,high \tag{8.4}$$

We would then take these estimates of In (r^2), and convert them to simply $\hat{r}$ by first exponentiating, and then taking the square root of what's left. To correct for heterogeneity, we would then weight all of the variables in our final *Growth* model, including the constant. Hence, reflecting on model (1) in Chapter 2, our new model would look like

$$\frac{y}{\hat{r}} = a_0 \frac{1}{\hat{r}} + a_1 \frac{x}{\hat{r}} + \frac{e}{\hat{r}}. \tag{8.5}$$

Completing this operation for the final regression from Chapter 7, and rerunning that regression, we get the output in Figure 8.3. For easier

File Edit Data Graphics Statistics User Window Help

Dynamic panel-data estimation, two-step system GMM

Group variable: **groups**		Number of obs	= 3758
Time variable : **t**		Number of groups	= 172
Number of instruments = **158**		Obs per group: min	= 2
Wald chi2(**9**) = **116.21**		avg	= 21.85
Prob > chi2 = **0.000**		max	= 41

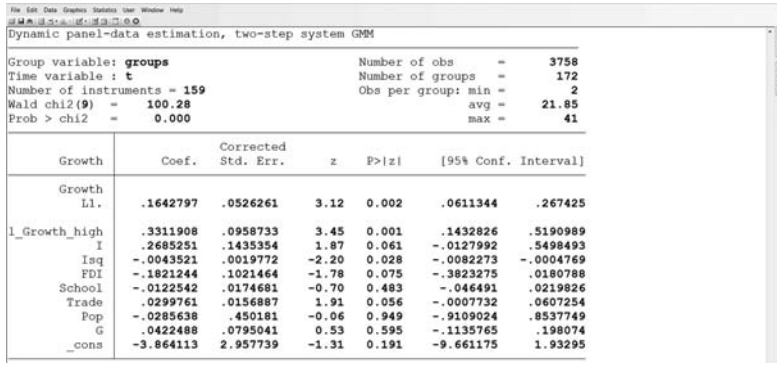
wGrowth	Coef.	Corrected Std. Err.	z	P>|z|	[95% Conf.	Interval]
wL1Growth	.1385385	.0534325	2.59	0.010	.0338128	.2432642
wl_Growth_high	.319041	.0886121	3.60	0.000	.1453644	.4927176
wI	-.0220514	.109085	-0.20	0.840	-.235854	.1917512
wIsq	-.0011041	.0015078	-0.73	0.464	-.0040593	.0018511
wFDI	-.157077	.1019572	-1.54	0.123	-.3569095	.0427556
wSchool	-.0055286	.0162229	-0.34	0.733	-.0373249	.0262677
wTrade	.0349019	.0139472	2.50	0.012	.0075658	.062238
wPop	-.3512925	.4302865	-0.82	0.414	-1.194638	.4920535
wG	.0092943	.0752447	0.12	0.902	-.1381825	.1567712

Figure 8.3 Regression corrected for variance heterogeneity

comparison, Figure 8.4 depicts the output from the final regression model generated in Chapter 7.

The "w" in front of each of the coefficient or variable names stands for that variable weighted by the $\hat{r}$ that we generated earlier. The reader should also keep in mind that Figure 8.4 reports a coefficient value for *_cons*, but Figure 8.3 doesn't. As discussed back in Chapter 7, *_cons* is the average value of the intercepts had they been estimated instead of subtracted out before hand by the within transformation. Therefore, when conducting the FGLS weighting procedure, there is no actual constant term to weight like there was for the cross-section case; it doesn't appear in the results of Figure 8.3 because it was purposely dropped when the regression was executed. I didn't want to add any confusion to the discussion of the output.

Now reporting on the results, not only are there substantial changes in the magnitudes of many of the coefficients, but also substantial changes in many of the coefficient's *p*-values when compared with the results from Chapter 7 in Figure 8.4. In fact, it seems that only *Trade* has an effect on *Growth* that is robust to the respecification of this model. The relatively large changes in coefficient magnitude are the main drawback of correcting for variance heterogeneity. Although asymptotically unbiased, after performing the FGLS procedure the coefficients do not maintain the nice finite sample properties that they had before the procedure was completed. To this end, the researcher must use discretion when evaluating



Dynamic panel-data estimation, two-step system GMM

Group variable: groups				Number of obs =	3758
Time variable : t				Number of groups =	172
Number of instruments = 159				Obs per group: min =	2
Wald chi2(9) = 100.28				avg =	21.85
Prob > chi2 = 0.000				max =	41

Growth	Coef.	Corrected Std. Err.	z	P>|z|	[95% Conf.	Interval]
Growth L1.	.1642797	.0526261	3.12	0.002	.0611344	.267425
l_Growth_high	.3311908	.0958733	3.45	0.001	.1432826	.5190989
I	.2685251	.1435354	1.87	0.061	-.0127992	.5498493
Isq	-.0043521	.0019772	-2.20	0.028	-.0082273	-.0004769
FDI	-.1821244	.1021464	-1.78	0.075	-.3823275	.0180788
School	-.0122542	.0174681	-0.70	0.483	-.046491	.0219826
Trade	.0299761	.0156887	1.91	0.056	-.0007732	.0607254
Pop	-.0285638	.450181	-0.06	0.949	-.9109024	.8537749
G	.0422488	.0795041	0.53	0.595	-.1135765	.198074
_cons	-3.864113	2.957739	-1.31	0.191	-9.661175	1.93295

Figure 8.4 Regression without correcting for variance heterogeneity

the benefits of variance heterogeneity correction, like that of correcting for substantial bias in *p*-values, versus the changes in the coefficient estimates themselves.

The rule of thumb I use is that if the magnitudes of the coefficients are important for forecasting specific levels of the left-hand-side variable, then I will probably ignore the heterogeneity because any change in the coefficient values may greatly change those forecasts. But if the actual values of the coefficients aren't important, and only the sign of the coefficient is important, then I will probably correct for variance heterogeneity.

To be honest, I have absolutely nothing to support my reasoning or justification for these bounds. It's simply a personal condition I apply to my own work. But it doesn't matter; as long as the researcher is aware of these issues and weighs the pros and cons of variance correction using the FGLS procedure, then one should be able to make an educated decision to perform it or not. However, this case is considerably more complicated as one of the relationships is quadratic in nature. Even though the investment part of the growth function continues to be concave downward, the new maximum exists at $I_{it} = -11$. This means that as a nation increases domestic investment, economic growth will fall at an increasing rate from the beginning! If we remember from Chapter 7, *I* increased growth up to about 33 percent of GDP, then decreased growth thereafter. Since we know without a doubt that investment increases growth at least over some range of investment (reference a standard production function, for

example), this latter result makes far more sense than the result we get after correcting for variance heterogeneity.

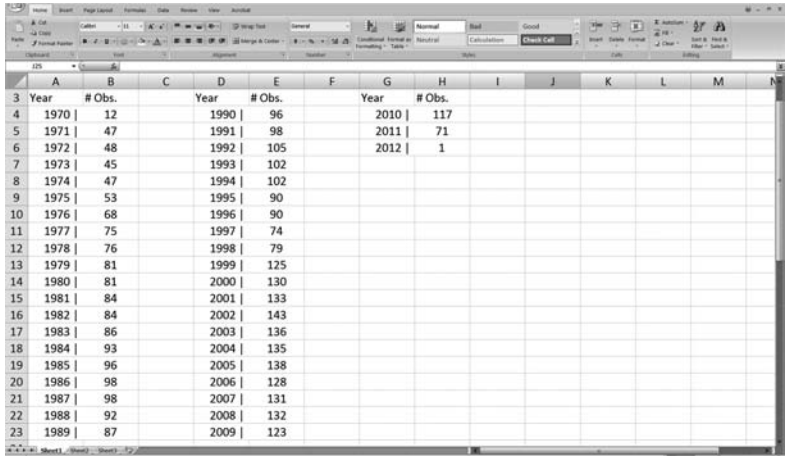
Given the inference we just conducted, I would conclude that correcting for variance heterogeneity in this case would not be the correct thing to do. I say this for two reasons. The first is simply because the amount of variation in the conditional variance that is explained by the income effects is quite low. When we look back at Figure 8.2, only about 2.5 percent of the variation in the conditional variance is explained with the inclusion of these dummy variables. The second reason I would forego the FGLS operation in this case is because the large changes to inference do not justify its use especially when so little heterogeneity is explained by these dummy variables. Having said that, once again, this situation is likely to be quite unique to the data set I am using. Any other permutation of this data set, much less using a completely different data set, might produce coefficient estimates that only change slightly, but produce *p*-values that change substantially. In cases such as this, it might behoove the researcher to use the respecified results. And even though coefficient estimates are considerably different in our particular case, at least we know that the standard errors are constant and conform to our original probabilistic assumptions we made at the beginning of our empirical work.

CHAPTER 9

Consistent and Balanced Panels

To reiterate the gist of the explanation given in Chapter 3, balanced panels simply mean that each *i* has the same number of *t* observations. In our case, this means that if one country covers 9 years' worth of data, then all countries cover 9 years' worth of data. Which years each country covers is not an issue with the concept of balanced panels; this is where consistent panels come into play. Consistency in panel data simply means that each of the balanced panels start in the same year and end in the same year. Therefore, you can have balanced but inconsistent panels, but you cannot have consistent but unbalanced panels.

There are perhaps three main reasons why researchers don't, or won't, recognize these concepts as actual misspecification. The first is the fact that researchers are only interested in drawing inference across *i* anyway, and not at all interested in the time dimensional inference from their estimates. Another is that by generating consistency, they may lose a large number of observations, and even a large number of individuals, hence dramatically changing their empirical experiment(s). And lastly, a researcher may not account for unbalanced and inconsistent panels because it is not a proper form of misspecification in a purely statistical sense. In other words, balancing of panels is not necessary to attain NIID. But researchers should indeed be aware that results obtained using unbalanced and inconsistent panels can generate false inference for the reasons outlined in Chapter 3. For the purpose of our exposition here, however, we will implement consistency in our panel data regardless of the loss in observations, groups, or both. Having said that, a loss of observations is certainly a concern that should be considered before taking on this misspecification issue.



	A	B	C	D	E	F	G	H	I	J	K	L	M
3	Year	# Obs.		Year	# Obs.		Year	# Obs.					
4	1970	12		1990	96		2010	117					
5	1971	47		1991	98		2011	71					
6	1972	48		1992	105		2012	1					
7	1973	45		1993	102								
8	1974	47		1994	102								
9	1975	53		1995	90								
10	1976	68		1996	90								
11	1977	75		1997	74								
12	1978	76		1998	79								
13	1979	81		1999	125								
14	1980	81		2000	130								
15	1981	84		2001	133								
16	1982	84		2002	143								
17	1983	86		2003	136								
18	1984	93		2004	135								
19	1985	96		2005	138								
20	1986	98		2006	128								
21	1987	98		2007	131								
22	1988	92		2008	132								
23	1989	87		2009	123								

Figure 9.1 Year-on-year frequency distribution of observations

Figure 9.1 is a frequency distribution of our panel data delineated by year. The left-hand side of each column represents the actual year of the observations, and the right-hand side of each column lists the number of countries that have observations for that year. The reader can see that with the exception of the two latest years of data, the closer one gets to the present the more countries per year one has. For instance, the year 1979 has 81 countries in it.

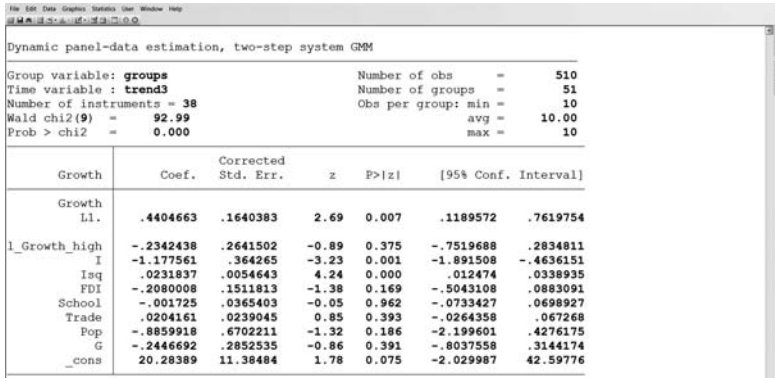
Currently, our regressions have included 172 countries. Obviously, none of the yearly cross sections have that many countries. Why is this? This is because many countries have data that either stops or begins toward the middle of the coverage of observations. This means that this frequency distribution isn't of much use. But, it does give us some indication of where to begin our balancing act.

Even though each type of panel data set is different, especially within the context of the empirical area (i.e., this is a cross-country macroeconomic data set, but the type of data for a microeconomist may have substantially different characteristics with regard to the i and t dimensions), in our specific case we would like to have at least 10 years' worth of data for each country. According to Figure 9.1, the number of countries starts to drop off significantly after about 2008; so to start, I'll drop all observations after that date. We also see that the number of yearly observations starts to substantially increase at about 1976—so I'll drop all observations

before that date. Then what I do is generate a variable with my econometric software that keeps track of the maximum number of observations per country (in my case with Stata, the command would incorporate the "egen *varname* = max(*trendvariable*)" function). I'll then use this variable to drop all countries that have less than the maximum number of observations that are equal to the spread between 1976 and 2008, or 33 years' worth of observations. I'll then run an arbitrary within regression to see how many countries are used in it, as well as how many observations there are per country. I'll then move the lower year from 1976 to 1977, and tell the maximum observation command to drop all countries with observations less than 32 years' worth of data, rerun the regression, and document the number of countries and observations per country this regression covered. I'll continue this procedure until (1) I reach the highest number of countries possible with more than 10 years' worth of data, or (2) I reach 10 years' worth of data for each country.

In our case, the best I could do is a data set with 10 years' worth of data covering 51 countries. If I let each country cover more years, I would have had fewer countries. If I wanted to have more countries, I would have had fewer years per country—as I said, it's a balancing act. It certainly becomes apparent why most researchers do not use consistent panels in their regressions. Doing so in our case has led to 121 fewer countries covered in our data set, and a loss of 3,248 observations! Furthermore, our results have changed substantially as we can see in Figure 9.2 compared with the final model from Chapter 7.

The reader will notice that no longer is the business-cycle memory for high-income countries different from low- and middle-income countries. This outcome actually makes sense as most countries with consistent data are likely to be relatively more developed than countries with sporadic data, and thereby have business cycles more like developed economies. On comparing these results with the final regression output from Chapter 7 shown in Figure 7.11, one will also find that the relationship between *I* and *Growth* remains quadratic, but has changed concavity. Before this relationship was concave downward with a maximum at $I_{it} = 33.5$; now this function is concave upward with a minimum at $I_{it} = 25.5$. Therefore, instead of *I* positively affecting *Growth* at a decreasing rate up to 33.5 percent of GDP, and negatively affecting *Growth* thereafter, it



Dynamic panel-data estimation, two-step system GMM

Group variable: **groups** Number of obs = **510**
Time variable : **trend3** Number of groups = **51**
Number of instruments = **38** Obs per group: min = **10**
Wald chi2(**9**) = **92.99** avg = **10.00**
Prob > chi2 = **0.000** max = **10**

Growth	Coef.	Corrected Std. Err.	z	P>|z|	[95% Conf.	Interval]
Growth L1.	.4404663	.1640383	2.69	0.007	.1189572	.7619754
l_Growth_high	-.2342438	.2641502	-0.89	0.375	-.7519688	.2834811
I	-1.177561	.364265	-3.23	0.001	-1.891508	-.4636151
Isq	.0231837	.0054643	4.24	0.000	.012474	.0338935
FDI	-.2080008	.1511813	-1.38	0.169	-.5043108	.0883091
School	-.001725	.0365403	-0.05	0.962	-.0733427	.0698927
Trade	.0204161	.0239045	0.85	0.393	-.0264358	.067268
Pop	-.8859918	.6702211	-1.32	0.186	-2.199601	.4276175
G	-.2446692	.2852535	-0.86	0.391	-.8037558	.3144174
_cons	20.28389	11.38484	1.78	0.075	-2.029987	42.59776

Figure 9.2 Regression using consistent panels

now negatively affects *Growth* at a decreasing rate up to 25.5 percent of GDP, and positively affects it at an increasing rate thereafter. And the final difference between this output and that in Figure 7.11 is the fact that no longer are the coefficients for *FDI* and *Trade* statistically significant, meaning that they no longer have a significant effect on *Growth*.

These changes in estimates are quite dramatic, which then begs the question—at this point in our exposition, should we have implemented a consistent panel procedure and compare it to earlier results? The answer is an emphatic no! The reason is simple. Unless the data set is balanced in its initial design, artificially balancing and creating consistent panels should only be done from the beginning of your research. Obviously, employing this method at this late stage is pointless, unless of course you just love to program econometric software. Is this really an important specification issue to consider then? Of course it is because as the reader has seen, the inference you would draw from a balanced and consistent panel would be dramatically different from an unbalanced panel. But, as mentioned before, sometimes the loss of data prevents one from using consistent panels in their research. Think about it this way. Assume you are a growth and development economist like myself. Is it more important to cover as many countries as possible in your empirical analyses and have as many observations as you can find, or is it more important to have consistency in a business cycle component across countries like we have here with all our data starting in 1999 and ending in 2008? Again, it simply depends on the question(s) the researcher is trying to answer.

CHAPTER 10

Dynamic Parametric Heterogeneity

Harking back to the explanation of dynamic parametric heterogeneity outlined in Chapter 3, typically researchers assume that all slope coefficients are constant over time. This means that we assume all relationships between x and y are also constant. In other words, given the simple model

$$y = a_0 + a_1(t)x + e, \tag{10.1}$$

researchers typically assume that

$$a_1(t) = a_1. \tag{10.2}$$

But instead, what if we actually had

$$\frac{\partial y}{\partial x} = a_1(t), \tag{10.3}$$

where

$$a_1(t) = b_0 + b_1 t. \tag{10.4}$$

Or, on the other hand, the marginal effect has a structural shift somewhere in time such that

$$a_1(t) = b_0 + b_1 D, \tag{10.5}$$

where D is a dummy variable representing some break in the marginal effect, perhaps delineated by a significant one-time event in the history of the relationship. In either of these cases the inference the researcher

would draw from their results would not be accurate if they assumed the relationship was constant when it obviously isn't. So, in testing for dynamic parametric heterogeneity, we are testing whether $b_1 = 0$. If it is, then this phenomena doesn't exist; if it isn't, then it does exist and should be modeled as such.

A simple type of regression known as a rolling window regression is sufficient to tell a researcher whether their coefficients are trending or shifting over time. The rolling windows methodology goes like this (much of it is from Edwards and Kasibhatla 2009). A rolling window algorithm is essentially a recursive-least-squares estimation that does not hold the initial observations stationary (Baum 2001; Spanos 1986). A regression of (10.1) is performed across all individuals, i, over a multiperiod interval starting at the beginning period of data. The coefficient value, $\widehat{a_1}(1)$, is then recorded. The modeler then lets the window roll one period such that the regression is run across all i over the same multiperiod interval but spanning the second period to the end of the multiperiod interval. The coefficient value, $\widehat{a_1}(2)$, is again recorded. The researcher then lets the window roll one more period so that the span now covers the third period to the end of the multiperiod interval, and records the coefficient estimate, $\widehat{a_3}(1)$. The process continues until all periods are exhausted and we are left with the full set of estimated coefficients

$$\widehat{a_1}(t) = \{a_1(1), a_1(2), a_1(3), \ldots . a_1(T)\}. \tag{10.6}$$

When using yearly data that is subject to cyclical behavior, such as economic growth, the rolling window technique works best if performed using as large an interval as possible. This will reduce much of the single period variance in the coefficients—that is, it will smooth out the results much like using seasonally adjusted data. It also does not place rather ad hoc weights on prior observations like an exponential smoother, or force a particular model form such as with a Holt–Winters smoother. This is particularly important since we are unaware of the underlying distribution within each panel across any particular time period. It is the coefficient estimates in (10.6) that we test for dynamic heterogeneity. The way we test for dynamic heterogeneity using these coefficients is to plot them over time along with their estimated standard errors. Visually, if at any

point(s) the lower (upper) bound of the standard errors crosses the upper (lower) bound, then the changes in the coefficient estimate is a statistically significant one.

Having outlined the procedure, I would normally start running the rolling window regressions and get on with the inference from the results. But there is one issue that must be considered; do we use balanced or consistent panels, or both? After very little thought, the answer is obviously to use consistent data. The reason lies in the fact that it is a dynamic operation with the purpose of investigating parametric stability. Let me explain this further.

What exactly is panel data? It is essentially a number of cross sections stacked one on top of another. In our case, each cross section contains exactly one year's worth of data. A rolling window procedure would then start at the first year and roll through until the end. So, would it make sense to start this procedure at different years for different *i*'s? This would mean that the first regression window would contain a different number of *i*'s than other windows would. If we are interested in measuring dynamic stability in coefficients estimates, we would naturally want to have exactly the same *i*'s in every window as well as those windows starting and ending in exactly the same years. To this end, we will test the most basic form of our regression model using a version of the consistent data set we generated in Chapter 9. The only misspecification issues we will incorporate into our beginning model will be dependent variable dynamics because it could be argued that unless we control of dynamics in the dependent variable, coefficient estimates would naturally change over time. For simplicity's sake, we will ignore all other forms of misspecification.

We also have one other issue to address regarding our particular modeling case and data set; and that is the number of observations per country. If we use exactly the data set generated in Chapter 9, we will only have 10 observations per country. Reviewing the description for performing the rolling window regression, we know that our window must contain multiple observations; I generally prefer at least five time observations for my windows. But then this would only yield six coefficient and standard error estimates. This makes it difficult to discern whether heterogeneity exists or not. Because of this, I will reformat the data set constructed in

Chapter 9 by sacrificing countries and overall observations for a longer time dimension for each country.

The data set we will use for this analysis will cover 18 countries with 360 observations. Each country has 21 years' worth of data from 1988 to 2008 inclusive. Because of losing one observation due to the inclusion of lagged *Growth* on the right-hand side, there are a total of 20 years' worth of data for each country that can be used for the rolling window method. In our case, 5 years' worth of data across the 18 countries will make up the window. The window will effectively start in 1989. We will run a regression spanning all countries from 1989 to 1993 and record the estimated coefficients and their standard errors. We then move this window 1 year to cover 1990 to 1994 and do the same recording. This continues until the upper end of the window reaches the year 2008. We will then plot these estimates and evaluate the graph for heterogeneity in the estimates over time.

Figure 10.1 lists the output from our basic regression using this data set. Again, we are ignoring all other misspecification issues except dependent variable dynamics. Optimally, however, a researcher would perform this analysis after at least correcting for the basic misspecification issues outlined in this book. Again, we don't do this here simply because the data set used for this example does not resemble that used previously in the book. So, to avoid redundancy, I will pretend that all the procedures performed earlier have already been applied.

The initial estimates tabled in Figure 10.1 tell us that just like earlier in the book, there does exist a small amount of business cycle memory as the coefficient for lagged *Growth* is statistically significant. Furthermore, it appears as though increases in *Pop* and *G* lower *Growth*, while increases in *Trade* increase *Growth*. Let us now move on to the rolling window results.

We will start with the rolling window plot of the coefficient estimate for *I* in Figure 10.2. I've embedded a reference line at zero to make it easier to determine areas of statistical significance. The upper and lower 90 percent confidence interval bounds are drawn with dashed lines, while the coefficient estimate itself is marked with a solid line that lies in between the two dashed lines.

It quickly becomes apparent that there is some dynamic heterogeneity in this parameter. Remember, we can distinguish a statistically

```
. reg Growth l.Growth I FDI School Trade Pop G

      Source |       SS       df       MS              Number of obs =     360
-------------+------------------------------           F(  7,   352) =    5.24
       Model |  302.604797     7  43.2292567            Prob > F      =  0.0000
    Residual |  2906.70287   352  8.25767861           R-squared     =  0.0943
-------------+------------------------------           Adj R-squared =  0.0763
       Total |  3209.30767   359  8.93957567           Root MSE      =  2.8736

------------------------------------------------------------------------------
      Growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Growth |
         L1. |   .1665339   .0512781     3.25   0.001     .0656839    .267384
             |
           I |   .0037742    .022149     0.17   0.865    -.0397869   .0473353
         FDI |  -.0072075   .0390898    -0.18   0.854    -.0840864   .0696714
      School |  -.0039577   .0077335    -0.51   0.609    -.0191674   .0112519
       Trade |   .0171971   .0047741     3.60   0.000     .0078078   .0265864
         Pop |  -.7650574   .3214416    -2.38   0.018    -1.397245  -.1328698
           G |  -.1068449   .0359016    -2.98   0.003    -.1774535  -.0362364
       _cons |   3.674047   1.365248     2.69   0.007     .9889784   6.359116
------------------------------------------------------------------------------
```

Figure 10.1 Base regression for dynamic parametric heterogeneity

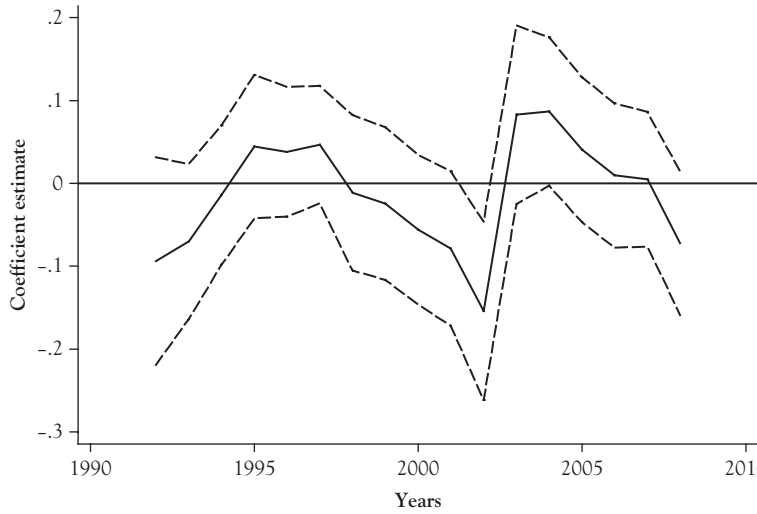


Figure 10.2 Rolling window coefficient estimate of I

significant amount of dynamic parametric heterogeneity by recognizing areas whereby the upper bound of the 90 percent confidence interval falls below the lower bound of the same interval, and of course vice versa. In Figure 10.2 we see this occur around the years 2002 and 2003. I mentioned that there is "some" evidence of heterogeneity as it is only the estimate in 2002 that causes this occurrence. Therefore, in this case, I would be inclined to ignore this instability—it simply doesn't last long enough to justify changing our model. Now we move on to the coefficient for *FDI*.

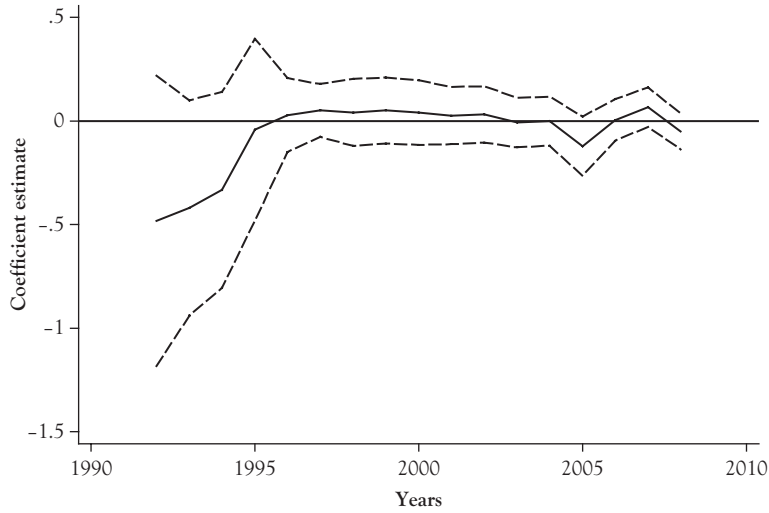


Figure 10.3 Rolling window coefficient estimate of FDI

When viewing Figure 10.3, the coefficient estimate for *FDI* seems to be very stable over time. With the exception of a large amount of variation around the beginning of globalization in the early 1990s, *FDI*'s effect on economic growth remains very consistent, although highly insignificant. Moving on to the coefficient for *School* in Figure 10.4, we find something quite interesting. Even though the static estimate from Figure 10.1 is about –0.004 and highly insignificant with a *p*-value of 0.609, there exists a statistically significant structural change in the coefficient's standard errors at about the year 2006. But, this is not enough to be captured by modeling this shift in the conditional mean. Interacting dummy variables that designate the periods before 2006 and from 2006 onward with the schooling variable still resulted in insignificance for the later period. We have to remember that the data has been smoothed out over a five-year period. This narrowing of the standard errors could be due to a one-period shock in the errors, which is not captured appropriately by the rolling window. And even though dummying out for the single year 2008 did produce statistical significance in the negative coefficient estimate, modeling a shock of just one period is inappropriate as you would be essentially modeling noise and not an empirical regularity. A researcher using later period data, however, would want to keep an eye on

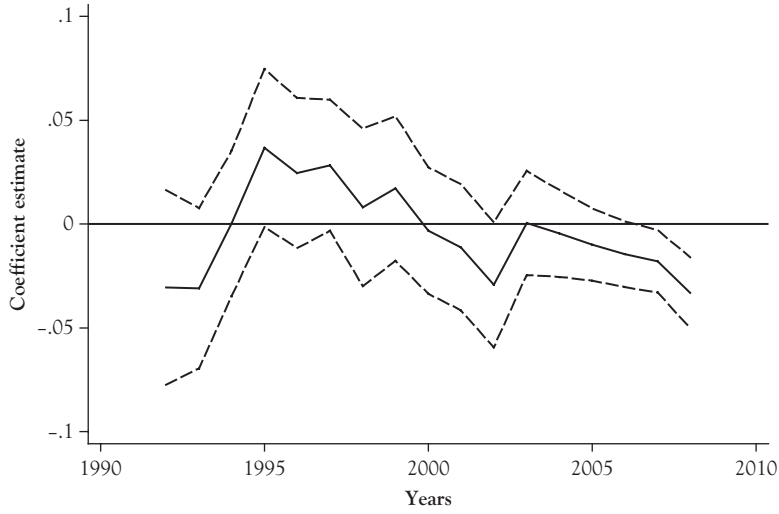


Figure 10.4 Rolling window coefficient estimate of School

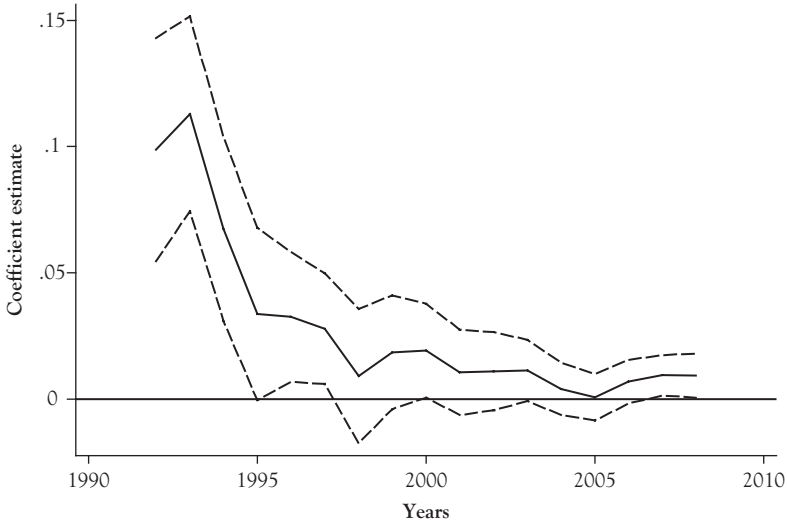


Figure 10.5 Rolling window coefficient estimate of Trade

this outcome. As more of these periods are added to the data, it could be the case that significance could be observed.

Evaluating the coefficient estimate for *Trade* in Figure 10.5, it's clear that the driving forces behind the significant coefficient in Figure 10.1 are the first 5 or 6 years of data. After that the coefficient estimate converges

toward zero, but for the most part remains significant. And even though it appears that many times after 1997, the lower bound of the 90 percent confidence interval is below zero implying insignificance, again, these are 5-year smoothed estimates; hence, we cannot be completely sure of significance until we actually model that area of the plot.

Modeling this area of heterogeneity by including a dummy variable into the regression that separates the years 1989 through 1994, from the remainder of the time period produces an estimate for the earlier period of 0.027 and a *p*-value of 0.000; and for the period 1995 to 2008 we get an estimate of 0.018 and a *p*-value of 0.000. Looking again at the plot, it seems as though the earlier period estimate should be even larger, but again, we can't see what the estimates where prior to the five-year average coefficient estimate beginning at 1992. They may be considerably lower pulling that periods' estimate downward. But are the coefficients we just estimated different from one another in a statistically significant sense? That is a completely separate question. Performing a test for the equality of the two estimates produces a *p*-value of 0.074 indicating that they are significantly different from one another and should be modeled as such.

Moving on to the coefficient for *Pop*, in Figure 10.6 we find relatively good stability from about 1995 onward. Furthermore, there is only

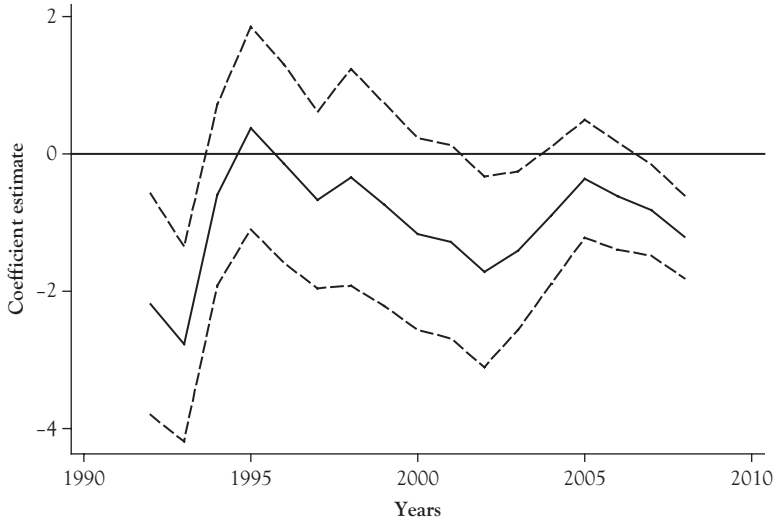
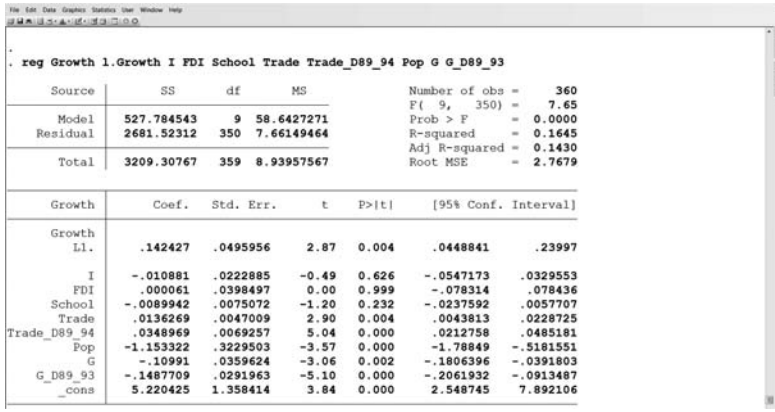


Figure 10.6 Rolling window coefficient estimate of Pop

one small area whereby the lower bound crosses the upper bound, and that is around the period 1993 through 1994. When delineating this in our regression we get two coefficients estimates that are over 0.400 apart; however, when testing whether this difference is significant, we get a *p*-value of 0.112 indicating that even though a large separation in magnitude exists, it is not a statistically significant one. Therefore, we must deem this coefficient to be a stable coefficient.

And finally we come to the rolling window estimation for the coefficient of *G*. The rolling window plot in Figure 10.7 shows substantial instability. Like the others, there is no real trending going on here, but simply structural shifts in estimates over time. It seems as though prior to 1998, the relationship between *G* and *Growth* is definitely negative. But then something occurs that reduces the coefficient estimate in absolute value, driving it toward zero. The bottom line is that one can obviously see that there is a difference in estimates over time. To once again highlight the deception that could occur by relying solely on the plot, however, placing a breakpoint anywhere other than at 1993/1994 results in nearly identical estimates of the two periods. Only when a dummy variable separating the period 1989 through 1993, from 1994 onward do we get a significantly different result. The coefficient estimate for the earlier

Figure 10.7 Rolling window coefficient estimate of G



```
. reg Growth l.Growth I FDI School Trade Trade_D89_94 Pop G G_D89_93

      Source |       SS       df       MS              Number of obs =     360
-------------+------------------------------           F(  9,   350) =    7.65
       Model |  527.784543     9  58.6427271            Prob > F      =  0.0000
    Residual |  2681.52312   350  7.66149464           R-squared     =  0.1645
-------------+------------------------------           Adj R-squared =  0.1430
       Total |  3209.30767   359  8.93957567           Root MSE      =  2.7679

------------------------------------------------------------------------------
      Growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Growth |
         L1. |    .142427   .0495956     2.87   0.004     .0448841    .23997
             |
           I |   -.010881   .0222885    -0.49   0.626    -.0547173    .0329553
         FDI |    .000061   .0398497     0.00   0.999     -.078314    .078436
      School |  -.0089942   .0075072    -1.20   0.232    -.0237592    .0057707
       Trade |    .0136269   .0047009     2.90   0.004     .0043813    .0228725
Trade_D89_94 |    .0348969   .0069257     5.04   0.000     .0212758    .0485181
         Pop |  -1.153322   .3229503    -3.57   0.000    -1.78849    -.5181551
           G |    -.10991   .0359624    -3.06   0.002    -.1806396   -.0391803
     G_D89_93 |  -.1487709   .0291963    -5.10   0.000    -.2061932   -.0913487
        _cons |   5.220425   1.358414     3.84   0.000     2.548745    7.892106
------------------------------------------------------------------------------
```

Figure 10.8 Regression correcting for dynamic parametric heterogeneity

period is –0.128, and for the latter period it is –0.088; a *p*-value testing the equality of the two coefficients is 0.053 indicating that the estimates for these periods are statistically different from one another.

It is at this point that we need to respecify our model and compare it to the results in Figure 10.1. In Figure 10.8, the *Trade* and *G* coefficients represent the effect these variables have on *Growth* for the periods 1995 onward and 1994 onward, respectively. The coefficient estimates in Figure 10.8 pertaining to *Trade_D89_94* and *G_D89_93* are the differences from the latter period estimates. The reader can see that they are both highly significant differences. But we also see that correcting for dynamic parametric heterogeneity more accurately models the data we have. The best indication of this is the fact that the adjusted R^2 for the model in Figure 10.1 was 0.076, but this same value for the model that generated the results in Figure 10.8 is nearly double that at 0.143! As the name of this book implies, this is a "better" model than it would have been had we ignored the misspecification of dynamic parametric heterogeneity.

Conclusion

It is quite apparent that if someone performed the typical sort of regression analysis whereby a researcher runs a simple linear regression, the inference they would draw from their models would be substantially different from the inference drawn from models that have been respecified as outlined in this book. This is even true when we only corrected for the basic misspecification issues outlined in items 1 through 5 in Chapter 1. This is perhaps why it has always bothered me why these topics are ignored by a large portion of the empirical research community; items 1 through 5 are so easy to test for and resolve that there is no reason to ignore them. Furthermore, in every case, the model that resulted after respecification fits the data better than before respecification. In other words, we ended up with a better model.

Again, I don't want to lose complete faith in the empirical research community, so I'll simply assume that there are good reasons why researchers do not broach these topics. But let's hope it's not because they are simply unaware of them (which means that their instructors weren't very thorough either), or are mining their results by relying on misspecified models simply because they give them the results they want.

Perhaps the most important objective that has been accomplished in this book, however, is the fact that all of our final models are "better" than our beginning models. They make sense within the context of the discipline, they fit the data better, and we didn't have to worry about issues specific to any discipline such as theoretical omitted variable bias. To this end, I hope that I have influenced the researchers who read this book to take the time and try to find a better model using this methodology; it has always worked for me and I'm sure it will work for you.

References

Anderson, T.W.; and C. Hsiao. "Formulation and Estimation of Dynamic Models Using Panel Data." *Journal of Econometrics* 18, no. 1 (January 1982), pp. 47–82.

Arellano, M.; and S. Bond. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies* 58, no. 2 (April 1991), pp. 277–297.

Arellano, M.; and O. Bover. "Another Look at the Instrumental Variable Estimation of Error Component Models." *Journal of Econometrics* 68, no. 1 (July 1995), pp. 29–52.

Baum, C.F. "Stata: The Language of Choice for Time Series Analysis?" *The Stata Journal* 1, no. 1 (2001), pp. 1–16.

Blundell, R.; and S. Bond. "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models." *Journal of Econometrics* 87, no. 1 (November 1998), pp. 115–143.

Breusch, T.S.; and A.R. Pagan. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* 47, no. 5 (September 1979), pp. 1287–1294.

Edwards, J.A. "*A Beginner's Guide to Economic Research and Presentation.*" New York, NY: Business Expert Press, 2013.

Edwards, J.A.; and K. Kasibhatla. "Dynamic Heterogeneity in Cross-Country Growth Relationships." *Economic Modelling* 26, no. 2 (March 2009), pp. 445–455.

Edwards, J.A.; and A. McGuirk. "Statistical Adequacy and the Reliability of Inference." *Econ Journal Watch* 1, no. 2 (August 2004), pp. 244–59.

Edwards, J.A.; A. Sams; and B. Yang."A Refinement in the Specification of Empirical Macroeconomic Models as an Extension to the EBA Procedure." *The BE Journal of Macroeconomics* 6, no. 2 (October 2006),.

McAleer, M. "Sherlock Holmes and the Search for Truth: A Diagnostic Tale." *Journal of Economic Surveys* 8, no. 4 (December 1994), pp. 317–370.

McAleer, M.;A. Pagan; and P.A. Volker. "What Will Take the Con Out of Econometrics?" *American Economic Review* 75, no. 3 (June 1985), pp. 293–307.

Park, R.E. "Estimation with Heteroskedastic Error Terms." *Econometrica* 34, no. 4 (1966).

Roodman, D. "*How to Do xtabond2: An Introduction to 'Difference' and 'System' GMM in Stata*." Center for Global Development Working Paper no. 103, December 2006.

Roodman, D. "A Note on the Theme of Too Many Instruments." *Oxford Bulletin of Economics and Statistics* 71, no.1 (January 2009), pp. 135–158.

Spanos, A. "*Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*." Cambridge: Cambridge University Press, 1999.

Spanos, A. "*Statistical Foundations of Econometric Modelling*." Cambridge: Cambridge University Press, 1986.

White, H. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica* 48, no. 4 (May 1980), pp. 817–838.

Index

OTHER TITLES FROM THE ECONOMICS COLLECTION

Philip Romero, The University of Oregon and Jeffrey Edwards,
North Carolina A&T State University, Editors

- *Managerial Economics: Concepts and Principles* by Donald Stengel
- *Your Macroeconomic Edge: Investing Strategies for the Post-Recession World* by Philip J. Romero
- *Working with Economic Indicators: Interpretation and Sources* by Donald Stengel
- *Innovative Pricing Strategies to Increase Profits* by Daniel Marburger
- *Regression for Economics* by Shahdad Naghshpour
- *Statistics for Economics* by Shahdad Naghshpour
- *How Strong Is Your Firm's Competitive Advantage?* by Daniel Marburger
- *A Primer on Microeconomics* by Thomas Beveridge
- *Game Theory: Anticipating Reactions for Winning Actions* by Mark L. Burkey
- *A Primer on Macroeconomics* by Thomas Beveridge
- *Economic Decision Making Using Cost Data: A Guide for Managers* by Daniel Marburger
- *The Fundamentals of Money and Financial Systems* by Shahdad Naghshpour
- *International Economics: Understanding the Forces of Globalization for Managers* by Paul Torelli
- *The Economics of Crime* by Zagros Madjd-Sadjadi
- *Money and Banking: An Intermediate Market-Based Approach* by William D. Gerdes
- *Monetary Policy within the IS-LM Framework* by Shahdad Naghshpour

Building Better Econometric Models Using Cross Section and Panel Data

Jeffrey A. Edwards

Many empirical researchers yearn for an econometric model that better explains their data. Yet these researchers rarely pursue this objective for fear of the statistical complexities involved in specifying that model. This book is intended to alleviate those anxieties by providing a practical methodology that anyone familiar with regression analysis can employ—a methodology that will yield a model that is both more informative and is a better representation of the data.

This book outlines simple, practical procedures that can be used to specify a model that better explains the data. Such procedures employ the use of purely statistical techniques performed upon a publicly available data set, which allows readers to follow along at every stage of the procedure. Using the econometric software Stata (though most other statistical software packages can be used as well), this book demonstrates how to test for model misspecification and how to respecify these models in a practical way that not only enhances the inference drawn from the results, but adds a level of robustness that can increase the researcher's confidence in the output generated. By following this procedure, researchers will be led to a better, more finely tuned empirical model that yields better results.

Dr. Jeffrey A. Edwards is a Professor of Economics at North Carolina Agricultural and Technical State University. He is the author of dozens of publications, an editor of the Economics Collection at Business Expert Press, an assistant editor for the Journal of Economics (MVEA), and sits on the advisory board for Applied Econometrics and International Development. He has a PhD in Economics with a field major in Econometrics.

ECONOMICS COLLECTION

Philip J. Romero and Jeffrey A. Edwards, *Editors*

businessexpert Press

www.businessexpertpress.com

ISBN: 978-1-60649-974-0

9 781606 499740 90000