

More about regression

Pär Nyman

April 9, 2026

This text is written as a natural continuation of Österman and Folke (2024). It follows a similarly non-mathematical approach, with the aim of helping you develop your intuition and your ability to reason as an applied researcher. At the same time, it assumes that you are familiar with the basics of regression analysis, so you are strongly encouraged to read that text before starting this one. Taken together, the two texts cover the content of the first three weeks, but see the syllabus for more detailed reading instructions.¹

The text is structured so that, with the exception of the first section, each section can be read independently. In other words, you can read the text in the order in which it is written, in the order in which the different topics are introduced in the lectures, or later use it to look up specific issues. However, it is probably a good idea to start with the first section, “Regression in applied research,” and then read it again once you have finished the rest of the text. In this section, we discuss modelling choices in more general terms, and many of the arguments made there are applicable to most of the situations described in the other sections.

This text was written with the intention of being improved between semesters. If you find any inconsistencies, if there are passages that you think are difficult to understand, or if there are arguments you disagree with, please write me an e-mail (par.nyman@statsvet.su.se) so that I can improve the text for next year’s students. Planned improvements currently include sections with Stata code, as well as more examples to make the text less dense.

1 Regression in applied research

In this introductory section, we provide background for the rest of the text. It contains some important points to which we will often return, but above all it presents our approach to applied regression analysis and explains why it may differ from what you would read in a statistics textbook.

If we know what the ‘true’ model, or data-generating process, looks like, it is often straightforward to specify a correct regression model. In applied research, however, we do not know what the true model looks like, and we usually face a large set of plausible model specifications. In this introductory section, we discuss how we navigate such situations and why model selection is not only about assumptions and statistical tests, but also about interpretability, presentation, and research ethics. We first go through the different assumptions on which our regression models usually rest and discuss their relative importance. As you will notice during the course, we take some of these

¹ If you have taken the C-level methods course at this department, you will probably notice that Österman and Folke (2024) cover much of the same material as that course.

assumptions very seriously, while others receive little attention. We then discuss the value of parsimony and why it may lead us to make assumptions even when we know they may be violated.

1.1 Assumptions

A statistics textbook usually explains which estimator is best, and what properties it has, given a set of assumptions. In applied research, however, it is rare to encounter situations in which all these assumptions are met. That does not mean they are unimportant, but for a practitioner it is essential to be able to assess the relative importance of different violations and to draw inferences from a range of models when there is uncertainty about which one is best.

In this course, we focus on one estimator: Ordinary Least Squares (OLS). We assume that you already have some prior knowledge of regression, and if you have ever conducted a linear regression analysis, you have probably used OLS. Depending on the courses you have taken before, however, you may or may not know that OLS is considered the optimal² estimator under a specific set of assumptions. There is no disagreement among statisticians about what these assumptions are, but they can be presented and organized in many different ways. How many assumptions one distinguishes depends on how they are grouped, but a linear model estimated with OLS is optimal when the following conditions are met:

1. The exogeneity assumption
 - (a) The dependent variable does not affect the independent variable.
 - (b) No relevant variables (i.e., confounders) have been omitted.
2. Linear relationship
3. No measurement error
4. Well-behaved residuals
 - (a) Error terms are independent.
 - (b) Error terms are normally distributed.
 - (c) Error terms have constant variance (homoskedasticity).
5. Technical issues
 - (a) The expected value of the residuals should be zero.
 - (b) There must be more observations than regressors.
 - (c) No perfect multicollinearity.
 - (d) No irrelevant regressors.

Of these assumptions, the first is arguably the most important. Strictly speaking, the OLS estimator requires the first three assumptions to hold in order to be unbiased. But whereas non-linearity and measurement error mainly tend to attenuate the estimated effect, violations of the exogeneity assumption can – and often do – bias the results in ways that render them useless. If you have ever heard the phrase “correlation

² More precisely, OLS is *unbiased*, *consistent*, and *more efficient* than other unbiased estimators when all relevant conditions are met. However, except for bias, we will not engage with these concepts in this text.

is not causation,” it was probably because someone was questioning the exogeneity assumption.

For the exogeneity assumption to hold, the dependent variable must not affect the independent variables, and if there are confounders (variables that affect both the independent variable of interest and the outcome), we must include them all, measure them without error, and correctly specify the functional form of their relationship with the other variables. Unfortunately, this assumption is rarely satisfied unless we are conducting an experiment, and it is therefore a major concern in almost all non-experimental analyses of causal questions. This assumption is sometimes referred to as the ‘zero conditional mean’ assumption.³

Over the last decades, social science has seen a strong shift toward design-based causal inference, by which we mean methods that help us satisfy the exogeneity assumption. In this course, we therefore devote several lectures to such techniques, and they are also covered in the coursebook (Angrist and Pischke 2014). The gold standard is the *controlled experiment*, in which the researcher delivers a treatment to a randomly assigned treatment group, but we will also discuss several kinds of *natural experiments*, which are situations not controlled by the researcher but that still share some of the attractive features of a controlled experiment. These techniques rely on their own specific sets of assumptions in order to satisfy the exogeneity assumption, and we will discuss them later in the course.

Yet most studies of causal questions do not have such a design. In some rare cases, we can still give our results a causal interpretation without too much concern. More often, however, we have to rely on control variables to reduce the omitted variable bias that results from violating the exogeneity assumption. Unfortunately, this practice is unlikely to eliminate the bias completely, and some would even argue that it may aggravate it. Making causal inferences from non-experimental designs will be discussed in Section 4 and in the lecture on control variables.

Because violations of the other assumptions tend to have less severe consequences, we cover them more briefly. The two assumptions that are most often raised in research seminars, aside from exogeneity, are linearity and the independence of residuals. Linearity is discussed in Section 6, and the independence of residuals in Section 2. We do not cover measurement error in this text, but it is discussed in Österman and Folke (2024, see the section on measurement error and the section on standard errors). As for the remaining assumptions, we have chosen not to discuss them in depth in this course. It is useful to have heard of them, but in our experience they rarely constitute a major problem in applied research.

1.2 The trade-off between parsimony and flexibility

The simpler our model is, the more assumptions we usually need to make. This may sound counterintuitive at first, but consider it this way: everything we do not model – every control variable we omit – is something we are implicitly assuming does not matter. So why do we make all these assumptions? Why should we ever assume a linear relationship when we could use a more flexible specification? And why do we not simply include all the variables we can find, along with different transformations and interactions, and let the data decide the rest? While it may be tempting to relax

³ It is given this name because it implies that the error term has an expected value of zero for all values of the independent variable. By ‘error term’, we do not mean the residuals (which often average to zero by definition), but rather the error term in the data-generating process. Because we think this may be confusing, we prefer to refer to it as the exogeneity assumption.

assumptions by adding new components to a model, increased flexibility always comes at a cost. So although much of the remainder of this text deals with situations in which we may want to extend our model, it is important to remember why parsimony also has value.

To begin with, statistical precision tends to decrease as we add more variables to a model. This is primarily because estimation uses only the variation that is unique to each variable (that is, variation that cannot be explained by the other variables). The more variables we include, and the more highly correlated they are, the worse our precision becomes. (Adding variables that are correlated with the outcome but not with the other independent variables can in fact increase precision, but that is not the main point here.) The phenomenon of strongly correlated predictors is usually referred to as *multicollinearity*. In forecasting and machine learning, too much flexibility relative to the available data is often referred to as overfitting. Regardless of the field, the underlying problem is that we are estimating parameters (i.e., slope coefficients) with too little effective information.

Moreover, it is often desirable to reduce the test of each theory to a single parameter per hypothesis, because this facilitates interpretation and presentation, makes it easier to compare results across specifications, and allows straightforward statistical tests. More generally, this also fits how most theories are formulated: they are meant to be simplifications of how the world works, not exact replicas of it. Naturally, this argument does not apply to models focused solely on prediction, because in that case our interest is not primarily in the regression coefficients.

Finally, many argue that researchers, deliberately or not, tend to choose models that produce certain results – *p-hacking* is the common term when such behaviour is used to obtain statistically significant findings. Complex models and non-standard choices increase the degrees of freedom available to the researcher, and thereby the opportunity to tweak the model until the desired results are obtained. For this reason, some readers are more persuaded by simple models, even when more complex choices can be justified.

Because of this trade-off, it is usually a good idea to present results from a set of different models, ranging from a very simple specification to more complex and data-demanding ones, which may sometimes be better visualized in a graph than in a regression table. By examining this range of models, we can provide a fuller picture of the evidence than if we only reported our main specification. In particular, if we can show that our key variable performs similarly in both simple and more complex models (i.e., that it is robust to different specifications), we are more likely to convince the reader – and more likely to convince ourselves – that the relationship is real.

From a presentation perspective, it can also be helpful to think of your results as consisting of two layers – one communicative and one exploratory (this terminology is our own, so there is no need to learn it). The task of the communicative layer is to convey your main message without distractions. This is the kind of content you would include in presentation slides or a short public summary, and all of it should also appear in the paper itself. In this layer, you should prioritize interpretation and theoretical relevance, and reduce the amount of information in order to highlight the main message. You may also use selected examples to illustrate your argument, such as a quote or a description of a specific event or development, even if these examples are more representative of your argument than of the full population under study. And do not be afraid to make assumptions, especially when they follow from your theory, because this is where you articulate your argument about how the world works.

The exploratory layer is where you make sure that the message in the communica-

tive layer fairly represents what the data actually look like. You are the main audience for this work, and most of it will never appear in the paper or even in the appendix. When you are not constrained by the space limitations of a paper or the limited attention span of your audience, you can relax assumptions, increase the flexibility of your models, and include more information in the graphs you produce. The results in this layer will rarely be as clear-cut as those in the communicative layer. That is to be expected. But if these results speak against the claims made in your communicative layer, it is important to revise the latter.

2 Statistical precision in regression analysis

Let us assume that we have conducted a survey experiment for the purpose of studying how economic incentives affect willingness to take a vaccine. One group of individuals was informed that the vaccine is free of charge, while the other group was told that they would be paid 100 SEK if they took the vaccine, and we then asked how willing they were to take it. The treatment assignment was randomized, so by comparing willingness to get vaccinated between these two groups, we obtain an unbiased estimate of how that willingness was affected by the monetary incentive. But even if this is a valid experiment, it is still possible that the people allocated to one group just happened, already before the treatment was delivered, to be more or less inclined to get vaccinated than the people allocated to the other group. How then do we know whether a difference between the two groups should be attributed to the treatment or to such sampling errors?

We never know for sure, because there is no upper bound on how large such sampling errors can be (although it is very unlikely, it is possible that all vaccine-takers were allocated to one group and all abstainers to the other). Fortunately, there are techniques we can use to deal with this uncertainty. For example, we can calculate how much sampling variability we should expect on average (which is approximately what a standard error captures), as well as how likely we would be to obtain a difference at least as large as the one observed, assuming that the treatment had no effect (which is what the p-value tells us).

Much of this is already covered in the text by Österman and Folke (2024). They provide a good introduction to the difference between random and systematic errors, what a standard error is, and how – as well as why – we can use it to calculate confidence intervals and p-values. Unless you already have a good understanding of these subjects, we recommend that you read Section 3.1 and Section 4 in Österman and Folke (2024) before taking on this text. You may also want to read the subsection on statistical certainty (in Section 5.1), although that material will also be covered here.

One point from Österman and Folke (2024) that we want to emphasize in particular is that everything in this section is about random errors (also called sampling errors), not about systematic errors (also called bias). When we talk about precision, we mean the expected distance between our estimated regression coefficient (β) and its expected value ($E[\beta]$).⁴ However, the expected value of our regression coefficient can be far from the causal effect we are interested in, for example because of severe omitted variable bias. In such cases, it is important not to mistake a precise estimate for a

⁴ The expected value is what the regression coefficient would be if we estimated it without random errors. It equals what the coefficient would be if we had an infinite number of observations, or what we would get if we could draw an infinite number of samples, estimate a regression coefficient in each sample, and then calculate the average coefficient across all samples.

good estimate. To maintain the distinction between random and systematic errors, I will add phrases such as ‘assuming our model is correct’ or ‘assuming no bias’ when making inferences about the causal effect, or alternatively restrict the inference to the coefficient’s expected value ($E[\beta]$), which equals the causal effect only in the absence of bias.

2.1 Making inference about the data-generating process

Statistical inference is usually taught in a setting where we draw a sample in which we can observe our quantity of interest, and then want to draw conclusions about the larger population from which the sample was drawn. For example, we may be interested in the Swedish unemployment rate or current voter support for the Social Democrats. We interview a small sample of the population about their labour-market status, or about what party they would vote for if there were an election today, and based on their answers we can estimate the share of unemployed people or Social Democratic voters in the larger population. Most readers of this text have probably encountered similar examples in previous courses, and it is primarily in this way that Österman and Folke (2024) discuss statistical inference.

However, when we conduct regression analysis, we typically use statistical inference to draw conclusions about the processes that generate the data we observe, rather than about some larger population. The reason is that if we are trying to answer a causal question – for example, how does labour-market status affect party attitudes? – we cannot directly observe the effect we are interested in, not even in the sample we study. We can observe whether a person is unemployed and ask her what party she would vote for, but we cannot observe the degree to which her party choice was affected by her labour-market status, because we cannot observe the counterfactual (how she would vote if she were not unemployed). This is sometimes referred to as the fundamental problem of causal inference. If we have a sufficient number of individuals, we can *estimate* the size of this effect (under a set of assumptions about what the relationship may look like), but this is still an estimate and not the true effect. Our standard errors then measure how precisely we have estimated the effect, and if we calculate a confidence interval, it gives us lower and upper bounds for the size of this effect (given a certain confidence level and assuming that our estimate is unbiased).⁵ In terms of formulas and calculations, what we do is the same as if we wanted to draw inference about a large population. The difference is simply that we need these techniques even when we are only interested in an effect within a sample.

2.2 Standard errors and statistical inference

When we conduct a regression analysis, we are presented with two kinds of standard errors. We get a root-mean-square error (Stata calls this Root MSE), which is a measure of how much the predicted values differ from the observed values, and we get a standard error for each coefficient. It is the latter that we refer to when we talk about standard errors in this text.

For each regression coefficient, its standard error tells us how precisely the coefficient is estimated. As an approximation, we can interpret the standard error as how much we should expect the estimated coefficient to deviate from its true value on average because of sampling variability. In other words, small standard errors

⁵ See Section 4.4 in Österman and Folke (2024) for a longer discussion of statistical inference about something other than a population.

indicate precise estimates. It cannot be emphasized enough, however, that standard errors are only a measure of random, not systematic, errors. Hence, they do not tell us anything about possible biases, and this ‘true value’ is only true given the model we have assumed, or given that our coefficient is estimated without bias.

More strictly, the standard error is an estimate of what the standard deviation would be if we could draw a large number of samples, run a regression on each one, collect the corresponding regression coefficient from each sample, and calculate the standard deviation of this distribution of coefficients. Statisticians would call this the estimated standard deviation of the coefficient’s sampling distribution. For reasons we will not go into here, we can assume that this sampling distribution is approximately normally distributed⁶, and in any normal distribution we know that 90/95/99 percent of the samples will produce a regression coefficient that lies no more than 1.65/1.96/2.58 standard deviations from the mean. Based on this property of the normal distribution, we can use the following techniques to draw inference about the effect we are interested in.

- We can calculate a confidence interval around our regression coefficient, stretching from 1.96 standard errors below the point estimate to 1.96 standard errors above it ($\beta \pm 1.96 \times SE_{\beta}$). Assuming our model is correct (i.e., the estimate is unbiased), such a confidence interval will include the true effect 95 percent of the time.
- We can calculate a p-value for the null hypothesis that there is no effect. The p-value then tells us the probability of estimating an effect at least this far from zero, assuming that both the model and the null hypothesis are correct. The p-value will be 0.1/0.05/0.01 if the regression coefficient is 1.65/1.96/2.58 standard errors from zero, but because it can take any value between 0 and 1, we let our statistical software calculate it for us.

It has become established practice to focus on whether the estimated effect is statistically significantly different from zero. We can do that either by checking whether the confidence interval (usually calculated at the 95 percent level) includes zero, or, equivalently, whether the p-value is smaller than the threshold value (usually 0.05). In regression tables and elsewhere, this is sometimes illustrated with asterisks (*), for example with one, two, or three asterisks denoting the 0.1, 0.05, and 0.01 levels. If none of these are available, for example because we are looking at a regression table without asterisks, it is common to check whether the regression coefficient is more than twice the size of the standard error (an approximation of 1.96) as a quick way of seeing whether it is statistically significant.

In many ways, it makes sense to be concerned with statistical significance. If precision is so poor, or the estimated effect so small, that we cannot rule out the possibility that the true effect is zero, then why should we care at all? However, this practice has caused many problems and has led large parts of the scientific community to question statistical testing more broadly. Some of these problems concern how individuals handle statistical inference. They may involve direct misunderstandings, but also an unbalanced emphasis on statistical significance relative to other considerations. Other problems are primarily rooted in the selection that occurs as researchers around the world conduct billions of statistical tests, only a small share of which can be published

⁶ Technically, we assume a t-distribution, but the critical values of the t-distribution converge toward the critical values of the normal distribution as the degrees of freedom increase.

and widely shared. If we direct our attention to the most interesting results, estimates with very large random errors are likely to be overrepresented in that selection. All of these problems are relevant in a course like this one, but since they are covered in Österman and Folke (2024, Section 4.4), I will not repeat them here. Even so, and even if this is largely a system-level problem, individual researchers still have a responsibility to avoid selecting research questions, samples, or model specifications based on whether the coefficient of interest happens to be statistically significant.

2.3 What determines precision?

The size of the standard errors depends on three factors: i) the degrees of freedom, which is the statistician’s way of saying the number of observations minus the number of estimated regression coefficients (including the constant); ii) the unexplained variation in the dependent variable, which is equivalent to the sum of squared residuals; and iii) the unexplained variation in the independent variable of interest, which equals the sum of squared residuals if this variable is regressed on the other independent variables. Formally, this can be written as follows:

$$SE_{\beta} = \sqrt{\frac{1}{n-k} \frac{\text{Unexplained variation in } y}{\text{Unexplained variation in } x}} = \sqrt{\frac{1}{n-k} \frac{\sum_1^n (y_i - \hat{y})^2}{\sum_1^n (x_i - \hat{x})^2}} \quad (1)$$

However, this is not a mathematics course, and you do not need to remember this equation or all the technical terms above. Nor will we derive the expression. Instead, I think it helps to see that most of us already have an intuitive sense that these three factors matter, or at least apply them in everyday reasoning, even if we rarely think about them in such formal terms. For example, if your friend told you that she gets a headache after drinking a full-bodied red wine, you would probably say that we cannot draw any conclusions from a single observation, and that if we wanted to get to the bottom of the hypothesis, we could ask her to keep daily notes on how she feels and what she eats and drinks. I think we can all agree that the more observations we have, the more confident we can be in our conclusions. You might also ask whether it is common for her to get headaches. Why? Because, intuitively, you know that the variation in y matters. If this happens regularly even when she does not drink red wine (that is, if there is a great deal of unexplained variation), then we need a larger number of observations before we can say whether there actually is a relationship with wine drinking. And finally, you might ask how often she drinks full-bodied red wine. If she does so only once per month, then she will need to keep a diary for a very long time before you have enough data to test the hypothesis, because there is so little variation in x . If she tends to combine full-bodied red wine with other alcohol as well as unhealthy food, and we want to distinguish the wine effect from the influence of alcohol and fat, then it is important that we also have data from days when she drinks only full-bodied red wine. That is why it is the *unique* variation we care about.

It follows from the equation, but also from the discussion above, that there are typically three ways in which we can improve the precision of our estimates:

- Increase the degrees of freedom by increasing the number of observations. Gathering more data is not always an option, but perhaps some observations were dropped because of missing values on a control variable. Dropping that variable, or imputing the missing values, could make a big difference. Technically, it also helps to reduce the number of regression coefficients, but in most situations

that margin is rather limited (note that we are only talking about the degrees of freedom here, not about the other ways in which removing variables may affect precision).

- Reduce the residual variation in the dependent variable. The most common way of doing this is probably to add independent variables that influence the outcome and are not affected by the independent variable of interest, but it can also be done by reducing measurement error in the dependent variable or removing variation that is due to chance or extraordinary circumstances.
- Increase the unique variation in the independent variable(s) of interest by removing correlated predictors. This can be done by removing control variables or using less flexible specifications for them (cf. Section 6), in which case it may come at the cost of a greater risk of omitted variable bias. But sometimes it is also possible to use a linear or unconditional main effect instead of something more flexible.

2.4 Clustered standard errors

Sometimes the data belong to different groups or clusters. These may be pupils clustered within classes, municipalities nested within larger regions, or so-called panel data, where an individual, a country, or some other unit has been measured at different points in time, just to name a few examples.

Observations tend to be more similar within clusters than between them. As a consequence, residuals tend to be positively correlated within clusters, which violates the assumption that error terms are independent (see Section 1.1). Unless accounted for, this within-cluster similarity can bias your standard errors so that the coefficients appear to be more precisely estimated than they really are. In other words, the estimated standard errors become too small, which means that the coefficient appears more precisely estimated than it actually is, and that you run the risk of falsely concluding that a coefficient is statistically significant.

Recall that the standard error for the slope coefficient is a measure of how much we would expect the coefficient to differ if we re-estimated the model on a new sample of observations. If we have a small sample, the standard error will be large, because we may have sampled observations that are not very representative of the true relationship between the variables. The more observations we have, the smaller such sampling errors are expected to be.

However, if the observations are not independent of one another, the number of observations exaggerates how much information there actually is in our sample. Imagine that we want to know how party choice differs between men and women. The more individuals we ask, the more reliable our estimate will be. But that is only true to the extent that more observations also mean more information. If we ask the same people multiple times, then we have multiplied the number of observations, but we have not added much new information. Or, for a more realistic example, if we ask groups of friends, they will not add as much information per individual as if they were independent of each other, because people tend to associate with others who are like themselves (or because they adopt the attitudes their friends hold). If we do not take these clusters into account when we – or Stata – calculate the standard errors, the standard errors will be biased downward (that is, the real uncertainty will be larger than the standard errors suggest). The size of this bias depends on how much variation there is within clusters compared to between them.

The standard way of handling this problem is to use clustered standard errors. What we do is define clusters, that is, the groups within which we suspect that observations are not independent of one another, and let our statistical software adjust the standard errors based on how much variation there is within clusters compared to between them. Let us return to the example. If friends are identical to each other, then all variation is found between clusters, and the effective number of observations is equal to the number of clusters. Conversely, if people are no more similar within clusters than between them, then the effective number of observations is equal to the number of individuals. Usually, the truth is somewhere in between, and this will be reflected in the standard errors provided by the software.

Let me round off the section on clustered standard errors with a few words of caution. The first warning is that if you have too few clusters, the standard errors will still be biased downward. A common rule of thumb is that you should have at least 40–50 clusters, but as usual, it is impossible to set an exact threshold for when the number of clusters is too small. First, the bias decreases continuously with the number of clusters, so the more clusters the better. Second, the bias is worse for unbalanced clusters (clusters of varying size), so while 30 clusters may be acceptable for balanced data, 70 could be too few in another setting. Third, what we should do in any given setting also depends on what alternatives we have. Using few clusters may be preferable to not clustering at all, but if you are choosing between two different variables on which to base the clustering, there are good arguments for picking the one with the most categories, if any of the variables has fewer than, say, 50 categories.

The American psychologist Abraham Maslow famously said that if all you have is a hammer, everything looks like a nail. In many fields, clustered standard errors are now used in much the same way as Maslow’s hammer. However, while they solve many common problems, they are not a universal solution. We will not cover any other techniques in this course, but as a general recommendation – especially for those of you who continue in academia – I think it is always a good idea to think carefully about how the data are correlated in your particular setting, and whether your chosen clusters are reasonable approximations of those patterns of correlation.

3 Control variables

In this section, we cover the basic principles of using control variables. We begin by discussing the different reasons why we may add independent variables to a model, because sometimes we include variables for purposes other than controlling for confounders. We then turn to suitable criteria for selecting control variables, assuming that controlling for confounders is our goal, as well as two special cases of control variables. We conclude the section by discussing how to apply these selection criteria in practice and how to evaluate the results.

The primary reason for using control variables is that the exogeneity assumption does not hold and that we believe we can reduce the resulting omitted variable bias by controlling for possible confounders. However, no matter how carefully we select control variables, in most situations the exogeneity assumption will still be violated, and sometimes adding control variables may even make things worse. In the next section, we therefore discuss the validity of the control-variable approach. When can it be convincing, and when is it doomed to fail? But let us start with the basics.

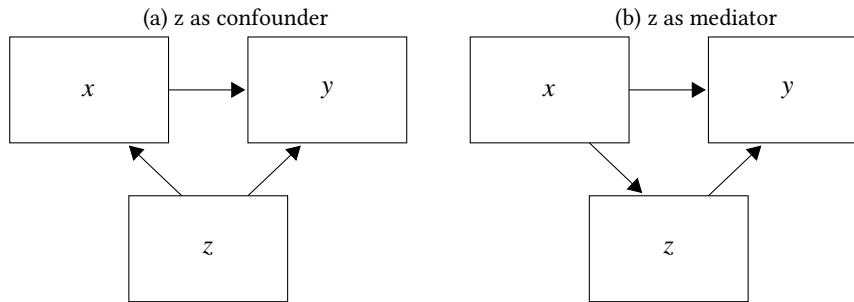


Figure 1: Confounder or mediator?

3.1 The purpose of including additional variables

There can be several reasons why we choose to add additional independent variables to a model. The most common motivation is that we are interested in a causal relationship between two variables (x and y) and want to isolate this effect from possible confounders. Remember, a confounder is a variable that affects both the independent variable of interest and the dependent variable, and excluding it from the model will result in omitted variable bias. When we include variables for this reason, we usually call them control variables.⁷ Most of this section will be about this control-variable approach to isolating causal relationships, but before we dig deeper, we will briefly consider three other reasons why one might include additional independent variables.

First, sometimes we add variables that are endogenous to – that is, potentially affected by – our independent variable of interest because we want to investigate *how* or *why* the dependent variable is affected by the independent variable of interest, and we suspect that z is an important channel, a *mediator*, through which x affects y . If the effect of x disappears when we introduce z , and we are confident that the causal direction runs from x to z rather than the other way around, then we conclude that the effect of x on y was mediated by z . Figure 1 illustrates the difference between a confounder and a mediator using a causal diagram. Such mediation analysis is well described in, and often attributed to, Baron and Kenny (1986). However, because causal relationships tend to be complex and multidirectional, it is usually difficult to identify mediators convincingly.

Second, sometimes we add independent variables because doing so improves the precision with which we estimate the effect of interest. For this to be the case, the added variables should be more strongly correlated with the outcome than with the independent variable, otherwise we are likely to worsen rather than improve precision. It is also important that the added variables are not endogenous to the independent variable of interest, or we may unintentionally control away part of the effect we are interested in. For these reasons, the practice of including variables to improve precision is particularly common in experiments, since we can then be certain that variables measured before the (random) assignment will be unrelated to the treatment variable.

⁷ Sometimes this term is used to denote all independent variables, regardless of their purpose, but in this text I use it specifically to refer to possible confounders. When talking about independent variables in general, I will refer to them as independent variables, covariates, right-hand-side variables, regressors, or predictors.

Third, although our interest often lies in studying some causal effect, this is not always the case, and other goals may require a larger set of variables. This is certainly true when making predictions, because it is rarely possible to generate decent forecasts using a single predictor. Moreover, sometimes our aim is to map out the different explanations for an outcome, compare the size of different effects, or test the joint explanatory power of a set of independent variables. Needless to say, such situations require at least a handful of variables.

3.2 Criteria for selecting control variables

It follows from the discussion in the previous subsection that the criteria for selecting independent variables depend entirely on their purpose. But assuming that we are interested in the effect of x on y and want a set of control variables in order to, hopefully, reduce suspected omitted variable bias, how do we know which variables to include?

The first criterion is that we should have some reason to suspect that the variable under consideration could act as a confounder, that is, affect both x and y . These suspicions can range from clear ideas about the exact causal relationship, to theories and conventions specific to a field of research, to variables that we include because they tend to influence many different things, or because they serve as proxies for a multitude of potential confounders.⁸

The second criterion is that the control variable should be exogenous to, that is, unaffected by, both x and y .⁹ If the variable is affected by x and influences y , then it is a channel through which the causal effect of interest is mediated. Controlling for this channel will remove part of the causal effect and bias the results toward zero. That is why Angrist and Pischke (2014) refer to such variables as ‘bad controls’. Bias can also arise when controlling for a variable that is affected by *both* x and y (or by causes of x and y). Such variables are often referred to as *colliders*. We avoid both kinds of bias by not controlling for, or otherwise conditioning on, variables that are themselves affected by x . Purposefully including endogenous variables can be interesting, and we briefly described such mediation analysis in the previous subsection, but we must not confuse it with the purpose we are concerned with here.

The third criterion is that the control variable cannot be too strongly correlated with the independent variable of interest. The reason is that only variation that is unique to each variable – that is, the difference between the observed value and what we would predict based on the value of the other variable – is used to estimate the regression coefficients. Therefore, the more strongly correlated the variables are, the less variation we have left at our disposal. The phenomenon of correlated regressors is called multicollinearity and is present in most quantitative studies. It would be meaningless to specify the level at which multicollinearity becomes problematic, because that depends entirely on how much variation there is in the data and on how important precise estimates are for your purposes. In an underpowered study, any level of multicollinearity is a problem, whereas in other studies it may be possible to ignore the issue altogether. If you still want to measure the level of multicollinearity in your model, this is usually done by having your statistical software calculate the

⁸ Alternatively, it could also be the case that we want to control away variation in the independent variable that we are not interested in. In such cases, omitting the control variable would bias our results even if it did not affect the dependent variable, because diluting our variation of interest with variation unrelated to the outcome would lead to attenuation bias.

⁹ Technically, if z is unrelated to y , then it does not matter if it is endogenous to x . However, if we knew that z and y were unrelated, then we would never include z as a control in the first place.

variance inflation factor for each independent variable.

A problem with multicollinearity as a criterion for inclusion is that variables that are highly correlated with the regressor of interest may also be the most relevant controls. Thus, even if leaving them out is sometimes our only option, it is an unsatisfactory solution because it means that we cannot rule out the possibility that it was the omitted variable, rather than the independent variable it was correlated with, that influenced the outcome. However, if the variables are conceptually similar, this need not threaten our interpretation of the results.

The fourth criterion concerns data quality. A control variable may be relevant in theory, but difficult to measure, or inadequately measured in the data available to you. And what good is a control variable if it does not measure what you are after? It may give the impression that you have controlled for something, while you in fact have controlled for something else. And what if your control variable has too many missing values? You could include it, but at the cost of throwing away many observations from the analysis; and you can impute the missing values before including the variable, in which case you would get to keep your observations, but at the cost of the control variable only partly controlling for what it should.

3.3 Fixed effects

Sometimes we do not think of controls as specific possible confounders, but rather as a way of controlling away variation that is more susceptible to omitted variable bias, while retaining the identifying variation that we find more credible. A particularly common and useful kind of control variable in this sense is what we call *fixed effects*, which can be used whenever our data consist of groups or clusters of observations. These may be pupils clustered within classes, municipalities nested within larger regions, or so-called panel data, where an individual, a country, or some other unit is observed at different points in time. We rarely refer to fixed effects as control variables, but they work in exactly the same way.

Let us say that we are analysing pupil performance, and that we have data collected from different school classes. In that case, there is variation *between* clusters as well as *within* them; for example, some classes perform better than others, but there are also differences among pupils within each class. If we believe that the variation between clusters is more susceptible to omitted variable bias, then we can restrict the model to exploiting only the within-cluster variation by using fixed effects, which is equivalent to controlling for the classes to which the pupils belong. This is no magic bullet, but it is often a reasonable thing to do. For example, if we have data across countries and over time, we may include country fixed effects to control for all time-invariant confounders, such as countries' histories and much of the institutional variation. In another setting, we may use family fixed effects to control for the variation shared by siblings (e.g., characteristics inherited from their parents), and rely only on the variation that differs between them. While it may sound tempting always to control for all these things, we must remember that the remaining variation may still be susceptible to omitted variable bias, sometimes even more so. And the more variation we control for, the less variation remains to estimate the effect, and the worse our precision becomes.

There are two ways of conceptualizing fixed effects. They yield the same results, so if you find one easier to understand than the other, you can simply stick to that version.¹⁰ One approach is to conduct a within transformation, which means that we

¹⁰ As described here, the two methods would provide slightly different standard errors and t-values unless

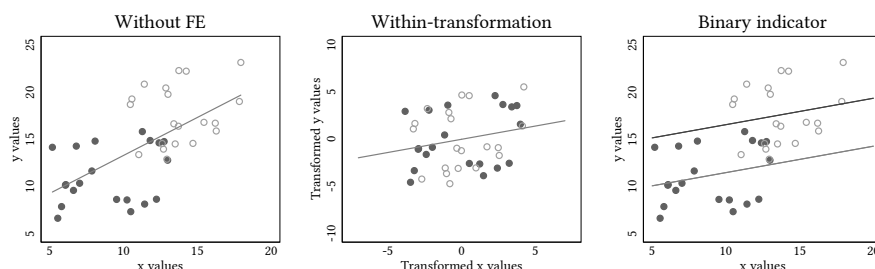


Figure 2: Illustrating fixed effects

first subtract the cluster mean from each variable value. The transformed variables then have a mean of zero in all clusters, which means that we have removed the variation *between* clusters (since the mean is now the same in all clusters), while retaining all variation *within* clusters (since all observations within a cluster have been transformed in the same way). We then use these transformed variables in our regression. The alternative is not to transform the data, but instead to include a binary indicator – so-called dummy variables – for each cluster (except one, which serves as the reference category) in the regression model. These indicators then control for cluster membership, which means that the remaining variation available to identify the effect consists only of within-cluster variation.

These two approaches can also be illustrated graphically. In Figure 2, the left-hand panel shows 40 observations divided into two equally sized clusters. There is clearly a positive relationship between the two variables in this panel, which is also illustrated by the steep slope of the regression line. In the middle panel, we have within-transformed both variables. As you can see, the pattern within each cluster is intact, but both clusters have been drawn toward the middle of the graph because both variables are now centred at zero. When we use only this within-cluster variation, the positive relationship is less pronounced than before, and the regression line is flatter. In the third panel, we instead add a binary indicator for one of the groups rather than transforming the data. This allows us to estimate separate intercepts for the two groups. As you can see, this method gives us exactly the same slope as the within transformation.

Whenever you use a fixed-effects model, remember that it is neither possible nor necessary to include variables that are constant within clusters. If you only had such variables as controls, or if the purpose of the model is prediction, then this is not a problem. All the variation that was previously captured by these variables is now captured by the fixed effects. However, it also means that you cannot use fixed effects if your interest lies in the effect of variables that differ only between clusters. For example, if you want to know how income inequality (dependent variable) differs across different types of welfare state (independent variable), it would not make sense to include country fixed effects.

In some settings, using fixed effects to study within-cluster variation has a more specific and interesting interpretation. For example, if our data consist of monozygotic twins, we can use twin-pair fixed effects to study only the variation between twins.

we adjust the degrees of freedom when using the within transformation. However, that is beyond the scope of this course, and it does not make the conceptualization incorrect.

Because monozygotic twins have the same genes, this allows us to study various relationships while holding all genetic confounders constant, something that is otherwise impossible in a non-experimental setting. This is sometimes referred to as a discordant MZ-twin design. Fixed effects are also used in the Difference-in-Differences design that we will cover later in the course. The attractive feature of that design is that, if the parallel trends assumption holds, fixed effects allow us to control for all time-invariant confounders and thus estimate the causal effect of an intervention. But all this is covered in Chapter 5 of Angrist and Pischke (2014).

3.4 Variable selection in practice

As you have probably suspected by now, few control variables meet all the criteria above. Therefore, adding more variables is usually a trade-off between controlling for additional confounders, on the one hand, and statistical precision, or the risk of including endogenous controls, on the other. Moreover, more controls do not necessarily mean less omitted variable bias, although it is admittedly easy to think of it that way.

The usual way of handling these trade-offs is to use a range of regression models, starting with a simple model with only a few controls, if any, and then sequentially adding sets of control variables, leaving the more problematic ones to the later specifications. Regression tables are often structured in this way, by adding more and more controls as we move from the first to the last column, but a similar progression can often be seen throughout a paper, with bivariate patterns reported early on and the most data-demanding specifications relegated to the appendix.

However, this does not mean that we should always move toward the most demanding specification when we add more controls or other forms of flexibility, because at some point precision becomes so poor that the analyses cease to be meaningful. Instead, we move back and forth around a main specification that the researcher believes strikes the best balance among the various trade-offs. This is sometimes referred to as the researcher's preferred specification. Even when this is not stated explicitly, it is often the case that one has to depart from a single model, rather than from a whole range of models, when reporting results, adding problematic control variables, or conducting other forms of robustness checks.

3.5 Evaluating the results

Even when we have a preferred model, it is important to evaluate a wide range of specifications, because we have no way of knowing what the best model looks like. This is especially true when we lack a strong design and rely on control variables to isolate the effect. What we want to know is whether the estimated effect is stable across a wide range of specifications. The more robust it is, the stronger the case for a causal interpretation. But sometimes the coefficient(s) of interest will change substantially as we add new variables. This may indicate that the original relationship was confounded by one or several of these controls – if the coefficient moves toward zero; otherwise, we may call the control variables suppressors – but it need not be so, and it is important to understand why the coefficient changed.

One thing to consider is whether the coefficient changed because the sample changed. Remember, if you add a control variable with missing data, the corresponding observations will be dropped. Whether the change in the coefficient was due to the change in the estimation sample or to the correlations among the variables is easy

to check by restricting the first model to the observations included in the second regression. If the loss of observations caused the change, then this is not really a reason for concern, even though it may mean that we cannot keep the new variable in subsequent models.¹¹

Other potential issues are more difficult to interpret. How did the new control variables affect the standard error? If precision became much worse, this may be a reason to stop adding variables and to place less weight on the poorly estimated specification than you otherwise would. And are you sure that the control variables are exogenous to your independent variable? If not, it could also be the case that you have controlled away part of the true causal effect – the part that operates through these mediators. In such situations, it is rarely possible to say which estimate we should believe, which is why it is so important to evaluate the full range of specifications and to be transparent about both the robustness of the results and the lack thereof.

Quite often, we will conclude that the relationship we were interested in appears to be spurious and confounded by some other variable. Sometimes this is a major disappointment, and we begin planning the next project. At other times, however, the finding may still be worth investigating. Showing that something does not have a causal effect can be just as relevant, especially if others claim that it does.

4 Causal arguments with observational data

A common typology in quantitative research distinguishes between studies focused on *description*, *prediction*, and *causal relationships*. The boundaries between these categories are not always sharp: the difference between a causal and a descriptive question is sometimes largely semantic, and professional forecasters know that understanding the causal relationships underlying their prediction models helps them make better predictions. Still, this trichotomy is often useful because it speaks to the assumptions we need to make.

Most quantitative studies analyse causal questions. That is, they are not only interested in a descriptive or predictive difference (e.g., whether turnout is higher among people with a high level of education), but also want to say something about the counterfactual outcome (e.g., what a person's turnout would be if we changed his or her level of education), or – synonymously – the causal effect (e.g., how education affects turnout). As discussed in Section 1.1, estimates of causal effects will be biased if the exogeneity assumption is violated. To satisfy this assumption, the social sciences have seen a major shift toward design-based causal inference. By this, we mean different strategies for obtaining exogenous variation in the independent variable of interest. Weeks 6–8 are devoted to such designs.

In this section, we discuss situations in which we study a causal question without an experimental design. So while it is true that correlation does not automatically imply causation, there are also occasions when it does. Either because it is reasonable to assume that the exogeneity assumption holds, or because we are studying a relationship in which the effect of interest is so large that the omitted variable bias is small by comparison. These two arguments often overlap, but we will discuss them separately in Sections 4.1 and 4.2.

¹¹ Technically, the change could also be coincidental: if the identifying variation changes dramatically, then we should expect the estimated coefficient to change, for the same reason that we would expect a difference if we re-estimated the analysis on a partly different sample. However, this point is rarely acknowledged.

However, in most non-experimental settings, omitted variable bias remains a serious threat. The traditional way of dealing with it is to add control variables, but this practice rarely allows us to dismiss the possibility that our estimates are so biased that the correlation we have found could be driven by something else, and therefore does not reflect a causal relationship. These challenges are discussed in Section 4.3, and in Section 4.4 we discuss how to adapt our terminology to express uncertainty about causality.

4.1 Situations where the independent variable is exogenous

Sometimes, we are interested in a relationship where we believe the exogeneity assumption to hold because the independent variable of interest is exogenous (i.e., unlikely to be affected either by the dependent variable or by another variable that independently affects the dependent variable).

This tends to be the case for some background characteristics of individuals (biological sex, ethnicity, birth order among siblings, etc.) as well as different kinds of natural phenomena, such as weather, temperature, or natural disasters. Note, however, that exogeneity is often conditional on certain sample restrictions. For example, if you want to estimate the causal effect of being the first-born child, you must condition on the number of siblings, or birth order will be endogenous to all family characteristics that affect the number of children.

Quite often, these are cases where the difference between a causal and a descriptive interpretation is largely semantic. For example, the causal question ‘does immigrant background reduce the probability of becoming a politician?’ is not necessarily harder to answer than the descriptive question ‘are immigrants underrepresented in politics?’, nor is it obviously more interesting. This is probably because causal interpretations are often more interesting for variables or events that can be changed (e.g., getting an education or implementing a reform), which exogenous variables cannot be.

We should also be aware that many variables we casually assume to be exogenous may still involve some endogeneity. For example, in private life it makes sense to treat the sex of one’s children as essentially random. However, we know that male fetuses are more vulnerable than female fetuses, which in turn means that the probability of having a boy is negatively correlated with all factors that are positively correlated with miscarriage. In Sweden, these effects are probably small, and this is therefore an example of overlap between the ‘almost exogenous’ and ‘the bias is probably small’ kinds of arguments.

4.2 Situations where the bias is small

Sometimes a causal interpretation is reasonable even if the exogeneity assumption is violated, because the possible bias is small relative to the size of the effect we are studying. In such situations, our estimates are still biased, but we are less concerned about that bias because it does not affect our main story.

This is the case when the effect we study is so large and sudden that it stands out from other sources of variation in the dependent variable. We encounter such situations all the time in daily life. Imagine that you have been out celebrating after completing a difficult exam in a methods course. If you wake up the next day with a headache, you would probably interpret this chain of events causally. Because you do not observe the counterfactual (how you would have felt if you had not celebrated), you cannot be entirely sure, but hopefully waking up with a headache is uncommon

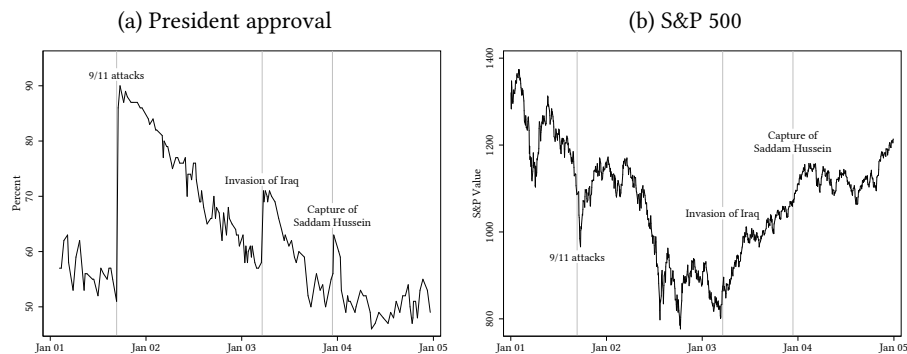


Figure 3: Is the effect large (left) or small (right) relative to the possible bias?

enough that it would be unreasonable not to assume that it was caused by the binge drinking.

Such situations are rarer in applied research, but they do occur. Suppose that we are studying so-called “rally around the flag” behaviour, meaning that support for the government or other leaders increases during times of crisis or external conflict. The left panel of Figure 3 shows what such data may look like, illustrated here with the approval rating of President George W. Bush during 2001–2005, with vertical lines denoting three major external conflicts: the 9/11 attacks, the American invasion of Iraq, and the capture of Saddam Hussein. The takeaway is that it seems obvious that the large spikes were indeed caused by these events – there does not appear to be anything else with a comparable effect on the approval rate, and we do not think it was the approval rate that caused these events. If we were to regress the week-to-week change in approval on an indicator for major events involving external enemies, we would likely obtain a biased estimate of this effect, because we should expect omitted variables to be related to both conflict and approval. However, that bias would be much smaller than the estimated effect.

For comparison, the right-hand panel shows how the S&P 500 (an American stock market index) developed over the same period. It is an interesting question how the stock market reacts to events like these. But, as the graph suggests, most of the variation in the dependent variable has nothing to do with them. This is a much more common situation. We can still do our best to estimate the effect, using control variables to isolate it from possible confounders, but we should not expect the analysis to be highly convincing.

Quite often, political reforms give rise to sudden changes in our variables. However, even if the reform itself is sudden, people may change their behaviour well before it is implemented. This may happen because the reform signals something, because people take the opportunity to act before implementation, or because some underlying factor caused both the reform and the behavioural change.

Imagine that we want to study how closing restaurants affects the spread of Covid-19. Before looking at the data, most people would probably agree that a sensible strategy would be to compare the reproduction number of the virus in the days before restaurants were closed with the days after. This is a short time window, and hopefully not too much else is happening at the same time. But take a look at Figure 4. It shows the number of UK restaurant bookings on the OpenTable platform, and how they

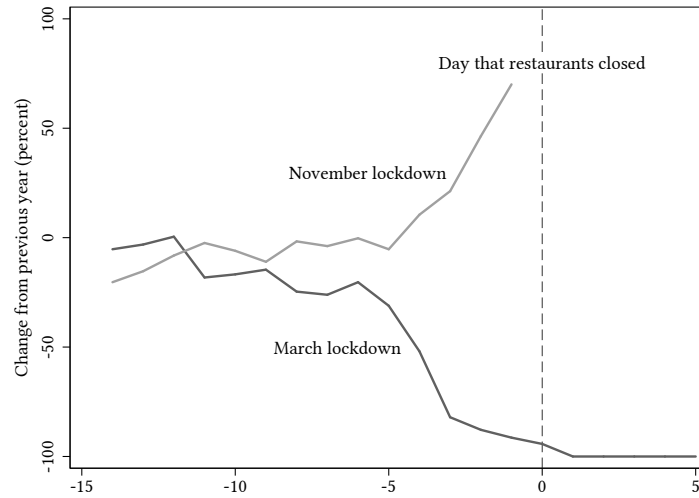


Figure 4: OpenTable bookings in the days before the lockdowns

The graph shows the number of restaurant bookings on OpenTable and illustrates two kinds of behavioural responses in anticipation of a reform.

developed before the March and November lockdowns.

In March, the number of restaurant guests fell before the lockdown – perhaps because the lockdown signalled how serious the situation was, or perhaps because people changed their behaviour in response to the increasing number of Covid cases. In November, the pattern was the opposite. People increased the number of restaurant visits in anticipation of the reform because they knew they would not be able to dine out for some time, and perhaps the willingness to distance voluntarily had declined since the spring. In both cases, a very simple before–after comparison would be biased, because behaviour changed a few days before the reform was implemented. Could we compare with an earlier period instead? Yes, probably. But the more time that passes, the greater the risk that other important factors will also affect how quickly the virus spreads.

4.3 Control variables

Unfortunately, the most common situation in applied research is that the exogeneity assumption is violated and that we suspect the resulting bias could be substantial. The standard procedure for dealing with such situations is to add control variables in order to isolate the relationship of interest from possible confounders (i.e., variables that affect both the independent and the dependent variable). In theory, we would eliminate omitted variable bias if we added all relevant controls, measured them without error, and correctly specified the functional form of their relationships with the dependent variable. In practice, this is not possible, because the systems we study are so complex that in most applications there is no way for us to know and measure all the relevant variables. Still, through careful selection of control variables, and by evaluating a range of different model specifications, we can at least try to reduce the size of this bias.

However, scholars have become increasingly sceptical of the practice of using

control variables to isolate causal effects. This scepticism is based on two insights. First, adding controls can either increase or decrease bias, and there is no way of knowing in advance which effect it will have – some omitted variables will bias the estimate downward, others upward, and it is only if we control for the dominant ones that the bias will be reduced. Second, if people want to find an effect, they are likely to experiment with different specifications until they do (so-called p-hacking). That does not necessarily imply dishonesty; it is deeply human to subconsciously search for arguments that rationalize a model specification consistent with the expected or desired result. Taken together, these two insights have led many scholars to question the validity of research that relies heavily on control variables.

Yet most social science research on causal questions does rely on control variables. And while there are situations in which it is virtually impossible to identify the causal effect of interest using controls, there are also many cases in which such a strategy can be fairly convincing. Rather than dismissing it altogether, it is more constructive to discuss how we can distinguish between relationships in which the proposed causal story is the most likely one and those in which it is not. At the end of Section 3, we discussed how to judge the robustness of results by evaluating a range of model specifications. Here, by contrast, we focus on the conditions that affect *a priori* validity (that is, what we can say before looking at the results). It should be emphasized that while these are valid arguments, they are typically not strong enough to convince a sceptic. Moreover, arguments of this kind are rarely made explicit in published research. I am not entirely sure why, but it may be better to view them as a way of training your own ability to judge regression results than as a guide to how you should sell your findings in a paper. In any case, there are at least three factors that affect the *a priori* possibility of interpreting regression coefficients as causal estimates when you have to rely on control variables.

The first thing to ask is how theoretically plausible the causal story is. The more surprising your results are, and the smaller we should expect the effect to be, the harder it will be to convince others. Does democracy lead to higher life satisfaction? This is probably something you could investigate, with some caveats, using a traditional control-variable approach. But if you want to study how life satisfaction is affected by the quality of public transportation, or some other marginal phenomenon, then be prepared for many people to dismiss your results as the product of omitted variable bias unless you have a very strong design. It is probably not a coincidence that research on peculiarities and irrational behaviour (e.g., the subfield of behavioural economics) is so dependent on experimental designs.

Second, some variables are more susceptible to bias than others. This can be because they are so strongly affected by difficult-to-measure factors that most of the identifying variation would come from those confounders even if we tried to control for them. For example, if you want to study whether drinking expensive wine increases life expectancy, it would be so difficult to control for all the class-related factors that affect both health and wine preferences that the exercise would be close to pointless; the effect of interest would be small relative to the remaining bias, and most of the variation in your independent variable is probably determined by class. You would be better off studying the effects of something that is less tightly linked to background characteristics, such as being injured in a workplace accident, having children at a school that politicians want to close, or living in a municipality that was hit hard by a recent storm. None of these variables is exogenous, but with a good choice of control variables we could probably end up with identifying variation (the variation left after the control variables have accounted for the rest) that makes a causal interpretation

more convincing than the alternatives, even if it is still far from conclusive.

A related issue is how important it is to distinguish the effect of the independent variable from that of other variables that are conceptually and empirically correlated with it. For example, even if it is difficult to isolate the effect of income from that of education, cultural capital, or occupation, this would be less of a problem if the findings are also interesting when interpreted more broadly as an effect of class or socioeconomic status.

Third, sometimes you can rely on other kinds of empirical evidence to support your interpretation. Perhaps interviews or an experiment could shed light on the assumed causal direction or the plausibility of your interpretation. This is a strong argument in favour of mixed methods, although one must be very careful when generalizing findings across contexts.

4.4 Expressing the uncertainty

When we run regressions, our statistical software provides us with coefficients, and it is our job to interpret them. By choosing our words carefully, we also signal how confident we are in a causal interpretation. If we use words like ‘effect’ or even ‘causal effect’ to describe our estimates, many readers will interpret that as a strong causal claim. The logic of this convention can be questioned (why would it not still be an effect estimate if it is biased?), but it is what it is, and I therefore recommend more neutral language, such as referring to the ‘regression coefficient’ or something similar. In the same spirit, if you want to be explicit about what you have actually learned, it is often helpful to write something like ‘we find an association between x and y , such that units with higher x tend to have higher y .’ If you think a causal interpretation is plausible but not secure, you can signal this by saying that ‘the pattern is consistent with the hypothesis that ...’ or that ‘I find suggestive evidence that ...’, while also making clear that alternative explanations remain possible (e.g., ‘although omitted variables or reverse causality cannot be ruled out’ or ‘although the broad confidence intervals show that the data are consistent with an effect no larger than ...’).

More important, of course, is that we are transparent about the assumptions that must hold for our coefficient to be an unbiased estimate of the causal effect, and that we discuss how concerned we should be about bias. This is often easier with design-based causal inference, because these techniques come with carefully specified assumptions that are sometimes even testable (e.g., ‘under the assumption of parallel trends, the estimated effect is ...’). But if you are in a situation where you have probably violated the exogeneity assumption, yet sincerely believe that a causal interpretation is still the most likely explanation for the relationship you have found, then explain why you think that is the case, while also discussing other possible explanations.

It has become relatively common – especially among people who only conduct experimental studies – to use the term ‘description’ for all analyses that do not allow design-based causal inference, and to claim that such approaches can never be used to assess a causal theory or hypothesis. I think this is misguided. Partly for conceptual reasons – it is better to keep estimation bias and the nature of the question as two separate issues – and partly because there are situations in which traditional regression on observational data is not all that susceptible to omitted variable bias. Moreover, in many social science settings, it is simply not feasible – for practical or ethical reasons – to study the question of interest experimentally. That does not mean we should stop studying those causal questions. At the same time, it can sometimes make sense to frame a study as addressing a descriptive question if the results would still be

interesting even if the relationship were not causal. For example, even if we cannot show that socioeconomic differences in political participation reflect a causal effect of income or education, they are still interesting because they indicate that some groups are more influential than others. However, if you have studied an inherently causal question, and only the causal interpretation is interesting, then you cannot solve the problem simply by calling it description.

5 Conditional effects and interaction variables

Sometimes we suspect that the effect of one independent variable depends on the value of another independent variable (researchers often say that the effect is conditioned or moderated by the third variable). Examples of such hypotheses might be that ‘immigration primarily affects support for radical right parties in areas with high unemployment’ or that ‘tax cuts are more common than tax hikes during election years, but not in autocracies’. Or, more generally, an increase in x is associated with an increase in y when condition z is met, but not when condition z is absent.¹²

The standard way to model such relationships is to include the product of the two variables as an additional variable. If our first two variables are x and z , then this interaction is written xz or $x \times z$. We will use the same notation throughout this section – and we will refer to the regression coefficients for the constitutive terms as b_1 and b_2 , and the interaction coefficient as b_3 – but note that these are simply arbitrary labels and may refer to something else in another equation.

$$y = a + b_1x + b_2z + b_3xz + e \quad (2)$$

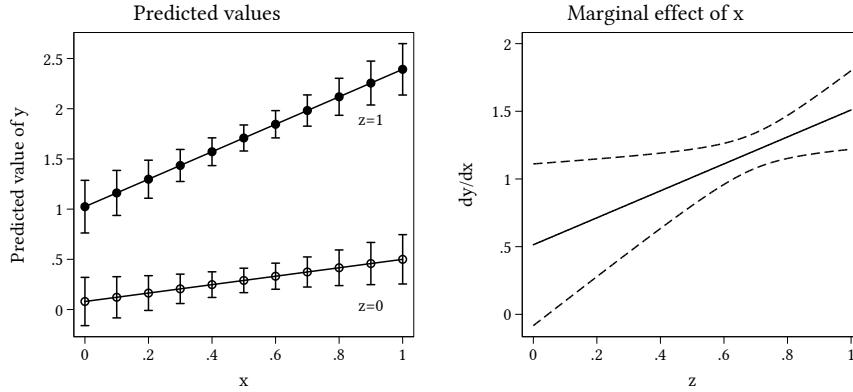
Interpreting models with interactions can be tricky, because the effect of your independent variable x is now captured by two different parameters: b_1 and b_3 . We could calculate the first derivative with respect to x , and conclude that when x increases by 1, y is predicted to increase by $b_1 + b_3 \times z$. Have you forgotten how to take derivatives? Do not panic. You can easily reach the same result using intuition alone. If x increases by 1, then b_1x increases by b_1 , b_3xz increases by b_3z , and the other terms (a , b_2z , and e) remain unaffected.

$$\frac{\partial y}{\partial x} = b_1 + b_3z \quad (3)$$

It is common to let z be a binary variable that can only take the value 0 or 1, partly because this makes interpretation much easier. In these cases, it is easy to see that the effect of an increase in x is either b_1 (when $z = 0$) or $b_1 + b_3$ (when $z = 1$), while b_3 tells us how much the effect when $z = 0$ differs from the effect when $z = 1$. When we use interactions in our models, it is often this difference that is of theoretical relevance, because it tells us to what extent the effect is conditioned on z .

When z is not binary, b_1 is still the effect of an increase in x when $z = 0$. However, $z = 0$ may not be a meaningful condition, and it may not even occur in the data. One solution is therefore to centre the conditioning variable around a substantively relevant value, most commonly its mean (that is, to subtract the mean so that the new mean becomes zero). In that case, b_1 can be interpreted as the effect of x when z is at its mean. Another approach is to rescale z so that it ranges from 0 (a theoretical or empirical

¹² In this text, we only discuss two-way interactions (that is, two variables interacted with each other). It is also possible to interact larger numbers of variables, or to include quadratic terms or other transformations in the interaction, but such models are more difficult to interpret.



The left-hand panel shows that the predicted value of y increases with x (both lines have a positive slope), as well as with z (the line where $z = 1$ has a larger intercept than the line where $z = 0$), and that the slope is steeper when $z = 1$. The right-hand panel shows that the marginal effect of x increases with z . The error bars in the left-hand panel and the dashed lines in the right-hand panel represent 95 percent confidence intervals.

Figure 5: Two ways of visualizing an interaction effect

minimum) to 1 (the corresponding maximum). These transformations do not change the fundamental properties of the model; they only change which value of z is treated as the baseline and therefore how the coefficients are most naturally interpreted.

When z is continuous, b_3 tells us how much the effect of x changes when z increases by one unit. In this case, the main object of interest is usually the *marginal effect* of x at substantively relevant values of z , rather than the effect at $z = 1$ in particular. For that reason, it is often more informative to present the implied effect of x across the observed range of z (for example, in a marginal-effects plot) than to focus on any single sum of coefficients.

5.1 Visual presentation

Interaction effects are often presented graphically. When the conditioning variable is a binary indicator, we can think of the model as two regression lines with different slopes (and since we include z as a standalone variable, they also have different intercepts). However, because we would get a new line for each value of z , this is no longer possible if z is a continuous variable. In such cases, it is better to show the effect of a change in x , which is usually referred to as the marginal effect of x and in this case equals $b_1 + b_3z$, over different values of z .

Figure 5 shows both of these kinds of graphs. The underlying data are simulated, which means that they were generated in statistical software. Simulations are a bit like playing God: we decide what the true model, or data-generating process, looks like, and then we let this model generate the data. In this simulation, we assumed that $y = 0.5x + 1xz + 1z + e$. Here, we first let z be a binary variable that takes either the value 0 or 1. This means that we can present the interaction as a graph of predicted values, which we do in the left-hand panel. According to the true model (which we only know because this is a simulation), we should expect the lower line to have a slope coefficient of 0.5 ($z = 0$), and the upper line to have a slope coefficient of 1.5 ($z = 1$). We should

not expect the slopes in the graph to be exactly equal to these values, because the simulation included an error term, but they are fairly close. In the right-hand panel, we have changed z into a continuous variable ranging from 0 to 1, which means that we must visualize the interaction using a marginal-effects graph. Instead of predicted values, the y-axis now shows the predicted change in y from a one-unit increase in x . As expected, it increases from 0.5 (the slope of the lower line in the left-hand panel) to 1.5 (the slope of the upper line in the left-hand panel).

5.2 Common mistakes and good practices

Even in published research, interaction models are often misinterpreted. Since the logic of interaction effects is arguably not very complicated, this is probably best explained by their theoretical relevance, which makes them appealing even to people who are not especially comfortable with regression analysis in general. There are a few frequent pitfalls that may lead to confusion. We therefore flag some of them, along with a few good practices, here.

Except in very rare cases, researchers should include the constitutive terms when using interaction models, unless these terms are perfectly collinear with other variables. In other words, if we have the interaction xz in our model, then we should also include the constitutive terms x and z . Excluding z from the model is equivalent to making the assumption that $b_2 = 0$, or that z has no effect on y when $x = 0$. If this assumption is false, the other coefficients – and most notably the interaction term – would suffer from omitted variable bias. And even if we suspect that it might be true, there is usually little to gain from excluding the constitutive term.

Another common mistake is to interpret the effect of the constitutive terms as unconditional average effects (that is, the predicted change in y from a one-unit increase in x), just as we would do if there were no interaction term, when they are in fact the marginal effects *when the conditioning variable is equal to zero*. Unless the conditioning variable has a mean of zero, you should therefore avoid talking about the effect of the constitutive term as ‘unconditional’, ‘main’, or ‘independent’.

A third mistake is to draw conclusions about substantively irrelevant situations because they are conditioned on z taking an impossible value. For example, if z cannot take the value zero, it does not make much sense to give a substantive interpretation of b_1 . Similarly, though less categorically, we must also be careful when making inferences about situations for which we have little data. Partly this is because they may be estimated with greater uncertainty, and partly because they tend to be less relevant. It is therefore common to place the marginal-effects graph above a histogram showing the relative frequency of the conditioning variable, so that the reader can see how common different values are.

A fourth issue to keep in mind is that the control variables you include in a model to isolate the effect of x may not be sufficient to isolate the effect of xz (control variables are rarely sufficient, but that is not the main point here). Instead of asking ourselves which variables may be related to both x and y , we must now ask which variables other than z may also condition the effect of x (while also being related to z). Sometimes we can come up with such examples, in which case we should include both those variables and their interaction with x in our model. One usual suspect is the x variable itself – if the effect is non-linear, and x is correlated with z , this non-linearity may bias the estimated effect of the interaction between x and z . But there are also standard procedures that we can use, either instead of or in combination with the variables we believe may be important. One such procedure is to interact z with all independent

variables in the model. If the conditioning variable is binary, this is equivalent to splitting the sample in two based on the value of this variable (in both cases we allow the effect of each predictor to differ between $z = 0$ and $z = 1$).

5.3 Further readings

If you would like a longer introduction to these issues, without necessarily going into greater depth, I recommend Brambor et al. (2006).

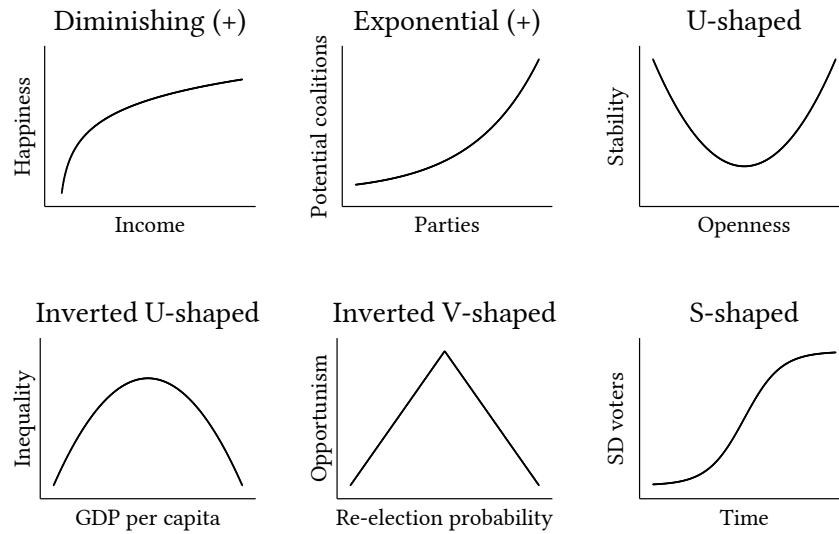
6 Non-linear relationships

In the linear regression model, it is assumed that a change in an explanatory variable, holding all other predictors constant, causes the same change in the dependent variable regardless of the initial values of the two variables. Put differently, we assume that the relationship between x and y can be characterized as a straight line. The constant slope of that line is equivalent to assuming that the effect is the same for all values of x and y .

However, sometimes there are reasons to expect that the relationship is far from linear, and at other times we want to allow for a non-linear relationship simply to see how much the estimated model differs from the linear one. In most such cases, this can be done within the framework of linear regression, simply by transforming the variables of interest. Thus, we still assume that y is a linear function of our independent variables, but these variables may include non-linear transformations of the original variables. For example, if we suspect that income has a diminishing effect on happiness, we could model this non-linearity by transforming the income variable (e.g., replacing it with its log) and then assume a linear relationship between happiness and logged income. It is also possible to transform the dependent variable, although that is less common.

In this section, we go through a number of non-linearities that can be modelled through such transformations. The examples are ordered by increasing flexibility. We begin with models that correspond to more or less clear assumptions about the functional form (i.e., the shape of the relationship), and then successively introduce more flexible specifications that are more sensitive to the data.

Which model we want to use depends on the situation at hand. Sometimes we have clear expectations about the functional form, perhaps based on common practice in previous research or because our theory predicts a certain kind of non-linear relationship. In these cases, it usually makes sense to use a simple model, partly because this is usually what theory leads us to expect (theory rarely predicts a relationship that follows a fourth-degree polynomial), and partly because the relationship is captured in one or two coefficients, which allows for intuitive interpretation and facilitates comparisons across specifications. All the functional forms presented in Figure 6 belong to this category. Some of them are more advanced than others, but they can all follow from quite simple theoretical models. At other times, we may not care about interpreting the functional form. This is usually the case when the variable is used as a control or when our objective is prediction rather than understanding a relationship. In such cases, we can use specifications that are as flexible as the data allow, because the loss of interpretability is irrelevant. Even then, however, linear specifications remain by far the most common. A third situation is that we want to test the validity of an assumed functional form by trying a more flexible specification. If the result still



From top left to bottom right, these graphs illustrate the ideas that i) income increases happiness, but with decreasing returns, ii) the number of possible coalitions increases exponentially with the number of parties, iii) political stability is higher in countries that are either very authoritarian or very democratic, iv) inequality increases in the first phase of growth but then begins to decrease, v) opportunistic policy-making increases as the probability of re-election approaches 0.5, and vi) if a party grows the way viruses spread – through interpersonal interactions – support will develop over time like an S-curve. Note that all of these relationships could be turned upside down if the effect had the opposite sign.

Figure 6: Examples of non-linear relationships

looks broadly like the functional form we initially assumed, that is a strong argument for keeping the simpler assumption. Because flexible specifications rely on a larger number of estimated parameters, which often lack a meaningful interpretation on their own, these models are usually presented and evaluated visually.

We now go through a number of common techniques for modelling non-linearities. However, it should be emphasized that the most important takeaway from this section is not the specific techniques described below, but rather the more general discussion of linearity summarized above. Read the subsections below as examples and applications of those broader points.

6.1 Logging variables for diminishing/exponential relationships

It is well known that economic prosperity has a positive but diminishing effect on public health. For example, a \$10 increase in daily income would mean a great deal for life expectancy in a poor country such as Zambia, whereas the effect in Sweden would be barely noticeable. A common explanation for this difference is that in Zambia, the additional income would to a much greater extent be used to cover basic needs such as food and medicine, whereas in Sweden the extra goods consumed would have little effect on people's health. These decreasing returns are illustrated in the left-hand panel of Figure 7. It is obvious that drawing a straight line through the dots in this scatter plot would be a poor representation of the data.

The standard solution for diminishing relationships like this one is to replace the

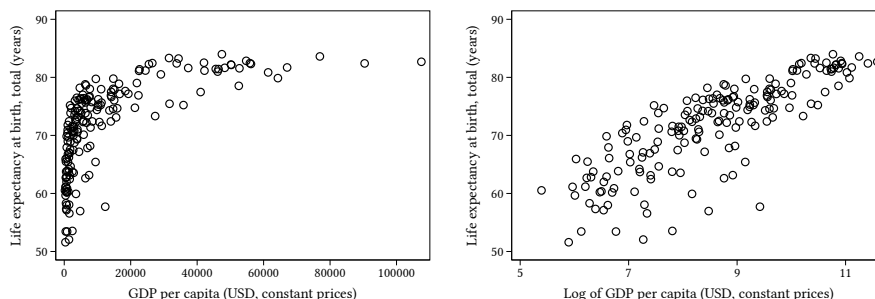


Figure 7: The relationship between GDP and life expectancy

independent variable with its logarithm. Such logs have the attractive property that a one-unit increase always corresponds to the same relative change. The most common logarithm has base 10, which means that a one-unit increase in the logged variable corresponds to a tenfold increase in the original variable. When we use a logged variable as an explanatory variable, we are therefore assuming that the outcome responds to relative changes ('if x increases by 100 percent') rather than absolute changes ('if x increases by ten dollars'). The choice to use a log transformation can be motivated both empirically ('it fits the data') and theoretically ('it is reasonable that doubling GDP produces the same change in life expectancy regardless of baseline GDP').

The right-hand panel of Figure 7 shows that the relationship looks much more linear after we transform the GDP variable. Perhaps it is not perfectly linear, but the attractive features of the log transformation – straightforward interpretation and the fact that it follows a common convention – mean that it is probably linear enough for our purposes.

It is also possible to use a log as the dependent variable, if we expect a certain increase in an independent variable to increase the outcome by some percentage. And if both sides are logged, the regression coefficient can be interpreted as an elasticity: if x increases by one percent, y is expected to increase by b percent.

6.2 Second-degree polynomials to model U-shaped relationships

Another form of non-linear relationship occurs when a value of the independent variable corresponds to a minimum or maximum value of the dependent variable. For example, income tends to increase with age up to a certain point, after which it begins to decline. A straight line would therefore be a poor summary of this relationship. Another classic example is Simon Kuznets's idea that the most unequal countries are those found in the middle of the income distribution. As a country undergoes industrialization and mechanizes agriculture, a surplus of labour keeps profits high and wages low. Eventually, however, this process leads to democratization and redistribution, which causes inequality to fall again.

Such U-shaped relationships are usually modelled as a second-degree polynomial. In the example above, the U is turned upside down, and such cases are often referred to as an 'inverted U-shape'. This corresponds to the assumption that there is some value of x at which y reaches a minimum or maximum, and that y increases or decreases as we move further away from that point. However, the functional form does not always

look symmetrical, because the minimum or maximum may be located far outside the observed range of the data. Technically, fitting a second-degree polynomial means including both the original variable x and the quadratic term x^2 in the regression equation.

$$y = a + b_1x + b_2x^2 + e \quad (4)$$

Although it is relatively easy to calculate where the minimum or maximum of such a curve is located, we recommend visualizing the shape of the relationship using a graph of predicted values ($\hat{y}$), or perhaps marginal effects ($b_1 + 2b_2x$). Remember to consider the distribution of your data when interpreting the graph. If all your data lie to the left or right of the minimum or maximum, it may not make much sense to describe the relationship as U-shaped.

When interpreting statistical significance for this kind of relationship, things are a little more complicated than usual. We are typically not interested in whether b_1 differs significantly from zero. If b_2 is statistically significant, that tells us that the relationship is indeed non-linear. And if we want to know whether there is any relationship between the independent and dependent variable, we need to test the joint null hypothesis that both coefficients are zero:

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0 \quad (5)$$

6.3 Quantile groups and other categories

Another way of dealing with possible non-linearities is to divide the data into groups and include a binary indicator for each group in the regression model, except one, which serves as the reference category. These categories may be quantile groups¹³ or other kinds of categories, such as age intervals like 15–24 (young), 25–64 (working age), and 65+ (elderly). An attractive feature of this approach is that we do not need to make any assumption at all about the functional form. It can therefore be a good choice either when your variable is naturally suited to a categorical specification – which is especially true for ordinal-scale variables – or when you simply do not want to impose assumptions about the functional form. The downside is that we must assume the effect to be constant within each group. And if we deal with that by making the groups smaller, we lose precision in the estimation of the different group effects.

If you have natural or theoretically guided categories, the number of categories will usually follow from the distinctions you care about. If not, or if you are using quantile groups, the appropriate number of groups depends on two things. First, the more observations you have, the more groups you can use without introducing too much sampling error within each group. More precisely, it is not just the number of observations that matters, but the amount of identifying variation, which also depends on the data and the model specification. Second, the appropriate number of groups depends on how much information you want to show: only the broad pattern, or also local deviations, at the cost of more sampling error.

¹³ Quantile groups are equally sized groups based on intervals of a continuous variable. Depending on how many groups there are, it is common to use the terms quartile (four groups), quintile (five groups), decile (ten groups), and percentile (one hundred groups). So, for example, the fourth quartile group is simply the top 25 percent of the observations, and the third decile group consists of the individuals between percentiles 20 and 30. One note on terminology: ‘quantile’ refers to the cut-point between groups, whereas ‘quantile group’, ‘bin’, ‘class’, or ‘interval’ refers to the observations between these cut-points. However, it is common – though somewhat sloppy – to use ‘quantile’ also for the groups themselves.

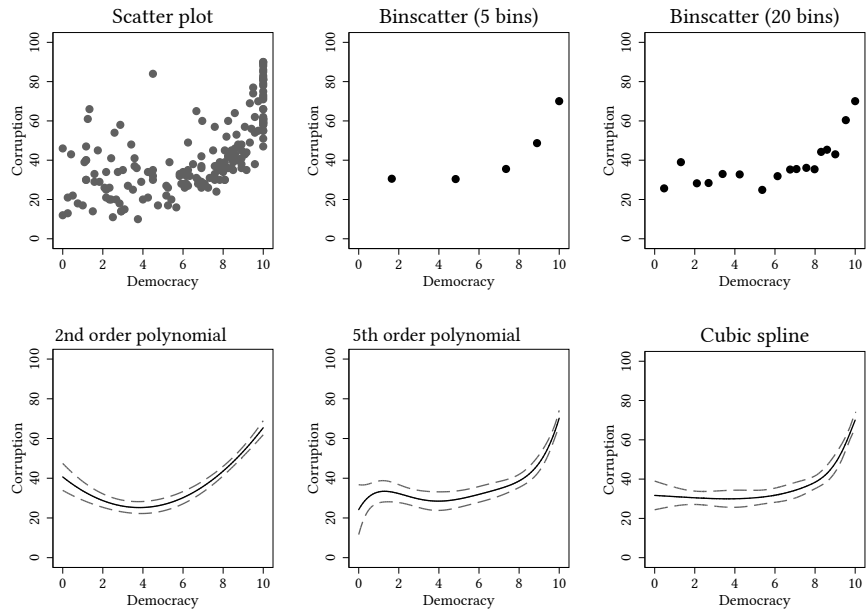


Figure 8: Describing the non-linear relationship between democracy and corruption

6.4 Polynomials and non-parametric smoothers

If we want a flexible model without grouping the data, we can also increase the complexity of the assumed functional form. The most straightforward alternative is simply to add higher-order powers of the independent variable, thereby fitting a third-, fourth-, or even fifth-degree polynomial. For most people, the regression coefficients in such models are impossible to interpret directly, so if interpretation matters, the relationship needs to be presented graphically.

In Figure 8, the non-linear relationship between the degree of democracy (measured using democracy indices from Freedom House and Polity) and the level of perceived corruption (measured using an index from Transparency International) is presented in six different ways. The top panels show a standard scatter plot, followed by binned scatterplots in which the data have been grouped into five or twenty quantile groups. The bottom panels show three different functions that can be used to model a non-linear relationship.

The main takeaway is that all approaches provide a reasonably good fit to the data. The five bins in the top-middle panel summarize the pattern nicely, but a sceptical reader may wonder whether the small number of bins hides valuable information. The binscatter with 20 bins suggests that this is not the case. And at a quick glance, the functions in the bottom panel probably look almost identical. This is not uncommon. For most datasets, different flexible models tend to yield fairly similar descriptions, especially when there are many non-clustered and evenly distributed data points.

But there are also differences across these curves that are typical of the different specifications, especially in the left tail. The second-degree polynomial indicates a rather high level of corruption at the lowest levels of democracy. This is because the simple specification is limited in the shapes it can take and, most importantly, must be

symmetrical around the minimum point: if there is a sharp increase to the right of the minimum point, there will be a similar increase to the left, and vice versa. In this case, the curve provides a reasonable fit to the data, although the behaviour in the left tail is perhaps not ideal.

The tendency is the opposite in the fifth-degree polynomial, where the curve turns downward in the left tail. Such sharp turns are common in higher-degree polynomials. This does not necessarily make the model a poor fit, especially not when it is used as a control, but these graphs often appear jumpy at the edges and can be seriously misleading if you try to extrapolate beyond the range of observed data.

This is why splines are so useful. We will not go into the details of how splines work, but their defining characteristic is that we estimate separate functions for different intervals of the data, and that these functions are joined at different cut-points (so-called ‘knots’). In the bottom-right panel of Figure 8, we use a restricted cubic spline, which is a third-degree polynomial that is constrained to be linear at the edges of the data distribution. As we can see, using a spline rather than a classic polynomial fit produces much less ‘jumping’ at very low levels of democracy. However, splines can be difficult to work with in Stata, and they are not always worth the extra effort.

If we only want to graph the relationship, and do not intend to use the flexible specification in a regression model, we can also use different kinds of non-parametric (or semi-parametric) methods. By non-parametric, we mean that we do not assume a function and estimate its coefficients. Instead, we calculate some local quantity using only the nearest observations for each point along the x-axis. The most familiar example is the moving average. If you are interested in such techniques, the Wikipedia articles on ‘Smoothing’ provide a decent starting point.

6.5 Linearity and the level of measurement

Linearity typically assumes that variables are measured on an interval or ratio scale. If one-unit differences in a variable are not proportional, it is unlikely that a one-unit increase in the independent variable would cause the same change in the dependent variable regardless of the initial values of the two variables.

However, this does not imply that we cannot include variables measured on a nominal or ordinal scale. It is quite common for relationships involving an ordinal variable to be approximately linear, in which case the linear model tends to work well. If this is not the case, and the variable is an independent variable, then we can estimate separate effects for each category, just as we did in Section 6.3. Below, we discuss situations in which the dependent variable is binary or ordinal.

Binary dependent variables

What should we do if the outcome variable is dichotomous? This simple question is answered differently across disciplines and even subfields, which once again shows that quantitative methods are sometimes shaped as much by convention as by statistics. Some argue that a linear model still works well, in which case we interpret the predicted values as probabilities, whereas others recommend logistic regression.

Proponents of logistic regression rightly argue that the true functional form cannot be linear, because linear models can predict probabilities below 0 or above 1. Advocates of OLS reply that the linear model does a good job of estimating the average marginal effect, which is usually the quantity we care about, and that logistic regression is more difficult to interpret because additional steps are needed to calculate marginal effects.

A reasonable conclusion from these arguments is that no model is perfect, and that the choice of model depends to a large extent on the purpose of the analysis. If we are interested in the effect that one variable has on another, the easy interpretation of OLS is a strong argument in its favour. This is especially true if we are only interested in the *average* marginal effect, because even if the linear model is not a good representation of the true functional form, it is still a well-behaved estimator of that average marginal effect. However, one should be cautious about drawing inferences for subsamples or extrapolating the results, because even if the estimated effect is a good estimate of the *average* marginal effect, the effect may differ substantially across groups. Remember that a non-linear relationship implies that the marginal effect is not constant. And if we include interaction variables in this model, we must acknowledge the risk that the interactions are in part capturing this non-linearity.

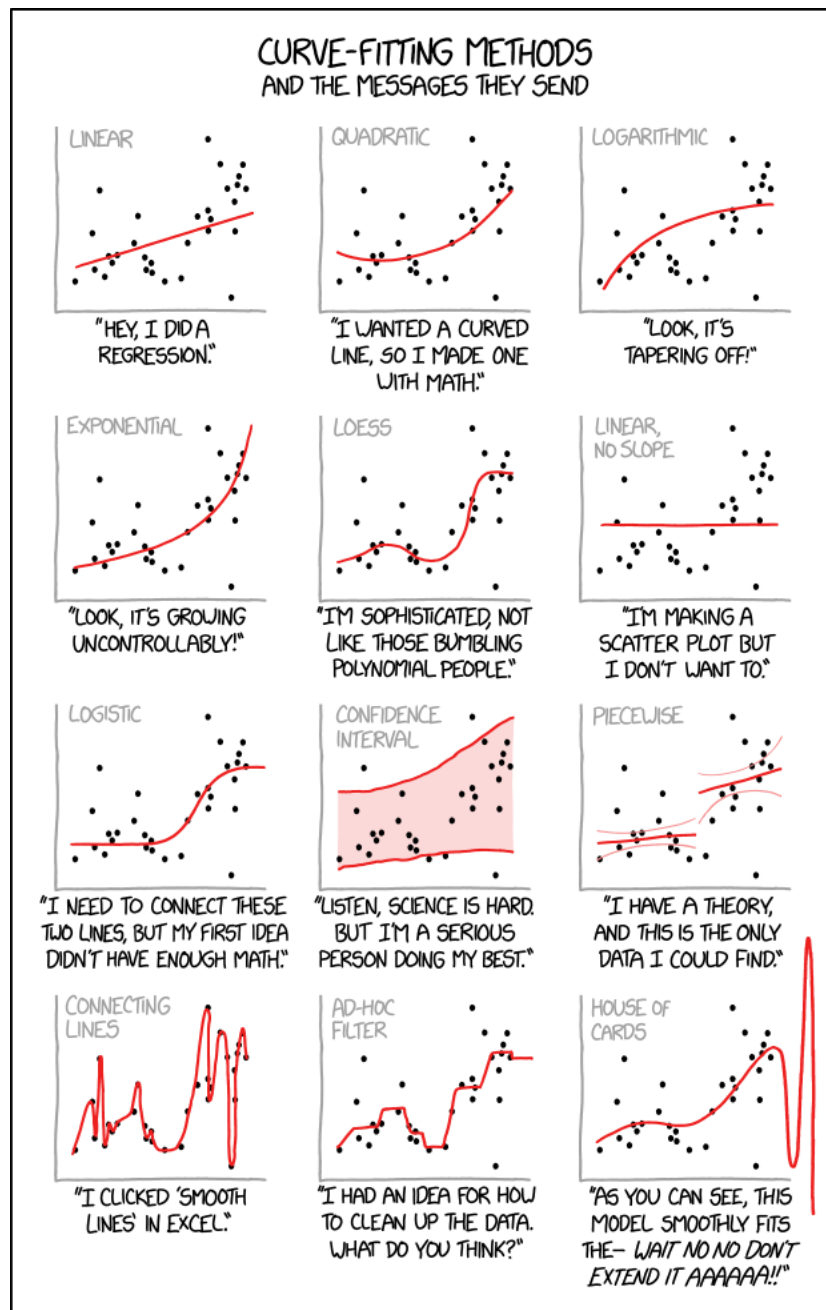
If, by contrast, our primary interest lies in prediction, and computing power is not an issue, then a logit (or probit) model makes more sense. This model also assumes a specific functional form, and there is no reason to believe that this is exactly what the true relationship looks like; however, it probably comes closer than the linear model does. We do not cover logistic regression in this course. If you think it is something you want to use, we recommend the short booklet by Pampel (2020). If you only want the basics and would prefer something in Swedish, SPSS-akuten has a good guide on the subject.

Ordinal dependent variables

It is common to have outcome variables that, strictly speaking, are measured on an ordinal scale. In principle, treating such variables as interval-scaled can distort the results, and in extreme cases even alter the substantive interpretation of the relationship. We should therefore be cautious about using dependent variables for which it is obvious that one-unit differences are not comparable across the scale. However, the kinds of ordinal scales commonly used in regression analyses – such as survey items where respondents answer on a scale from 1 to 5 or from 0 to 10 – tend to work quite well as outcomes in a linear model (Labovitz 1967; Norman 2010).

If you have a dependent variable that is, strictly speaking, measured on an ordinal scale, but it seems plausible to approximate it as an interval scale (i.e., to assume that all one-unit differences are proportional), then there is usually little reason for concern, and there are probably other aspects of your model that deserve more attention. However, if you are concerned, or if you want to quantify the importance of this linearity assumption, there are a few things you could do. First, you could use a binned scatter plot or something similar to evaluate whether the bivariate relationship looks approximately linear. Second, you could actually test how sensitive your results are to violations of the interval-scale assumption.¹⁴ Third, the most common way of handling

¹⁴ I have never seen this done, but I think there are two kinds of analyses that could be very convincing. First, if your independent variable is binary, then you can show how your outcome is distributed conditional on the independent variable. If all the high values are more common for one of the variable values, and all the low values are more common for the other, then we can confidently say that the average is higher in the first group. But if the pattern varies a great deal across adjacent values, then the linearity assumption is central to the interpretation. Second, you could simulate new data in which you allow different violations of the interval-scale assumption, and then show how sensitive your results are to these violations. For example, assume that each step is either 1 (as you assumed) or twice as large, and test all possible combinations ranging from one step being twice as large to all steps except one being twice as large. If *all* steps are twice as large, their relative size is unchanged. Run a new regression for each combination and present the resulting coefficients in a histogram, or zoom in on the steps that matter most for your interpretation.



problematic ordinal-scale outcomes is to use some form of *ordinal regression*. We do not go through that here, because it is more complicated than it is useful for our purposes, but the Wikipedia article of the same name is a good place to start.

7 Outliers and influential observations

Individual observations that differ substantially from the rest of the data are called outliers. Such observations can be interesting in their own right, for example as cases for a case study, but they may also affect our regression results much more than other observations do. In fact, removing a single observation can sometimes fundamentally alter the results. In this section, we examine the conditions under which an outlier affects the results, how to detect such observations, and what we should do once we have identified important outliers.

Let us begin by going through different kinds of ‘outlierness’. First, an observation is said to have *leverage* if it has an extreme value on the independent variable. If its value on the dependent variable is what we would expect given the regression line (or rather the regression line estimated without this observation), then the observation does not affect the slope coefficient. However, it does improve statistical precision, and thus reduce standard errors, more than other observations.

Second, if an observation has an unusual value on the dependent variable given what we would expect based on its values on the independent variables, we call this a *vertical outlier*, a *regression outlier*, or an observation with *discrepancy*. Such observations tend to have large residuals (by definition, if we exclude them when estimating the regression), but their variable values do not need to be extreme in themselves.

To affect the regression line, an observation must have both leverage and discrepancy. We call such cases *influential observations*, and you can visualize their impact by imagining that they pull the regression line toward themselves as if they were magnetic.

7.1 Identifying outliers

It is always a good idea to familiarize yourself with the data before you begin specifying models. Which tools to use depends on personal preferences as well as on the nature of the data, but there is no need to make things complicated. In all its simplicity, a standard scatter plot will indicate whether there are observations with extreme values on your two most important variables, as well as whether there are observations that lie far from the (bivariate) regression line. You can also sort the dataset by your key variables – and later by the residuals, for that matter – and inspect the observations at the top and bottom of the distribution. How extreme are they? Why do you think such extreme values appear in this country or that municipality? And do they correspond to what you observe on other variables?

There are also more formal procedures for identifying outliers. The appropriate statistic depends on what kind of outlier we are interested in. In our experience, we often want to know which observations had the greatest influence on a specific regression coefficient. The statistic that measures this kind of influence is called *DFBETAs*. For each observation, it shows how much a parameter would change if we excluded that specific observation from the analysis, standardized by dividing the change by the parameter’s standard error. However, there are other statistics available if we want to measure leverage, discrepancy, or overall influence on the model.¹⁵ I will

¹⁵ We can measure leverage through *hat values* (a measure of how far an observation’s values on the independent variables are from those of the other observations) and discrepancy through *studentized residuals* (for each observation, we calculate what the residual would have been if the observation had not affected the regression line, and then standardize it). If we want to know how much each observation

not go into the mathematics here, but all these statistics are easily obtained in Stata.

Sometimes we are interested in how influential different clusters of observations are. For example, if we have data that span both countries and years, we may want to know how sensitive the results are to the exclusion of a single country, that is, a set of observations. In such situations, the best approach is usually to write a section of your do-file in which you exclude countries one by one and store the regression results.

It should also be mentioned that we often ignore outliers altogether. Sometimes this is a mistake, but it can also be because our data make outliers unlikely. For example, if we have a survey with a large number of respondents, and all our key variables are constrained to a small number of possible values, then we can be fairly confident that whatever results we find will not be driven by a handful of observations.

7.2 Dealing with outliers

Sometimes it is an unusual measurement error that has caused an observation to become an outlier. In such cases, it is obvious that we should either replace the erroneous value with the correct one or exclude the observation from the analysis. In doing so, it is important to try to understand what caused the error, whether there may be other incorrect values that we do not observe, and whether excluding the cases we can identify may induce some kind of selection bias.

In most cases, outliers represent correct but unusual values. We should then direct our attention to two questions: i) how sensitive are the results to the exclusion of one or a few outliers, and ii) are the observations driving the results consistent with your theory and assumed mechanisms?

To answer the first question, you can exclude the most influential observations one by one and write something about how this affects your results. If the results change a great deal, you should be transparent about that and discuss whether it is a problem that they are so sensitive to the exclusion of these outliers. Unfortunately, there is seldom a better solution than simply being open about the fact that the results are not very robust in this respect.

To answer the second question, you should identify the most influential observations that also have DFBETAs with the same sign as your main effect (if they have the opposite sign, they are pushing your results toward zero). Who are they? What do you know about these observations? On their face, do they seem representative of the pattern you are describing? Are there other plausible explanations for this particular outcome that are missing from your regression model? The usefulness of such an analysis depends on the kind of data you have and is, of course, greater when your observations come with a lot of contextual information than when you are dealing with anonymous data points about which you know very little.

To some extent, the problem of outliers reflects the large weight that OLS gives to large deviations from the regression line. An alternative to excluding outliers may therefore be to use an estimator that down-weights the most influential observations. This is especially relevant when your data are distributed in such a way that removing one outlier would merely lead to another observation taking its place. *Robust regression* is an umbrella term for such estimators (though the term is sometimes also used for other kinds of robustness checks). We will not go into the details here, but it is useful to know that such methods exist if you run into outlier problems that are difficult to handle.

influences the results more generally, we can calculate *Cook's distance*, or *Cook's D*.

8 Working with panel data

Panel data arise when we observe the same units repeatedly over time. In other words, each observation is located both in *space* (or across units) and in *time*. Typical examples include country-year data, such as the Quality of Government (QoG) data used frequently in this course; data on Swedish municipalities observed across electoral terms; or survey data in which the same respondents are interviewed more than once. Compared to a purely cross-sectional dataset, panel data give us more information about both differences between units and changes within the same unit over time. At the same time, this structure creates several methodological challenges that do not arise, or arise less clearly, in ordinary cross-sectional analysis.

Some related terms are worth keeping apart. If we follow a single unit over time, we usually speak of a *time series*. If we follow many units over time, we usually speak of *panel data*. A common distinction within panel data is between *time-series cross-sectional* (TSCS) data and *longitudinal* data. In broad terms, both are forms of panel data, but TSCS usually refers to repeated observations on fixed aggregate units such as countries, states, or municipalities, with a fairly long time dimension. By contrast, longitudinal data more often refer to repeated observations on individuals or other micro-level units, where the number of units typically exceeds the number of time points by a wide margin. In practice, the boundary is not always sharp. Many commonly used macro-level datasets are analyzed as if they were panel data even when the underlying observations are partly built from repeated cross-sections rather than the same individuals being followed over time, as is often the case for aggregated outcomes such as country-level unemployment rates.

To make this structure explicit, we usually index observations with one subscript for the unit and one for time. A simple panel-data model can therefore be written as

$$y_{i,t} = a + b_1x_{i,t} + b_2z_i + e_{i,t} \quad (6)$$

where i denotes the unit (for example, a country, municipality, firm, or individual) and t denotes time. The dependent variable $y_{i,t}$ and the independent variable $x_{i,t}$ both vary across units and over time, whereas z_i varies across units but is constant over time.

The three subsections that follow focus on the main implications of this data structure. The first discusses the inferential challenges created by dependence in the residuals and how these affect standard errors. The second turns to identification, and in particular to how variation across units and over time can be used – or restricted – in order to make causal interpretation more credible. The third discusses time-series dynamics, that is, how repeated observations over time force us to think about lagged effects, persistence, and the timing of causal relationships.

8.1 Inferential challenges of panel data

Conventional standard errors are only correct under the assumption that all residuals are independent of one another. In panel data, however, this assumption is rarely plausible. Residuals for a given unit at time t are often correlated with residuals for the same unit at earlier time points, such as $t - 1$ or $t - 2$. This form of temporal dependence is usually referred to as *autocorrelation* or *serial correlation*. It is likely to arise when there is unmodelled persistence in the dependent variable or in important omitted variables.

Residuals may also be correlated *across* units at the same point in time. In other words, residuals observed in a given year may be correlated across countries, firms,

or individuals. This is often called *contemporaneous correlation* or, in some contexts, *spatial correlation*. For example, if we are studying an economic outcome, we would expect common global shocks to affect many countries in the same year, generating contemporaneous correlation. We might also expect this correlation to be especially strong among countries that are geographically close, closely linked through trade, or have similar industrial structures.

Unless these forms of dependence are taken into account, the standard errors will be incorrect, in most cases understating the true uncertainty surrounding the estimated regression coefficients. There is no universally correct way to adjust standard errors, because the appropriate adjustment depends on the residual dependence structure in the data. The most common approach is to cluster the standard errors at the panel, or unit, level (for example, by country or individual), which allows for arbitrary correlation within units over time and thereby addresses serial correlation (see also Section 2.4). This approach can be extended to two-way clustering, which adjusts for correlation both within units over time and within time periods across units, for example by clustering by both country and year.

Two additional points are worth noting when working with clustered standard errors. First, if the number of clusters is small (say, only a few dozen, or less), conventional cluster-robust standard errors can be unreliable (see Cameron and Miller 2015, for a discussion of the problem and possible solutions). Second, in more design-based research – not least difference-in-differences, which can be seen as a particular way of analyzing panel data – the standard recommendation is to cluster the standard errors at the same level at which the treatment is assigned. This means that if treatment varies at the state, municipality, school, or firm level, then residual dependence induced by that treatment is also likely to operate at that level, and the clustering should match it.

In this course, we will not use any adjustments other than clustering. This is primarily because clustering is the current standard in much applied work, not because it is necessarily superior to all alternatives (such as panel-corrected standard errors, which students of this course encounter in the texts by Beck and Katz). In practice, researchers should always ask whether clustering adequately captures the dependence structure they expect in the residuals, or whether the setting calls for a different solution.

8.2 Identification with variation across space and time

The previous section focused on inference: more specifically, on how residual dependence in panel data affects standard errors and how we can adjust them. This subsection addresses a different issue, namely identification. Even if we compute the standard errors correctly, we still need to ask what kind of variation in the data is being used to estimate the relationship of interest, and whether that variation is likely to support a credible causal interpretation.

A central feature of panel data is that they vary along two dimensions: across units and over time. This means that regression coefficients may be identified partly from differences *between* units and partly from changes *within* the same unit over time. For example, if we study countries over time, some of the variation comes from the fact that some countries have higher values than others on average (between-variation), while some comes from the fact that the same country changes from one year to the next (within-variation). These two sources of variation are often substantively quite different, and they may differ in how vulnerable they are to omitted variable bias.

8.2.1 Fixed effects

As discussed earlier in the text (see Section 3.3), one common way of restricting the analysis to within-unit variation is to use *unit fixed effects*. In a panel-data setting, unit fixed effects control for all omitted variables that are constant over time within each unit. If the units are countries, this includes all stable country-specific characteristics, whether observed or unobserved. In that sense, fixed effects are often attractive because they allow us to net out persistent differences between units and instead estimate the relationship using only changes within the same unit over time.

Formulated in this way, unit fixed effects can sound like an obvious improvement. But they are not a free lunch. First, fixed effects only improve identification if the within-unit variation is less vulnerable to omitted variable bias than the variation between units. This is often a plausible assumption, but it is not guaranteed. Time-varying confounders may still bias the estimates, and in some settings the remaining within-unit variation may even be more difficult to interpret than the between-unit variation that has been removed.

Second, unit fixed effects will often reduce statistical precision. By absorbing all stable differences between units, they remove a substantial part of the variation in the data. The regression coefficients are then estimated using a smaller pool of variation, which typically leads to larger standard errors (compare the discussion of precision in Section 2.3). In other words, fixed effects may reduce one source of bias, but they often do so at the cost of noisier estimates.

Despite these two limitations, the most common strategy in applied panel-data research is to include unit fixed effects. The reason is that stable differences between units are often especially difficult to interpret causally. Countries, firms, schools, and individuals differ from one another in many ways that are hard to observe, measure, or model directly. Restricting the analysis to within-unit change is therefore often seen as a reasonable baseline, even when it is not a complete solution to the identification problem.

However, whether unit fixed effects are the most appropriate baseline depends on the empirical setting. They are especially common in country-year data, where units often differ in persistent and difficult-to-measure ways, and where many outcomes change only gradually over time. They are often less central in settings such as election-period data with fast-moving policy outcomes, where the relevant variation may be concentrated in shorter-term shifts rather than in stable cross-unit differences. Their use also depends on model specification. Unit fixed effects are particularly common in stationary models estimated in levels, whereas they are often less central in first-differenced specifications, which already remove all time-invariant unit-level factors, and in dynamic models with lagged dependent variables, where part of the persistence is modeled directly rather than absorbed by the fixed effects.

First differences

A closely related way of focusing on within-unit variation is to estimate the model in *first differences*. The idea is simple: instead of explaining the level of $y_{i,t}$, we explain the change from one period to the next, $\Delta y_{i,t} = y_{i,t} - y_{i,t-1}$. If we start from a model with an unobserved unit-specific component c_i ,

$$y_{i,t} = a + bx_{i,t} + c_i + e_{i,t}, \quad (7)$$

then taking first differences yields

$$\Delta y_{i,t} = b \Delta x_{i,t} + \Delta e_{i,t}, \quad (8)$$

because the time-invariant unit component c_i drops out. In this sense, first differences remove the same class of time-invariant omitted variables as unit fixed effects.

First differences and fixed effects therefore often tell a similar substantive story, but they emphasize slightly different identifying variation. Fixed effects estimate relationships in levels using deviations from each unit's mean, whereas first differences estimate relationships in changes from one period to the next. First differences may be attractive when the outcome is best understood as a change process or when long-run cross-unit differences are hard to interpret. On the other hand, differencing can amplify noise and is less suitable when most meaningful variation occurs at low frequencies rather than year-to-year fluctuations. As always, the choice should be guided by theory and by a clear idea of what variation is substantively meaningful.

Random effects

An alternative to unit fixed effects is to use *random effects*. Like pooled OLS, a random-effects model makes use of both variation *between* units and variation *within* units over time. Unlike pooled OLS, however, it does not simply ignore the panel structure. Instead, it explicitly models the fact that observations from the same unit are linked by a common unit-specific component, and it estimates the model using generalized least squares (GLS) rather than ordinary least squares. Under the standard random-effects assumptions – in particular, that the unit-specific component is uncorrelated with the regressors – this makes the estimator more efficient than pooled OLS, because it takes the panel structure into account rather than treating repeated observations on the same unit as if they were fully independent. Put intuitively, the estimator gives relatively more weight to variation that adds new information and relatively less weight to observations that largely reproduce the same underlying unit-specific pattern.

Compared to fixed effects, the advantage of random effects is different. Under the same assumptions, random effects is more statistically efficient than fixed effects because it uses more of the available variation in the data. Fixed effects discard all stable between-unit variation and rely only on within-unit change. Random effects, by contrast, retain both within-unit and between-unit variation. This often improves statistical precision and also makes it possible to estimate the effects of variables that are constant, or nearly constant, within units over time. The cost is that random effects require a stronger assumption: the unobserved unit-specific component must be uncorrelated with the regressors. If that assumption fails, the random-effects estimator is generally inconsistent, whereas fixed effects remains consistent under weaker conditions.

Random effects is often described as a compromise between pooled OLS and fixed effects. Strictly speaking, it is best understood as a GLS estimator that combines the *between* and *within* estimators. But the intuition is similar. Instead of removing the full unit mean, as fixed effects does, random effects removes only a fraction of it. It therefore remains closer to pooled OLS when the unit-specific component is small, and moves closer to fixed effects when that component is large. In this sense, random effects can be understood as a partial de-meaning approach that lies between complete pooling and full fixed-effects transformation.

In practice, the use of random effects varies across disciplines and subfields. In some areas – especially those working with classic micro-panels or multilevel data – random

effects plays a prominent role. In most applied social-science work, however, pooled OLS and especially fixed effects are more commonly used as baseline specifications. This reflects both substantive caution and disciplinary convention: researchers are often reluctant to rely on the stronger assumptions required by random effects unless they have a clear reason to do so.

A common textbook procedure for choosing between fixed and random effects is the *Hausman test*. Strictly speaking, it is not a general test between OLS, fixed effects, and random effects, but a test of whether the difference between the fixed-effects and random-effects estimates is large enough to cast doubt on the random-effects assumptions. The intuition is straightforward: if the random-effects assumption is correct, then both estimators should be consistent, but random effects should be more efficient; if that assumption is violated, the random-effects estimator becomes inconsistent, while fixed effects remains consistent under weaker assumptions. A large discrepancy between the two estimates therefore suggests that the unobserved unit-specific component is correlated with the regressors, in which case fixed effects is usually preferred.

At the same time, the test has important limitations. A statistically insignificant difference does not show that the difference is truly zero; it may simply reflect limited statistical power. Nor does the test evaluate the substantive trade-off between bias and variance: in some settings, an estimator with a small amount of bias may still be preferable if it offers a substantial gain in precision. Finally, the logic of the test treats the fixed-effects estimator as the benchmark, but fixed effects can itself be biased if the model is misspecified, for example because of time-varying omitted variables or problematic dynamics. The Hausman test should therefore be seen as a useful diagnostic, not as a mechanical decision rule.

Time

In panel data, it is often useful to control not only for differences *between* units, but also for factors that vary *over time* in ways that affect many or all units at once. One reason is that units are often exposed to common shocks, such as recessions, wars, pandemics, or changes in commodity prices. Another is that many outcomes are shaped by broader temporal trends, such as technological change, demographic shifts, or gradual changes in social norms. If such time-varying influences are related to the explanatory variables of interest but not adequately captured by the model, they may bias the estimated coefficients.

The logic of *time fixed effects* is directly analogous to that of unit fixed effects. Just as unit fixed effects control for all omitted factors that are constant over time within each unit, time fixed effects control for all omitted factors that are common to all units within a given time period. In a country-year dataset, for example, year fixed effects absorb everything that affects all countries in a particular year in the same way, whether observed or unobserved. This can be useful when we believe that the most important confounding factors are period-specific rather than unit-specific.

In principle, time fixed effects can be used on their own, just as unit fixed effects can. In practice, however, they are more often combined with unit fixed effects in a *two-way fixed effects* specification. Such a model controls both for all stable differences between units and for all common shocks in each time period. In that sense, it addresses two important sources of omitted variable bias at once: time-invariant heterogeneity across units and period-specific influences shared across units. For this reason, two-way fixed effects is a very common baseline specification in applied work with panel data. After

adding both unit and time fixed effects, the remaining identifying variation comes from deviations *within* units over time relative to the common time pattern—that is, from changes in a unit’s x that are not shared by all units in the same period. In many applications, this is a demanding requirement: once stable between-unit differences and common year-to-year movements have been absorbed, relatively little variation may remain for estimating the effect precisely, and standard errors can become large.

At the same time, time fixed effects are not always the only, or the best, way of controlling for temporal variation. A less restrictive alternative is to include a *linear time trend*. Whereas time fixed effects allow each period to have its own separate intercept, a linear trend assumes only that the outcome (or the omitted variables affecting it) follows a common straight-line development over time. This uses fewer degrees of freedom and may therefore preserve more precision, but it is also more restrictive: if the true time pattern is not well approximated by a straight line, the adjustment may be inadequate.

With more fine-grained time data, it may also be relevant to allow for non-linear trends or to adjust for seasonal patterns. For example, monthly or quarterly data may exhibit recurring within-year fluctuations that should be modeled explicitly. However, such adjustments are less central for the types of panel data considered in this text, where the time dimension is often measured in years and the main concern is usually broader common shocks and long-run trends rather than short-run seasonal variation.

Content:

8.3 Time-series dynamics

When data are collected repeatedly over time for the same unit – such as an individual, country, or stock – we refer to them as *time-series* or *dynamic* data. These may take the form of a single time series, or of multiple parallel time series observed for many units, as in panel data or time-series cross-sectional data. Although the discussion in this section is relevant to both settings, it is written primarily with the latter in mind. Researchers working with a single time series often draw on a more specialized toolbox – namely, time-series analysis – and tend to approach these issues somewhat differently.

Compared to cross-sectional data, a key difference is that time-series data require us to model dynamics. By this, we mean making assumptions about how quickly variables affect one another and how persistent those effects are. If we assume an immediate effect – that is, if a one-unit increase in x is expected to produce an immediate change of b in y , and if x then returns to its previous value, y is expected to do the same – we would write the outcome for unit i at time t ($y_{i,t}$) as a function of the independent variable observed at the same point in time ($x_{i,t}$):

$$y_{i,t} = a + bx_{i,t} + e_{i,t} \quad (9)$$

There are two basic ways of relaxing this assumption. We can either include values of the independent variable from earlier points in time, or do the same for the dependent variable. Both possibilities are discussed below. For ease of presentation, I assume that t is measured in years, although the same logic applies if time is measured in days, weeks, or some other interval.

Time-series data also raise a number of additional issues that are not discussed here, such as stationarity and serial correlation in the residuals. If you want to specialize in this area, it is a good idea to consult the short subsection on further readings.

8.3.1 Lagging the independent variable

The first possibility is to assume that the independent variable has both an immediate effect (b_1) and an effect in the following year (b_2), after which the effect returns to zero. In the regression equation, this is achieved by including the lagged variable $x_{i,t-1}$, which represents the value of the independent variable (x) in the preceding year ($t - 1$). This is the simplest form of a so-called Finite Distributed Lag (FDL) model. If we believe the effect persists for two or three years as well, we can easily add more lags to the model ($x_{i,t-2}$, $x_{i,t-3}$, etc.).

$$y_{i,t} = a + b_1x_{i,t} + b_2x_{i,t-1} + e_{i,t} \quad (10)$$

Once lagged values of x are included, it is no longer sufficient to focus on a single regression coefficient. If x increases permanently by one unit, y is assumed to change immediately by b_1 , and from $t + 1$ onward by $b_1 + b_2$ (relative to $t - 1$). If x instead increases temporarily by one unit, the expected immediate change in y is b_1 , but at $t + 1$ the effect becomes b_2 , after which it is assumed to disappear.

8.3.2 Lagging the dependent variable

The other possibility is to include a lagged dependent variable (y_{t-1}) in the model. This is a very common specification and is usually referred to simply as a Lagged Dependent Variable (LDV) model. Why might we assume that the dependent variable affects its own future values? One useful way to think about this is as a form of inertia in a slow-moving system. For example, if y is a policy outcome (such as spending), we might expect policymaking to be incremental. Another interpretation is that the lagged dependent variable captures the combined influence of omitted lagged effects from many known and unknown variables. In the equation below, λ is used as the coefficient for y_{t-1} in order to distinguish it clearly from the other coefficients.

$$y_{i,t} = a + \lambda y_{i,t-1} + bx_{i,t} + e_{i,t} \quad (11)$$

In most cases, λ will take a value between 0 and 1. Values close to zero mean that the dependent variable is only weakly shaped by its past, so changes in the independent variables have nearly their full effect immediately. Values close to 1, by contrast, mean that the dependent variable is highly persistent, so even a temporary change in the independent variables can have a long-lasting effect on the outcome.

These dynamics are illustrated in Figure 9. The left-hand panel shows the expected change in y if x temporarily increases by one unit, while the right-hand panel shows the expected change produced by a similar permanent increase in x .

As long as λ is smaller than one ($|\lambda| < 1$), the long-run effect of a *temporary* change in the independent variable will always be zero, because the outcome tends to return to its equilibrium. By contrast, the long-run effect of a *permanent* change in x can be calculated using the following expression:

$$\hat{y}_{i,\infty} - \hat{y}_{i,t} = \frac{b}{1 - \lambda} \quad (12)$$

This makes clear that if $\lambda = 0$, the long-run effect is simply b , just as in models without lags. As λ approaches one, the long-run effect becomes increasingly large relative to b . If $\lambda \geq 1$, the model is likely to be misspecified, and you should consult the sections on *stationarity* in the suggested readings below.

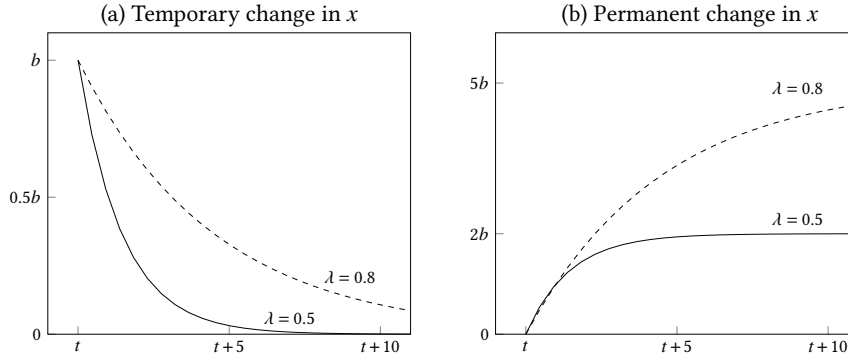


Figure 9: Predicted change in outcome in the LDV model

A final caution concerns dynamic models that include both unit fixed effects and a lagged dependent variable. In such models, the lagged dependent variable is mechanically related to the omitted variables that remain after the fixed-effects transformation. As a result, the estimated coefficient on $y_{i,t-1}$ can be biased when the time dimension is short. The problem becomes less important as the number of time periods grows, but it can be non-negligible in panels with few time points.

8.3.3 Lagging both variables

In many applications, it is plausible that both of the mechanisms discussed above are present at the same time. The dependent variable may be persistent, so that past values of y help predict current outcomes, while the independent variable may also have effects that unfold gradually over time. In such cases, it is often useful to include lags of both the dependent and the independent variable in the same model. This type of specification is usually referred to as an *Autoregressive Distributed Lag* (ADL) model.

The simplest version combines one lag of the dependent variable with one lag of the independent variable:

$$y_{i,t} = a + \lambda y_{i,t-1} + b_1 x_{i,t} + b_2 x_{i,t-1} + e_{i,t} \quad (13)$$

This model combines the logic of the LDV and FDL specifications. The coefficient λ captures persistence in the dependent variable, while b_1 and b_2 describe the immediate and delayed effects of the independent variable. In other words, the current value of y is allowed to depend both on its own recent past and on both current and past values of x .

As in the previous subsections, the coefficients should not be interpreted in isolation. The coefficient b_1 captures the immediate effect of a one-unit increase in $x_{i,t}$, holding the lagged variables constant. But because the model also includes $y_{i,t-1}$, part of this initial effect may persist into future periods through the autoregressive term. Likewise, the lagged coefficient b_2 contributes to the effect in the following period. For this reason, ADL models are usually interpreted in terms of short-run and long-run effects rather than as a set of unrelated coefficients.

If x increases permanently by one unit, the long-run effect in this simple ADL model is given by:

$$\hat{y}_{i,\infty} - \hat{y}_{i,t} = \frac{b_1 + b_2}{1 - \lambda} \quad (14)$$

This expression shows that the long-run effect depends both on the total effect of the current and lagged values of x , and on how persistent the dependent variable is. If λ is close to zero, the long-run effect will be close to $b_1 + b_2$. If λ is large, the long-run effect becomes substantially larger, because changes in y carry forward over time.

ADL models are often attractive because they are flexible and can represent several plausible dynamic processes within a single specification. At the same time, they are also more demanding to estimate and interpret than simpler models. Beck and Katz (2011) emphasize that the ADL specification is a useful starting point because it nests several common dynamic models. For example, the ADL(1,1) reduces to the LDV model when $b_2 = 0$, and to a first-order FDL model when $\lambda = 0$, which makes it possible to test whether these simplifying assumptions are supported by the data. As always, the choice of specification should also be guided by substantive theory and by an intuition for how quickly we expect causes to affect outcomes.

Unbalanced panels and missing data

Many real-world panel datasets are *unbalanced*: some units are observed for more time periods than others, and even within a unit the series may contain gaps. Unbalanced panels are not a problem in themselves, but they become important whenever missingness is systematic. If observations are missing in ways that are related to the outcome or to the explanatory variables, estimates may be biased, and the interpretation of “within-unit change” becomes less straightforward.

Missing data are particularly consequential in dynamic specifications. When a model includes lagged variables, each missing observation can mechanically remove additional time periods from the estimation sample. For example, if we include $y_{i,t-1}$, then an observation for unit i at time t can only be used if $y_{i,t-1}$ is observed as well. Gaps *within* a unit’s time series are therefore more damaging than missing observations at the beginning or end of the series. In practice, it is always a good idea to inspect how the sample changes as soon as lags are introduced, and to be explicit about the resulting estimation sample.

8.3.4 Further readings

It is helpful to think of two broad approaches to time-series data. This text is written within a tradition in which regression models are kept as simple as possible, and dynamics are modeled sparingly unless theory provides a clear reason to include them. The emphasis in this tradition is usually on understanding a particular causal relationship. Beck and Katz (2011) is a useful guide for readers working in this vein. The other approach is more heavily influenced by statistics and by a model-building tradition in which one starts with a highly flexible specification and then uses statistical tests to determine which regressors to retain. In that tradition, prediction and model construction as such are at least as important as estimating the effect of a single variable. Most introductory texts on time-series analysis – which are typically written within this tradition – cover broadly similar material.

References

- Angrist, Joshua D and Jörn-Steffen Pischke (2014). *Mastering 'metrics: The path from cause to effect*. Princeton University Press.
- Baron, Reuben M and David A Kenny (1986). "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations." *Journal of personality and social psychology* 51.6, 1173.
- Beck, Nathaniel and Jonathan N Katz (2011). "Modeling dynamics in time-series–cross-section political economy data". *Annual Review of Political Science* 14, 331–352.
- Brambor, Thomas, William Roberts Clark, and Matt Golder (2006). "Understanding interaction models: Improving empirical analyses". *Political analysis*, 63–82.
- Cameron, A Colin and Douglas L Miller (2015). "A practitioner's guide to cluster-robust inference". *Journal of human resources* 50.2, 317–372.
- Labovitz, Sanford (1967). "Some observations on measurement and statistics". *Social Forces* 46.2, 151–160.
- Norman, Geoff (2010). "Likert scales, levels of measurement and the "laws" of statistics". *Advances in health sciences education* 15.5, 625–632.
- Pampel, Fred C (2020). *Logistic regression: A primer*. Vol. 132. SAGE Publications, Incorporated.
- Österman, Marcus and Olle Folke (2024). *Getting started with quantitative analysis: descriptive statistics, statistical inference and regression analysis*.