9 Explanatory and Omitted Variables

9.1 INTRODUCTION

No other model misspecification has attracted more attention than the choice of explanatory variables. Beginning with the first researcher who added a control variable to her baseline model and reported both models, scholars have long since argued that the true set of explanatory variables remains unknown and that it is likely that relevant regressors are omitted. Failure to include all relevant confounding variables, i.e. variables that have a causal effect on the dependent variable, results in a biased estimate for the variable of interest if covariance between the vector of omitted variables and the dependent variable is correlated with the variable of interest. Contrary to randomized controlled experiments, which can render the influence of confounding variables (though not of those that condition the treatment effect) irrelevant by increasing sample size to infinity, in observational data with no control over treatment status omitted variable bias does not disappear as sample size grows to infinity.

The potential exclusion of relevant variables has received notably more attention than the potential inclusion of irrelevant variables. However, the erroneous inclusion of irrelevant variables threatens the validity of inferences just as does the omission of variables (Clarke 2005). At the very least, including irrelevant variables will decrease the efficiency of estimations. Some cling to the belief that inefficiency "only" implies inflated standard errors and can therefore be neglected unless sample size is small or the data exhibits little variation. This view ignores the fact that researchers obtain a single point estimate per model. Erroneously included variables increase the sampling variance of the estimates, which implies that the expected deviation of an estimate from the truth increases. In this sense, inefficiency has the same effect as bias. Reducing efficiency means increasing the influence of noise on estimation outcomes. Inefficiency – just like bias – increases the probability of invalid inferences (King, Keohane, and Verba 1994;

Plümper and Troeger 2007, 2011). Accordingly, researchers should worry about erroneously included variables.

In reality, including all relevant confounding variables, but excluding all irrelevant variables, is not merely easier said than done: it is outright impossible with observational data. With limited information, the optimal number of variables to include in an estimation model may fall short of the number of variables contained in the true data-generating process. To make matters worse, even if we ignored the fact that, given limited information, models have to trade off simplicity versus generality, the inclusion of additional variables does not necessarily reduce bias even if the additional variable is part of the data-generating process. It is not difficult to construct a simple data-generating process with more than two omitted variables that results in an increase in bias if we add one omitted variable and a reduction in bias when both omitted variables are added.

We show how robustness tests help researchers uncertain about the choice of regressors by exploring the stability of estimated effects toward plausible changes to the set of explanatory variables. We start by discussing observed control variables, regarding which researchers are uncertain whether to include or exclude them. Both the exclusion of potential confounders and the inclusion of irrelevant variables threaten the validity of inferences on the effect of the variable of interest. We argue that estimation models will never include all the right and exclude all the wrong variables. We suggest that baseline models should contain only control variables known or suspected to exert a strong effect on the dependent variable, accompanied by robustness test models addressing uncertainty about the set of explanatory variables.

We then move to unknown and unobserved omitted variables potentially confounding the effect of the main variable of interest. We argue that the standard solution to "unobserved heterogeneity" in the form of differencing the data or unit fixed effects does not reduce bias if model misspecifications other than the omission of time-invariant unobservables represent a more important inferential threat. We suggest a number of tests that are less costly than differencing or unit fixed effects and at the same time are more flexible since they can also deal with time-varying unobserved confounders.

9.2 INCLUSION AND EXCLUSION OF CONTROL VARIABLES

Theories are typically focused on identifying a single or at most a few causal mechanisms and cannot be expected to provide a full account of the data-generating process. Theories typically aim at simplifying complex relationships instead of seeking to provide a full account of the determinants of a certain phenomenon – and if they do so, they fail. In other words, theories are not intended to guide the selection of explanatory variables. As a consequence, empirical models continue to be chosen in a haphazard way. As Leamer (1983: 34) has observed, the "standard set" of control variables tends to be arbitrary since it is "often based on whatever list the first researcher happened to select." Other variable selection algorithms may even be worse: what gave regression analysis of observational data a bad name is the possibility of selecting variables based on the desired result. Researchers typically provide only limited justification of why they include and why they exclude a variable that other scholars have included. Many right-hand side variables seem to have been selected based on common sense, tradition, or – perhaps – desired results. If common sense were a good scientific adviser, social scientists would be paid considerably less. Tradition – or path-dependence – seems to be a reasonable strategy, but it perpetuates model misspecification and severely reduces competition between models, which likely hampers scientific progress.

The fear of omitted variable bias induces some to include a long list of *potential* determinants as control variables – sometimes called the "kitchen sink" or "garbage can" approach. Some of these variables will be irrelevant and cause inefficiency. Others capture the same or a very similar causal factor, which creates bias and reduces the efficiency of the estimate. For example, political scientists know that the choice of different political institutions is correlated. Therefore, researchers need to be careful not to include institutional variables that are not independent of each other. One of the more blatant examples would be including both the electoral system and the number of parties in government or political fragmentation in the estimation model. Even if the explanatory variables are sufficiently independent of each other, they can still be highly correlated with each other in small to medium-sized samples containing limited information, which will decrease the efficiency of estimations and may result in biased estimates.

Political methodologist Christopher Achen's (in)famous rule of three, which in a nutshell denounces regression analyses with more than three regressors as "meaningless" (Achen 2002: 446), at least when no formal model structures the investigation, stems from these concerns. Achen seems to argue that researchers should construct empirical models that closely resemble the theoretical model – and ignore the data-generating process. However, this in turn downplays the need to control for all potential confounders in regression analysis of observational studies. Including "too many" control variables and including "too few" variables can equally threaten causal inferences (Clarke 2005). Naturally, the same goes for wrongly included explanatory variables.

Control variables bridge the gap between the theoretical model and the true data-generating process. Control variables, in other words, ought to move empirical models closer to the data-generating process and away from the stylized causal effects of theories. Models will nevertheless never match the complexity of the true data-generating process.[1] With limited information to analyze, researchers should not try to exactly model the true data-generating process. Instead, a good empirical model balances omitted variable bias and inefficiency. If researchers are interested in the effect of one or a few variables of interest, controls only need to be included if they are correlated with the variables of interest. The omission of a variable that is not correlated with the variable of interest does not bias the estimate of the effect of the variable of interest – regardless of how strongly it influences the dependent variable. This logic makes clear why two empirical models that explain the same phenomenon should be specified somewhat differently when the variable of interest differs. It also clarifies why researchers have to be reluctant to interpret the effect of control variables: a model that isolates the variable of interest from the influence of confounders does not need to isolate controls from the influence of confounders. However, in non-linear models some bias is inevitable even when omitted determinants of the outcome are uncorrelated with any of the right-hand side variables (Wooldridge 2010: 584f.).

Rather than trying to build all complications into a single model, we suggest that researchers conduct robustness tests to check whether estimates from a relatively simple baseline model are robust to systematically added complications. Following our suggestion, the baseline model would only contain those control variables known or suspected to exert the strongest effect on the dependent variable.

Explanatory variables tests are almost as old as regression analysis: adding or removing explanatory variables checks whether estimates are robust toward changing the covariance structure of the model. Adding or removing variables changes both the efficiency of an estimate and the part of the variance of an explanatory variable that is not correlated with the variance of another explanatory variable included in the model.

As always, any model known to be misspecified cannot function as a robustness test. A consequence of this rule is that researchers must take care not to additionally include variables through which other

1 As Hsiao (2003: 8) explains: "In explaining individual behavior, one may extend the list of factors ad infinitum. It is neither feasible nor desirable to include all factors affecting the outcome of all individuals in a model specification, since the purpose of modelling is not to mimic the reality but to capture the essential forces affecting the outcome. It is typical to leave out those factors that are believed to have insignificant impacts or are peculiar to certain individuals."

explanatory variables exert their impact on the dependent variable (King, Keohane, and Verba 1994: 173). This is not a straightforward rule given that in the social sciences many phenomena co-determine each other or – to exaggerate – everything has an impact on everything else. For example, the level of economic development will partly determine the political regime type, and vice versa, but both economic development and political regime type can be valid determinants of famine mortality (Plümper and Neumayer 2009).

Testing robustness toward dropping some of the regressors contained in the baseline model can provide another robustness test but requires even more thought than adding further control variables. Given that in the social sciences many variables are not entirely independent of each other, the effect of one variable has to be interpreted as conditional on the set of other right-hand side variables included in the model. For example, a civil war increases mortality rates directly and indirectly through increasing food scarcity. Dropping a variable measuring food scarcity from the baseline model can be justified as a robustness test, but researchers must keep in mind that the effect of civil war tested in the baseline and robustness test model differs across the two models. In one model the effect is conditional on food scarcity, in the other model any direct effect of food scarcity is assumed to be absent. This raises the question whether food scarcity should ever have been included in the baseline model if researchers are interested in the total (direct and indirect) effect of civil war on mortality since the estimation will fail to account for civil war partly determining food scarcity, thereby partly affecting mortality through its effect on food scarcity. There is no easy answer to this type of question since social scientists are necessarily uncertain about the correct set of regressors to be included in an estimation model.

9.3 UNKNOWN AND UNOBSERVED OMITTED VARIABLES

The standard econometric argument for the existence of "unobserved heterogeneity" is that some factors influencing the outcome cannot be observed. Unobserved heterogeneity, by definition, is unobserved and cannot be directly captured by control variables. Known unobserved factors can be indirectly and approximately captured by proxy variables. Yet, if a factor cannot be observed, the quality of a proxy cannot be known. Even more challengingly, omitted variables need not be "known unobservables" but can be "unknown observables" and "unknown unobservables." These variables are omitted because scholars do not even know that they are part of the data-generating process.

Econometric theory has developed a simple "solution" to the problem of omitted variables in panel or cross-sectional time-series data: assume that all omitted variables are time-invariant and employ unit fixed-effects or first-differences estimation, which throw away the between-variation of all included variables to isolate their parameter estimate from the bias resulting from omitted time-invariant variables. This provides no alternative to robustness testing. The solution only works if all omitted unobserved variables are indeed time-invariant and not just assumed to be time-invariant. In other words: the apparent solution determines the description of the problem, rather than the problem determining the solution. An undesirable side-effect has been the relative neglect of omitted unobserved *time-varying* variables.

Proponents of fixed-effects or first-differences estimation could argue that the between-variation of omitted unobserved variables tends to be larger than the omitted over-time variation within units. Even if this were true, differencing or unit fixed effects would reduce overall bias only if the existence of omitted time-invariant variables dominated all other model misspecifications. The two techniques are likely to perform worse than a model that ignores the bias from omitted time-invariant variables if other model misspecifications dominate (Plümper and Troeger 2016), most importantly dynamic misspecification. This is not ignorable since dynamic model misspecifications are common (De Boef and Keele 2008). For example, consider a trended dependent variable, trended independent variables, and an omitted trended variable. Differencing and unit fixed-effects estimates will be spurious, because the within-variation of interest is trended and will be correlated with the omitted trended variable. As another example, measurement error may become exacerbated if the between-variation is dropped under the plausible assumption that the true values are highly correlated over time but measurement error is randomly distributed over time in each unit. If so, moving from “levels” to “changes in levels” intensifies measurement error by lowering the signal-to-noise ratio (Bound et al. 2001: 3714).

Given that these techniques mitigate one potential model misspecification at the expense of exacerbating the impact of other potential model misspecifications, the widespread use of fixed-effects models appears problematic and unwarranted. Between-variation is information. Differencing and fixed-effects estimation eliminate valuable information. In fact, both techniques eliminate more information – all the between-variation – than they would do in an optimal world, in which they would merely eliminate the variance of the regressors correlated with the omitted time-invariant variables. This leads to a loss of efficiency, which can be substantial if the within-variation is low and the between-variation high. In addition, both techniques implicitly change the hypotheses tested in subtle ways. For example, the hypothesis that individuals with higher income have a higher propensity to buy certain goods differs from the hypothesis that an increase in income results in a higher propensity to buy certain goods. A theory that makes predictions about the effect of x on the between-variation in y cannot be tested.

Let us not be misunderstood: we are not arguing that differencing or unit fixed effects should be avoided under all circumstances. There will be conditions where they are warranted to make meaningful inferences about short-term adjustments to changes, and we have used them in our own research. For example, Plümper and Neumayer (2013) – aided by substantial within-variation in all variables – take out all level effects to identify the effect of co-payment schemes on changes in mortality rates. Likewise, Gibbons, Neumayer, and Perkins (2015) include year fixed effects and fixed effects for all combinations of subjects, universities, and entry tariffs, without which it would be impossible to identify an effect of student satisfaction on subject-specific student applications to universities. Rather than condemning these techniques per se, we are criticizing the “fixed effects by default” attitude that persists in many research areas of the social sciences – in some more so than in others.

9.4 ROBUSTNESS TESTS FOR POTENTIALLY OMITTED VARIABLES

Depending on the specification of the baseline model, robustness tests for unobserved variables assess whether the baseline model estimates are robust toward either eliminating a part of the variation that might be correlated with the unobserved variables or toward accounting for unobserved variables in a plausible alternative way. We start our discussion with two robustness tests that remain squarely within traditional thinking and that assume, entirely unrealistically, that unobserved variables are strictly time-invariant.

The first test is of the robustness limit type: the *between-variation test* increases the between-variation that is dropped from all variables (including the dependent variable). Researchers start with a pooled-ordinary least squares (OLS) model (or its equivalent), and then eliminate the between-variation in 10-percent steps, where the transformation of a variable x is $x_{it} - \sqrt{\lambda_s}\bar{x}_i$ (and similar for all other variables including the dependent variable), where λ_s $(0.0 \leq \sqrt{\lambda_s} \leq 1.0)$ denotes the degree to which between-variation is eliminated. Thus, in order to eliminate the between-variation in 10-percent steps, we need to successively increase λ_s from 0 to 1, which means increasing $\sqrt{\lambda_s}$ from 0 to 0.316, 0.447, 0.548, 0.632, 0.707, 0.755, 0.837, 0.894, 0.949 and, finally, 1, at which point we have reached the unit fixed-effects specification.

Table 9.1: Between-variation Test

Degree of de-meaning ($\sqrt{\lambda_s}$)	**Coefficient (s.e.)**	**ρ short term**	**ρ long term**
0.0 (0.000)	0.00960*** (0.00219)	baseline	baseline
0.1 (0.316)	0.0108*** (0.00242)	0.89	0.87
0.2 (0.447)	0.0112*** (0.00259)	0.84	0.80
0.3 (0.548)	0.0113*** (0.00275)	0.81	0.76
0.4 (0.632)	0.0110*** (0.00293)	0.81	0.74
0.5 (0.707)	0.0103*** (0.00313)	0.82	0.74
0.6 (0.775)	0.00904** (0.00339)	0.79	0.74
0.7 (0.837)	0.00694* (0.00372)	0.64	0.68
0.8 (0.894)	0.00406 (0.00406)	0.37	0.48
0.9 (0.949)	0.00115 (0.00438)	0.17	0.26
1.0 (1.000)	−0.000138 (0.00455)	0.11	0.19

Note: Coefficients show the effect of pre-tax income inequality on the Gini coefficient in longevity. All other independent variables of baseline model included but effects not reported.

Statistically significant at *0.1, ** at 0.05, *** at 0.01 level.

In Neumayer and Plümper (2016b) we analyze the effect of pre-tax income inequality as well as income redistribution (the absolute difference between pre-tax and post-tax income inequality) on inequality in longevity, that is, the inequality in the number of years individuals in a country live. We analyze a pooled cross-sectional time-series sample of up to 28 countries over the period 1974 to 2011. The OLS baseline model does not include unit fixed effects. We have two main variables of interest, but, for the purpose of illustrating this robustness test, let us focus on pre-tax income inequality, for which table 9.1 shows estimated coefficients (the robustness analysis is almost identical for the other variable of main interest). Since the baseline model includes the lagged dependent variable, we analyze the degrees of robustness ρ for both short-term and long-term effects. The top of the table starts with the pooled-OLS baseline model (zero degree

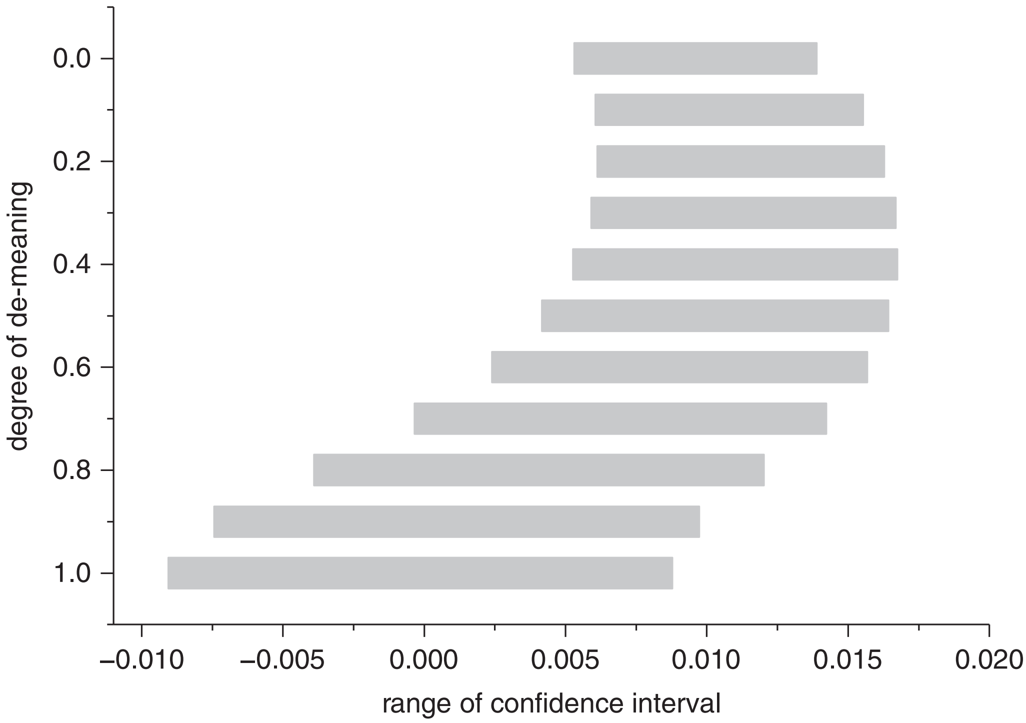

Figure 9.1: Visualization of the Between-variation Test

of de-meaning), increasingly eliminating the between-variation in the data by successively de-meaning all variables in 10-percent steps, arriving eventually at the unit fixed-effects model as variables are completely de-meaned. The estimated effect remains stable until at least 70 percent of between-variation is dropped. Beyond this point, the estimated effect declines and even switches signs with full de-meaning in the unit fixed-effects model. The degrees of robustness decline to 0.11 in the short term and 0.19 in the long term in this model.

In figure 9.1, we display the same information as in table 9.1 to demonstrate the stability of the estimated coefficients. It demonstrates with greater clarity that the estimates remain very stable unless we remove at least two-thirds of the between-variation.

A second test does not drop the between-variation from the estimation, but includes group fixed effects based on a grouping of units. This *group-wise fixed-effects test* eliminates the between-group variation but leaves the between-variation within groups intact. Ideally, the grouping of units is theoretically informed.[2] Where this is not possible, group membership can be simply based on locational or relational information. In comparative cross-country research, for example, a simple proxy for group heterogeneity

2 See for example Cederman, Buhaug, and Rød (2009: 519).

Table 9.2: Groupwise Fixed-effects Test (replacement type)

	m1: baseline	m2: group fixed effects	ρ
Longevity inequality (t–1)	0.867***	0.853***	0.872
	(0.0305)	(0.0367)	
Life expectancy	–0.000306***	–0.000358**	0.765
	(0.000103)	(0.000162)	
GDP per capita (ln)	0.000234	–0.000253	0.790
	(0.000392)	(0.000349)	
Health expenditure relative to GDP (ln)	0.00124**	0.000548	0.673
	(0.000468)	(0.000503)	
Alcohol consumption per capita (ln)	6.64e–05	0.000113	0.859
	(0.000120)	(0.000153)	
Lung cancer mortality rate	0.000916	0.000983	0.895
	(0.000664)	(0.000802)	
External cause mortality rate	0.00273***	0.00282***	0.923
	(0.000556)	(0.000611)	
Pre-tax income inequality	0.00960***	0.00603***	0.637
	(0.00219)	(0.00207)	
Income redistribution	–0.00940***	–0.00645**	0.782
	(0.00249)	(0.00248)	
Observations	476	476	
Number of countries	28	28	

Note: Dependent variable is the Gini coefficient of longevity. OLS estimation. Year-specific fixed effects included. Standard errors adjusted for clustering on countries in parentheses.

Statistically significant at * 0.1, ** at 0.05, *** at 0.01 level.

can be based on the World Bank's definition of macro-regions or on Huntington's (1996) definition of civilizations.

For our illustrative example, we rely on Böhm et al.'s (2013) classification, which groups countries into types of healthcare systems according to the private, societal or state organization of the regulation, financing, and provision of health care, giving us seven group dummies in total. Model m1 in table 9.2 presents results for the baseline model and the robustness test model m2 with these group fixed effects included. We only show estimated degrees of robustness for short-term effects. The estimated coefficients of our two main explanatory variables decrease by about 30 percent in size when the group fixed effects are included. Nevertheless, the estimated degrees of robustness are high.

As an alternative, a robustness limit variant of this test stepwise increases the number of exclusive groups and thus the number of group fixed effects. The more exclusive groups are formed, the more the model resembles the unit fixed-effects model, with the latter being reached as the number of exclusive groups reaches N–1. If, as will often be the case, the categorization of units into groups is uncertain, researchers can base group membership on a cluster analysis of the *between*-variation in the substantive regressors, thereby ensuring that all units exclusively cluster in one group. Researchers can either vary the specification of the number of groups into which units are clustered or they can vary the specification of the similarity value used in the cluster analysis.

In our illustrative example, we employ a Ward's linkage cluster analysis (1963) to sort countries into increasingly larger number of groups. We decrease the dissimilarity measure starting from 51 in steps of 2 until we reach the unit fixed-effects model at dissimilarity measure 1. Table 9.3 shows results whenever an increase in the number of groups is the consequence of the decrease in dissimilarity measure applied to the cluster analysis. Results are very robust up to and including a dissimilarity value of 5, which sorts the 28 countries into 21 groups. Lower dissimilarity values and consequently more groups result in low degrees of robustness.

As we have argued above, the assumption that all omitted variables are time-invariant is usually not supported in real data. Researchers should therefore explore the robustness of their baseline model to changing assumptions about omitted unobserved variables that can vary over time. We suggest a robustness limit type test which investigates to what extent the omitted variable needs to be correlated with the variable of interest and the dependent variable to render the estimated effect for the variable of interest non-robust. A similar robustness test has been proposed by Frank (2000), who suggests generating a random placeholder for an omitted confounder and stepwise increasing the correlation between the placeholder and the dependent variable.

With panel data, the *correlated artificial variable test* can be specified in three ways. The first variant just uses the time-invariant components of x and y and creates a placebo variable z which has a specified covariance structure with the between-variation of x and y, $\overline{x}_i$ and $\overline{y}_i$. The second variant uses only the overall correlation with the within-variation of x and y, $x_{it} - \overline{x}_i$ and $y_{it} - \overline{y}_i$ (this is a test about omitted common trends), and the third variant uses the correlation with both observed variables x and y. Of course, in cross-sectional data only the first variant and in time-series data only the second variant is possible.

A further robustness test that works for both time-invariant and time-varying unobserved variables can be based on what is known as a spatial-

Table 9.3: Groupwise Fixed-effects Test (robustness limit type)

Maximum degree of dissimilarity (no. of groups)	Coefficient (s.e.)	ρ short term	ρ long term
51 (4)	0.00960*** (0.00219)	0.96	0.96
41 (5)	0.00914*** (0.00220)	0.94	0.95
39 (6)	0.00920*** (0.00216)	0.95	0.95
35 (7)	0.00916*** (0.00216)	0.95	0.95
25 (8)	0.00812*** (0.00196)	0.92	0.94
21 (11)	0.00870*** (0.00221)	0.93	0.94
15 (13)	0.0103*** (0.00326)	0.80	0.83
13 (15)	0.00998*** (0.00352)	0.77	0.81
9 (18)	0.0105** (0.00428)	0.67	0.70
7 (19)	0.00776* (0.00386)	0.68	0.70
5 (21)	0.00709* (0.00411)	0.62	0.66
3 (24)	0.00216 (0.00363)	0.19	0.29
1 (28)	−0.000138 (0.00470)	0.11	0.19

Note: Coefficients show the effect of pre-tax income inequality on the Gini coefficient in longevity. All other independent variables of baseline model included but effects not reported.

Statistically significant at * 0.1, ** at 0.05, *** at 0.01 level.

error model. The *spatial-error test* exploits (and depends on) the fact that in many research designs omitted variables will be spatially correlated. To estimate a spatial-error model, analysts first predict the residuals by estimating the baseline model without spatial-error component. They can then create a spatial-error component by weighting the residuals of other units with a measure of "closeness" between the unit under observation and these other units as weights (see chapter 14 for more details). This could be a measure of geographical contiguity or inverse distance, but it could also be some other measure of connectivity or relatedness. If the assumption that

Table 9.4: Spatial-error Test

	m1: baseline	m3: spatial-error variable included	ρ
Longevity inequality (t–1)	0.867***	0.869***	0.946
	(0.0305)	(0.0310)	
Life expectancy	–0.000306***	–0.000296***	0.947
	(0.000103)	(0.000104)	
GDP per capita (ln)	0.000234	0.000201	0.949
	(0.000392)	(0.000392)	
Health expenditure relative to GDP (ln)	0.00124**	0.00125**	0.952
	(0.000468)	(0.000464)	
Alcohol consumption per capita (ln)	6.64e–05	6.53e–05	0.952
	(0.000120)	(0.000119)	
Lung cancer mortality rate	0.000916	0.000922	0.953
	(0.000664)	(0.000657)	
External cause mortality rate	0.00273***	0.00270***	0.947
	(0.000556)	(0.000562)	
Pre-tax income inequality	0.00960***	0.00956***	0.955
	(0.00219)	(0.00214)	
Income redistribution	–0.00940***	–0.00920***	0.950
	(0.00249)	(0.00248)	
Spatial-error variable		0.214**	
		(0.104)	
Observations	476	476	
Number of countries	28	28	

Note: Dependent variable is the Gini coefficient of longevity. OLS estimation. Year-specific fixed effects included. Standard errors adjusted for clustering on countries in parentheses.

Statistically significant at * 0.1, ** at 0.05, *** at 0.01 level.

omitted variables spatially correlate holds, the spatial-error control variable included in the robustness test model reduces the influence of omitted variables.

Model m3 includes a spatial-error variable based on geographical contiguity as the connectivity variable for the spatial-weights matrix (see table 9.4). The coefficient of the spatial-error variable is positive, as one would expect if residuals are spatially clustered among contiguous countries. The other estimates are very robust, which suggests that the omission of a spatial-error variable, in this case at least, is essentially inconsequential.

In other research projects, the test could affect the estimated degrees of robustness more strongly.

9.5 CONCLUSION

The continuing development of apparent econometric solutions to the problem of variable choice demonstrates that a silver bullet method for selecting all the right and none of the wrong explanatory variables does not exist. Empirical research will therefore continue to omit the right and include the wrong variables. Like any other method, robustness tests will not lead to the correct set of regressors. But that is neither their purpose nor our intention. The main idea of robustness tests is to analyze whether unknown and known but unobservable factors exist which could render estimates non-robust and potentially invalidate inferences.

Due to their strong emphasis on unknown potential confounders, structured permutation and robustness limit type tests are best suited for analyzing the effects of uncertainty about the set of explanatory variables. In other words, many of the robustness tests discussed here stand in the tradition of Leamer's sensitivity testing and Rosenbaum's bounds test. Yet, researchers need to avoid drawing inferences from models which are known to be misspecified. This not only requires care in the selection of substantive variables; it also means that techniques that eliminate a specific kind of variation from the data, like the unit fixed-effects model, should not be employed as the default option. Such models can make sense either as baseline or robustness test models in some limited number of research projects but represent models known to be misspecified in many others.