Robustness Checks and Robustness Tests in Applied Economics

Halbert White Xun Lu
Department of Economics
University of California, San Diego

June 18, 2010

Abstract

A common exercise in empirical studies is a "robustness check," where the researcher examines how certain "core" regression coefficient estimates behave when the regression specification is modified by adding or removing regressors. If the coefficients are plausible and robust, this is commonly interpreted as evidence of structural validity. Here, we study when and how one can infer structural validity from coefficient robustness and plausibility. As we show, there are numerous pitfalls, as commonly implemented robustness checks give neither necessary nor sufficient evidence for structural validity. Indeed, if not conducted properly, robustness checks can be completely uninformative or entirely misleading. We discuss how critical and non-critical core variables can be properly specified and how non-core variables for the comparison regression can be chosen to ensure that robustness checks are indeed structurally informative. We provide a straightforward new Hausman (1978)-type test of robustness for the critical core coefficients, additional diagnostics that can help explain why robustness test rejection occurs, and a new estimator, the Feasible Optimally combined GLS (FOGLeSs) estimator, that makes relatively efficient use of the robustness check regressions. A new procedure for Matlab, *testrob*, embodies these methods.

1 Introduction

A now common exercise in empirical studies is a "robustness check," where the researcher examines how certain "core" regression coefficient estimates behave when the regression specification is modified in some way, typically by adding or removing regressors. Leamer (1983) influentially advocated investigations of this sort, arguing that "fragility" of regression coefficient estimates is indicative of specification error, and that sensitivity analyses (i.e., robustness checks) should be routinely conducted to help diagnose misspecification.

Such exercises are now so popular that standard econometric software has modules designed to perform robustness checks automatically; for example, one can use the STATA commands *rcheck* or *checkrob*. A finding that the coefficients don't change much is taken to be evidence that these coefficients are "robust."[1] If the signs and magnitudes of the estimated regression coefficients are also plausible, this is commonly taken as evidence that the estimated regression coefficients can be reliably interpreted as the true causal effects of the associated regressors, with all that this may imply for policy analysis and economic insight.

Examples are pervasive, appearing in almost every area of applied econometrics. For example, of the 98 papers published in *The American Economic Review* during 2009, 76 involve some data analysis. Of these, 23 perform a robustness check along the lines just described, using a variety of estimators suitable to the data, such as ordinary least squares, logit, instrumental variables, or panel methods (Adams, Einav, and Levin, 2009, Alfaro and Charlton, 2009, Angelucci and De Giorgi, 2009, Angrist and Lavy, 2009, Ashraf, 2009, Boivin, Giannoni, and Mihov, 2009, Cai, Chen, and Fang, 2009, Chen and Li 2009, Chetty, Looney, and Kroft, 2009, Dobkin and Nicosia 2009, Forbes and Lederman, 2009, Hargreaves Heap and Zizzo, 2009, Hendel, Nevo, and Ortalo-Magné, 2009, Lavy, 2009, Leaver, 2009, Makowsky and Stratmann, 2009, Mas and Moretti, 2009, Matsusaka, 2009, Miller, 2009, Oberholzer-Gee and Waldfogel, 2009, Sialm, 2009, Spilimbergo, 2009, Urquiola and Verhoogen, 2009).

But when and how can evidence of coefficient robustness and plausibility support the inference of structural validity? Our purpose here is to address this question in substantive detail. For maximum clarity and practical relevance, we consider a simple linear regression context. We show that even in this familiar framework, there are many opportunities to go astray. Some of these pitfalls can be avoided by paying attention to properties of regression analysis that should be well known. But if they are, they are often ignored in practice. This neglect can have serious adverse consequences when checking robustness, so it is useful to emphasize these. Other pitfalls have not previously been recognized; these bear critically on properly interpreting empirical results, especially with regard to robustness checks.

It should be well known, but it is too often forgotten, that is *not* necessary that all estimated coefficients make economic sense. See, e.g., Stock and Watson (2007, pp.478-479), who stress this point. For easy reference and because it forms part of the foundation for the analysis to follow, we spell out why requiring coefficient signs and magnitudes to make economic sense is necessary only for a certain subset of the regressors with particular relevance for robustness checks: the *critical* core variables are precisely those whose effects are of primary interest and whose coefficients should be plausible. As should also be well known, plausibility of regression coefficients is not sufficient to permit attribution of causal effects, regardless of robustness. For

[1] This usage of "robust" should not be confused with the concept of robustness in the statistics literature, which refers to the insensitivity of an estimator to adding or removing sample observations, typically extreme in some way.

easy reference, we also briefly spell this out here.

Robustness *is* necessary for valid causal inference, in that the coefficients of the critical core variables should be insensitive to adding or dropping variables, under appropriate conditions. But several pertinent questions have not so far been adequately addressed. Specifically, which variables besides the critical core variables should also be core variables? Which non-core variables should one add or remove? Should the latter satisfy certain conditions, or is it valid to compare regressions dropping or including any non-core variables? We show that the choice of core variables and the non-core variables included or excluded must obey specific restrictions dictated by the underlying structural equations. Ignoring these restrictions can lead to fundamentally flawed economic inference. We propose straightforward analysis and statistical procedures for distinguishing core and non-core variables.

Most importantly, why perform just a "check," when it should be easy to conduct a true test of the hypothesis that the relevant coefficients don't change? At the least, this would permit the researcher to determine objectively whether the coefficients had changed too much. The specification testing principles articulated in Hausman's (1978) landmark work apply directly. Accordingly, we give a straightforward robustness *test* that turns informal robustness checks into true Hausman (1978)-type structural specification tests.

Suppose we find that the critical core coefficients are not robust. Does a robustness check or test provide insight into the reason for the failure? Are there additional steps that could be taken to gain further insight? We show how and why robustness can fail, and we discuss methods that can be used to gain deeper insight.

Or, suppose we do find that the critical core coefficients are robust and suppose we have other good reasons to believe that these validly measure economic effects of interest. Then there are multiple consistent estimators, and we would like to obtain the most precise of these. But which set of regressors should we use? A larger set or a smaller set? Or does it matter? We show that the choice of efficient regressors for causal inference is a deeper and context-dependent question that presents interesting opportunities for further investigation. We also propose relatively efficient feasible generalized least squares (FGLS) methods and a new estimator, the Feasible Optimally combined GLS (FOGLeSs) estimator, that can provide superior inference in practice.

In addressing each of these questions, we provide succinct answers and, where relevant, easily implemented procedures. Apart from some (important) details specific to robustness checking, our robustness test is standard; similarly, our FGLS and FOGLeSs estimators are straightforward. The contribution of this paper is therefore not some completely novel method. Rather, this paper is about the proper use of now common procedures that seem to be widely misapplied. Proper application of these, together with the new practical methods embodied here in our Hausman test-based procedure *testrob*, could considerably strengthen the validity and reliability of structural inference in economics. To make these methods easy to apply, Matlab code for *testrob* is freely available at `www.econ.ucsd.edu/~xunlu/code`.

2 The Structural Data Generating Process

Economic theory justifies claims that an outcome or response of interest, Y, is structurally generated as

$$Y = r(D, Z, U),$$

where r is the unknown structural function, D represents observed causes of interest, and Z and U are other drivers of Y, where Z is observed and U is not. In particular, U represents not just "shocks," but all factors driving Y that are too costly or too difficult to observe precisely.

Our interest attaches to the effects of D on Y. For example, let Y be wage, let D be schooling, let Z be experience, and let U be unobserved ability. Then we are interested in the effects of schooling on wage. Or let Y be GDP growth, let D be lagged GDP growth and lagged oil price changes, let Z be other observed drivers of GDP growth (e.g., monetary policy), and let U represent unobserved drivers of growth, including shocks. In this case, we are considering a component of a structural VAR system, and we are interested in the effects of lagged GDP growth and lagged oil prices on GDP growth. When interest truly attaches to the effects of *all* observed drivers of Y, we assign these to D and omit Z.

For maximum clarity and practical relevance, we assume here that r is linear, so

$$Y = D'\beta_o + Z'\alpha_o + U, \tag{1}$$

where β_o and α_o represent the effects of D and Z on Y respectively. Later, we discuss some consequences of nonlinearity. Because D is the cause of interest, we are interested primarily in β_o; we may have little or no interest in α_o. Note that U has now become a scalar; in this linear case, we can view U as representing a linear combination of unobserved drivers of Y, without essential loss of generality.

A standard assumption that identifies β_o is exogeneity: (D, Z) is independent of U, written $(D, Z) \perp U$, using the notation of Dawid (1979) (henceforth "D79"). In fact, this identifies both β_o and α_o. When we are mainly interested in β_o, we do not need such a strong assumption. A weaker condition identifying β_o is a conditional form of exogeneity:

$$D \perp U \mid X, \tag{2}$$

that is, D is independent of U given X, where $X = (Z', W')'$ is a vector of "control variables" or "covariates." This condition is actually much weaker than exogeneity, since it not only permits

D and U to be correlated, but it also permits X to be correlated with U. A classic example is when W is an IQ score proxying for unobserved ability, as in Griliches (1977). Below and in the appendix, we discuss in detail where the W's come from.

A still weaker condition identifying β_o is conditional mean independence, $E(U \mid D, X) = E(U \mid X)$; for example, Stock and Watson (2007) use this assumption in their outstanding textbook. Wooldridge (2002, pp. 607-608) has an excellent discussion comparing this to (2). An even weaker (in fact necessary) requirement is conditional non-correlation, $E([D - \zeta^* X]U) = 0$, where $\zeta^* \equiv E(DX)E(XX')^{-1}$ is the matrix of regression coefficients for the regression of D on X. Here, we work with (2) for three main reasons. First, this permits us to closely link our analysis to the relevant literature on treatment effect estimation, which has substantive implications for choosing W. Second, estimator efficiency is a key focus here, and (2) has straightforward and important useful implications for efficiency, discussed below. Third, this condition ensures valid structural inference in contexts extending well beyond linear regression (e.g., White and Chalak, 2009).

3 Making Economic Sense

The conditional exogeneity relation (2) is key to determining the *critical* core variables, i.e., those variables whose coefficients should make economic sense and be robust. To see this, we begin by considering the regression of Y on D and X, $E(Y \mid D, X)$. Defining the regression error $\varepsilon \equiv Y - E(Y \mid D, X)$, we have the regression equation

$$Y = E(Y \mid D, X) + \varepsilon.$$

This is purely a predictive relation, with no necessary economic content. Observe that U and ε are distinct objects. Whereas ε is simply a prediction error with no causal content for Y, U represents unobserved drivers of Y.

The underlying structure does create a relationship between U and ε, however. Using the structural relation and conditional exogeneity, we have

$$\begin{aligned} E(Y \mid D, X) &= E(D'\beta_o + Z'\alpha_o + U \mid D, X) = D'\beta_o + Z'\alpha_o + E(U \mid D, X) \\ &= D'\beta_o + Z'\alpha_o + E(U \mid X). \end{aligned}$$

The final equality holds because $D \perp U \mid X$ implies $E(U \mid D, X) = E(U \mid X)$. It follows that $\varepsilon = U - E(U \mid X)$. Thus, ε is a function of U and X, an important fact that we rely on later.

We can now easily see which regression coefficients should make economic sense and which should not. Suppose for simplicity that

$$E(U \mid X) = X'\delta^* = Z'\delta_z^* + W'\delta_w^*. \tag{3}$$

Then the regression equation has the form

$$\begin{aligned} Y &= D'\beta_o + Z'\alpha_o + X'\delta^* + \varepsilon \\ &= D'\beta_o + Z'(\alpha_o + \delta_z^*) + W'\delta_w^* + \varepsilon \\ &= D'\beta_o + X'\gamma^* + \varepsilon, \end{aligned}$$

say, where $\gamma^* \equiv ((\alpha_o + \delta_z^*)', \delta_w^{*\prime})'$. Because $E(\varepsilon \mid D, X) = 0$, the regression coefficients β_o and γ^* can be consistently estimated by ordinary or generalized least squares under standard conditions.

Because β_o represents the effects of the causes of interest, D, its signs and magnitudes clearly should make economic sense. But the coefficient on Z, the other observed drivers of Y, is $\gamma_z^* \equiv \alpha_o + \delta_z^*$. This is a mix of causal (α_o) and predictive (δ_z^*) coefficients. Even though Z drives Y, the presence of δ_z^* ensures there is no reason whatsoever that these coefficients should make economic sense. The same is true for the coefficients of W. These are purely predictive, so their signs and magnitudes have no economic content.

These facts should be well known. Indeed, Stock and Watson (2007, pp.478-479) give a very clear discussion of precisely this point in their undergraduate textbook. But this deserves emphasis here for two related reasons. First, researchers and reviewers of research submitted for publication often overlook this. Reviewers may take researchers to task for "puzzling" or "nonsense" coefficients that are not the primary focus of interest; and researchers, anticipating such reactions (correctly or not), often (mis)direct significant effort to specification searches that will preempt such attacks. This often includes robustness checks for all core coefficients.

This motivates the second reason for emphasizing this, directly relevant to our focus here: Because only D should have economically sensible coefficients, only D's coefficients should be subject to robustness checking or testing. Thus, we call D *critical* core variables. Researchers should not robustness check all core coefficients, only the critical core coefficients. Similarly, reviewer demands that all regression coefficients be robust and make economic sense are unnecessary and unrealistic. Indeed, this stifles valid research and provides incentives for researchers to "cook" their results until naively demanding reviewers will find them palatable.

It is also easy see here why regression coefficients "making sense" is not sufficient to permit attribution of causal effects: if either linearity (eq.(1) or eq.(3)) or conditional exogeneity (eq.(2)) fails, then the regression equation just gives the optimal linear prediction equation

$$Y = D'\beta^* + X'\gamma^* + \varepsilon,$$

where the optimal linear prediction coefficients are

$$\begin{pmatrix} \beta^* \\ \gamma^* \end{pmatrix} \equiv \begin{bmatrix} E(DD') & E(DX') \\ E(XD') & E(XX') \end{bmatrix}^{-1} \begin{pmatrix} E(DY) \\ E(XY) \end{pmatrix},$$

and $\varepsilon \equiv Y - D'\beta^* + X'\gamma^*$ is the optimal linear prediction error. These coefficients could be anything; some of them could even make economic "sense." But without the structural content of linearity and conditional exogeneity, this is, as MacBeth might say, a tale told by an idiot, full of sound and fury, signifying nothing but predictive optimality. Researchers and reviewers comforted by plausible regression coefficients but with little or no other evidence of correct structural specification may be easily led astray.

4 Robustness

To determine whether one has estimated effects of interest, β_o, or only predictive coefficients, β^*, one can check or test robustness by dropping or adding covariates. As we show in the next three sections, however, there are important restrictions on which variables one may include or exclude when examining robustness. This means that letting STATA loose with *checkrob*, for example, is potentially problematic, as this module estimates a set of regressions where the dependent variable is regressed on the core variables (included in all regressions) and *all possible combinations* of other (non-core) variables, without regard to these restrictions.

We begin by considering what happens with alternative choices of covariates, say X_1 and X_2. In particular, suppose we take $X_1 = (Z', W')'$ as above and form $X_2 = W$ by dropping Z. In the latter case, suppose conditional exogeneity holds with $D \perp U \mid W$. Now

$$E(Y \mid D, X_2) = E(D'\beta_o + Z'\alpha_o + U \mid D, X_2) = D'\beta_o + E(Z'\alpha_o \mid D, W) + E(U \mid D, W).$$

Parallel to the prior analysis, we have $E(U \mid D, W) = E(U \mid W)$, and we take $E(U \mid W) = W'\delta_w^*$. But we must also account for $E(Z'\alpha_o \mid D, W)$. This is a function of D and W, so suppose for simplicity that $E(Z' \mid D, W) = D'\zeta_d^* + W'\zeta_w^*$. This gives

$$\begin{aligned} E(Y \mid D, X_2) &= D'\beta_o + D'\zeta_d^* \, \alpha_o + W'\zeta_w^* \, \alpha_o + W'\delta_w^* \\ &= D'(\beta_o + \zeta_d^* \, \alpha_o) + W'(\zeta_w^* + \delta_w^* \, \alpha_o). \end{aligned}$$

Now the coefficient on D is not β_o but $(\beta_o + \zeta_d^* \, \alpha_o)$. If robustness held, these coefficients would be identical. But these coefficients generally differ.[2] A robustness check or test based on core variables D and comparing regressions including and excluding Z while including covariates W would therefore generally signal non-robustness of the D coefficients, despite the validity of the regression including Z and W for estimating β_o. This shows that robustness checks or tests generally should *not* drop observed drivers Z of Y. That is, the Z's should be core variables for

[2] No difference occurs when $\zeta_d^* \alpha_o = 0$. For this, it suffices that $\alpha_o = 0$ (Z does not drive Y) or that $\zeta_d^* = 0$ (e.g., $D \perp Z \mid W$). In what follows, we use the qualifier "generally" to implicitly recognize these exceptions or other similar special circumstances.

robustness checks or tests, in the sense that they generally should be included along with D in the regressions used for robustness checking.

On the other hand, even when Z is included as a core variable, dropping one or more elements of W could lead to failure of conditional exogeneity, again signaling non-robustness, despite the validity of the regression including Z and W for estimating β_o.

Robustness is nevertheless necessary for valid causal inference, provided we have (at least) two alternate choices of covariates, say X_1 and X_2, both of which contain Z and both of which ensure conditional exogeneity of D :

$$D \perp U \mid X_1 \qquad \text{and} \qquad D \perp U \mid X_2. \tag{4}$$

A similar situation has been considered in a related treatment effect context by Hahn (2004) and by White and Lu (2010b). Nevertheless, those papers focus on a different issue, estimator efficiency; we will return to this below.

The reasoning above now gives two regression equations of similar form:

$$\begin{aligned} Y &= D'\beta_o + X_1'\gamma_1^* + \varepsilon_1 \\ Y &= D'\beta_o + X_2'\gamma_2^* + \varepsilon_2, \end{aligned}$$

where $(\gamma_1^*, \varepsilon_1)$ and $(\gamma_2^*, \varepsilon_2)$ are defined analogously to their counterparts above. Because conditional exogeneity holds for both sets of covariates, we can estimate β_o consistently from both regressions. That is, whether we include D and X_1 in the regression or D and X_2, we will get similar critical core variable coefficient estimates, as both estimate the same thing, β_o. This is the essence of the robustness property, and we have just shown that it is a consequence of correct structural specification, hence necessary.

Notice that even though Z is included in both regressions, its coefficients generally differ between the two. In the first regression, the Z coefficient is $(\alpha_o + \delta_{1z}^*)$, say; it is $(\alpha_o + \delta_{2z}^*)$ in the second. This reflects the fact that Z is playing a predictive role for U in these regressions, and this role changes when different W's are involved. Non-robustness of the Z coefficients does not signal structural misspecification. For this reason, the Z's are *non-critical* core variables.

When the structural assumptions fail, we still have optimal prediction regressions, say,

$$\begin{aligned} Y &= D'\beta_1^* + X_1'\gamma_1^* + \varepsilon_1 \\ Y &= D'\beta_2^* + X_2'\gamma_2^* + \varepsilon_2. \end{aligned}$$

A robustness test gains power against structural misspecification from the fact that when the structural assumptions fail, we typically have that β_1^* and β_2^* differ. Note, however, that such tests are not guaranteed to have good power, as β_1^* and β_2^* can be similar or even equal in the presence of structural misspecification, as when $X_j = (Z', W_j')$ and $E(DW_j')$, $E(ZW_j')$ are close or equal to zero, $j = 1, 2$. We provide an example in Appendix I.

These properties are precisely those permitting application of a Hausman (1978)-type test. Below, we specify a Hausman-style robustness test based on differences in critical core coefficient estimates, such as $\hat{\beta}_{1n} - \hat{\beta}_{2n}$, where $\hat{\beta}_{1n}$ and $\hat{\beta}_{2n}$ are estimators of β_1^* and β_2^*, respectively. When this test rejects, this implies that any or all of the following maintained hypotheses fail: (i) $Y = D'\beta_o + X'\gamma^* + \varepsilon$ (regression linearity); (ii) $D \perp U \mid X_1$ (conditional exogeneity of D w.r.t. X_1); (iii) $D \perp U \mid X_2$ (conditional exogeneity of D w.r.t. X_2). Regression linearity can fail due to either structural nonlinearity (failure of $Y = D'\beta_o + Z'\alpha_o + U$) or predictive nonlinearity (failure of $E(U \mid X) = X'\delta^* = Z'\delta_z^* + W'\delta_w^*$) or both. Robustness tests are therefore non-specific; either exogeneity failures or nonlinearity may be responsible for rejection. In particular, if rejection is due to the failure of (iii) only, then consistent estimation of causal effects β_o is still possible, as $\beta_1^* = \beta_o$. The robustness test rejects because of a misleading choice of comparison covariates. A similar situation holds if rejection is due to the failure of (ii) only, as then $\beta_2^* = \beta_o$.

To avoid rejections due to a misleading choice of comparison covariates, it is helpful to gain a better understanding of where the covariates W come from. We take this up next.

5 Selecting Covariates

So far, we have seen that the covariates X should contain Z. They may also contain additional variables W, as $X \equiv (Z', W')'$. We now consider how these additional variables may arise. First, we consider how W may be chosen to ensure the validity of covariates X, that is, $D \perp U \mid X$. Then we consider how the core and non-core covariates potentially useful for examining robustness may be chosen.

5.1 Valid Covariates

If strict exogeneity holds, i.e., $(D, Z) \perp U$, then $D \perp U \mid Z$ (e.g., by lemma 4.3 of D79). In this case, $X = Z$ is a valid choice of covariates, and we need not include other variables W in the core covariates. Nevertheless, we still need valid non-core covariates W to perform a robustness check, so it is important to understand how these can arise. Further, when D or Z are *endogenous* (e.g., correlated with U), as in the classic example where schooling (D) may be correlated with ability (U), then W can play a crucial role by providing core covariates that ensure $D \perp U \mid (Z, W)$.

With $X = (Z', W')'$, the condition $D \perp U \mid X$ says that given X, D contains no predictively useful information for U. It also says that given X, U contains no predictively useful information for D. For the first criterion, we want X to be a good predictor of U, so that D has nothing more to contribute. For the second, we want X to be a good predictor of D, so that U has nothing more to contribute. Either or both of these properties may ensure $D \perp U \mid X$.

Appendix II contains a detailed discussion based on these heuristics supporting the conclusion that W can either contain proxies for U, observed drivers of D, or proxies for unobserved drivers of D. It should not contain outcomes driven by D. One can begin, then, by selecting an initial vector of covariates W satisfying these conditions. By themselves, these criteria do not guarantee $D \perp U \mid X$. On the other hand, however, W may indeed ensure this. We now discuss how to determine which of these possibilities holds and how this relates to specifying core and non-core covariates for robustness checking and testing.

5.2 Core Covariates

Suppose W ensures $D \perp U \mid X$. Then W may contain more information than is minimally needed to identify β_o. To obtain core covariates, we therefore seek a subvector W_1 of W such that $X_1 = (Z', W_1')'$ generates what Heckman and Navarro-Lozano (2004) call the *minimum relevant information set*, the smallest information set giving $D \perp U \mid X_1$. Note that W_1 could have dimension zero; we could also have $W_1 = W$.

Let $X = (X_1', W_2')'$. Given $D \perp U \mid X$, lemma 4.3 of D79 yields two conditions that imply $D \perp U \mid X_1$, so that W_1 is core and W_2 is not. One condition is $D \perp W_2 \mid X_1$. The other is $U \perp W_2 \mid X_1$.

First, consider $D \perp W_2 \mid X_1$. This condition involves only observables, so it is straightforward to investigate empirically. When the stochastic dependence can be fully captured by linear regression, one can regress D on $X = (Z', W')'$ to potentially reveal a subvector W_1 of W such that $D \perp U \mid X_1$. The matrix of regression coefficients for this regression is

$$\zeta^* = E(DX')E(XX')^{-1}.$$

Partitioning ζ^* ($k_0 \times (k_z + k_w)$, say) as $\zeta^* \equiv (\zeta_z^*, \zeta_w^*)$, suppose that there is (after a possible permutation) a partition of W, $W = (W_1', W_2')'$, such that, with $\zeta_w^* \equiv (\zeta_{w_1}^*, \zeta_{w_2}^*)$, we have $\zeta_{w_2}^* = 0$ (a $k_0 \times k_{w_2}$ matrix, say). Then $D \perp W_2 \mid (Z, W_1)$, so we can rule out W_2 as core covariates.

This suggests a straightforward, practical regression-based method for isolating non-core covariates: one can examine the sample regression coefficients from the regressions of each element of D on $X = (Z', W')'$ and identify non-core variables W_2 as those having estimated coefficients close to zero in every such regression. One may proceed heuristically or conduct more formal inference or model selection. Whatever the approach, one should err on the side of keeping rather than dropping variables, as retaining valid but non-essential covariates is much less costly than dropping true core covariates, as the former does not render estimates of β_o inconsistent, whereas the latter does.

In the fully general (nonlinear) case, analogous but more elaborate procedures can reveal such a W_1. For simplicity, we restrict attention here to the linear case.

Now consider whether we can identify further elements of W_1 as non-core. For this, we seek a partition of W_1, $W_1 = (W_{11}', W_{12}')'$ such that $U \perp W_{12} \mid X_{11}$, where $X_{11} = (Z', W_{11}')'$. If such a partition exists, we can rule out W_{12} as core variables, since we then have $D \perp U \mid X_{11}$. Since U is unobservable, we cannot proceed directly, as above. Nevertheless, we might try to use regression residuals $\varepsilon_{11} \equiv Y - E(Y \mid D, X_{11})$ to find such a partition. This turns out to be problematic, as we now discuss.

Specifically, observe that $D \perp U \mid X_{11}$ implies that $\varepsilon_{11} = U - E(U \mid X_{11})$. It follows by lemmas 4.1 and 4.2 of D79 that $U \perp W_{12} \mid X_{11}$ then implies $\varepsilon_{11} \perp W_{12} \mid X_{11}$. Now suppose we test and reject $\varepsilon_{11} \perp W_{12} \mid X_{11}$. Then $U \not\perp W_{12} \mid X_{11}$, so it is not clear whether W_{12} is core or not. Alternatively, suppose we fail to reject $\varepsilon_{11} \perp W_{12} \mid X_{11}$. Because this generally does not imply $U \perp W_{12} \mid X_{11}$, we again cannot make a determination. Properties of the regression residuals ε_{11} do not provide definitive insight. Instead, however, one might use knowledge of the underlying structure as discussed in Section 5.1 and in Section 5.3 to justify dropping or retaining elements of W_1. Specifying non-core elements W_{12} of W_1 satisfying $U \perp W_{12} \mid X_{11}$ unavoidably requires some judgement.

For practical purposes, then, we recommend using core covariates W_1 determined by checking $D \perp W_2 \mid X_1$ as described above, adjusted by deleting any elements for which $U \perp W_{12} \mid X_{11}$ is plausible a priori. Because we recommend erring on the side of inclusion, this may result in a set of core covariates that might be larger than absolutely necessary; but these should generally be quite adequate for robustness checking and testing.

For reference later, we note that an important consequence of $D \perp U \mid (X_1, W_2)$ and $D \perp W_2 \mid X_1$ following from lemma 4.3 of D79 is that $D \perp (U, W_2) \mid X_1$. Similarly, we have that $D \perp U \mid (X_1, W_2)$ and $U \perp W_2 \mid X_1$ imply $(D, W_2) \perp U \mid X_1$.

5.3 Non-core Covariates

We began with an initial set of covariates W and described how to extract core covariates; for convenience, we now just denote these W_1 and also write $X_1 = (Z', W_1')'$. To use a robustness check or test to see whether W_1 does in fact ensure $D \perp U \mid X_1$ and therefore identify β_o, we require suitable non-core covariates for the comparison regressions.

We may already have some non-core covariates, namely any components of W not in W_1. But this is not guaranteed, as we may have $W_1 = W$. Even if we do have some initial non-core covariates, it can be useful to find others. These can enhance the power of robustness tests; they can also enhance the efficiency of the estimator for β_o.

There are various ways to construct additional valid covariates. One effective way is to find covariates X_2 such that $D \perp U \mid X_1$ implies $D \perp U \mid X_2$. White and Chalak (2010, prop. 3) provide a relevant result, showing that if $D \perp U \mid X_1$ and if

$$W_2 = q(X_1, U, V), \quad \text{where} \quad D \perp V \mid (U, X_1)$$

for some unknown structural function q, then with $X_2 = (X_1', W_2')'$, we have

$$D \perp U \mid X_2.$$

A leading example occurs when W_2 is a proxy for U different than W_1, subject to measurement errors V, say $W_2 = q(U, V)$. If the measurement errors are independent of (D, U, X_1), as is often plausible, then $D \perp V \mid (U, X_1)$ also holds.

Because W_2 can be a vector here, any subvector of W_2 has the same properties. It follows that when W_2 is constructed in this way, automated methods, along the lines of STATA's *rcheck* and *checkrob* modules, that treat D and X_1 as core variables and W_2 as non-core (now taking $X = X_1$ in forming W_2), will yield a potentially useful set of comparison regressions.

White and Lu (2010b) provide various structural configurations sufficient for certain unconfoundedness assumptions used by Hahn (2004) in studying the efficiency of treatment effect estimators. These structures can inform the choice of both initial and additional covariates; here we focus on the latter. Although D is a binary scalar in Hahn (2004) and White and Lu (2010b), this is not required here: D can be a vector with categorical or continuous elements.

A condition sufficient for one of Hahn's unconfoundedness assumptions (his A.3(a)) is

$$D \perp (U, W_2) \mid X_1.$$

White and Lu (2010b) give two different underlying structures for which this holds. The example above from White and Chalak (2010, prop. 3) is another case where this holds. With $X_2 = (X_1', W_2')'$, lemma 4.2 of D79 gives the required condition (4),

$$D \perp U \mid X_1 \quad \text{and} \quad D \perp U \mid X_2.$$

Note that $D \perp (U, W_2) \mid X_1$ was implied for our initial covariates when $D \perp W_2 \mid X_1$. The covariates identified by White and Lu's (2010b) structures may be the same or different than those obtained from the initial covariates.

A condition sufficient for Hahn's other assumption (A.3(b)) is

$$(D, W_2) \perp U \mid X_1.$$

White and Lu (2010b) give two different underlying structures for which this holds. With $X_2 = (X_1', W_2')'$, this also implies

$$D \perp U \mid X_1 \qquad \text{and} \qquad D \perp U \mid X_2.$$

Note that $(D, W_2) \perp U \mid X_1$ was implied for our initial covariates when $U \perp W_2 \mid X_1$. The covariates identified by White and Lu's (2010b) structures may be the same or different than those obtained from the initial covariates.

More generally, suppose that

$$(D, W_{21}) \perp (U, W_{22}) \mid X_1.$$

With $W_2 = (W_{21}', W_{22}')'$ and $X_2 = (X_1', W_2')'$, lemma 4.2 of D79 again gives

$$D \perp U \mid X_1 \qquad \text{and} \qquad D \perp U \mid X_2.$$

In these last three examples, we have $X_2 = (X_1', W_2')'$, as in the first example. The structure of these results also ensures that a potentially useful comparison regression will result with core variables D and X_1 and any subvector of W_2 as non-core.

These examples are not necessarily the only possibilities, but they illustrate different ways to obtain additional non-core covariates potentially useful for robustness checking and testing. Significantly, the underlying economic structure plays a key role in determining both the non-critical core variables, X_1, and the non-core covariates, W_2. By failing to account for this structure, one may improperly specify core and non-core variables, making it easy to draw flawed economic inferences. Indeed, our characterization of some of these examples as yielding "potentially" useful non-core variables is meant to signal that these might *not* be useful after all, in the sense that their use in robustness checks or robustness tests provides no information about structural misspecification. The next section more explicitly identifies some of the pitfalls.

6 A Robustness Test for Structural Misspecification

Performing a robustness test is a completely standard procedure, so it is somewhat mystifying that such tests have not been routinely conducted, even if not focused on just the critical core coefficients. Here, we describe such a procedure, embodied in our *testrob* module. A Matlab routine for this can be found at `www.econ.ucsd.edu/~xunlu/code`. We hope its availablility will help make such tests standard.

To describe the robustness test, we first specify the robustness check regressions. Let the sample consist of n observations (Y_i, D_i, Z_i, W_i), $i = 1, ..., n$. The core variables are those appearing in the core regression: the critical core variables, D_i, and the non-critical core variables, $X_{1i} = (Z_i', W_{1i}')'$, where W_{1i} is a subvector of W_i. The remaining elements of W_i are the non-core variables. There are $J-1$ comparison regressions that include D_i and $X_{ji} = (X_{1i}', W_{ji}')'$, $j = 2, ..., J$, where W_{ji} is a subvector of the non-core elements of W_i. Not all subvectors of the non-core variables need to appear in the comparison regressions.

The ordinary least squares (OLS) robustness check regression estimators are

$$\hat{\delta}_{jn} \equiv \begin{pmatrix} \hat{\beta}_{jn} \\ \hat{\gamma}_{jn} \end{pmatrix} \equiv \begin{bmatrix} \mathbf{D}'\mathbf{D} & \mathbf{D}'\mathbf{X}_j \\ \mathbf{X}_j'\mathbf{D} & \mathbf{X}_j'\mathbf{X}_j \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{D}' \\ \mathbf{X}_j' \end{pmatrix} \mathbf{Y}, \qquad j = 1, ..., J,$$

where $\mathbf{D}$ is the $n \times k_0$ matrix with $k_0 \times 1$ rows D_i', $\mathbf{X}_j$ is the $n \times k_j$ matrix with $k_j \times 1$ rows X_{ji}', and $\mathbf{Y}$ is the $n \times 1$ vector with elements Y_i.

Letting $\hat{\delta}_n \equiv (\hat{\delta}_{1n}', ..., \hat{\delta}_{Jn}')'$, it follows under mild conditions (in particular, without assuming correct specification) that

$$\sqrt{n}(\hat{\delta}_n - \delta^*) \xrightarrow{d} N(0, M^{*-1} V^* M^{*-1}),$$

where $\delta^* \equiv (\delta_1^{*\prime}, ..., \delta_J^{*\prime})'$ and $\delta_j^* \equiv (\beta_j^{*\prime}, \gamma_j^{*\prime})'$, with optimal prediction coefficients β_j^* and γ_j^* as defined above, and where M^* and V^* are given by

$$M^* \equiv diag(M_1^*, ..., M_J^*) \qquad \text{and} \qquad V^* \equiv [V_{kj}^*],$$

where, for $k, j = 1, ..., J$,

$$M_j^* \equiv \begin{bmatrix} E(DD') & E(DX_j') \\ E(X_jD') & E(X_jX_j') \end{bmatrix} \qquad \text{and} \qquad V_{kj}^* \equiv \begin{bmatrix} E(\varepsilon_k\varepsilon_j DD') & E(\varepsilon_k\varepsilon_j DX_j') \\ E(\varepsilon_k\varepsilon_j X_kD') & E(\varepsilon_k\varepsilon_j X_kX_j') \end{bmatrix}.$$

See Chalak and White (2009) for regularity conditions in this context.

The critical core coefficient robustness hypothesis is

$$H_o : \Delta\, S\, \delta^* = 0,$$

where S is a selection matrix that selects $2 \le K \le J$ subvectors β_j^* from $\delta^* \equiv (\delta_1^{*\prime}, ...\delta_J^{*\prime})'$ and Δ is the $(K-1)k_0 \times Kk_0$ differencing matrix

$$\Delta = \begin{bmatrix} I & -I & 0 & & 0 \\ I & 0 & -I & & 0 \\ & & & \ddots & \\ I & 0 & 0 & & -I \end{bmatrix}.$$

The proper choice of subvectors selected by S is crucial. We discuss this below.

Letting $R \equiv \Delta\, S$, the robustness test statistic is

$$\mathcal{R}_{Kn} \equiv n\ \hat{\delta}_n'\, R'\ [R\hat{M}_n^{-1}\hat{V}_n\hat{M}_n^{-1}R']^{-1}\ R\,\hat{\delta}_n,$$

where $\hat{M}_n$ and $\hat{V}_n$ are consistent estimators of M^* and V^*, respectively, and it is assumed that $RM^{*-1}V^*M^{*-1}R'$ is nonsingular. The *testrob* routine estimates V^* under the assumption that the regression errors $\varepsilon_i \equiv (\varepsilon_{1i}, ..., \varepsilon_{Ji})'$ are uncorrelated across i.

As is standard, under H_o,

$$\mathcal{R}_{Kn} \xrightarrow{d} \chi^2_{(K-1)k_0},$$

where $\chi^2_{(K-1)k_0}$ denotes the chi-squared distribution with $(K-1)k_0$ degrees of freedom[3]. One rejects critical core coefficient robustness at the α level if $\mathcal{R}_{Kn}$ exceeds the $1-\alpha$ percentile of the $\chi^2_{(K-1)k_0}$ distribution. Because this test is a completely standard parametric test, it has power against local alternatives at rate $n^{-1/2}$.

Under the global alternative, $H_A : \Delta\, S\, \delta^* \neq 0$, the test is consistent; that is, it will eventually detect any departure of β_1^* from any selected β_j^*. This may leave some misspecified alternatives undetected, as explained above; also, rejection is non-specific, as it can signal regression non-linearity and/or failures of conditional exogeneity. Nevertheless, the structure generating the non-core variables affects the alternatives detected. For example, with non-core variables generated as in the first example of Section 5.3, the test will detect the failure of core conditional exogeneity, i.e., $D \perp U \mid X_1$, or the failure of $D \perp V \mid (U, X_1)$. (Since V can be anything, the assumption that $W_2 = q(X_1, U, V)$ is essentially without loss of generality.)

In fact, the way the non-core variables are generated plays a key role in determining the choice of comparison regressions. To see why, recall that $\varepsilon_j = U - E(U \mid X_j)$. Suppose that $X_k = (X_j', W_{jk}')'$ and that

$$U \perp W_{jk} \mid X_j.$$

Then

$$\varepsilon_k = U - E(U \mid X_k) = U - E(U \mid X_j, W_{jk}) = U - E(U \mid X_j) = \varepsilon_j.$$

It follows that using comparison regressions that include X_j as well as W_{jk} or its components introduces linear dependencies that increase the opportunities for singularity of $RM^{*-1}V^*M^{*-1}R'$ (hence $R\hat{M}_n^{-1}\hat{V}_n\hat{M}_n^{-1}R'$). We thus recommend caution in specifying comparison regressions of this sort.

[3] In practice, the number of degrees of freedom may need adjustment to accommodate certain linear dependencies among the elements of $R\hat{\delta}_n$. With linear dependencies, the statistic becomes $\mathcal{R}_{Kn} \equiv n\ \hat{\delta}_n'\, R'\ [R\hat{M}_n^{-1}\hat{V}_n\hat{M}_n^{-1}R']^{-}\ R\,\hat{\delta}_n$, where $[R\hat{M}_n^{-1}\hat{V}_n\hat{M}_n^{-1}R']^{-}$ denotes a suitable generalized inverse of $R\hat{M}_n^{-1}\hat{V}_n\hat{M}_n^{-1}R'$, and $\mathcal{R}_{Kn} \xrightarrow{d} \chi^2_{k^*}$, where $k^* \equiv \mathrm{rk}(R\hat{M}_n^{-1}\hat{V}_n\hat{M}_n^{-1}R')$.

As discussed in section 5, when $(D, W_{21}) \perp (U, W_{22}) \mid X_1$, we have $D \perp U \mid X_1, s(W_{21}), s(W_{22})$ for any subvector $s(W_{21})$ of W_{21} and subvector $s(W_{22})$ of W_{22} by D79 Lemma 4.2. But including all subvectors of (W_{11}, W_{22}) may introduce linear dependencies into $RM^{*-1}V^*M^{*-1}R'$. Thus, *testrob* carefully assesses and handles non-singularities in $R\hat{M}_n^{-1}\hat{V}_n\hat{M}_n^{-1}R'$.

Also, when there are multiple sets of covariates, say, X_j and X_k such that $D \perp U \mid X_j$ and $D \perp U \mid X_k$, it is important to recognize that it is generally *not* valid to use the combined vector (i.e., (X_j, X_k)) for robustness checking, as there is no guarantee that $D \perp U \mid (X_j, X_k)$ holds. Algorithms that include all possible subsets of the non-core variables, such as *rcheck* or *checkrob*, ignore this restriction, leading to the likely appearance of misleading non-robustness in the critical core coefficients.

7 Gaining Insight into Robustness Rejections

When the robustness test rejects, it would be helpful to know why. Is it regression nonlinearity or is it failure of conditional exogeneity? There is an extensive literature on testing for neglected nonlinearity in regression analysis, ranging from Ramsey's (1969) classic RESET procedure to modern neural network or random field tests. (See, for example, Lee, White and Granger, 1993; Hamilton, 2001; and Dahl and Gonzalez-Rivera, 2003.) These references, like those to follow, barely scratch the surface, but they can at least help in locating other relevant work. Such tests are easy and should be routinely performed; for example, one can use STATA's *reset* procedure.

If the nonlinearity test rejects, one may seek more flexible specifications. The relevant literature is vast, ranging from use of more flexible functional forms including algebraic or trigonometric polynomials (e.g. Gallant, 1982); partially linear models (Engle, Granger, Rice, and Weiss, 1986; see Su and White (2010) for a partial review); local polynomial regression (e.g., Cleveland, 1979; Ruppert and Wand, 1994); and artificial neural networks (e.g., White, 2006a).

On the other hand, if robustness is rejected but not linearity, then we have apparent evidence of the failure of conditional exogeneity. But the robustness test is sufficiently nonspecific that it is helpful to employ further diagnostics to understand what may be driving the robustness rejection. At the heart of the matter is whether $D \perp U \mid X_1$ holds for the core regression. There are now many approaches to testing this in the literature. Early procedures were developed for the case of scalar binary D. See Rosenbaum (1987), Heckman and Hotz (1989), and the review by Imbens (2004). Recently White and Chalak (2010) have given tests complementary to these that permit D to be a vector of categorical or continuous variables. White and Lu (2010a) give straightforward tests for conditional exogeneity that involve regressing transformed regression residuals on functions of the original regressors. In the next section, we discuss a special case involving squared residuals that emerges naturally in the context of efficient estimation.

Given the realities of the review and publication process in economics, it is perhaps not surprising that the literature contains plenty of robustness checks, but not nearly as much in the way of more extensive specification analysis. By submitting only results that may have been arrived at by specification searches designed to produce plausible results passing robustness checks, researchers can avoid having reviewers point out that this or that regression coefficient doesn't make sense or that the results might not be robust. And if this is enough to satisfy naive reviewers, why take a chance? Performing further analyses that could potentially reveal specification problems, such as nonlinearity or exogeneity failure, is just asking for trouble.

Of course, this is also a recipe for shoddy research. Specification searches designed to achieve the appearance of plausibility and robustness have no place in a scientific endeavor. But even in the absence of such data snooping, it must be recognized that robustness checks are not tests, so they cannot have power. And, although robustness *tests* may have power, they are non-specific and may thus may lack power in specific important directions. By deploying true specification testing methods, such as the Hausman-style robustness test of the previous section, the further tests described in this section, or any of the methods available in the vast specification testing literature to examine whether findings supported by robustness checks really do hold up, researchers can considerably improve the quality of empirical research. Reviewers of work submitted for publication can encourage this improvement rather than hindering it by focusing on critical core coefficients only and by reorienting their demands toward *testing* robustness and specification.

8 Estimator Efficiency with Correct Structural Specification

Suppose the robustness test does not reject, and that other specification tests (e.g., nonlinearity, conditional exogeneity) do not indicate specification problems. Then one has multiple consistent estimators for β_o, and one must decide which to use. Should one just use the core regression or instead use a regression including non-core variables? Or should one use some other estimator?

The criterion that resolves this question is estimator efficiency. We seek the most precise estimator of the critical core coefficients, as this delivers the tightest confidence intervals and the most powerful tests for the effects of interest. Recent work of Hahn (2004) shows that which estimator is efficient is not immediately obvious. As Hahn (2004) and White and Lu (2010b) further show, this depends on the exact nature of the underlying structure.

8.1 Efficiency Considerations

To understand the main issues in the present context, recall that the asymptotic covariance matrix for $\hat{\delta}_{nj} \equiv (\hat{\beta}_{nj}', \hat{\gamma}_{nj}')'$, $avar(\hat{\delta}_{nj})$, is $M_j^{*-1} V_j^* M_j^{*-1}$, where

$$M_j^* \equiv \begin{bmatrix} E(DD') & E(DX_j') \\ E(X_j D') & E(X_j X_j') \end{bmatrix} \quad \text{and} \quad V_j^* \equiv \begin{bmatrix} E(\varepsilon_j^2 DD') & E(\varepsilon_j^2 DX_j') \\ E(\varepsilon_j^2 X_j D') & E(\varepsilon_j^2 X_j X_j') \end{bmatrix}.$$

Suppose for simplicity that $\sigma_j^2 \equiv E(\varepsilon_j^2 \mid D, X_j)$ does not depend on D or X_j. Since this assumption is unrealistic, we remove it later; for now, it enables us to expose the key issues. With constant σ_j^2, we have the classical result that

$$M_j^{*-1} V_j^* M_j^{*-1} = \sigma_j^2 \begin{bmatrix} E(DD') & E(DX_j') \\ E(X_j D') & E(X_j X_j') \end{bmatrix}^{-1}.$$

A little algebra shows that for the critical core coefficients $\hat{\beta}_{nj}$, we have

$$avar(\hat{\beta}_{nj}) \equiv \sigma_j^2 \, E(\eta_j \eta_j')^{-1},$$

where $\eta_j \equiv D - \zeta_j^* X_j$ is the vector of residuals from the linear regression of D on X_j, with regression coefficient matrix $\zeta_j^* \equiv E(DX_j') E(X_j X_j')^{-1}$. We now see the essential issue: estimator efficiency is determined by a trade-off between the residual variance of the Y regression, σ_j^2, and the residual covariance matrix of the D regression, $E(\eta_j \eta_j')$. Adding regressors beyond those of the core regression may either increase or decrease $avar(\hat{\beta}_{nj})$, depending on whether the included non-core variables reduce σ_j^2 enough to offset the the corresponding increase in $E(\eta_j \eta_j')^{-1}$ resulting from the reduction in $E(\eta_j \eta_j')$.

The high-level intuition is that additional non-core covariates should be included if they tend to act as proxies for U (reducing σ_j^2) and should not be included if they tend to act as drivers of D or proxies for unobserved drivers of D (reducing $E(\eta_j \eta_j')$). Indeed, when for some W^*, $W^* \perp U \mid X_j$ holds, adding elements of W^* to X_j can only increase $E(\eta_j \eta_j')^{-1}$, as adding W^* leaves σ_j^2 unchanged.

A general investigation of the efficiency bounds and of estimators attaining these bounds relevant to selecting covariates in the current context is a fascinating topic. The work of Hahn (2004) for the case of scalar binary D is a good start, but the issues appear sufficiently challenging that a general analysis is beyond our scope here. Instead, we discuss practical methods that can in the meantime provide concrete benefits in applications.

One practical approach is to select the most efficient estimator among the robustness check regressions by simply comparing their asymptotic covariance matrix estimators. This is particularly straightforward for the estimators used in the robustness test, as their asymptotic covariances are already estimated as part of the test.

But one can do better, because the efficiency of the estimators used in the robustness check can be improved by using generalized least squares (GLS) instead of OLS. Further, the nature of these GLS corrections is of independent interest, as, among other things, this bears on the general analysis of efficient estimators.

8.2 GLS for the Robustness Check Regressions

In line with our previous discussion, from now on we consider robustness check regressions where X_j contains X_1. For regression j with independent or martingale difference regression errors ε_{ji}, the GLS correction involves transforming each observation by $\sigma_{ji}^{-1} = E(\varepsilon_{ji}^2 \mid D_i, X_{ji})^{-1/2}$. Recall that $D \perp U \mid X_j$ implies $\varepsilon_j = U - E(U \mid X_j)$. Then

$$\begin{aligned} E(\varepsilon_j^2 \mid D, X_j) &= E([U - E(U \mid X_j)]^2 \mid D, X_j) \\ &= E(U^2 \mid D, X_j) - 2E(U \mid D, X_j)E(U \mid X_j) + E(U \mid X_j)^2 \\ &= E(U^2 \mid X_j) - E(U \mid X_j)^2, \end{aligned}$$

where the final equality holds because $D \perp U \mid X_j$. We see that with conditional exogeneity, conditional heteroskedasticity only depends on X and does not depend on D. Also, we see that conditional heteroskedasticity is typical, as X is chosen so that X and U are dependent. GLS is thus typically required for estimator efficiency.

The fact that $E(\varepsilon_j^2 \mid D, X_j)$ does not depend on D has useful implications. First, it suggests another way to test conditional exogeneity. Second, it simplifies the GLS computations. We first take up GLS estimation and then discuss specification testing.

Typically, the conditional variances $\sigma_{ji}^2 = E(\varepsilon_{ji}^2 \mid D_i, X_{ji})$ are unknown; given suitable estimators $\hat{\sigma}_{ji}$, a feasible GLS (FGLS) estimator for robustness check regression j is

$$\tilde{\delta}_{jn} \equiv \begin{pmatrix} \tilde{\beta}_{jn} \\ \tilde{\gamma}_{jn} \end{pmatrix} \equiv \begin{bmatrix} \tilde{\mathbf{D}}'\tilde{\mathbf{D}} & \tilde{\mathbf{D}}'\tilde{\mathbf{X}}_j \\ \tilde{\mathbf{X}}_j'\tilde{\mathbf{D}} & \tilde{\mathbf{X}}_j'\tilde{\mathbf{X}}_j \end{bmatrix}^{-1} \begin{pmatrix} \tilde{\mathbf{D}}' \\ \tilde{\mathbf{X}}_j' \end{pmatrix} \tilde{\mathbf{Y}},$$

where $\tilde{\mathbf{D}}$ is the $n \times k_0$ matrix with $k_0 \times 1$ rows $\tilde{D}_i' \equiv D_i'/\hat{\sigma}_{ji}$, $\tilde{\mathbf{X}}_j$ is the $n \times k_j$ matrix with $k_j \times 1$ rows $\tilde{X}_{ji}' = X_{ji}'/\hat{\sigma}_{ji}$, and $\tilde{\mathbf{Y}}$ is the $n \times 1$ vector with elements $\tilde{Y}_i = Y_i/\hat{\sigma}_{ji}$. Regression linearity ensures that $\tilde{\delta}_{jn}$ is consistent for δ_j^*. Conditions ensuring that FGLS estimators are asymptotically equivalent to GLS for nonparametric choice of $\hat{\sigma}_{ji}$ are given by Robinson (1987), Stinchcombe and White (1991), and Andrews (1994), among others.

An interesting possibility discussed by Stinchcombe and White (1991) is to estimate σ_{ji}^2 using artificial neural networks. This entails estimating a neural network regression model, such as

$$\hat{\varepsilon}_{ji}^2 = \sum_{\ell=1}^{q} \psi(\tau_{j\ell 0} + X_{ji}'\tau_{j\ell})\, \theta_{j\ell} + \upsilon_{ji}, \tag{5}$$

where $\hat{\varepsilon}_{ji} \equiv Y_i - D_i'\hat{\beta}_{jn} - X_{ji}'\hat{\gamma}_{jn}$ is the estimated residual for observation i from robustness check regression j, $q = q_n$ is the number of "hidden units" (with $q_n \to \infty$ as $n \to \infty$), ψ is the hidden unit "activation function," e.g., the logistic CDF or PDF, $(\tau_{j\ell 0}, \tau_{j\ell}', \theta_{j\ell}')$, $\ell = 1, ..., q$, are parameters to be estimated, and υ_{ji} is the regression error. When ψ is suitably chosen, nonlinear estimation can be avoided by choosing $(\tau_{j\ell 0}, \tau_{j\ell}')'$ randomly; see White (2006a). The estimates $\hat{\sigma}_{ji}^2$ are the fitted values from this regression. Observe that this estimation is simplified by the fact that D_i need not be included on the right-hand side above. This conserves degrees of freedom, offering significant benefits for nonparametric estimation.

The FGLS critical core coefficient estimator $\tilde{\beta}_{nj}$ now has

$$avar(\tilde{\beta}_{nj}) \equiv E(\tilde{\eta}_j \tilde{\eta}_j')^{-1},$$

where $\tilde{\eta}_j \equiv (D - \tilde{\zeta}_j^* X_j)/\sigma_j$ and $\tilde{\zeta}_j^* \equiv E(DX_j'/\sigma_j^2)E(X_j X_j'/\sigma_j^2)^{-1}$. If j^* indexes the efficient estimator, then for all $j \neq j^*$, $avar(\tilde{\beta}_{nj}) - avar(\tilde{\beta}_{nj^*})$ will be positive semi-definite. The covariance matrices can be estimated in the obvious way, and one can search for an estimator with the smallest estimated covariance. In *testrob*, the covariance matrices are estimated using the heteroskedasticity-consistent method of White (1980), to accommodate the fact that the FGLS adjustment might not be fully successful.

8.3 Combining FGLS Estimators

Even if one finds an FGLS estimator with smallest covariance matrix, there is no guarantee that this estimator makes efficient use of the sample information. Usually, one can find a more efficient estimator by optimally combining the individual FGLS estimators. To describe this estimator, we note that, parallel to the OLS case, asymptotic normality holds for FGLS, i.e.,

$$\sqrt{n}(\tilde{\delta}_n - \delta^*) \xrightarrow{d} N(0, \tilde{M}^{*-1}\tilde{V}^*\tilde{M}^{*-1}),$$

under mild conditions, with

$$\tilde{M}^* \equiv diag(\tilde{M}_1^*, ..., \tilde{M}_J^*) \qquad \text{and} \qquad \tilde{V}^* \equiv [\tilde{V}_{kj}^*],$$

where, for $k, j = 1, ..., J$,

$$\tilde{M}_j^* \equiv \begin{bmatrix} E(DD'/\sigma_j^2) & E(DX_j'/\sigma_j^2) \\ E(X_j D'/\sigma_j^2) & E(X_j X_j'/\sigma_j^2) \end{bmatrix} \qquad \text{and} \qquad \tilde{V}_{kj}^* \equiv \begin{bmatrix} E(\tilde{\varepsilon}_k \tilde{\varepsilon}_j DD'/\sigma_k\sigma_j) & E(\tilde{\varepsilon}_k \tilde{\varepsilon}_j DX_j'/\sigma_k\sigma_j) \\ E(\tilde{\varepsilon}_k \tilde{\varepsilon}_j X_k D'/\sigma_k\sigma_j) & E(\tilde{\varepsilon}_k \tilde{\varepsilon}_j X_k X_j'/\sigma_k\sigma_j) \end{bmatrix},$$

with $\tilde{\varepsilon}_j \equiv (Y_j - D'\beta_o - X_j'\gamma_j^*)/\sigma_j$.

Letting S be the $Jk_0 \times \dim(\delta^*)$ selection matrix that extracts $\tilde{\beta}_n \equiv (\tilde{\beta}_{n1}', ..., \tilde{\beta}_{nJ}')'$ from $\tilde{\delta}_n$ (so $\tilde{\beta}_n = S\,\tilde{\delta}_n$), the asymptotic normality result for $\tilde{\beta}_n$ can be represented as an artificial regression

$$\sqrt{n}\,\tilde{\beta}_n = \sqrt{n}\,\mathcal{I}\,\beta_o + \upsilon,$$

where the $Jk_0 \times k_0$ matrix of artificial regressors is $\mathcal{I} \equiv \boldsymbol{\iota} \otimes I_{k_0}$, where $\boldsymbol{\iota}$ is the $J \times 1$ vector of ones and I_{k_0} is the identity matrix of order k_0; and υ is the $Jk_0 \times 1$ artificial regression error, distributed as $N(0, \Omega^*)$ with $\Omega^* \equiv S\tilde{M}^{*-1}\tilde{V}^*\tilde{M}^{*-1}S'$. For simplicity, we assume that Ω^* is nonsingular; if not, we simply drop comparison regressions until nonsingularity holds.

The Feasible Optimally combined GLS (FOGLeSs) estimator is obtained by applying FGLS to this artificial regression:

$$\tilde{\beta}_n^* = (\mathcal{I}'\hat{\Omega}^{*-1}\mathcal{I})^{-1}\mathcal{I}'\hat{\Omega}^{*-1}\tilde{\beta}_n,$$

where $\hat{\Omega}^*$ is a suitable consistent estimator of Ω^*. This estimator satisfies

$$\sqrt{n}(\tilde{\beta}_n^* - \beta_o) \xrightarrow{d} N(0, (\mathcal{I}'\Omega^{*-1}\mathcal{I})^{-1}).$$

One can consistently estimate the asymptotic covariance matrix $(\mathcal{I}'\Omega^{*-1}\mathcal{I})^{-1}$ in a number of obvious ways.

The estimator $\tilde{\beta}_n^*$ is optimal in the sense that for any other combination estimator $\tilde{\beta}_n^\dagger = A\tilde{\beta}_n$ such that $\tilde{\beta}_n^\dagger \xrightarrow{p} \beta_o$, where A is a nonstochastic $k_0 \times Jk_0$ weighting matrix, we have that $avar(\tilde{\beta}_n^\dagger) - avar(\tilde{\beta}_n^*)$ is positive semi-definite.

Of course, this is only a relative efficiency result. That is, $\tilde{\beta}_n^*$ makes relatively efficient use of the FGLS comparison regressions, but it need not be fully efficient. This is nevertheless useful, given that the relevant efficiency bound and the estimator attaining this bound are presently unknown. But note that if one of the comparison FGLS regressions is relatively (or even fully) efficient, then $A^* \equiv (\mathcal{I}'\Omega^{*-1}\mathcal{I})^{-1}\mathcal{I}'\Omega^{*-1}$ will signal this by taking the form

$$A^* = [\mathbf{0}, ..., \mathbf{0}, I_{k_o}, \mathbf{0}, ..., \mathbf{0}],$$

where $\mathbf{0}$ is a $k_0 \times k_0$ zero matrix, and I_{k_0} appears in the j^*th position when $\tilde{\beta}_{nj^*}$ is the relatively (fully) efficient estimator. It will thus be interesting to inspect the sample estimator of A^* to see if it approximates this form. If so, one may prefer just to use the indicated estimator.

Given that we are considering making efficient use of information contained in a system of related equations, one might consider stacking the robustness check regressions and trying to construct an estimator analogous to the Seemingly Unrelated Regression (SUR) estimator. As it turns out however, this strategy breaks down here. The problem is that a necessary condition for the consistency of this estimator here is that $E(\varepsilon_j \mid D, Z, W) = 0$, $j = 1, ..., J$, but this generally fails for the robustness check comparison regressions. Instead, we have only that $E(\varepsilon_j \mid D, X_j) = 0$, which generally does not imply $E(\varepsilon_j \mid D, Z, W) = 0$.

To close this section, we note that using some estimator other than the FOGLeSs estimator, for example, the core regression OLS estimator, gives inefficient and therefore potentially flawed inferences about the critical core coefficients.

8.4 Further Specification Analysis

Observe that the FGLS and the FOGLeSs estimators are consistent for the same coefficients as the corresponding OLS estimators. Any apparent divergence among these estimators potentially signals misspecification (i.e., regression nonlinearity or failure of conditional exogeneity). These differences can be formally tested using a standard Hausman (1978) test. As this is completely standard, we do not pursue this here.

As noted above, the fact that $E(\varepsilon_j^2 \mid D, X_j)$ does not depend on D suggests a convenient and informative way to test conditional exogeneity. Specifically, for any j, one can estimate a neural network regression of the form

$$\hat{\varepsilon}_{ji}^2 = D_i'\lambda_j + \sum_{\ell=1}^{q} \psi(\tau_{j\ell 0} + X_{ji}'\tau_{j\ell})\,\theta_{j\ell} + \upsilon_{ji}. \tag{6}$$

Under the null of conditional exogeneity, $H_{0j} : D \perp U \mid X_j$, we have $\lambda_j = 0$. Full details of testing $\lambda_j = 0$ in this way can be found in White and Lu (2010a, section 6.3).

The natural point in the process to do these tests is immediately after performing the robustness test and before implementing FGLS. These results provide important diagnostics about where specification problems may lie that one should be aware of before going further. They also can be quickly adjusted to use in implementing FGLS.

The diagnostic information relates to whether $D \perp U \mid X_j$. Suppose the robustness test does not reject. If we nevertheless reject H_{0j}, this signals a specification problem that the robustness test may lack power to detect. On the other hand, if the robustness test did reject, then this signals a possible reason for the rejection.

This diagnostic for the core regression, $j = 1$, is especially informative. First, we can test H_{01} even when a robustness test is not possible (for example, due to singularity of $RM^{*-1}V^*M^{*-1}R'$). If we reject H_{01}, then we have evidence that the core covariates are insufficient to identify β_o; this is, after all, the paramount issue for the study at hand. Further, regardless of whether or not we reject H_{01}, the comparison regression diagnostics indicate which comparison regressions cannot be validly used in the FOGLeSs estimator. If for any j we reject H_{0j}, then we have evidence that $\beta_j^* \neq \beta_o$, so there is little point to computing the FGLS estimator $\tilde{\beta}_{nj}$ and including it in obtaining the FOGLeSs estimator. If we do have some comparison regressions with covariates that identify β_o, we can just use these in the FOGLeSs estimator, regardless of the outcome of the robustness test. Examining the comparison regression diagnostics may help in finding covariates that do identify β_o, as we will not reject H_{0j} for these regressions. Care must be taken, however, because $E(\varepsilon_j^2 \mid D, X_j)$ not depending on D does not imply $D \perp U \mid X_j$; ideally, judgement informed by structural insight will be applied in such cases.

9 A Step-by-Step Summary

This section summarizes the robustness checking and testing process discussed here. We discuss the modeling steps required prior to using *testrob*, followed by those of the *testrob* module. We also offer some supplemental comments. The process is interactive, requiring the researcher to make a number of decisions along the way.

We emphasize that what follows is not intended as a definitive implementation of robustness checking and testing as advocated here. Instead, it is designed to be a prototypical template that can be conveniently used in applications and that can serve as an accessible basis for variants that may be more sophisticated or better suited to specific contexts.

Before using *testrob,* we recommend that the researcher take the following steps:

1. Identify an outcome of interest, Y, and potential causes of interest, D. The latter are the critical core variables, and it is important to designate them as such, both in interpreting and reporting the results. This helps ensure that attention is properly focused on the coefficients whose plausibility and robustness matter and not focused on coefficients irrelevant to the main research concerns.

2. Formally or informally model not only the economic structure relating D to Y, but also that determining D. From the structure determining Y, one can ascertain the actual or potential drivers of Y other than D. Those drivers that can be accurately observed are non-critical core variables, Z. Those that cannot are the unobservables, U. Similarly, from the structure for D, one can ascertain its observed and unobserved drivers; call these Q and V, respectively.

3. Specify valid covariates, W. As discussed above, these can be proxies for U, observed drivers Q of D, and proxies for unobserved drivers V of D. They should not be variables driven by D. This specification should be comprehensive and in accord with Section 5.

   Several subgroups of covariates must be separately designated for *testrob*: W_0, the initial covariates, which will be used to construct the core covariates; and $L > 0$ groups of non-core covariates, W_ℓ^*. The latter can be either groups of non-core covariates to be used *as is* in the robustness check comparison regression(s), or they can be groups of covariates and their subsets to be used in the comparison regressions. A "flag" to *testrob* specifies whether a group is to be used as is, or as a group together with specified subsets. The groups are specified to *testrob* by designating indexes of the W vector. The indexes of W_0 should not overlap with those of the W_ℓ^*'s. The indexes of the W_ℓ^*'s can overlap but should not coincide.

We comment that there is no necessary virtue in proliferating comparison regressions, as a single comparison regression can signal structural misspecification. Having many comparison regressions can dilute the power of the robustness test, especially by introducing near singularities into the needed asymptotic covariance matrix. Also, the required computing time increases (possibly substantially) as a function of the number of comparison regressions. One possible benefit to having more comparison regressions is improved efficiency for the FOGLeSs estimator, but this can also be achieved by judicious choice of non-core covariates, particularly proxies for U that impact its conditional variance.

With these preparations, one is ready to invoke *testrob*. Appendix III contains the dialog for a *testrob* application, following the steps specified next.

4. The first task for *testrob* is to identify core covariates. One option is to declare that W_0 contains no non-core covariates. This sets $X_1 = (Z', W_0')'$ as the vector of non-critical core covariates, and one proceeds to the next step. Otherwise, *testrob* regresses each element of D on $X_0 = (Z', W_0')'$. These results are presented to the user, together with a list of the variables in decreasing order of the $p-$value for the chi-squared statistic for the hypothesis that the coefficients of that variable are jointly zero. Recall that non-core variables will have approximately zero coefficients in all regressions, so the top variables in this list are those most plausible to be non-core. The *testrob* process then queries the user as to which, if any covariates to treat as non-core. Denote the remaining initial covariates $\tilde{W}_0$. At this point, one can declare that $\tilde{W}_0$ contains no non-core covariates, which sets $X_1 = (Z', \tilde{W}_0')'$ as the vector of core covariates. One then proceeds to the next step.

   If one proceeds with the current step, the *testrob* process then queries the user as to which additional covariates to treat as non-core. As described in Section 5.2, the researcher might use knowledge about the underlying structure to decide this. The remaining variables, W_1, are used to construct the non-critical core covariates, $X_1 = (Z', W_1')'$.

   The non-core variables from this step, denoted W_0^*, are treated in the same way as the user-specified core variables.

5. Next, *testrob* conducts the specified comparison regressions and the robustness test for the critical core coefficients. The non-core covariates in the comparison regressions include the W_ℓ^*'s specified by the user, together with the non-core variables W_0^* identified in step 4. These results are presented to the user, together with information about which comparison regressions had to be dropped to achieve a non-singular robustness test covariance matrix. To provide further insight, *testrob* also computes and reports diagnostics for regression

nonlinearity and conditional exogeneity for each retained comparison regression. The conditional exogeneity test is that described in Section 8.4. The nonlinearity test is a neural network test for neglected nonlinearity described in Lee, White, and Granger (1993).

6. If the results of step 5 suggest serious structural misspecification, the researcher may decide to terminate the *testrob* process and rethink his or her model specification. If, however, the results support regression linearity and conditional exogeneity for some subset of the comparison regressions, one may decide to continue. For this, the process queries the user as to which comparison regressions to use. Given these, *testrob* estimates the squared residual equation (5), computes the FGLS regressions, and reports the results to the user.

7. Finally, *testrob* computes the FOGLeSs estimator and associated statistics, reports these to the user, and terminates.

10 An Illustrative Example

We illustrate the *testrob* procedure using the data set analyzed by Pérez-González (2006). This is a rich dataset to which Pérez-González applies methods that correspond closely to our discussion in the introduction. Pérez-González is interested in the impact of inherited control on firm performance. He uses data from 335 management transitions of publicly traded U.S. corporations to examine whether firms with familially related incoming chief executive officers (CEOs) underperform in terms of operating profitability relative to firms with unrelated incoming CEOs. Thus, in this application, D is a binary variable that equals 1 if the incoming CEO is related to the departing CEO, to the founder, or to a large shareholder by blood or marriage and that equals 0 otherwise. Pérez-González uses operating return on assets (OROA) as a measure of firm performance. Specifically, Y here is the difference in OROA calculated as the three-year average after succession minus the three-year average before succession. Precise definitions of D and Y are given in Tables $1(a)$ and $1(b)$, respectively.

Table 1 (a): Outcome of interest

	Name	Description
Y	Difference in OROA	the three year average of industry- and performance- adjusted OROA after CEO transitions *minus* the three year average before CEO transitions[4]

Table 1 (b): Potential cause of interest (critical core variable)

	Name	Description
D	Family CEO	= 1 if the incoming CEO of the firm is related by blood or marriage to the departing CEO, to the founder, or to a larger shareholder = 0 otherwise

As stated in Pérez-González (2006, p.1559), "the main argument against family succession in publicly traded firms is that competitive contest for top executive positions would rarely result in a family CEO". Nevertheless, Pérez-González (2006, p.1560) also argues that family succession may benefit firm performance, for example, by "reducing agency problems," "facilitating firm specific investment," "easing cooperation and the transmission of knowledge within organizations," or "having a long-term focus that unrelated chief executives lack". Thus, it is of interest to examine the impact of family succession on firm performance empirically.

Table 1 (c): Covariates

	Name	Description
W (1)	Ln sales	logarithm of sales one year prior to the CEO transition
W (2)	Industry-adjusted OROA	industry adjusted OROA one year prior to the CEO transition
W (3)	Industry-adjusted M-B	industry adjusted market-to-book (M-B) ratio one year prior to the CEO transition
W (4)	Board ownership	the fraction of ownership held by officers and directors
W (5)	Family directors	= 1 if the fraction of family to total directors is higher than the median in the sample = 0 otherwise
W (6)	Mean pre-transition industry and performance adjusted OROA	three year pre-transition average of the industry- and performance- adjusted OROA
W (7)	Less selective college	= 1 if the college attended by the incoming CEO is not in "very competitive" or higher in Barron's ranking = 0 otherwise
W (8)	Graduate school	= 1 if the incoming CEO attended a graduate school = 0 otherwise
W (9)	Age promoted	the age when the incoming CEO is appointed
W (10)	Woman	= 1 if the incoming CEO is a woman = 0 otherwise
W (11)	Positive R&D expenses	= 1 if the firm reported positive R&D expenses the year prior to the CEO transition = 0 otherwise
W (12)	Nonretirements	= 1 if the departing CEO was not reported to leave the firm due to a "retirement" = 0 otherwise
W (13)	Early succession	= 1 if the departing CEO left his position before 65 = 0 otherwise
W (14)	Departing CEO remains as chairman	= 1 if the departing CEO continued as chairman after the CEO transition = 0 otherwise
W (15)	CEO ownership	the ownership share of the incoming CEO
W (16) – W (34)	Year dummy, 1981-1999	year dummies

Although Pérez-González does not include error-free measures of other causes Z of Y, he does include a large number of covariates to control for unobserved drivers. We list these in Table 1(c).

Next, we demonstrate implementing *testrob* as described above and present the results. Appendix III contains the corresponding *testrob* dialog.

1. As discussed above, Y is the "Difference of OROA". D is "Family CEO".

2. Many factors affect OROA during the CEO transition period, for example, the size of the firm, how the firm performed in the past, the board characteristics, and how the firm invested in R&D. We assume that we do not observe these other true drivers of OROA, but observe their proxies as discussed in step 3 below. Thus we let Z be empty. There are also many factors influencing the CEO hiring decision, for example, characteristics of the incoming CEO, such as age, gender, and education. Also, as stated in Pérez-González (2006, p.1578), "previous studies have shown that firm performance, size and board characteristics affect firms' hiring and firing decisions, as well as selection of internal relative to external candidates." For example, small firms may have difficulty in hiring competent unrelated managers. A departing CEO who overperforms relative to other firms in the same industry, may "have power and (influence to name an offerspring as CEO)."

3. We further classify the available covariates into different groups. We use "Ln sales" to proxy firm size, "Industry adjusted OROA", "Industry-adjusted M-B," and "Mean pre-transition industry- and performance- adjusted OROA" to proxy for the firm's past performance. We use "board ownership" and "Family directors" to proxy for board characteristics. We use "Positive R&D expense" to proxy for the firm's R&D expenditure. Further, as pointed out in Pérez-González (p.1582), "CEO separation conditions or the age at which the departing CEO retires, may reveal information about the state of affairs of a corporation that is not captured by firm characteristics". We use "Nonretirements," "Early succession," "Departing CEO remains as chairman," and "CEO ownership" to represent this. Further, we use "Less selective college," "Graduate school," "Age promoted,"

and "Woman" to proxy the incoming CEO's characteristics. Table 2 summarizes.

Table 2: Covariate classification

Covariate group	Covariates
Firm size	$W(1)$
Firm's past performance	$W(2), W(3), W(6)$
Board characteristics	$W(4), W(5)$
Firm's R&D expenditure	$W(11)$
Departing CEO's separation conditions and incoming CEO's ownership	$W(12), W(13), W(14), W(15)$
Incoming CEO's characteristics	$W(7), W(8), W(9), W(10)$

As we discussed in step 2, firm size, firm's past performance, and firm's board characteristics affect both OROA during the CEO transition period and the selection of the CEO. Therefore, we include proxies for these among the core covariates. We also include year dummies as core covariates. Thus, the initial core covariates are $W_0 = \{W(1), W(2), W(3), W(4), W(5), W(6), W(16) - W(34)\}$.[5] We use the remaining three groups in the comparison regressions. That is, $W_1^* = \{W(7), W(8), W(9), W(10)\}$, $W_2^* = \{W(11)\}$, and $W_3^* = \{W(12), W(13), W(14), W(15)\}$. To avoid proliferating comparison regressions, we do not use subsets of W_1^*, W_2^*, or W_3^*.

4. In this step, we identify core covariates. We do not declare that the initial core covariates contain no non-core covariates. Thus, *testrob* regresses D on $X_0 = W_0$. Table 3 below presents the results.

Table 3: Regression of D on X_0

Covariates	$W(3)$	$W(1)$	$W(6)$	$W(4)$	$W(2)$	$W(5)$
$p-$values	0.973	0.885	0.766	0.471	0.334	0.000

Note: The results for year dummies are not reported. We include these in all regressions.

$W(3), W(1), W(6), W(4)$, and $W(2)$ all have high $p-$values. We could drop all of these from the core covariates. To be conservative, however, we decide to drop $\{W(3), W(1), W(6)\}$. Thus the core covariates are now $W_0 = \{W(2), W(4), W(5)\}$. The new non-core covariates $W_0^* = \{W(1), W(3), W(6)\}$ become a new comparison group. We only use the full group W_0^* in the comparison regressions.

[5] This set of covariates is exactly that used in column 1 in Table 9 in Pérez-González (2006, p.1581).

5. We now have 5 sets of comparison groups, as shown in Table 4.

Table 4: Sets of comparison regressions

Set	Regressors
Set 1	$\{D\} \cup W_0$
Set 2	$\{D\} \cup W_0 \cup W_0^*$
Set 3	$\{D\} \cup W_0 \cup W_1^*$
Set 4	$\{D\} \cup W_0 \cup W_2^*$
Set 5	$\{D\} \cup W_0 \cup W_3^*$

Testrob performs a singularity check; all five sets are retained. Then *testrob* implements the robustness test and performs linearity and conditional exogeneity tests for each set. Table 5 presents the results.

Table 5: Robustness and diagnostic tests

Robustness test p-value			0.768		
Diagnostic tests	Set 1	Set 2	Set 3	Set 4	Set 5
Linearity test p-values	0.635	0.756	0.644	0.451	0.279
Conditional exogeneity test p-values	0.248	0.885	0.958	0.551	0.651

6. It appears that there is no structural misspecification detected in step 5. We thus decide to continue, using all five comparison groups. *Testrob* now estimates the coefficient on the critical core variable for each covariate group and also computes the FOGLeSs regression. Table 6 presents the results:

Table 6: FGLS and FOGLeSs results

	Set 1	Set 2	Set 3	Set 4	Set 5	FOGLeSs
Estimate	-0.0286	-0.0327	-0.0210	-0.0286	-0.0291	-0.0292
Robust standard errors	0.0083	0.0075	0.0091	0.0078	0.0096	0.0062
Asymptotic $t-$statistic	-3.4481	-4.3603	-2.2930	-3.6740	-3.0148	-4.6557
FOGLeSs weight	0.1855	0.3873	0.1319	0.1897	0.1057	-

Finally, *testrob* terminates.

This formally demonstrates that the results of Pérez-González (2006) are robust to different sets of covariates. The FOGLeSs estimate of the coefficient for the critical core variable, Family CEO, is -0.0292 with a $t-$statistic of -4.6557.

11 Summary and Concluding Remarks

Although robustness checks are common in applied economics, their use is subject to numerous pitfalls. If not implemented properly, they may be uninformative; at worst, they can be entirely misleading. Here, we discuss these pitfalls and provide straightforward methods that preserve the diagnostic spirit underlying robustness checks. We distinguish between critical and non-critical core variables, and we discuss how these can be properly specified. We also discuss how non-core variables for the comparison regressions can be chosen to ensure that robustness checks are indeed structurally informative.

Our formal robustness test is a Hausman-style specification test. We supplement this with diagnostics for nonlinearity and exogeneity that can help in understanding why robustness test rejection occurs or in identifying invalid sets of covariates. The Feasible Optimally combined GLS (FOGLeSs) estimator provides a relatively efficient combination of the estimators compared in performing the robustness test. A new Matlab procedure, *testrob*, freely available at `www.econ.ucsd.edu/~xunlu/code` embodies these methods. Our hope is that the ready availability of this procedure will encourage researchers to subject their analyses to more informative methods for identifying potential misspecification, strengthening the validity and reliability of structural inference in economics.

12 Appendix I: Equality of β_1^* and β_2^* Inconsistent for β_o

Consider vectors X_1 and $X_2 = (X_1, W)$, where X_1 contains Z. Suppose that $D \not\perp U \mid X_1$ and $D \not\perp U \mid X_2$, so that both X_1 and X_2 are invalid covariates. Further, suppose $W \perp D \mid X_1$. Letting conditional expectations be linear for simplicity, we have

$$\begin{aligned} E(Y|D, X_2) &= D'\beta_o + Z'\alpha_o + E(U \mid D, X_1, W) \\ &= D'\beta_o + Z'\alpha_o + D'\zeta_d^* + X_1'\zeta_{x_1}^* + W'\zeta_w^* \\ &= D'\left(\beta_o + \zeta_d^*\right) + Z'\alpha_o + X_1'\zeta_{x_1}^* + W'\zeta_w^* \end{aligned}$$

$$\begin{aligned} E(Y|D, X_1) &= E\left[E(Y|\, D, X_2)|\, D, X_1\right] \\ &= D'\left(\beta_o + \zeta_d^*\right) + Z'\alpha_o + X_1'\zeta_{x_1}^* + E\left[W'\zeta_w^*\middle|\, D, X_1\right] \\ &= D'\left(\beta_o + \zeta_d^*\right) + Z'\alpha_o + X_1'\zeta_{x_1}^* + X_1'\delta_{x_1}^*\zeta_w^* \\ &= D'\left(\beta_o + \zeta_d^*\right) + Z'\alpha_o + X_1'\left(\zeta_{x_1}^* + \delta_{x_1}^*\zeta_w^*\right) \end{aligned}$$

This implies robustness of the D coefficients, despite the invalidity of both X_1 and X_2.

13 Appendix II: Choosing Valid Covariates

To examine the how valid covariates W can be chosen, we invoke Reichenbach's (1956) principle of common cause: if two variables are correlated[6], then one causes the other or there is a

[6] For succinctness in what follows, we use "correlation" loosely to mean any stochastic dependence.

third common cause of both. Chalak and White (2008) give formal conditions, applicable here, ensuring the validity of this principle. Here, we want X correlated with U and/or with D, so we focus on the properties of W that may enhance this correlation enough to yield $D \perp U \mid X$. What follows summarizes a more extensive discussion in White (2006b).

First, consider the relation between W and U. If these are correlated, then W causes U, U causes W, or there is an underlying common cause. Each possibility is useful in suggesting choices for W; however, these possibilities can be simplified. ($a.i$) If W is a cause of U, then by substitution, it is also a cause of Y, typically together with some other unobservables, say V. Now W can be considered a component of Z, and the issue is whether or not $D \perp V \mid Z$. This takes us back to the original situation. Thus, the possibility of substitution ensures that, without loss of generality, we can adopt the convention that W does not cause U. ($a.ii$) If U is a cause of W, then W acts as a proxy for U, as in the classic case where IQ score acts as a proxy for unobserved ability. Note that W's role is purely predictive for U, as W is not a structural factor determining Y. ($a.iii$) If there is an observable common cause of U and W, then by substitution we can include it in Z, just as in case ($a.i$). W becomes redundant in this case. If there is an *unobservable* common cause of U and W, say V, then by substitution we arrive back at case ($a.ii$). With these conventions, the single relevant case is that W is a proxy for U, caused by U. Such proxies are a main source of covariates W. An interesting possibility in time-series applications is that W may contain not only lags but also *leads* of variables driven by U, as discussed by White and Kennedy (2009) and White and Lu (2010a).

Next, consider the relation between W and D. If these are correlated, then W causes D, D causes W, or there is an underlying common cause. Again, the possibilities simplify. ($b.i$) The causes W of D are another key source of potential covariates. ($b.ii$) The case where D causes W is problematic. As discussed by Rosenbaum (1984) and by Heckman and Navarro-Lozano (2004), including regressors (W) driven by the cause of interest (D) in a regression with D generally leads to inconsistent estimates of the effects of interest. Such regressions are not an appropriate basis for examining robustness, so we rule out variables caused by D as covariates. ($b.iii$) If there is an observable common cause of D and W, then, by substitution, that common cause is a cause of D, and we are back to case ($b.i$) by replacing the original W with its observable cause in common with D. If there is an *unobservable* common cause of D and W, say V, then W acts as a proxy for V, analogous to case ($a.ii$).

Summarizing, we see that W can either contain proxies for U, observed drivers of D, or proxies for unobserved drivers of D. It should not contain outcomes driven by D.

14 Appendix III: *testrob* Results for the Example of Section 10

Welcome to testrob (c) 2010 Halbert White and Xun Lu.

When reporting results from this procedure, please cite:

White, H. and X. Lu (2010), "Robustness Checks and Robustness Tests in Applied Economics,"

UCSD Department of Economics Discussion Paper.

Please enter your dependent variable Y: Y

Please enter your critical core covariates D: D

Please enter your structurally relevant core covariates Z: []

Please enter ALL other covariates W: W

The next step is to choose the initial core covariates W_ 0.

Please use the column numbers of W to indicate the columns you choose.

For example, enter [1 3 5] to indicate that columns 1, 3, and 5 of W are the initial core covariates.

Please enter the initial core covariates from W: [1:6 16:34]

The next step is to choose groups of non-core covariates.

How many groups of non-core covariates? 3

Next, for each group, please use the column numbers of W to indicate the columns you choose.

For example, enter [1 3 5] to indicate that columns 1, 3 and 5 of W are the first non-core covariates.

For each group, please also indicate how to use the group.

For example, suppose the group is [1 3 5].

Enter '0' to indicate using the whole group (i.e. [1 3 5]).

Enter '1' to indicate using the whole group plus all subsets of the group with 1 element

(i.e. [1 3 5] [1] [3] [5]).

Enter '-1' to indicate using the whole group plus all subsets of the group with 3-1=2 elements

(i.e. [1 3 5] [1 3] [3 5] [1 5]).

Enter '2' to indicate using the whole group plus all subsets of the group with 1 or 2 elements

(i.e. [1 3 5] [1] [3] [5] [1 3] [3 5] [1 5]).

Enter '-2' to indicate using the whole group plus all subsets of the group with 3-1=2 or 3-2=1 elements

(i.e. [1 3 5] [1 3] [3 5] [1 5] [1] [3] [5]).

Please enter the index of non-core covariates for Group 1: [7:10]

Please indicate how to use the group: 0

Please enter the index of non-core covariates for Group 2: [11]

Please indicate how to use the group: 0

Please enter the index of non-core covariates for Group 3: [12:15]

Please indicate how to use the group: 0

The next step is to identify core covariates from the initial core covariates W_0.

Do you want to declare that the initial core covariates contain no non-core covariates?

Please enter 1 for 'Yes' or 0 for 'No': 0

Testrob will regress each element of D on (Z,W_0).

For which elements of (Z,W_0), do you want to see the p-values?

Please press 'enter' to see all the elements. Otherwise enter the column numbers: [1:6]

Here are the results of regressing each element of D on (Z,W_0):

column number	3	1	6	4	2	5
p-values	0.97289	0.8847	0.76559	0.47161	0.33388	0

(High p-values indicate non-core covariates.)

Do you want to redeclare the core covariates?

Please enter 1 for 'Yes' or 0 for 'No': 1

Please enter the column number of the new core covariates: [2 4 5 16:34]

The remaining elements from the initial core covariates are used to form a new group W_0_star for robustness testing.

Please indicate how to use the non-core covariates W_0_star: 0

Testob has performed a singularity check.

The following sets of variables are retained for robustness testing:

Set 1:

Set 2: 1 3 6

Set 3: 7 8 9 10

Set 4: 11

Set 5: 12 13 14 15

The testrob robustness test p-value is: 0.76821

The p-value of the linearity test for Set 1 is : 0.63531

The p-value of the conditional exogeneity test for Set 1 is : 0.248

The p-value of the linearity test for Set 2 is : 0.75629

The p-value of the conditional exogeneity test for Set 2 is : 0.885

The p-value of the linearity test for Set 3 is : 0.64388

The p-value of the conditional exogeneity test for Set 3 is : 0.958

The p-value of the linearity test for Set 4 is : 0.45134

The p-value of the conditional exogeneity test for Set 4 is : 0.551

The p-value of the linearity test for Set 5 is : 0.27894

The p-value of the conditional exogeneity test for Set 5 is : 0.651

Do you want to continue testrob? Please enter 1 for 'Yes' and 0 for 'No': 1

Do you want to drop any of the 5 sets of covariates above?

Please enter 1 for 'Yes' or 0 for 'No': 0

Feasible GLS results for the critical core variables using Set 1:

GLS estimate	-0.028622
GLS robust standard error	0.0083009
t-stat	-3.4481

Feasible GLS results for the critical core variables using Set 2:

GLS estimate	-0.032672
GLS robust standard error	0.0074929
t-stat	-4.3603

Feasible GLS results for the critical core variables using Set 3:

GLS estimate	-0.020954
GLS robust standard error	0.0091384
t-stat	-2.293

Feasible GLS results for the critical core variables using Set 4:

GLS estimate	-0.028627
GLS robust standard error	0.0077918
t-stat	-3.674

Feasible GLS results for the critical core variables using Set 5:

GLS estimate	-0.029084
GLS robust standard error	0.0096473
t-stat	-3.0148

FOGLeSs estimation results for the critical core variables:

FOGLeSs estimate	-0.029229
standard error	0.0062782
t-stat	-4.6557

The coefficient weights for FOGLeSs:

set number	1	2	3	4	5
weight	0.18547	0.38728	0.13186	0.18973	0.10566

Do you want to see the results displayed in tables?

Please enter 1 for 'Yes' or 0 for 'No': 0

End of testrob.

References

[1] Adams, William, Liran Einav, and Jonathan Levin. 2009. "Liquidity Constraints and Imperfect Information in Subprime Lending." American Economic Review 99:49-84.

[2] Alfaro, Laura, and Andrew Charlton. 2009. "Intra-Industry Foreign Direct Investment." American Economic Review 99:2096-2119.

[3] Andrews, Donald. 1994. "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity." Econometrica 62:43-72

[4] Angelucci, Manuela, and Giacomo De Giorgi. 2009. "Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption?" American Economic Review 99:486-508.

[5] Angrist, Joshua, and Victor Lavy. 2009. "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial." American Economic Review 99:1384-1414.

[6] Ashraf, Nava. 2009. "Spousal Control and Intra-Household Decision Making: An Experimental Study in the Philippines." American Economic Review 99:1245-1277.

[7] Barron's Educational Series, Inc. 1980. Barron's Profiles of American Colleges. Hauppauge: Barron's Educational Series, Inc.

[8] Boivin, Jean, Marc P Giannoni, and Ilian Mihov. 2009. "Sticky Prices and Monetary Policy: Evidence from Disaggregated US Data." American Economic Review 99:350-384.

[9] Cai, Hongbin, Yuyu Chen, and Hanming Fang. 2009. "Observational Learning: Evidence from a Randomized Natural Field Experiment." American Economic Review 99:864-882.

[10] Chalak, Karim, and Halbert White. 2008. "Independence and Conditional Independence in Structural Systems." UCSD Department of Economics Discussion Paper.

[11] Chalak, Karim, and Halbert White. 2009. "An Extended Class of Instrumental Variables for the Estimation of Causal Effects." UCSD Department of Economics Discussion Paper.

[12] Chen, Yan, and Sherry Xin Li. 2009. "Group Identity and Social Preferences." American Economic Review 99:431-457.

[13] Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." American Economic Review 99:1145-1177.

[14] Cleveland, William S. 1979. "Robust Locally Weighted Regression and Smoothing Scatterplots." Journal of the American Statistical Association 74:829-836.

[15] Dahl, Christian M., and Gloria González-Rivera. 2003. "Testing for Neglected Nonlinearity in Regression Models Based on the Theory of Random Fields." Journal of Econometrics 114:141-164.

[16] Dawid, A. P. 1979. "Conditional Independence in Statistical Theory." Journal of the Royal Statistical Society. Series B (Methodological) 41:1-31.

[17] Dobkin, Carlos, and Nancy Nicosia. 2009. "The War on Drugs: Methamphetamine, Public Health, and Crime." American Economic Review 99:324-349.

[18] Engle, Robert F., C. W. J. Granger, John Rice, and Andrew Weiss. 1986. “Semiparametric Estimates of the Relation Between Weather and Electricity Sales.” Journal of the American Statistical Association 81:310-320.

[19] Forbes, Silke Januszewski, and Mara Lederman. 2009. “Adaptation and Vertical Integration in the Airline Industry.” American Economic Review 99:1831-1849.

[20] Gallant, A. Ronald. 1982. “Unbiased Determination of Production Technologies.” Journal of Econometrics 20:285-323.

[21] Griliches, Zvi. 1977. “Estimating the Returns to Schooling: Some Econometric Problems.” Econometrica 45:1-22.

[22] Hahn, Jinyong. 2004. “Functional Restriction and Efficiency in Causal Inference.” The Review of Economics and Statistics 86:73-76.

[23] Hamilton, James D. 2001. “A Parametric Approach to Flexible Nonlinear Inference.” Econometrica 69:537-573.

[24] Hargreaves Heap, Shaun P, and Daniel John Zizzo. 2009. “The Value of Groups.” American Economic Review 99:295-323.

[25] Hausman, J. A. 1978. “Specification Tests in Econometrics.” Econometrica 46:1251-1271.

[26] Heckman, James J., and V. Joseph Hotz. 1989. “Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training.” Journal of the American Statistical Association 84:862-874.

[27] Heckman, James, and Salvador Navarro-Lozano. 2004. “Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models.” The Review of Economics and Statistics 86:30-57.

[28] Hendel, Igal, Aviv Nevo, and François Ortalo-Magné. 2009. “The Relative Performance of Real Estate Marketing Platforms: MLS versus FSBOMadison.com.” American Economic Review 99:1878-1898.

[29] Imbens, Guido W. 2004. “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review.” Review of Economics and Statistics 86:4-29.

[30] Lavy, Victor. 2009. “Performance Pay and Teachers’ Effort, Productivity, and Grading Ethics.” American Economic Review 99:1979-2011.

[31] Leamer, Edward E. 1983. “Let’s Take the Con Out of Econometrics.” American Economic Review 73:31-43.

[32] Leaver, Clare. 2009. “Bureaucratic Minimal Squawk Behavior: Theory and Evidence from Regulatory Agencies.” American Economic Review 99:572-607.

[33] Lee, Tae-Hwy, Halbert White, and Clive W. J. Granger. 1993. “Testing for Neglected Nonlinearity in Time Series Models : A Comparison of Neural Network Methods and Alternative Tests.” Journal of Econometrics 56:269-290.

[34] Makowsky, Michael D, and Thomas Stratmann. 2009. “Political Economy at Any Speed: What Determines Traffic Citations?” American Economic Review 99:509-527.

[35] Mas, Alexandre, and Enrico Moretti. 2009. “Peers at Work.” American Economic Review 99:112-145.

[36] Matsusaka, John G. 2009. “Direct Democracy and Public Employees.” American Economic Review 99:2227-2246.

[37] Miller, Nathan H. 2009. “Strategic Leniency and Cartel Enforcement.” American Economic Review 99:750-768.

[38] Oberholzer-Gee, Felix, and Joel Waldfogel. 2009. “Media Markets and Localism: Does Local News en Español Boost Hispanic Voter Turnout?” American Economic Review 99:2120-2128.

[39] Pérez-González, Francisco. 2006. “Inherited Control and Firm Performance.” American Economic Review 96:1559-1588.

[40] Ramsey, J. B. 1969. “Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis.” Journal of the Royal Statistical Society. Series B (Methodological) 31:350-371.

[41] Reichenbach, Hans. 1956. The Direction of Time. Berkeley: University of California Press..

[42] Robinson, P. M. 1987. “Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form.” Econometrica 55:875-891.

[43] Rosenbaum, Paul R. 1984. “The Consequences of Adjusting for a Concomitant Variable That Has Been Affected by Treatment,” Journal of the Royal Statistical Society, Series A 147:656-666.

[44] Rosenbaum, Paul R. 1987. “The Role of a Second Control Group in an Observational Study.” Statistical Science 2:292-306.

[45] Ruppert, D., and M. P. Wand. 1994. “Multivariate Locally Weighted Least Squares Regression.” The Annals of Statistics 22:1346-1370.

[46] Sialm, Clemens. 2009. “Tax Changes and Asset Pricing.” American Economic Review 99:1356-1383.

[47] Spilimbergo, Antonio. 2009. “Democracy and Foreign Education.” American Economic Review 99:528-543.

[48] Stock, James H., and Mark W. Watson. 2007. Introduction to Econometrics. Boston: Addison-Wesley.

[49] Su, Liangjun, and Halbert White. 2010. “Testing Structural Change in Partially Linear Models.” Econometric Theory (forthcoming).

[50] Urquiola, Miguel, and Eric Verhoogen. 2009. “Class-Size Caps, Sorting, and the Regression-Discontinuity Design.” American Economic Review 99:179-215.

[51] White, Halbert. 1980. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," Econometrica 48: 817-838

[52] White, Halbert. 2006(a). “Approximate Nonlinear Forecasting Methods,” in G Elliott, CWJ Granger, and A Timmermann, eds, Handbook of Economics Forecasting. New York: Elsevier, 460-512.

[53] White, Halbert. 2006(b). “Time-series Estimation of the Effects of Natural Experiments.” Journal of Econometrics 135:527-566.

[54] White, Halbert and Karim Chalak. 2009. “Identifying Effects of Endogenous Causes in Nonseparable Systems Using Covariates,” UCSD Department of Economics Discussion Paper.

[55] White, Halbert, and Karim Chalak. 2010. “Testing a Conditional Form of Exogeneity.” UCSD Department of Economics Discussion Paper.

[56] White, Halbert, and Pauline Kennedy. 2009. “Retrospective Estimation of Causal Effects Through Time,” in J. Castle and N. Shephard, eds., The Methodology and Practice of Econometrics: A Festschrift in Honour of David Hendry. Oxford: Oxford University Press, 59-87.

[57] White, Halbert, and Xun Lu. 2010(a). "Granger Causality and Dynamic Structural Systems." Journal of Financial Econometrics 8:193-243.

[58] White, Halbert, and Xun Lu. 2010(b). "Causal Diagrams for Treatment Effect Estimation with Application to Efficient Covariate Selection." UCSD Department of Economics Discussion Paper.

[59] White, Halbert, and Maxwell Stinchcombe. 1991. "Adaptive Efficient Weighted Least Squares with Dependent Observations." in W Stahel and S Weisberg, eds, Directions in Robust Statistics and Diagnostics, IMA Volumes in Mathematics and Its Applications. New York: Springer-Verlag, 337-364.

[60] Wooldridge, Jeffrey M. 2002. Econometric Analysis of Cross Section and Panel Data. Cambridge: MIT Press.