Poorly Measured Confounders are More Useful on the Left Than on the Right

Jörn-Steffen Pischke
LSE

Hannes Schwandt*
Princeton University

October 2014

Abstract

Researchers frequently test identifying assumptions in regression based research designs (which include e.g. instrumental variables or differences-in-differences models) by adding additional control variables on the right hand side of the regression. If such additions do not affect the coefficient of interest (much) a study is presumed to be reliable. We caution that such invariance may result from the fact that many observed variables are poor measures of the potential underlying confounders. In this case, a more powerful test of the identifying assumption is to put the variable on the left hand side of the candidate regression. We provide relevant derivations for the estimators and test statistics involved, as well as power calculations, which can help applied researchers to interpret their findings. We illustrate these results in the context of various strategies which have been suggested to identify the returns to schooling.

*This paper builds on ideas in our paper "A Cautionary Note on Using Industry Affiliation to Predict Income," NBER WP18384. We thank Alberto Abadie, Josh Angrist, and Brigham Frandsen for helpful comments.

1 Introduction

Research on causal effects depends on implicit identifying assumptions, which typically form the core of a debate about the quality and credibility of a particular research design. In regression or matching based strategies, this is the claim that variation in the regressor of interest is as good as random after conditioning on a sufficient set of control variables. In instrumental variables models it is the exclusion restriction. In panel or differences-in-differences designs it is the parallel trends assumption, possibly after suitable conditioning. The credibility of a design can be enhanced when researchers can show explicitly that potentially remaining sources of selection bias have been eliminated. This is often done through some form of balancing or falsification tests.

The research designs mentioned above can all be thought of as variants of regression strategies. If the researcher has access to a candidate confounder, tests for the identifying assumption take two canonical forms. The confounder can be added as a control variable on the right hand side of the regression. The identifying assumption is confirmed if the estimated causal effect of interest is insensitive to this variable addition–we call this the coefficient comparison test. Alternatively, the candidate confounder can be placed on the left hand side of the regression instead of the outcome variable. A zero coefficient on the causal variable of interest then confirms the identifying assumption. This is analogous to the balancing test typically carried out using baseline characteristics or pre-treatment outcomes in a randomized trial, and frequently used in regression discontinuity designs.

Researchers often rely on one or the other of these tests. We argue that the balancing test, using the candidate confounder on the left hand side of the regression, is generally more powerful. This is particularly the case when the available variable is a noisy measure of the true underlying confounder. The attenuation due to measurement error often implies that adding the candidate variable on the right hand side as a regressor does little to eliminate any omitted variables bias. The same measurement error does comparatively less damage when putting this variable on the left hand side. Regression strategies work well in finding small but relevant amounts of variation in noisy dependent variables.

These two testing strategies are intimately related through the omitted variables bias formula. The omitted variables bias formula shows that the coefficient comparison test involves two regression parameters, while the balancing test only involves one of these two. If the researcher has a strong prior that the added regressor ought to matter for the outcome under study then the balancing test will provide the remaining information necessary to assess the research design. This maintained assumption is the ultimate source of the superior power of the balancing test. However, we show that meaningful differences emerge only when there is some substantial amount of measurement error in the added regressor in practice. We derive the biases in the relevant estimators in Section 3.

A second point we are making is that the two strategies both lead to explicit statistical tests. The balancing test is a simple t-test used routinely by researchers. When adding a covariate on the right hand side, comparing the coefficient of interest across the two regressions can be done using a generalized Hausman test. In practice, we haven't seen this test carried out in applied papers, where researchers typically just eye-ball the results. We provide the relevant test statistics and discuss how they behave under measurement error in Section 4. We also show how this test is simple to implement for varying identification strategies. We demonstrate the superior power of the balancing test under a variety of scenarios in Section 4.2.

While we stress the virtues of the balancing test, this does not mean that the explicit coefficient comparison is without value, even when the added regressor is measured with error. Suppose the added candidate regressor, measured correctly, is the only confounder. In this case, the same model with classical measurement error in this variable is under-identified by one parameter. The reliability ratio of the mismeasured variable is a natural metric for this missing parameter, and we show how researchers can point identify the parameter of interest with assumptions about the measurement error.

The same result can be used to place bounds on the parameter of interest using ranges of the measurement error, or the amount of measurement error necessary for a zero effect can be obtained. We feel that this is a simple and useful way of combing the results from the three regressions underlying all the estimation and testing here in terms of a single metric.

The principles underlying the points we are making are not new but the consequences do not seem to be fully appreciated in much applied work. Griliches (1977) is a classic reference for the issues arising when regression controls are measured with error. Like us, Griliches' discussion is framed around the omitted variables bias arising in linear regressions, the general framework used most widely in empirical studies.

This explicit measurement error scenario has played relatively little role in subsequent discussions of omitted variables bias. Rosenbaum and Rubin (1983) discuss adjustments for an unobserved omitted variable in a framework with a binary outcome, binary treatment, and binary covariate by making assumptions about the parameters in the relevant omitted variables bias formula. Imbens (2003) extends this analysis by transforming the unknown parameters to partial correlations rather than regression coefficients, which he finds more intuitive. This strand of analysis needs assumptions about two unknown parameters because no information about the omitted regressor is available. We assume that the researcher has a noisy measure available, which is enough to identify one of the missing parameters.

Battistin and Chesher (2014) is closely related as it discusses identification in the presence of a mismeasured covariate. They go beyond our analysis in focusing on measurement error in non-linear models. Like in the literature following Rosenbaum and Rubin (1983) they discuss identification given assumptions about the missing parameter, namely the degree of measurement error in the covariate. While we show these results for the linear case, we go beyond the Battistin and Chesher analysis in our discussion of testing. This should be of interest to applied researchers who would like to show that they cannot reject identification in their sample given a measure for a candidate confounder.

Altonji, Elder and Taber (2005) discuss an alternative but closely related approach to the problem. Applied researchers have often argued that relative stability of regression coefficients when adding additional controls provides evidence for credible identification. Implicit in this argument is the idea that other confounders not controlled for are similar to the controls just added to the regression. The paper by Altonji, Elder and Taber (2005) formalizes this argument. In practice, adding controls will typically move the coefficient of interest somewhat if it is not much. Oster (2014) extends the Altonji, Elder and Taber work by providing precise conditions for point identification in this case. The approach in these papers relies on an assumption about how the omitted variables bias due to the observed regressor is related to omitted variables bias due to unobserved confounders.

The unobserved confounder in this previous work can be thought of as the source of measurement error in the covariate which is added to the regression in our analysis. For example, in our empirical example below, we use mother's education as a measure for family background but this variable may only capture a small part of all the relevant family background information, a lot of which may be orthogonal to mother's education. Since our discussion of inference and testing covers this case, our framework is a useful starting point for researchers who are willing to make the assumptions in Altonji, Elder and Taber (2005) or Oster (2014).

Griliches (1977) uses estimates of the returns to schooling, which have formed a staple of labor economics ever since, as example for the methodological points he makes. We use Griliches' data from the National Longitudinal Survey of Young Men to illustrate our results in section 5. In addition to Griliches (1977) this data set has been used in a well known study by Card (1995). It is well suited for our purposes because the data contain various test score measures which can be used as controls in a regression strategy (as investigated by Griliches, 1977), a candidate instrument for college attendance (investigated by Card, 1995), as well as a myriad of other useful variables on individual and family background. The empirical results support our theoretical claims.

2 A Simple Framework

Consider the following simple framework starting with a regression equation

$$y_i = \alpha^s + \beta^s s_i + e_i^s \tag{1}$$

where y_i is an outcome like log wages, s_i is the causal variable of interest, like years of schooling, and e_i is a regression residual. The researcher proposes this short regression model to be causal. This might be the case because the data come from a randomized experiment, so the simple bivariate regression is all we need. More likely, the researcher has a particular research design applied to observational data. For example, in the case of a regression strategy controlling for confounders, y_i and s_i would be residuals from regressions of the original variables on the chosen controls. In the case of panel data or differences-in-differences designs the controls are sets of fixed effects. In the case of instrumental variables s_i would be the predicted values from a first stage regression. In practice, (1) encompasses a wide variety of empirical approaches.

Now consider the possibility that the estimate β^s from (1) may not actually be causal. There may be a candidate confounder x_i, so that the causal effect of s_i on y_i would only be obtained conditional on x_i, as in the long regression

$$y_i = \alpha + \beta s_i + \gamma x_i + e_i \tag{2}$$

and the researcher would like to probe whether this is a concern. For example, in the returns to schooling context, x_i might be some remaining part of an individual's earnings capacity which is also related to schooling, like ability and family background.

Researchers who find themselves in a situation where they start with a proposed causal model (1) and a measure for a candidate confounder x_i typically do one of two things: They either regress x_i on s_i and check whether s_i is significant, or they include x_i on the right hand side of the original regression, and check whether the estimate of β changes materially when x_i is added to the regression of interest. The first strategy constitutes a test for "balance," a standard check for successful randomization in an experiment. In principle, the second strategy has the advantage that it goes beyond testing whether (1) qualifies as a causal regression. If β changes appreciably this suggests that the original estimate β^s is biased. However, the results obtained with x_i as an additional control should be closer to the causal effect we seek to uncover. In particular, if x_i were the only relevant confounder and if we measure it without error, the β estimate from the controlled regression is the causal effect of interest. In practice, there is usually little reason to believe that these two conditions are met, and hence a difference between β and β^s again only indicates a flawed research design.

The relationship between these two strategies is easy to see. Write the regression of x_i on s_i, which we will call the balancing regression, as

$$x_i = \delta_0 + \delta s_i + u_i. \tag{3}$$

The change in the coefficient β from adding x_i to the regression (1) is given by the omitted variables bias formula

$$\beta^s - \beta = \gamma\delta \tag{4}$$

where δ is the coefficient on s_i in the balancing regression. The change in the coefficient of interest β from adding x_i consists of two components, the coefficient γ on x_i in (2) and the coefficient δ from the balancing regression.

Here we consider the relationship between the two approaches: the balancing test, consisting of an investigation of the null hypothesis

$$H_0 : \delta = 0, \tag{5}$$

compared to the inspection of the coefficient movement $\beta^s - \beta$. The latter strategy of comparing β^s and β is often done informally but it can be formalized as a statistical test of the null hypothesis

$$H_0 : \beta^s - \beta = 0, \tag{6}$$

which we will call the coefficient comparison test. We have not seen this carried out explicitly in applied research. From (4) it is clear that (6) amounts to

$$H_0 : \beta^s - \beta = 0 \Leftrightarrow \gamma = 0 \text{ or } \delta = 0.$$

This highlights that the two approaches formally test the same hypothesis under the maintained assumption $\gamma \neq 0$. We may often have a strong sense that $\gamma \neq 0$; i.e. we are dealing with a variable x_i which we believe affects the outcome, but we are unsure whether it is related to the regressor of interest s_i. In this case, both tests would seem equally suitable. Nevertheless, in other cases γ may be zero, or we may be unsure. In this case the coefficient comparison test seems to dominate because it directly addresses the question we are after, namely whether the coefficient of interest β is affected by the inclusion of x_i in the regression.

Here we make the point that the balancing test adds valuable information particularly when the true confounder is measured with error. In general, x_i may not be easy to measure. If the available measure for x_i contains classical measurement error, the estimate of γ in (2) will be attenuated, and the comparison $\beta^s - \beta$ will be too small (in absolute value) as a result. The estimate of δ from the balancing regression is still unbiased in the presence of measurement error; this regression simply loses precision because the mismeasured variable is on the left hand side. Under the maintained assumption that $0 < \gamma < \infty$, the balancing test is more powerful than the coefficient comparison test. In order to make these statements precise, we collect results for the estimators of the relevant parameters and test statistics for the case of classical measurement error in the following section.

3 Estimators in the Presence of Measurement Error

The candidate variable x_i is not observed. Instead, the researcher works with the mismeasured variable

$$x_i^m = x_i + m_i. \tag{7}$$

The measurement error m_i is classical, i.e. $E(m_i) = 0$, $Cov(x_i, m_i) = 0$. As a result, the researcher compares the regressions

$$\begin{aligned} y_i &= \alpha^s + \beta^s s_i + e_i^s \\ y_i &= \alpha^m + \beta^m s_i + \gamma^m x_i^m + e_i^m. \end{aligned} \tag{8}$$

Notice that the short regression does not involve the mismeasured x_i, so that $\beta^s = \beta + \gamma\delta$ as before. However, the coefficients β^m and γ^m are biased now and are related to the coefficients from (2) in the following way:

$$\begin{aligned} \beta^m &= \beta + \gamma\delta\frac{1-\lambda}{1-R^2} = \beta + \gamma\delta\theta \qquad (9) \\ \gamma^m &= \gamma\frac{\lambda - R^2}{1-R^2} = \gamma(1-\theta) \end{aligned}$$

where R^2 is the R^2 of the regression of s_i on x_i^m and

$$\lambda = \frac{Var(x_i)}{Var(x_i^m)}$$

is the reliability of x_i^m. It measures the amount of measurement error present as the fraction of the variance in the observed x_i^m, which is due to the signal in the true x_i. λ is also the attnuation factor in a simple bivariate regression on x_i^m. An alternative way to parameterize the amount of measurement error is

$$\theta = \frac{1-\lambda}{1-R^2} = \frac{\sigma_m^2}{\sigma_u^2 + \sigma_m^2}.$$

$1 - \theta$ is the multivariate attenuation factor. Recall that u_i is the residual from the balancing regression (3).

Notice that

$$x_i^m = \delta_0^m + \delta^m s_i + \ u_i + m_i, \tag{10}$$

and hence $\lambda > R^2$. As a result

$$\begin{aligned} 0 &< \frac{1-\lambda}{1-R^2} < 1 \\ 0 &< \frac{\lambda - R^2}{1-R^2} < \lambda. \end{aligned}$$

θ is an alternative way to parameterize the degree of measurement error in x_i compared to λ and R^2. The θ parameterization uses only the variation in

x_i^m which is orthogonal to s_i. This is the part of the variation in x_i^m which is relevant to the estimate of γ^m in regression (8), which also has s_i as a regressor. θ turns out to be a useful parameter in many of the derivations that follow.

The coefficient β^m is biased but less so than β^s. In fact, β^m lies between β^s and β. The estimate γ^m is attenuated; the attenuation is bigger than in the case of a bivariate regression of y_i on x_i^m without the regressor s_i if x_i^m and s_i are correlated ($R^2 > 0$).

These results highlight a number of issues. The gap $\beta^s - \beta^m$ is too small compared to the desired $\beta^s - \beta$, directly affecting the coefficient comparison test. In addition, γ^m is biased towards zero. Ceteris paribus, this is making the assessment of the hypothesis $\gamma = 0$ more difficult. Finally, the balancing regression (10) with the mismeasured x_i^m involves measurement error in the dependent variable and therefore no bias in the estimate of δ^m, i.e. $\delta^m = \delta$, but simply a loss of precision.

The results here can be used to think about identification of β in the presence of measurement error. Rearranging (9) yields

$$\begin{aligned} \gamma &= \gamma^m \frac{1 - R^2}{\lambda - R^2} \\ \beta &= \beta^m - \delta\gamma^m \frac{1 - \lambda}{\lambda - R^2}. \end{aligned} \tag{11}$$

Since R^2 is observed from the data this only involves the unknown parameter λ. If we are willing to make an assumption about the measurement error we are able to point identify β. Even if λ is not known precisely, (11) can be used to bound β for a range of plausible reliabilities. Alternatively, (9) can be used to derive the value of λ for which $\beta = 0$. These calculations are similar in spirit to the ones suggested by Oster (2014) in her setting.

4 Inference

In this section, we consider how conventional standard errors and test statistics for the quantities of interest are affected in the homoskedastic case.[1] We present the theoretical power functions for the two alternative test statistics; derivations are in the appendix. The power results are extended to the heteroskedastic case and non-classical measurement error in simulations. Our basic conclusions are very robust in all these different scenarios.

Start with the coefficient $\widehat{\delta}^m$ from the balancing regression:

$$\begin{aligned} se\left(\widehat{\delta}^m\right) &= \frac{1}{\sqrt{n}} \frac{\sqrt{\sigma_u^2 + \sigma_m^2}}{\sigma_s} \\ &= \frac{\sigma_u}{\sqrt{n}\sigma_s\sqrt{1-\theta}} \\ &= \frac{se\left(\widehat{\delta}\right)}{\sqrt{1-\theta}}. \end{aligned}$$

It is easy to see that the standard error is inflated compared to the case with no measurement error. The t-test

$$t_{\delta^m} = \frac{\widehat{\delta}^m}{se\left(\widehat{\delta}^m\right)}$$

remains consistent because m_i is correctly accounted for in the residual of the balancing regression (10), but the t-statistic is smaller than in the error free case. This means the null hypothesis (5) is rejected less often. The test is less powerful than in the error free case.

We next turn to $\widehat{\gamma}^m$, the coefficient on the mismeasured x_i^m in (8). The estimate of γ is of interest since it determines the coefficient movement $\beta^s - \beta = \gamma\delta$ in conjunction with the result from the balancing regression. The

[1] See the appendix for the precise setup of the model. The primitive disturbances are s_i, u_i, e_i, and m_i, which we assume to be uncorrelated with each other. Other variables are determined by (3), (2), and (7).

standard error for $\widehat{\gamma}^m$ is

$$\begin{aligned} se\left(\widehat{\gamma}^m\right) &= \frac{1}{\sqrt{n}}\frac{\sqrt{Var\left(e_i^m\right)}}{\sqrt{Var\left(\widetilde{x}_i^m\right)}} \\ &= \frac{1}{\sqrt{n}}\sqrt{\frac{\gamma^2\theta\sigma_u^2+\sigma_e^2}{\sigma_u^2+\sigma_m^2}} \\ &= \frac{\sqrt{1-\theta}}{\sqrt{n}}\left(\left|\gamma\right|\sqrt{\theta}+\frac{\sigma_e}{\sigma_u}\right) \\ &= \sqrt{1-\theta}se\left(\widehat{\gamma}\right)+\frac{\sqrt{\left(1-\theta\right)\theta}}{\sqrt{n}}\left|\gamma\right|. \end{aligned}$$

The standard error for $\widehat{\gamma}^m$ involves two terms: the first term is an attenuated version of the standard error for $\widehat{\gamma}$ from the corresponding regression with the correctly measured x_i, while the second term depends on the value of γ. The parameters in the two terms are not directly related, so $se\left(\widehat{\gamma}^m\right) \gtrless se\left(\widehat{\gamma}\right)$. Measurement error does not necessarily inflate the standard error here.

The two terms have a simple, intuitive interpretation. Measurement error biases the coefficient γ^m towards zero, the attenuation factor is $1-\theta$. The standard error is attenuated in the same direction; this is the first term $\sqrt{1-\theta}se\left(\widehat{\gamma}\right)$. The second term $\left|\gamma\right|\sqrt{\left(1-\theta\right)\theta}/\sqrt{n}$ is related to the fact that the residual variance $Var\left(e_i^m\right)$ is larger when there is measurement error. The increase in the variance is related to the true γ, which enters the residual. But attenuation matters here as well, so this term is inverse U-shaped in θ and is greatest when $\theta = 0.5$.

The t-statistic is

$$\begin{aligned} t_{\gamma^m} &= \frac{\widehat{\gamma}^m}{se\left(\widehat{\gamma}^m\right)} \\ &\rightarrow \sqrt{1-\theta}\frac{\sqrt{n}\gamma}{\left(\left|\gamma\right|\sqrt{\theta}+\frac{\sigma_e}{\sigma_u}\right)} \end{aligned}$$

which is smaller than t_γ for the error free case. Compared to the balancing test statistic t_{δ^m}, measurement error reduces t_{γ^m} relatively more, namely due to the term $\left|\gamma\right|\sqrt{\theta}$ in the denominator. This is due to the fact that measurement error in regressor both attenuates the relevant coefficient towards zero as well introducing additional variance into the residual. The upshot from this discussion is that classical measurement error makes the assessment of whether $\gamma = 0$ comparatively more difficult compared to the assessment whether $\delta = 0$.

Finally, consider the quantity $\beta^s - \beta^m$, which enters the coefficient comparison test,

$$Var\left(\widehat{\beta}^s - \widehat{\beta}^m\right) = Var\left(\widehat{\beta}^s\right) + Var\left(\widehat{\beta}^m\right) - 2Cov\left(\widehat{\beta}^s, \widehat{\beta}^m\right).$$

There are various things to note about this expression. $Var\left(\widehat{\beta}^s\right)$ and $Var\left(\widehat{\beta}^m\right)$ cannot be ranked. Adding an additional regressor may increase or lower the standard error on the regressor of interest. Secondly, the covariance term reduces the sampling variance of the coefficient comparison test. The covariance term will generally be sizeable compared to the $Var\left(\widehat{\beta}^s\right)$ and $Var\left(\widehat{\beta}^m\right)$ because the regression residuals e_i^s and e_i^m will be highly correlated. In fact,

$$Cov\left(e_i^s, e_i^m\right) = \gamma^2\theta\sigma_u^2 + \sigma_e^2$$

and

$$Cov\left(\widehat{\beta}^s, \widehat{\beta}^m\right) = \frac{1}{n}\frac{\gamma^2\theta\sigma_u^2 + \sigma_e^2}{\sigma_s^2}.$$

We show that the covariance term is closely related to the sampling variance of the short regression coefficient

$$Var\left(\widehat{\beta}^s\right) = \frac{1}{n}\frac{\gamma^2\sigma_u^2 + \sigma_e^2}{\sigma_s^2}.$$

Because the covariance term gets subtracted, looking at the standard errors of $\widehat{\beta}^s$ and $\widehat{\beta}^m$ alone can be very misleading about the precision of the coefficient comparison.

Putting everything together

$$Var\left(\widehat{\beta}^s - \widehat{\beta}^m\right) = \frac{1}{n}(1-\theta)\left(\gamma^2\frac{\sigma_u^2}{\sigma_s^2} + \theta\delta^2\gamma^2 + \delta^2\frac{\sigma_e^2}{\sigma_u^2}\right).$$

Setting $\theta = 0$, it is easy to see that, like $Var\left(\widehat{\gamma}^m\right)$, $Var\left(\widehat{\beta}^s - \widehat{\beta}^m\right)$ has both an attenuation factor as well as an additional positive term compared to

$Var\left(\widehat{\beta}^s - \widehat{\beta}\right)$. Measurement error may therefore raise or lower the sampling variance for the coefficient comparison test.

The coefficient comparison test itself can be formulated as a t-test as well, since we are interested in the movement in a single parameter.

$$t_{(\beta^s - \beta^m)} = \frac{\widehat{\beta}^s - \widehat{\beta}^m}{\sqrt{\frac{1}{n}\left(1-\theta\right)\left(\gamma^2\frac{\sigma_u^2}{\sigma_s^2} + \theta\delta^2\gamma^2 + \delta^2\frac{\sigma_e^2}{\sigma_u^2}\right)}}.$$

Note that

$$\beta^s - \beta^m = \delta\gamma^m = \delta\gamma\left(1-\theta\right)$$

so that

$$\begin{aligned} t_{(\beta^s - \beta^m)} &\to \frac{\delta\gamma\left(1-\theta\right)}{\sqrt{\frac{1}{n}\left(1-\theta\right)\left(\gamma^2\frac{\sigma_u^2}{\sigma_s^2} + \theta\delta^2\gamma^2 + \delta^2\frac{\sigma_e^2}{\sigma_u^2}\right)}} \\ &= \sqrt{1-\theta}\frac{\sqrt{n}\delta\gamma}{\sqrt{\gamma^2\frac{\sigma_u^2}{\sigma_s^2} + \theta\delta^2\gamma^2 + \delta^2\frac{\sigma_e^2}{\sigma_u^2}}} \end{aligned}$$

As in the case of t_{γ^m}, measurement error lowers this t-statistic. Comparing $t_{(\beta^s - \beta^m)}$ and t_{δ^m} we note that $t_{(\beta^s - \beta^m)}$ is affected more by measurement error:

$$t_{\delta^m} = \sqrt{1-\theta}t_{\delta},$$

while $t_{(\beta^s - \beta^m)}$ is subject both to the attenuation factor $\sqrt{1-\theta}$ and to the additional variance term $\theta\delta^2\gamma^2$. This suggests that under the maintained hypothesis $\gamma \neq 0$, the balancing test will be more powerful than the coefficient comparison test. This result itself is not surprising; after all it ought to be easier to test $\delta = 0$ while maintaining $\gamma \neq 0$, compared to testing the compound hypothesis $\delta = 0$ or $\gamma = 0$. Below we show that the differences in power between the tests can be substantial when there is a lot of measurement error in x_i^m. Before we do so, we briefly note how the coefficient comparison test can be implemented in practice.

4.1 Implementing the Coefficient Comparison Test

The balancing test is a straightforward t-test, which regression software calculates routinely. We noted that the coefficient comparison test is a generalized Hausman test. Regression software will typically calculate this as well if it allows for seemingly unrelated regression estimation (SURE). SURE takes $Cov\left(e_i^s, e_i^m\right)$ into account and therefore facilitates the test. In Stata, this is implemented via the `suest` command. Generically, the test would take the following form:

```
reg y s
est store reg1
reg y s x
est store reg2
suest reg1 reg2
test[reg1_mean]s=[reg2_mean]s
```

The test easily accommodates covariates or can be carried out with the variables `y`, `s`, and `x` being residuals from a previous regression (hence facilitating large numbers of fixed effects though degrees of freedom may have to be adjusted in this case).

As far as we can tell, the Stata `suest` or `3reg` commands don't work for the type of IV regressions we might be interested in here. An alternative, which also works for IV, is to take the regressions (1) and (2) and stack them:

$$\begin{bmatrix} y_i \\ y_i \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha^s \\ \alpha \end{bmatrix} + \begin{bmatrix} s_i & 0 \\ 0 & s_i \end{bmatrix} \begin{bmatrix} \beta^s \\ \beta \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & x_i \end{bmatrix} \begin{bmatrix} 0 \\ \gamma \end{bmatrix} + \begin{bmatrix} e_i^s \\ e_i \end{bmatrix}.$$

Testing $\beta^s - \beta = 0$ is akin to a Chow test across the two specifications (1) and (2). Of course, the data here are not two subsamples but the original data set duplicated. To take account of this and allow for the correlation in the residuals across duplicates, it is crucial to cluster standard errors on the observation identifier i.

4.2 Power comparisons

The ability of a test to reject when the null hypothesis is false is described by the power function of the test. The power functions here are functions of d, the values the parameter δ might take. Using the results from the previous section, the power function for a 5% critical value of the balancing test is

$$Power_{t_{\delta^m}}(d) = 1 - \Phi\left(1.96 - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right) + \Phi\left(-1.96 - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right)$$

while the power function for the coefficient comparison test is

$$Power_{t_{(\beta^s-\beta^m)}}(d;\gamma) = 1 - \Phi\left(1.96 - d\frac{\sqrt{n}\gamma(1-\theta)}{\sqrt{V_\beta(d;\gamma)}}\right) + \Phi\left(-1.96 - d\frac{\sqrt{n}\gamma(1-\theta)}{\sqrt{V_\beta(d;\gamma)}}\right)$$

where

$$V_\beta(d;\gamma) = (1-\theta)\left(\frac{\gamma^2\sigma_u^2}{\sigma_s^2} + \theta d^2\gamma^2 + \frac{d^2\sigma_e^2}{\sigma_u^2}\right).$$

Note that the power function for the balancing test does not involve the parameter γ. Nevertheless, for $0 < \gamma < \infty$ it can be written as

$$Power_{t_\delta}(d) = 1 - \Phi\left(1.96 - d\frac{\sqrt{n}\gamma(1-\theta)}{\sqrt{V_\delta(d;\gamma)}}\right) + \Phi\left(-1.96 - d\frac{\sqrt{n}\gamma(1-\theta)}{\sqrt{V_\delta(d;\gamma)}}\right).$$

where

$$V_\delta(d;\gamma) = (1-\theta)\frac{\gamma^2\sigma_u^2}{\sigma_s^2}.$$

It is hence apparent that $V_\beta(d;\gamma) > V_\delta(d;\gamma)$, i.e. the coefficient comparison test has a larger variance. As a result

$$Power_{t_\delta}(d) > Power_{t_{(\beta^s-\beta^m)}}(d;\gamma).$$

In practice, this result may or may not be important, so we illustrate it with a number of numerical examples. Table 1 displays the parameter values as well as the implied values of the R^2 of regression (8). The values were chosen so that for intermediate amounts of measurement error in x_i^m the R^2s are reflective of regressions fairly typical of those in applied microeconomics, for example, a wage regression.

In Figure 1, we plot the power functions for both tests for three different magnitudes of the measurement error. The first set involves the power functions with no measurement error. The power functions can be seen to increase quickly with d, and both tests reject with virtual certainty as d reaches values of 1. The balancing test is slightly more powerful but this difference is small, and only visible in the figure for a small range of d.

The second set of power functions corresponds to a reliability ratio for x_i^m of $\lambda = 0.5$. Measurement error of that magnitude visibly affects the power of both tests. The balancing test still rejects with certainty for $d > 1.5$ while the power the coefficient comparison test flattens out around a value of 0.93. This discrepancy becomes even more pronounced with a $\lambda = 0.25$. The power of the coefficient comparison test does not rise above 0.6 in this case, while the balancing test still rejects with a probability of 1 at values of d slightly above 2.[2]

The results in Figure 1 highlight two important things. There are parameter combinations where the balancing test has substantially more power than the coefficient comparison test. On the other hand, there are other regions where the power of the two tests is very similar, for example, the region where $d < 0.5$ in Figure 1. In these cases, both tests perform very similar but, of course, specific results may differ in small samples. Hence, in a particular application, the coefficient comparison test may reject when the balancing test doesn't.

The homoskedastic case with classical measurement error might be highly stylized and not correspond well to the situations typically encountered in empirical practice. We therefore explore some other scenarios as well using simulations. Figure 2 shows the original theoretical power functions for the case with no measurement error from Figure 1. It adds empirical rejection

[2] Power for the cc test can actually be seen to start to decline as d increases. This comes from the fact that the amount of measurement error is parameterized in terms of the reliability λ of x_i^m. For a constant reliability the amount of measurement error increases with d. We felt that thinking about the reliability is probably the most natural way for applied researchers to think about the amount of measurement error they face in a variable.

rates from simulations with heteroskedastic errors u_i and e_i of the form

$$\begin{aligned}
\sigma_{u,i}^2 &= \left(\frac{e^{|s_i|}}{1+e^{|s_i|}}\right)^2 \sigma_{0u}^2 \\
\sigma_{e,i}^2 &= \left(\frac{e^{|s_i|}}{1+e^{|s_i|}}\right)^2 \sigma_{0e}^2.
\end{aligned}$$

We chose the baseline variances σ_{0u}^2 and σ_{0e}^2 so that $\overline{\sigma}_u^2 = 3$ and $\overline{\sigma}_e^2 = 30$ to match the variances in Figure 1. All test statistics employ robust standard errors. We plot the rejection rates for data with no measurement error and for the more severe measurement error given by a reliability ratio $\lambda = 0.25$.[3] As can be seen in Figure 2, both the balancing and the coefficient comparison tests lose some power when the residuals are heteroskedastic compared to the homoskedastic baseline. Otherwise, the results look very similar to those in Figure 1. Heteroskedasticity does not seem to alter the conclusions appreciatively.

Next, we explore mean reverting measurement error (Bound et al., 1994). We generate measurement error as

$$m_i = \kappa x_i + \mu_i$$

and set $\kappa = -0.5$. Notice that x_i^m can now be written as

$$x_i^m = (1+\kappa)\,\delta_0 + (1+\kappa)\,\delta s_i + \ (1+\kappa)\,u_i + \mu_i,$$

so that this parameterization directly affects the coefficient in the balancing regression, which will be smaller than δ for a negative κ. At the same time, the residual variance in this regression is also reduced for a given reliability ratio.[4] Figure 3 demonstrates that the power of both tests deteriorates even for moderate amounts of measurement error now but the coefficient comparison test is still most affected.

[3] We did 25,000 replications in these simulations, and the underlying regressions have a 100 observations.

[4] Note that fixing λ, σ_μ^2 is given by

$$\sigma_\mu^2 = \frac{1-\lambda\,(1+\kappa)}{\lambda} Var\,(x_i)\,.$$

The case of mean reverting measurement error captures a variety of ideas, including the one that we may observe only part of a particular concept. Imagine we would like to include in our regression a variable $x_i = w_{1i} + w_{2i}$, where w_{1i} and w_{2i} are two orthogonal variables. We observe $x_i^m = w_{1i}$. For example, x_i may be family background, w_{1i} is mother's education and other parts of family background correlated with it, and w_{2i} are all relevant parts of family background, which are uncorrelated with mother's education. As long as selection bias due to w_{1i} and w_{2i} is the same, this amounts to the mean reverting measurement error formulation above. This scenario is also isomorphic to the model studied by Oster (2014). See the appendix for details.

5 Empirical Analysis

We illustrate the theoretical results in the context of estimating the returns to schooling using data from the National Longitudinal Survey of Young Men. This is a panel study of about 5,000 male respondents interviewed from 1966 to 1981. The data set has featured in many prominent analyses of the returns to education, including Griliches (1977) and Card (1995). We use the NLS extract posted by David Card and augment it with the variable on body height measured in the 1973 survey. We estimate regressions similar to eq. (2). The variable y_i is the log hourly wage in 1976 and s_i is the number of years of schooling reported by the respondent in 1976. Our samples are restricted to observations without missing values in any of the variables used in a particular table or set of tables.

We start in Table 2 by presenting simple OLS regressions controlling for experience, race, and region of residence. The estimated return to schooling is 7.5%. This estimate is unlikely to reflect the causal effect of education on income because important confounders, which influence both education and income simultaneously such as ability or family background, are not controlled for.

In columns (2) to (5) we include variables which might proxy for the respondent's family background. In column (2) we include mother's education, in column (3) whether the household had a library card when the respondent was 14, and in column (4) we add body height measured in inches. Each of these variables is correlated with earnings and the coefficient on education moves moderately when these controls are included. Mother's education captures an important component of a respondent's family background. The library card measure has been used by researchers to proxy for important parental attitudes (e.g. Farber and Gibbons, 1996). Body height is a variable determined by parents' genes and by nutrition and disease environment during childhood. It is unlikely a particularly powerful control variable but it is predetermined and correlated with family background, self-esteem, and ability (e.g. Persico, Postlewaite, and Silverman, 2004; Case and Paxson, 2008). The return to education falls by .1 to .2 log points when these controls are added. In column (5) we enter all three variables simultaneously. The coefficients on the controls are somewhat attenuated and the return to education falls slightly further to 7.1%.

It might be tempting to conclude from the relatively small change in the estimated returns to schooling that this estimate might safely be given a causal interpretation. We provide a variety of evidence that this conclusion is unlikely to be a sound one. Below the estimates in columns (2) to (5), we display the p-values from the coefficient comparison test, comparing each of the estimated returns to education to the one from column (1). Although the coefficient movements are small, the tests all reject at the 5% level, and in columns (4) and (5) they reject at the 1% level.

The results in columns 6 to 8, where we regress maternal education, the library card, and body height on education demonstrates this worry. The education coefficient is positive and strongly significant in all three regressions, with t-values ranging from 4.4 to 13.1. The magnitudes of the coefficients are substantively important. It is difficult to think of these results as causal effects: the respondent's education should not affect predetermined proxies of family background. Instead, these estimates reflect selection bias. Individuals with more education have significantly better educated mothers, were more likely to grow up in a household with a library card, and experienced more body growth when young. Measurement error leads to attenuation bias when these variables are used on the right-hand side which renders them fairly useless as controls. The measurement error does not matter for the estimates in columns 6-8, and these are informative about the role of selection. Comparing the p-values at the bottom of the table to the corresponding ones for the coefficient comparison test in columns 2 to 4 demonstrates the superior power of the balancing test.

Finally, we report a number of additional results in the table. The R^2 from regression of education on the added regressor (mother's education, the library card, or height) is an ingredient necessary for the calculations that follow. Next, we report the values for β if the added regressor was the only remaining source of omitted variables bias, assuming various degrees of measurement error. These calculations are based on equation (11). Since the idea that any of the candidate controls by themselves would identify the return in these bare bones wage equations does not seem particularly believable we will discuss these results in the context of Table 3.

In Table 3 we repeat the same set of regressions including a direct measure for ability, the respondent's score on the Knowledge of the World of Work test (KWW), a variable used by Griliches (1977) as a proxy for ability. The sample size is reduced due to the exclusion of missing values in the test score. Estimated returns without the KWW score are very similar to those in the original sample. Adding the KWW score reduces the coefficient on education by almost 20%, from 0.075 to 0.061. Adding maternal education, the library card, and body height does very little now to the estimated returns to education. The coefficient comparison test indicates that none of the small changes in the returns to education are significant. Controlling for the KWW scores has largely knocked out the library card effect but done little to the coefficients on maternal education and body height. The relatively small and insignificant coefficient movements in columns (2) to (5) suggest that the specification controlling for the KWW score might solve the ability bias problem.

Columns (6)-(8), however, show that the regressions with the controls on the left hand side still mostly result in significant education coefficients even when the KWW score is in the regression. This suggests that the estimated returns in columns 1-5 might also still be biased by selection. The estimated coefficients on education for the three controls are on the order of half their value from Table 1, and the body height measure is now only significant at the 10% level. Particularly the relationship between mother's and own education is still sizable, and this measure still indicates the possibility of important selection.

The calculations at the bottom of the table based on equation (11) also confirm that mother's education might potentially pick up variation due to an important confounder. These calculations assume that mother's education is the only omitted control in column (1) while acknowledging that the available measure might contain a lot of noise compared to the correct control. With the moderate amounts of measurement error implied by a reliability of 0.75 or 0.5 the returns to education coefficient still moves fairly little when adding mother's education. For a reliability of 0.5 the return remains .058 compared to .061 without controlling for mother's education. If the reliability is only 0.25 the return falls more strongly to 0.055. In order for the entire estimated return in column (1) to be explained by omitted variables bias due to mother's education the reliability needs to be as low as 0.05, as can be seen in the last row.

These numbers highlight a lot of curvature in the relationship between the reliability and the implied return to education. Figure 4 illustrates this for the case of the mother's education variable. It becomes clear that the return changes little for reliabilities above 0.25 but then falls precipitously for more severe measurement error. If we believe that mother's education captures family background poorly enough there is a lot room for bias from this source.

Looking at the columns (3) and (4) we can see that the same isn't true for the library card and body height measures. Here the returns relationship is essentially flat over the range of reliabilities as low as 0.25. Reliabilities as low as 0.01 are necessary for a zero return. This confirms that these variables have lost most of their power as confounders once KWW is controlled in the regressions. The flat relationship between the reliability and returns is due to the fact that both δ and γ^m are lower for the library card and body height in Table 3 compared to Table 2. We don't claim here that adding these variables to the regressions with the KWW score would be a suitable identification strategy in any case. Rather, we see the implied β calculation for different reliabilities as an intuitive measure summarizing the impact of the relevant values of δ, γ^m, and the R^2 between years of education and the added regressor.

While the KWW score might be a powerful control it is likely also measured with substantial error. Griliches (1977) proposes to instrument this measure with an IQ testscore variable, which is also contained in the NLS data, to eliminate at least some of the consequences of this measurement error. In Table 4 we repeat the schooling regressions with IQ as instrument for the KWW score. The coefficient on the KWW score almost triples, in line with the idea that an individual test score is a very noisy measure of ability. The education coefficient now falls to only about half its previous value from 0.061 to 0.034. This might be due to positive omitted variable bias present in the previous regressions which is eliminated by IQ-instrumented KWW (although there may be other possible explanations for the change as well). Both the coefficient comparison tests and the balancing tests indicate no evidence of selection any more. This is due to a combination of lower point estimates and larger standard errors. The contrast between tables 3 and 4 highlights the usefulness of the balancing test: it warns about the Table 3 results, while the coefficient comparison test delivers insignificant differences in either case.

Finding an instrumental variable for education is an alternative to control strategies, such as using test scores. In Table 5 we follow Card's (1995) analysis and instrument education using distance to the nearest college, while dropping the KWW score.[5] We use the same sample as in Table 2, which

[5] We use a single dummy variable for whether there is a four year college in the county,

differs from Card's sample.[6] Our IV estimates of the return to education are slightly higher than in Table 2 but a lot lower than in Card (1995) at around 8%. The IV returns estimates are noisy, never quite reaching a t-statistic of 2. Columns 1-5 of Table 5 show that the IV estimate on education, while bouncing around a bit, does not change significantly when maternal education, the library card, or body height are included. In particular, if these three controls are included at the same time in column (5) the point estimate is clearly indistinguishable from the unconditional estimate in column (1).

IV regressions with pre-determined variables on the left hand side can be thought of as a test for the exclusion restriction or random assignment of the instruments. Unfortunately, in this case the selection regressions in columns (6)-(8) are also much less precise and as a result less informative. The coefficients in the regressions for mother's education and body height have the wrong sign but confidence intervals cover anything ranging from zero selection to large positive amounts. Only the library card measure is large, positive, and significant around the 6% level, warning of some remaining potential for selection even in the IV regressions. While the data do not speak clearly in this particular case this does not render the methodology any less useful.

6 Conclusion

Using predetermined characteristics as dependent variables offers a useful specification check for a variety of identification strategies popular in empirical economics. We argue that this is the case even for variables which might be poorly measured and are of little value as control variables. Such variables should be available in many data sets, and we encourage researchers to perform such "balancing" tests more frequently. We show that this is a more powerful strategy than adding the same variables on the right hand side of the regression as controls and looking for movement in the coefficient

and we instrument experience and experience squared by age and age squared.

[6] We restrict Card's sample to non-missing values in maternal education, the library card, and body height.

of interest.

We have illustrated our theoretical results with an application to the returns to education. Taking our assessment from this exercise at face value, a reader might conclude that the results in Table 4, returns around 3.5%, can safely be regarded as causal estimates. Of course, this is not the conclusion reached in the literature, where much higher IV estimates like those in Table 5 are generally preferred (see e.g. Card, 2001 or Angrist and Pischke, 2015, chapter 6). This serves as a reminder that the discussion here is focused on sharpening one particular tool in the kit of applied economists; it is not a miracle cure for all ills.

The balancing test and other statistics we discuss here are useful to gauge selection bias due to observed confounders, even when they are potentially measured poorly. It does not address any other issues which may also haunt a successful empirical investigation of causal effects. One possible issue is measurement error in the variable of interest, which is also exacerbated as more potent controls are added. Griliches (1977) shows that a modest amount of measurement error in schooling may be responsible for the patterns of returns we have displayed in Tables 2 to 4. Another issue, also discussed by Griliches, is that controls like test scores might themselves at least be partly influenced by schooling, which would make them bad controls. For all these reasons, IV estimates of the returns may be preferable.

There are other issues we have sidestepped in our analysis. Our discussion has focused on the case where a researcher has a single regressor or a small set of such regressors available for addition to a candidate regression. But sometimes we might be interested in the robustness of the original results when a large number of regressor are added. An example would be a differences-in-differences analysis in a state-year panel, where the researcher is interested in checking whether the results are robust to the inclusion of state specific trends. The balancing test seems to be of little use in this case. In fact, the analysis in Hausman (1978) and Holly (1982) highlights that the coefficient comparison (Hausman) test may be particularly powerful in some cases where many regressors are added.[7] Whether the principles of the balancing test can be harnessed in a fruitful way for such scenarios is a useful avenue for future research.

References

[1] Angrist, Joshua and Jörn-Steffen Pischke (2015) *Mastering Metrics. The Path from Cause to Effect*, Princeton: Princeton Univeristy Press.

[2] Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber (2005) "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, vol. 113, no. 1, Februrary, 151-184.

[3] Battistin, Erich and Andrew Chesher (2014), "Treatment Effect Estimation with Covariate Measurement Error," *Journal of Econometrics*, vol 178, no. 2, February, 707-715.

[4] Bound, John et al. (1994) "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data," *Journal of Labor Economics*, vol. 12, no. 3, 345-368.

[5] Card, David (1995) "Using Geographic Variations in College Proximity to Estimate the Returns to Schooling," in *Aspects of Labor Market Behavior: Essays in Honor of John Vanderkamp*, L. N. Christofides, E. K. Grand, and R. Swidinsky, eds., Toronto: University of Toronto Press.

[6] Card, David (2001) "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, vol. 69, September, 1127–1160.

[7] Case, Anne, and Christina Paxson (2008) "Stature and Status: Height, Ability, and Labor Market Outcomes, " *Journal of Political Economy*, vol. 116, no. 3, 499-532.

[7] See also McKinnon (1992) section II.9 for a more intuitive discussion of these issues.

[8] Farber, Henry S., and Robert Gibbons (1996) "Learning and Wage Dynamics," *The Quarterly Journal of Economics*, vol. 111, no. 4, 1007-1047.

[9] Griliches, Zvi (1977) "Estimating the Returns to Schooling - Some Econometric Problems," *Econometrica*, vol. 45, January, 1-22.

[10] Hausman, Jerry (1978) "Specification Tests in Econometrics," *Econometrica*, vol. 46, no. 6, November, 1251-1272.

[11] Holly, Alberto (1982) "A Remark on Hausman's Specification Test," *Econometrica*, vol. 50, no. 3, May, 749-759.

[12] Imbens, Guido W. (2003) "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review Papers and Proceedings*, vol. 93, no. 2, May, 126-132.

[13] MacKinnon, James G. (1992) "Model Specification Tests and Artificial Regressions," *Journal of Economic Literature*, vol. 30, no. 1, March, 102-146.

[14] Oster, Emily (2014) "Unobservable Selection and Coefficient Stability: Theory and Evidence," mimeographed, University of Chicago, January 23.

[15] Persico, Nicola, Andrew Postlewaite and Dan Silverman (2004) "The Effect of Adolescent Experience on Labor Market Outcomes: The Case of Height," *Journal of Political Economy*, vol. 112, no. 5, 1019-1053.

[16] Rosenbaum, Paul R. and Donald B. Rubin (1983) "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society. Series B*, vol. 45, no. 2, 212-218.

Figure 1: Theoretical Rejection Rates

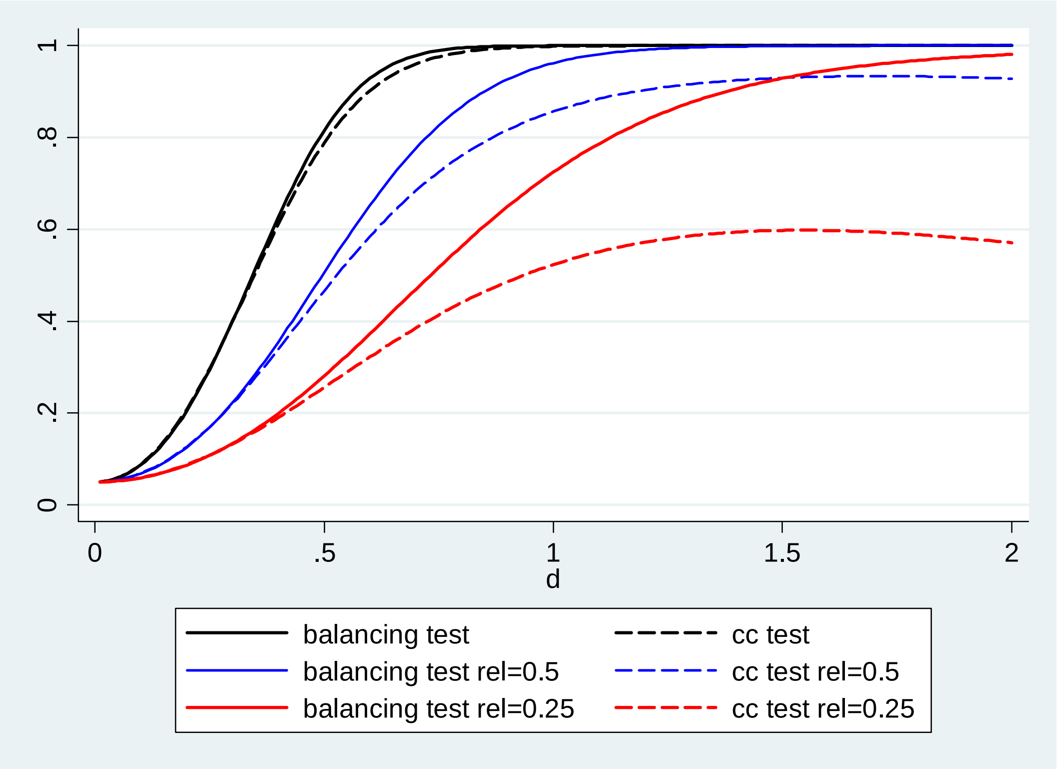

Figure 2: Simulated Rejection Rates with Heteroskedasticity

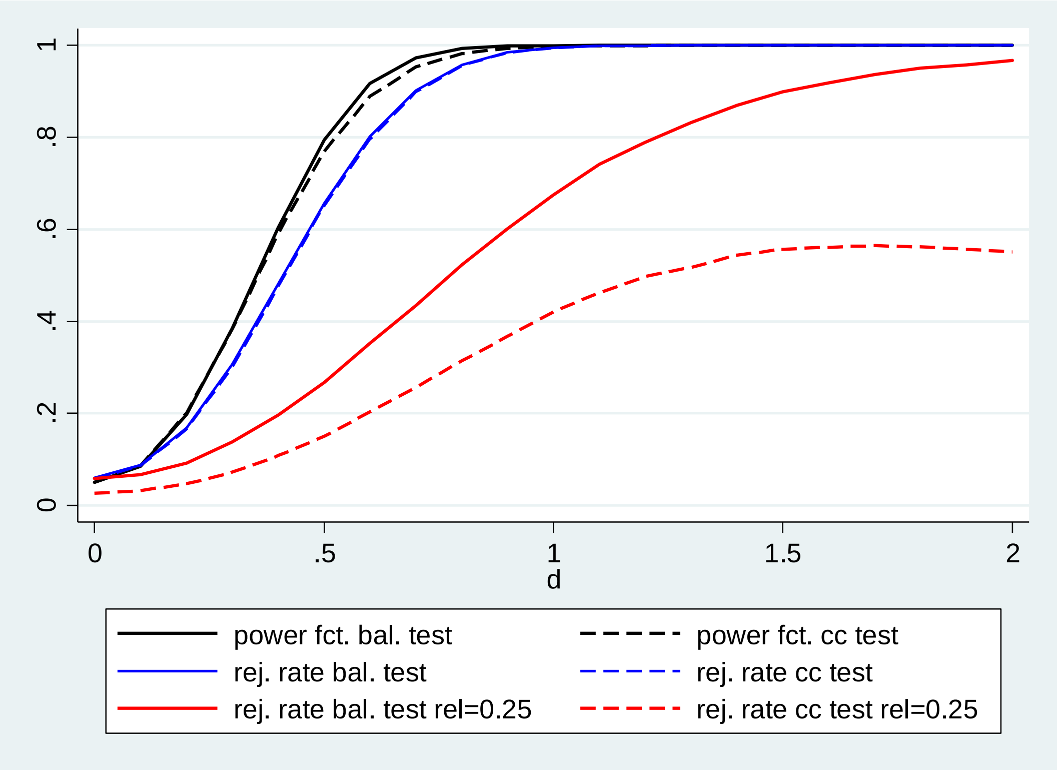

Figure 3: Simulated Rejection Rates with Mean Reverting Measurement Error

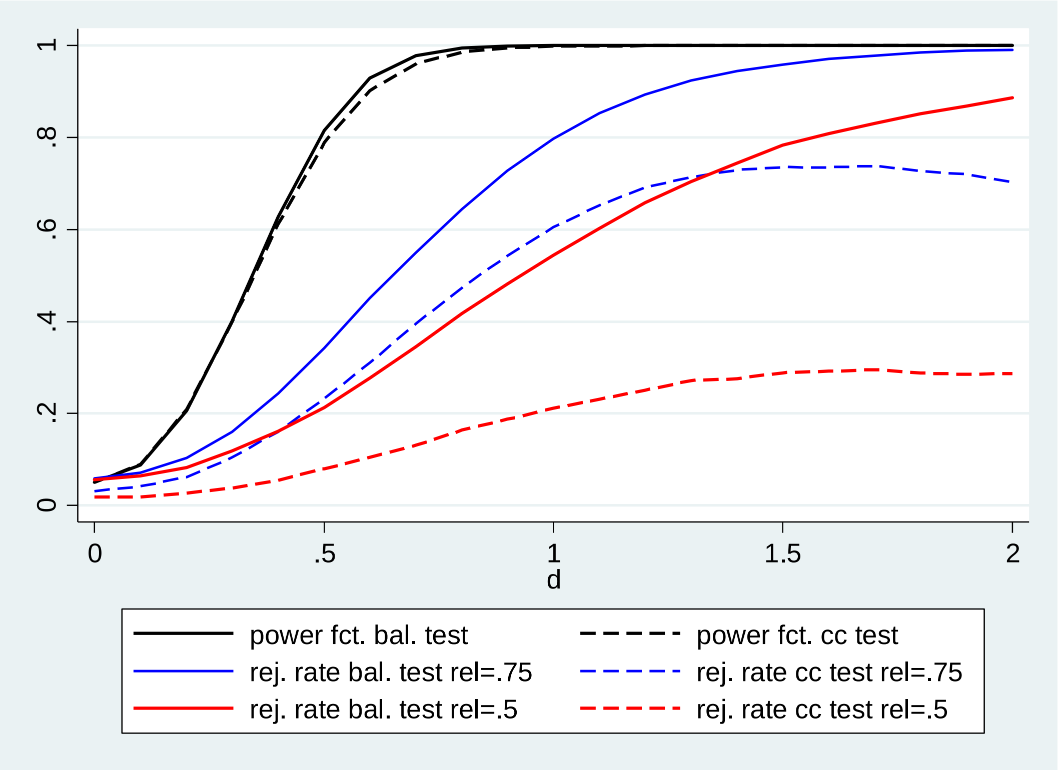

Figure 4: Implied βs for Different Values the Reliability of Mother's Education

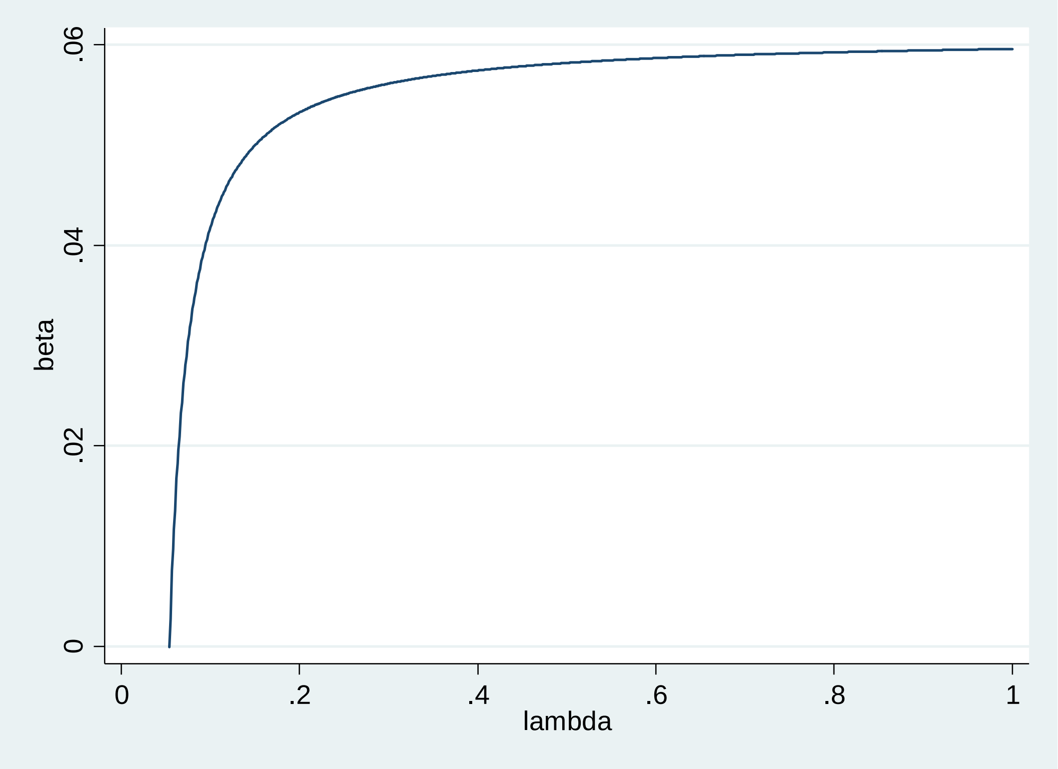

Table 1: Parameters for Power Calculations and Implied R^{2}s

$\sigma_s^2 = 1$ $\sigma_u^2 = 3$ $\sigma_e^2 = 30$		$\gamma = 3$ $n = 100$	
	R^2		
d	$\lambda = 1$	$\lambda = 0.5$	$\lambda = 0.25$
0	0.47	0.24	0.12
0.5	0.49	0.26	0.15
1.0	0.55	0.31	0.22
1.5	0.61	0.39	0.32
2.0	0.68	0.47	0.42

Table 2: Baseline Regressions for Returns to Schooling and Specification Checks

	Log hourly earnings					Mother's years of education	Library card at age 14	Body height in inches
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education	0.0751	0.0728	0.0735	0.0740	0.0710	0.3946	0.0371	0.1204
	(0.0040)	(0.0042)	(0.0040)	(0.0040)	(0.0042)	(0.0300)	(0.0040)	(0.0273)
Mother's years of education		0.0059			0.0044			
		(0.0029)			(0.0030)			
Library card at age 14			0.0428		0.0361			
			(0.0183)		(0.0184)			
Body height in inches				0.0090	0.0084			
				(0.0027)	(0.0027)			
p-values								
Coefficient comparison test		0.045	0.023	0.010	0.002			
Balancing test						0.000	0.000	0.000
R^2 of education on added regressor		0.073	0.032	0.007				
Implied β for reliability = 0.75		0.0719	0.0730	0.0737				
reliability = 0.50		0.0700	0.0718	0.0729				
reliability = 0.25		0.0628	0.0680	0.0707				
Reliability that implies β=0		0.102	0.052	0.022				

N = 2,500 in all regressions. Heteroskedasticity robust standard errors in parentheses. All regressions control for experience, experience-squared, indicators for black, for southern residence and residence in an SMSA in 1976, indicators for region in 1966 and living in an SMSA in 1966.

Table 3: Regressions for Returns to Schooling and Specification Checks Controlling for the KWW Score

	Log hourly earnings					Mother's years of education	Library card at age 14	Body height in inches
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education	0.0609	0.0596	0.0608	0.0603	0.0591	0.2500	0.0133	0.0731
	(0.0059)	(0.0060)	(0.0059)	(0.0059)	(0.0060)	(0.0422)	(0.0059)	(0.0416)
KWW score	0.0070	0.0068	0.0069	0.0069	0.0067	0.0410	0.0076	0.0145
	(0.0015)	(0.0016)	(0.0016)	(0.0015)	(0.0016)	(0.0107)	(0.0016)	(0.0117)
Mother's years of education		0.0053			0.0048			
		(0.0037)			(0.0037)			
Library card at age 14			0.0097		0.0045			
			(0.0215)		(0.0216)			
Body height in inches				0.0078	0.0075			
				(0.0034)	(0.0034)			
p-values								
Coefficient comparison test		0.163	0.652	0.158	0.085			
Balancing test						0.000	0.025	0.079
R^2 of education on added regressor		0.033	0.006	0.002				
Implied β for reliability = 0.75		0.0591	0.0607	0.0602				
reliability = 0.50		0.0582	0.0607	0.0598				
reliability = 0.25		0.0550	0.0604	0.0586				
Reliability that implies β=0		0.054	0.008	0.011				

N = 1,773 in all regressions, due to missing values in IQ. Heteroskedasticity robust standard errors in parentheses. All regressions control for experience, experience-squared, indicators for black, for southern residence and residence in an SMSA in 1976, indicators for region in 1966 and living in an SMSA in 1966.

Table 4: Regressions for Returns to Schooling and Specification Checks Instrumenting the KWW Score

	Log hourly earnings					Mother's years of education	Library card at age 14	Body height in inches
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education	0.0340	0.0339	0.0342	0.0343	0.0345	0.0234	0.0168	-0.0486
	(0.0139)	(0.0139)	(0.0138)	(0.0139)	(0.0138)	(0.0952)	(0.0134)	(0.0998)
KWW score instrumented by IQ		0.0195	0.0200	0.0194	0.0191	0.1496	0.0060	0.0728
		(0.0063)	(0.0063)	(0.0062)	(0.0064)	(0.0422)	(0.0060)	(0.0449)
Mother's years of education		0.0028			0.0026			
		(0.0039)			(0.0039)			
Library card at age 14			-0.0130		-0.0154			
			(0.0245)		(0.0243)			
Body height in inches				0.0070	0.0069			
				(0.0034)	(0.0034)			
p-values								
Coefficient comparison test		0.818	0.635	0.636	0.552			
Balancing test						0.806	0.212	0.626
R^2 of education on added regressor		0.000	0.001	0.004				
Implied β for reliability = 0.75		0.0339	0.0343	0.0344				
reliability = 0.50		0.0338	0.0344	0.0347				
reliability = 0.25		0.0337	0.0349	0.0353				
Reliability that implies β=0		0.002	-0.006	-0.006				

N = 1,773 in all regressions, due to missing values in IQ. Heteroskedasticity robust standard errors in parentheses. All regressions control for experience, experience-squared, indicators for black, for southern residence and residence in an SMSA in 1976, indicators for region in 1966 and living in an SMSA in 1966.

Table 5: Regressions for Returns to Schooling and Specification Checks Instrumenting Schooling by Proximity to College

	Log hourly earnings					Mother's years of education	Library card at age 14	Body height in inches
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education instrumented by college proximity	0.0816 (0.0431)	0.0818 (0.0417)	0.0778 (0.0518)	0.0845 (0.0418)	0.0822 (0.0466)	-0.0952 (0.3594)	0.1015 (0.0542)	-0.3658 (0.3681)
Mother's years of education		0.0030 (0.0143)			0.0012 (0.0140)			
Library card at age 14			0.0367 (0.0886)		0.0237 (0.0581)			
Body height in inches				0.0081 (0.0044)	0.0079 (0.0032)			
p-values								
Coefficient comparison test		0.873	0.686	0.380	0.908			
Balancing test						0.791	0.061	0.321

N = 2,500 in all regressions. Heteroskedasticity robust standard errors in parentheses. All regressions control for experience, experience-squared, indicators for black, for southern residence and residence in an SMSA in 1976, indicators for region in 1966 and living in an SMSA in 1966.

7 Appendix

7.1 Power Functions

7.1.1 The Balancing Test

The desired balancing regression is

$$x_i = \delta_0 + \delta s_i + \ u_i,$$

however, x_i is measured with error

$$x_i^m = x_i + m_i.$$

Effectively, we run the balancing regression

$$x_i^m = \delta_0^m + \delta^m s_i + \ u_i + m_i.$$

The test statistic for the null hypothesis that the balancing coefficient δ is zero is

$$t_{\delta^m} = \frac{\widehat{\delta}^m}{se\left(\widehat{\delta}^m\right)} = \frac{\widehat{\delta}^m}{\frac{1}{\sqrt{n}}\frac{\sqrt{\sigma_u^2+\sigma_m^2}}{\sigma_s}}$$

Define

$$\begin{aligned} \theta &= \frac{\sigma_m^2}{\sigma_u^2+\sigma_m^2} \\ &\Rightarrow \ \sigma_u^2+\sigma_m^2 = \frac{\sigma_u^2}{1-\theta} \end{aligned}$$

Hence

$$t_{\delta^m} = \widehat{\delta}^m \frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}.$$

The rejection probability is

$$\begin{aligned}
\Pr\left(\left|t_{\delta^m}\right| > C\right| H_1) &= \Pr\left(t_{\delta^m} > C\right| H_1) + \ \Pr\left(t_{\delta^m} < -C\right| H_1) \\
&= \Pr\left(\left.\frac{\widehat{\delta}^m}{se\left(\widehat{\delta}^m\right)} > C\right| H_1\right) + \Pr\left(\left.\frac{\widehat{\delta}^m}{se\left(\widehat{\delta}^m\right)} < -C\right| H_1\right) \\
&= \Pr\left(\left.\frac{\widehat{\delta}^m - d}{se\left(\widehat{\delta}^m\right)} > C - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right| H_1\right) \\
&\quad + \Pr\left(\left.\frac{\widehat{\delta}^m - d}{se\left(\widehat{\delta}^m\right)} < -C - \frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right| H_1\right) \\
&\xrightarrow{d} 1 - \Phi\left(C - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right) + \Phi\left(-C - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right)
\end{aligned}$$

This is the power function of the balancing test

$$Power_{t_\delta}(d) = 1 - \Phi\left(1.96 - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right) + \Phi\left(-1.96 - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right).$$

7.1.2 The Coefficient Comparison Test

The short and long regressions are

$$\begin{aligned}
y_i &= \alpha^s + \beta^s s_i + e_i^s \\
y_i &= \alpha + \beta s_i + \gamma x_i + e_i,
\end{aligned}$$

and

$$x_i = \delta_0 + \delta s_i + \ u_i.$$

Adding measurement error in x_i:

$$x_i^m = x_i + m_i,$$

we have

$$\begin{aligned}
y_i &= \alpha^s + \beta^s s_i + e_i^s \\
y_i &= \alpha^m + \beta^m s_i + \gamma^m x_i^m + e_i^m \\
x_i^m &= \delta_0 + \delta s_i + \ u_i + m_i.
\end{aligned}$$

Treat s_i, u_i, e_i, and m_i as the underlying disturbances which in turn will determine x_i, y_i and e_i^s. Because e_i is a residual uncorrelated with s_i and

x_i, it follows that $Cov\left(e_i, u_i\right) = 0$. We normalize s_i to a mean zero variable. Hence,

$$\begin{array}{c} s_i \\ u_i \\ e_i \\ m_i \end{array} \sim \left(\left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \end{array} \right], \left[\begin{array}{cccc} \sigma_s & 0 & 0 & 0 \\ 0 & \sigma_u & 0 & 0 \\ 0 & 0 & \sigma_e & 0 \\ 0 & 0 & 0 & \sigma_m \end{array} \right] \right).$$

We want to test $\beta^s - \beta^m = 0$. Of course

$$\beta^s - \beta^m = \delta \gamma^m,$$

and we will assume $\gamma \neq 0$, so that

$$\beta^s - \beta^m = 0 \Leftrightarrow \delta = 0.$$

The test statistic is

$$t_\beta = \frac{\widehat{\beta}^s - \widehat{\beta}^m}{\sqrt{Var\left(\widehat{\beta}^s\right) + Var\left(\widehat{\beta}^m\right) - 2Cov\left(\widehat{\beta}^s, \widehat{\beta}^m\right)}},$$

which is asymptotically standard normal. The sampling variances are

$$\begin{aligned} Var\left(\widehat{\beta}^m\right) &= \frac{1}{n} \frac{Var\left(e_i^m\right)}{Var\left(\widetilde{s}_i^m\right)} \\ Var\left(\widehat{\beta}^s\right) &= \frac{1}{n} \frac{Var\left(e_i^s\right)}{\sigma_s^2}, \end{aligned}$$

which we will now derive in terms of the underlying parameters.

We start by deriving $Var\left(\widehat{\beta}^m\right)$. $\widetilde{s}_i^m$ is given by

$$s_i = \pi_0 + \pi_1 x_i^m + \widetilde{s}_i^m$$

and

$$\begin{aligned} Var\left(x_i^m\right) &= \delta^2 \sigma_s^2 + \sigma_u^2 + \sigma_m^2 \\ Cov\left(x_i^m, s_i\right) &= \delta \sigma_s^2 \end{aligned}$$

so

$$\begin{aligned} \sigma_s^2 &= \pi_1^2 Var\left(x_i^m\right) + Var\left(\widetilde{s}_i^m\right) \\ \sigma_s^2 &= \frac{Cov\left(x_i^m, s_i\right)^2}{Var\left(x_i^m\right)^2} Var\left(x_i^m\right) + Var\left(\widetilde{s}_i^m\right) \\ &= \frac{\delta^2 \sigma_s^4}{\delta^2 \sigma_s^2 + \sigma_u^2 + \sigma_m^2} + Var\left(\widetilde{s}_i^m\right) \\ Var\left(\widetilde{s}_i^m\right) &= \frac{\sigma_s^2 \left(\sigma_u^2 + \sigma_m^2\right)}{\delta^2 \sigma_s^2 + \sigma_u^2 + \sigma_m^2} \end{aligned}$$

Next, we need $Var\left(e_i^m\right)$. Define the reliability

$$\lambda = \frac{Var\left(x_i\right)}{Var\left(x_i^m\right)} = \frac{\delta^2\sigma_s^2 + \sigma_u^2}{\delta^2\sigma_s^2 + \sigma_u^2 + \sigma_m^2}$$

and the R^2 of the regression of s_i on x_i^m

$$\begin{aligned} R^2 &= 1 - \frac{Var\left(\widetilde{s}_i^m\right)}{\sigma_s^2} \\ &= 1 - \frac{\sigma_u^2 + \sigma_m^2}{\delta^2\sigma_s^2 + \sigma_u^2 + \sigma_m^2} = \frac{\delta^2\sigma_s^2}{\delta^2\sigma_s^2 + \sigma_u^2 + \sigma_m^2}, \end{aligned}$$

Then

$$\begin{aligned} \beta^m &= \beta + \gamma\delta\frac{1-\lambda}{1-R^2} \\ &= \beta + \gamma\delta\frac{\frac{\sigma_m^2}{\delta^2\sigma_s^2+\sigma_u^2+\sigma_m^2}}{\frac{\sigma_u^2+\sigma_m^2}{\delta^2\sigma_s^2+\sigma_u^2+\sigma_m^2}} = \beta + \gamma\delta\frac{\sigma_m^2}{\sigma_u^2 + \sigma_m^2}, \end{aligned}$$

and

$$\begin{aligned} \gamma^m &= \gamma\frac{\lambda - R^2}{1-R^2} \\ &= \gamma\frac{\frac{\sigma_u^2}{\delta^2\sigma_s^2+\sigma_u^2+\sigma_m^2}}{\frac{\sigma_u^2+\sigma_m^2}{\delta^2\sigma_s^2+\sigma_u^2+\sigma_m^2}} = \gamma\frac{\sigma_u^2}{\sigma_u^2 + \sigma_m^2}. \end{aligned}$$

Using

$$\theta = \frac{\sigma_m^2}{\sigma_u^2 + \sigma_m^2}$$

we have

$$\begin{aligned} \beta^m &= \beta + \gamma\delta\theta \\ \gamma^m &= \gamma\left(1-\theta\right) \end{aligned}$$

Using these results in

$$\begin{aligned} y_i &= \alpha^m + \beta^m s_i + \gamma^m x_i^m + e_i^m \\ &= \alpha^m + \left(\beta + \gamma\delta\theta\right)s_i + \gamma\left(1-\theta\right)x_i^m + e_i^m \\ &= \left(\alpha^m + \gamma\left(1-\theta\right)\delta_0\right) + \left(\beta + \gamma\delta\right)s_i + \gamma\left(1-\theta\right)\left(u_i + m_i\right) + e_i^m \end{aligned}$$

$$\begin{aligned} y_i &= \alpha + \beta s_i + \gamma\left(\delta_0 + \delta s_i + \ u_i\right) + e_i \\ &= \left(\alpha + \gamma\delta_0\right) + \left(\beta + \gamma\delta\right)s_i + \gamma u_i + e_i \end{aligned}$$

Matching residuals yields

$$\begin{aligned}
\gamma u_i + e_i &= \gamma\left(1-\theta\right)\left(u_i + m_i\right) + e_i^m \\
e_i^m &= \gamma\theta u_i - \gamma\left(1-\theta\right) m_i + e_i \\
Var\left(e_i^m\right) &= \gamma^2\theta^2\sigma_u^2 + \gamma^2\left(1-\theta\right)^2\sigma_m^2 + \sigma_e^2 \\
&= \gamma^2\left(\left(\frac{\sigma_m^2}{\sigma_u^2+\sigma_m^2}\right)^2\sigma_u^2 + \left(\frac{\sigma_u^2}{\sigma_u^2+\sigma_m^2}\right)^2\sigma_m^2\right) + \sigma_e^2 \\
&= \gamma^2\theta\sigma_u^2 + \sigma_e^2
\end{aligned}$$

So

$$\begin{aligned}
Var\left(\widehat{\beta}^m\right) &= \frac{1}{n}\frac{Var\left(e_i^m\right)}{Var\left(\widetilde{s}_i^m\right)} \\
&= \frac{1}{n}\frac{\gamma^2\theta\sigma_u^2+\sigma_e^2}{\frac{\sigma_s^2\left(\sigma_u^2+\sigma_m^2\right)}{\delta^2\sigma_s^2+\sigma_u^2+\sigma_m^2}} \\
&= \frac{1}{n}\left(\frac{\delta^2\left(1-\theta\right)}{\sigma_u^2} + \frac{1}{\sigma_s^2}\right)\left(\gamma^2\theta\sigma_u^2+\sigma_e^2\right)
\end{aligned}$$

and similarly we can derive

$$\begin{aligned}
Var\left(\widehat{\gamma}^m\right) &= \frac{1}{n}\frac{Var\left(e_i^m\right)}{Var\left(\widetilde{x}_i^m\right)} \\
&= \frac{1}{n}\frac{\gamma^2\theta\sigma_u^2+\sigma_e^2}{\sigma_u^2+\sigma_m^2} \\
&= \frac{1-\theta}{n}\left(\gamma^2\theta + \frac{\sigma_e^2}{\sigma_u^2}\right)
\end{aligned}$$

Now we derive $Var\left(\widehat{\beta}^s\right)$, which does not involve the mismeasured x_i. Comparing the short and the long regression, the relationship between the residuals is

$$\begin{aligned}
y_i &= \alpha^s + \beta^s s_i + e_i^s \\
&= \alpha^s + \left(\beta+\gamma\delta\right) s_i + e_i^s \\
y_i &= \alpha + \beta s_i + \gamma\left(\delta_0 + \delta s_i + \ u_i\right) + e_i \\
&= \alpha + \gamma\delta_0 + \left(\beta+\gamma\delta\right) s_i + \gamma u_i + e_i \\
e_i^s &= \gamma u_i + e_i,
\end{aligned}$$

and hence

$$Var\left(e_i^s\right) = \gamma^2\sigma_u^2 + \sigma_e^2,$$

so

$$Var\left(\widehat{\beta}^s\right)=\frac{1}{n}\frac{Var\left(e_i^s\right)}{\sigma_s^2}=\frac{1}{n}\frac{\gamma^2\sigma_u^2+\sigma_e^2}{\sigma_s^2}.$$

Finally, we derive $Cov\left(\widehat{\beta}^s,\widehat{\beta}\right)$. Using

$$\begin{aligned}\widehat{\beta}^s-\beta^s &= \frac{\sum e_i^s s_i}{\sum s_i^2}\\ \widehat{\beta}^m-\beta^m &= \frac{\sum e_i^m \widetilde{s}_i^m}{\sum\left(\widetilde{s}_i^m\right)^2}\end{aligned}$$

we have

$$\sqrt{n}\left[\begin{array}{c}\sum e_i^s s_i\\ \sum e_i^m\widetilde{s}_i^m\end{array}\right]\xrightarrow{d}N\left(0,\left[\begin{array}{cc}E\left[\left(e_i^s\right)^2 s_i^2\right] & E\left[e_i^s e_i^m s_i\widetilde{s}_i^m\right]\\ E\left[e_i^s e_i^m s_i\widetilde{s}_i^m\right] & E\left[\left(e_i^m\right)^2\left(\widetilde{s}_i^m\right)^2\right]\end{array}\right]\right).$$

In addition, using

$$\begin{aligned}p\lim\frac{1}{n}\sum s_i^2 &= \sigma_s^2\\ p\lim\frac{1}{n}\sum\left(\widetilde{s}_i^m\right)^2 &= Var\left(\widetilde{s}_i^m\right),\end{aligned}$$

by Slutsky's theorem

$$\begin{aligned}Cov\left(\widehat{\beta}^s,\widehat{\beta}\right) &= \frac{1}{n}\frac{E\left[e_i^s e_i^m s_i\widetilde{s}_i^m\right]}{\sigma_s^2 Var\left(\widetilde{s}_i^m\right)}\\ &= \frac{1}{n}\frac{E\left[E\left(e_i^s e_i^m|s_i,\widetilde{s}_i^m\right)s_i\widetilde{s}_i^m\right]}{\sigma_s^2 Var\left(\widetilde{s}_i^m\right)}\\ &= \frac{1}{n}\frac{Cov\left(e_i^s,e_i^m\right)Var\left(\widetilde{s}_i^m\right)}{\sigma_s^2 Var\left(\widetilde{s}_i^m\right)}\\ &= \frac{1}{n}\frac{Cov\left(e_i^s,e_i^m\right)}{\sigma_s^2}.\end{aligned}$$

Using our earlier result that

$$y_i=\left(\alpha^m+\gamma\left(1-\theta\right)\delta_0\right)+\left(\beta+\gamma\delta\right)s_i+\gamma\left(1-\theta\right)\left(u_i+m_i\right)+e_i^m$$

and comparing this to the short regression

$$y_i=\alpha^s+\beta^s s_i+e_i^s,$$

we have

$$e_i^s=\gamma\left(1-\theta\right)\left(\ u_i+m_i\right)+e_i^m.$$

Note that $u_i + m_i$ is the residual from a regression of x_i^m on s_i, we have

$$Cov\left(e_i^s, e_i^m\right) = Var(e_i^m) = \gamma^2\theta\sigma_u^2 + \sigma_e^2$$

and hence

$$Cov\left(\widehat{\beta}^s, \widehat{\beta}^m\right) = \frac{1}{n}\frac{\gamma^2\theta\sigma_u^2 + \sigma_e^2}{\sigma_s^2}.$$

Returning to the test statistic

$$t_\beta = \frac{\widehat{\beta}^s - \widehat{\beta}^m}{\sqrt{Var\left(\widehat{\beta}^s\right) + Var\left(\widehat{\beta}^m\right) - 2Cov\left(\widehat{\beta}^s, \widehat{\beta}^m\right)}}$$

we first derive

$$\begin{aligned}
\frac{1}{n}V_\beta\left(d;\gamma\right) &= Var\left(\widehat{\beta}^s\right) + Var\left(\widehat{\beta}^m\right) - 2Cov\left(\widehat{\beta}^s, \widehat{\beta}^m\right) \\
&= \frac{1}{n}\frac{\gamma^2\sigma_u^2 + \sigma_e^2}{\sigma_s^2} + \frac{1}{n}\frac{\delta^2\sigma_s^2 + \sigma_u^2 + \sigma_m^2}{\sigma_s^2\left(\sigma_u^2 + \sigma_m^2\right)}\left(\gamma^2\theta\sigma_u^2 + \sigma_e^2\right) - 2\frac{1}{n}\frac{\gamma^2\theta\sigma_u^2 + \sigma_e^2}{\sigma_s^2} \\
&= \frac{1}{n}\left(\frac{\left(\gamma^2\left(1-2\theta\right)\sigma_u^2 - \sigma_e^2\right)\left(\sigma_u^2 + \sigma_m^2\right) + \left(\delta^2\sigma_s^2 + \sigma_u^2 + \sigma_m^2\right)\left(\gamma^2\theta\sigma_u^2 + \sigma_e^2\right)}{\sigma_s^2\left(\sigma_u^2 + \sigma_m^2\right)}\right) \\
&= \frac{1}{n}\left(1-\theta\right)\left(\frac{\gamma^2\sigma_u^2}{\sigma_s^2} + \theta\delta^2\gamma^2 + \frac{\delta^2\sigma_e^2}{\sigma_u^2}\right)
\end{aligned}$$

Note that

$$\beta^s - \beta^m = \delta\gamma^m = \delta\gamma\left(1-\theta\right)$$

so the power function of the coefficient comparison test is

$$Power_{t_\beta}\left(d;\gamma\right) = 1 - \Phi\left(1.96 - d\frac{\sqrt{n}\gamma\left(1-\theta\right)}{\sqrt{V_\beta\left(d;\gamma\right)}}\right) + \Phi\left(-1.96 - d\frac{\sqrt{n}\gamma\left(1-\theta\right)}{\sqrt{V_\beta\left(d;\gamma\right)}}\right).$$

7.2 Comparison with Oster (2014)

Oster's (2014) formulation of the causal regression takes the form

$$y_i = \alpha + \beta s_i + w_{1i} + w_{2i} + e_i,$$

where w_{1i} is an observed covariate and w_{2i} is an unobserved covariate, uncorrelated with w_{1i}. To map this into our setup, think of the true x_i as capturing both w_{1i} and w_{2i}, i.e. $x_i = w_{1i} + w_{2i}$. Furthermore,

$$\frac{Cov(s_i, w_{1i})}{\sigma_1^2} = \frac{Cov(s_i, w_{2i})}{\sigma_2^2},$$

where σ_1^2 and σ_2^2 are the variances of w_{1i} and w_{2i}, respectively. Then, Oster's regression can be written as

$$y_i = \alpha + \beta s_i + x_i + e_i,$$

which is our regression with $\gamma = 1$. Our observed $x_i^m = w_{1i}$, so measurement error $m_i = -w_{2i}$. Measurement error here is mean reverting with $\sigma_\mu^2 = 0$ and $\kappa = \left(\sigma_1 - \sqrt{\sigma_1^2 + \sigma_2^2}\right) / \sqrt{\sigma_1^2 + \sigma_2^2}$.