cemmap

centre for microdata methods and practice

Testing for the presence of measurement error in Stata

Young Jun Lee
Daniel Wilhelm

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP47/19

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

An ESRC Research Centre

Testing for the Presence of Measurement Error in Stata

Young Jun Lee
Department of Economics
University of Copenhagen
Copenhagen, Denmark
yjl@econ.ku.dk

Daniel Wilhelm
Department of Economics
UCL, CeMMAP
London, UK
d.wilhelm@ucl.ac.uk

Abstract. In this paper, we describe how to test for the presence of measurement error in explanatory variables. First, we discuss the test of such hypotheses in parametric models such as linear regressions and then introduce a new *Stata* command [R] **dgmtest** for a nonparametric test proposed in Wilhelm (2018). To illustrate the new command, we provide Monte Carlo simulations and an empirical application to testing for measurement error in administrative earnings data.

Keywords: st0001, nonparametric test, measurement error, measurement error bias

1 Introduction

In this paper, we describe how to test for the presence of measurement error in explanatory variables. Specifically, consider an outcome Y (e.g. earnings) that depends on an explanatory variable X^* (e.g. schooling). We do not observe X^* directly, but only two variables X and Z that are related to X^*. We suspect X to be an error-contaminated measurement of X^* (e.g. schooling as reported in a survey) and Z is a variable related to X^*, perhaps an instrument (e.g. distance to college) or a repeated measurement (e.g. schooling as reported in another survey). The hypothesis of no measurement error in X is

$$H_0\colon P(X = X^*) = 1, \tag{1}$$

In the schooling example, testing H_0 could be useful as a first-step model specification test to tell the researcher whether measurement error is an important feature of the data that should be modelled. However, testing H_0 may also be of direct economic interest because, for example, the null of no measurement error can often be shown to be implied by the absence of frictions in a structural economic model (e.g. Chetty (2012), Wilhelm (2018)). A test of H_0 can therefore be interpreted as a test of the absence of such frictions.

In a finite sample, we may not be able to detect measurement error even though X is in fact mismeasured. The reason is that measurement errors might be small relative to the overall sampling noise. In this sense, we can interpret the test of H_0 as finding out whether measurement error is severe enough for the data to tell the difference between models with and without measurement error.

In this paper, we describe how to test for the presence of measurement error without imposing any parametric restrictions and, in fact, without requiring the model to be identified. Both of these aspects are important for empirical practice. First, when testing for measurement error it is important to allow for nonlinearities in the relationship of Y and X^* because measurement error in X can make the relationship appear nonlinear when it isn't and make it appear linear when it isn't (Chesher (1991)). To disentangle measurement error from nonlinearities therefore requires a procedure that can allow for nonlinearities. Second, nonparametric measurement error models are identified only under fairly strong conditions and their estimation involves complicated procedures such as Fourier transforms and operator inversions (Schennach (2013, 2016), Hu (2017)). However, Wilhelm (2018) shows that testing for the presence of measurement error does not require identification of the model and is thus possible without such strong assumptions. In particular, the test is able to detect a wide range of nonclassical measurement error models, i.e. models in which the measurement error depends on the true, latent variable. Another byproduct of avoiding identification of the model is that complicated estimation techniques are not necessary. In fact, the test we describe only employs standard nonparametric regression techniques.

The null hypothesis depends on the latent variable X^* and thus cannot directly be tested. In Section 2, this paper therefore first describes how to convert the null hypothesis into a testable restriction in terms of the observable variables Y, X, Z in a simple example, a linear regression model. In this model, H_0 can easily be tested using existing *Stata* commands following Hausman (1978). Section 3 then describes the extension of such ideas to the nonparametric framework as recently proposed by Wilhelm (2018). We introduce the new *Stata* command [R] **dgmtest** that implements a test of H_0 without imposing any parametric restrictions. Section 4 reports the results of Monte Carlo simulations for this new command and Section 5 concludes with an empirical example in which we show how to test for measurement error in administrative earnings data.

Related Literature Mahajan (2006) proposes a test for the presence of measurement error when the explanatory variable X^* and the observed measure X are binary. There are also some existing tests for the presence of measurement error in parametric models that require identification and consistent estimators of the model: Hausman (1978), Chesher (1990), Chesher et al. (2002), Hahn and Hausman (2002), and Hu (2008). Related to Hausman (1978), in empirical work it is common to estimate linear regressions by OLS and IV, and then attribute a difference in the two estimates to the presence of measurement error, treating the IV estimate as the consistent and unbiased one. Of course, this strategy is valid only if the true relationship of interest is in fact linear, the measurement error is classical, and the model is identified. None of these assumptions are required in the nonparametric approach described in the present paper.

In principle, one could imagine constructing a test for the presence of measurement error by comparing an estimator of the model that accounts for the possibility of measurement error with one that ignores it, similar in spirit to the work by Durbin (1954), Wu (1973), and Hausman (1978). If the difference between the two is statistically significant, then one could conclude that this is evidence for the presence of measurement error. However, this strategy would require identification and consistent estimation of the measurement error model, which leads to overly strong assumptions, the necessity of solving ill-posed inverse problems in the continuous variable case, and potentially highly variable estimators. These difficulties can all be avoided by the nonparametric approach described in this paper.

2 Linear Regression Model

Consider the linear regression model for an outcome Y and an explanatory variable X^*, assuming for simplicity that there are no further regressors (the extension to the presence of additional controls is straightforward and discussed below),

$$Y = \alpha + \beta X^* + \varepsilon, \qquad E(\varepsilon X^*) = 0. \tag{2}$$

Instead of X^*, we observe a measurement X of X^* and an instrumental variable (IV) Z which depends on X^* (i.e. $E(X^* Z) \neq 0$), but is excluded from the outcome equation (i.e. $E(\varepsilon Z) = 0$). Testing for the presence of measurement error in this context is straightforward (Hausman (1978)). Under the null of no measurement error OLS consistently estimates β, but under the alternative of some measurement error it is inconsistent. The IV estimator, however, is consistent under both the null and the alternative. Therefore, one can simply compute both estimators and compare them. If their difference is statistically significant, that indicates the presence of measurement error.

To better understand the connection to the nonparametric test described in the next section it might be instructive to note that the test based on the difference of OLS and IV estimators is equivalent to testing significance in an expanded regression. To see this, suppose there is no measurement error in X, then

$$Y = \alpha + \beta X + \varepsilon, \qquad E(\varepsilon X) = 0.$$

Therefore, when regressing Y onto both X and Z, the exclusion of the IV implies that the coefficient of Z must be zero[1], i.e. we test the hypothesis of no measurement error by instead testing

$$\bar{\gamma} = 0 \tag{3}$$

in the regression

$$Y = \bar{\alpha} + \bar{\beta} X + \bar{\gamma} Z + \bar{\varepsilon}.$$

In conclusion, we have shown that the null of no measurement error, (1), implies (3) in the linear regression model. The only assumption for this to be true is that (2) holds and that the IV is excluded from the outcome equation, i.e. $E(\varepsilon Z) = 0$. Therefore, a rejection of the restriction (3) implies a rejection of the hypothesis of no measurement error, (1).

[1] Hausman (1978) suggests running a regression of Y on X and the projection $\hat{X}$ of X onto Z. Then, H_0 implies that the coefficient of $\hat{X}$ must be zero.

However, without further assumptions, failing to reject (3) does not necessarily imply failing to reject the null of no measurement error, (1). Suppose $X = X^* + \eta_X$, so that η_X represents the measurement error in X. If the measurement error in X is assumed to be classical (i.e. it is uncorrelated with the latent regressor, $E(X^*\eta_X) = 0$), uncorrelated with the regression error, $E(\varepsilon\eta_X) = 0$, and some further regularity conditions hold, then it is easy to see that the null hypothesis H_0 not only implies but is, in fact, also implied by (3). Therefore, failing to reject (3) may be interpreted as failing to reject H_0 and rejecting (3) may be interpreted as rejecting H_0.

Consider the following simulated example that illustrates the finite sample performance of the test by Hausman (1978). First, we simulate data without measurement error in the regressor ($X = X^*$):

```
. set obs 200
. gen double z = rnormal(0,1)
. gen double u = rnormal(0,0.5)
. gen double e = rnormal(0,0.5)
. gen xs = 0.5*z + u
. gen x = xs
. gen y = xs + e
```

Then we regress Y on X and Z (and a constant):

```
. reg y x z

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(2, 197)       =    190.55
       Model |  110.874464         2   55.437232   Prob > F        =    0.0000
    Residual |  57.3145743       197  .290936925   R-squared       =    0.6592
-------------+----------------------------------   Adj R-squared   =    0.6558
       Total |  168.189038       199  .845171047   Root MSE        =    .53939

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.052337   .0747708    14.07   0.000      .904883    1.199791
           z |  -.0218478   .0585315    -0.37   0.709    -.1372765     .093581
       _cons |  -.0184318   .0381422    -0.48   0.629    -.0936512    .0567876
------------------------------------------------------------------------------
```

to find that Z is not significant at any reasonable confidence level (p-value is 0.709). Therefore, we fail to reject the null of no measurement error as expected. Now, we generate a measurement error-contaminated regressor ($X \neq X^*$):

```
. gen double eta = rnormal(0,0.5)
. gen x = xs + eta
```

Again, we regress Y on X and Z (and a constant):

```
. reg y x z

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(2, 197)       =    116.36
       Model |  91.0844045         2  45.5422023   Prob > F        =    0.0000
    Residual |  77.1046338       197   .39139408   R-squared       =    0.5416
-------------+----------------------------------   Adj R-squared   =    0.5369
       Total |  168.189038       199  .845171047   Root MSE        =    .62561
```

```
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .5981834   .0608373     9.83   0.000     .4782075    .7181593
           z |   .2439546   .0578135     4.22   0.000     .1299419    .3579674
       _cons |  -.0331463   .0442702    -0.75   0.455    -.1204506     .054158
------------------------------------------------------------------------------
```

to find that now Z is significant at every reasonable confidence level (p-value is 0.000). Therefore, we strongly reject the null of no measurement error.

In the presence of additional, correctly measured, controls in the regression model, we would proceed exactly as above except that we would include the additional controls in the regression command.

3 Nonparametric Model – the New dgmtest Command

While the approach to testing H_0 in the previous section is straightforward and intuitive, its validity relies on strong assumptions: linearity in the outcome equation and classical measurement error in X. Since nonlinearities in the regression equation and measurement error in X may manifest themselves in similar ways (Chesher (1991)), it is important to allow for nonlinearities in the relationship between Y and X^* when testing for measurement error. In addition, a large literature has documented that measurement error in economic data is rarely classical (see the survey by Bound et al. (2001), for example). In this section, we describe how to test H_0 in nonlinear models with nonclassical measurement error.

Suppose the variable Z is related to X^*, but together with the measurement X is excluded from the outcome model in the sense that

$$E(Y|X^*, X, Z) = E(Y|X^*) \qquad \text{a.s.} \tag{4}$$

i.e. they can affect outcomes only through the true explanatory variable X^*. Then it is easy to see that, under H_0, Z must also be excluded from the outcome equation conditional on the observed X:

$$E(Y|X, Z) = E(Y|X) \qquad \text{a.s..} \tag{5}$$

Unlike H_0, this is a restriction that depends only on observables and can directly be tested without making any parametric assumptions about how the conditional mean of Y depends on X^*. The test by Delgado and Gonzalez Manteiga (2001) introduced in the next subsection and implemented in the new *Stata* command [R] **dgmtest**, for instance, directly tests the restriction (5). Because of the above argument it can be interpreted as a test of the original null of interest, the null of no measurement error in (1).

The exclusion restriction (4) is standard in the literature on identification and estimation of measurement error models (Carroll et al. (2006), Chen et al. (2011), Schennach

(2013, 2016), Hu (2017)) and has already been justified in a wide range of empirical applications. Since the assumption is central to the validity of the test for measurement error, we now provide a few examples.

Consider a generic production problem in which Y is an output that is produced from a vector of inputs X^*. The inputs are measured by the vector X and Z provides an alternative vector of measurements. In this context, the exclusion restriction is often a natural assumption as it requires the "true" inputs X^* to be the factors that matter for production, not the measurements (X, Z). Therefore, conditional on knowing X^*, the measurements X and Z should not provide any additional information about the output Y. Cunha et al. (2010), Heckman et al. (2013), Attanasio et al. (2015), Attanasio et al. (2017) are examples of empirical papers in the skill formation literature that have justified the exclusion restriction in this fashion. The same argument also applies to many other production problems in which inputs are difficult to measure (e.g. Olley and Pakes (1996)).

In the empirical part of Wilhelm (2018) and in Section 5 below, Y, X, Z are three measurements of earnings, but Y and (X, Z) come from two different data sources, one from a survey and the other from an administrative dataset. We then argue the exclusion restriction holds because the error in Z has a very different origin from the error in Y, at least conditional on X^*.

There are many other empirical applications that also impose the exclusion restriction (4): for instance, Altonji (1986) studies labor supply, Kane and Rouse (1995) and Kane et al. (1999) the returns to education, Card (1996) the effect of unions on the wage structure, Hu et al. (2013) auctions with unobserved heterogeneity, Feng and Hu (2013) unemployment dynamics, and Arellano et al. (2017) earnings dynamics.

Wilhelm (2018) actually shows that, under additional assumptions, H_0 not only implies, but is also implied by the observable restriction (5). Therefore, failing to reject (5) may be interpreted as failing to reject H_0 and rejecting (5) may be interpreted as rejecting H_0.

The main assumptions required for this equivalence result are, first, the exclusion restriction (4), second, a relevance condition that ensures Z is sufficiently strongly related to X^* and, third, monotonicity of the conditional mean function $x^* \mapsto E(Y|X^* = x^*)$.

To satisfy the relevance condition we need to be able to find two values of Z, say z_1, z_2, such that the probability mass functions of $X^*|Z = z_1$ and $X^*|Z = z_2$ do not cross more than once. This assumption is testable under the additional assumption that X and X^* are sufficiently strongly monotonically related because, in that case, we must also have that the probability mass functions of $X|Z = z_1$ and $X|Z = z_2$ do not cross more than once (see Appendix A.3 in Wilhelm (2018)). Finally, monotonicity of the relationship between the outcome and the explanatory variable is a weak assumption that is often directly implied by economic theory, e.g. when the conditional mean $E(Y|X^* = x^*)$ is a production, cost, or utility function. Examples can be found in Matzkin (1994), Olley and Pakes (1996), Cunha et al. (2010), Blundell et al. (2012, 2016), Kasy (2014), Wilhelm (2015), Hoderlein et al. (2016), Chetverikov and Wilhelm

(2017), among many others.

We now heuristically explain why the exclusion restriction, the relevance condition and the monotonicity condition together guarantee equivalence of H_0 and (5). We have already argued why H_0 implies (5) under the exclusion restriction, so we only need to show that the reverse holds as well.

Consider the special case when X^*, X are continuously distributed and X^*, X, Z are scalars. Suppose the observable implication (5) holds. Then, for any two values z_1, z_2, we have $E[Y|X, Z = z_1] = E[Y|X, Z = z_2]$. Then, by the exclusion restriction,

$$\int E[Y|X^*] \, d(P_{X^*|X,Z=z_1} - P_{X^*|X,Z=z_2}) = 0.$$

If $E[Y|X^* = \cdot]$ is differentiable, then integration by parts yields

$$\int \left(P_{X^*|X=x,Z=z_1}(x^*) - P_{X^*|X=x,Z=z_2}(x^*)\right) \frac{\partial E[Y|X^* = x^*]}{\partial x^*} dx^* = 0. \tag{6}$$

We want to show that this equation implies the null hypothesis H_0. On the contrary, assume that this is not the case. To generate a contradiction, we want to ensure that (6) does not hold under the alternative H_1. This is the case, for example, when $E[Y|X^* = \cdot]$ is monotone (and not constant) and $P_{X^*|X=x,Z=z_2}$ first-order stochastically dominates (FOSD) $P_{X^*|X=x,Z=z_1}$ (and they are not equal) under H_1. The relevance condition of Wilhelm (2018) ensures that this FOSD holds. The monotonicity assumption, on the other hand, implies that the derivative of the conditional expectation does not change sign (and is nonzero somewhere) and the dominance condition implies that the difference of the conditional distributions is nonnegative (and positive somewhere). In conclusion, the integral in (6) is nonzero under H_1, yielding the desired contradiction, so the null of no measurement error must hold. For more details on the exact assumptions and arguments, see Wilhelm (2018).

In some applications, Z may be excluded from the outcome equation only after conditioning on some additional, correctly measured controls W, i.e. the exclusion restriction (4) is replaced by

$$E(Y|X^*, W, X, Z) = E(Y|X^*, W) \qquad \text{a.s.}. \tag{7}$$

This additional conditioning on W is necessary, for example, in cases in which W determines both Y and (X, Z). Under (7), the null hypothesis H_0 then implies

$$E(Y|X, W, Z) = E(Y|X, W) \qquad \text{a.s.}. \tag{8}$$

The null hypothesis is, in fact, equivalent to (8) under conditions similar to those required for the equivalence of H_0 and (5). In the implementation of the test we allow for two types of additional controls, say $W = (W_1, W_2)$, where the vector W_1 is included in the conditional mean in a nonseparable fashion and the vector W_2 is additively separable and linear:

$$E(Y|X, W) = g(X, W_1) + \pi' W_2 \tag{9}$$

for some function g and some vector of coefficients π. (5)

There exist many nonparametric tests of the conditional mean independence in (5) and (8), for example Gozalo (1993), Fan and Li (1996), Delgado and Gonzalez Manteiga (2001), Mahajan (2006), and Huang et al. (2016). Therefore, any of those could be used for nonparametrically testing for the presence of measurement error. In the presence, of several additional covariates W, the curse of dimensionality may however cause fully nonparametric tests to be infeasible. We therefore recommend the semiparametric, partially linear model in (9) as a more practical approach in such cases.

In the following subsections we introduce a new *Stata* command [R] **dgmtest** that implements the test by Delgado and Gonzalez Manteiga (2001). This test has some desirable properties such as relatively simple implementation and its ability to detect alternatives at the $\sqrt{n}$-rate.

3.1 The Test by Delgado and Gonzalez Manteiga (2001)

We briefly describe the approach by Delgado and Gonzalez Manteiga (2001) for testing the conditional mean independence (5). There are many other reasons why one might want to test such a restriction and the test for the presence of measurement error as described in this paper is only one of these. To simplify the description, we focus on the case in which there are no additional controls W.

The authors rewrite the null hypothesis of conditional mean independence, (5), as

$$E[T(X,Z)] = 0,$$

where

$$T(x,z) := E\left[f_X(X)\{Y - E(Y|X)\}1\{X \leq x\}1\{Z \leq z\}\right],$$

$1\{A\}$ is equal to one if the event A holds, zero otherwise, and f_X is the density of X. Given a random sample $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ from the distribution of (Y, X, Z), consider the empirical analogue $T_n(x,z)$ of $T(x,z)$:

$$T_n(x,z) := \frac{1}{n^2}\sum_i \sum_j \frac{1}{h} K\left(\frac{X_i - X_j}{h}\right)(Y_i - Y_j)\, 1\{X_i \leq x\}1\{Z_i \leq z\}.$$

where h is a bandwidth parameter and K a kernel function. Delgado and Gonzalez Manteiga (2001) propose two test statistics: the Cramér-von Mises statistic $T_n := n\sum_{i=1}^n T_n(X_i, Z_i)^2$ and the Kolmogorov-Smirnov statistic $T_n := \sup_{x,z,y} |\sqrt{n}T_n(x,z)|$. Critical values of the test are computed using the bootstrap procedure described in Delgado and Gonzalez Manteiga (2001).

Testing the version with additional controls, (8), is a simple extension of the above test. In the presence of additively separable controls W_2, we perform the test in two steps. First, we compute an estimator $\hat{\pi}$ of π as in Robinson (1988). Then, we apply Delgado and Gonzalez Manteiga (2001)'s test as described above, replacing Y_i by $Y_i - \hat{\pi}' W_{2i}$.

3.2 Syntax

The [R] **dgmtest** command implements the test by Delgado and Gonzalez Manteiga (2001). The syntax of the command is as follows:

```
dgmtest depvar expvar [if] [in] [, qz(#) qw2(#) teststat(string)
   kernel(string) bootdist(string) bw(#) bootnum(#) ngrid(#) qgrid(#)]
```

The two required arguments of the command are *depvar* (the outcome variable Y) and *expvar* (a list of variables containing all elements of X, W_1, W_2, and Z). Therefore, *expvar* should consist of at least two variables in which case the first is taken to be X and the second to be Z. If there are more than two variables, then the options *qz* and *qw2* determine which variables in the list are X, W_1, W_2, and Z. For instance, if *expvar* contains 6 variables, *qz* equals the default value of 1, and *qw2* is equal to 2, then the first 3 variables in the list are interpreted as (X, W_1) (which one is X and which one is W_1 does not matter as the test treats both types of variables exactly the same), the fourth and fifth variables are interpreted as W_2, and the sixth variable as Z.

3.3 Options

We now describe the options of the command. If options are left unspecified, the command runs on the default settings.

`qz(integer)` is the dimension of Z. The default is 1.

`qw2(integer)` is the dimension of W_2. The default is 0, which means there are no additional controls W_2.

`teststat(string)` is the type of test statistic to be used: `CvM` and `KS` represent the Cramér-von Mises and Kolmogorov-Smirnov statistics, respectively. The default is `CvM`.

`kernel(string)` is the kernel function. The default kernel is the Epanechnikov kernel (`epanechnikov`). Alternatively, we can choose one among two other Epanechnikov kernels order of 2 and 4 with the support $[-1, 1]$ (`epan2` and `epan4`), biweight kernel (`biweight`), Gaussian kernel (`normal`), rectangle kernel (`rectangle`), and triangular kernel (`triangular`).

`bootdist(string)` is the distribution of the bootstrap multiplier variable. Following Delgado and Gonzalez Manteiga (2001), it should have a zero mean and unit variance. The default is `mammen` in Härdle and Mammen (1993), which is the two point distribution attaching masses $\left(\sqrt{5}+1\right)/2\sqrt{5}$ and $\left(\sqrt{5}-1\right)/2\sqrt{5}$ to the points $-\left(\sqrt{5}-1\right)/2$ and $\left(\sqrt{5}+1\right)/2$, respectively. Alternatively, we can choose the Rademacher distribution (`rademacher`) or the continuous uniform distribution on $\left(-\sqrt{3}, \sqrt{3}\right)$ (`uniform`).

`bw(real)` is the bandwidth h, taken to be the same for every component of (X, W_1).

The default is $n^{-1/3q}$, which is a rule of thumb in Delgado and Gonzalez Manteiga (2001), where n is the sample size and q the dimension of (X, W_1).

`bootnum(integer)` is the number of bootstrap samples for the computation of the test's critical value. The default is 500.

`ngrid(integer)` is the number of equally spaced grid points used to compute the supremum of the Kolmogorov-Smirnov statistic, if that statistic is chosen via the option `teststat`. The default is 0, which means that the sample serves as the grid. Choosing 0 is required for calculating the exact Kolmogorov-Smirnov statistic, but it is a burden when we perform a simulation with a large sample, so one might want to choose a positive number smaller than the sample size in that case. The user need not specify this if `CvM` is used for `teststat`.

`qgrid(real)` is a quantile probability between 0 and 1 to set the min and max values of the grid points in the previous option. If `qgrid` is smaller than 0.5, the min value is the `qgrid`-quantile and the max value is the (1-`qgrid`)-quantile. The default is 0, so that in that case the grid ranges from the min to the max value in the sample. The user need not specify this if `CvM` is used for `teststat`.

3.4 Saved Results

The command `dgmtest` generates the following results in `e()`:

Scalars			
`e(N)`	number of observations	`e(btpv)`	bootstrap p-value
`e(dimXW1)`	dimension of (X, W_1)	`e(btcv1)`	1% bootstrap critical value
`e(dimW2)`	dimension of W_2	`e(btcv5)`	5% bootstrap critical value
`e(dimZ)`	dimension of Z	`e(btcv10)`	10% bootstrap critical value
`e(stat)`	scalar value of the test statistic	`e(ngrid)`	number of grid points
`e(bootnum)`	number of bootstrap samples	`e(qgrid)`	quantile probability for min or max values of grid points
`e(bw)`	bandwidth h		
Macros			
`e(cmd)`	`dgmtest`	`e(teststat)`	type of test statistic
`e(title)`	nonparametric significance test	`e(bootdist)`	distribution of bootstrap multiplier variable
`e(kernel)`	type of kernel function		

3.5 A Simple Example

Consider again the simple simulated example from Section 2. First, perform the nonparametric test for measurement error on the correctly measured explanatory variable, using the default settings of the `dgmtest` command:

```
. dgmtest y xs z

----------------------------------------------------
 Delgado and Manteiga test
----------------------------------------------------

H0: E[Y | X,W1,Z] = E[Y | X,W1]

----- parameter settings -----
```

```
Test statistic: CvM (default)
Kernel: epanechnikov (default)
bw = n^(1/3q) (default)
bootstrap multiplier distribution: mammen (default)

 number of observations: 200
 bandwidth: .17099759
 dimension of (X,W1): 1
 dimension of W2: 0
 dimension of Z: 1
 number of bootstrap samples: 500

----- test results -----

 CvM = .0023243
 bootstrap critical value at 1%: .01183001
 bootstrap critical value at 5%: .00939969
 bootstrap critical value at 10%: .00781435
 p(CvM < CvM*) = .812
```

The p-value of the Cramér-von Mises version of the test is 0.812 which means we fail to reject the null of no measurement error at all reasonable confidence levels. Now, we perform the test on mismeasured explanatory variable, again using the default settings of the command:

```
. dgmtest y x z

-----------------------------------------------------
 Delgado and Manteiga test
-----------------------------------------------------

H0: E[Y | X,W1,Z] = E[Y | X,W1]

----- parameter settings -----

Test statistic: CvM (default)
Kernel: epanechnikov (default)
bw = n^(1/3q) (default)
bootstrap multiplier distribution: mammen (default)

 number of observations: 200
 bandwidth: .17099759
 dimension of (X,W1): 1
 dimension of W2: 0
 dimension of Z: 1
 number of bootstrap samples: 500

----- test results -----

 CvM = .01688708
 bootstrap critical value at 1%: .01369306
 bootstrap critical value at 5%: .01035709
 bootstrap critical value at 10%: .00813346
 p(CvM < CvM*) = .002
```

As expected the nonparametric test detects the measurement error and strongly rejects the null of no measurement error (p-value is 0.002) at all reasonable confidence levels.

4 Monte Carlo Simulation

In this section, we present a small simulation study investigating the finite sample performance of the measurement error test.

We consider the following outcome equation

$$Y = X^{*2} + \frac{1}{2}X^{*} + N\left(0, \sigma_{\varepsilon}^{2}\right) \tag{10}$$

with different models for the measurement system:

$$\begin{array}{lll}
\text{Model I}: & X = X^{*} + D \cdot N\left(0, \sigma_{ME}^{2}\right), & Z = X^{*} + N\left(0, 0.3^{2}\right); \\
\text{Model II}: & X = X^{*} + D \cdot N\left(0, \sigma_{ME}^{2}\right) e^{-\left|X^{*}-0.5\right|}, & Z = X^{*} + N\left(0, 0.3^{2}\right); \\
\text{Model III}: & X = X^{*} + D \cdot N\left(0, \sigma_{ME}^{2}\right) e^{-\left|X^{*}-0.5\right|}, & Z = X^{*} + N\left(0, 0.3^{2}\right) e^{-\left|X^{*}-0.5\right|}; \\
\text{Model IV}: & X = X^{*} + D \cdot N\left(0, \sigma_{ME}^{2}\right), & Z = -\left(X^{*}-1\right)^{2} + N\left(0, 0.2^{2}\right).
\end{array}$$

The value for σ_{ε} is 0.5 for the models I, II, and III, and 0.2 for the model IV. In all four models, $X^{*} \sim U\left[0,1\right]$ and the random variable D is Bernoulli$(1-\lambda)$, where $1-\lambda$ is the probability of measurement error in X occurring. $1-\lambda = 0$ means there is no measurement error in X, which represents the null hypothesis. To generate alternatives, we increase $1-\lambda$ on a grid up to one. We also vary the standard deviation of the measurement error in X, σ_{ME}, in $\{0.2, 0.5, 1\}$. Therefore, alternatives get closer to the null as we decrease $1-\lambda$ and/or σ_{ME}. We also vary the sample size $n \in \{200, 500\}$, but all models are simulated on 1,000 Monte Carlo samples. Following Delgado and Gonzalez Manteiga (2001), we use the bandwidth rule-of-thumb value $n^{-1/3}$. Simulation results for different choices of bandwidths, which are not presented here, are very similar.

The Cramér-von Mises statistics are generated by

```
. dgmtest Y X Z, kernel(epan2) bootnum(100)
```

The Kolmogorov-Smirnov test statistics with 10 grid points are generated by

```
. dgmtest Y X Z, teststat(KS) kernel(epan2) bootnum(100) ngrid(10) qgrid(0.05)
```

Table 1 shows the rejection frequencies of the test. Overall the test controls size well and possesses power against all alternatives. These findings are consistent with the Monte Carlo simulation results in Wilhelm (2018).

Table 1: Rejection frequencies from the simulation experiment.

		$n = 200$					$n = 500$				
σ_{ME}	$1 - \lambda$	0	0.25	0.5	0.75	1	0	0.25	0.5	0.75	1
Model I											
0.2			0.164	0.377	0.600	0.789		0.270	0.677	0.922	0.983
0.5	CvM	0.049	0.394	0.853	0.981	0.995	0.049	0.777	0.996	1.000	1.000
1.0			0.319	0.847	0.994	0.999		0.683	0.997	1.000	1.000
0.2			0.143	0.316	0.538	0.711		0.242	0.611	0.875	0.973
0.5	KS	0.051	0.377	0.836	0.973	0.996	0.054	0.697	0.995	1.000	1.000
1.0			0.314	0.813	0.988	0.998		0.652	0.996	1.000	1.000
Model II											
0.2			0.123	0.240	0.374	0.537		0.190	0.436	0.717	0.886
0.5	CvM	0.049	0.322	0.767	0.956	0.992	0.049	0.630	0.986	1.000	1.000
1.0			0.370	0.876	0.996	0.998		0.755	0.999	1.000	1.000
0.2			0.111	0.211	0.322	0.490		0.166	0.380	0.642	0.856
0.5	KS	0.051	0.287	0.713	0.934	0.986	0.054	0.567	0.974	1.000	1.000
1.0			0.357	0.845	0.990	0.998		0.698	0.995	1.000	1.000
Model III											
0.2			0.149	0.312	0.512	0.706		0.235	0.591	0.852	0.963
0.5	CvM	0.051	0.399	0.876	0.986	1.000	0.055	0.782	0.997	1.000	1.000
1.0			0.472	0.952	1.000	1.000		0.875	1.000	1.000	1.000
0.2			0.127	0.287	0.429	0.632		0.201	0.523	0.819	0.950
0.5	KS	0.050	0.376	0.848	0.983	0.998	0.053	0.736	0.996	1.000	1.000
1.0			0.446	0.952	0.998	1.000		0.844	1.000	1.000	1.000
Model IV											
0.2			0.586	0.941	0.997	1.000		0.938	1.000	1.000	1.000
0.5	CvM	0.076	0.912	1.000	1.000	1.000	0.061	1.000	1.000	1.000	1.000
1.0			0.828	1.000	1.000	1.000		0.999	1.000	1.000	1.000
0.2			0.464	0.889	0.990	0.998		0.847	0.999	1.000	1.000
0.5	KS	0.062	0.898	1.000	1.000	1.000	0.052	0.998	1.000	1.000	1.000
1.0			0.802	1.000	1.000	1.000		0.999	1.000	1.000	1.000

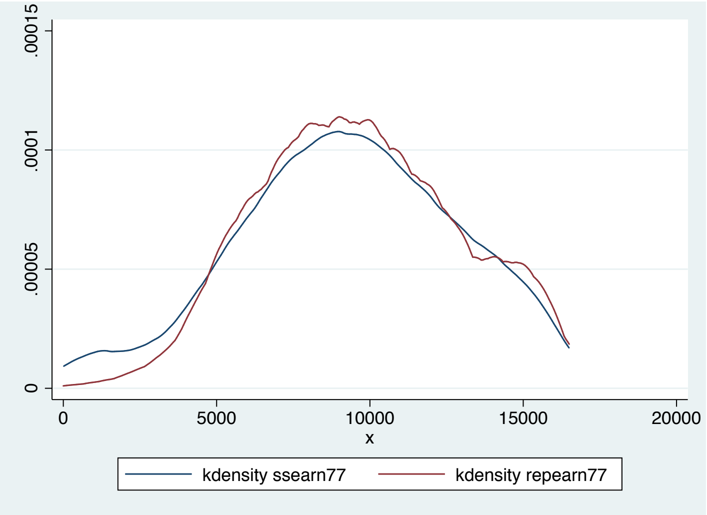

Figure 1: Nonparametric density estimates of administrative earnings ("ssearn77") and survey earnings ("repearn77") in 1977, using cross-validated bandwidths.

5 Example: Testing for the Presence of Measurement Error in Administrative Earnings Data

In this section, we test for measurement error in the U.S. Social Security Administration's measure of earnings. While measurement error in survey responses is a widespread concern that has occupied a large literature (Bound et al. (2001)), only recently empirical researchers have emphasized concerns about the reliability of administrative data (e.g. Fitzenberger et al. (2006), Kapteyn and Ypma (2007), Abowd and Stinson (2007), Groen (2011)).

The data come from the 1978 Current Population Survey-Social Security Earnings Records Exact Match File. The sample selection is similar to Wilhelm (2018) except that we only consider white singles of age between 25 and 60 who work full time the full year. The sample size is 2,683 individuals. The dataset contains a survey measure of earnings in 1977 (`repearn77`) from the CPS and two administrative measures of earnings in 1977 and in 1976 (`ssearn77` and `ssearn76`), the earnings records of the social security administration. We denote by Y the survey measure and by X and Z the administrative measures in 1977 and 1976, respectively. A test for the presence of measurement error in X as in H_0 is then a test of the presence of measurement error in administrative earnings in 1977.

Figure 1 shows nonparametric density estimates of survey and administrative earnings. Figure 2 plots the nonparametric density estimate of the difference between administrative and survey earnings. There is substantial probability mass within USD $\pm 1,000$ which are large deviations relative to the maximum earnings in the sample

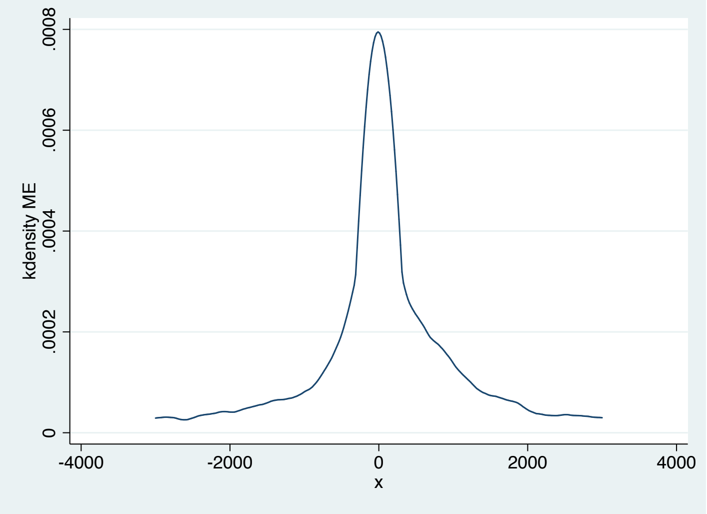

Figure 2: Nonparametric density estimate of the difference in administrative and survey earnings in 1977, using a cross-validated bandwidth.

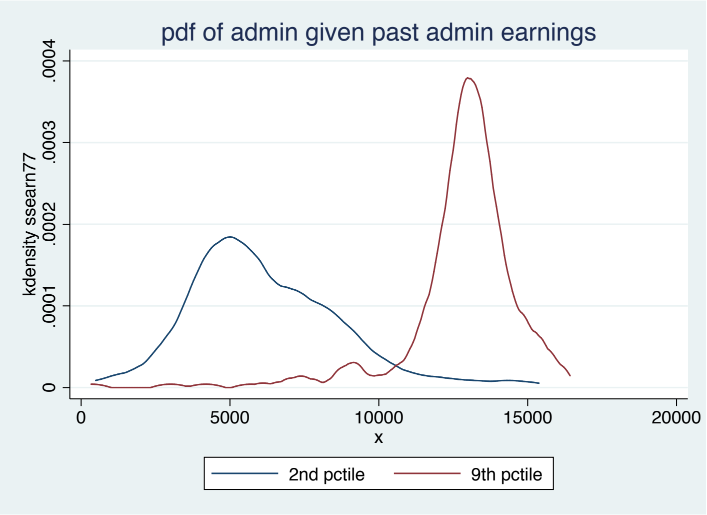

Figure 3: Nonparametric estimate of the conditional density of administrative earnings in 1977 given lagged administrative earnings being in the 10th or 90th percentile. Bandwidths are chosen by cross-validation.

(USD $16,500$).

The exclusion restriction (4) is likely to hold in this context because the measurement errors in survey and administrative earnings come from very different sources (see the

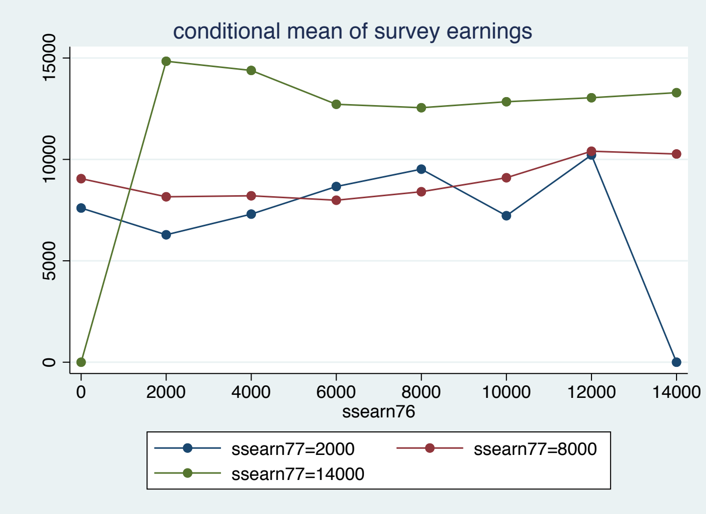

Figure 4: Nonparametric estimate of $E[Y|X,Z]$, where Y is survey earnings in 1977, X and Z are administrative earnings in 1977 and 1976, respectively. Bandwidths are chosen by cross-validation.

more detailed discussion in Wilhelm (2018)). To assess the relevance of the second measurement Z, which here is lagged administrative earnings, we plot the density of administrative earnings in 1977 given those in 1976. Figure 3 shows this density for those individuals with lagged earnings in the 10th and 90th percentile of the 1976 earnings distribution. The graph shows that the second measurement Z, lagged administrative earnings, shifts the earnings distribution in the next period to the right as we go from the 10th to the 90th percentile. In particular, the two densities seem to cross only once, which is consistent with the relevance condition that is needed for the equivalence of H_0 and the observable restriction (5).

Figure 4 shows nonparametric estimates of the conditional mean $E(Y|X=x, Z=z)$ as a function of z for three values of x. If there was no measurement error in X, then (5) implies that this conditional mean should not vary with z. The graphs suggests that there is some variation in that dimension, particularly for small and large values of earnings, but the graph does not contain any information about whether this variation is statistically significant. We therefore now discuss the results of the formal test of H_0.

The test is performed using the new command [R] **dgmtest** with its default settings except we increase the number of bootstrap samples to 5,000:

```
. dgmtest repearn77 ssearn77 ssearn76, bootnum(5000)

-----------------------------------------------------
 Delgado and Manteiga test
-----------------------------------------------------
```

Table 2: Test results.

	p-value	test stat.	cval 1%	cval 5%	cval 10%	h	sample size
full sample	0.026	0.512	0.631	0.418	0.336	0.072	2,682
males	0.141	0.276	0.597	0.404	0.321	0.102	944
females	0.100	0.318	0.583	0.407	0.319	0.083	1,738
< highschool	0.080	0.290	0.759	0.759	0.111	0.169	206
highschool	0.210	0.143	0.616	0.459	0.210	0.091	1,329
> highschool	0.072	0.818	1.504	0.922	0.721	0.096	1,147

```
H0: E[Y | X,W1,Z] = E[Y | X,W1]

----- parameter settings -----

Test statistic: CvM (default)
Kernel: epanechnikov (default)
bw = n^(1/3q) (default)
bootstrap multiplier distribution: mammen (default)

 number of observations: 2682
 bandwidth: .07197479
 dimension of (X,W1): 1
 dimension of W2: 0
 dimension of Z: 1
 number of bootstrap samples: 5000

----- test results -----

 CvM = .51238949
 bootstrap critical value at 1%: .63053938
 bootstrap critical value at 5%: .41803533
 bootstrap critical value at 10%: .33279162
 p(CvM < CvM*) = .0262
```

The test produces a p-value of 0.0262 so we reject the null of no measurement error in administrative earnings at high confidence levels. Table 2 shows the test results for the full sample as well as for subsamples with the same gender and education. The p-values for the low and high education groups are about 8% and 7%, which is some evidence for the presence of measurement error, but weaker than in the full sample. For individuals in the middle education group there is no evidence of measurement error. Similarly, we cannot reject the null on the subsamples of males and females. Of course, the sample sizes on the subsamples are significantly smaller than on the full sample, so it may be harder to reject the null for that reason.

6 Concluding Remarks

This paper describes how to test for the presence of measurement error in covariates. While in linear regression models with classical measurement error, testing the null of no measurement error can be carried out using simple linear regression techniques, this paper introduces the new command [R] **dgmtest** which implements a nonparametric test that doesn't rely on linearity nor on the measurement error (if there is any) to be classical.

The command is an implementation of the Delgado and Gonzalez Manteiga (2001) test of conditional mean independence, a hypothesis that might be of interest in applications other than testing for the presence of measurement error.

7 References

Abowd, J. M., and M. H. Stinson. 2007. Estimating Measurement Error in SIPP Annual Job Earnings: A Comparison of Census Survey and SSA Administrative Data. Technical report.

Altonji, J. G. 1986. Intertemporal Substitution in Labor Supply: Evidence from Micro Data. *The Journal of Political Economy* 94(3): S176–S215.

Arellano, M., R. Blundell, and S. Bonhomme. 2017. Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework. *Econometrica* 85(3): 693–734.

Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio-Codina. 2015. Estimating the Production Function for Human Capital: Results from a Randomized Control Trial in Colombia. Working Paper 20965, NBER.

Attanasio, O., C. Meghir, and E. Nix. 2017. Human Capital Development and Parental Investment in India. Technical report.

Blundell, R., J. Horowitz, and M. Parey. 2016. Nonparametric Estimation of a Nonseparable Demand Function under the Slutsky Inequality Restriction. *Review of Economics and Statistics* forthcoming.

Blundell, R., J. L. Horowitz, and M. Parey. 2012. Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quantitative Economics* 3(1): 29–51.

Bound, J., C. Brown, and N. Mathiowetz. 2001. Measurement Error in Survey Data. In *Handbook of Econometrics*, ed. J. J. Heckman and E. E. Leamer, vol. V, 3705–3843. Elsevier Science B.V.

Card, D. 1996. The Effect of Unions on the Structure of Wages: A Longitudinal Analysis. *Econometrica* 64(4): 957–979.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall, New York.

Chen, X., H. Hong, and D. Nekipelov. 2011. Nonlinear Models of Measurement Errors. *Journal of Economic Literature* 49(4): 901–937.

Chesher, A. 1990. On Assessing the Effect and Detecting the Presence of Measurement Error. Technical report, University of Bristol.

———. 1991. The Effect of Measurement Error. *Biometrika* 78(3): 451–462.

Chesher, A., M. Dumangane, and R. J. Smith. 2002. Duration response measurement error. *Journal of Econometrics* 111(2): 169 – 194.

Chetty, R. 2012. Bounds on Elasticities With Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply. *Econometrica* 80(3): 969–1018.

Chetverikov, D., and D. Wilhelm. 2017. Nonparametric Instrumental Variable Estimation Under Monotonicity. *Econometrica* 85(4): 1303–1320.

Cunha, F., J. J. Heckman, and S. M. Schennach. 2010. Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica* 78(3): 883–931.

Delgado, M. A., and W. Gonzalez Manteiga. 2001. Significance Testing in Nonparametric Regression Based on the Bootstrap. *The Annals of Statistics* 29(5): 1469–1507.

Durbin, J. 1954. Errors in Variables. *Review of the International Statistical Institute* 22(1/3): 23–32.

Fan, Y., and Q. Li. 1996. Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms. *Econometrica* 64(4): 865–890.

Feng, S., and Y. Hu. 2013. Misclassification Errors and the Underestimation of the US Unemployment Rate. *American Economic Review* 103(2): 1054–70.

Fitzenberger, B., A. Osikominu, and R. Völter. 2006. Imputation Rules to Improve the Education Variable in the IAB Employment Subsample. *Schmollers Jahrbuch (Zeitschrift für Wirtschafts- und Sozialwissenschaften / Journal of the Applied Social Sciences)* 126(3): 405–436.

Gozalo, P. L. 1993. A Consistent Model Specification Test for Nonparametric Estimation of Regression Function Models. *Econometric Theory* 9(3): 451–477.

Groen, J. A. 2011. Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures. Technical report.

Hahn, J., and J. Hausman. 2002. A New Specification Test for the Validity of Instrumental Variables. *Econometrica* 70(1): 163–189.

Härdle, W., and E. Mammen. 1993. Comparing Nonparametric Versus Parametric Regression Fits. *The Annals of Statistics* 21(4): 1926–1947.

Hausman, J. A. 1978. Specification Tests in Econometrics. *Econometrica* 46(6): 1251–1271.

Heckman, J., R. Pinto, and P. Savelyev. 2013. Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review* 103(6): 2052–2086.

Hoderlein, S., H. Holzmann, M. Kasy, and A. Meister. 2016. Erratum Instrumental Variables with Unrestricted Heterogeneity and Continuous Treatment. *The Review of Economic Studies* forthcoming.

Hu, Y. 2008. Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution. *Journal of Econometrics* 144: 27–61.

———. 2017. The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics. *Journal of Econometrics* 200(2): 154 – 168.

Hu, Y., D. McAdams, and M. Shum. 2013. Identification of first-price auctions with non-separable unobserved heterogeneity. *Journal of Econometrics* 174(2): 186–193.

Huang, M., Y. Sun, and H. White. 2016. A Flexible Nonparametric Test for Conditional Independence. *Econometric Theory* 32(6): 1434–1482.

Kane, T. J., and C. E. Rouse. 1995. Labor-Market Returns to Two- and Four-Year College. *The American Economic Review* 85(3): 600–614.

Kane, T. J., C. E. Rouse, and D. Staiger. 1999. Estimating Returns to Schooling When Schooling is Misreported. Working Paper 7235, NBER.

Kapteyn, A., and J. Y. Ypma. 2007. Measurement Error and Misclassification: A Comparison of Survey and Administrative Data. *Journal of Labor Economics* 25(3): 513–551.

Kasy, M. 2014. Instrumental Variables with Unrestricted Heterogeneity and Continuous Treatment. *The Review of Economic Studies* 81(4): 1614.

Mahajan, A. 2006. Identification and Estimation of Regression Models with Misclassification. *Econometrica* 74(3): 631–665.

Matzkin, R. L. 1994. Restrictions of Economic Theory in Nonparametric Methods. In *Handbook of Econometrics*, ed. R. F. Engle and D. L. McFadden, vol. IV, 2523–2558. Elsevier Science B.V.

Olley, G. S., and A. Pakes. 1996. The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica* 64(6): 1263–1297.

Robinson, P. M. 1988. Root-N-Consistent Semiparametric Regression. *Econometrica* 56(4): 931–954.

Schennach, S. M. 2013. Measurement Error in Nonlinear Models — A Review. In *Advances in Economics and Econometrics*, vol. 3. Cambridge University Press.

———. 2016. Recent Advances in the Measurement Error Literature. *Annual Review of Economics* 8: 341–377.

Wilhelm, D. 2015. Identification and Estimation of Nonparametric Panel Data Regressions with Measurement Error. Working Paper CWP34/15, cemmap.

———. 2018. Testing for the Presence of Measurement Error. Working Paper CWP45/18, CeMMAP.

Wu, D.-M. 1973. Alternative Tests of Independence between Stochastic Regressors and Disturbances. *Econometrica* 41(4): 733–750.

About the authors

Young Jun Lee is a post-doctoral researcher in the Department of Economics at University of Copenhagen.

Daniel Wilhelm is a Lecturer in Economics at University College London, and a staff member of the ESRC Centre for Microdata Methods and Practice (CeMMAP), a research fellow at the Institute for Fiscal Studies and the Centre for Research and Analysis of Migration (CReAM).