

Overidentification Tests and Causality: A Second Response to Roodman and Morduch¹

Mark M. Pitt

Abstract: After Pitt (2011) pointed out the flaws in the RM replication effort, Roodman subsequently notes “*that when we fix our regressions, they continue to fail tests of the assumptions needed to infer causality. So improving the match to the original greatly strengthens our conclusion that this study does not convincingly demonstrate an impact of microcredit on poverty.*” This claim is based on RM’s tests of overidentifying restrictions, which the current response demonstrates are fundamentally flawed. New results presented below provide strong support to the hypothesis that microfinance causally improves the lives of participants.

Recently, David Roodman and Jonathan Morduch [2009] (henceforth RM) have written a paper that claims that the “headline result”, as they call it, from the Pitt and Khandker (1998) (henceforth PK) paper published in a 1998 issue of *The Journal of Political Economy* (“The Impact of Group-Based Credit on Poor Households in Bangladesh: Does the Gender of Participants Matter?”) cannot be replicated with the data. The headline result they refer to is that “annual household consumption expenditure increases 18 taka for every 100 additional taka borrowed by women...compared with 11 taka for men.” (RM p. 980). Very recently, I responded (Pitt 2011) by pointing out econometric/data mistakes made by RM and conclusively demonstrated that, after correcting these errors, the PK results hold even when estimated with Roodman’s *cmp* Stata module. David Roodman concurs, writing in his blog of March 31, 2011 that “Pitt’s response has exposed an important mistake in our work. With his fixes, we now match their key results extremely well”.

This second response to RM covers the issue of overidentification tests and causation. RM present six Sargan tests and two Hansen J tests of overidentification based upon what they describe as the “closest 2SLS [two-stage least squares] analog” to the PK model. RM suggest that the test statistics they calculate lead them to conclude that the excluded instruments in PK are invalid and that no causal inference can be made. As these test statistics suffered from the same error with data that affected their replication of the maximum likelihood estimates of PK, Roodman has re-computed them and posted these new estimates in his blog concluding that they “Strengthen(s) our main conclusion, which is about lack of proof of *causality*,” and that “when we fix our regressions, they continue to fail tests of the assumptions needed to infer causality. So

¹ Thanks to David Roodman for his useful comments on a preliminary version of this paper.

improving the match to the original greatly strengthens our conclusion that this study does not convincingly demonstrate an *impact* of microcredit on poverty.”

The validity of the PK instruments is an important substantive issue that was not well addressed in the PK paper. It was first raised by Morduch (1998), resulting in a partial response by Pitt (1999). The RM paper’s most useful service has been to raise this issue anew and bring formal tests of overidentification into the discussion. Unfortunately, RM make serious errors in their calculation of the overidentification test statistics and consequently in their claims about causation.

This response will demonstrate that all six of the Sargan tests that they present have been incorrectly calculated, and are not valid tests of overidentification. In addition, it will be shown that the errors made by Roodman in calculating these tests will, in general, lead to over-rejection of the null hypothesis that the overidentifying restrictions hold. Furthermore, the two Hansen J tests that Roodman presents are also suspect. Below, new evidence of the robustness and validity of the overidentifying restrictions in the PK model are presented. This evidence strongly suggests that the overidentifying restrictions of PK are valid.

Why are the Sargan tests presented by RM and in the Roodman blog wrong?

If an equation is overidentified, we may test whether the excluded instruments are appropriately independent of the error process. A test of overidentifying restrictions regresses the residuals from an IV or 2SLS regression on all instruments. Under the assumption of iid errors, this is known as a Sargan test, and is routinely produced by 2SLS estimation procedures in software packages such as Stata.

The Sargan tests of overidentifying restrictions as originally presented in RM and re-calculated in Roodman’s blog posting are wrong for two important reasons:

(1) The Sargan test is invalid when estimating sample-weighted regressions, as in PK and RM. In general, if sampling is based on exogenous sampling then weights are not needed. However, if sampling is based on endogenous variables (or both exogenous and endogenous variables), then sampling weights are required to achieve consistency. As RM explicitly recognize, the PK sample is a choice-based sampling design and that sampling weights are required in order to obtain consistent estimates of the parameters of the PK model no matter how it is estimated. Even if the underlying distribution of the errors in a population is homoskedastic, a heteroskedasticity-robust White (also known as Huber-White sandwich) covariance matrix is required to get consistent standard errors (Manski and Lerman, 1977). The sample weights make the sampled data appear as if the errors are not iid, and consequently, Sargan tests, which require

iid errors, are false. Fortunately, applying sample weights as PK (and RM) do result in consistent parameter estimates (Manski and Lerman). It is just the test statistics that are affected.

In Stata, for example, whenever the *pweight* (probability weight) option is specified for the *regress*, *tobit*, *ivregress2* (two-stage least-squares), *ivreg2*, *probit*, *logit*, and every other estimation command that allows for *pweights*, Stata automatically and only reports appropriate robust (White) standard errors. In the case of Stata's 2SLS command *ivregress*, it will report the Sargan test when the *pweights* (or robust) option is not specified, but does not report the Sargan test when *pweights* are specified because it knows that the Sargan test is wrong with *pweights*. Likewise, the popular *ivreg2* program of Baum and Shaeffer will not report a Sargan test for models with *pweights*. Only the user-contributed *xtabond2* command written by David Roodman ignores the slings and arrows of statistical theory to report a Sargan test with *pweights*. Any test statistic based upon these non-robust standard errors, including Sargan tests of overidentification, is wrong and has no statistical validity. This is not a question of which overidentification test the researcher prefers, it is simply a question of whether a test statistic is wrong or right from the perspective of statistical theory. The effect of not using robust statistics to compute tests of overidentification is to over-reject the null hypothesis that the instruments are valid. For example, 500 replications of estimating a PK-type model with synthetic data (with 5000 observations per replication) and *pweights* results in Sargan tests rejecting the validity of the instruments 80% of the time at the .05 level even though the instruments are valid by construction.

(2)The two-stage least squares (2SLS) “analog” of the PK model that Roodman estimates induces heteroskedasticity in its design, invalidating the Sargan test. Even with a random sample that does not require weighting, Roodman's method of estimating the PK model with 2SLS necessarily introduces heteroskedasticity thus invalidating the iid assumption required for the Sargan test. In the PK model, households that are precluded from having any choice to participate in a micro-credit program either because they are “nontarget” due to their wealth, or because they live in a village without a credit program, have deterministically zero levels of microcredit. These non-choice households are the key to parameter identification in PK. The maximum likelihood method used by PK, as well as the maximum likelihood estimates of RM estimated with Roodman's *cmp* Stata command, treat deterministically zero microcredit correctly. However, the 2SLS setup of Roodman treats them essentially as stochastic. These deterministically zero dependent variables are in the first-stage equation along with the stochastic choice credit variable. The problem is that the error variance of the deterministic observations is zero – that is what deterministic means – while the error variance of the stochastic observation is some positive number. Even if there is a common (homoskedastic) variance for the observations with microcredit choice, this setup introduces heteroskedasticity by lumping in the observations with zero variance. Consequently, the 2SLS errors are not iid and the Sargan test is wrong.

Once again in this case, not using robust statistics to compute tests of overidentification leads to over-rejecting of the null hypothesis that the instruments are valid. (The mass point at zero credit for observations with choice to join a program will also generate heteroskedasticity, as linear 2SLS is essentially a linear probability model applied to a limited dependent variable.) (Roodman has now acknowledged in an update to his blog that the Sargan tests reported in RM and in his blog are “not quite right” as a consequence of the sampling weighting in the regressions).

What are the issues with Roodman’s calculation of Hansen J tests?

If the errors in a 2SLS regression are not iid, then Hansen’s J statistic is the appropriate test of the overidentifying restrictions. RM only calculate the Hansen J test for the full estimation dataset because they state that pooling multi-round data for the same set of households will likely result in errors that are not iid, failing to discern that sample weighting and the construction of their 2SLS method also imply that the errors are not iid. Nonetheless, there are still issues with RM’s method of calculating the Hansen J tests, although none as serious as with his Sargan tests.

(1)The problem of zero variances for the non-choice sample may not fully be alleviated even with a Hansen test based upon a robust covariance matrix. The justification of the Hansen J test is based upon statistical theorems that may not hold when the error of the residuals for a subset of observations has a zero variance. Consider that all the non-choice observations in the second-stage equation are essentially exogenous observations, as they are simply regressions of household consumption on only exogenous variables (credit is exogenously and deterministically zero for them). Endogeneity in the PK model is a function of who you are and varies by observation – I have never seen this as a property in any other IV model that I know. It is far from apparent to me that you can just plunk the PK model into the classical 2SLS format and just assume that all the statistical theory derived for 2SLS models still works when not all of the observations have strictly positive variances and are endogenous. Maybe all of the statistical theory still holds, but anyone willing to undertake statistical inference of the PK model using classical statistical principles needs to at least note the issue and assure the reader that it does not matter. To be clear, PK have themselves suggested that 2SLS can recover the parameters of the underlying data generating process, even if it is of the PK form. However, Roodman and Morduch need to formally prove that their method of computing the Hansen J test has the statistical properties that they implicitly claim for it.

(2)The instrument set used in the 2SLS setup of RM is not the same as in PK’s paper or in RM when estimated with Roodman’s *cmp*, so RM’s Hansen test is not formally a test of the PK overidentifying restrictions. In particular, in PK and RM, the identifying restrictions for endogenous female credit consists only of female choice interacted with the exogenous variables and village dummies. Likewise, in PK and RM, the identifying restriction for endogenous male

credit consists only of male choice interacted with the exogenous variables and village dummies. However, in the RM 2SLS setup, the male choice interactions are included in the instrument set for female credit, and the female choice interactions are included in the instrument set for male credit. Perhaps this is a reasonable model (it gets noted in PK), but it is not what PK do, and it is not even what RM do, so this expanded set of instruments cannot be part of a test of PK's overidentifying restrictions. After reading the preliminary version of this paper, Roodman has responded privately to me that this seems to be criticizing the reduced-form equations for not being structural. Nonetheless, one cannot know how much any such difference in specification may matter. I suspect it does not matter much at all.

(3) If there is heteroskedasticity, for example heteroskedasticity induced by sample weights or the inclusion of deterministic observations, then 2SLS is not asymptotically efficient, but GMM is. The use of GMM does come with a price. Efficient GMM estimators can have poor small sample properties. In particular, Wald tests on the parameters tend to over-reject the null. If one iterates the GMM, constantly updating the weight matrix, the estimates have the same asymptotic distribution, however, there is evidence that more iterations improves finite-sample performance. The asymptotic distribution does not depend on how many iterations are done. The empirical question is: does iterative GMM matter in testing overidentifying restrictions in PK? I took Roodman's Stata do file posted on his blog, as well as his estimation dataset, and simply iterated GMM 100 iterations. The Hansen J test that he reports in his updated blog (and which is confirmed in my alteration of his code) has a p-value of 0.0233, which indicates a problem with the overidentifying restrictions at the .05 level but not the .01 level of significance. The iterated GMM has a p-value of 0.211, almost ten times as large, and indicates that the overidentifying restriction are valid at any of the critical values typically used.

A Note on RM versus Roodman's Newly Blogged Results

All of the above discussion clearly applies to the RM paper. As far as I can tell, it also applies to Roodman's re-computed estimates, posted in his blog, that correct the mistakes pointed out in Pitt(2011). Roodman notes in his blog that readers should "take these thoughts as preliminary." None of the replication errors identified in Pitt (2011) or in blog postings subsequent to that have provided any information about faults that I claim affect Roodman's calculation of overidentification tests until this current post of mine. For example, in his March 31 blog posting presenting the re-estimated overidentification tests, Roodman says: "In addition to homoskedasticity, the Pitt-Khandker LIML regressions assume no cross-household error correlation; thus the only deviation from sphericity that they allow is serial correlation. This makes errors i.i.d. within survey rounds and makes the Sargan test valid for 2SLS regressions restricted to individual rounds. The Hansen test is required for regression that pool all three rounds. " This is the very line of argument that is disputed in this paper. [Note that this morning Roodman added the following update to his blog post: : [Update: this is not quite right. The

regressions are weighted for sampling---"pweights" in Stata parlance---which effectively introduces heteroskedasticity and invalidates the Sargan test even in cross-sections. But Hansen, now shown for all columns, remains valid and corroborates Sargan.]”

Moreover, the PK paper did not do any 2SLS estimation or calculate or perform any tests of overidentification, so there has never been any code of ours to share or post. The calculation of tests of overidentifying restriction was initiated by RM, not PK.

Some New Tests of Instrument Validity

The issue of the validity of the instruments was a theme in Morduch’s 1998 paper. In particular, Morduch is concerned with the interactions of land with all of the other exogenous regressors because there may be “systematic differences between the landless and landed in, say, the impact of age on income.” This is certainly a valid concern for if it were true the instruments would not be valid and the results reported in PK might not be considered causal. In my 1999 response to Morduch, I addressed that concern by re-estimating the model with these interactions. Table 6 of Pitt (1999) presents estimates of the effects of program credit, by program and gender, on the log of household per capita income, allowing for interactions between land ownership and all of the exogenous regressors and interactions between land ownership and all of the thana fixed effects. There are three villages in each thana in the sample design, and all three villages in each thana have the same credit program (one of BRDB, BRDB, or Grameen Bank). All 18 exogenous regressors and the 29 thana dummy variables are interacted with land and are included in the consumption equation. This is actually the most “crucial” interaction that could be included since it is the land related eligibility rule that is the source of identification in the model. Consequently, these interactions are the ones most likely to destroy identification if in fact it is the linearity of the consumption function that is driving the identification of credit effects. Based on these estimates that allowed for interactions with landholding, I concluded in Pitt(1999) that the “bottom line” qualitative results of PK still hold –there are positive and statistically significant effects of female credit program participation on household consumption, and much smaller and generally statistically insignificant effects of male credit program participation on household consumption, as in PK.

However, I did not present tests of overidentification in that paper, so it seems worth revisiting. Table 1 presents the estimates of the PK model as reported in the PK paper in column (1), and as re-estimated using *cmp* in column (2) (see Pitt (2011) for the reasons why these two estimates may differ). The estimates first presented in Pitt (1999) are presented again in column (3) of Table 1, only reestimated with Roodman’s *cmp* command. The column (3) estimates add the full set of interactions of landholding with all the exogenous variables from the first-stage equation plus interactions of landholding with thana-level dummy variables in the second-stage

(household consumption) equation. As in Pitt (1999), the “headline results” still hold. Women’s credit has statistically significant and positive effects on household consumption, although these effects are slightly attenuated with the addition of 47 additional control variables.

The key result to report here is that a regression of the *cmp* residuals from the household consumption equation on the set of identifying restrictions (an F-test) has a p-value of 0.794, that is, the identifying restrictions do not explain the unexplained portion of the household consumption equation, strongly suggesting their validity. In the 2SLS framework, the Hansen J test of overidentification can simply be computed as this F-test times (the number of identifying instruments) , and has a chi-squared distribution with degrees of freedom equal to the number of overidentifying restrictions. Although this method of computing a test of overidentifying restriction might not be exactly correct in this framework, the calculation yields a “statistic” with a p-value of 0.75. This p-value is not even close to conventional levels of significance, providing strong evidence of the validity of the instruments and the positive causal effects of microfinance in this model.

Summary

RM’s replication of PK concludes that (1) PK is not replicable, indeed that the signs of credit on household consumption are negative rather than positive, and (2) overidentification tests invalidate the instrument set used by PK so that no causal inference can be drawn. In my two response papers, I show that there are serious faults in both the RM replication and in their calculation of test of overidentification which coincidentally (1) reverse the sign of parameters, and (2) lead to the over-rejection of the overidentifying restrictions. The PK parameters results hold and as this paper has demonstrated, they can safely be considered causal.

Table 1

Estimates of the Impact of Credit on Per Capita Expenditure

Explanatory Variables	Log of Weekly Total Expenditure per Capita		
	PK published (1)	PK data using Roodman's <i>cmp</i> program (3)	Pk data with land interactions and <i>cmp</i> program
Amount borrowed by female from BRAC	.0394 (4.237)	.0443 (4.78)	.0372 (4.15)
Amount borrowed by male from BRAC	.0192 (1.593)	.0093 (0.52)	.0109 (0.81)
Amount borrowed by female from BRDB	.0402 (3.813)	.0458 (4.30)	.0388 (3.77)
Amount borrowed by male from BRDB	.0233 (1.936)	.0128 (0.70)	.0151 (1.14)
Amount borrowed by female from GB	.0432 (4.249)	.0420 (4.80)	.0358 (4.19)
Amount borrowed by male from GB	.0179 (1.431)	.0072 (0.45)	.0073 (0.62)
No. of observations	5218	5218	5218

Note: Figures in parentheses are asymptotic t-ratios clustered at the household level.

Sources:

Col (1): PK Table 2, page 981;

Col (2): My new estimates using the data I sent to RM (*PK estimation dataset*) and Roodman's program *cmp*;

Col (3): Same as column (2) but the second-stage (household consumption) equation also includes full interactions of landholding with all the exogenous variables from the first-stage equation and interactions of landholding with thana-level dummy variables. This column is a replication of column (2) in Table 6 of Pitt (1999).