On Mis-measured Binary Regressors: New Results and Some Comments on the Literature

Francis J. DiTraglia Camilo Garcia-Jimeno

University of Pennsylvania

October 21, 2015

Abstract

Abstract goes here.

Keywords: Instrumental variables, Measurement error, Endogeneity, Binary regressor, Partial Identification

JEL Codes: C10, C18, C25, C26

1 Introduction

This paper studies the use of a valid instrument to identify the causal effect of an endogeneous, binary treatment that is subject to non-differential measurement error in a non-parametric regression model with additively separable errors. Although a relevant case for applied work, this setting has received little attention in the literature. The only existing result appears in an important paper by **?**, who proves identification by relying on a discrete instrument that takes on at least two values. Here we begin by showing this result is incorrect. To do so, we first provide a convenient notational framework within which to situate the problem. Using this framework we show that the proof in Appendix A.2 of **?** leads to a contradiction. Throughout the paper, **?** maintains an assumption (Assumption 4) which he calls the "Dependency Condition." This assumption requires that the instrumental variable be relevant. When extending his result for an exogenous treatment to the more general case of an endogenous one in a model with additively separable errors, however, he must impose an additional condition on the model (Equation 11), which turns out to imply the lack of a first-stage, violating the Dependency Condition.

Since one cannot impose the condition in Equation 11 of **?**, we go on to study the prospects for identification in this model more broadly. We consider two possibilities. First, borrowing an idea from **?**, we explore whether expanding the support of the instrument, so that it takes on more than two values, yields identification. Allowing the instrument to take on additional values increases the number of available moment conditions. We show, however, that these additional moments cannot point identify the treatment effect. This holds true regardless of how many (finite) values the instrument takes on.

We then consider a new source of identifying information in the form of a conditional homoskedasticity assumption. In particular, we suppose that the conditional *variance* of the regression error term given the instrument is constant. While stronger than the usual mean independence assumption, this assumption holds automatically in a randomized controlled trial or a genuine natural experiment. To the best of our knowledge, this source of information has not been exploited in the extant literature on instrumental variables. We show that this assumption leads to a novel partial identification result that is easy to implement in practice and can be applied regardless of the number of values that the instrument takes on. Moreover, it can be used to obtain point identification in some special cases that nevertheless may be empirically relevant.

The remainder of this paper is organized as follows. In section 2 we discuss the literature in relation to our framework. Section 3 then lays out the econometric model, its assumptions, and our notational framework. Section 4 then presents two of our main results: A proof of **?**'s incorrect claim of identification under non-differential misclassification with a valid instrument and an endogenous treatment, and our proof of the generic un-identification of this model. Section 5 then presents our partial identification results relying on a conditional homoskedasticity assumption, a test for the presence of misclassification, and a simulation exercise illustrating its usefulness. Section 6 concludes.

2 Related Literature

Many treatments of interest in applied work are binary. To take a particularly prominent example, consider treatment status in a randomized controlled trial. Even if the randomization is pristine, which yields a valid binary instrument (the offer of treatment), subjects may select into treatment based on unobservables, and given the many real-world complications that arise in the field, measurement error may be an important concern. As is well known, instrumental variables (IV) based on a single valid instrument suffices to recover the treatment effect in a linear model with a single endogenous regressor subject to classical measurement errors. As is less well known, classical measurement error is in fact impossible when the regressor of interest is binary: because a true 1 can only be mis-measured as a 0 and a true 0 can only be mis-measured as a 1, the measurement error must be *negatively* correlated with the true treatment status (**??**).

Measurement error in a binary regressor is usually called *mis-classification*. The simplest form of mis-classification is so-called *non-differential* measurement error. In this case, conditional on true treatment status, and possibly a set of exogenous covariates, the measurement error is assumed to be uncorrelated with all other variables in the system. Even under this comparatively mild departure from classical measurement error, the IV estimator is inconsistent (**??**). Moreover, the probability limit of the IV estimator does not depend on whether the treatment is endogenous or not (**?**).

When the treatment is in fact *exogenous*, however, a valid instrument suffices to recover the treatment effect using a non-linear GMM estimator. **?** and **?** more-or-less simultaneously pointed this out in a setting in which *two* alternative measures of treatment are available, both subject to non-differential measurement error. In essence, one measure serves as an instrument for the other although the estimator is quite different from IV.[1] Subsequently, **?** correctly note that an instrumental variable can take the place of one of the measures of treatment in a linear model with an exogenous treatment, allowing one to implement a variant of the GMM estimator proposed by **?** and **?**. However, as we will show below, the assumptions required to obtain this result are are stronger than **?** appear to realize: the usual IV assumption that the instrument is mean independent of the regression error is insufficient for identification. **?** extends the results of **?** and **?** to a more general nonparametric regression setting using a binary instrument in place of one of the treatment measures. Although unaware of **?**, **?** makes the correct assumption over the instrument and treatment to guarantee identification of the conditional mean function. When the treatment is in fact exogenous, this coincides with the treatment effect. **?** provides a related identification result in the same model as **?** under different assumptions. In particular, the variable that plays the role of the "instrument" need not satisfy the exclusion restriction provided that it does not interact with the treatment and takes on at least three distinct values.

Much less is known about the case in which the treatment, in addition to

[1]Ignoring covariates, the observable moments in this case are the joint probability distribution of the two binary treatment measures and the conditional means of the outcome variable given the two measures. Although the system is highly non-linear, it can be manipulated to yield an explicit solution for the treatment effect provided that the true treatment is exogenous.

suffering from non-differential measurement error, is also endogenous. Only two papers consider this case. **?** briefly discuss the prospects for identification in this setting. Although they do not provide a formal proof they argue, in the context of their parametric linear model, that the treatment effect is unlikely to be identified unless one is willing to impose strong and somewhat unnatural conditions. The second paper that considers this case is **?**. He extends his main result to the case of an endogenous treatment, providing an explicit proof of identification under the usual IV assumption in a model with additively separable errors. Although their discussion does not apply to the non-parametric case, **?**'s intuition turns out to be right: **?**'s proof is incorrect, as we prove below using a convenient notational framework introduced in the following section.

3 The Model and Assumptions

Let T^* be a binary indicator of true treatment status, possibly endogenous, $\mathbf{x}$ be a vector of exogenous covariates, and y be an outcome of interest where

$$y = h(T^*, \mathbf{x}) + \varepsilon \tag{1}$$

and ε is mean zero. Since T^* is potentially endogenous, $\mathbb{E}[\varepsilon|T^*, \mathbf{x}]$ may not be zero. Now let z be a discrete instrumental variable with support set $\{z_k\}_{k=1}^K$ satisfying the usual instrumental variables assumption, namely $\mathbb{E}[\varepsilon|z, \mathbf{x}] = 0$. We assume throughout that z is a relevant instrument for T^*, in other words

$$\mathbb{P}(T^* = 1|z_j, \mathbf{x}) \neq \mathbb{P}(T^* = 1|z_k, \mathbf{x}), \quad \forall k \neq j. \tag{2}$$

Our goal is to estimate the average treatment effect (ATE) function

$$\tau(\mathbf{x}) = h(1, \mathbf{x}) - h(0, \mathbf{x}). \tag{3}$$

We maintain throughout that $\tau(\mathbf{x}) \neq 0$. If it were zero, this would imply that T^* is irrelevant for y which can be directly tested regardless of whether any mis-classification is present and regardless of whether T^* is endogenous.[2]

Now, suppose we observe not T^* but a noisy measure T polluted by non-differential measurement error. In particular, we assume that

$$\mathbb{P}(T = 1|T^* = 0, z, \mathbf{x}) = \alpha_0(\mathbf{x}) \tag{4}$$
$$\mathbb{P}(T = 0|T^* = 1, z, \mathbf{x}) = \alpha_1(\mathbf{x}) \tag{5}$$

and additionally that

$$\mathbb{E}[\varepsilon|T^*, T, z, \mathbf{x}] = \mathbb{E}[\varepsilon|T^*, z, \mathbf{x}] \tag{6}$$

Equations 4–5 amount to the assumption that z and T are conditionally independent given $(T^*, \mathbf{x})$. In other words, z provides no additional information about the process that causes T to be mis-classified above that already contained in T^* and $\mathbf{x}$. In contrast, we allow for the possibility that the measurement error process *does* depend on the exogenous covariates $\mathbf{x}$. Equation 6 states that, given knowledge of true treatment status, the instrument and the exogenous covariates, the *observed* treatment status contains no information about the mean of the regression error term. The assumptions on the measurement error process contained in Equations 4–6 are standard in the literature. Another standard assumption is the condition

$$\alpha_0(\mathbf{x}) + \alpha_1(\mathbf{x}) < 1 \tag{7}$$

which rules out the possibility that $1 - T$ is a better measure of T^* than T is, and vice-versa. This condition is standard in the literature (**?????**). Without it, the treatment effect would only be identified up to sign.

[2]This is because, as we will see below, the Wald Estimator is identified and is proportional to the treatment effect. This estimator exists provided that we have a valid and relevant instrument that takes on at least two values.

	$z = 1$	$z = 1$	$\ldots$	$z = K$
$T = 0$	$\bar{y}_{01}$ p_{01}	$\bar{y}_{02}$ p_{02}	$\ldots$	$\bar{y}_{0K}$ p_{0K}
$T = 1$	$\bar{y}_{11}$ p_{11}	$\bar{y}_{12}$ p_{12}	$\ldots$	$\bar{y}_{1K}$ p_{1K}

Table 1: Observables, using the shorthand $p_{0k} = q_k(1 - p_k)$ and $p_{1k} = q_k p_k$.

Our arguments below, like those of **?** and **?**, proceed by holding the exogenous covariates *fixed* at some level $\mathbf{x}_a$. As such, there is no loss of generality from suppressing dependence on $\mathbf{x}$ in our notation. It should be understood throughout that any conditioning statements are evaluated at $\mathbf{x} = \mathbf{x}_a$. To this end let $c = h(0, \mathbf{x}_a)$ and define $\beta = h(1, \mathbf{x}_a) - h(0, \mathbf{x}_a)$. Using this notation, Equation 1 can be re-expressed as a simple linear model, namely

$$y = \beta T^* + u \tag{8}$$

where we define $u = c + \varepsilon$, an error term that need not be mean zero. In the context of Equation 8 the only observable information consists of the moments of y, conditional on T, z, the conditional probabilities of T given z, and the marginal probabilities of z. For the moment, following the existing literature, we will restrict attention to the conditional mean of y. Let $\bar{y}_{t,k}$ denote $\mathbb{E}[y|T = t, z = z_k]$, let p_k denote $\mathbb{P}(T = 1|z = z_k)$ and let $q_k = \mathbb{P}(z = z_k)$. Table 1 depicts the observable moments for this problem.

The observed cell means $\bar{y}_{tk}$ depend on a number of unobservable parameters which we now define. Let m^*_{tk} denote the conditional mean of u given $T^* = t$ and $z = z_k$, $\mathbb{E}[u|T^* = t, z = z_k]$, and let p^*_k denote $\mathbb{P}(T^* = 1|z = z_k)$. These quantities are depicted in Table 2. By the Law of Total Probability and the definitions of p_k and p^*_k,

$$\begin{aligned} p_k &= \mathbb{P}(T = 1|z = z_k, T^* = 0)(1 - p^*_k) + \mathbb{P}(T = 1|z = z_k, T^* = 1)p^*_k \\ &= \alpha_0(1 - p^*_k) + (1 - \alpha_1)p^*_k \end{aligned}$$

	$z=1$	$z=1$	$\dots$	$z=K$
$T^*=0$	m^*_{01} / p^*_{01}	m^*_{02} / p^*_{02}	$\dots$	m^*_{0K} / p^*_{0K}
$T^*=1$	m^*_{11} / p^*_{11}	m^*_{12} / p^*_{12}	$\dots$	m^*_{1K} / p^*_{1K}

Table 2: Unobservables, using the shorthand $p^*_{0k} = q_k(1-p^*_k)$ and $p^*_{1k} = q_k p^*_k$.

since the misclassification probabilities do not depend on z by Equations 4–5. Rearranging,

$$p^*_k = \frac{p_k - \alpha_0}{1-\alpha_0-\alpha_1}, \quad 1 - p^*_k = \frac{1-p_k-\alpha_1}{1-\alpha_0-\alpha_1}. \tag{9}$$

Equation 9 implies that p^*_k is observable up to knowledge of the mis-classification rates α_0, α_1 since p_k is observable. Thus, the full set of parameters needed to characterize the model in Equation 8 consists of $\beta, \alpha_0, \alpha_1$ and the conditional means of u, namely m^*_{tk} for a total of $2K+3$ parameters. In contrast, there are only $2K$ available moment conditions, namely:

$$\hat{y}_{0k} = \frac{\alpha_1(p_k-\alpha_0)(\beta+m^*_{1k}) + (1-\alpha_0)(1-p_k-\alpha_1)m^*_{0k}}{1-\alpha_0-\alpha_1} \tag{10}$$

$$\hat{y}_{1k} = \frac{(1-\alpha_1)(p_k-\alpha_0)(\beta+m^*_{1k}) + \alpha_0(1-p_k-\alpha_1)m^*_{0k}}{1-\alpha_0-\alpha_1} \tag{11}$$

by the Law of Iterated Expectations, where the observables on the left hand side are defined according to $\hat{y}_{0k} = (1-p_k)\bar{y}_{0k}$ and $\hat{y}_{1k} = p_k\bar{y}_{1k}$. Notice that the observable "weighted" cell mean $\hat{y}_{tk}$ depends on both m^*_{tk} *and* $m^*_{1-t,k}$ since the cell in which $T=t$ from Table 1 is in fact a mixture of both the cells $T^*=0$ and $T^*=1$ from Table 2, for a particular column k.

Clearly we have fewer equations than unknowns. What additional restrictions could we consider imposing on the system? In a very interesting paper, **?** proposes using a three-valued "instrument" that does *not* satisfy the exclusion restriction. By assuming instead that there is no *interaction* between the instrument and the treatment, he is able to prove identification of the treatment effect. Using our notation it is very easy to see why and how **?**'s argument works. His moment conditions are equivalent to Equations 10 and 11 with the additional restriction that $m^*_{0k} = m^*_{1k}$ for all $k = 1, \ldots, K$. This leaves the number of equations unchanged at $2K$, but reduces the number of unknowns to $K+3$. The smallest K for which $K+3$ is at least as large as $2K$ is 3, which makes it clear why **?**'s proof requires that the "instrument" take on at least three values.

Unlike **?**, we, along with **?** and others, assume that z satisfies the exclusion restriction. This implies a different constraint on the m^*_{tk} from Table 2. Since $u = c + \varepsilon$, $\mathbb{E}[\varepsilon|z] = 0$ implies that

$$\mathbb{E}[u|z] = E[u] = c. \tag{12}$$

By the Law of Iterated Expectations, this can be expressed as

$$(1 - p^*_k)m^*_{0k} + p^*_k m^*_{1k} = c$$

for all $k = 1, \ldots, K$. This restriction imposes that a particular weighted sum over the rows of a given column of Table 2 takes the same value *across* columns. Using Equation 9 and rearranging gives

$$\frac{(1 - p_k - \alpha_1)m^*_{0k}}{1 - \alpha_0 - \alpha_1} = c - \frac{(p_k - \alpha_0)m_{1k^*}}{1 - \alpha_0 - \alpha_1}$$

which we can substitute into Equations 10 and 11 to yield

$$\hat{y}_{0k} = \alpha_1(p_k - \alpha_0)\left(\frac{\beta}{1 - \alpha_0 - \alpha_1}\right) + (1 - \alpha_0)c - (p_k - \alpha_0)m^*_{1k} \tag{13}$$

$$\hat{y}_{1k} = (1 - \alpha_1)(p_k - \alpha_0)\left(\frac{\beta}{1 - \alpha_0 - \alpha_1}\right) + \alpha_0 c + (p_k - \alpha_0)m^*_{1k}. \tag{14}$$

Equations 13 and 14 also make it clear why the IV estimator is inconsistent in the face of non-differential measurement error, and that this inconsistency does not depend on the endogeneity of the treatment, as noted by **?**. Adding together Equations 13 and 14 yields

$$\hat{y}_{0k} + \hat{y}_{1k} = c + (p_k - \alpha_0)\left(\frac{\beta}{1 - \alpha_0 - \alpha_1}\right)$$

completely eliminating the m^*_{1k} from the system. Taking the difference of the preceding expression expression evaluated at two different values of the instrument, z_k and z_ℓ, and rearranging

$$\mathcal{W} = \frac{(\hat{y}_{0k} + \hat{y}_{1k}) - (\hat{y}_{0\ell} + \hat{y}_{1\ell})}{p_k - p_\ell} = \frac{\beta}{1 - \alpha_0 - \alpha_1} \tag{15}$$

which is the well-known Wald IV estimator, since $\hat{y}_{0k} + \hat{y}_{1k} = \mathbb{E}[y|z = z_k]$.

Imposing Equation 12 replaces the K unknown parameters $\{m^*_{0k}\}_{k=1}^K$ with a single parameter c, leaving us with the same $2K$ equations but only $K + 4$ unknowns. When $K = 2$ (a binary instrument) we have 4 equations and 6 unknowns. So how can one identify β in this case? The literature has imposed additional assumptions which, using our notation, can once again be mapped into restrictions on the m^*_{tk}. **?**, **?**, and **?** make a *joint* exogeneity assumption on (T^*, z), namely $\mathbb{E}[\varepsilon|T^*, z] = 0$. Notice that this is strictly stronger than assuming that the instrument is valid and the treatment is exogenous. In our notation, this joint exogeneity assumption is equivalent to imposing $m^*_{tk} = c$ for all t, k. This reduces the parameter count to 4 regardless of the value of K. Thus, when the instrument is binary, we have exactly as many equations as unknowns. The arguments in **?**, **?**, and **?** are all equivalent to solving Equations 13 and 14 for β under the added restriction that $m^*_{1k} = c$, establishing identification for this case. **?** use the same argument in a linear model with a potentially continuous instrument, but impose only the weaker conditions that the treatment is exogenous and the instrument is valid. Nevertheless, a crucial step in their derivation implicitly assumes the stronger joint exogeneity assumption used by **?**, **?** and **?**. Without this assumption, their proof does not in fact go through.

If one wishes to allow for an endogenous treatment, clearly the joint exogeneity assumption $m^*_{tk} = c$ is unusable: we are back to $2K$ equations in $K + 4$ unknowns. Based on the identification arguments described above, there would seem to be two possible avenues for identification of the treatment effect when a valid instrument is available. A first possibility would be to impose alternative conditions on the m^*_{tk} that are compatible with an endogenous treatment. If z is binary, two additional restrictions would suffice to equate the counts of moments and unknowns. This is the route followed by **?** in his proof of identification with a binary instrument and endogenous treatment. His Equation (11), expressed in our notation, amounts to adding two cross-column restrictions in Table 2: $m^*_{11} = m^*_{12}$ and $m^*_{01} = m^*_{02}$. Another possibility, suggested by **?**'s approach, would be to rely on an instrument that takes on more than two values. Following this approach would suggest a 4-valued instrument, the smallest value of K for which $2K = K+4$. In the following section we present two of our main results: first **?**'s approach leads to a contradiction, and second, regardless of the value of K, β is unidentified.

4 Nonidentification

4.1 Mahajan's Approach

Here we show that **?**'s proof of identification for an endogenous treatment is incorrect. The problem is subtle so we give his argument in full detail. We continue to supress dependence on the exogenous covariates $\mathbf{x}$.

The first step of the argument is to show that if one could recover the conditional mean function of y given T^*, then a valid and relevant binary instrument would suffice to identify the treatment effect.

Assumption 1 (Mahajan A2)**.** *Suppose that* $y = c + \beta T^* + \varepsilon$ *where* *(i)* $\mathbb{E}[\varepsilon|z] = 0$

(ii) $\mathbb{P}(T^* = 1|z_k) \neq \mathbb{P}(T^* = 1|z_\ell)$ *for all* $k \neq \ell$

(iii) $\mathbb{P}(T = 1|T^* = 0, z) = \alpha_0$, $\mathbb{P}(T = 0|T^* = 1, z) = \alpha_1$

(iv) $\alpha_0 + \alpha_1 < 1$

Result 1 (Mahajan A2)**.** *Under Assumption 1, knowledge of the mis-classification error rates* α_0, α_1 *suffices to identify* β*.*

Proof of Proposition 1. Since z is a valid instrument that does not influence the mis-classification probabilities

$$\mathbb{E}[y|z_k] = c + \beta\mathbb{E}[T^*|z_k] + \mathbb{E}[\varepsilon|z_k] = c + \beta p_k^* = c + \beta\left(\frac{p_k - \alpha_0}{1 - \alpha_0 - \alpha_1}\right)$$

by Equation 9. Since p_k is observed, and z takes on two values, this is a system of two linear equations in c, β provided that α_0, α_1 are known. A unique solution exists if and only if $p_1 \neq p_2$. □

In his Theorem 1, **?** proves that α_0, α_1 can in fact be identified under the following assumptions.[3]

Assumption 2 (Mahajan A1)**.** *Define* $\nu = y - \mathbb{E}[y|T^*]$ *so that by construction we have* $\mathbb{E}[\nu|T^*] = 0$*. Assume that*

(i) $\mathbb{E}[\nu|T^*, T, z] = 0$.[4]

(ii) $\mathbb{P}(T^* = 1|z_k) \neq \mathbb{P}(T^* = 1|z_\ell)$ *for all* $k \neq \ell$

(iii) $\mathbb{P}(T = 1|T^* = 0, z) = \alpha_0$, $\mathbb{P}(T = 0|T^* = 1, z) = \alpha_1$

(iv) $\alpha_0 + \alpha_1 < 1$

[3]Technically, one additional assumption is required, namely that the conditional mean of y given T^* and any covariates would be identified if T^* were observed.

[4]This is **?**'s Equation (I).

(v) $\mathbb{E}[y|T^* = 0] \neq \mathbb{E}[y|T^* = 1]$

Theorem 1 (Mahajan Theorem 1)**.** *Under Assumptions 2, the error rates α_0, α_1 are identified as is the conditional mean function $\mathbb{E}[y|T^*]$.*

Proof of Theorem 1. See **?** Appendix A.1. □

Notice that the identification of the error rates in Theorem 1 does not depend on the interpretation of the conditional mean function $\mathbb{E}[y|T^*]$. If T^* is an exogenous treatment, the conditional mean coincides with the treatment effect; if it is endogenous, this is not the case. Either way, the meaning of α_0, α_1 is unchanged: these parameters simply characterize the misclassification process. Based on this observation, **?** claims that he can rely on Theorem 1 to identify α_0, α_1 and thus the causal effect β when the treatment is endogenous via Result 1. To do this, he must build a bridge between Assumption 1 and Assumption 2 that allows T^* to be endogenous. **?** does this by imposing one additional assumption: Equation 11 in his paper.

Assumption 3 (Mahajan Equation 11)**.** *Let $y = c + \beta T^* + \varepsilon$ where $\mathbb{E}[\varepsilon|T^*]$ may not be zero and suppose that*

$$\mathbb{E}[\varepsilon|T^*, T, z] = \mathbb{E}[\varepsilon|T^*].$$

Result 2. *Suppose that $y = c + \beta T^* + \varepsilon$ where $E[\varepsilon|z] = 0$ and define the unobserved projection error $\nu = y - \mathbb{E}[y|T^*]$. Then Assumption 3 implies that $E[\nu|T^*, T, z] = 0$, which is Assumption 2(i).*

Proof of Result 2. Taking conditional expectations of the causal model,

$$\mathbb{E}[y|T^*] = c + \beta T^* + \mathbb{E}[\varepsilon|T^*]$$

which implies that

$$\nu = y - c - \beta T^* - \mathbb{E}[\varepsilon|T^*] = \varepsilon - \mathbb{E}[\varepsilon|T^*].$$

Now, taking conditional expectations of both sides given T^*, T, z, we see that

$$\begin{aligned}\mathbb{E}[\nu|T^*, T, z] &= \mathbb{E}[\varepsilon|T^*, T, z] - \mathbb{E}\left[\mathbb{E}\left(\varepsilon|\, T^*\right)|\, T, T^*, z\right] \\ &= \mathbb{E}[\varepsilon|T^*, T, z] - \mathbb{E}\left[\varepsilon|\, T^*\right] = 0\end{aligned}$$

by Assumption 3, since $\mathbb{E}[\varepsilon|T^*]$ is (T^*, T, z)–measurable. □

To summarize, **?**'s claim is equivalent to the proposition that under Assumptions 1(i), 2(ii)–(v), and 3, β is identified even if T^* is endogenous. Although Result 1, Theorem 1 and Result 2 are all correct, **?**'s claim is not. While Assumption 3 does guarantee that Assumption 2(i) holds, when combined with Assumption 1(i) it also implies that 2(ii) fails if T^* is endogenous.

Proposition 1 (Lack of a First Stage)**.** *Suppose that Assumptions 1(i) and 3 hold and $\mathbb{E}[\varepsilon|T^*] \neq 0$. Then $\mathbb{P}(T^* = 1|z_1) = \mathbb{P}(T^* = 1|z_2)$, violating Assumption 2(ii).*

Proof of Proposition 1. By the Law of Iterated Expectations,

$$\mathbb{E}[\varepsilon|T^*, z] = \mathbb{E}_{T|T^*,z}\left[\mathbb{E}\left(\varepsilon|T^*, T, z\right)\right] = \mathbb{E}_{T|T^*,z}\left[\mathbb{E}\left(\varepsilon|T^*\right)\right] = \mathbb{E}\left[\varepsilon|T^*\right] \tag{16}$$

where the second equality follows from Assumption 3 and the final equality comes from the fact that $\mathbb{E}[\varepsilon|T^*]$ is (T^*, z)–measurable. Using our notation from above let $u = c + \varepsilon$ and define $m^*_{tk} = \mathbb{E}[u|T^* = t, z = z_k]$. Since c is a constant, by Equation 16 we see that $m^*_{01} = m^*_{02}$ and $m^*_{11} = m^*_{12}$. Now, by Assumption 1(i) we have $\mathbb{E}[\varepsilon|z] = 0$ so that $\mathbb{E}[u|z_1] = \mathbb{E}[u|z_2] = c$. Again using iterated expectations,

$$\begin{aligned}\mathbb{E}\left[u|z_1\right] = \mathbb{E}_{T^*|z_1}\left[\mathbb{E}\left(u|T^*, z_1\right)\right] = (1 - p^*_1)m^*_{01} + p^*_1 m^*_{11} = c \\ \mathbb{E}\left[u|z_2\right] = \mathbb{E}_{T^*|z_2}\left[\mathbb{E}\left(u|T^*, z_2\right)\right] = (1 - p^*_2)m^*_{02} + p^*_2 m^*_{12} = c\end{aligned}$$

The preceding two equations, combined with $m^*_{01} = m^*_{02}$ and $m^*_{11} = m^*_{12}$ imply that $p^*_1 = p^*_2$ unless $m^*_{01} = m^*_{11} = m^*_{02} = m^*_{12} = c$. But this four-way equality is ruled out by the assumption that $\mathbb{E}[\varepsilon|T^*] \neq 0$. $\square$

Add discussion of Mahajan's proof. In particular, explain why he needs there to be a first stage!

4.2 Generic Lack of Identification

We have seen that **?**'s approach cannot identify β when the treatment is endogenous: Assumption 3 in fact implies that the instrument is *irrelevant*. But this alone does not establish that a valid instrument is insufficient to identify β when the treatment is endogenous. In particular, our equation counts from above appear to suggest that a valid instrument that takes on at least four values might suffice for identification. Unfortunately, this is not the case as we now show.

Proposition 2 (Lack of Identification)**.** *Suppose that Assumption 1 holds and additionally that $\mathbb{E}[\varepsilon|T^*, T, z] = \mathbb{E}[\varepsilon|T^*, z]$ (non-differential measurement error). Then regardless of how many values z takes on, generically β is unidentified based on the observables contained in Table 1.*

Proof of Proposition 2. The assumptions of this proposition are the same as those used to derive Equations 13 and 14. These expressions, for $k = 1, \ldots, K$ constitute the full set of available moment conditions. To establish lack of identification, we derive a parametric relationship between β and the other model parameters such that, varying β along this parametric relationship, the observables $(\hat{y}_{0k}, \hat{y}_{1k})$ are unchanged for all k.

Recall from the discussion preceding Equation 15 that the Wald estimator $\mathcal{W} = \beta/(1 - \alpha_0 - \alpha_1)$ is identified in this model so long as K is at least 2.

Rearranging, we find that:

$$\begin{aligned}\alpha_0 &= (1-\alpha_1) - \beta/\mathcal{W}\\ (p_k - \alpha_0) &= p_k - (1-\alpha_1) + \beta/\mathcal{W}\\ 1-\alpha_0 &= \alpha_1 + \beta/\mathcal{W}\end{aligned}$$

Substituting these into Equations 13 and 14 and summing the two, we find, after some algebra, that

$$\hat{y}_{0k} + \hat{y}_{1k} + \mathcal{W}(1-p_k) = c + \beta + \mathcal{W}\alpha_1.$$

Since the left-hand side of this expression depends only on observables and the identified quantity $\mathcal{W}$ this shows that the right-hand side is itself identified in this model. For simplicity, we define $\mathcal{Q} = c + \beta + \mathcal{W}\alpha_1$. Since $\mathcal{W}$ and $\mathcal{Q}$ are both identified, varying either *necessarily* changes the observables, so we must hold both of them constant. We now show that Equations 13 and 14 can be expressed in terms of $\mathcal{W}$ and $\mathcal{Q}$. Conveniently, this eliminates α_0 from the system. After some algebra,

$$\hat{y}_{0k} = \alpha_1(\mathcal{Q} - m^*_{1k}) + \beta(c - m^*_{1k})/\mathcal{W} + (1-p_k)\left[m^*_{1k} - \mathcal{W}\alpha_1\right] \tag{17}$$

$$\hat{y}_{1k} = (1-\alpha_1)\mathcal{Q} + \beta(m^*_{1k} - c)/\mathcal{W} - (1-p_k)\left[m^*_{1k} + \mathcal{W}(1-\alpha_1)\right] \tag{18}$$

Now, rearranging Equation 18 we see that

$$\mathcal{Q} - \hat{y}_{1k} - \mathcal{W}(1-p_k) = \alpha_1(\mathcal{Q} - m^*_{1k}) + \beta(c - m^*_{1k})/\mathcal{W} + (1-p_k)\left[m^*_{1k} - \mathcal{W}\alpha_1\right] \tag{19}$$

Notice that the right-hand side of Equation 19 is the *same* as that of Equation 17 and that $\mathcal{Q} - \hat{y}_{1k} - \mathcal{W}(1-p_k)$ is precisely $\hat{y}_{0k}$. In other words, given the constraint that $\mathcal{W}$ and $\mathcal{Q}$ must be held fixed, we only have *one* equation for each value that the instrument takes on. Finally, we can solve this equation for m^*_{1k} as

$$m^*_{1k} = \frac{\mathcal{W}(\hat{y}_{0k} - \alpha_1 \mathcal{Q}) - \beta(\mathcal{Q} - \beta - \mathcal{W}\alpha_1) + \mathcal{W}^2(1-p_k)\alpha_1}{\mathcal{W}(1-p_k-\alpha_1) - \beta} \tag{20}$$

using the fact that $c = \mathcal{Q} - \beta - \mathcal{W}\alpha_1$. Equation 20 is a manifold parameterized by (β, α_1) that is *unique* to each value that the instrument takes on. Thus, by adjusting $\{m^*_{1k}\}_{k=1}^K$ according to Equation 20 we are free to vary β while holding all observable moments fixed. □

Notice that Equation 20 breaks if $p_k = \alpha_0$. This is because it would reduce the pair of equations to an expression kill the m^*_{1k} term in the pair of equations for this particular value of k. This seems to be related to **?**

Add some discussion of what this proof means, why there is a lack of identification intuitively, why adding more values for the instrument doesn't help, etc.

5 Identification by Homoskedasticity?

Having show that the moment conditions from Table 1 do not identify β regardless of the value of K, we now consider using the conditional *variance* of y. Exploiting this information requires us to make an additional assumption.

Assumption 4 (Homoskedasticity)**.** $\mathbb{E}[\varepsilon^2|z] = \mathbb{E}[\varepsilon^2]$

When combined with the usual IV assumption, $\mathbb{E}[u|z] = 0$, Assumption 4 implies that $Var(\varepsilon|z) = Var(\varepsilon)$. Whether this assumption is reasonable, naturally, depends on the application. When z is the offer of treatment in a randomized controlled trial, for example, Assumption 4 holds automatically as a consequence of the randomization. Similarly, in studies based on a "natural" rather than a controlled experiment, one typically argues that the instrument is not merely uncorrelated with the error term but *independent* of it, so that Assumption 4 follows.

Proposition 3. *Under the conditions of Proposition 2 and Assumption 4, $(\alpha_0 - \alpha_1)$ is identified.*

Proof of Proposition 3. Let s_k^2 denote the conditional variance of y given that $z = z_k$. Then we have:

$$s_k^2 = \beta^2 \sigma^2_{T^*|z_k} + \sigma^2_\varepsilon + 2\beta\sigma_{T^*,\varepsilon|z_k} \tag{21}$$

where $\sigma^2_{T^*|z_k}$ denotes the conditional variance of T^* given that $z = z_k$, $\sigma_{T^*,\varepsilon|z_k}$ denotes the conditional covariance between T^* and ε given that $z = z_k$ and we have used the fact that $Var(\varepsilon|z)$ does not depend on z by Assumption 4. Taking differences across two values z_k and z_ℓ of the instrument, we have

$$s_k^2 - s_\ell^2 = \beta^2\left(\sigma^2_{T^*|z_k} - \sigma^2_{T^*|z_\ell}\right) + 2\beta\left(\sigma_{T^*,\varepsilon|z_k} - \sigma_{T^*,\varepsilon|z_\ell}\right). \tag{22}$$

Now, by the definition of p_k^* and Equation 9,

$$\sigma^2_{T^*|z_k} = p_k^*(1-p_k^*) = \frac{(p_k - \alpha_0)(1 - p_k - \alpha_1)}{(1-\alpha_0-\alpha_1)^2} \tag{23}$$

which implies that

$$\left(\sigma^2_{T^*|z_k} - \sigma^2_{T^*|z_\ell}\right) = \frac{p_k(1-p_k) - p_\ell(1-p_\ell) + (\alpha_0 - \alpha_1)(p_k - p_\ell)}{(1-\alpha_0-\alpha_1)^2}. \tag{24}$$

Turning our attention to the conditional covariances, note that

$$\sigma_{T^*,\varepsilon|z_k} = E_{T^*|z}\left[E\left(T^*\varepsilon|T^* z_k\right)\right] = p_k^*(m_{1k}^* - c) \tag{25}$$

since $E[\varepsilon|z] = 0$. Thus, by Equation 9,

$$\left(\sigma_{T^*,\varepsilon|z_k} - \sigma_{T^*,\varepsilon|z_\ell}\right) = \frac{(p_k - \alpha_0)(m_{1k}^* - c) - (p_\ell - \alpha_0)(m_{1\ell}^* - c)}{1 - \alpha_0 - \alpha_1} \tag{26}$$

Now, substituting $\mathcal{W} = \beta/(1-\alpha_0-\alpha_1)$, we find that

$$\begin{aligned} s_k^2 - s_\ell^2 = \mathcal{W}^2 \left[p_k(1-p_k) - p_\ell(1-p_\ell) + (\alpha_0-\alpha_1)(p_k-p_\ell)\right] \\ +2\mathcal{W}\left[(p_k-\alpha_0)(m_{1k}^*-c) - (p_\ell-\alpha_0)(m_{1\ell}^*-c)\right] \end{aligned}$$

Now, define the observable

$$\widetilde{\mathcal{W}}_{k\ell} = \frac{\mathbb{E}[yT|z_k] - \mathbb{E}[yT|z_\ell]}{p_k - p_\ell}$$

Some algebra shows that, under our assumptions,

$$\begin{aligned} (p_k-\alpha_0)(m_{1k}^*-c) - (p_\ell-\alpha_0)(m_{1\ell}^*-c) = \\ (p_k-p_\ell)\left[\widetilde{W}_{k\ell} - \mathbb{E}[y] - \mathcal{W}\left\{(1-p) + (\alpha_0-\alpha_1)\right\}\right] \end{aligned}$$

where $p = E[T]$. Substituting this into our expression for $s_k^2 - s_\ell^2$ allows us to solve for $\alpha_0 - \alpha_1$ in terms of observables, specifically:

$$\alpha_0 - \alpha_1 = (2p - 1 - p_k - p_\ell) + \frac{2(\widetilde{W}_{k\ell} - \mathbb{E}[y])}{\mathcal{W}} - \frac{s_k^2 - s_\ell^2}{(p_k-p_\ell)\mathcal{W}^2}.$$

□