MONASH University

MONASH BUSINESS SCHOOL

Department of Economics

ISSN number 1441-5429

Discussion number 04/19

A new criterion for selecting valid instruments

Ratbek Dzhumashev[1] and Ainura Tursunalieva[2]

Abstract:
This paper develops a new criterion for selecting a valid instrumental variable (IV). This criterion imposes directional restrictions on an instrument in-formed by the sign of the covariance between the error term and the endogenous variable cov(X,u). We show that a valid IV for the case when cov(X,u) > 0 is not suitable for the case when cov(X,u) < 0, due to the differences in vector orientation. If an IV satisfies the traditional requirements but violates the orientation criterion stemming from the sign of cov(X,u), then the IV estimates exacerbate the bias observed in the ordinary least squares (OLS) method, rather than correcting it. To address this issue, we develop a simple test for the validity of an IV on the basis of the orientation criterion. The test can be applied to non-binary IVs.

Keywords: two-stage least squeare estimation, instrumental variable, vailidty test
JEL Codes: C26, C18

1 Introduction

In this paper, we develop a new criterion for selecting a valid IV for regression analysis. We argue that to choose a valid IV, it is important to take into account the sign of *cov(X,u)*, where *X* is the variable to be instrumented and *u* is the residual term of the given regression equation. The conventional approach to selecting a valid instrument, *Z*, only requires the following two conditions:

1. *Z* is uncorrelated with the disturbance *u*, so that *cov(Z,u)* = 0, and
2. *Z* sufficiently strongly correlated with *X* after controlling for other independent variables.

[1] Department of Econmoics, Monash University. Ratbek.dzhumashev@monash.edu
[2] Department of Statistics, Data Science and Epidemiology, School of Health Sciences, Swinburne University of Technology. aitursunalieva@swinburne.edu.au

However, our analysis supported by empirical evidence show that in addition to the traditional selection requirements imposed on a valid instrument, it is important to consider the sign of $cov(X,u)$. In other words, an IV, valid for a regression model with $cov(X,u) > 0$, is not valid for a regression model with $cov(X,u) < 0$. We find that when an IV satisfies the two traditional conditions (the exclusion restriction for an IV is assumed to hold on theoretical grounds), but violates the orientation criterion stemming from the sign of $cov(X,u)$, the IV estimates would exacerbate the bias observed in the ordinary least squares (OLS) method, instead of correcting it. This aspect of the IV methods has not been investigated in the literature.

The main contribution of the paper is that it demonstrates that a valid IV needs to satisfy an additional orientation criterion based on the sign of $cov(X,u)$, and develops a simple test for checking instrument validity on the basis of the orientation criterion.

In what follows, we present the rationale for this additional criterion for instrument selection and demonstrate the implications of this criterion, using examples of IV estimations.

2 A vector representation of an IV

Consider the classical endogeneity problem in a simple uni-variate linear regression model:

$$Y_i = \beta X_i + u_i, \tag{1}$$

where Y_i is the dependent variable, X_i is the independent variable, u_i is the disturbance term, and $cov(X,u) \neq 0$. The IV method is employed to overcome the endogeneity problem in Equation (1). To implement the IV method, we need to find an instrument Z that satisfies the following two conditions: $cov(Z,u) = 0$, and Z is partially and sufficiently strongly correlated with X after controlling for other independent variables.

A graphical representation of OLS method, when $cov(X,u) > 0$, is illustrated in

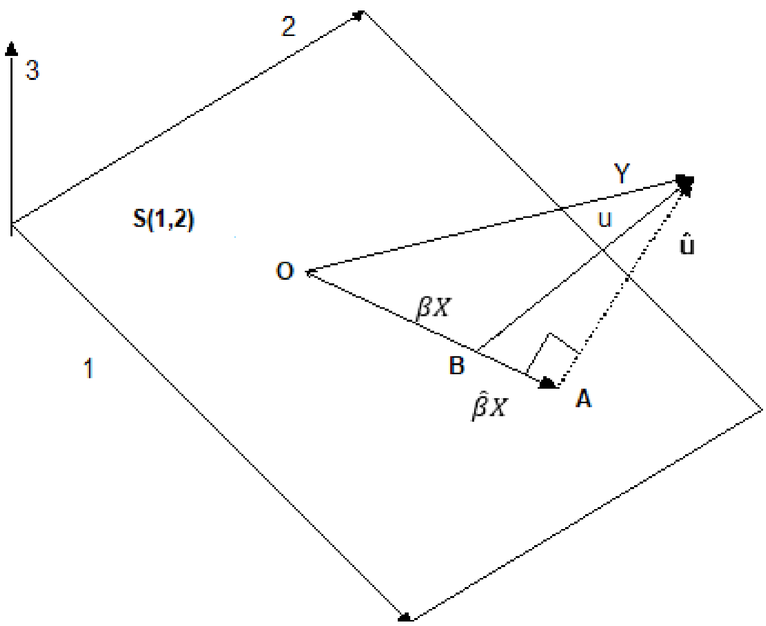


Figure 1: OLS with cov(X,u)>0

Figure 1.[1] In this case, the independent variable X (or a vector sum of regressors) is defined on the plane $S(1,2)$ spanned by vectors 1 and 2. The linear projection of Y on X yields a model that can be written as a projection decomposition:

$$Y_i = \hat{\beta} X_i + \hat{u}_i.$$

In this case, the estimated parameter $\hat{\beta}$ is biased because the disturbance term u is not independent of X.

In Figure 1, the true βX is given by the distance OB, whereas the value of $\hat{\beta} X$ is given by the distance OA. Geometrically, it implies that u is not orthogonal to X, and thus, $cov(X,u) > 0$ for this case. Note that the residual term $\hat{u}$ is orthogonal to $X \in S(1,2)$ because by definition $\hat{\beta} X_i$ is the projection of Y on X since the residual term $\hat{u}$ is orthogonal to X and thus, it is different from the disturbance term u. A problem of this nature is addressed by using IV methods. For example, a valid instrumental variable $Z \perp u$ used in the IV method can be found in the plane orthogonal to the vector u.

[1] See Davidson and MacKinnon (2009, pp.54-56) for a discussion of the geometry of the OLS method.

To understand the intuition of solving the endogeneity problem using an IV, consider the illustration in Figure 2[2]. Note that as the elements of a linear projection of Y on X, the vectors X, Y and u are coplanar (lie on the same plane). In general, an IV vector does not have to be coplanar with Y and X. However, a linear projection of such a vector on the plane spanned by X and Y will preserve the property of being orthogonal to u. Therefore, we can assume that the IV vector, Z, is also coplanar with X and Y, and proceed to the discussion of the geometric interpretation using the two-stage linear squares (TSLS) method.

The true projection βX should be equal to OB, but due to $cov(X,u) > 0$, the OLS method results in $\hat{\beta}X$ being equal to OA. Here, when $cov(X,u) > 0$, the OLS regression estimate based on the linear projection of Y on X includes a bias of magnitude that is equal to the distance BA. If we find a variable Z, such that $cov(Z,u) = 0$, then the projection of X on Z ($\hat{X} = \alpha Z$) will be given by OD, while the projection of Y on Z ($\hat{\beta}_{IV}\hat{X} = \hat{\beta}_{IV}\alpha Z$) will be given by OC. Using the rules for similar triangles, $\frac{OB}{OP} = \frac{OC}{OD}$. Since $\frac{OB}{OP} = \beta$ and $\frac{OC}{OD} = \frac{\beta_{IV}\alpha Z}{\alpha Z} = \beta_{IV} = \beta$, we can determine the true value of β.

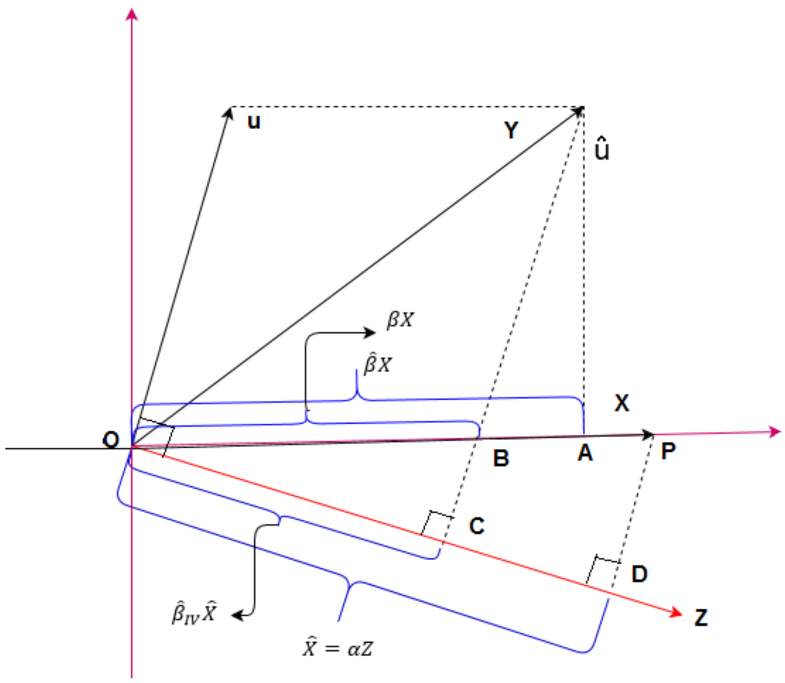


Figure 2: Illustration of the OLS and TSLS, $cov(X,u) > 0$

[2]See Butler (2016) for using such illustrations to describe the two-stage least squares (TSLS) method.

Now, based on the geometric configuration of vectors that constitute the estimates from the OLS and TSLS methods, we develop a new approach that captures the directional differences of regressors, dependent variable and an IV. We express the directional differences of the vectors by using the angular measures between them. Using these angular measures, we formulate an additional criterion for selecting a valid IV. Henceforth, we refer to this condition as the " IV-orientation" criterion.

2.1 Angular measures

In linear algebra, we compare two vectors through their inner product (Towsley et al., 2011). It can be established that

$$cov(X,Y) = SD(X)SD(Y)cos(\theta),$$

where $SD(\cdot)$ is the standard deviation of a given variable, θ is the angle between the two vectors X and Y. By definition the correlation between X and Y defined by

$$\rho(X,Y) = \frac{cov(X,Y)}{SD(X)SD(Y)}.$$

Thus, $\rho(X,Y) = cos(\theta)$, and hence, the angle $\theta(X,Y) = cos^{-1}(\rho(X,Y))$ can be used as a measure of correlation between two vectors. Here, we denote by $\theta(\cdot,\cdot)$ the function that measures an angle between two vectors. This function satisfies the symmetry property, thus, $\theta(X,Y) = \theta(Y,X)$.

Using the fact that we have a vector representation of the dependent variable given as:

$$\vec{Y} = \beta\vec{X} + \vec{u},$$

we can relate the angles $\theta(Z,u)$ and $\theta(X,u)$ to other angular measures of this relationship. It is known that a vector addition given as $\vec{Y} = \beta\vec{X} + \vec{u}$ can be expressed in the component form as:

$$\vec{Y} = \langle Y_1, Y_2\rangle = \beta\langle X_1, X_2\rangle + \langle u_1, u_2\rangle = \langle \beta X_1 + u_1, \beta X_2 + u_2\rangle,$$

where in $\langle\cdot,\cdot\rangle$ stands for the components of a vector found as a projections on the horizontal and vertical axis. By choosing the horizontal axis to be parallel to $\vec{X}$, the components are expressed as follows:

$$\langle Y_1, Y_2\rangle = \langle ||\vec{Y}||\cos\theta(Y,X), ||\vec{Y}||\sin\theta(Y,X)\rangle,$$

$$\langle u_1, u_2\rangle = \langle ||\vec{u}||\cos\theta(X,u), ||\vec{u}||\sin\theta(X,u)\rangle,$$

$$\langle X_1, X_2\rangle = \langle ||\vec{X}||, 0\rangle.$$

Here $||\vec{\cdot}||$ stands for the lengths of a vector. Then, it is straightforward to establish that

$$||\vec{u}||\sin\theta(X,u) = ||\vec{Y}||\sin\theta(Y,X).$$

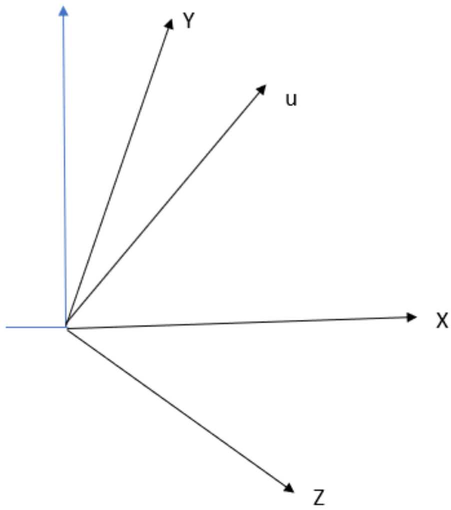
This equation can be written as

$$\theta(Y,X) = \arcsin\left(\frac{||\vec{u}||}{||\vec{Y}||}\sin\theta(X,u)\right).$$

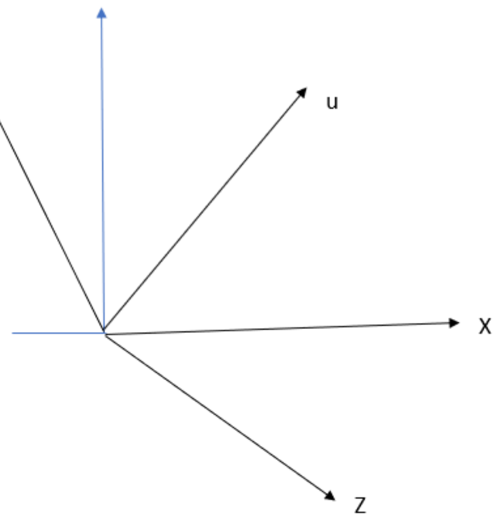
This result implies that $\theta(X,u) > 0$ and $\theta(Y,X) > 0$ are dependent angles.

Next, we determine the relative orientation of the vectors in the regression relationship. Since the orientation of the IV depends on the orientation of the disturbance term, u, we need to consider the case when $cov(X,u) > 0$ separate from the case when it is $cov(X,u) < 0$. To see the reason behind this this point, we consider the following. If $cov(X,u) > 0$, then u and Z should differ in their orientation more than u and X do. That is, the angle between u and Z must be greater than the angle between u and X. This is because when $cov(X,u) > 0$ as in Figure 2, the angle between u and X is less than 90°: $\theta(X,u) < 90°$. Assume we use as an IV vector Z such that the angle between Z and u is defined as $\theta(Z,u) < \theta(X,u)$. Since we know that $X \not\perp u$, therefore, such vector Z is not orthogonal to the disturbance term u, as the angle between such Z and u is smaller than the angle between X and u. Thus, if we used such a vector as an IV, we have *a priori* violation of the orthogonality condition for an IV; hence, $cov(Z,u) \neq 0$. Therefore, when $cov(X,u) > 0$, only a vector Z with orientation satisfying the condition $\theta(Z,u) > \theta(X,u)$ does not rule out the orthogonality of Z and u. In a similar fashion, we argue about the case, when $cov(X,u) < 0$.

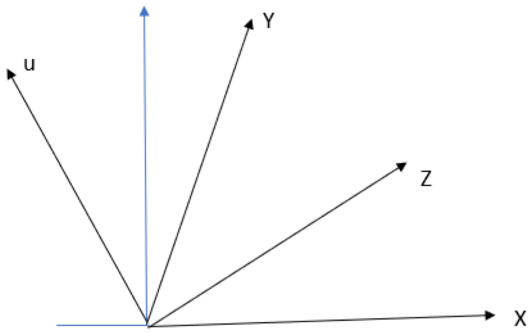
In light of the above discussion, we consider the angular relationships between X, Y and Z for two cases when $cov(X,u) > 0$ and $cov(X,u) < 0$.



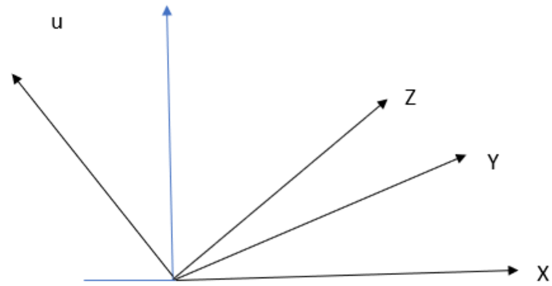
(a) $cov(X,u) > 0$, $cov(X,Y) \gtreqless cov(X,Z)$



(b) $cov(X,u) > 0$, $cov(X,Y) \gtreqless cov(X,Z)$



(c) $cov(X,u) < 0$, $cov(X,Y) < cov(X,Z)$



(d) $cov(X,u) < 0$, $cov(X,Y) > cov(X,Z)$

Figure 3: Different cases of vector configuration

The findings stemming from the above discussion are summarised in the following proposition.

Proposition 2.1 1. *If $cov(X,u) > 0$, then for a valid instrument Z the following angular measure conditions hold:*

$$\theta(Y,Z) > \theta(X,Z) \text{ and } \theta(Y,Z) > \theta(Y,X).$$

2. *If $cov(X,u) < 0$ and $cov(X,Y) > cov(X,Z)$, then for a valid instrument Z the following angular measure conditions hold:*

$$\theta(Y,Z) < \theta(Y,X) \textit{ and } \theta(X,Z) < \theta(Y,X).$$

3. *If $cov(X,u) < 0$ and $cov(X,Y) < cov(X,Z)$, then for a valid instrument Z the following angular measure conditions hold:*

$$\theta(X,Z) > \theta(Y,X) \textit{ and } \theta(X,Z) > \theta(Y,Z).$$

Proof 1. Case $cov(X,u) > 0$ and $cov(X,Y) \gtrless cov(X,Z)$.

The case when $cov(X,u) > 0$ is presented graphically in Figure 3, panels (a) and (b). Note that the orientation of Y is not crucial, in this case. Since $\theta(u,Z) > \theta(u,X)$ and in addition, both $\theta(Z,u)$ and $\theta(X,u)$ are angles between the same vector u for vectors X and Z, we can write

$$\theta(u,Z) \leq \theta(u,X) + \theta(X,Z).$$

Note that the above equality holds when vectors X,Y and Z are coplanar. We also know that $\theta(Y,X) = \arcsin\left(\frac{\|\vec{u}\|}{\|\vec{Y}\|}\sin\theta(X,u)\right)$, thus $\theta(u,Z) > \theta(u,X)$ implies that $\theta(Y,Z) > \theta(Y,X)$. Therefore, when $cov(X,u) > 0$, we state that, for these vectors, the following holds: $\theta(Y,Z) \leq \theta(Y,X) + \theta(X,Z)$. Then given that$\theta(Y,Z) > \theta(Y,X)$ and $\theta(Y,X) > 0$, we have $\theta(Y,Z) > \theta(X,Z)$. This configuration of vectors also implies that: $\theta(Y,X) \gtrless \theta(X,Z)$.

2. Two sub-cases of the condition, when $cov(X,u) < 0$.

Sub-case (1): $cov(X,u) < 0$ and $cov(X,Y) < cov(X,Z)$. This case is illustrated in Figure 3, panel (c).

In this case, $cov(X,u) < 0$ implies that $\theta(X,u) > \theta(Z,u)$. In addition, $cov(X,Y) < cov(X,Z)$ implies that $\theta(Y,X) > \theta(X,Z)$. Therefore, $\theta(Y,X) \leq \theta(X,Z) + \theta(Y,Z)$. Given that $\theta(Y,X) > \theta(X,Z)$ and $\theta(Y,Z) > 0$, we have

$$\theta(X,Z) < \theta(Y,X).$$

In addition, this configuration of vectors implies that:

$$\theta(Y,Z) \gtrless \theta(Y,X).$$

Sub-case (2): $cov(X,u) < 0$ and $cov(X,Y) > cov(X,Z)$. This case is illustrated in Figure 3, panel (d).

In this case, $cov(X,u) < 0$ implies that $\theta(X,u) > \theta(Z,u)$. In addition, $cov(X,Y) > cov(X,Z)$ implies that $\theta(Y,X) < \theta(X,Z)$. Therefore, $\theta(X,Z) \leq \theta(Y,X) + \theta(Y,Z)$. Given that $cov(X,Y) > cov(X,Z)$ and $\theta(Y,X) > 0$, we have

$$\theta(X,Z) > \theta(Y,Z).$$

This configuration of vectors also implies that:

$$\theta(Y,X) \gtrless \theta(Y,Z).\blacksquare$$

The main purpose of this proposition is that the orientation of the vectors given by the angles between pairs of vectors varies depending of whether $cov(X,u) > 0$ or $cov(X,u) < 0$. Specifically, when $cov(X,u) < 0$, one of the two conditions are failed, contrary to the case when $cov(X,u) > 0$, in which both conditions are satisfied. This result indicates that the orientation of a valid IV must be different for the case when $cov(X,u) < 0$ than for the case when $cov(X,u) > 0$. A formal definition of the additional criterion for selecting an IV is that the additional orientation criterion should be used to ensure the selection of a valid instrument.

Corollary 2.2

1. *If $cov(X,u) > 0$, then a valid instrument Z should satisfy the orientation criterion given as the following angle conditions between the vectors: (a) $\theta(Y,Z) > \theta(X,Y)$ and (b) $\theta(Y,Z) > \theta(X,Z)$.*

2. *If $cov(X,u) < 0$, and $cov(Y,X) < cov(X,Z)$ then a valid instrument Z should fail condition (a);*

3. *if $cov(X,u) < 0$, and $cov(Y,X) > cov(X,Z)$ then a valid instrument Z should fail condition (b).*

Proof Verified using Proposition 2.1.$\blacksquare$

Since all the conditions given in Corollary 2.2 are observable, we can use these conditions in selection of an IV given the sign of $cov(X,u)$. Based on Corollary 2.2, we formulate the following instrument validity test procedure.

A procedure for testing instrument validity using the orientation criterion

1. Determine the sign of $cov(X,u)$ based on theoretical considerations of the relationship between Y and X.

2. For vectors Y, X and Z given respectively as the dependent, regressor and IV of the regression model, calculate the angles: $\theta(Y,Z)$, $\theta(X,Y)$ and $\theta(X,Z)$.

3. Given the sign of $cov(X,u)$, apply the condition stated in Corollary 2.2 and determine, whether the selected IV passes the orientation criterion. That is,

   - if $cov(X,u) > 0$ then the null is that $\theta(Y,Z) > \theta(X,Y)$ and $\theta(Y,Z) > \theta(X,Z)$ hold;

   - if $cov(X,u) < 0$, and $cov(Y,X) < cov(X,Z)$ then the null is that $\theta(Y,Z) < \theta(X,Y)$ and $\theta(Y,Z) > \theta(X,Z)$ hold;

   - if $cov(X,u) < 0$, and $cov(Y,X) > cov(X,Z)$ then the null is that $\theta(Y,Z) > \theta(X,Y)$ and $\theta(Y,Z) < \theta(X,Z)$ hold.

2.2 The IV orientation criterion and the OLS bias

We show that ignoring the orientation criterion based on the sign of $cov(X,u)$ in IV estimations exacerbates the bias in the OLS estimation. This point is formulated as the following lemma.

Lemma 2.3 *Using an IV that violates the IV-orientation criterion exacerbates the magnitude of the bias stemming from the OLS estimation in the presence of the endogeneity problem.*

Proof This case is illustrated in Figure 2.2. The angle between Y and X is ψ, the angle between u and X is θ, the angle between X and Z is α, and the angle between X and Z_1 is γ. Thus, from the OLS projection of Y on X we have $\hat{\beta} = \frac{|Y|\cos(\psi)}{|X|}$, whereas the true parameter is given by $\beta = \frac{|Y|\cos(\psi)-|u|\cos(\theta)}{|X|}$. When $cov(X,u) = 0$, we have $\theta = 90°$ and $\cos(\theta) = 0$; thus, in this case the OLS results in unbiased estimate of β. In this case illustrated in Figure 4, however, $cov(X,u) > 0$, hence, $\cos(\theta) > 0$. Therefore, $\hat{\beta} > \beta$; that is the OLS estimate has upward bias.

Now, consider the case of an IV given by Z. We can verify that

$$\beta_{IV} = \frac{|Y|\cos(\psi+\alpha)}{|X|\cos(\alpha)} = \frac{|Y|}{|X|}(\cos(\psi) - \sin(\psi)\tan(\alpha)).$$

Thus, a valid IV corrects the upward bias and when $|Y|\sin(\psi)\sin(\alpha) = |u|\cos(\theta\cos(\alpha)$ holds, $\beta_{IV} = \beta$ is obtained.

If we use Z_1 (see Figure 4) as an IV for this case then, we can verify that we obtain

$$\beta^*_{IV} = \frac{|Y|}{|X|}(\cos(\psi) + \sin(\psi)\tan(\gamma)).$$

Since, γ and $\psi < 90°$, $\sin(\psi)\tan(\gamma) > 0$. Thus, $\beta^*_{IV} > \hat{\beta}$. This implies that using an IV that violates the orientation-criterion results in estimates that have a greater bias than in the OLS estimation. ■

The above result shows that selecting an IV depending on the sign of $cov(X,u)$ is crucial for avoiding the use of an invalid IV.

3 A numerical illustration

First, we test the conclusions about the importance of accounting for the sign of $cov(X,u)$ to choose a valid instrument by considering TSLS estimations based on artificial data. The advantage of this approach is that we know the true values of the disturbance term as we control the generation of dependent variables.

First, as a vector X, we use the variable *lwage* from the dataset used in Mroz (1987). Then, we calculated two version of the disturbance terms: $u'_p = 0.5X + e$ and $u'_n = -0.5X + e$, where e has a standard normal distribution, $N(0,1)$. Thus, we know that $cov(X,u_p) > 0$ and $cov(X,u_n) < 0$. To ensure that $E[u_p] = E[u_n] = 0$, we then demeaned both u'_p and u'_n: $u_i = (u'_i - mean(u'_i)), i \in \{p,n\}$. Given u_p and u_n, we

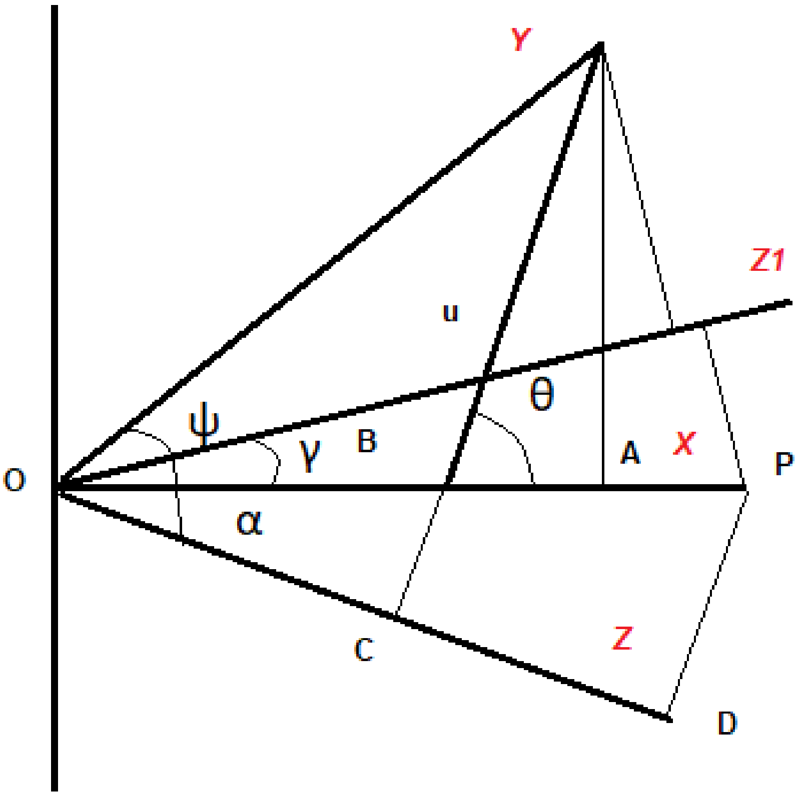


Figure 4: IVs with the correct and wrong orientations

generated our dependent variables: $Y_p = 1 + 2X + u_p$ and $Y_n = 1 + 2X + u_n$. When we regress Y_p or Y_n on X using the OLS method, we do not obtain the true value of the coefficient $\beta = \frac{E[Y]}{E[X]} = 2$ because the condition $cov(X, u_i) = 0$ does not hold. To use the TSLS method, we generated two instruments: Z_p, such that $\lim_{i\to\infty} cov(Z_p(i), u_p(i)) = 0$, and Z_n, such that $\lim_{i\to\infty} cov(Z_n(i), u_n(i)) = 0$.

Now, we can check whether these IVs satisfy the sign condition for $cov(X, u)$. Let us start with Y_p. Since the error term u_p in $Y_p = 1 + 2X + u_p$ has positive covariance with X, the following angle conditions $\theta(Y_p, Z_p) > \theta(X, Y_p)$ and $\theta(Y_p, Z_p) > \theta(X, Z_p)$ should hold. Our calculations give $\theta(Y_p, Z_p) = 61.45° > \theta(X, Y_p) = 15.19°$ and $\theta(Y_p, Z_p) = 61.45° > \theta(X, Z_p) = 58.66°$. Thus, Z_p satisfies the angle conditions for a valid IV, when $cov(X, u) > 0$. If instead, when we used the wrong instrument Z_n, we would have $\theta(Y_p, Z_n) = 48.01° > \theta(X, Y_p) = 15.19°$ and $\theta(Y_p, Z_n) = 48.01° < \theta(X, Z_n) = 48.33°$. Thus, the second angle condition fails in this case. This implies that Z_n is not a valid instrument for the specification given by $Y_p = 1 + 2X + u_p$.

Table 1: Case with a positive covariance

	Dependent variable: Y_p		
	OLS	*IV* Z_p	*IV* Z_n
X	2.522***	2.000***	3.000***
	(0.069)	(0.252)	(0.221)
Constant	0.378***	1.000***	−0.190
	(0.095)	(0.304)	(0.268)
Weak instruments		0	0
Wu-Hausman		0.03	0.02
Observations	428	428	428
R^2	0.760	0.728	0.733
Adjusted R^2	0.760	0.727	0.733
Residual Std. Error (df = 426)	1.025	1.093	1.082
F Statistic	1,352.419*** (df = 1; 426)		

Note: $cov(X, u_p) > 0$, $cov(X, u_n) < 0$ — *p<0.1; **p<0.05; ***p<0.01

Z_p satisfies the orientation criterion, Z_n fails it.

Table 2: Case with a negative covariance

	Dependent variable: Y_n		
	OLS	*IV* Z_n	*IV* Z_p
X	1.522***	2.000***	1.000***
	(0.069)	(0.221)	(0.252)
Constant	1.568***	1.000***	2.190***
	(0.095)	(0.268)	(0.304)
Weak instruments		0	0
Wu-Hausman		0.03	0.02
Observations	428	428	428
R^2	0.536	0.483	0.473
Adjusted R^2	0.535	0.482	0.472
Residual Std. Error (df = 426)	1.025	1.082	1.093
F Statistic	492.649*** (df = 1; 426)		

Note: $cov(X, u_p) > 0$, $cov(X, u_n) < 0$ — *p<0.1; **p<0.05; ***p<0.01

Z_n satisfies the orientation criterion, Z_p fails it.

Now, let us estimate a TSLS regression for $Y_p = 1 + 2X + u_p$, using Z_p as an instrument. Since, by design, the variable satisfies $cov(Z_p, u_p) = 0$ and the orientation criterion for the case when $cov(X, u) > 0$, we are able to exactly estimate the true parameter $\beta = 2$ (see Table 1). If we use the invalid instrument Z_n, we end up with entirely wrong estimates of the parameters (see Table 1, IV Z_n case). If we compare covariances of these instruments with X, then $cov(X, Z_p) = 0.311 < cov(X, Z_n) = 0.356$. Thus, a high correlation between Z_n and X does not imply that Z_n is a better (stronger) instrument than Z_p in this case. Analogously, if we estimate $Y_n = 1 + 2X + u_n$, where $cov(X, u_n) < 0$, then using Z_n as an instrument renders the correct estimates, while using Z_p as an instrument results in incorrect estimates (Table 2).

The important point of the above exercise is that in both examples we see that using an IV that fails the orientation criterion magnifies the bias observed in the OLS estimates.

3.1 Applying the IV orientation criterion to statistical data

In this section, we consider the problem found in Card (1995), who used wage and education data for a sample of men in 1976 to estimate the return to education given by the following equation:

$$lwage = \beta_0 + \beta \cdot educ + \mathbb{X}' \cdot \gamma + u, \tag{2}$$

where $\mathbb{X}' = exper + expersq + black + south + smsa + smsa66 + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668$.

Card (1995) used a dummy variable that indicates whether a subject grew up near a four-year college ($nearc4$) as an IV for the level of education ($educ$). Control variables used in equation (2): experience ($exper$), a black dummy variable ($black$), dummy variables for living in an Standard Metropolitan Statistical Area ($smsa$) and living in the South ($south$), and a full set of regional dummy variables and an SMSA dummy for where the man was living in 1966 ($reg662 ... reg668$). In our estimation, we use $nearc2$ as an additional instrument, an indicator for whether a subject grew up near a two-year college. In this case, we know that the endogeneity stems from omitting the variable measuring the ability level of an individual, which is positively correlated with $educ$. Thus, we have $cov(educ, u) > 0$, in this case. This implies that the OLS overestimates the effect of education on wages because of the omitted variable bias.

Table 3: The effect of education on wages (Card, 1995).

	Dependent variable: lwage			
	OLS	*IV* $nearc2$, $nearc4$	*IV* $fatheduc$, $motheduc$	*IV* artificial instruments
	(1)	(2)	(3)	(4)
educ	0.075^{***}	0.157^{***}	0.102^{***}	0.054^{***}
	(0.003)	(0.053)	(0.013)	(0.008)
exper	0.085^{***}	0.119^{***}	0.101^{***}	0.076^{***}
	(0.007)	(0.023)	(0.010)	(0.007)
expersq	-0.002^{***}	-0.002^{***}	-0.002^{***}	-0.002^{***}
	(0.0003)	(0.0003)	(0.0004)	(0.0003)
black	-0.199^{***}	-0.123^{**}	-0.155^{***}	-0.218^{***}
	(0.018)	(0.052)	(0.026)	(0.020)
south	-0.148^{***}	-0.143^{***}	-0.123^{***}	-0.149^{***}
	(0.026)	(0.028)	(0.032)	(0.026)
smsa	0.136^{***}	0.101^{***}	0.124^{***}	0.146^{***}
	(0.020)	(0.032)	(0.025)	(0.020)
smsa66	0.026	0.015	0.028	0.029
	(0.019)	(0.022)	(0.023)	(0.020)
Constant	4.739^{***}	3.340^{***}	4.230^{***}	5.099^{***}
	(0.072)	(0.895)	(0.223)	(0.142)
Other controls: smsa66, reg662, ...reg668				
Weak instruments		0	0	0
Wu-Hausman		0.085	0.05	0
Sargan		0.26	0.15	0.157
Observations	3,010	3,010	2,220	3,010
Adjusted R^2	0.296	0.166	0.255	0.288
Residual Std. Error	0.372 (df = 2994)	0.405 (df = 2994)	0.380 (df = 2204)	0.375 (df = 2994)
Note:				*p<0.1; **p<0.05; ***p<0.01

Surprisingly, in the TSLS results that used the original instruments (see Table 3), the estimate of the return to education is greater than in the OLS case. Especially when we use $nearc2$ and $nearc4$ as instruments, the effect of education is more than two times greater than the same effect in the OLS. It is easy to explain this counterintuitive result, if we apply the new 3-step procedure of the IV-orientation criterion (discussed on p. 10) to the original IVs.

1. As discussed above, given the omitted variable problem, we have $cov(X,u) > 0$.

2. We calculate the angles $\theta(lwage, IV)$, $\theta(educ, IV)$, and $\theta(lwage, educ)$, where $IV \in \{nearc2, nearc4\}$.

3. From Corollary 2.2 we know that when $cov(X,u) > 0$, then a valid instrument for the wage equation should satisfy both of the following angle conditions between the vectors: $\theta(lwage, IV) > \theta(lwage, educ)$ and $\theta(lwage, IV) > \theta(educ, IV)$.

4. We can verify that both $nearc2$ and $nearc4$ fail the angle conditions for a valid instrument for the case when $cov(X,u) > 0$ (see Table 4).

Table 4: The angle values between $Y = lwage$, $X = educ$, and IV

	Instrumental variables					
	(2)		*(3)*		*(4)*	
	$nearc4$	$nearc2$	$fatheduc$	$motheduc$	Z_1	Z_2
$\theta(lwage, IV)$	33.8°	48.1°	17.7°	14.4°	30.8°	29.82°
$\theta(educ, IV)$	34.3°	48.8°	16.3°	13.9°	28.1°	26.7°
$\theta(lwage, educ)$	10.8°	10.8°	9.1°	9.1°	9.1°	9.1°
Observations	3010	3010	2220	2220	1600	1600

However, the education level of a subject's father ($fatheduc$) and the education level of a subject's mother ($motheduc$), both satisfy this condition. In the latter specification (3) in Table 3, the estimate of the return to education, β, is smaller than specification (2) in Table 3, nevertheless its value is greater than the OLS estimation. In addition, according to the Wu–Hausman test, the estimate of β from specifications (2) and (3) may not be statistically different from the OLS estimate of β at a 5%–significance level.

We notice that, although, $fatheduc$ and $motheduc$ satisfy the IV orientation criterion, they still do not fully correct the bias of the OLS estimation. We explain this in the following way. If $ability$ and $educ$ are highly correlated, then $cov(X,u)$ must be high, which implies that the angle between u and $educ$ should be considerably smaller than the angle between Z and $educ$. From this perspective, although $fatheduc$ and $motheduc$ have correct orientation, they still significantly deviate from the vector that is truly orthogonal to the error term u and statistically not significantly different from the endogenous variable $educ$. The weak Wu–Hausman test result for specification (3) supports this conjecture, as the TSLS estimate of β is not significantly different from the OLS estimate. Therefore, the TSLS estimate of the return to education, using $fatheduc$ and $motheduc$ as IVs, may not be reliable.

Following this rationale, we conjecture that an instrument, satisfying the IV-orientation criterion and has an angle with $educ$ that is greater than the angles between the existing IVs and $educ$ and thus, results on strong rejection of the Wu–Hausman test, will lead to a TSLS estimate of the return to education to be smaller than the OLS estimate. To test this hypothesis, we have generated two variables (Z_1 and Z_2) that satisfy the angle conditions and have a larger angle difference from $educ$ than $fatheduc$ does (see Table 4). The estimate of the return to education, in this case, is lower than the OLS estimate; thus, these artificial instruments remove the effect of $ability$ (see Table 3), at least partially. In addition, the TSLS estimate of β is statistically different from the OLS estimate at a 1%–significance level according to the Wu–Hausman test. Although these instruments may not be meaningful in the usual sense, this experiment nevertheless indicates that if we are able to find such instruments, the estimate of the return to education would be in line with the logic of the omitted variable effect. In sum, this example demonstrates that neglecting the sign of $cov(X, u)$ in selecting IVs can result in using invalid instruments, which would lead to obtaining biased estimates.

It should be noted that using IVs that satisfy the orientation criterion will not guarantee that the bias in the OLS estimate will be removed, albeit it certainly reduces this bias compared to estimates based on IVs that violate the orientation criterion. However, knowing the direction of the bias in the OLS estimate and using this information together with the orientation criterion allows us to select the IVs that not only satisfy the orientation criterion, but also results in a lower bias than the OLS estimate.

4 Conclusion

This paper shows that for correct identification in the two-stage least square estimation, valid IVs should satisfy an extra orientation criterion in addition to the conventional ones. This criterion depends on the sign of the covariance between the error term and the endogenous variable: $cov(X, u)$. We find that when an IV satisfies only the two traditional conditions but violates the IV-orientation criterion, the IV estimates exacerbates the bias observed in the the ordinary least squares (OLS) method, instead of correcting it. This paper develops a rule to guide selection of IVs depending on the sign of $cov(X, u)$. Using simulated data along with observed data, the paper demonstrates empirical evidence about the implications of ignoring this additional orientation criterion for IVs.

References

Butler, R. J. (2016). The simple geometry of correlated regressors and iv corrections. *International Journal of Statistics in Medical Research 5*, 182–188.

Davidson, R. and J. G. MacKinnon (2009). *Econometric Theory and Methods* (Second ed.). Oxford University Press, New York, USA.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica 55*(4), 765–799.

Towsley, A., J. Pakianathan, and D. H. Douglass (2011). Correlation angles and inner products: Application to a problem from Physics. *ISRN Applied Mathematics 2011*(Article ID 323864), 1–12.