Simulating Instrumental Variables and Identifying Causality between High School Completion and Wage Outcomes

Jerry Hong

October 8, 2024

Abstract

In this discussion, the primary focus was on the use of instrumental variables (IV) and the construction of the Wald estimator in econometric modeling. We began by setting up a data-generating process (DGP) to study the relationship between high school completion, wages, and both observed and unobserved covariates, incorporating endogeneity concerns. To address this endogeneity—particularly the correlation between high school completion and unobserved factors like ability—we introduced three instruments: Z_1 and Z_2 (randomly assigned) and Z_3 (partially dependent on unobserved ability).

Using simulated data through R, we estimated model parameters through Ordinary Least Squares (OLS) and IV regressions. We performed separate IV regressions using Z_1, Z_2, Z_3, and combinations thereof to evaluate the causal impact of high school completion on wages. Testing for the correlation between the instruments and covariates ensured the instruments' validity. We emphasized the importance of robust standard errors and used the `coeftest` function from the `lmtest` package for statistical inference. We then create a table of each of the parameter estimates to see how they compare with their true values from the DGP, demonstrating any possible strength of IV methods for addressing endogeneity in estimating returns to education.

Additionally, we introduced the Wald estimator, an IV estimator particularly useful when instruments are binary. We outlined steps for constructing the Wald estimator, discussing the importance of conditioning on observed covariates for consistency and efficiency.

1 Initialize the Data Generating Process

We will set up the Data Generating Process (DGP) to simulate the estimated returns on wages from attending high school with the accompanied probabilities and instrumental variables. We will initialize with 1,000 observations. We set the true values of the parameters as 1. A random seed of 123 is set for replication.

1.1 Setting Up the Parameters and Covariates

Consider the following DGP:

$$\log(Y_t) = \tau T_i + \beta_0 + X_i'\beta + A_i'\gamma + \varepsilon_i,$$

where Y_i is Wages, $T_i = 1[\text{Person finished high school}]$, X_i is an observed covariate (e.g., parents' education), A_i is an unobserved covariate (e.g., ability), and ε_i is the error term. For simplicity, assume that X, A, and ε are all standard normal, i.e., distributed $N(0,1)$; and set $\tau = \beta_0 = \beta = \gamma = 1$. We also have three scholarships, Z_1, Z_2, and Z_3, which affect the probability of attendance in high school. Z_1 and Z_2 are randomly assigned with probability 0.5. Z_3 is almost exogenous - it has some randomness but some dependence on ability:

$$Z_3 = 1[\varepsilon_3 + A_i > 0],$$

where $\varepsilon_3 \sim N(0,3)$. T is positively correlated with both X and A, as well as with the instruments. The high school attendance equation is:

$$T_i = 1[5 \cdot Z_{1i} + 0.01 \cdot Z_{2i} + Z_{3i} + X_i + 10 \cdot A_i + \varepsilon_T > 0],$$

where $\varepsilon_T \sim N(0,2)$. The table 1 below shows the summary statistics after 1,000 observations.

Table 1: Summary Statistics

Statistic	Y	X	A	Z1	Z2	Z3	T
Min.	-4.584	-2.810	-3.048	0.000	0.000	0.000	0.000
1st Qu.	0.190	-0.628	-0.653	0.000	0.000	0.000	0.000
Median	1.661	0.009	0.055	0.000	0.000	0.000	1.000
Mean	1.656	0.016	0.042	0.486	0.499	0.494	0.618
3rd Qu.	3.064	0.665	0.753	1.000	1.000	1.000	1.000
Max.	8.109	3.241	3.390	1.000	1.000	1.000	1.000

2 Running Various Models with Instrumental Variables

We will be running five different models: one with OLS and the others using different combinations of instrument variables.

- OLS
- IV instrumenting with $Z1$
- IV instrumenting with $Z2$
- IV instrumenting with $Z3$
- IV instrumenting with $Z1$ and $Z2$

2.1 Parameter Estimated Results

For each model, we will extract the appropriate coefficients and see how they compare to their true parameter values.

Table 2: Parameter Estimates

	Dependent variable: Logarithm of Wages				
	OLS	*Instrumental Variable*			
	(1)	(2)	(3)	(4)	(5)
High School (T)	0.921***	0.385	−1.151	0.533	0.374
	(0.098)	(0.391)	(4.037)	(0.845)	(0.390)
Observed Covariate (X)	0.979***	0.979***	0.980***	0.979***	0.979***
	(0.031)	(0.032)	(0.038)	(0.032)	(0.032)
Unobserved Covariate (A)	1.056***	1.253***	1.814	1.198***	1.257***
	(0.047)	(0.146)	(1.477)	(0.311)	(0.146)
Constant	1.027***	1.350***	2.275	1.260**	1.356***
	(0.067)	(0.238)	(2.432)	(0.510)	(0.237)
R^2	0.772	0.766	0.670	0.769	0.765
Adjusted R^2	0.772	0.765	0.669	0.768	0.765

Note: Each model has 1,000 observations *p<0.1; **p<0.05; ***p<0.01

Overall, OLS produces the strongest and significant results at the 1% level, suggesting a strong and positive effect of high school completion on wages. It also produces the coefficients that are closest to the true parameter values set in the DGP. The IV models generally produce mixed results of varying significance, which raises potential challenges of the validity and reliability of identifying causal effect due to challenges of endogeneity and instruments used. As all models produce a significant

and consistent positive impact for the observed covariate, they are important in wage determination. However, the variation throughout all models suggest further investigation and checks for robustness to ensure reliability of these findings, especially the incorporation of the instrumental variables.

2.2 Determining an Instrument is Correlated with the Covariate

We will test if the instrument Z_1 is statistically correlated with the covariate X. This will be done through an OLS regression with robust standard errors. The table 3 below are the results.

Table 3: OLS Regression Results: Instrument (Z1) on Covariate (X)

Coefficients	**Estimate**	**Std. Error**	**t-value**	**Pr**$>|t|$
Intercept	0.086	0.044	1.975	0.049 *
Instrument (Z1)	-0.144	0.063	-2.303	0.022 *
Residual standard error	0.990 (998 DF)			
Multiple R-squared	0.0053			
Adjusted R-squared	0.0043			
F-statistic	5.302 on 1 and 998 DF	p-value: 0.022		

*Note: *p<0.1; **p<0.05; ***p<0.01*

According to above results, the estimate of the instrument Z_1 is -0.144, with a standard error of 0.063 and a statistically significant p-value of 0.022 (at the 5% significance level). This implies that on average, individuals with $Z_1 = 1$ have lower values of X as in parents' education. As this is statistically significant at 5% significance level suggests that this relationship is not through random chance. With an R-squared value of 0.0053, this indicates that the instrument of choice explains only a small portion (0.53%) in the variance of X. Thus, while significant, this relationship is very small in magnitude.

The exclusion restriction in IV analysis states that the instrument should affect the outcome (wages) only through the treatment (high school completion) and not through unobserved variables. Since Z_1 is negatively correlated with X, this violates the exclusion restriction. Thus, this instrument may affect wages not just through T, but also through X. In the context of developing countries, where access to education are influenced by numerous social and economic factors, Z_1 may not be a valid instrument given its correlation with the unobservable X. As such, this analysis suggests evidence against the validity of the exclusion restriction for Z_1. Because it may influence wages through other pipelines, this instrument is not likely to satisfy the strict exclusion restriction needed for a convincing analysis of returns to education. Thus, in this context, we should ere caution when using Z_1 in IV analysis.

2.3 Simulate with 100,000 Observations

Here, we will rerun the DGP but with 100,000 observations. The table 4 below is the summary statistics.

Table 4: Summary Statistics with 100,000 Observations

Statistic	**Y**	**X**	**A**	**Z1**	**Z2**	**Z3**	**T**
Min.	-6.632	-4.382	-4.289	0.000	0.000	0.000	0.000
1st Qu.	0.245	-0.671	-0.668	0.000	0.000	0.000	0.000
Median	1.662	0.002	0.005	0.000	0.000	0.000	1.000
Mean	1.618	0.002	0.005	0.499	0.500	0.499	0.613
3rd Qu.	3.012	0.676	0.677	1.000	1.000	1.000	1.000
Max.	9.683	4.323	4.124	1.000	1.000	1.000	1.000

From the above results, we see the summary statistics are a little more precise to the normal distribution, mainly the instruments and the treatment variable. As far as the variables Y, X, and A, the means decreased considerably while expanding their tail values.

2.4 The Wald Estimator

The Wald Estimator can be used to estimate the causal effect of a treatment when dealing with observational data. It takes the ratio of the simple difference outcomes for some instrument Z and the first-stage compliance coefficient. In this context, we want to identify the causal impact of an endogenous variable like education on the outcome of wages using instruments that are correlated with the endogenous variable but uncorrelated with the error term. The Wald Estimator will utilize the instruments to isolate the variation in the exogenous treatment variable (high school completion) on the outcome of wages. We will calculate the expectations from a sub-sample of the instrument Z_1 and the treatment variable.

Before we start, it is important to understand whether we should condition on X. We assess on whether we should based on its consistency and efficiency. Consistency refers to whether an estimator converges to the true parameter value as we increase sample size. Because we increase from 1,000 to 100,000 observations, we suspect that the Wald Estimator would be consistent towards the true parameter value. Regarding efficiency, it shows how well the estimator uses the provided information, where the smallest variance makes the estimator most efficient. Overall, conditioning on X can help improve consistency without sacrificing efficiency, especially if the observable characteristics might be correlated with both the instrument and the treatment variables.

We setup the Wald Estimator by taking the ratio of the difference in the expectations of the instrument over the difference in the expectations of the treatment. The other method is to use *ivreg* to run an IV regression model and extract the coefficient of the treatment. Ideally, both methods should return the same Wald Estimator.

For each method, both returned the same Wald Estimator of **1.095**, which is almost in line with the true treatment estimate of 1.

Disclaimer

The figures and models presented in this document have been generated with the assistance of generative artificial intelligence tools. However, the findings are reviewed and verified by the author(s) to ensure accuracy and clarity.