Advanced Statistics & Programming

Sanjana Aneja

November 25, 2025

Contents

1 Instrumental Variable Analysis

1.1 Instrumental variable theory

1. When the causal variable of interest (education) is endogenous, that is, there are unobserved factors that affect it, then the OLS of wages on education is biased as education is not independent of the error term (Assumption A.3). Here, including a valid instrument (IV) as a control would only state the effect of education on wages holding the IV fixed. Meaning it would do nothing to address the covariation between education and the error term. Thus, IVs are not just another covariate, but are used to create clean variation in the problematic variable, which is then used through IV/2SLS analysis to identify causality.

2. Let us suppose that an OLS regression of wages on education is performed. Linear regression assumes that there should be mean independence $E[\varepsilon_i \mid X_i] = 0$ (A.3). However, there are certain biases that could violate this assumption, affecting the estimated effect of education on wages.
   (i) Omitted variable bias: Omitted or unobserved variables, such as a person's ability, can affect both years of education and wages. If education is positively correlated with the omitted variable ability, that is, more able students earn more and study longer, then $E[\varepsilon_i \mid X_i] \neq 0$ and the OLS estimates are biased. They are biased as OLS attributes a part of ability's effect to the β of education.
   (ii) Endogenous treatment: Under endogenous treatment, the bias is that X (treatment group) is a choice that partly depends on unobserved factors that also affect wages (outcome variable). Thus, these effects of unobserved factors get captured by ε, making selection into X dependent on ε and violating assumption A.3. Thus, here because treatment choice and ε move together, the OLS is biased and inconsistent.
   (iii) Sample/selection bias: When analysing wages, data is often observed only for the individuals who are working. If education affects selection into employment and this selection is related to ϵ, then the regression happens on a non-random sample, violating A.3. For example, certain low-educated individuals with low unobserved wages who do not work are not recorded in the data. The low-educated people who are observed are those who are better paid and employed. Thus, there is positive selection (better wages) in the low-educated group, which can shrink the observed wage difference and bias the OLS esti-

mate.

3. Let Z be the geographic proximity to a college, which is used as an instrumental variable (IV) in the regression of wages on education. For identification of a valid IV, we need (i) relevance: $E[X \mid Z = 1] \neq \mathbb{E}[X \mid Z = 0]$ and (ii) cleanness (exogeneity): $E[\varepsilon \mid Z = 1] = \mathbb{E}[\varepsilon \mid Z = 0]$. However, some concerns could render proximity as an IV invalid.
(i) Relevance: Proximity is invalid if it barely shifts years of education in the sample. For instance, if an individual lives far away from a college but has fast commuting options to the college, then proximity is rendered irrelevant. This means that Z is not strongly correlated to X, and does not differ among different groups.
(ii) Cleanness: Proximity to a college could be dependent on factors such as parental resources, an advantage that directly increases wages. For instance, an individual could secure a high-paying job through high-earning parents with strong social networks. Thus, here, proximity Z is correlated to the error term. This means Z is not independent of the error term, making it invalid.

1.2 Instrumental Variable Application

1. Table 4 presents the results of the summary statistics of 15 relevant variables highlighted in Table 2 of the assignment.
The results showcase that first, the dependent variable *lwage* has a mean of 6.262 and a S.D. of 0.444. This mean corresponds to the mean of the *wage* variable of about $577. Taking the log of wages has stabilised the skewness of the wage data, allowing us to approximate percentage changes in wages per 1-year change in education. Second, the main explanatory variable *educ* has a mean of 13.623, a S.D. of 2.677, and a range of 1-18 years. Educ measures the years of schooling per observation and is potentially an endogenous causal variable whose effect on wages is the core of this analysis. Third, the instrumental variable *nearc4* serves as a binary indicator of whether an individual lived near a 4-year college in 1966. It has a mean of 0.682 and a S.D. of 0.466, which means that 68.2% of observations in the data lived near a 4-year college. Similarly, *nearc2* serves as a binary indicator of whether an individual lived near a 2-year college in 1966. It has a mean of 0.441 and a S.D. of 0.497, which means that 44.1% of observations in the data lived near a 2-year college. In this sense, nearc2 captures the alternative local access to post-secondary education compared to

nearc4. Lastly, *exper* captures the years of labor market experience calculated through using columns age and educ (age - educ - 6 = exper). It has a mean of 8.856, S.D. of 4.142, and a wide range (0-23), meaning on average, experience is around 9 years but can range widely.

Table 1: Summary Statistics for relevant variables

Statistic	N	Mean	St. Dev.	Min	Max
id	3,010	2,581.749	1,500.539	2	5,225
nearc4	3,010	0.682	0.466	0	1
nearc2	3,010	0.441	0.497	0	1
age	3,010	28.120	3.137	24	34
educ	3,010	13.263	2.677	1	18
lwage	3,010	6.262	0.444	4.605	7.785
wage	3,010	577.282	262.958	100	2,404
fatheduc	2,320	10.003	3.721	0	18
motheduc	2,657	10.348	3.180	0	18
IQ	2,061	102.450	15.424	50	149
married	3,003	2.271	2.067	1	6
momdad14	3,010	0.789	0.408	0	1
sinmom14	3,010	0.101	0.301	0	1
step14	3,010	0.039	0.193	0	1
exper	3,010	8.856	4.142	0	23

2. An instrument is considered relevant if $E[X \mid Z = 1] \neq \mathbb{E}[X \mid Z = 0]$.To check for the relevance of the instrument, we first regress the endogenous variable against only the instrument, and apply a partial F-test to see whether nearc4 is correlated with educ in the raw data (fs_min). Then we regress educ against all exogenous variables plus the instruments, and apply the same partial F-test to test $H_0 : \beta_{\text{IV},1} = \cdots = \beta_{\text{IV},L} = 0$ against H_1 : not all $\beta_{\text{IV},\ell}$ equal to 0 (fs_full). In this case, the exogenous variables included are those that were fixed (predetermined) before the child's schooling choice. Variables including *married*, *experience*, and *IQ* cannot be safely predetermined and can bias results. Thus, the model analysis is performed under the assumption that only age, family structure at 14, and parents' education are relevant controls.
The results of the partial F-test (ANOVA) for the first model, fs_min, show that the observed $F = 63.912 (p < 0.001)$. This rejects the null-

hypothesis $H0$ that $\beta_{nearc4} = 0$, implying that nearc4 significantly contributes to educ, in the presence of no exogenous variables. The results of the second partial F-test for the fs_full model show the observed F-statistic is equal to 10.82 ($p < 0.01$), which again implies that nearc4 significantly contributes to educ conditional on exogenous variables. Thus, nearc4 is relevant both unconditionally (fs_min) and conditionally (fs_full).

Furthermore, the graphical representation of a boxplot of educ by nearc 4 derives the same outcomes. Figure 2 indicates that individuals living near a 4-year college have a higher mean and median than those living far away from one. Thus, Figure 2, along with the regression and subsequent partial F-test results, showcases that *nearc4* is a relevant instrument for *educ*.

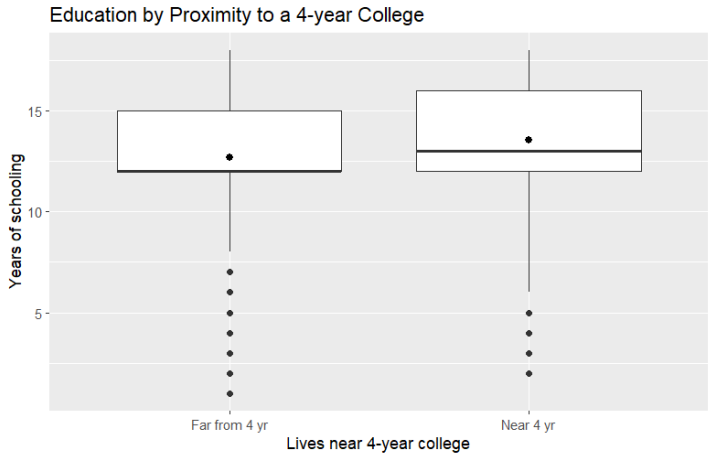

Figure 1: Boxplot of educ by nearc4

3. The IV regression analysis uses nearc4 to isolate the exogenous variation in educ (years of schooling) to estimate its causal effect on lwage (log hourly wages in cents). Table 5 presents the summary of two models with and without robust SEs. First, the base just-identified model rsltA (no controls) has $\hat{\beta}_{educ} = 0.188$, which is highly significant ($p < 0.001$), and has a S.E. of 0.026. As the outcome is in logs, one additional year of education is associated with approximately a 20.7% increase in wages.

   Adding relevant pre-determined controls to this model (age, parental education, family structure at 14) gives us rsltB. This model has a

$\hat{\beta}_{educ} = 0.270$ and is again significant ($p = 0.003$). This means that there is a larger causal effect in rsltB compared to rsltA, implying that one additional year of education is associated with around a 31.0% increase in wages. However, in rsltB, the S.E. increases to 0.090 as the first stage under controls only uses the variation in educ that remains after conditioning on those controls. Regardless, the estimate of $\hat{\beta}_{educ}$, in both model rsltA and rsltB, remains significant, with rsltB implying a larger causal effect in wages.

Table 5 showcases that using robust SEs does not critically affect the inferences in either the base (A: Robust) or the controlled model (B: Robust). In rsltA, both the conventional and robust SE for educ are 0.026 and provide the same highly significant estimate. Similarly, in rsltB provides the same 0.090 SE in both conventional and robust models. The other coefficients in the model only show minor deviation in SE, but no significant change. Thus, using heteroskedastic robust SEs does not change the statistical inferences, meaning an additional year of education remains significant in both the baseline and controlled IV regressions.

Table 2: IV regression analysis: Base and Control model (Conv vs Robust SEs)

	Dependent variable:			
	Log hourly wage (lwage)			
	A: Conv.	A: Robust	B: Conv.	B: Robust
	(1)	(2)	(3)	(4)
Constant	3.767***	3.767***	2.386***	2.386***
	(0.349)	(0.347)	(0.789)	(0.773)
educ	0.188***	0.188***	0.270***	0.270***
	(0.026)	(0.026)	(0.090)	(0.090)
age			0.032***	0.032***
			(0.007)	(0.007)
fatheduc			−0.046**	−0.046**
			(0.021)	(0.021)
motheduc			−0.036*	−0.036*
			(0.019)	(0.019)
momdad14			0.161	0.161
			(0.168)	(0.135)
sinmom14			0.219	0.219
			(0.237)	(0.235)
step14			0.227	0.227
			(0.217)	(0.192)
Observations	3,010	3,010	2,220	2,220
R^2	-0.574	-0.574	-1.250	-1.250
Adjusted R^2	-0.574	-0.574	-1.257	-1.257
Residual Std. Error	0.557 (df = 3008)	0.557 (df = 3008)	0.661 (df = 2212)	0.661 (df = 2212)
Note:				*p<0.1; **p<0.05; ***p<0.01

The results of IV using nearc4 and nearc2 for both the base and the controlled model are summarised in Table 7 (Appendix 2). In the base model, with nearc2, the estimate of educ is much larger ($\hat{\beta}_{educ} = 0.343$), implying a 41% increase in wages per additional school year. However, though the estimate is significant, the first-stage weak-instrument F value is only 6.76, implying nearc2 is potentially a weaker instrument. After adding controls in the IV regression, nearc2 essentially loses all first-stage power ($F \approx 0.09$). Then the second-stage coefficient of educ becomes negative and insignificant ($\hat{\beta}_{educ} = -2.508$). This cements the idea that nearc2 is a weak instrument, as once relevant exogenous factors are controlled for, the instrument no longer explains the variation in years of schooling. Thus, nearc2 is not a valid instrument and nearc4 should be used as the instrument in the controlled IV regression model.

4. Following the previous analysis, OLS estimates are obtained for both the base (no controls) and the controlled model. (i) OLS Base model:

The $\hat{\beta}_{educ} = 0.0521$, an additional year of schooling leads to around a 5.3% increase in wages compared to 20.7% in IV base model. (ii) OLS controlled model: $\hat{\beta}_{educ} = -0.033$, an additional year leads to around 3.3% increase in wages compared to 31.0% in IV controlled model. Across both models, IV leads to much larger estimates than OLS. This is consistent with the lecture's discussion of biased and inconsistent OLS estimates due to an endogenous treatment.

To formally test this difference, the Wu-Hausman exogeneity test is used. In the base model, the Wu-Hausman test is equal to 48.45 ($p < 0.001$) which is highly significant. Similarly, in the controlled model, the Wu-Hausman test is equal to 19.38 ($p < 0.001$), which is again highly significant. Thus, in both models, we reject the exogeneity of educ (H0) and prefer the use of IV/2SLS estimated models over the OLS models.

As only one instrument variable is used for the endogenous treatment, the IV is just-identified, so over-identification cannot be formally tested. Over-identification will only be an issue if more instruments than the endogenous regressor are added, for instance, nearc4 and nearc2. Then a Sargan-Hansen X^2 test can check the cleanness of the additional instruments. When a Sargan X^2 test is performed on a model using both nearc2 and nearc4, the test value is equal to 10.581 ($p = 0.001$). This means that at least one instrument in the model is invalid (nearc2 in this case). Thus, for this IV analysis, nearc4 is the only valid instrument.