EC413 Applied Econometrics Assignment Instrumental Variables

Word Count: 1477

Student Number: 202133230

1) Regression Specification & Data Discussion

The data provided for this report is from the 1976 National Longitudinal Survey of Young Men in the US, consisting of 34 variables and 3010 observations. A shortlist of variables was created based on intuition that will be discussed in further detail as well as a correlation matrix. Following data cleaning, the shortlist data frame consisted of 10 variables and 1600 complete observations, representing 53.2% of the full dataset

$$lwage = a + High\_edu + Age + Moth + Fath + Singm + Black$$

$$lwage = a + High\_edu + Age + Moth + Black$$

Where *lwage* are the logged wages of candidates - the dependent variable of interest. *High_edu* is the explanatory variable of interest, taking a value of 0 if years of education is below 12 i.e. the candidate did not complete high school. And 1 when the opposite is true. The following variables are the proposed controls that if left out of the model, could lead to an omitted variable problem. *Age* and *Moth* are controls for the age of the candidate and the years of the candidate's mother's education respectively. *Fath* and *Singm* are the years of the candidate's father's education and a dummy variable representing whether the candidate was raised by a single mother at age 14. *Black* is a dummy for the race of the candidate.

The intuition behind the initial choice of controls are as follows: Age was the first choice as wages are expected to rise with experience as candidates move into more senior positions over the course of their careers. Yet experience was excluded from the control set because years of education, a variable the instrument can affect, is a parameter in the experience variable. Thus cannot be used as a control as it would violate the exclusion restriction.

Years of education of parents could influence wages as more educated parents can better inform candidates about the job market and guide them through the opportunities available. This may lead to better career decisions over the course of at least the early careers of candidates. A channel through which the higher education attendance rate is influenced could be the parent's own expectation on higher education attendance being projected onto the candidate, as well as parenting styles that encourage curiosity and study (Davis-Kean, 2005).

The inclusion of the single mother and race dummy variables were to capture and control for the effect of family structure and racial disparities on income in the 1960-70s. But as shown in Figure 1.1, *Fath* and *Singm* had insignificant beta estimates at a 5% level and were subsequentially dropped from the final specification to maintain simplicity. Doing so decreased the AIC of the model from 1508.9 to 1505.3, further supporting the motion

to exclude them. Thus resulting in the final specification to explain the relationship between higher education attendance and wages.

To improve on this, including regions as dummy variables could make the model more robust as regional differences in prosperity that are omitted from this model could be captured and significance measured.

2) Estimating Regression

$$lwage = 4.737 + 0.856\, High\_edu + 0.0494\, Age - 0.0285\, Moth - 0.129\, Black$$

Using a two stage least squares approach the above regression was estimated, the results of which are summarised in Figure 1.2. With the Beta estimates shown above and standard errors of; 0.144, 0.35, 0.004, 0.019 and 0.046 respectively, all estimates but Mother's education were significant at least at a 5% level. The model suggests that attending higher education is associated with an 85% increase in expected wages. This is an enormous increase that is particularly strange. It is not supported by a simple check of taking the average difference between the set that attended higher education and those that didn't that results in a more reasonable 31% increase. Suggesting the model is biased, suffering from an overestimation of the true effect of higher education on wages. Evidently some level of omitted variable bias has crept into the model. This can be avoided in the future by using a more robust selection criteria for the choice of controls. Measurement Error and reverse causation are unlikely to be an issue in this case.

3) Instrumental Variable

The first assumption necessary for the IV is that it is relevant and not only intuitively, but also mechanically a good predictor of the endogenous variable in question. This was tested by ensuring there is a significant first stage where the beta estimate is significant as well as an F-statistic greater than 10 – though this benchmark has more recently been brought into question. With reference to the first stage shown in Figure 1.4, there is a highly significant beta estimate of 0.082 indicating that living near a 4yr college is associated with an increase of 8% likelihood of attending higher education. This makes the instrument a weak predictor of attending higher education further emphasised by the low first stage F-statistic of 10.75, only barely above the benchmark. This raises the concerns of a weak instrument problem as the correlation strength in the first stage is low.

The next assumption is that of independence, stating that the instrument must be uncorrelated with the errors of the structural model to isolate movements in the explanatory variable in question while filtering out other factors. If this assumption is

violated the instrument becomes endogenous and invalid. As shown in Figure1.4 there is no concern surrounding this as the instrument's correlation with the errors of the model is essentially 0.

Exclusion Restriction is another assumption, stating that the only channel through which living near a college will affect wages is through the effect it has on likelihood to attend higher education. Intuition suggests that the only possible violation of this is that if more prosperous areas have more colleges, then by living near a college a candidate is expected to have a higher wage regardless of if they attended higher education. This idea of colleges not being uniformly distributed across the country will be touched on again. Due to lack of concrete evidence against it, it will be assumed that the exclusion restriction stands.

Figure 1.5 summarises the balance test results. The importance of this is to check if the instrument is as good as randomly assigned. If the assignment i.e. living near a 4yr college was correlated and therefore a predictor for certain attributes, then it would not be a source of endogeneity. The results show that it is only balanced in the factor of race due to the low and insignificant beta estimate of 0.001. The IQ estimate is significantly different from 0 at a 1% level suggesting the instrument is not randomly distributed across this attribute. But due to the relatively low estimate of 2 IQ points this can largely be ignored. What cannot be ignored is the large and highly significant estimate across the region attribute. Living near a 4yr college is associated with a 20% lower chance of living in the south during 1966. This is a problem for the assumption that the treatment is as good as randomly assigned. Though if there are less colleges in the south compared to other districts, even if treatment was randomly assigned, it would be expected that if one lives near a college, they are less likely to be from the south.

The first stage regresses the instrument, near 4yr college on the explanatory variable of interest; *Higher_edu*, the summary of which has been discussed above. This weak first stage, being the denominator in the calculation of the final 2SLS beta estimate, is partially the cause of the overestimation.

The second stage is estimated by regressing the fitted values from the first stage as the explanatory variable as opposed to the actual series. Doing so isolates exogenous variation in the explanatory allowing its true effect on the dependent to be measured. The large beta estimate of 85% is the ratio between the reduced form estimate and the aforementioned first stage.

4) Extension

Splitting the data set on regions lowered the predictive power of the rural model especially, as the much lower sample size results in larger standard errors and fuzzier

estimates. Before comparing the explanatory variable betas, it is important to note that the first stage f-statistic in both cases have fallen well below the threshold for a valid instrument so readers should be sceptical of any conclusions drawn from these models' validity.

Figures 2.1 and 3.1 show betas of -0.347 and 0.482 for rural and metropolitan respectively. The rural estimate being very counter-intuitive suggesting wages decrease with attendance in higher education. The metropolitan estimate is in line with the general model but with a lower, more reasonable marginal effect of 48% increase in expected wages. The relationship from the generalised dataset holds in the case the metropolitan region likely due to this making up 2/3 of the data set. But the standard errors, in the case of the rural, are too large for a conclusion to be made with any degree of confidence.

References

Davis-Kean, P.E. (2005). The Influence of Parent Education and Family Income on Child Achievement: The Indirect Role of Parental Expectations and the Home Environment. *Journal of Family Psychology*, 19(2), pp.294–304.

Nlsinfo.org. (2020). *Geographic Residence & Neighborhood | National Longitudinal Surveys*. [online] Available at: https://www.nlsinfo.org/content/cohorts/nlsy79/topical-guide/household/geographic-residence-neighborhood-composition [Accessed 19 Oct. 2024].

Norris J.N (2024). Topic 1 Instrumental Variable Analysis Lecture SlidesEC413: Applied Econometrics. University of Strathclyde Business School [Accessed: 19 Oct. 2024]

Appendix

Section 1: Full Data Set

N = 1600

Figure 1.1: Summary of initial Model

Coefficients	Estimate	Standard Error	T-value	P-value	Significance Level
(Intercept)	4.795211	0.104927	45.7	< 2.00E-16	***
X	0.08226	0.021625	3.804	0.000148	***
Age	0.048484	0.003307	14.66	< 2.00E-16	***
Fath	0.002989	0.003616	0.826	0.408651	
Moth	0.010258	0.004244	2.417	0.015748	*
Singm	0.033753	0.146752	0.23	0.818122	
Black	-0.152325	0.033234	-4.583	4.93E-06	***

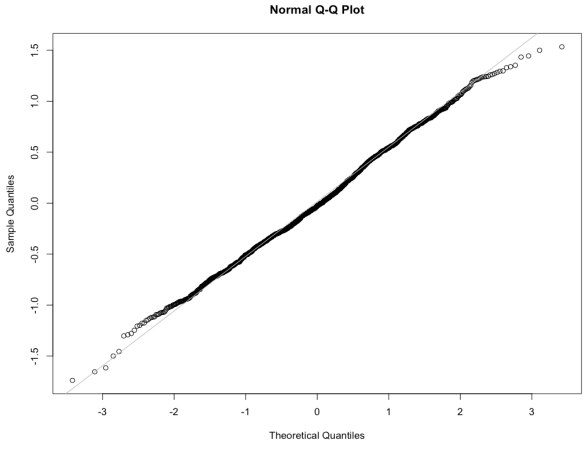
Residuals				
Min	Q1	Median	Q3	Max
-1.6327	-0.22946	0.02071	0.25026	1.20718

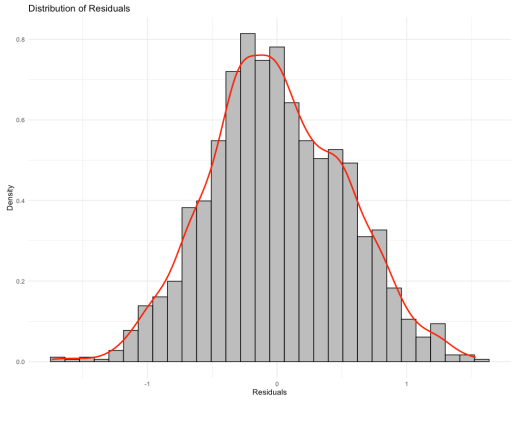
AIC	AIC (after dropping insignificant parameters)
1508.896	1505.636

Figure 1.2: Summary of Two Stage Least Squares Estimates

Coefficients	Estimate	Standard Error	T-value	P-value	Significance Level
(Intercept)	4.737373	0.143716	32.963	< 2.00E-16	***
X	0.85583	0.349254	2.45	1.44E-02	*
Age	0.04942	0.004492	11.002	< 2.00E-16	***
Moth	-0.028475	0.018982	-1.5	1.34E-01	
Black	-0.129173	0.046156	-2.799	0.00519	**

Residuals				
Min	Q1	Median	Q3	Max
-1.73852	-0.34786	-0.02865	0.37556	1.53344





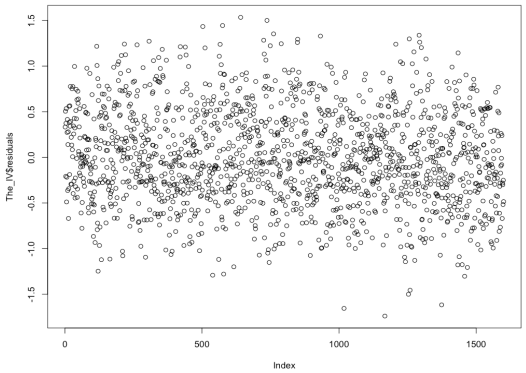


Figure 1.3: Reduced Form

Coefficients	Estimate	Standard Error	T-value	P-value	Significance Level
(Intercept)	4.77995	0.104168	45.887	< 2.00E-16	***
Z	0.070447	0.021262	3.313	9.42E-04	***
Age	0.047956	0.003305	14.511	< 2.00E-16	***
Moth	0.016134	0.003493	4.619	4.17E-06	***
Black	-0.158668	0.032978	-4.811	1.64E-06	***

Residuals				
Min	Q1	Median	Q3	Max
-1.64591	-0.2258	0.01877	0.25132	1.1897

Figure 1.4: The First Stage

Coefficients	Estimate	Standard Error	T-value	P-value	Significance Level
(Intercept)	0.04975	0.123019	0.404	0.68597	
Z	0.082315	0.025109	3.278	0.00107	**
Age	-0.00171	0.003903	-0.438	0.66143	
Moth	0.052123	0.004125	12.635	< 2.00E-16	***
Black	-0.034463	0.038946	-0.885	0.37635	

Residuals				
Min	Q1	Median	Q3	Max
-0.9199	-0.5046	0.2752	0.3442	0.9981

AIC	F-Statistic
10.75	10.75

Instrument Correlation with testing regression Errors

1.28E-15

Figure 1.5: Balance Testing: Z Estimate Results across various Data Attributes

Attribute	Instrument Beta Estimate	Standard Error	T-value	P-value	Significance Level
IQ	2.0654	0.7403	2.79	0.00533	**
Race	0.001718	0.016138	0.106	0.915	
Reigion	-0.2019983	0.0241691	-8.358	< 2.00E-16	***

Section 2: Rural Split

N= 400

Figure2.1: Summary of Two Stage Least Squares Estimates

Coefficients	Estimate	Standard Error	T-value	P-value	Significance Level
(Intercept)	4.739971	0.390594	12.135	< 2.00E-16	***
X	-0.346646	1.612254	-0.215	8.30E-01	
Age	0.039677	0.007658	5.181	3.53E-07	***
Moth	0.050745	0.117298	0.433	6.66E-01	
Black	-0.0778	0.262876	-0.296	0.767	

Residuals				
Min	Q1	Median	Q3	Max
-1.808997	-0.242128	-0.001279	0.243648	1.130227

Figure2.2: The Reduced Form

Coefficients	Estimate	Standard Error	T-value	P-value	Significance Level
(Intercept)	4.815119	0.19849	24.259	< 2.00E-16	***
Z	-0.008774	0.036923	-0.238	0.8123	
Age	0.040293	0.006386	6.31	7.51E-10	***
Moth	0.025514	0.006464	3.947	9.36E-05	***
Black	-0.132934	0.076911	-1.728	8.47E-02	.

Residuals				
Min	Q1	Median	Q3	Max
-1.55185	-0.20327	0.02612	0.23177	1.02941

AIC
339.4541

Figure 2.3: The First Stage

Coefficients	Estimate	Standard Error	T-value	P-value	Significance Level
(Intercept)	-0.216785	0.246613	-0.879	0.3799	
Z	0.025311	0.045875	0.552	0.5814	
Age	-0.001777	0.007934	-0.224	0.8229	
Moth	0.072788	0.008031	9.064	<2e-16	***
Black	0.15905	0.095557	1.664	0.0968	.

Residuals				
Min	Q1	Median	Q3	Max
-0.8306	-0.4551	0.1002	0.386	1.0921

AIC	F-Statistic
513.1204	0.3044

Instrument Correlation with testing regression Errors
7.66E-15

Figure 2.4: The Balance Test

Attribute	Instrument Beta Estimate	Standard Error	T-value	P-value	Significance Level
IQ	2.0654	0.7403	2.79	0.00533	**
Race	0.001718	0.016138	0.106	0.915	
Region	-0.2019983	0.0241691	-8.358	< 2.00E-16	***

Section 3: Metropolitan Split

N= 1200

Figure3.1: Summary of Two Stage Least Squares Estimates

Coefficients	Estimate	Standard Error	T-value	P-value	Significance Level

(Intercept)	4.772142	0.156573	30.479	<2e-16	***
X	0.482233	0.37504	1.286	1.99E-01	
Age	0.050518	0.004229	11.945	<2e-16	***
Moth	-0.010112	0.017384	-0.582	5.61E-01	
Black	-0.145831	0.051361	-2.839	0.0046	**

Figure3.2: The Reduced Form

Coefficients	Estimate	Standard Error	T-value	P-value	Significance Level
(Intercept)	4.858483	0.120267	40.397	< 2.00E-16	***
Z	0.038629	0.027204	1.42	0.15587	
Age	0.049439	0.003778	13.087	< 2.00E-16	***
Moth	0.011216	0.004063	2.76	5.86E-03	**
Black	-0.187453	0.036079	-5.196	2.40E-07	***

Residuals				
Min	Q1	Median	Q3	Max
-1.59406	-0.22189	0.02035	0.25298	1.17569

AIC
1124.445

Figure 3.3: The First Stage

Coefficients	Estimate	Standard Error	T-value	P-value	Significance Level
Z	0.080104	0.032068	2.498	0.0126	*
Age	-0.002236	0.004453	-0.502	0.6156	
Moth	0.044228	0.00479	9.234	<2e-16	***
Black	-0.086311	0.04253	-2.029	0.0426	*

Residuals				
Min	Q1	Median	Q3	Max
-0.9064	-0.5414	0.266	0.3243	0.8836

AIC	F-Statistic
1519.248	6.24

Instrument Correlation with testing regression Errors
1.84E-15

Figure 3.4: The Balance Test

Attribute	Instrument Beta Estimate	Standard Error	T-value	P-value	Significance Level
IQ	1.4576	1.031	1.414	0.158	
Race	-0.0007652	0.0218029	-0.035	0.972	
Region	-0.1986091	0.0311226	-6.38E+00	2.50E-10	***

Section 4: Summary Statistics

Figure 4.1

		Min	Q1	Median	Mean	Q3	Max
	Total	53	95	105	103.9	114	149
IQ	Rural	60	94	103	102.1	112	138
	Metropolitan	53	95	105	104.4	115	149
	Total	0	9	12	11.03	12	18
Mother Education	Rural	1	9	12	10.85	12	18
	Metropolitan	0	10	12	11.09	12	18
	Total	0	8	12	10.61	12	18
Father Education	Rural	0	8	10	9.895	12	18
	Metropolitan	0	8	12	10.85	12	18
	Total	24	26	28	28.18	30	34
Age	Rural	24	26	27	28.03	30	34
	Metropolitan	24	26	28	28.23	30.25	34
	Total	8	12	14	14.13	16	18
Education	Rural	8	12	13	13.75	16	18
	Metropolitan	9	12	14	14.25	16	18
	Total	4.718	6.082	6.369	6.343	6.608	7.785
Logged Wages	Rural	4.852	5.961	6.216	6.209	6.474	7.244
	Metropolitan	4.718	6.136	6.423	6.388	6.645	7.785

Figure 4.2