Problem set 4: Weighted Least Squares and Instrumental Variables Estimators

September 18, 2013

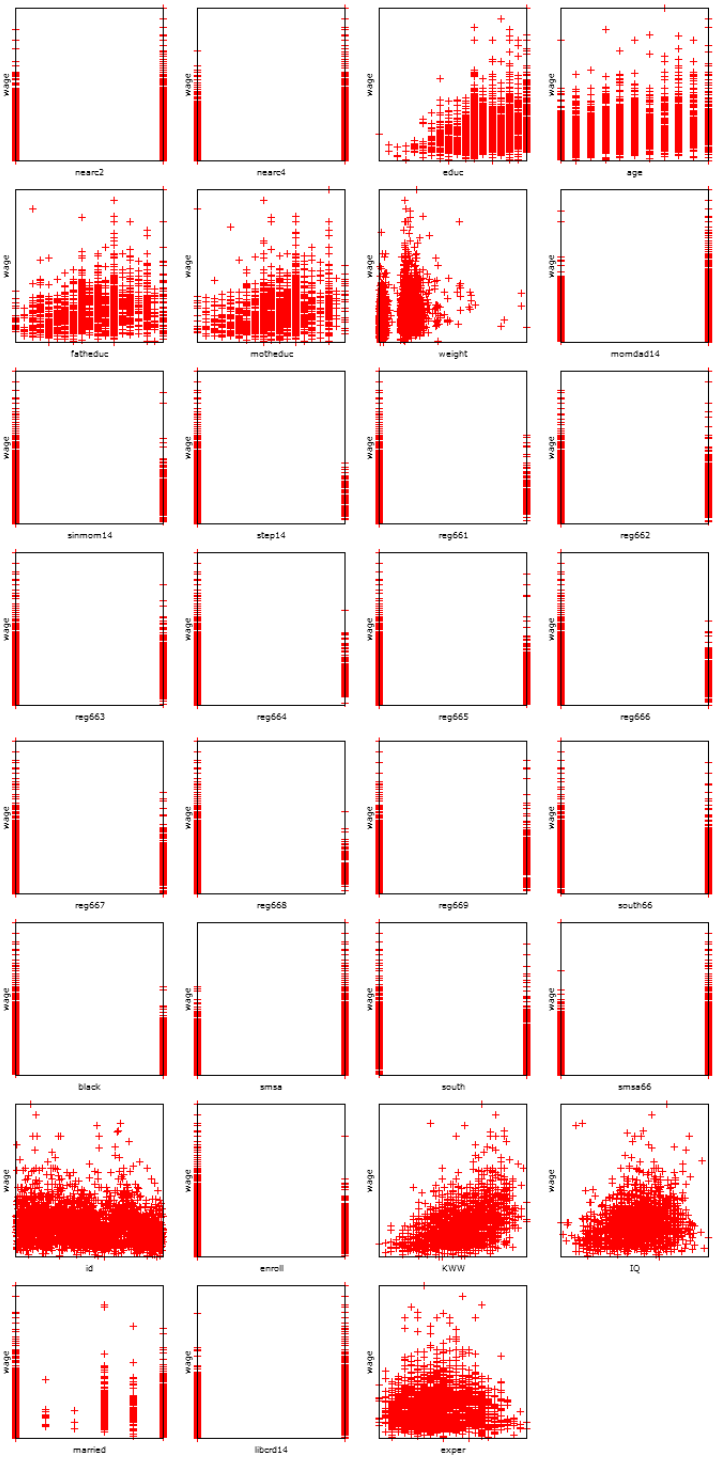
1 Introduction

These answers accompany the Youtube lecture series, and roughly corresponds to videos 125-155, covering Weighted Least Squares and Instrumental Variables analysis.

2 Returns to education - practical

2.1 Weighted Least Squares

1. (a) The scatter graphs from this question are displayed below. In my view the variables which are most correlated with hourly wages are: *educ, fatheduc, motheduc, sinmom14, step14, reg661 (etc. other regions), black, enroll, kww,* and *exper*.



(b) The results of this regression are shown below. The interpretation of the constant here is the wage that would be achieved without any years of education completed.

Model X: OLS, using observations 1–3010
Dependent variable: wage

	Coefficient	Std. Error	*t*-ratio	p-value
const	183.949	23.1039	7.9618	0.0000
educ	29.6554	1.70751	17.3677	0.0000

Mean dependent var	577.2824	S.D. dependent var	262.9583
Sum squared resid	1.89e+08	S.E. of regression	250.7309
R^2	0.091139	Adjusted R^2	0.090837
$F(1, 3008)$	301.6365	P-value(F)	1.83e–64
Log-likelihood	−20898.39	Akaike criterion	41800.78
Schwarz criterion	41812.80	Hannan–Quinn	41805.10

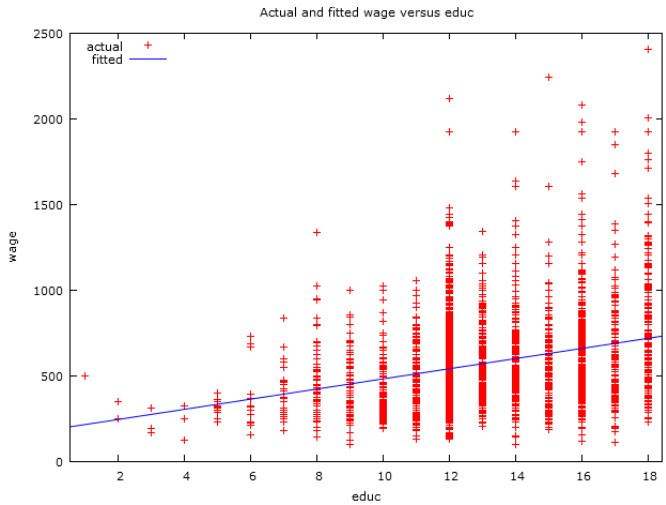
(c) The R-squared is about 9%. This means that a linear model with education only explains a minority of variation in wage away from its mean.

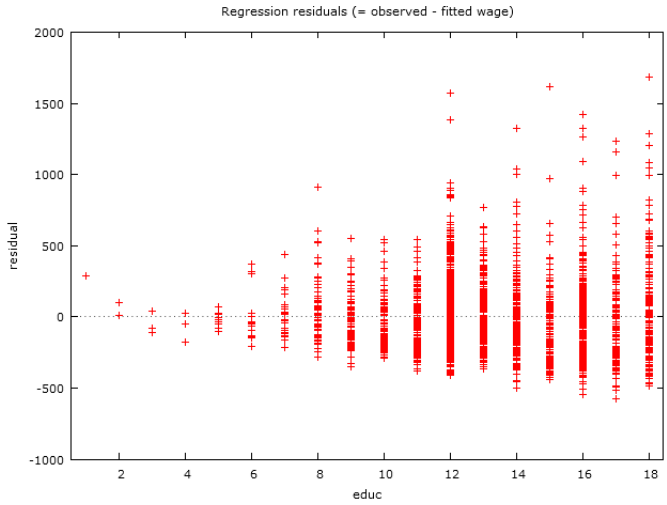
(d) The results of a Breusch-Pagan test are shown below; we reject the null of no heteroscedasticity at the 5% level. It appears that there is heteroscedasticity across the *educ* variable. This is to be expected; when an individual has low levels of education they are limited in choice of career, whereas when they have higher levels, their choice is greater.

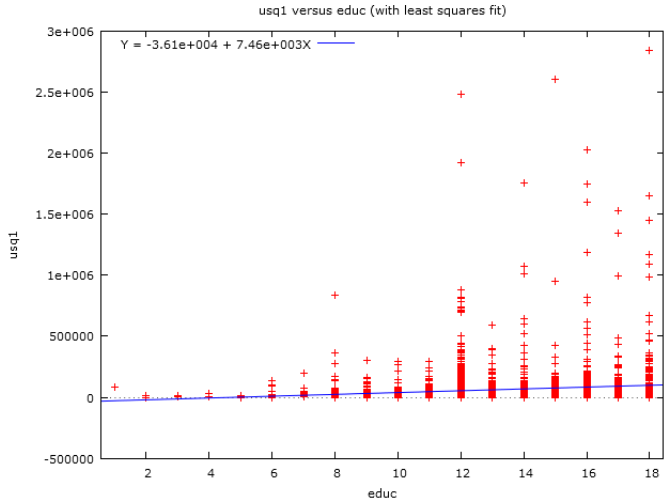
	coefficient	**std. error**	**t-ratio**	**p-value**
const	-0.575	0.231	-2.488	0.0129
educ	0.119	0.017	6.951	4.42E-12

Test statistic:	LM = 152.025
with p-value =	0.00000

(e) There are three different ways (possibly more) of graphically display this heteroscedasticity. From each of the plots it is obvious that there is an increasing error as *educ* increases.







The problems are: OLS estimators are no longer the most efficient, unbiased linear estimators; the standard errors reported by statistical programs (including Gretl) are no longer valid (they are calculated assuming homoscedastic errors). To see an in-depth discussion of these issues go to a Youtube video here: `http://tinyurl.com/oymszed`

(f) Re-running the regression using the log of *wage* on *educ*, we find:

Model X: OLS, using observations 1–3010
Dependent variable: lwage

	Coefficient	Std. Error	*t*-ratio	p-value
const	5.57088	0.0388295	143.4705	0.0000
educ	0.0520942	0.00286971	18.1531	0.0000

Mean dependent var	6.261832	S.D. dependent var	0.443798
Sum squared resid	534.1263	S.E. of regression	0.421389
R^2	0.098737	Adjusted R^2	0.098437
$F(1, 3008)$	329.5368	P-value(F)	5.77e–70
Log-likelihood	−1668.765	Akaike criterion	3341.530
Schwarz criterion	3353.550	Hannan–Quinn	3345.853

The coefficient on *educ*, 0.052, is the % increase in wage from a 1 year increase

in education. So a 1 year increment of *educ* causes on average, a 5.2% increase in wages.

(g) The adjusted R-squares is one potential way to compare these non-nested models. It is slightly higher in the logged model, suggestive that this may be a better specification. In part we prefer this specification since it makes all the coefficients now have the interpretation of, '% increase in *wage* for a 1 unit increase in...'. Also for models where more explanatory variables are added it is likely that a multiplicative model for wages will be a more realistic fit than a purely linear one.

(h) Testing for the presence of heteroscedasticity in the logged model we find that there is no evidence of heteroscedasticity across *educ*. See the results below. So in this sense logging the dependent variable has had the effect of reducing the dispersion sufficiently so as to remove the heteroscedasticity.

	coefficient	**std. error**	**t-ratio**	**p-value**
const	0.854	0.138	6.199	6.45E-10
educ	0.011	0.010	1.079	0.2806
Test statistic:	LM = 1.3024			
with p-value =	0.254			

(i) The regression of *wage* on *educ* using the robust-standard error option in Gretl (can be ticked in the OLS model box - the one where you choose your independent and dependent variables). Note that the coefficient estimates are exactly the same - the *robust standard errors* option only affects the standard errors. Note that these are both higher in this regression than in the original. This is because the standard errors in the original are incorrect - they fail to account for heteroscedasticity.

Model X: OLS, using observations 1–3010
Dependent variable: wage
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	Std. Error	*t*-ratio	p-value
const	183.949	22.5810	8.1462	0.0000
educ	29.6554	1.77588	16.6990	0.0000

Mean dependent var	577.2824	S.D. dependent var	262.9583
Sum squared resid	1.89e+08	S.E. of regression	250.7309
R^2	0.091139	Adjusted R^2	0.090837
$F(1, 3008)$	278.8574	P-value(F)	6.14e–60
Log-likelihood	−20898.39	Akaike criterion	41800.78
Schwarz criterion	41812.80	Hannan–Quinn	41805.10

(j) The results of the fGLS regression are shown below. The coefficient on education is also slightly lower than it was before.

Model X: Heteroskedasticity-corrected, using observations 1–3010

Dependent variable: wage

	Coefficient	Std. Error	t-ratio	p-value
const	189.650	20.0983	9.4361	0.0000
educ	29.2190	1.55911	18.7409	0.0000

Statistics based on the weighted data:

Sum squared resid	12089.39	S.E. of regression	2.004764
R^2	0.104554	Adjusted R^2	0.104257
$F(1, 3008)$	351.2213	P-value(F)	3.30e–74
Log-likelihood	−6363.539	Akaike criterion	12731.08
Schwarz criterion	12743.10	Hannan–Quinn	12735.40

Statistics based on the original data:

Mean dependent var	577.2824	S.D. dependent var	262.9583
Sum squared resid	1.89e+08	S.E. of regression	250.7336

(k) The standard errors on the estimates are slightly lower, than in the OLS case. We should prefer the estimates produced from the fGLS since they should get closer to the true parameter more frequently (since they are more efficient).

2.2 Instrumental variables estimation

(l) There are likely a number of omitted factors which are correlated with *educ*. This means that *educ* is likely endogenous, causing it to be both biased and inconsistent. An example of an omitted factor might be experience. This is likely correlated with education and with wage. (Of course we have that variable, so we could include it, but for our current discussion this is relevant.) Another relevant variable might be some measure of effort/drive.

(m) *Enroll* is definitely correlated with education, but the problem is enrollment is likely correlated with the omitted factors. Enrollment is likely correlated with experience, effort/drive and a host of other omitted factors which influence wage. In that sense it is certainly not exogenous.

(n) The distance an individual lives from college is likely a relatively good instrument. It should have the property that it affects *educ*, because being farther away will make an individual less likely to stay in school. It also is likely uncorrelated with the omitted factors (important in determining wage) in our regression. Perhaps one could argue that individuals/families most interested in schooling would choose to live not far from school, but in my view this is likely a quite weak effect.

(o) We should use the instrument which is most correlated with *educ*. In this case it is the *nearc4* variable. This variable explains about 2% of the variance in *educ*, opposed to *near2* which explains less than 1%. (To see this have a look at the R-squared values from the two regressions below).

Model X: OLS, using observations 1–3010

Dependent variable: educ

	Coefficient	Std. Error	t-ratio	p-value
const	13.1509	0.0651894	201.7341	0.0000
nearc2	0.255258	0.0981804	2.5999	0.0094

Mean dependent var	13.26346	S.D. dependent var	2.676913
Sum squared resid	21513.74	S.E. of regression	2.674355
R^2	0.002242	Adjusted R^2	0.001910
$F(1, 3008)$	6.759438	P-value(F)	0.009371
Log-likelihood	−7230.966	Akaike criterion	14465.93
Schwarz criterion	14477.95	Hannan–Quinn	14470.25

Model 8: OLS, using observations 1–3010
Dependent variable: educ

	Coefficient	Std. Error	t-ratio	p-value
const	12.6980	0.0856416	148.2692	0.0000
nearc4	0.829019	0.103699	7.9945	0.0000

Mean dependent var	13.26346	S.D. dependent var	2.676913
Sum squared resid	21113.48	S.E. of regression	2.649360
R^2	0.020805	Adjusted R^2	0.020480
$F(1, 3008)$	63.91186	P-value(F)	1.84e–15
Log-likelihood	−7202.702	Akaike criterion	14409.40
Schwarz criterion	14421.42	Hannan–Quinn	14413.73

(p) See the above regression (with *nearc4* on *educ*) for the correct first stage regression. I would argue that the instrument is sufficiently strong, since its correlation with *educ* likely significantly outweighs its correlation with omitted variables.

(q) The regression of log *wage* on the estimated *educ* from the first stage, *educhat*, is shown below.

Model X: OLS, using observations 1–3010
Dependent variable: lwage

	Coefficient	Std. Error	t-ratio	p-value
const	3.76747	0.274330	13.7334	0.0000
educhat	0.188063	0.0206744	9.0964	0.0000

Mean dependent var	6.261832	S.D. dependent var	0.443798
Sum squared resid	576.7756	S.E. of regression	0.437890
R^2	0.026772	Adjusted R^2	0.026448
$F(1, 3008)$	82.74453	P-value(F)	1.65e–19
Log-likelihood	−1784.381	Akaike criterion	3572.762
Schwarz criterion	3584.781	Hannan–Quinn	3577.084

The return to 1 incremental year of education from OLS estimates, about a 5% increase in weekly wages, is far smaller than that suggested by instrumental variables estimators, at around 19%!

(r) The replicated results using TSLS are shown below.

Model X: TSLS, using observations 1–3010
Dependent variable: lwage
Instrumented: educ
Instruments: const nearc4

	Coefficient	Std. Error	z	p-value
const	3.76747	0.348862	10.7993	0.0000
educ	0.188063	0.0262913	7.1530	0.0000

Mean dependent var	6.261832	S.D. dependent var	0.443798
Sum squared resid	932.7532	S.E. of regression	0.556858
R^2	0.098737	Adjusted R^2	0.098437
$F(1, 3008)$	51.16576	P-value(F)	1.06e–12
Log-likelihood	−27180.59	Akaike criterion	54365.18
Schwarz criterion	54377.20	Hannan–Quinn	54369.50

Hausman test –
Null hypothesis: OLS estimates are consistent
Asymptotic test statistic: $\chi^2(1) = 48.4992$
with p-value = 3.30423e-012

Weak instrument test –
First-stage $F(1, 3008) = 63.9119$

(s) The null hypothesis under the Hausman test is that the OLS estimates are consistent. The Chi-squared statistic associated with this test was found to be 48.5, with a p value far less than the 5% level required. This is suggestive that OLS estimates are not consistent. To be honest we should have expected this result after the OLS estimates differed significantly from those of instrumental variables estimates.

(t) The errors on the estimated slope coefficient from instrumental variables estimates are roughly ten times those of OLS! This is typical of instrumental variables estimators.

(u) Estimating the model via 2SLS, using both *nearc2* and *nearc4* as instruments for *educ*. The return on *educ* is found to be a little higher, at 20%.

Model X: TSLS, using observations 1–3010
Dependent variable: lwage
Instrumented: educ
Instruments: const nearc4 nearc2

	Coefficient	Std. Error	z	p-value
const	3.63019	0.352717	10.2921	0.0000
educ	0.198413	0.0265814	7.4644	0.0000

Mean dependent var	6.261832	S.D. dependent var	0.443798
Sum squared resid	995.7549	S.E. of regression	0.575357
R^2	0.098737	Adjusted R^2	0.098437
Chi-square(1)	55.71674	p-value	8.37e–14

Hausman test –
Null hypothesis: OLS estimates are consistent
Asymptotic test statistic: $\chi^2(1) = 58.9117$
with p-value = 1.64908e-014

Sargan over-identification test –
Null hypothesis: all instruments are valid
Test statistic: LM = 3.41949
with p-value = $P(\chi^2(1) > 3.41949) = 0.0644308$

Weak instrument test –
First-stage $F(2, 3007) = 33.3943$

(v) The Sargan over-identification test, which has the null hypothesis that all instruments are valid, indicates (albeit mildy) that there may be over-identification here. This may be indicative of the fact that one of the instruments may not be valid. Non-validity here means that it may be correlated with the error term. Also if we actually looked at the first stage regression we can see that *nearc2* is only weakly correlated with *educ*, after taking into account *nearc4*.

Model X: OLS, using observations 1–3010
Dependent variable: educ

	Coefficient	Std. Error	*t*-ratio	p-value
const	12.6396	0.0923765	136.8268	0.0000
nearc2	0.164967	0.0979305	1.6845	0.0922
nearc4	0.808066	0.104411	7.7393	0.0000

Mean dependent var	13.26346	S.D. dependent var	2.676913
Sum squared resid	21093.57	S.E. of regression	2.648551
R^2	0.021728	Adjusted R^2	0.021078
$F(2, 3007)$	33.39427	P-value(F)	4.53e–15
Log-likelihood	−7201.282	Akaike criterion	14408.56
Schwarz criterion	14426.59	Hannan–Quinn	14415.05

It is hard to chose between this last regression and the situation where a single instrumental variable is used. I would probably prefer the situation where only *nearc4* is used, since the extra instrument, *nearc2* is only weakly correlated with *educ* after taking account of the former, and the finite sample bias of a 2SLS estimator is increasing in the number of instruments.

(w) I won't try to improve on the results of the paper from Card (1995), but provide it here for you to compare your results and (more importantly) methodology with. `http://davidcard.berkeley.edu/papers/geo_var_schooling.pdf`

3 Theory

2. (a) There is likely reverse causality acting, whereby low growth perhaps makes a country more susceptible to civil war. Due to this *civil* is endogenous and hence OLS estimators for the effect of civil war on growth will be biased and inconsistent.

(b) The variable *landlocked* cannot be used as an instrument since it is significant in the regression equation, we need another exogenous variable!

(c) *mountains* is likely a reasonably instrument for *civil*; a number of studies have shown that countries with mountainous terrain are more likely to undergo civil wars; also (unless the region is completely overrun with mountainous terrain) I would suggest that this variable is likely not highly correlated with growth.

(d) The first stage regression of the correct form is shown below.

$$civil = \delta_0 + \delta_1 landlocked + \delta_2 mountains + \eta$$

(e) No the researcher should not. Although the results produced via this methodology can be better in a few circumstances, they more often that not lead to problems. TSLS estimators are consistent where least squares are used in both stages, independent of whether the first stage dependent variable is non-continuous, and does not require that the specification be exactly right. Whereas a danger of using a nonlinear first stage such as logit or probit, is that in order for consistent estimates to be produced, the first stage has to be exactly right.

(f) This suggests that the researcher should use this variable *ethnic* as an extra instrument in the first stage of estimation. In my view it is likely that this variable is exogenous from growth, meaning that it is, in principle, a valid instrument.

(g) The null hypothesis from this test is that OLS estimators are consistent. A Chi-squared value of 56.74 is way above the critical value for a chi-squared statistic with two degrees of freedom; suggesting OLS estimates are inconsistent.

3. (a) An OLS estimator will be unbiased, but inefficient because there is a violation of the heteroscedasticity assumption.

(b) A better estimator is a GLS (specifically weighted least squares), since it will be unbiased and more efficient. Actually it will be the most efficient, unbiased, linear estimator. In other words BLUE. The weights applied to both sides of the equation would be $\frac{1}{\sqrt{log(x)\sigma^2}}$, meaning the appropriately estimated equation will take the form:

$$\frac{y}{\sqrt{log(x)\sigma^2}} = \frac{\alpha}{\sqrt{log(x)\sigma^2}} + \beta\frac{x}{\sqrt{log(x)\sigma^2}} + \frac{\varepsilon}{\sqrt{log(x)\sigma^2}}$$

(c) We can prove it is BLUE by appealing to the Gauss-Markov conditions. Since we know that $\mathbb{E}[\varepsilon|x] = 0$, we know similarly that $\mathbb{E}[\frac{\varepsilon}{\sqrt{log(x)\sigma^2}}|\frac{x}{\sqrt{log(x)\sigma^2}}] = 0$. The only thing we need to prove is that the error is homoskedastic. To do this we just need to derive the conditional variance of the error.

$$Var(\frac{\varepsilon}{\sqrt{log(x)\sigma^2}}|\frac{x}{\sqrt{log(x)\sigma^2}}) = \frac{Var(\varepsilon|x)}{log(x)\sigma^2} = \frac{log(x)\sigma^2}{log(x)\sigma^2} = 1$$

Hence we have proved that the Weighted Least Squares estimator is homoscedastic.