

Partialing Out and Effects of Orthogonality in Multiple Regression

Read in the data set *partial.dta* from the course web site.

Consider the simple regression of weekly pay (grosspay) on age

```
reg grsswk age
```

Source	SS	df	MS			
Model	27599238.2	1	27599238.2	Number of obs =	16216	
Residual	1.0325e+09	16214	63680.8603	F(1, 16214) =	433.40	
				Prob > F =	0.0000	
				R-squared =	0.0260	
				Adj R-squared =	0.0260	
Total	1.0601e+09	16215	65379.0136	Root MSE =	252.35	

grsswk	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	3.569728	.1714711	20.82	0.000	3.233625	3.90583
_cons	180.2082	6.887326	26.17	0.000	166.7083	193.7081

Suppose add a variable (random) that is orthogonal to age

[can check orthogonality by looking at the correlation coefficient. Can see correlation between age and random is very low – as it should be since random is a randomly generated variable]

```
. corr grsswk age yrsed random
(obs=16195)
```

	grsswk	age	yrsed	random
grsswk	1.0000			
age	0.1613	1.0000		
yrsed	0.1854	0.7728	1.0000	
random	0.0013	-0.0008	-0.0000	1.0000

The regression estimate of age is virtually unchanged – as it should be since given

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2\beta_2$$

orthogonality $\Rightarrow (X_1'X_2) = 0$ and so the OLS estimate of age in this multiple regression should be the same as the OLS estimate in the simple regression above (which it is, almost)

```
. reg grsswk age random
```

Source	SS	df	MS			
Model	27600144.3	2	13800072.1	Number of obs =	16216	
Residual	1.0325e+09	16213	63684.7321	F(2, 16213) =	216.69	
				Prob > F =	0.0000	
				R-squared =	0.0260	
				Adj R-squared =	0.0259	
Total	1.0601e+09	16215	65379.0136	Root MSE =	252.36	

grsswk	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	3.569739	.1714764	20.82	0.000	3.233626	3.905851
random	.8196194	6.871395	0.12	0.905	-12.64907	14.28831
_cons	179.7979	7.698603	23.35	0.000	164.7078	194.888

Note that a multiple regression involving a variable that is correlated with age **does** change the estimated effect of age, the direction of which is determined by a) the correlation of the new variable with age and the direct effect of the new variable (years of education) on pay

```
. reg grsswk age yrsed
```

Source	SS	df	MS			
Model	37249783.7	2	18624891.8	Number of obs =	16195	
Residual	1.0219e+09	16192	63109.106	F(2, 16192) =	295.12	
				Prob > F =	0.0000	
				R-squared =	0.0352	
Total	1.0591e+09	16194	65401.5332	Adj R-squared =	0.0351	
				Root MSE =	251.22	

grsswk	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.9941416	.2691275	3.69	0.000	.4666221	1.521661
yrsed	1.736265	.140194	12.38	0.000	1.461469	2.011061
_cons	247.6549	8.763275	28.26	0.000	230.4779	264.8318

(so the estimated effect on age in the multiple regressions is **smaller** than in the simple regression because the correlation between age and living in London is **positive**)