

Example of Effects of Omission of Relevant Variables

u psldat

You decide to run a simple regression of log hourly pay on years of work experience

reg lhw exper

Source	SS	df	MS			
Model	52.3263364	1	52.3263364	Number of obs =	17321	
Residual	6113.83639	17319	.353013245	F(1, 17319) =	148.23	
Total	6166.16272	17320	.356014014	Prob > F =	0.0000	
				R-squared =	0.0085	
				Adj R-squared =	0.0084	
				Root MSE =	.59415	

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	.0045364	.0003726	12.17	0.000	.0038061	.0052668
_cons	1.857851	.0092112	201.70	0.000	1.839796	1.875905

and then decide to include a dummy variable for whether the individual is a graduate

reg lhw exper grad

Source	SS	df	MS			
Model	854.145433	2	427.072717	Number of obs =	17321	
Residual	5312.01729	17318	.306733878	F(2, 17318) =	1392.32	
Total	6166.16272	17320	.356014014	Prob > F =	0.0000	
				R-squared =	0.1385	
				Adj R-squared =	0.1384	
				Root MSE =	.55384	

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	.0079872	.0003538	22.57	0.000	.0072936	.0086807
grad	.6144253	.0120174	51.13	0.000	.59087	.6379807
_cons	1.691547	.0091817	184.23	0.000	1.67355	1.709544

The coefficient on experience increases significantly. Why?

The algebra of omitted variables tells us that

$$E(\hat{\mathbf{b}}_{\text{exper}}^{2 \text{ var model}}) = \hat{\mathbf{b}}_{\text{exper}}^{3 \text{ var model}} + (X_1'X_1)^{-1}X_1'X_2 \hat{\mathbf{b}}_{\text{grad}} \quad (1)$$

so that the OLS estimate of experience in the 2 variable model equals the OLS coefficient on experience in the full model plus a correction factor which is equal to the coefficient from a regression of graduate status on experience, $(X_1'X_1)^{-1}X_1'X_2$, multiplied by the OLS coefficient on graduate in the full model

Can test this by regressing graduate on experience

reg grad exper

Source	SS	df	MS			
Model	80.2003462	1	80.2003462	Number of obs =	17321	
Residual	2123.91997	17319	.122635254	F(1, 17319) =	653.97	
				Prob > F =	0.0000	
				R-squared =	0.0364	
				Adj R-squared =	0.0363	
Total	2204.12032	17320	.127258679	Root MSE =	.35019	

grad	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	-.0056162	.0002196	-25.57	0.000	-.0060467	-.0051857
_cons	.2706645	.0054291	49.85	0.000	.2600229	.2813061

and using (1) above to give

```
display 0.0079872+(-.0056162*.61444253)
.00453637
```

which is the coefficient on experience in the 2 variable model.

(1) also explains why the coefficient on experience in the unrestricted model is more positive. a) Experience and graduate status are negatively correlated (there are relatively more graduates among younger workers) - see the regression coefficient on experience in the auxillary regression above b) graduates earn more. The product of these two effects is negative. Not controlling for both these effects exerts a downward bias on experience in the restricted model.

Generalising

reg lhw exper exper2

Source	SS	df	MS			
Model	336.103556	2	168.051778	Number of obs =	17321	
Residual	5830.05917	17318	.336647371	F(2, 17318) =	499.19	
				Prob > F =	0.0000	
				R-squared =	0.0545	
				Adj R-squared =	0.0544	
Total	6166.16272	17320	.356014014	Root MSE =	.58021	

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	.0420931	.0013438	31.32	0.000	.0394592	.044727
exper2	-.0008484	.0000292	-29.03	0.000	-.0009056	-.0007911
_cons	1.567028	.0134628	116.40	0.000	1.54064	1.593417

Now (1) implies that the coefficients in the restricted model equal the coefficients in the full model plus a correction factor where now $(X_1'X_1)^{-1}X_1'X_2$ is a k_1 by k_2 matrix of coefficients equal to the ols coefficients from a regression of each of the X_2 variables on all the X_1 variables.

In this case $k_1 = 2$ (experience, experience²) and $k_2 = 1$

So $(X_1'X_1)^{-1}X_1'X_2$ is 2×1 (one row for each of the coefficients in the b vector in the restricted model ie experience, experience²).

The full regression (including graduate status) is

```
reg lhw exper exper2 grad
```

Source	SS	df	MS			
Model	1149.17149	3	383.057162	Number of obs = 17321		
Residual	5016.99124	17317	.289714803	F(3, 17317) = 1322.19		
				Prob > F = 0.0000		
				R-squared = 0.1864		
				Adj R-squared = 0.1862		
Total	6166.16272	17320	.356014014	Root MSE = .53825		

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	.0463079	.0012491	37.07	0.000	.0438595	.0487563
exper2	-.0008651	.0000271	-31.91	0.000	-.0009182	-.0008119
grad	.6187622	.0116801	52.98	0.000	.595868	.6416563
_cons	1.393823	.01291	107.96	0.000	1.368518	1.419128

Since X_2 only contains 1 variable (graduate), need only consider one regression (of graduate status on the X_1 variables experience and experience²)

```
reg grad exper exper2
```

Source	SS	df	MS			
Model	80.4878852	2	40.2439426	Number of obs = 17321		
Residual	2123.63243	17318	.122625732	F(2, 17318) = 328.19		
				Prob > F = 0.0000		
				R-squared = 0.0365		
				Adj R-squared = 0.0364		
Total	2204.12032	17320	.127258679	Root MSE = .35018		

grad	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	-.0068117	.000811	-8.40	0.000	-.0084013	-.005222
exper2	.000027	.0000176	1.53	0.126	-7.56e-06	.0000616
_cons	.2799218	.0081253	34.45	0.000	.2639954	.2958482

Using (1) the OLS coefficient on experience in the restricted model is

```
display 0.0463079+(-.0068117*.6187622)
```

```
.04209308
```

and the OLS coefficient on experience² in the restricted model is

```
display -0.0008651+(-.000027*.6187622)
```

```
-.00088181
```

marvellous.