

Getting the Model Specification Right

How do we know that the estimated model is the “right” one?
(coefficients unbiased and efficient)

What are the consequences of using the wrong model?

Getting the Model Specification Right

How do we know that the estimated model is the “right” one?
(coefficients unbiased and efficient)

What are the consequences of using the wrong model?

Involves investigating

1) Correct functional form
(use logs or levels, squared terms, inverses etc)

Getting the Model Specification Right

How do we know that the estimated model is the “right” one?
(coefficients unbiased and efficient)

What are the consequences of using the wrong model?

Involves investigating

- 1) Correct functional form
(use logs or levels, squared terms, inverses etc)
- 2) Whether Gauss-Markov assumptions needed for OLS hold

Getting the Model Specification Right

How do we know that the estimated model is the “right” one?
(coefficients unbiased and efficient)

What are the consequences of using the wrong model?

Involves investigating

- 1) Correct functional form
(use logs or levels, squared terms, inverses etc)
- 2) Whether Gauss-Markov assumptions needed for OLS hold
- 3) Choice of variables to include in the model

Omission of (Relevant) Variables

Omission of (Relevant) Variables

Suppose the true model is given by

$$\text{True: } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \quad (1)$$

Omission of (Relevant) Variables

Suppose the true model is given by

$$\text{True: } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \quad (1)$$

But instead you estimate

$$\text{Estimate: } y = \beta_0 + \beta_1 X_1 + v \quad (2)$$

Omission of (Relevant) Variables

Suppose the true model is given by

$$\text{True: } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \quad (1)$$

But instead you estimate

$$\text{Estimate: } y = \beta_0 + \beta_1 X_1 + v \quad (2)$$

(ie imposing the restriction that $\beta_2=0$, X_2 has no explanatory power, when in fact not true)

Omission of (Relevant) Variables

Suppose the true model is given by

$$\text{True: } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \quad (1)$$

But instead you estimate

$$\text{Estimate: } y = \beta_0 + \beta_1 X_1 + v \quad (2)$$

(ie imposing the restriction that $\beta_2=0$, X_2 has no explanatory power, when in fact not true)

We now know that OLS on (2) gives

$$\hat{\beta}_1 = \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)}$$

Omission of (Relevant) Variables

Suppose the true model is given by

$$\text{True: } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \quad (1)$$

But instead you estimate

$$\text{Estimate: } y = \beta_0 + \beta_1 X_1 + v \quad (2)$$

(ie imposing the restriction that $\beta_2=0$, X_2 has no explanatory power, when in fact not true)

We now know that OLS on (2) gives

$$\hat{\beta}_1 = \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)}$$

when the estimate should be (if used the true model)

$$\hat{\beta}_1 = \frac{\text{Cov}(X_1, Y)\text{Var}(X_2) - \text{Cov}(X_2, Y)\text{Cov}(X_1, X_2)}{\text{Var}(X_1)\text{Var}(X_2) - [\text{Cov}(X_1, X_2)]^2}$$

So unless $\text{Cov}(X_1, X_2) = 0$

(and if it is then X_1 and X_2 are said to be “orthogonal” ie included & omitted variables uncorrelated)

So unless $\text{Cov}(X_1, X_2) = 0$

(and if it is then X_1 and X_2 are said to be “orthogonal” ie included & omitted variables uncorrelated)

then the estimates you get from the two models will be different

So unless $\text{Cov}(X_1, X_2) = 0$

(and if it is then X_1 and X_2 are said to be orthogonal ie included & omitted variables uncorrelated)

then the estimates you get from the two models will be different

Does this matter if OLS is supposed to give unbiased efficient estimates?

So unless $\text{Cov}(X_1, X_2) = 0$

(and if it is then X_1 and X_2 are said to be orthogonal ie included & omitted variables uncorrelated)

then the estimates you get from the two models will be different

Does this matter if OLS is supposed to give unbiased efficient estimates?

Estimate is unbiased only if $E(\hat{\beta}_1^{2 \text{ variable}}) = \beta_1$

and since the estimate for the estimated model (2) can always be written as

$$\hat{\beta}_1^{2 \text{ variable}} = \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)}$$

So unless $\text{Cov}(X_1, X_2) = 0$

(and if it is then X_1 and X_2 are said to be orthogonal ie included & omitted variables uncorrelated)

then the estimates you get from the two models will be different

Does this matter if OLS is supposed to give unbiased efficient estimates?

Estimate is unbiased only if $E(\hat{\beta}_1^{2 \text{ variable}}) = \beta_1$

and since the estimate for the estimated model (2) can always be written as

$$\hat{\beta}_1^{2 \text{ variable}} = \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)} \quad \text{sub. in for TRUE } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

So unless $\text{Cov}(X_1, X_2) = 0$

(and if it is then X_1 and X_2 are said to be orthogonal ie included & omitted variables uncorrelated)

then the estimates you get from the two models will be different

Does this matter if OLS is supposed to give unbiased efficient estimates?

Estimate is unbiased only if $E(\hat{\beta}_1^{2 \text{ variable}}) = \beta_1$

and since the estimate for the estimated model (2) can always be written as

$$\hat{\beta}_1^{2 \text{ variable}} = \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)} \quad \text{sub. in for TRUE } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

$$\hat{\beta}_1^{2 \text{ variable}} = \frac{\text{Cov}(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u)}{\text{Var}(X_1)}$$

Using rules on covariances

$$\hat{\beta}_1^{2 \text{ variable}} = \frac{\text{Cov}(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u)}{\text{Var}(X_1)} \quad \text{becomes}$$

Using rules on covariances

$$\hat{\beta}_1^{2 \text{ variable}} = \frac{\text{Cov}(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u)}{\text{Var}(X_1)} \quad \text{becomes}$$

$$\hat{\beta}_1^{2 \text{ var}} = \frac{1}{\text{Var}(X_1)} [\text{Cov}(X_1, \beta_0) + \text{Cov}(X_1, \beta_1 X_1) + \text{Cov}(X_1, \beta_2 X_2) + \text{Cov}(X_1, u)]$$

Using rules on covariances

$$\hat{\beta}_1^{2 \text{ variable}} = \frac{\text{Cov}(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u)}{\text{Var}(X_1)} \quad \text{becomes}$$

$$\hat{\beta}_1^{2 \text{ var}} = \frac{1}{\text{Var}(X_1)} [\text{Cov}(X_1, \beta_0) + \text{Cov}(X_1, \beta_1 X_1) + \text{Cov}(X_1, \beta_2 X_2) + \text{Cov}(X_1, u)]$$

$$\hat{\beta}_1^{2 \text{ var}} = 0 + \beta_1 \frac{\text{Cov}(X_1, X_1)}{\text{Var}(X_1)} + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} + \frac{\text{Cov}(X_1, u)}{\text{Var}(X_1)}$$

Using rules on covariances

$$\hat{\beta}_1^{2 \text{ variable}} = \frac{\text{Cov}(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u)}{\text{Var}(X_1)} \quad \text{becomes}$$

$$\hat{\beta}_1^{2 \text{ var}} = \frac{1}{\text{Var}(X_1)} [\text{Cov}(X_1, \beta_0) + \text{Cov}(X_1, \beta_1 X_1) + \text{Cov}(X_1, \beta_2 X_2) + \text{Cov}(X_1, u)]$$

$$\hat{\beta}_1^{2 \text{ var}} = 0 + \beta_1 \frac{\text{Cov}(X_1, X_1)}{\text{Var}(X_1)} + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} + \frac{\text{Cov}(X_1, u)}{\text{Var}(X_1)}$$

and taking expectations (to get bias)

Using rules on covariances

$$\hat{\beta}_1^{2 \text{ variable}} = \frac{\text{Cov}(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u)}{\text{Var}(X_1)} \quad \text{becomes}$$

$$\hat{\beta}_1^{2 \text{ var}} = \frac{1}{\text{Var}(X_1)} [\text{Cov}(X_1, \beta_0) + \text{Cov}(X_1, \beta_1 X_1) + \text{Cov}(X_1, \beta_2 X_2) + \text{Cov}(X_1, u)]$$

$$\hat{\beta}_1^{2 \text{ var}} = 0 + \beta_1 \frac{\text{Cov}(X_1, X_1)}{\text{Var}(X_1)} + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} + \frac{\text{Cov}(X_1, u)}{\text{Var}(X_1)}$$

and taking expectations (to get bias)

$$E(\hat{\beta}_1^{2 \text{ var}}) = \beta_1 + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} \neq \beta_1$$

Using rules on covariances

$$\hat{\beta}_1^{2 \text{ variable}} = \frac{\text{Cov}(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u)}{\text{Var}(X_1)} \quad \text{becomes}$$

$$\hat{\beta}_1^{2 \text{ var}} = \frac{1}{\text{Var}(X_1)} [\text{Cov}(X_1, \beta_0) + \text{Cov}(X_1, \beta_1 X_1) + \text{Cov}(X_1, \beta_2 X_2) + \text{Cov}(X_1, u)]$$

$$\hat{\beta}_1^{2 \text{ var}} = 0 + \beta_1 \frac{\text{Cov}(X_1, X_1)}{\text{Var}(X_1)} + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} + \frac{\text{Cov}(X_1, u)}{\text{Var}(X_1)}$$

and taking expectations (to get bias)

$$E(\hat{\beta}_1^{2 \text{ var}}) = \beta_1 + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} \neq \beta_1$$

So OLS is biased when omit relevant variables and the sign of bias depends on

Using rules on covariances

$$\hat{\beta}_1^{2 \text{ variable}} = \frac{\text{Cov}(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u)}{\text{Var}(X_1)} \quad \text{becomes}$$

$$\hat{\beta}_1^{2 \text{ var}} = \frac{1}{\text{Var}(X_1)} [\text{Cov}(X_1, \beta_0) + \text{Cov}(X_1, \beta_1 X_1) + \text{Cov}(X_1, \beta_2 X_2) + \text{Cov}(X_1, u)]$$

$$\hat{\beta}_1^{2 \text{ var}} = 0 + \beta_1 \frac{\text{Cov}(X_1, X_1)}{\text{Var}(X_1)} + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} + \frac{\text{Cov}(X_1, u)}{\text{Var}(X_1)}$$

and taking expectations (to get bias)

$$E(\hat{\beta}_1^{2 \text{ var}}) = \beta_1 + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} \neq \beta_1$$

So OLS is biased when omit relevant variables and the sign of bias depends on

a) the covariance between the variables, $\text{Cov}(X_1, X_2)$

Using rules on covariances

$$\hat{\beta}_1^{2 \text{ variable}} = \frac{\text{Cov}(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u)}{\text{Var}(X_1)} \quad \text{becomes}$$

$$\hat{\beta}_1^{2 \text{ var}} = \frac{1}{\text{Var}(X_1)} [\text{Cov}(X_1, \beta_0) + \text{Cov}(X_1, \beta_1 X_1) + \text{Cov}(X_1, \beta_2 X_2) + \text{Cov}(X_1, u)]$$

$$\hat{\beta}_1^{2 \text{ var}} = 0 + \beta_1 \frac{\text{Cov}(X_1, X_1)}{\text{Var}(X_1)} + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} + \frac{\text{Cov}(X_1, u)}{\text{Var}(X_1)}$$

and taking expectations (to get bias)

$$E(\hat{\beta}_1^{2 \text{ var}}) = \beta_1 + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} \neq \beta_1$$

So OLS is biased when omit relevant variables and the sign of bias depends on

- a) the covariance between the variables, $\text{Cov}(X_1, X_2)$
- b) the sign of the effect β_2 of the extra variable, X_2 , on y
(if $\beta_2 = 0$ shouldn't be in model in 1st place)

Given

$$E(\hat{\beta}_1^2 \text{var}) = \beta_1 + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} \neq \beta_1$$

Given

$$E(\hat{\beta}_1^2) = \beta_1 + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} \neq \beta_1$$

If $\beta_2 > 0$ and $\text{Var}(X_1) > 0$

(which it will be since variances are always positive)

Given

$$E(\hat{\beta}_1^{2 \text{ var}}) = \beta_1 + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} \neq \beta_1$$

If $\beta_2 > 0$ and $\text{Var}(X_1) > 0$

(which it will be since variances are always positive)

then $\hat{\beta}_1^{2 \text{ var}} > \beta_1$ if $\text{Cov}(X_1, X_2) > 0$
 $\hat{\beta}_1^{2 \text{ var}} < \beta_1$ if $\text{Cov}(X_1, X_2) < 0$

Given

$$E(\hat{\beta}_1^{2 \text{ var}}) = \beta_1 + \frac{\beta_2 \text{Cov}(X_1, X_2)}{\text{Var}(X_1)} \neq \beta_1$$

If $\beta_2 > 0$ and $\text{Var}(X_1) > 0$

(which it will be since variances are always positive)

then $\hat{\beta}_1^{2 \text{ var}} > \beta_1$ if $\text{Cov}(X_1, X_2) > 0$
 $\hat{\beta}_1^{2 \text{ var}} < \beta_1$ if $\text{Cov}(X_1, X_2) < 0$

Hence can tell **sign** of bias by looking at change in value of estimate in simple and multiple model

Equally important: Not only are estimates biased in presence of omitted variables, but can show that standard errors (hence t, F values etc) are also biased

Equally important: Not only are estimates biased in presence of omitted variables, but can show that standard errors (hence t, F values etc) are also biased

So it would seem that it is important to include as many variables on the right hand side in order to avoid specification error bias

Equally important: Not only are estimates biased in presence of omitted variables, but can show that standard errors (hence t, F values etc) are also biased

So it would seem that it is important to include as many variables on the right hand side in order to avoid specification error bias

But...

Example: the data set wages.dta has information on the log of hourly wages, age and education (among other things)

A simple regression of the log of hourly wages on age gives

```
. reg lhw age
```

Source	SS	df	MS			
Model	.180178687	1	.180178687	Number of obs =	5709	
Residual	1598.49274	5707	.280093348	F(1, 5707) =	0.64	
Total	1598.67291	5708	.280075843	Prob > F =	0.4226	
				R-squared =	0.0001	
				Adj R-squared =	-0.0001	
				Root MSE =	.52924	

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0005209	.0006495	0.80	0.423	-.0007523	.0017941
_cons	1.877335	.0277215	67.72	0.000	1.82299	1.931679

because pay is determined by things other than age there is omitted variable bias in these estimates (both coefficients and standard errors are likely to be wrong)

If now add a dummy variable for graduate status

```
. reg lhw age grad
```

Source	SS	df	MS			
Model	265.349213	2	132.674607	Number of obs =	5709	
Residual	1333.3237	5706	.23367047	F(2, 5706) =	567.79	
Total	1598.67291	5708	.280075843	Prob > F =	0.0000	
				R-squared =	0.1660	
				Adj R-squared =	0.1657	
				Root MSE =	.48339	

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0005209	.0006495	0.80	0.423	-.0007523	.0017941
grad	.0005209	.0006495	0.80	0.423	-.0007523	.0017941
_cons	1.877335	.0277215	67.72	0.000	1.82299	1.931679

age		.0013775	.0005938	2.32	0.020	.0002134	.0025415
grad		.6033545	.0179107	33.69	0.000	.5682427	.6384663
_cons		1.751178	.0255957	68.42	0.000	1.701001	1.801355

Can see, size of estimate doubles and becomes statistically significant

Also since sign on graduate is positive, reasoning above tells us that age and graduate status are *negatively* correlated. Can see this by looking at simple correlation coefficient

```
. corr age grad if e(sample)
(obs=5709)
```

	age	grad
age	1.0000	
grad	-0.0428	1.0000

Inclusion of Irrelevant Variables

Suppose instead that include more variables than are needed

Inclusion of Irrelevant Variables

Suppose instead that include more variables than are needed

True: $y = \beta_0 + \beta_1 X_1 + u$ (1)

Inclusion of Irrelevant Variables

Suppose instead that include more variables than are needed

$$\text{True: } y = \beta_0 + \beta_1 X_1 + u \quad (1)$$

$$\text{Estimate: } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \quad (2)$$

(Irrelevant means that the variable X_2 has no explanatory power)

Inclusion of Irrelevant Variables

Suppose instead that include more variables than are needed

$$\text{True: } y = \beta_0 + \beta_1 X_1 + u \quad (1)$$

$$\text{Estimate: } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \quad (2)$$

(Irrelevant means that the variable X_2 has no explanatory power so model (2) is failing to impose the restriction that $\beta_2=0$,

Inclusion of Irrelevant Variables

Suppose instead that include more variables than are needed

$$\text{True: } y = \beta_0 + \beta_1 X_1 + u \quad (1)$$

$$\text{Estimate: } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \quad (2)$$

(Irrelevant means that the variable X_2 has no explanatory power so model (2) is failing to impose the restriction that $\beta_2=0$,

OLS on (2) gives

$$\hat{\beta}_1 = \frac{\text{Cov}(X_1, Y)\text{Var}(X_2) - \text{Cov}(X_2, Y)\text{Cov}(X_1, X_2)}{\text{Var}(X_1)\text{Var}(X_2) - [\text{Cov}(X_1, X_2)]^2}$$

Inclusion of Irrelevant Variables

Suppose instead that include more variables than are needed

$$\text{True: } y = \beta_0 + \beta_1 X_1 + u \quad (1)$$

$$\text{Estimate: } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \quad (2)$$

(Irrelevant means that the variable X_2 has no explanatory power so model (2) is failing to impose the restriction that $\beta_2=0$,

OLS on (2) gives

$$\hat{\beta}_1 = \frac{\text{Cov}(X_1, Y)\text{Var}(X_2) - \text{Cov}(X_2, Y)\text{Cov}(X_1, X_2)}{\text{Var}(X_1)\text{Var}(X_2) - [\text{Cov}(X_1, X_2)]^2}$$
$$\neq \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)}$$

Inclusion of Irrelevant Variables

Suppose instead that include more variables than are needed

$$\text{True: } y = \beta_0 + \beta_1 X_1 + u \quad (1)$$

$$\text{Estimate: } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \quad (2)$$

(Irrelevant means that the variable X_2 has no explanatory power so model (2) is failing to impose the restriction that $\beta_2=0$,

OLS on (2) gives

$$\hat{\beta}_1 = \frac{\text{Cov}(X_1, Y)\text{Var}(X_2) - \text{Cov}(X_2, Y)\text{Cov}(X_1, X_2)}{\text{Var}(X_1)\text{Var}(X_2) - [\text{Cov}(X_1, X_2)]^2} \neq \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)}$$
$$\neq \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)}$$

However in this case can show OLS estimate of β_1 will *not* be biased

Why?

Why?

Just give an intuitive proof (see Dougherty for formal proof)

Why?

Just give an intuitive proof (see Dougherty for formal proof)

Given $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v$ (2)

Why?

Just give an intuitive proof (see Dougherty for formal proof)

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \quad (2)$$

Since true effect of β_2 is zero and we know OLS gives unbiased estimates of true values

Why?

Just give an intuitive proof (see Dougherty for formal proof)

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \quad (2)$$

Since true effect of β_2 is zero and we know OLS gives unbiased estimates of true values

then would expect, on average, the OLS estimate of β_2 in (2) should also be zero

Why?

Just give an intuitive proof (see Dougherty for formal proof)

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \quad (2)$$

Since true effect of β_2 is zero and we know OLS gives unbiased estimates of true values

then would expect, on average, the OLS estimate of β_2 in (2) should also be zero

If it does not then it is only the result of chance. Its presence in the model does not affect the bias of the other variables

Why?

Just give an intuitive proof (see Dougherty for formal proof)

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \quad (2)$$

Since true effect of β_2 is zero and we know OLS gives unbiased estimates of true values

then would expect, on average, the OLS estimate of β_2 in (2) should also be zero

If it does not then it is only the result of chance. Its presence in the model does not affect the bias of the other variables

but

Why?

Just give an intuitive proof (see Dougherty for formal proof)

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \quad (2)$$

Since true effect of β_2 is zero and we know OLS gives unbiased estimates of true values

then would expect, on average, the OLS estimate of β_2 in (2) should also be zero

If it does not then it is only the result of chance. Its presence in the model does not affect the bias of the other variables)

but

will be inefficient, since in 3 variable model

$$\text{Var}(\hat{\beta}_1) = \frac{s^2}{N * \text{Var}(X)} * \frac{1}{1 - r_{X_1 X_2}^2} \neq \frac{s^2}{N * \text{Var}(X)}$$

Why?

Just give an intuitive proof (see Dougherty for formal proof)

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \quad (2)$$

Since true effect of β_2 is zero and we know OLS gives unbiased estimates of true values

then would expect, on average, the OLS estimate of β_2 in (2) should also be zero

If it does not then it is only the result of chance. Its presence in the model does not affect the bias of the other variables)

but will be inefficient, since in 3 variable model

$$\text{Var}(\hat{\beta}_1) = \frac{s^2}{N * \text{Var}(X)} * \frac{1}{1 - r_{X_1 X_2}^2} \neq \frac{s^2}{N * \text{Var}(X)}$$

so including extra irrelevant variables has a cost in terms of larger standard errors (smaller t, F values) than otherwise (and type II error)

Example

Example

Using the data set `smokes.dta`, suppose we are interested in the association between smoking and pay. A regression of the log of hourly wages on a dummy variable to indicate whether or not someone smokes (`smokes`) and a continuous variable to indicate the number of cigarettes smoked each week (`quant`) and controls for age and gender gives the following

```
. reg lhw age female smokes quant
```

Source	SS	df	MS			
Model	235.697543	4	58.9243857	Number of obs =	10061	
Residual	3478.78205	10056	.345940936	F(4, 10056) =	170.33	
Total	3714.47959	10060	.369232564	Prob > F =	0.0000	
				R-squared =	0.0635	
				Adj R-squared =	0.0631	
				Root MSE =	.58817	

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0051492	.0004697	10.96	0.000	.0042286	.0060698
female	-.2326228	.011761	-19.78	0.000	-.2556766	-.2095689
smokes	-.1190897	.0236027	-5.05	0.000	-.1653558	-.0728236
quant	-.000528	.000208	-2.54	0.011	-.0009356	-.0001203
_cons	6.828719	.021466	318.12	0.000	6.786641	6.870796

Now add an indicator for how many cigarettes are smoked at the weekend – Hard to believe that should have true effect on wages but the effect is

```
. reg lhw age female smokes quant quantw
```

Source	SS	df	MS			
Model	235.758451	5	47.1516902	Number of obs =	10061	
Residual	3478.72114	10055	.345969283	F(5, 10055) =	136.29	
Total	3714.47959	10060	.369232564	Prob > F =	0.0000	
				R-squared =	0.0635	
				Adj R-squared =	0.0630	
				Root MSE =	.58819	

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----	-------	-----------	---	------	----------------------	--

age		.0051537	.0004698	10.97	0.000	.0042328	.0060746
female		-.2325969	.0117616	-19.78	0.000	-.2556521	-.2095418
smokes		-.1214821	.0242826	-5.00	0.000	-.169081	-.0738833
quant		-.0006898	.0004382	-1.57	0.115	-.0015489	.0001692
quantwke		.0011567	.0027568	0.42	0.675	-.0042472	.0065606
_cons		6.82852	.0214721	318.02	0.000	6.78643	6.870609

Can see (compared with original specification) that standard error on smoking dummy in particular is much larger (t value is halved) by inclusion of tar level
Irrelevant variables – particularly if they are strongly correlated with other rhs variables can inflate standard errors, (introduce unnecessary multicollinearity)

```
corr quant quantw
(obs=30049)
-----+-----
      quant |      1.0000
quantwke |      0.9674      1.0000
```

Can see (compared with original specification) that standard error on *quant* variable is much larger (t value now insignificant)

Specification Analysis

OLS is biased when omit relevant variables and the sign of bias depends on

- a) the covariance between the variables, $Cov(X_1, X_2)$
- b) the sign of the effect β_2 of the extra variable, X_2 , on y
(if $\beta_2 = 0$ shouldn't be in model in 1st place)

$$E(\hat{\beta}_1^2 \text{var}) = \beta_1 + \frac{\beta_2 Cov(X_1, X_2)}{Var(X_1)} \neq \beta_1$$

Equally important: Not only are estimates biased in presence of omitted variables, but can show that standard errors (hence t, F values etc) are also biased

Including extra irrelevant variables leaves coefficient estimates unbiased but has a cost in terms of larger standard errors (smaller t, F values) than otherwise (and type II error)

N.B. It may seem confusing that omitting relevant variables can cause bias when we use OLS and including irrelevant variables will not. Isn't OLS always supposed to give unbiased estimates?

N.B. It may seem confusing that omitting relevant variables can cause bias when we use OLS and including irrelevant variables will not. Isn't OLS always supposed to give unbiased estimates?

Remember OLS will only give unbiased estimates if all 4 of the Gauss-Markov assumptions hold (see earlier notes).

N.B. It may seem confusing that omitting relevant variables can cause bias when we use OLS and including irrelevant variables will not. Isn't OLS always supposed to give unbiased estimates?

Remember OLS will only give unbiased estimates if all 4 of the Gauss-Markov assumptions hold (see earlier notes).

The case of omitted variables violates the assumption $\text{Cov}(X,u) = 0$

N.B. It may seem confusing that omitting relevant variables can cause bias when we use OLS and including irrelevant variables will not. Isn't OLS always supposed to give unbiased estimates?

Remember OLS will only give unbiased estimates if all 4 of the Gauss-Markov assumptions hold (see earlier notes).

The case of omitted variables violates the assumption $\text{Cov}(X,u) = 0$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v$$

N.B. It may seem confusing that omitting relevant variables can cause bias when we use OLS and including irrelevant variables will not. Isn't OLS always supposed to give unbiased estimates?

Remember OLS will only give unbiased estimates if all 4 of the Gauss-Markov assumptions hold (see earlier notes).

The case of omitted variables violates the assumption $\text{Cov}(X,u) = 0$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v$$

$$y = \beta_0 + \beta_1 X_1 + (\beta_2 X_2 + v)$$

N.B. It may seem confusing that omitting relevant variables can cause bias when we use OLS and including irrelevant variables will not. Isn't OLS always supposed to give unbiased estimates?

Remember OLS will only give unbiased estimates if all 4 of the Gauss-Markov assumptions hold (see earlier notes).

The case of omitted variables violates the assumption $\text{Cov}(X,u) = 0$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v$$

$$y = \beta_0 + \beta_1 X_1 + (\beta_2 X_2 + v)$$

$$y = \beta_0 + \beta_1 X_1 + u \quad \text{where } u = \beta_2 X_2 + v$$

since the omitted variable(s) X_2 now form part of the residual and if they are correlated with the included variables X_1 then $\text{Cov}(X_1, X_2) = \text{Cov}(X,u) \neq 0$

N.B. It may seem confusing that omitting relevant variables can cause bias when we use OLS and including irrelevant variables will not. Isn't OLS always supposed to give unbiased estimates?

Remember OLS will only give unbiased estimates if all 4 of the Gauss-Markov assumptions hold (see earlier notes).

The case of omitted variables violates the assumption $\text{Cov}(X,u) = 0$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v$$

$$y = \beta_0 + \beta_1 X_1 + (\beta_2 X_2 + v)$$

$$y = \beta_0 + \beta_1 X_1 + u \quad \text{where } u = \beta_2 X_2 + v$$

since the omitted variable(s) X_2 now form part of the residual, if they are correlated with the included variables X_1 then $\text{Cov}(X_1, X_2) = \text{Cov}(X,u) \neq 0$

In contrast in the case of inclusion of irrelevant variables the X_2 are not part of the residual and if all the other Gauss-Markov assumptions hold the OLS will be unbiased (though in this case it will be inefficient because the X_2 variables should not be in the model).

Testing for Functional Form

To test formally whether should have included extra variables (strictly higher order terms of the included variables like squares or cubed values) then do the Ramsey Regression Specification Error Test (RESET)

Testing for Functional Form

To test formally whether should have included extra variables (strictly higher order terms of the included variables like squares or cubed values) then do the Ramsey Regression Specification Error Test (RESET)

Given chosen model

1) Estimate:
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Testing for Functional Form

To test formally whether should have included extra variables (strictly higher order terms of the included variables like squares or cubed values) then do the Ramsey Regression Specification Error Test (RESET)

Given chosen model

1) Estimate: $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

2) save predicted (fitted) values : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$

notice that predicted value is a weighted average of *all* the right hand side variables with weights given by size of coefficients

Testing for Functional Form

To test formally whether should have included extra variables (strictly higher order terms of the included variables like squares or cubed values) then do the Ramsey Regression Specification Error Test (RESET)

Given chosen model

1) Estimate: $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

2) save predicted (fitted) values : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$

predicted value is a weighted average of all the right hand side variables with weights given by size of coefficients

It follows that *higher order* powers of this predicted variable are averages of higher order powers of all the X variables

(eg age^2 , $\text{years of education}^2$)

Idea then is to add these higher powers of the predicted value to the original equation

Idea then is to add these higher powers of the predicted value to the original equation

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + y^{\wedge 2} + y^{\wedge 3} + \dots + y^{\wedge k} + v$$

- rather than add the squares and cubes etc of ALL the X variables

Idea then is to add these higher powers of the predicted value to the original equation

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + y^{\wedge 2} + y^{\wedge 3} + \dots + y^{\wedge k} + v$$

- rather than add the squares and cubes etc of ALL the X variables

How many extra terms is arbitrary – should check robustness of result to variation in number

Idea then is to add these higher powers of the predicted value to the original equation

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + y^{\wedge 2} + y^{\wedge 3} + \dots + y^{\wedge k} + v$$

- rather than add the squares and cubes etc of ALL the X variables

How many extra terms is arbitrary – should check robustness of result to variation in number

4) Use F test for subset of variables to test whether these extra variables

$y^{\wedge 2}$ $y^{\wedge 3}$... $y^{\wedge k}$ are jointly significant

Idea then is to add these higher powers of the predicted value to the original equation

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + y^{\wedge 2} + y^{\wedge 3} + \dots + y^{\wedge k} + v$$

- rather than add the squares and cubes etc of ALL the X variables

How many extra terms is arbitrary – should check robustness of result to variation in number

4) Use F test for subset of variables to test whether these extra variables

$$y^{\wedge 2} \ y^{\wedge 3} \ \dots \ y^{\wedge k} \quad \text{are jointly significant}$$

5) Reject null of **no** functional form mis-specification

$$\text{if estimated } F > F_{\text{critical}}$$

Example Misspecification Testing

The data set `food_2.dta`, contains information on food expenditure and the age of the household head in a sample of British adults. You decide to examine the impact of age on food expenditure

```
. reg food age
```

Source	SS	df	MS			
Model	19877.8218	1	19877.8218	Number of obs =	340	
Residual	599501.744	338	1773.6738	F(1, 338) =	11.21	
				Prob > F =	0.0009	
				R-squared =	0.0321	
				Adj R-squared =	0.0292	
				Root MSE =	42.115	
Total	619379.565	339	1827.07836			

food	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	-.4486191	.1340079	-3.35	0.001	-.7122136	-.1850247
_cons	86.93161	7.253069	11.99	0.000	72.66477	101.1984

Simple regression of food expenditure on age suggests a negative association between age and food expenditure

However a scatter plot suggests that the relationship between age and food expenditure is non-linear – so it may be that this simple specification suffers from omitted variable bias.

To test formally for this use the RESET test

```
. predict fhat /* this is how you get predicted values in Stata */  
(option xb assumed; fitted values)
```

```
. g fhat2=fhat^2          /* now square and cube these fitted values */
. g fhat3=fhat^3
```

```
reg food age fhat2
```

Source	SS	df	MS			
Model	87958.6589	2	43979.3294	Number of obs =	340	
Residual	531420.906	337	1576.91664	F(2, 337) =	27.89	
Total	619379.565	339	1827.07836	Prob > F =	0.0000	
				R-squared =	0.1420	
				Adj R-squared =	0.1369	
				Root MSE =	39.71	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
food						
age	-15.20347	2.249123	-6.76	0.000	-19.62756	-10.77938
fhat2	-.259274	.0394594	-6.57	0.000	-.3368919	-.1816562
_cons	1918.255	278.7967	6.88	0.000	1369.854	2466.655

and test for joint significance of these extra variables (what2 what3) using the F test

$$F = \frac{RSS_{restrict} - RSS_{unrestrict}}{RSS_{unrestrict} / N - K_{unrestrict}} \sim F(J, N - K_{unrestrict})$$

$$= \frac{599502 - 531421}{531421 / 340 - 3} \sim F(1, 340 - 3)$$

$$= 43.2$$

From F tables, critical value at 5% level $F(1, 337) = F(1, \infty) = 3.92$

So estimated $F > F_{critical}$

Still **reject** null that model is correctly specified – so need to look for more variables or change functional form.

N.B. Stata does this test automatically but adds a different number of powers of the predicted values to the regression – how many? Test can be sensitive to number of powers used.

```
. ovtest
```

```
Ramsey RESET test using powers of the fitted values of food
Ho: model has no omitted variables
      F(3, 335) =      15.58
      Prob > F =      0.0000
```

So the Ramsey RESET test suggests that there may be omitted higher order powers of age

Adding the square of age

```
reg food age age2
```

Source	SS	df	MS	Number of obs = 340		
Model	87958.728	2	43979.364	F(2, 337) =	27.89	
Residual	531420.837	337	1576.91643	Prob > F =	0.0000	
-----				R-squared =	0.1420	
Total	619379.565	339	1827.07836	Adj R-squared =	0.1369	
-----				Root MSE =	39.71	
food	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	5.019477	.8417381	5.96	0.000	3.363754	6.6752
age2	-.0521813	.0079416	-6.57	0.000	-.0678025	-.03656
_cons	-41.10639	20.65161	-1.99	0.047	-81.7287	-.4840842

```
ovtest
```

```
Ramsey RESET test using powers of the fitted values of food
Ho: model has no omitted variables
      F(3, 334) =      1.37
      Prob > F =      0.2526
```

which suggests that no more higher order powers are needed (but that age^2 and possibly age^3 are needed in the model). N.B. this test should not really be used to test whether other variables (like gender) should be included in the model.

```
. reg food age age2 female
```

Source	SS	df	MS	Number of obs =	340
Model	105466.689	3	35155.5629	F(3, 336) =	22.98
Residual	513912.877	336	1529.50261	Prob > F =	0.0000
Total	619379.565	339	1827.07836	R-squared =	0.1703
				Adj R-squared =	0.1629
				Root MSE =	39.109

food	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	4.664694	.8355931	5.58	0.000	3.021041 6.308347
age2	-.0487368	.0078872	-6.18	0.000	-.0642514 -.0332222
female	-14.75513	4.361143	-3.38	0.001	-23.33371 -6.176544
_cons	-27.02573	20.76021	-1.30	0.194	-67.86208 13.81062

```
ovtest
```

```
Ramsey RESET test using powers of the fitted values of food
```

```
Ho: model has no omitted variables
```

```
F(3, 333) = 1.22
Prob > F = 0.3019
```

Use and Interpretation of Dummy Variables

A simple regression of the log of hourly wages on age gives

```
. reg lhwage age
```

Source	SS	df	MS			
Model	75.4334757	1	75.4334757	Number of obs =	12098	
Residual	3873.61564	12096	.320239388	F(1, 12096) =	235.55	
				Prob > F =	0.0000	
				R-squared =	0.0191	
				Adj R-squared =	0.0190	
Total	3949.04911	12097	.326448633	Root MSE =	.5659	

lhwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0070548	.0004597	15.348	0.000	.0061538	.0079558
_cons	1.693719	.0186945	90.600	0.000	1.657075	1.730364

Now introduce a male dummy variable (1= male, 0 otherwise) as an **intercept dummy**. This specification says the slope effect (of age) is the same for men and women, but that the intercept (or the **average difference** in pay between men and women) is different

```
. reg lhw age male
```

Source	SS	df	MS			
Model	264.053053	2	132.026526	Number of obs =	12098	
Residual	3684.99606	12095	.304671026	F(2, 12095) =	433.34	
				Prob > F =	0.0000	
				R-squared =	0.0669	
				Adj R-squared =	0.0667	
Total	3949.04911	12097	.326448633	Root MSE =	.55197	

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0066816	.0004486	14.89	0.000	.0058022	.0075609
male	.2498691	.0100423	24.88	0.000	.2301846	.2695537
_cons	1.583852	.0187615	84.42	0.000	1.547077	1.620628

Model is $\text{Ln}W = b_0 + b_1\text{Age} + b_2\text{male}$

so constant, b_0 , measures the intercept of default group (women) with age set to zero and $b_0 + b_2$ is the intercept for men

The model assumes these differences are constant at any age so we can interpret the coefficient as the average difference in earnings between men and women

Hence

average wage difference between men and women
 $= (b_0 - (b_0 + b_2)) = -b_2 = 25\%$ more on average

Note that if we define a dummy variables as female (1= female, 0 otherwise) then

```
. reg lhwage age female
```

Source	SS	df	MS			
Model	264.053053	2	132.026526	Number of obs =	12098	
Residual	3684.99606	12095	.304671026	F(2, 12095) =	433.34	
Total	3949.04911	12097	.326448633	Prob > F =	0.0000	
				R-squared =	0.0669	
				Adj R-squared =	0.0667	
				Root MSE =	.55197	

lhwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0066816	.0004486	14.894	0.000	.0058022	.0075609
female	-.2498691	.0100423	-24.882	0.000	-.2695537	-.2301846
_cons	1.833721	.0190829	96.093	0.000	1.796316	1.871127

The coefficient estimate on the dummy variable is the same but the sign of the effect is reversed (now negative). This is because the reference (default) category in this regression is now men

Model is now $\text{Ln}W = b_0 + b_1\text{Age} + b_2\text{female}$

so constant, b_0 , measures average earnings of default group (men)
and $b_0 + b_2$ is average earnings of women

So now

average wage difference between men and women
 $= (b_0 - (b_0 + b_2)) = b_2 = -25\%$ less on average

Hence it does not matter which way the dummy variable is defined as long as you are clear as to the appropriate reference category.

Now consider an **interaction term** – multiply slope variable (age) by dummy variable.

Model is now $\text{LnW} = b_0 + b_1\text{Age} + b_2\text{Female*Age}$

This means that slope effect is different for the 2 groups

$$\begin{aligned} d\text{LnW}/d\text{Age} &= b_1 \text{ if female}=0 \\ &= b_1 + b_2 \text{ if female}=1 \end{aligned}$$

```
. g femage=female*age          /* command to create interaction term */

. reg lhwage age femage
Source |          SS      df      MS                Number of obs =   12098
-----+-----+-----+-----+-----+-----+-----
Model |   283.289249      2   141.644625          F( 2, 12095) =   467.35
Residual |  3665.75986 12095    .3030806          Prob > F      =    0.0000
-----+-----+-----+-----+-----+-----
Total |  3949.04911 12097    .326448633          R-squared     =    0.0717
                                           Adj R-squared =    0.0716
                                           Root MSE    =    .55053

-----+-----+-----+-----+-----+-----
lhwage |          Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----+-----+-----+-----+-----
   age |   .0096943   .0004584     21.148   0.000     .0087958   .0105929
 femage |  -.006454   .0002465    -26.188   0.000    -.0069371  -.005971
  _cons |   1.715961   .0182066     94.249   0.000     1.680273   1.751649
```

So effect of 1 extra year of age on earnings
 = .0097 if male
 = (.0097 - .0065) if female

Normally would include both an intercept and a slope dummy variable in the same regression to decide whether differences were caused by differences in intercepts (and therefore unconnected with the slope variables) or the slope variables

```
. reg lhwage age female femage
```

Source	SS	df	MS			
Model	283.506857	3	94.5022855	Number of obs =	12098	
Residual	3665.54226	12094	.303087668	F(3, 12094) =	311.80	
Total	3949.04911	12097	.326448633	Prob > F =	0.0000	
				R-squared =	0.0718	
				Adj R-squared =	0.0716	
				Root MSE =	.55053	

lhwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0100393	.0006131	16.376	0.000	.0088376	.011241
female	.0308822	.0364465	0.847	0.397	-.0405588	.1023233
femage	-.0071846	.0008968	-8.012	0.000	-.0089425	-.0054268
_cons	1.701176	.0252186	67.457	0.000	1.651743	1.750608

In this example the average differences in pay between men and women appear to be driven by factors which cause the slopes to differ (ie the rewards to extra years of experience are much lower for women than men)

- Note that this model is equivalent to running separate regressions for men and women – since allowing both intercept and slope to vary

Example of Dummy Variable Trap

Suppose interested in estimating the effect of (5) different qualifications on pay

A regression of the log of hourly earnings on dummy variables for each of the 5 education categories gives the following output

```
. reg lhwage age postgrad grad highint low none
```

Source	SS	df	MS			
Model	932.600688	5	186.520138	Number of obs =	12098	
Residual	3016.44842	12092	.249458189	F(5, 12092) =	747.70	
-----				Prob > F =	0.0000	
-----				R-squared =	0.2362	
-----				Adj R-squared =	0.2358	
Total	3949.04911	12097	.326448633	Root MSE =	.49946	

lhwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.010341	.0004148	24.931	0.000	.009528	.0111541
postgrad	(dropped)					
grad	-.0924185	.0237212	-3.896	0.000	-.1389159	-.045921
highint	-.4011569	.0225955	-17.754	0.000	-.4454478	-.356866
low	-.6723372	.0209313	-32.121	0.000	-.7133659	-.6313086
none	-.9497773	.0242098	-39.231	0.000	-.9972324	-.9023222
_cons	2.110261	.0259174	81.422	0.000	2.059459	2.161064

Since there are 5 possible education categories

(postgrad, graduate, higher intermediate, low and no qualifications)

5 dummy variables exhaust the set of possible categories and the sum of these 5 dummy variables is always one for each observation in the data set.

Observation	constant	postgrad	graduate	higher	low	noquals	Sum
1	1	1	0	0	0	0	1
2	1	0	1	0	0	0	1
3	1	0	0	0	0	1	1

Given the presence of a constant using 5 dummy variables leads to pure multicollinearity, (the sum=1 = value of the constant)

Solution: drop one of the dummy variables. Then sum will no longer equal one for **every** observation in the data set.

Observation	constant	postgrad	graduate	higher	low	Sum of dummies
1	1	1	0	0	0	1
2	1	0	1	0	0	1
3	1	0	0	0	0	0

Doesn't matter which one you drop, though convention says drop the dummy variable corresponding to the most common category. However changing the "default" category does change the coefficients, since all dummy variables are measured relative to this default reference category

Example: Dropping the postgraduate dummy (which Stata did automatically before when faced with the dummy variable trap) just replicates the above results. All the education dummy variables pay effects are measured relative to the missing postgraduate dummy variable (which effectively is now picked up by the constant term)

```
. reg lhw age grad highint low none
```

Source	SS	df	MS			
Model	932.600688	5	186.520138	Number of obs =	12098	
Residual	3016.44842	12092	.249458189	F(5, 12092) =	747.70	
Total	3949.04911	12097	.326448633	Prob > F =	0.0000	
				R-squared =	0.2362	
				Adj R-squared =	0.2358	
				Root MSE =	.49946	

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.010341	.0004148	24.93	0.000	.009528	.0111541
grad	-.0924185	.0237212	-3.90	0.000	-.1389159	-.045921
highint	-.4011569	.0225955	-17.75	0.000	-.4454478	-.356866
low	-.6723372	.0209313	-32.12	0.000	-.7133659	-.6313086
none	-.9497773	.0242098	-39.23	0.000	-.9972324	-.9023222
_cons	2.110261	.0259174	81.42	0.000	2.059459	2.161064

So coefficients on education dummies are all negative since all categories earn less than the default group of postgraduates

However changing the default category to the no qualifications group gives

```
. reg lhw age postgrad grad highint low
```

Source	SS	df	MS			
Model	932.600688	5	186.520138	Number of obs =	12098	
Residual	3016.44842	12092	.249458189	F(5, 12092) =	747.70	
Total	3949.04911	12097	.326448633	Prob > F =	0.0000	
				R-squared =	0.2362	
				Adj R-squared =	0.2358	
				Root MSE =	.49946	

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.010341	.0004148	24.93	0.000	.009528	.0111541
postgrad	.9497773	.0242098	39.23	0.000	.9023222	.9972324
grad	.8573589	.0189204	45.31	0.000	.8202718	.894446
highint	.5486204	.0174109	31.51	0.000	.5144922	.5827486
low	.2774401	.0151439	18.32	0.000	.2477555	.3071246
_cons	1.160484	.0231247	50.18	0.000	1.115156	1.205812

and now the coefficients are all positive (relative to those with no quals.)

Dummy Variables and Policy Analysis

One important use of a regression is to try and evaluate the “treatment effect” of a policy intervention.

Usually this means comparing outcomes for those affected by a policy then “event”),

Eg a law on banning cars in central London – creates a “treatment” group, (eg those who drive in London) and those not, (the “control” group).

In principle one could set up a dummy variable to denote membership of the treatment group (or not) and run the following regression

$$\ln W = a + b * \text{Treatment Dummy} + u \quad (1)$$

Problem: a single period regression of the dependent variable on the “treatment” variable as in (1) will **not** give the desired treatment effect.

This is because there may always have been a different value for the treatment group even before the policy intervention took place. If there are systematic differences between treatment and control groups then a simple comparison of the behaviour of the two will give a biased estimate of the “effect of treatment on the treated” – the coefficient b.

The idea then is to try and purge the regression estimate of all these potential behavioural and environmental differences.

Do this by looking at the **change** in the dependent variable for the two groups, (the “**difference in differences**”) over the period in which the policy intervention took place.

The idea is then to compare the change in Y for the treatment group who experienced the shock (subset t) with the change in Y of the control group who did not, (subset c).

Change for Treatment group

$$[Y_t^2 - Y_t^1] = \text{Effect of Policy} + \text{other influences}$$

Change for control group

$$[Y_c^2 - Y_c^1] = \text{Effect of other influences}$$

So $[Y_t^2 - Y_t^1] - [Y_c^2 - Y_c^1] = \text{Effect of Policy}$

In practice this estimator can be obtained from cross-section data from 2 periods – one observed before a program was implemented and the other in the period after.

$\text{Ln}W_1 = a_1 + b_1 \text{Treatment Dummy Variable}_1$	Period Before
$\text{Ln}W_2 = a_2 + b_2 \text{Treatment Dummy Variable}_2$	Period After

The coefficients b_1 and b_2 give the differential impact of the treatment group on wages in each period. The difference between these two coefficients gives the “difference in difference” estimator – the change in the treatment effect following an intervention.

Note however that there is no standard error associated with this method. This can be obtained by combining (pooling) the data over both years and running the following regression.

$$\text{LnW} = a + a_2 \text{Year}_2 + b_1 \text{Treatment Dummy} + b_2 \text{Year}_2 * \text{Treatment Dummy}$$

Where now a is the average wage of the control group in the base year,
 a_2 , is the average wage of the control group in the second year,
 b_1 gives the difference on wages between treatment and control group in the base year
 b_2 is the “difference in difference” estimator – the additional change in wages for the treatment group relative to the control in the second period.

If $\text{Year}_2=0$ and $\text{Treatment Dummy} = 0$, $\text{LnW} = a$

If $\text{Year}_2=0$ and $\text{Treatment Dummy} = 1$, $\text{LnW} = a + b_1$

If $\text{Year}_2=1$ and $\text{Treatment Dummy} = 0$, $\text{LnW} = a + a_2$

If $\text{Year}_2=1$ and $\text{Treatment Dummy} = 1$, $\text{LnW} = a + a_2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a_2 + b_1 + b_2) - (a + b_1) = a_2 + b_2$$

and the change in wages for the control group is

$$(a + a_2) - (a) = a_2$$

so the “difference in difference” estimator

= Change in wages for treatment – change in wages for control

$$= (a_2 + b_2) - (a_2) = b_2$$

Example: In April 2000 the UK government introduced the Working Families Tax Credit aimed at increasing the income in work relative to out of work for groups of traditionally low paid individuals with children. In addition financial help was also given toward child care.

If successful the scheme could have been expected to increase the hours worked of those who benefited most from the scheme- namely single parents. By comparing hours of worked for this group before and after the change with a suitable control group, it should be possible to obtain a difference in difference estimate of the policy effect.

The following example uses other single childless women as a control group.

```

. tab year, g(y)
  /* set up year dummies. Stata will create two dummy variables
     y1=1 if year=1998, = 0 otherwise
     y2=1 if year=2000, = 0 otherwise */

. g lonepy2=lonep*y2          /* create interaction variable */

. reg hours lonep if year==98

-----+-----
Source |      SS      df      MS                Number of obs =   29026
-----+-----
Model | 1159891.90      1 1159891.90            F( 1, 29024) = 3041.43
Residual | 11068703.6 29024  381.363824            Prob > F      =  0.0000
-----+-----
Total | 12228595.5 29025  421.312507            R-squared     =  0.0949
                                           Adj R-squared =  0.0948
                                           Root MSE     = 19.529

-----+-----
hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
lonep | -13.14152   .2382905   -55.15  0.000   -13.60858   -12.67446
_cons |  27.88671   .1436816   194.09  0.000    27.60509    28.16834

. reg hours lonep if year==2000

-----+-----
Source |      SS      df      MS                Number of obs =   28369
-----+-----
Model |  969891.29      1  969891.29            F( 1, 28367) = 2905.13
Residual | 9470465.62 28367  333.855029            Prob > F      =  0.0000
-----+-----
Total | 10440356.9 28368  368.032886            R-squared     =  0.0929
                                           Adj R-squared =  0.0929
                                           Root MSE     = 18.272

-----+-----
hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
lonep | -12.10205   .2245309   -53.90  0.000   -12.54214   -11.66195
_cons |  26.56678   .1368139   194.18  0.000    26.29861    26.83494

```

The coefficient on lone parents gives the difference in average hours worked between lone parents and the control group for the relevant year.

Comparing the lone parent coefficient across periods, lone parents worked 13 hours less than other single women in 1998 before the policy, ($27.9 - 13.1 = 14.8$ hours for single parents on average) and

12 hours less than other single women immediately after the introduction of WFTC, (26.6-12.1 = 14.5 hours for lone parents in 2000, on average).

So the change (difference in difference)

$$\begin{aligned}
 &= -13.1 - (-12.1) = 1.0 \\
 &= (\text{Hours}^{\text{LonePar}}_{2000} - \text{Hours}^{\text{LonePar}}_{1998}) - (\text{Hours}^{\text{Single}}_{2000} - \text{Hours}^{\text{Single}}_{1998}) \\
 &= (14.5 - 14.8) - (26.6 - 27.9) = -0.3 - (-0.7) = 1.0
 \end{aligned}$$

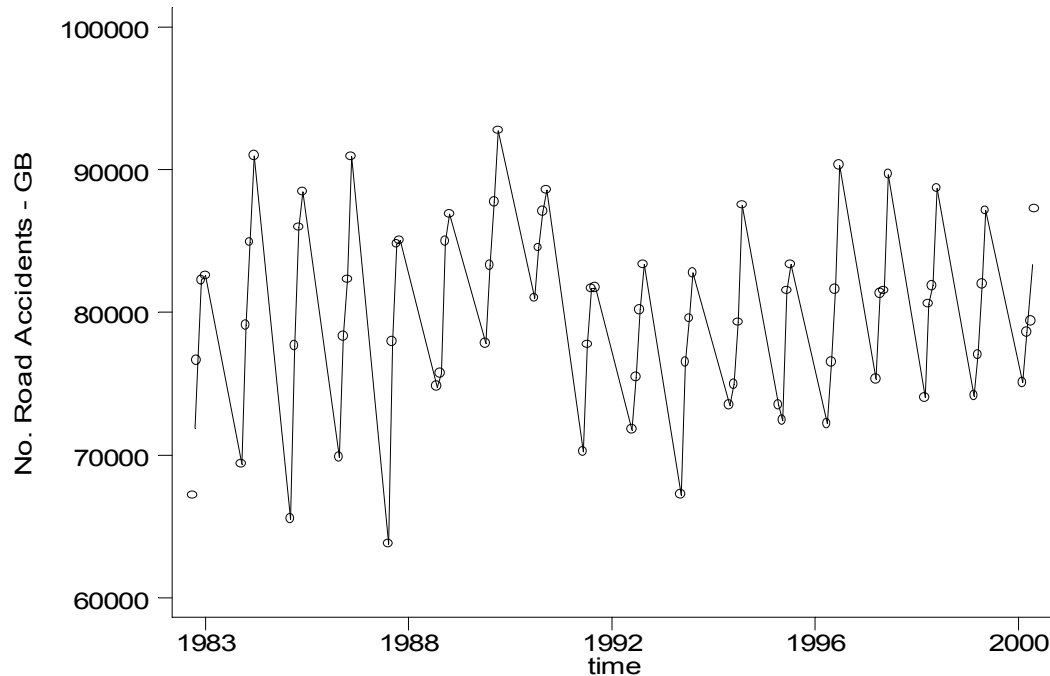
which suggests lone parents worked relatively about 1 hour more as a result of the policy. (Note that hours worked actually fall for both groups, they just fall less for lone parents).

To obtain standard errors, pool the data and estimate the following

```
. reg hours y2 lonep lonepy2
```

Source	SS	df	MS			
Model	2145163.25	3	715054.418	Number of obs = 57395		
Residual	20539169.2	57391	357.881362	F(3, 57391) = 1998.02		
-----				Prob > F = 0.0000		
Total	22684332.5	57394	395.238744	R-squared = 0.0946		
hours	Coef.	Std. Err.	t	P> t	Adj R-squared = 0.0945	
-----				Root MSE = 18.918		
y2	-1.319938	.1985909	-6.65	0.000	[95% Conf. Interval]	
lonep	-13.14152	.2308375	-56.93	0.000	-1.709177	-.9306989
lonepy2	1.039477	.3276099	3.17	0.002	-13.59396	-12.68908
_cons	27.88671	.1391877	200.35	0.000	.3973598	1.681594
					27.6139	28.15952

Using Dummy Variables to capture Seasonality in Data



The data set accidents.dta contains quarterly information on the number of road accidents in the UK from 1983 to 2000

The graph shows that road accidents vary more **within** than **between** years

Can use dummy variables to pick out and control for seasonal variation in data.

Can see seasonal influence from a regression of number of accidents on 3 dummy variables (1 for each quarter minus the default category – which is the 4th quarter)

```
list acc year quart q1 q2 q3          /* list data */
```

	acc	year	quart	q1	q2	q3
1.	67135	1983	Q1	1	0	0
2.	76622	1983	Q2	0	1	0
3.	82277	1983	Q3	0	0	1
4.	82550	1983	Q4	0	0	0
5.	69362	1984	Q1	1	0	0
6.	79124	1984	Q2	0	1	0

```

. reg acc q1 q2 q3

```

Source	SS	df	MS			
Model	2.2572e+09	3	752388623	Number of obs =	72	
Residual	777899883	68	11439704.2	F(3, 68) =	65.77	
Total	3.0351e+09	71	42747405.0	Prob > F =	0.0000	
				R-squared =	0.7437	
				Adj R-squared =	0.7324	
				Root MSE =	3382.3	

acc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
q1	-15080.83	1127.421	-13.38	0.000	-17330.57	-12831.1
q2	-9083.889	1127.421	-8.06	0.000	-11333.62	-6834.155
q3	-4386.278	1127.421	-3.89	0.000	-6636.011	-2136.544
_cons	87088.39	797.2071	109.24	0.000	85497.59	88679.19

Regression of accident numbers on quarterly dummies (q4=winter is default given by constant term at 87088 accidents, on average in the 4th quarter) shows accidents are significantly less likely to happen outside winter

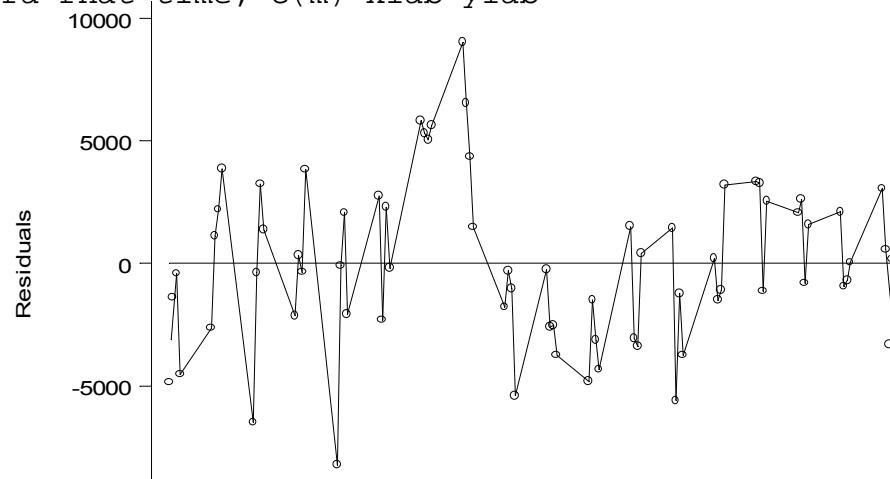
Saving residual values after netting out the influence of the seasons gives **“seasonally adjusted”** accident data (better guide to underlying trend)

Do this with following command after a regression

```

. predict rhat, resid
/* saves the residuals in a new variable with the name "rhat" */
. gra rhat time, c(m) xlab ylab

```



Graph shows that once seasonality accounted for, there is little evidence in a change in the number of road accidents over time.

Can also use seasonal dummy variables to check whether an apparent association between variables is in fact caused by seasonality in the data

```
. reg acc du
```

Source	SS	df	MS			
Model	236050086	1	236050086	Number of obs =	71	
Residual	2.6325e+09	69	38151620.6	F(1, 69) =	6.19	
Total	2.8685e+09	70	40978741.5	Prob > F =	0.0153	
				R-squared =	0.0823	
				Adj R-squared =	0.0690	
				Root MSE =	6176.7	

acc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
du	-4104.777	1650.228	-2.49	0.015	-7396.892	-812.662
_cons	79558.78	768.3058	103.55	0.000	78026.06	81091.51

The regression suggests a negative association between the change in the unemployment rate and the level of accidents

(a 1 percentage point rise in the unemployment rate leads to a fall in the number of accidents by 4104 if this regression is to be believed)

Might this be in part because seasonal movements in both data series are influencing the results (the unemployment rate also varies seasonally, typically higher in q1 of each year)

```
. reg acc du q2-q4
```

Source	SS	df	MS			
Model	2.1275e+09	4	531865433	Number of obs =	71	
Residual	741050172	66	11228032.9	F(4, 66) =	47.37	
Total	2.8685e+09	70	40978741.5	Prob > F =	0.0000	
				R-squared =	0.7417	
				Adj R-squared =	0.7260	
				Root MSE =	3350.8	

acc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
du						
q2						
q3						
q4						
_cons						

du		-1030.818	1009.324	-1.02	0.311	-3045.999	984.3627
q2		5132.594	1266.59	4.05	0.000	2603.766	7661.422
q3		10093.64	1174.291	8.60	0.000	7749.089	12438.18
q4		14353.92	1212.479	11.84	0.000	11933.13	16774.72
_cons		72488.21	834.607	86.85	0.000	70821.87	74154.56

Can see if add quarterly seasonal dummy variables then apparent effect of unemployment disappears.