Lecture 13. Use and Interpretation of Dummy Variables

Stop worrying for 1 lecture and learn to appreciate the uses that "dummy variables" can be put to

Using dummy variables to measure average differences

Using dummy variables when more than 2 discrete categories

Using dummy variables for policy analysis

Using dummy variables to net out seasonality

Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

D = 1 if the criterion is satisfied

Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

D = 1 if the criterion is satisfied
D = 0 if not

Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

D = 1 if the criterion is satisfied
D = 0 if not

Eg. Male/Female

Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

D = 1 if the criterion is satisfied
D = 0 if not

Eg. Male/Female
so that the dummy variable "Male" would be coded
1 if male

Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

D = 1 if the criterion is satisfied
D = 0 if not

Eg. Male/Female
so that the dummy variable "Male" would be coded
1 if male
and 0 if female

Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

D = 1 if the criterion is satisfied
D = 0 if not

Eg. Male/Female
so that the dummy variable “Male” would be coded
1 if male
and 0 if female

(though could equally create another variable “Female” coded 1 if female and 0 if male)

Example: Suppose we are interested in the gender pay gap

Example: Suppose we are interested in the gender pay gap

Model is $LnW = b_0 + b_1 Age + b_2 Male$

where Male = 1 or 0

Example: Suppose we are interested in the gender pay gap

Model is $LnW = b_0 + b_1 Age + b_2 Male$

where Male = 1 or 0

For men therefore the predicted wage

$$\hat{LnW}_{men} = \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2 * (1)$$

Example: Suppose we are interested in the gender pay gap

Model is $\qquad LnW = b_0 + b_1Age + b_2Male$

where Male = 1 or 0

For men therefore the predicted wage

$$\widehat{LnW}_{men} = \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2 * (1)$$

$$= \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2$$

Example: Suppose we are interested in the gender pay gap

Model is $LnW = b_0 + b_1 Age + b_2 Male$

where Male = 1 or 0

For men therefore the predicted wage

$$\widehat{LnW}_{men} = \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2 * (1)$$

$$= \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2$$

For women

$$\widehat{LnW}_{women} = \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2 * (0)$$

Example: Suppose we are interested in the gender pay gap

Model is $LnW = b_0 + b_1 Age + b_2 Male$

where Male = 1 or 0

For men therefore the predicted wage

$$\hat{LnW}_{men} = \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2 * (1)$$

$$= \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2$$

For women

$$\hat{LnW}_{women} = \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2 * (0)$$

$$= \hat{b}_0 + \hat{b}_1 Age$$

Remember that OLS predicts the mean or average value of the dependent variable

$$\overline{\hat{Y}} = \overline{Y}$$

(see lecture 2)

So in the case of a regression model with log wages as the dependent variable, $LnW = b_0 + b_1 Age + b_2 Male$

the average of the fitted values equals the average of log wages

$$\overline{\widehat{Ln(W)}} = \overline{LnW}$$

Remember that OLS predicts the mean or average value of the dependent variable

$$\overline{\hat{Y}} = \overline{Y}$$

(see lecture 2)

Remember that OLS predicts the mean or average value of the dependent variable

$$\overline{\hat{Y}} = \overline{Y}$$

(see lecture 2)

So in the case of a regression model with log wages as the dependent variable, $LnW = b_0 + b_1 Age + b_2 Male$

Remember that OLS predicts the mean or average value of the dependent variable

$$\overline{\hat{Y}} = \overline{Y}$$

(see lecture 2)

So in the case of a regression model with log wages as the dependent variable, $LnW = b_0 + b_1 Age + b_2 Male$

the average of the fitted values equals the average of log wages

$$\overline{\widehat{Ln(W)}} = \overline{LnW}$$

So the (average) difference in pay between men and women is then

$LnW^{men} – LnW^{women}$

So the (average) difference in pay between men and women is then

$$\text{LnW}^{men} - \text{LnW}^{women} = \overline{\widehat{LnW}}_{men} - \overline{\widehat{LnW}}_{women}$$

So the (average) difference in pay between men and women is then

$$\text{LnW}^{men} - \text{LnW}^{women} = \overline{\hat{LnW}}_{men} - \overline{\hat{LnW}}_{women}$$

$$= \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2 - \hat{b}_0 + \hat{b}_1 Age$$

The (average) difference in pay between men and women is then

$$LnW^{men} - LnW^{women} = \overline{\hat{LnW}}_{men} - \overline{\hat{LnW}}_{women}$$

$$= \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2 - \hat{b}_0 + \hat{b}_1 Age$$

$$= \hat{b}_2$$

which is just the coefficient on the male dummy variable

The (average) difference in pay between men and women is then

$$\text{LnW}^{\text{men}} - \text{LnW}^{\text{women}} = \overline{\widehat{LnW}}_{men} - \overline{\widehat{LnW}}_{women}$$

$$= \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2 - \hat{b}_0 + \hat{b}_1 Age$$

$$= \hat{b}_2$$

which is just the coefficient on the male dummy variable

It also follows that the 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 (average) difference in pay between men and women is then

$$LnW^{men} - LnW^{women} = \bar{\hat{LnW}}_{men} - \bar{\hat{LnW}}_{women}$$

$$= \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2 - \hat{b}_0 + \hat{b}_1 Age + \hat{b}_2$$

$$= \hat{b}_2$$

which is just the coefficient on the male dummy variable

So the coefficients on dummy variables measure the average difference

between the group coded with the value “1”

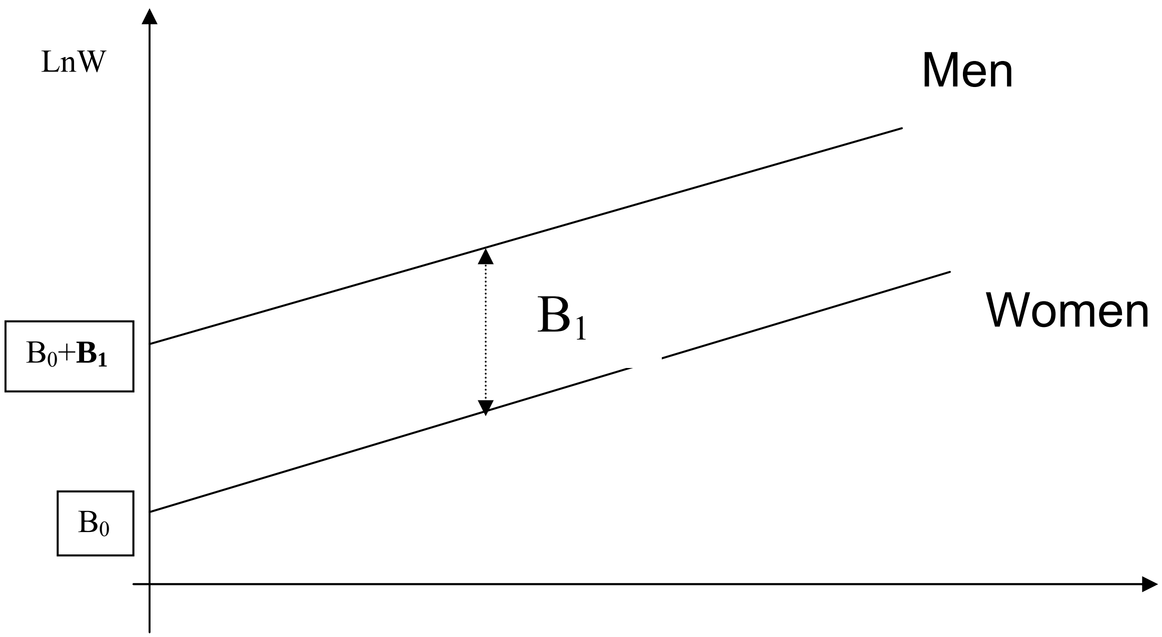
and the group coded with the value “0” (the “default” or “base group”)

It also follows that the constant, b_0, now measures the notional value of the dependent variable (in this case log wages) of the default group (in this case women) with age set to zero

and $b_0 + b_2$ is the intercept and notional value of log wages at age zero for men

So to measure ***average*** difference between two groups

$$LnW = \beta_0 + \beta_1 \text{Group Dummy}$$



A simple regression of the log of hourly wages on age using the data set ps4data.dta gives

```
. reg lhwage age
  Source |       SS       df       MS              Number of obs =   12098
---------+------------------------------           F(  1, 12096) =  235.55
   Model |  75.4334757     1  75.4334757           Prob > F      =  0.0000
Residual |  3873.61564 12096  .320239388           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              Number of obs =   12098
-------------+------------------------------           F(  2, 12095) =  433.34
       Model |  264.053053     2  132.026526           Prob > F      =  0.0000
    Residual |  3684.99606 12095  .304671026           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
```

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              Number of obs =   12098
---------+------------------------------           F(  2, 12095) =  433.34
   Model |  264.053053     2  132.026526           Prob > F      =  0.0000
Residual |  3684.99606 12095  .304671026           R-squared     =  0.0669
---------+------------------------------           Adj R-squared =  0.0667
   Total |  3949.04911 12097  .326448633           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 $LnW = b_0 + b_1 Age + b_2 female$

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

So now

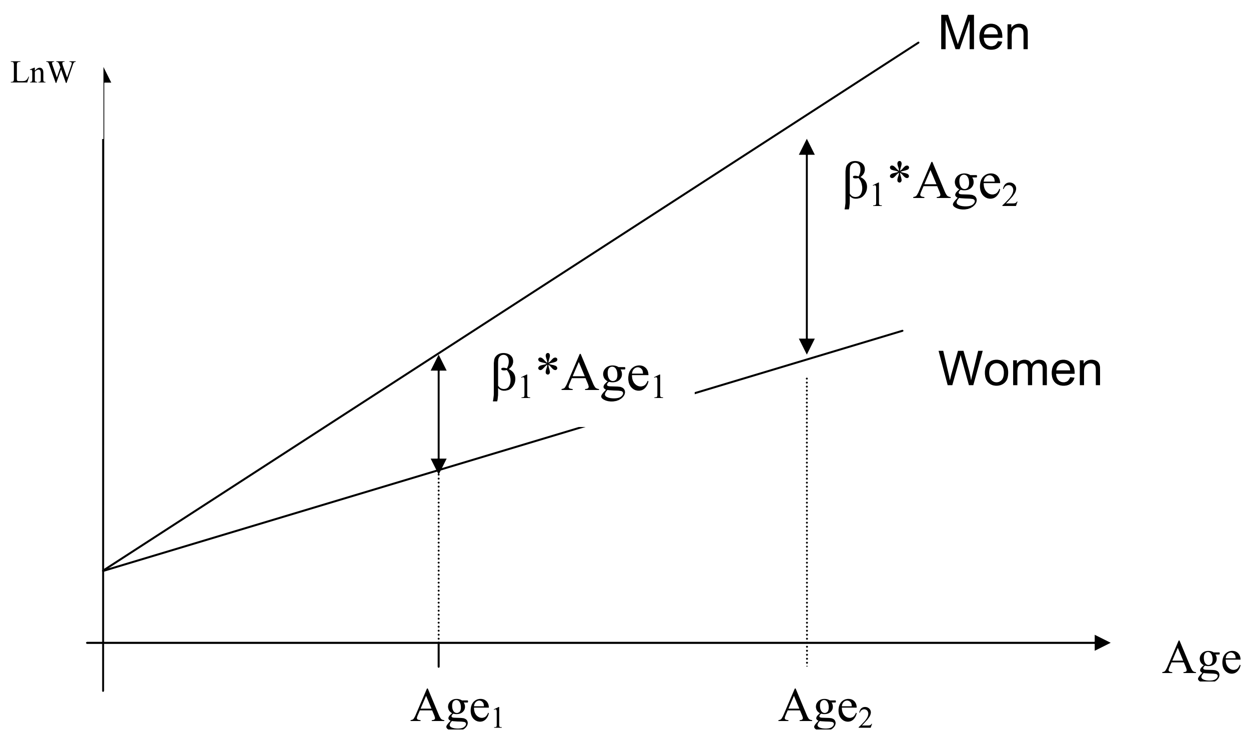
average wage difference between men and women

$=(b_0 – (b_0 + b_2) = b_2 = -25\%$ 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.

2) To measure ***Difference in Slope Effects*** between two groups

LnW = β_0 + β_1Group Dummy*Slope Variable



(Dummy Variable Interaction Term)

We need to consider an **interaction term** – multiply slope variable (age) by dummy variable.

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

Model is now $LnW = b_0 + b_1 Age + b_2 Female*Age$

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

Model is now $LnW = b_0 + b_1Age + b_2Female*Age$

This means that (slope) age effect is different for the 2 groups

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

Model is now $LnW = b_0 + b_1 Age + b_2 Female*Age$

This means that (slope) age effect is different for the 2 groups

$dLnW/dAge = b_1$ if female = 0

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

Model is now $LnW = b_0 + b_1 Age + b_2 Female*Age$

This means that (slope) age effect is different for the 2 groups

$dLnW/dAge = b_1$ if female = 0

$= b_1 + b_2$ if female = 1

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

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

Can include both an intercept and a slope dummy variable in the same regression to decide whether differences were caused by differences in intercepts or the slope variables

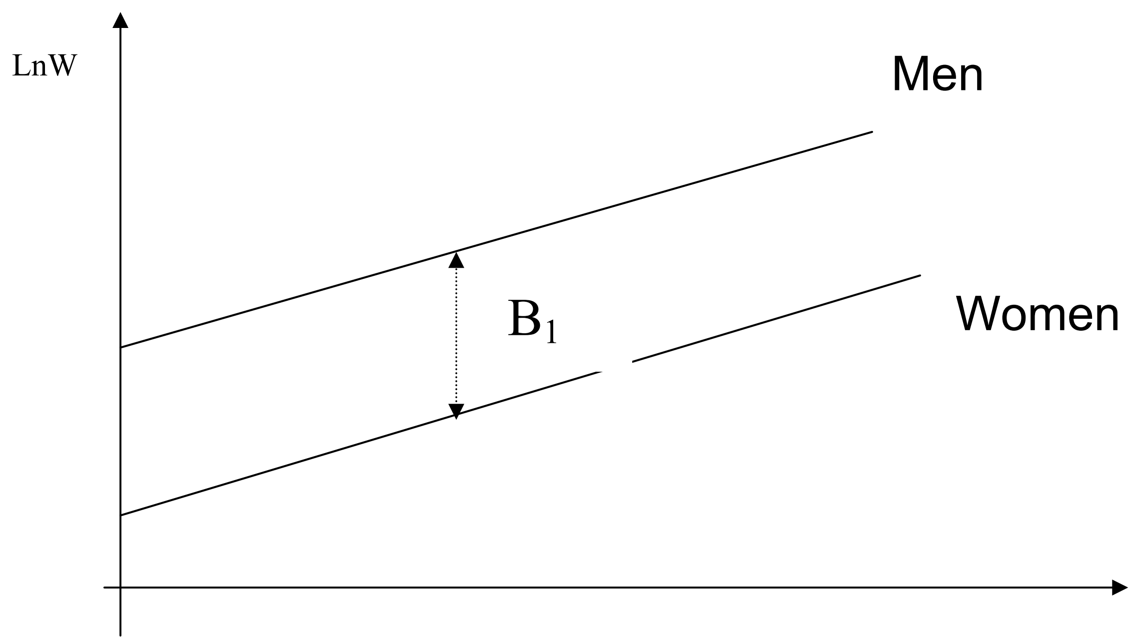
```
. reg lhwage age female femage
  Source |       SS       df       MS                  Number of obs =   12098
---------+------------------------------               F(  3, 12094) =  311.80
   Model |  283.506857     3  94.5022855               Prob > F      =  0.0000
Residual |  3665.54226 12094  .303087668               R-squared     =  0.0718
---------+------------------------------               Adj R-squared =  0.0716
   Total |  3949.04911 12097  .326448633               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

Using & Understanding Dummy Variables

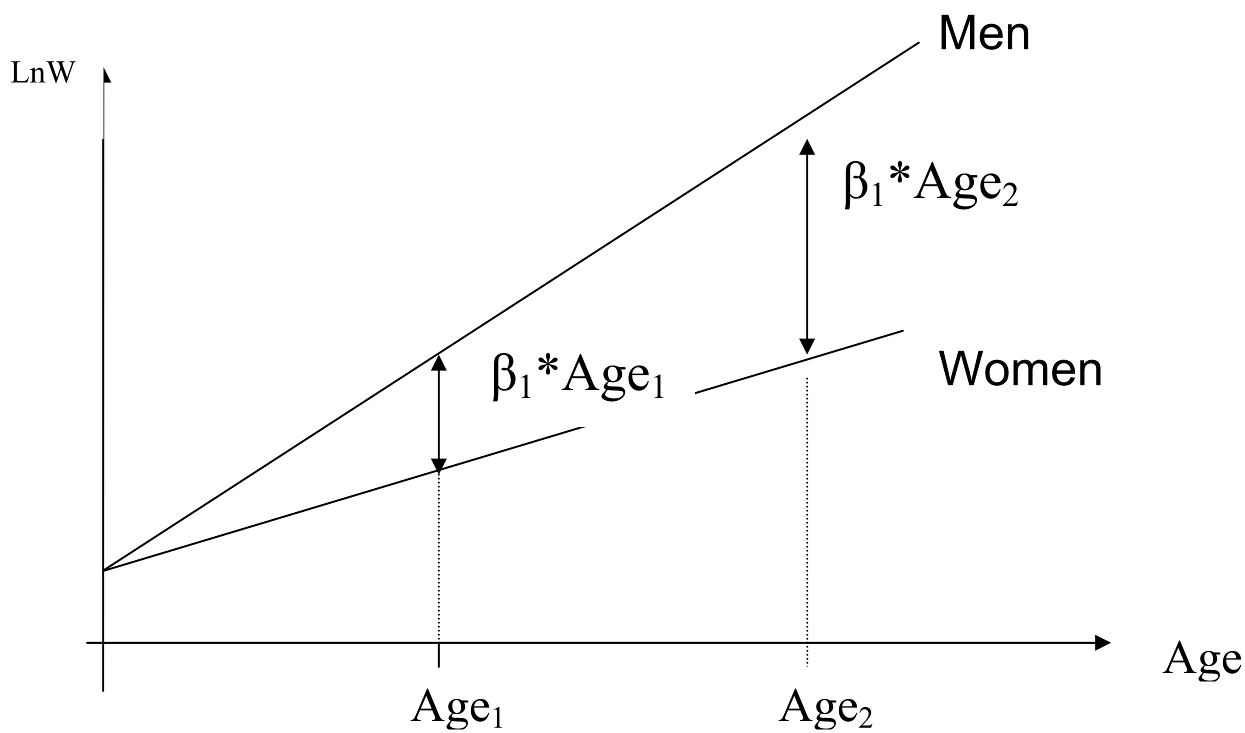
To measure ***average*** difference between two groups

$$LnW = \beta_0 + \beta_1 \text{Group Dummy}$$



2) To measure ***Difference in Slope Effects*** between two groups

LnW = β_0 + β_1Group Dummy*Slope Variable



(Dummy Variable Interaction Term)

The Dummy Variable Trap

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

The Dummy Variable Trap

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

The Dummy Variable Trap

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

Eg if no. groups = 3 (North, Midlands, South)

The Dummy Variable Trap

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

Eg if no. groups = 3 (North, Midlands, South)

Then define

D_{North} = 1 if live in the North, 0 otherwise

The Dummy Variable Trap

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

Eg if no. groups = 3 (North, Midlands, South)

Then define

D_{North} = 1 if live in the North, 0 otherwise
$D_{Midlands}$ = 1 if live in the Midlands, 0 otherwise

The Dummy Variable Trap

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

Eg if no. groups = 3 (North, Midlands, South)

Then define

D_{North} = 1 if live in the North, 0 otherwise
$D_{Midlands}$ = 1 if live in the Midlands, 0 otherwise
D_{South} = 1 if live in the South, 0 otherwise

The Dummy Variable Trap

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

Eg if no. groups = 3 (North, Midlands, South)

Then define

D_{North} = 1 if live in the North, 0 otherwise
$D_{Midlands}$ = 1 if live in the Midlands, 0 otherwise
D_{South} = 1 if live in the South, 0 otherwise

However

The Dummy Variable Trap

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

Eg if no. groups = 3 (North, Midlands, South)

Then define

D_{North} = 1 if live in the North, 0 otherwise
$D_{Midlands}$ = 1 if live in the Midlands, 0 otherwise
D_{South} = 1 if live in the South, 0 otherwise

However

As a rule should always include ***one less*** dummy variable in the model than there are categories, otherwise will introduce multicolinearity into the model

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 5 education categories gives the following output

```
. reg lhwage age postgrad grad highint low none
  Source |       SS       df       MS                  Number of obs =   12098
---------+------------------------------               F(  5, 12092) =  747.70
   Model |  932.600688     5  186.520138               Prob > F      =  0.0000
Residual |  3016.44842 12092  .249458189               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 in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

Since in this example there are 5 possible education categories
(postgrad, graduate, higher intermediate, low and no qualifications)

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

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

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

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

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

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum
1	1	1	0	0	0	0	

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

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

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum
1	1	1	0	0	0	0	1

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

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

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum
1	1	1	0	0	0	0	1
2	1	0	1	0	0	0	

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

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

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum
1	1	1	0	0	0	0	1
2	1	0	1	0	0	0	1

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

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

Obs.	Constant	postgrad	grad	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

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

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

Obs.	Constant	postgrad	grad	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 multicolinearity, becuse the sum=1 which is the same as the value of the constant)

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

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

Obs.	Constant	postgrad	grad	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 multicolinearity, (the sum=1 = value of the constant)

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

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

Obs.	Constant	postgrad	grad	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 multicolinearity, (the sum=1 = value of the constant)

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

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

Obs.	constant	postgrad	grad	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 multicolinearity, (the sum=1 = value of the constant)

So can’t include all 5 dummies and the constant in the same model

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

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

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	higher	low
1	1	1	0	0	0
2	1	0	1	0	0
3	1	0	0	0	0

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

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	higher	low	Sum of dummies
1	1	1	0	0	0	
2	1	0	1	0	0	
3	1	0	0	0	0	

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

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	higher	low	Sum of dummies
1	1	1	0	0	0	1
2	1	0	1	0	0	
3	1	0	0	0	0	

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

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	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	

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

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	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

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

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	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

and so the sum is no longer collinear with the constant

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

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	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

and so the sum is no longer collinear with the constant

Doesn't matter which one you drop, though convention says drop the dummy variable corresponding to the most common category.

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

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	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

and so the sum is no longer collinear with the constant

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              Number of obs =   12098
-------------+------------------------------           F(  5, 12092) =  747.70
       Model |  932.600688     5  186.520138           Prob > F      =  0.0000
    Residual |  3016.44842 12092  .249458189           R-squared     =  0.2362
-------------+------------------------------           Adj R-squared =  0.2358
       Total |  3949.04911 12097  .326448633           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
```

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

Changing the default category to the no qualifications group gives

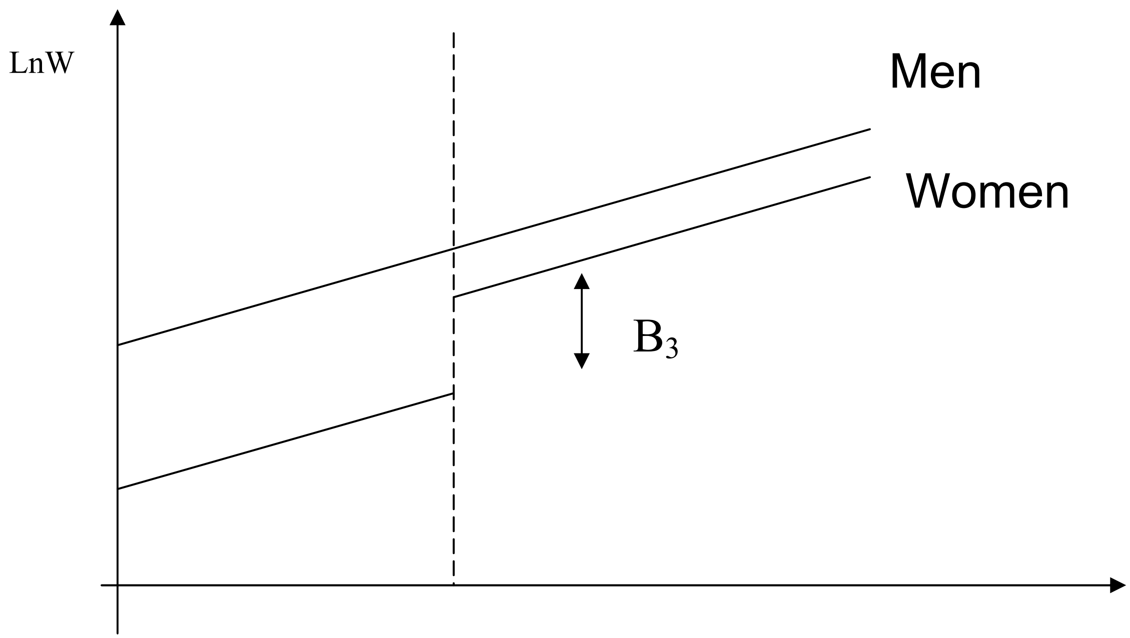
```
. reg lhw age postgrad grad highint low
      Source |       SS       df       MS              Number of obs =   12098
-------------+------------------------------           F(  5, 12092) =  747.70
       Model |  932.600688     5  186.520138           Prob > F      =  0.0000
    Residual |  3016.44842 12092  .249458189           R-squared     =  0.2362
-------------+------------------------------           Adj R-squared =  0.2358
       Total |  3949.04911 12097  .326448633           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 practical use of a regression is to try and evaluate the "treatment effect" of a policy intervention.

3) To Measure Effects of ***Change in the Average Behaviour*** of two groups, one subject to a policy the other not (the Difference-in-Difference Estimator)



Treatment/Policy affects only **a sub-section** of the population
Eg A drug, EMA, Change in Tuition Fees, Minimum Wage

and may lead to a change in behaviour for the treated group - as captured by a change in the intercept (or slope) **after** the intervention (treatment) takes place

Dummy Variables and Policy Analysis

One important practical 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 that is of concern to economists

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

Other examples targeted tax cuts, minimum wages,

area variation in schooling practices, policing

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

$$LnW = a + b*Treatment\ Dummy + u \qquad (1)$$

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

$$LnW = a + b*\text{Treatment Dummy} + u \qquad (1)$$

where Treatment = 1 if exposed to a treatment = 0 if not

```
reg price newham  if time>3 & (newham==1 |  croydon==1)
reg price newham  if time<=3 & (newham==1 |  croydon==1)
reg price newham after  afternew if time>3 & (newham==1 |   croydon==1)
```

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

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

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

ie. Could estimate

$$LnW = a + b*Treatment\ Dummy + u$$

in the period before any treatment took place

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

ie. Could estimate

$$LnW = a + b*Treatment\ Dummy + u$$

in the period before any treatment took place

but what ever the effect observed it cannot be caused by the treatment since these events are observed before the treatment took place.

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

ie. Could estimate

$$LnW = a + b*Treatment\ Dummy + u$$

in the period before any treatment took place

but what ever the effect observed it cannot be caused by the treatment since these events are observed before the treatment took place.

The idea instead is then to compare the **change** in Y for the treatment group who experienced the shock with the change in Y of the control group who did not

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

ie. Could estimate

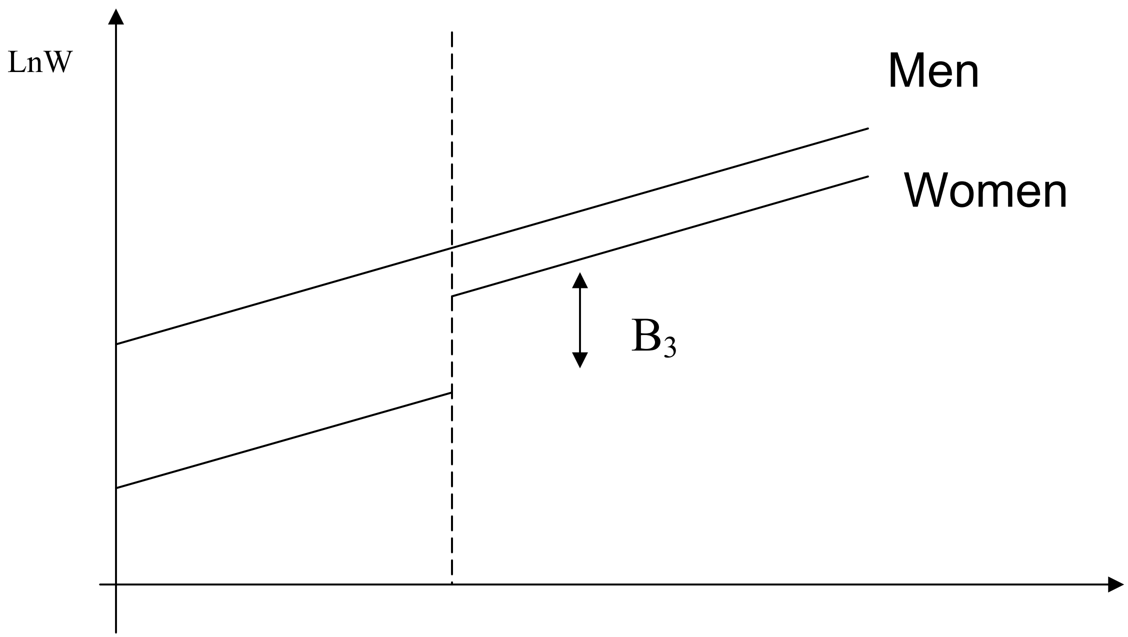
$$LnW = a + b*Treatment\ Dummy + u$$

in the period before any treatment took place

but what ever the effect observed it cannot be caused by the treatment since these events are observed before the treatment took place.

The idea instead is then to compare the **change** in Y for the treatment group who experienced the shock with the change in Y of the control group who did not

- the **"difference in difference estimator"**


LnW
Men
Women
B_3

If the change for Treatment group reflects both

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

If the change for Treatment group reflects both

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

but the change for control group is only caused by

If the change for Treatment group reflects both

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

but the change for control group is only caused by

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

If the change for Treatment group reflects both

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

but the change for control group is only caused by

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

then the difference in differences

If the change for Treatment group reflects both

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

but the change for control group is only caused by

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

then the difference in differences

$$[Y_t^2 - Y_t^1] - [Y_c^2 - Y_c^1] =$$

$$\text{Effect of Policy} + \text{other influences}$$

If the change for Treatment group reflects both

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

but the change for control group is only caused by

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

then the difference in differences

$$[Y_t^2 - Y_t^1] - [Y_c^2 - Y_c^1] =$$

$$\text{Effect of Policy} + \text{other influences} - \text{Effect of other influences}$$

If the change for Treatment group reflects both

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

but the change for control group is only caused by

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

then the difference in differences

$$[Y_t^2 – Y_t^1] - [Y_c^2 – Y_c^1] =$$

$$\text{Effect of Policy} + \text{other influences} - \text{Effect of other influences}$$

$$= \text{Effect of Policy}$$

If the change for Treatment group reflects both

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

but the change for control group is only caused by

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

then the difference in differences

$$[Y_t^2 - Y_t^1] - [Y_c^2 - Y_c^1] =$$

$$\text{Effect of Policy} + \text{other influences} - \text{Effect of other influences}$$

$$= \text{Effect of Policy}$$

(assuming the effect of other influences is the same for both treatment and control groups)

If the change for Treatment group reflects both

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

but the change for control group is only caused by

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

then the difference in differences

$$[Y_t^2 - Y_t^1] - [Y_c^2 - Y_c^1] =$$

$$\text{Effect of Policy} + \text{other influences} - \text{Effect of other influences}$$

$$= \text{Effect of Policy}$$

(assuming the effect of other influences is the same for both treatment and control groups)

Hence the need to try and choose a control group that is similar to the treatment group (apart from the experience of the treatment)

In practice this estimator can be obtained by combining (pooling) the data over the periods before and after and running the following regression

$$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After{*}Treatment\ Dummy$$

In practice this estimator can be obtained by combining (pooling) the data over the periods before and after and running the following regression

$$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After{*}Treatment\ Dummy$$

where now

After is a dummy variable

= 1 if data observed after the treatment

= 0 if data observed before the treatment

In practice this estimator can be obtained by combining (pooling) the data over the periods before and after and running the following regression

$LnW = a + a_2After + b_1Treatment\ Dummy + b_2After*Treatment\ Dummy$

where now

After is a dummy variable
= 1 if data observed after the treatment
= 0 if data observed before the treatment

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 change in wages for the treatment group relative to the control in the second period.

Why ?
$LnW = a + a_2After + b_1Treatment\ Dummy + b_2After*Treatment\ Dummy$

$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After{*}Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After{*}Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2{*}0 + b_1{*}0 + b_2{*}0$$

$$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After*Treatment\ Dummy$$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$
$$= a$$

$$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After*Treatment\ Dummy$$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$
$$= a$$

Similarly

If After =0 and Treatment Dummy = 1, $LnW = a + b_1$

$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After*Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$
$$= a$$

Similarly

If After =0 and Treatment Dummy = 1, $LnW = a + b_1$

If After =1 and Treatment Dummy = 0, $LnW = a + a_2$

$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After*Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$
$$= a$$

Similarly

If After =0 and Treatment Dummy = 1, $LnW = a + b_1$

If After =1 and Treatment Dummy = 0, $LnW = a + a_2$

If After =0 and Treatment Dummy = 1, $LnW = a + a2 + b_1 + b_2$

$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After*Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$
$$= a$$

Similarly

If After =0 and Treatment Dummy = 1, $LnW = a + b_1$
If After =1 and Treatment Dummy = 0, $LnW = a + a_2$
If After =0 and Treatment Dummy = 1, $LnW = a + a2 + b_1 + b_2$

So the change in wages for the treatment group is

$LnW = a + a_2After + b_1Treatment\ Dummy + b_2After*Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$
$$= a$$

Similarly

If After =0 and Treatment Dummy = 1, $LnW = a + b_1$

If After =1 and Treatment Dummy = 0, $LnW = a + a_2$

If After =0 and Treatment Dummy = 1, $LnW = a + a2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a2 + b_1 + b_2)$$

$LnW = a + a_2After + b_1Treatment\ Dummy + b_2After*Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$
$$= a$$

Similarly

If After =0 and Treatment Dummy = 1, $LnW = a + b_1$

If After =1 and Treatment Dummy = 0, $LnW = a + a_2$

If After =0 and Treatment Dummy = 1, $LnW = a + a2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a2 + b_1 + b_2) – (a + b_1)$$

$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After*Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$
$$= a$$

Similarly

If After =0 and Treatment Dummy = 1, $LnW = a + b_1$
If After =1 and Treatment Dummy = 0, $LnW = a + a_2$
If After =0 and Treatment Dummy = 1, $LnW = a + a2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a2 + b_1 + b_2) – (a + b_1) = a_2 + b_2$$

$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After*Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$
$$= a$$

Similarly

If After =0 and Treatment Dummy = 1, $LnW = a + b_1$
If After =1 and Treatment Dummy = 0, $LnW = a + a_2$
If After =0 and Treatment Dummy = 1, $LnW = a + a2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a2 + b_1 + b_2) – (a + b_1) = a_2 + b_2$$

and the change in wages for the control group is

$$(a + a2\ ) – (a\ )$$

$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After*Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$
$$= a$$

Similarly

If After =0 and Treatment Dummy = 1, $LnW = a + b_1$
If After =1 and Treatment Dummy = 0, $LnW = a + a_2$
If After =0 and Treatment Dummy = 1, $LnW = a + a2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a2 + b_1 + b_2) – (a + b_1) = a_2 + b_2$$

and the change in wages for the control group is

$$(a + a2) – (a) = a_2$$

$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After*Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$

$= a$

Similarly

If After =0 and Treatment Dummy = 1, LnW = $a + b_1$

If After =1 and Treatment Dummy = 0, LnW = $a + a_2$

If After =0 and Treatment Dummy = 1, LnW = $a + a2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a2 + b_1 + b_2) – (a + b_1) = a_2 + b_2$$

and the change in wages for the control group is

$$(a + a2) – (a) = a_2$$

so the "difference in difference" estimator

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

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

$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After*Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$
$$= a$$

Similarly

If After =0 and Treatment Dummy = 1, $LnW = a + b_1$

If After =1 and Treatment Dummy = 0, $LnW = a + a_2$

If After =0 and Treatment Dummy = 1, $LnW = a + a2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a2 + b_1 + b_2) – (a + b_1) = a_2 + b_2$$

and the change in wages for the control group is

$$(a + a2) – (a) = a_2$$

so the "difference in difference" estimator

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

$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After*Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$
$$= a$$

Similarly

If After =0 and Treatment Dummy = 1, $LnW = a + b_1$
If After =1 and Treatment Dummy = 0, $LnW = a + a_2$
If After =0 and Treatment Dummy = 1, $LnW = a + a2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a2 + b_1 + b_2) – (a + b_1) = a_2 + b_2$$

and the change in wages for the control group is

$$(a + a2\ ) – (a\ ) = a_2$$

so the “difference in difference” estimator
= Change in wages for treatment – change in wages for control

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

$LnW = a + a_2 After + b_1 Treatment\ Dummy + b_2 After*Treatment\ Dummy$

If After=0 and Treatment Dummy = 0

$$LnW = a + a_2*0 + b_1*0 + b_2*0$$
$$= a$$

Similarly

If After =0 and Treatment Dummy = 1, $LnW = a + b_1$
If After =1 and Treatment Dummy = 0, $LnW = a + a_2$
If After =0 and Treatment Dummy = 1, $LnW = a + a2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a2 + b_1 + b_2) – (a + b_1) = a_2 + b_2$$

and the change in wages for the control group is

$$(a + a2) – (a) = a_2$$

so the "difference in difference" estimator
= Change in wages for treatment – change in wages for control

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

Halifax press release

House prices in East End rise after Olympic win

Friday 2nd February 2007

With 2,000 days to go to the start of the 2012 Olympics, new research from Halifax Estate Agents shows that house prices in three London postal districts close to the site of 2012 Olympics games have risen by more than 15%, or at least £35,000, since London's winning bid was announced.

Across London, house prices have risen by 15% since Q2 2005.

The best performance has been in Leytonstone (E11), which saw a 23% (£50,714) increase in its average house price since mid 2005 followed by Hackney (E8) with a 21% (£48,578) increase and Clapton (E5) 18% (£38,895) rise.

Seven areas close to the Olympic site recorded house price increases of more than 10% since Q2 2006, while all areas close to the Games site have seen at least a £15,000 rise in their average house price. (Table 1)

Stratford (E15), the focal point for Olympic construction activity, saw an 8% (£16,801) increase in its average price since June 2005 to £225,652.

Previous Olympic host cities have seen strengthening house prices

Each of the previous four host cities (Barcelona, Atlanta, Sydney, Athens) have seen house prices rise by more than the national average over the five year period in the run-up to the Olympic games, the main period of Olympic related development activity. These host cities averaged house price increases of 66% over five years against an average rise in host nation house prices of 47% - a differential of 19 percentage points. (Table 2)

Key Findings

London (See Table 1)

- **The postal district near the Olympic site with the highest average house price is Leytonstone (E11)** - £275,827. The next most expensive Olympic areas are Hackney (E8) £274,948 and Clapton (E5) £258,394

Example In July 2005, London won the rights to host the 2012 Olympic games. Shortly afterward there were media reports that house prices were rising "fast" in areas close to the Olympic site. Can evaluate whether this was true by using Newham as the Treatment area (the borough in which the Olympic site is located) and a similar London borough further away from the site as a "control".

The data set *olympics.dta* has monthly data on house prices over time in Newham & Hounslow. The dummy variable "newham" takes the value 1 if the house price observation is from Newham and 0 if not. The dummy variable "after" takes the value 1 if the month was after the Olympic announcement and 0 otherwise. The interaction term "afternew"

g afternew=after*newham

takes the value 1 only if the month is after the event and the observation is in Newham. The coefficient on this term will be the difference-in-difference estimator (the differential effect of the Olympic bid on house prices in newham relative to the control area of croydon.

```
. reg price after if newham==1
      Source |       SS       df       MS              Number of obs =      81
-------------+------------------------------           F(  1,    79) =   38.09
       Model |  3.8272e+10     1  3.8272e+10           Prob > F      =  0.0000
    Residual |  7.9385e+10    79  1.0049e+09           R-squared     =  0.3253
-------------+------------------------------           Adj R-squared =  0.3167
       Total |  1.1766e+11    80  1.4707e+09           Root MSE      =   31700
------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       after |   53378.93   8649.343     6.17   0.000     36162.84    70595.02
       _cons |     165035   3962.462    41.65   0.000     157147.9      172922
```

House prices were indeed higher in Newham after the Olympic announcement, but…

```
. reg price after if hounslow==1
      Source |       SS       df       MS              Number of obs =      80
-------------+------------------------------           F(  1,    78) =   36.75
       Model |  2.9394e+10     1  2.9394e+10           Prob > F      =  0.0000
    Residual |  6.2388e+10    78   799846080           R-squared     =  0.3203
-------------+------------------------------           Adj R-squared =  0.3115
       Total |  9.1782e+10    79  1.1618e+09           Root MSE      =   28282
------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
       after |   46857.27   7729.537     6.06   0.000     31468.94     62245.6
       _cons |   205399.7    3563.14    57.65   0.000     198306.1    212493.4
```

they were also higher in Hounslow

Moreover the annual rate of growth of house prices (approximated by the log of the 12 month change) was not significantly different in the after period

```
. reg dlogp after if newham==1
      Source |       SS       df       MS              Number of obs =      72
-------------+------------------------------           F(  1,    70) =    0.74
       Model |  .025162236     1  .025162236           Prob > F      =  0.3931
    Residual |  2.38491217    70  .034070174           R-squared     =  0.0104
-------------+------------------------------           Adj R-squared = -0.0037
       Total |   2.4100744    71   .03394471           Root MSE      =  .18458

------------------------------------------------------------------------------
       dlogp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       after |  -.0440185   .0512209    -0.86   0.393    -.1461754    .0581385
       _cons |   .0868633   .0248889     3.49   0.001      .037224    .1365027
```

and the difference-in-difference analysis confirms that there was no differential house price growth between the two areas. It seems claims of a house price effect were exaggerated.

```
reg logp after newham afternew if newham==1 | hounslow==1
      Source |       SS       df       MS              Number of obs =     161
-------------+------------------------------           F(  3,   157) =   40.90
       Model |  3.70463956     3  1.23487985           Prob > F      =  0.0000
    Residual |  4.74020159   157  .030192367           R-squared     =  0.4387
-------------+------------------------------           Adj R-squared =  0.4280
       Total |  8.44484115   160  .052780257           Root MSE      =  .17376

------------------------------------------------------------------------------
        logp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       after |   .2172745   .0474896     4.58   0.000     .1234735    .3110755
      newham |  -.2308987   .0308383    -7.49   0.000    -.2918101   -.1699872
    afternew |   .0868757   .0671047     1.29   0.197    -.0456688    .2194202
       _cons |   12.22069   .0218916   558.24   0.000     12.17745    12.26393
```

Using Dummy Variables to capture Seasonality in Data

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

Key time series

National accounts aggregates

Last updated: 23/01/08

Seasonally adjusted

	£ million		Indices (2003 = 100)						
	At current prices		Value indices at current prices		Chained volume indices			Implied deflators[3]	
	Gross domestic product (GDP) at market prices	Gross value added (GVA) at basic prices	GDP at market prices[1]	GVA at basic prices	Gross national disposable income at market prices[2]	GDP at market prices	GVA at basic prices	GDP at market prices	GVA at basic prices
	YBHA	ABML	YBEU	YBEX	YBFP	YBEZ	CGCE	YBGB	CGBV
2002	1,055,793	937,323	94.4	94.3	97.1	97.3	97.3	97.0	97.0
2003	1,118,245	993,507	100.0	100.0	100.0	100.0	100.0	100.0	100.0
2004	1,184,296	1,051,934	105.9	105.9	103.4	103.3	103.3	102.6	102.5
2005	1,233,976	1,096,629	110.3	110.4	104.2	105.2	105.2	104.9	104.9
2006	1,303,573	1,158,871	116.6	116.6	105.8	108.2	108.3	107.7	107.7
2007						111.6	111.7		
2002 Q1	259,054	229,737	92.7	92.5	95.9	96.4	96.5	96.1	95.9
2002 Q2	262,774	233,372	94.0	94.0	96.2	97.0	96.9	96.9	97.0
2002 Q3	265,836	236,103	95.1	95.1	98.3	97.7	97.6	97.4	97.4
2002 Q4	268,129	238,111	95.9	95.9	98.2	98.2	98.1	97.7	97.7
2003 Q1	272,953	242,612	97.6	97.7	99.4	98.8	98.8	98.9	98.9
2003 Q2	277,119	246,427	99.1	99.2	98.9	99.3	99.3	99.8	99.9
2003 Q3	281,996	250,492	100.9	100.9	100.0	100.4	100.4	100.4	100.5
2003 Q4	286,177	253,976	102.4	102.3	101.7	101.5	101.6	100.9	100.7
2004 Q1	288,912	256,106	103.3	103.1	101.9	102.2	102.2	101.1	100.9
2004 Q2	295,066	262,094	105.5	105.5	103.2	103.1	103.2	102.3	102.3
2004 Q3	297,941	264,732	106.6	106.6	103.0	103.5	103.5	102.9	103.0
2004 Q4	302,377	269,002	108.2	108.3	105.4	104.1	104.2	103.9	104.0

Using Dummy Variables to capture Seasonality in Data

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

The idea is to include a set of dummy variables for each quarter (or month or day) which will then net out the average change in a variable resulting from any seasonal fluctuations

Using Dummy Variables to capture Seasonality in Data

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

The idea is to include a set of dummy variables for each quarter (or month or day) which will then net out the average change in a variable resulting from any seasonal fluctuations

$$Y_t = b_0 + b_1Q1 + b_2Q2 + b_3Q3 + b_4X + u_t$$

Using Dummy Variables to capture Seasonality in Data

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

The idea is to include a set of dummy variables for each quarter (or month or day) which will then net out the average change in a variable resulting from any seasonal fluctuations

$$Y_t = b_0 + b_1Q1 + b_2Q2 + b_3Q3 + b_4X + u_t$$

Hence the coefficient on the quarterly dummy Q1
=1 if data belong to the 1^{st} quarter of the year (Jan-Mar)
= 0 otherwise

Using Dummy Variables to capture Seasonality in Data

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

The idea is to include a set of dummy variables for each quarter (or month or day) which will then net out the average change in a variable resulting from any seasonal fluctuations

$$Y_t = b_0 + b_1Q1 + b_2Q2 + b_3Q3 + b_4X + u_t$$

Hence the coefficient on the quarterly dummy Q1
=1 if data belong to the 1^{st} quarter of the year (Jan-Mar)
= 0 otherwise

gives the level of Y in the 1^{st} quarter of the year relative to the constant (Q4 level of Y) averaged over all Q1 observations in the data set

Using Dummy Variables to capture Seasonality in Data

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

The idea is to include a set of dummy variables for each quarter (or month or day) which will then net out the average change in a variable resulting from any seasonal fluctuations

$$Y_t = b_0 + b_1Q1 + b_2Q2 + b_3Q3 + b_4X + u_t$$

Hence the coefficient on the quarterly dummy Q1
=1 if data belong to the 1st quarter of the year (Jan-Mar)
= 0 otherwise

gives the level of Y in the 1st quarter of the year relative to the constant (Q4 level of Y) averaged over all Q1 observations in the data set

Series net of seasonal effects are said to be “seasonally adjusted”

It may also be useful to model an economic series as a combination of seasonal and a trend component

It may also be useful to model an economic series as a combination of seasonal and a trend component

$$Y_t = b_0 + b_1Q1 + b_2Q2 + b_3Q3 + b_4Trend + u_t$$

It may also be useful to model an economic series as a combination of seasonal and a trend component

$$Y_t = b_0 + b_1 Q1 + b_2 Q2 + b_3 Q3 + b_4 Trend + u_t$$

where Trend = 1 in year 1

It may also be useful to model an economic series as a combination of seasonal and a trend component

$$Y_t = b_0 + b_1Q1 + b_2Q2 + b_3Q3 + b_4Trend + u_t$$

where Trend = 1 in year 1
= 2 in year 2

It may also be useful to model an economic series as a combination of seasonal and a trend component

$$Y_t = b_0 + b_1Q1 + b_2Q2 + b_3Q3 + b_4Trend + u_t$$

where Trend =1 in year 1

= 2 in year 2

= T in year T

It may also be useful to model an economic series as a combination of seasonal and a trend component

$$Y_t = b_0 + b_1Q1 + b_2Q2 + b_3Q3 + b_4Trend + u_t$$

where Trend =1 in year 1
= 2 in year 2

= T in year T

since $dY_t/dTrend = b_4$

given that the coefficient measures the unit change in y for a unit change in the trend variable

and the units of measurement in this case are years

It may also be useful to model an economic series as a combination of seasonal and a trend component

$$Y_t = b_0 + b_1Q1 + b_2Q2 + b_3Q3 + b_4Trend + u_t$$

where Trend =1 in year 1
= 2 in year 2

= T in year T

since $dY_t/dTrend = b_4$

given that the coefficient measures the unit change in y for a unit change in the trend variable

and the units of measurement in this case are years

then in the model above the trend term measures the annual change in the Y variable net of any seasonal influences

Policy, guidance and research

Aviation
Crime and public transport
Economics and appraisal
Transport evaluation
Freight
Railways
Regional and local transport
› **Road safety**
› **Strategy, targets and performance**
Road Safety Act 2006
Child road safety
Demonstration and partnership projects
Driver and rider safety
Economic assessment
Local authority guidance
Research
Speed management
Road safety partnership grant scheme
Road Safety Grant Challenge Fund Criteria
Road safety grants
Webchat with the road safety minister
Government's response to the

Tomorrow's roads: safer for everyone

› **Back to contents**

Print this page **Print all pages** **Download PDF**

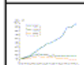
Chapter 1 - Introduction

Road accidents

1.1 Road accidents cause immense human suffering. Every year, around 3,500 people are killed on Britain's roads and 40,000 are seriously injured. In total, there are over 300,000 road casualties, in nearly 240,000 accidents, and about fifteen times that number of non-injury incidents. This represents a serious economic burden; the direct cost of road accidents involving deaths or injuries is thought to be in the region of £3billion a year.

1.2 Nevertheless, Britain has had - relatively speaking - remarkable success in reducing road casualties. And this is despite the vast growth in traffic since the beginning of the last century. In 1930 there were only 2.3 million motor vehicles in Great Britain, but over 7,000 people were killed in road accidents. Today, there are over 27 million vehicles on our roads but far fewer road deaths.

Indices of traffic and casualties: 1949-1998



1.3 In 1987 a target was set to reduce road casualties by one-third by 2000 compared with the average for 1981-85. We have more than achieved this target for reducing deaths and serious injuries. Road deaths have fallen by 39% and serious injuries by 45% and we are now one of the safest countries in Europe and indeed the world. However, there has not been any such steep decline in the number of accidents, nor in the number of slight injuries, although improvements in vehicle design have helped to reduce the severity of injuries to car occupants.

The new targets

1.4 There is no reason for us to be complacent. We know we can reduce road casualties still further. That is why we are setting a new 10-year target and launching this new road safety strategy. We need new targets to help everyone to focus on achieving a further substantial improvement in road safety over the next 10 years. By 2010 we want to achieve, compared with the average for 1994-98:

- a 40% reduction in the number of people killed or seriously injured in road accidents;

See also

› Child road safety: achieving the 2010 target - Full report (PDF 432 kb)

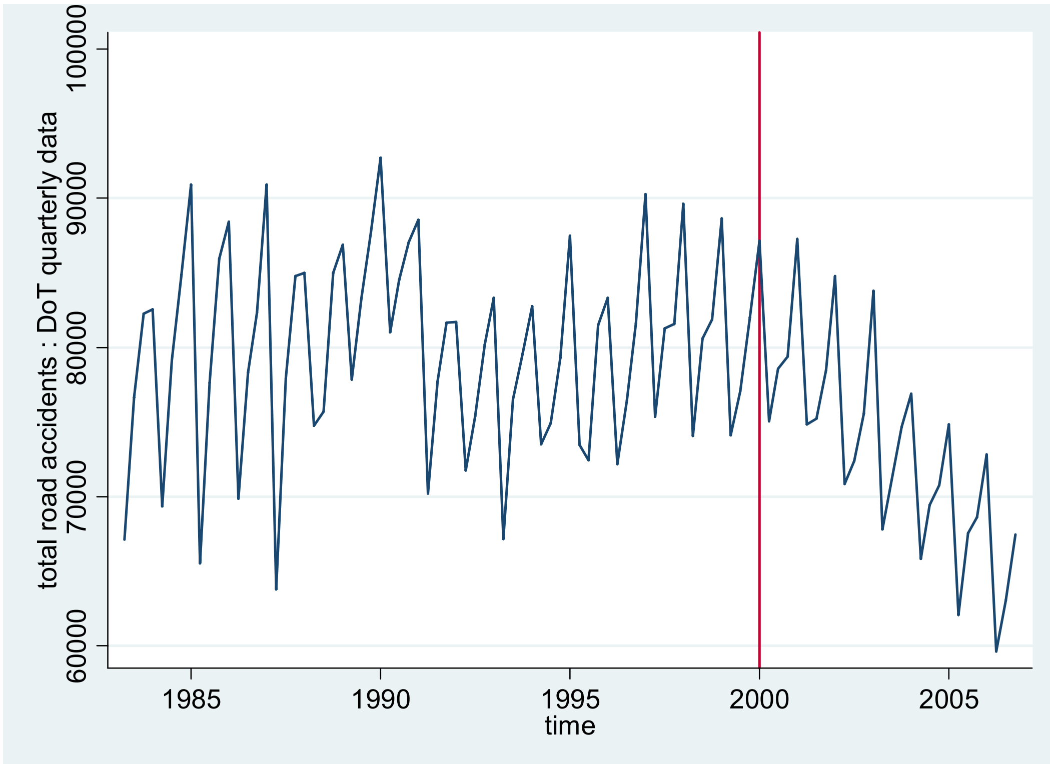
The new targets

1.4 There is no reason for us to be complacent. We know we can reduce road casualties still further. That is why we are setting a new 10-year target and launching this new road safety strategy. We need new targets to help everyone to focus on achieving a further substantial improvement in road safety over the next 10 years. By 2010 we want to achieve, compared with the average for 1994-98:

- a 40% reduction in the number of people killed or seriously injured in road accidents;
- a 50% reduction in the number of children killed or seriously injured; and
- a 10% reduction in the slight casualty rate, expressed as the number of people slightly injured per 100 million vehicle kilometres.

The data set accidents.dta (on the course web site) contains quarterly information on the number of road accidents in the UK from 1983 to 2006

```
twoway (line acc time, xline(2000) )
```



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

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 time Q1 Q2 Q3 Q4, clean

         acc   year   quart      time   Q1   Q2   Q3   Q4
  1.   67135   1983      Q1   1983.25    1    0    0    0
  2.   76622   1983      Q2    1983.5    0    1    0    0
  3.   82277   1983      Q3   1983.75    0    0    1    0
  4.   82550   1983      Q4      1984    0    0    0    1
  5.   69362   1984      Q1   1984.25    1    0    0    0
  6.   79124   1984      Q2    1984.5    0    1    0    0
```

A regression of road accident numbers on quarterly dummies (q4=winter is default given by constant term at 85249 accidents, on average in the 4th quarter) shows accidents are significantly less likely to happen outside the fourth quarter (October-December). On average there are 14,539 fewer accidents in the first quarter of the year than in the last

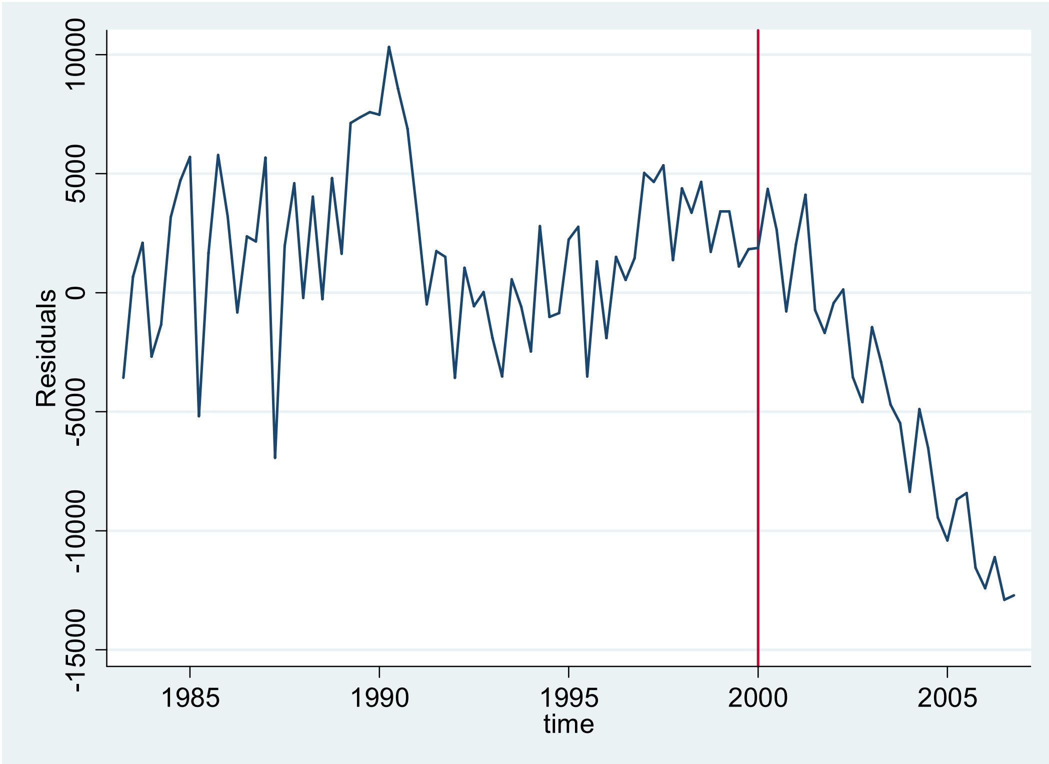
```
. reg acc Q1 Q2 Q3

      Source |       SS       df       MS              Number of obs =      95
-------------+------------------------------           F(  3,    91) =   34.16
       Model |  2.6976e+09     3   899214117           Prob > F      =  0.0000
    Residual |  2.3957e+09    91  26326242.3           R-squared     =  0.5296
-------------+------------------------------           Adj R-squared =  0.5141
       Total |  5.0933e+09    94  54184365.9           Root MSE      =  5130.9
------------------------------------------------------------------------------
         acc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Q1 |  -14539.44   1497.179    -9.71   0.000     -17513.4   -11565.48
          Q2 |  -9292.567   1497.179    -6.21   0.000    -12266.53   -6318.604
          Q3 |  -5074.609   1497.179    -3.39   0.001    -8048.572   -2100.646
       _cons |   85249.61   1069.869    79.68   0.000     83124.45    87374.77
------------------------------------------------------------------------------
```

Saving residual values after netting out the influence of the seasons is the basis for the production of **“seasonally adjusted”** data (better guide to underlying trend), used in many official government statistics.

Can get a sense of how this works with the following command after a regression

```
. predict rhat, resid
/* saves the residuals in a new variable with the name “rhat” */
```



Graph of the residuals is much smoother than the original series – it should be since much of the seasonality has been taken out by the dummy variables. The graph also shows that once seasonality accounted for, there is little evidence in a change in the number of road accidents over time until the year 2000

To model both seasonal and trend components of an economic series, simply include both seasonal dummies and a time trend in the regression model

$$Y_t = b_0 + b_1Q_1 + b_2Q_2 + b_3Q_3 + b_4TREND + u_t$$

. reg acc Q1 Q2 Q3 year

```
      Source |       SS       df       MS              Number of obs =      95
```

```
-------------+------------------------------           F(  4,    90) =   45.39
       Model |  3.4052e+09     4   851308410           Prob > F      =  0.0000
    Residual |  1.6881e+09    90  18756630.6           R-squared     =  0.6686
-------------+------------------------------           Adj R-squared =  0.6538
       Total |  5.0933e+09    94  54184365.9           Root MSE      =  4330.9
------------------------------------------------------------------------------
         acc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Q1 |  -14340.33   1264.153   -11.34   0.000    -16851.79   -11828.87
          Q2 |  -9093.455   1264.153    -7.19   0.000    -11604.92   -6581.995
          Q3 |  -4875.497   1264.153    -3.86   0.000    -7386.958   -2364.037
        year |  -398.2231   64.83547    -6.14   0.000    -527.0301   -269.4161
       _cons |   879306.5   129285.1     6.80   0.000     622459.1     1136154
------------------------------------------------------------------------------
```

Can see that there is a downward trend in road accidents (of around 400 a year over the whole sample period) net of any seasonality. Could also use dummy variable interactions to test whether this trend is stronger after 2000. How?

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              Number of obs =      71
-------------+------------------------------           F(  1,    69) =    6.19
       Model |   236050086     1   236050086           Prob > F      =  0.0153
    Residual |  2.6325e+09    69  38151620.6           R-squared     =  0.0823
-------------+------------------------------           Adj R-squared =  0.0690
       Total |  2.8685e+09    70  40978741.5           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              Number of obs =      71
-------------+------------------------------           F(  4,    66) =   47.37
       Model |  2.1275e+09     4   531865433           Prob > F      =  0.0000
    Residual |   741050172    66  11228032.9           R-squared     =  0.7417
-------------+------------------------------           Adj R-squared =  0.7260
       Total |  2.8685e+09    70  40978741.5           Root MSE      =  3350.8

------------------------------------------------------------------------------
         acc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          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.