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; North/South

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

Model is $LnW = b_0 + b_1Age + b_2Male$

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                  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 $LnW = b_0 + b_1Age + b_2female$

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 $LnW = b_0 + b_1 Age + b_2 Female*Age$

This means that slope 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 (and therefore unconnected with the slope variables) 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

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                  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 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 multicolinearity, (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              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
```

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              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 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

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

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

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

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.

$LnW_1 = a_1 + b_1\text{Treatment Dummy Variable}_1$ Period Before
$LnW_2 = a_2 + b_2\text{Treatment Dummy Variable}_2$ 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.

$LnW = a + a_2 Year_2 + b_1 \text{Treatment Dummy} + b_2 Year_2 * \text{Treatment Dummy}$

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 $Year_2=0$ and Treatment Dummy = 0, LnW = a
If $Year_2=0$ and Treatment Dummy = 1, LnW = $a + b_1$
If $Year_2=1$ and Treatment Dummy = 0, LnW = $a + a_2$
If $Year_2=0$ 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

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

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
-------------+------------------------------           F(  1, 29024) = 3041.43
       Model |  1159891.90     1  1159891.90           Prob > F      =  0.0000
    Residual |  11068703.6 29024  381.363824           R-squared     =  0.0949
-------------+------------------------------           Adj R-squared =  0.0948
       Total |  12228595.5 29025  421.312507           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
-------------+------------------------------           F(  1, 28367) = 2905.13
       Model |   969891.29     1   969891.29           Prob > F      =  0.0000
    Residual |  9470465.62 28367  333.855029           R-squared     =  0.0929
-------------+------------------------------           Adj R-squared =  0.0929
       Total |  10440356.9 28368  368.032886           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)

$= -13.1 - (-12.1) = 1.0$

$= (Hours^{LonePar}_{2000} - Hours^{LonePar}_{1998}) - (Hours^{Single}_{2000} - Hours^{Single}_{1998})$

$= (14.5-14.8) - (26.6 - 27.9) = -0.3 - (-0.7) = 1.0$

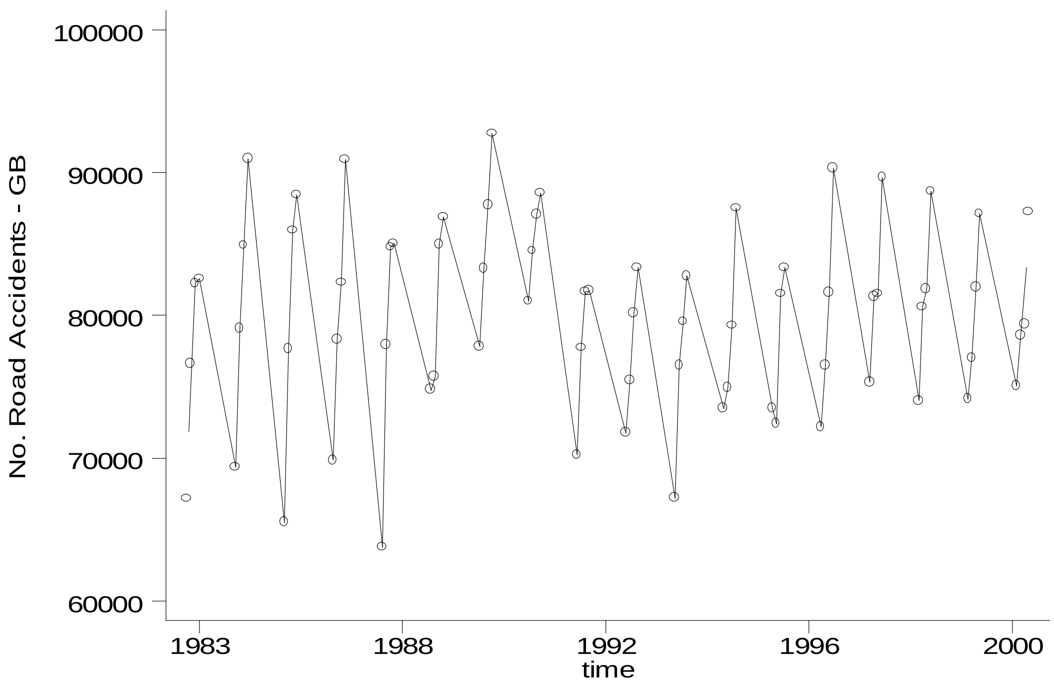
which suggests lone parents worked relatively about 1 hour more as a result of the policy. (Note that hours worked actually fall for both groups, they just fall less for lone parents).

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

```
. reg hours y2 lonep lonepy2

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

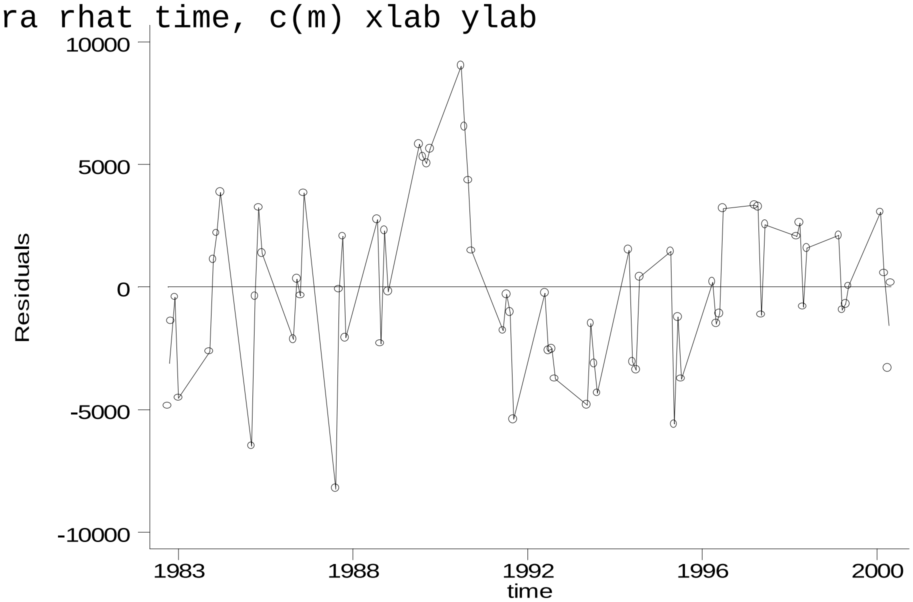
```
. reg acc q1 q2 q3
      Source |       SS       df       MS              Number of obs =      72
-------------+------------------------------           F(  3,    68) =   65.77
       Model |  2.2572e+09     3   752388623           Prob > F      =  0.0000
    Residual |   777899883    68  11439704.2           R-squared     =  0.7437
-------------+------------------------------           Adj R-squared =  0.7324
       Total |  3.0351e+09    71  42747405.0           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              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.